What's in a name? Probabilistic inference of religious community from South Asian names

Raphael Susewind

October 19, 2013

This document is the author's accepted manuscript version of this paper, incorporating any revisions agreed during the peer review process. Some differences between this version and the published version may remain. You are advised to consult the publisher's version if you wish to cite from it:

Susewind, R. (2015). What's in a name? Probabilistic inference of religious community from South Asian names. Field Methods, 27(4), 319-332.

http://dx.doi.org/10.1177/1525822X14564275

Postprint of an article forthcoming in Field Methods 27(3). Please quote publisher's version!

What's in a name? Probabilistic inference of religious community from South Asian names

Raphael Susewind

Affiliation

Raphael Susewind (mail@raphael-susewind.de) is Doctoral Candidate in Social Anthropology at Bielefeld University and Associate of the Contemporary South Asia Studies Programme at the University of Oxford.

Acknowledgments

I thank Björn Alpermann, Kurt Salentin, Neelanchan Sircar, the editor of this journal, and three anonymous reviewers for their helpful suggestions; Gilles Verniers for his dataset on Muslim legislators in Uttar Pradesh; and Santhosh Thottingal for the IndicSoundex technology. I would also like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work.

Abstract

Fine-grained data on religious communities is often considered sensitive in South Asia, and consequently remains inaccessible. Yet without such data, statistical research on communal relations and group-based inequality remains superficial, hampering the development of appropriate policy measures to prevent further social exclusion on the basis of religion. The open-source algorithm introduced in this article provides a workaround by probabilistically exploiting the communal connotations of names; it transforms name lists – which are readily available – into a new source of demographic data. The algorithm proves highly accurate in

identifying Muslim population shares in Uttar Pradesh, India's most populous state, but could arguably be employed more widely across South Asia; it potentially enables more detailed analyses in economics, development studies and political science as well as better sampling procedures in sociology and anthropology. This paper describes the algorithm, evaluates its accuracy, reflects on ethical implications, and introduces a sample dataset.

Religion-based inequalities are of growing concern in several South Asian countries, and in particular in India, where the Sachar Committee report recently revealed the extent of social exclusion suffered by the country's Muslim minority (Sachar et al. 2006). In order to contextualise, qualify, and deepen the report's findings, more fine-grained studies are urgently required, using both qualitative and quantitative methods (Basant and Shariff 2010: 2; Gayer and Jaffrelot 2012: 13). To better establish the material context of poverty, it would for instance be insightful to study landholding patterns among Muslims or the extent of their inclusion in welfare schemes, but unfortunately, neither BPL (below the poverty line) lists nor land records contain religious categorizations. To see how electoral dynamics might increase and/or divert political pressure for change, it would be useful to know the religion composition of the electorate as well as the religious identity of elected representatives – yet such figures remain unavailable. To understand how education could advance the socioeconomic condition of minorities, one would want to analyse admission pattern in schools and universities – alas, these are not tabulated by religion either; even basic literacy rates on aggregate district level are only broken down by religion since the 2001 Census. Consequently, detailed studies of the dynamics of vertical inequality across various categories of horizontal diversity frequently get derailed by data scarcity, no matter whether one wants to study them quantitatively or merely intends to identify promising field sites for qualitative research. Ultimately, the empirical basis on which policy decisions could be made thus remains fairly coarse, slowing attempts to make growth inclusive in South Asian societies.

Name lists are in contrast frequently available on higher resolution than other data; in India for instance one can easily access electoral rolls on election booth level, BPL lists on block level, land records on street level, university admission databases on departmental level, etc.

Inspired by the observation that names around the world have social and religious connotations, this paper introduces an open-source computer algorithm that attempts to infer population shares of minorities by probabilistically exploiting the religious connotations of names on such extensive lists. While software lacks access to many additional clues on which people base their assessment of others (such as dress, language, spatial context, etc.), it does possess the advantage of scalability. Given sufficiently distinct name distributions, such software is capable of probabilistically minimising errors across large datasets, resulting in reliable statistical inference. In the present context, this approach is the only feasible option to generate spatially and temporally fine-grained disaggregate statistics on religious demography across South Asia.

So what's in a name? After briefly reviewing earlier attempts to infer religious demography from name lists, this paper describes the algorithm's probabilistic logic and the name reference list used for classification, evaluates its accuracy against an independent 'gold standard', introduces a sample dataset and suggests further potential applications, while also reflecting on ethical limitations and pitfalls. On balance, I conclude that inferring religious community from South Asian names is a statistically and ethically viable second-best option for studies of religious demography as long as primary data remains inaccessible for political or practical reasons.

EXISTING TECHNIQUES

Research in the anthropology of kinship, demography, and public health has long established that 'contemporary name frequency distributions retain distinct geographic, social and ethno-

cultural patterning that can be exploited to understand population structure' (Mateos, Longley, and O'Sullivan 2011: e22943). In the South Asian context, too, names signify social structure in multiple ways: surnames often bear regional, caste or occupational connotations, given names can suggest religious community and both frequently allow educated guesses about class. While none of these significations are definite and all will vary with social and spatial context, most names are distinct enough for a reliable approximation of the religious composition of large name lists.

Some recent studies successfully used this colouring of names to enable more fine-grained analyses of social or political issues involving religion in South Asia, and in India in particular. The Sachar Committee mentioned earlier relied on name classification to identify Muslims on government payrolls and lists of beneficiaries in centrally sponsored development schemes (Sachar et al. 2006). Field et al. (2008) and Galonnier (2012) classified names on the electoral rolls of Ahmedabad and Aligarh to establish the extent of Muslims' residential segregation. Jaffrelot and Kumar (2009) explored the religious and caste background of legislators, again partly relying on name matching. Likewise, Bhalotra, Clots-Figueras, and Iyer (n.d.) used name matching to analyse the role of legislators' religious affiliation in communal violence. Most of these studies however matched names to religious community by hand, often relying on several local experts – a strategy not feasible for larger datasets.

In the wider field of computational linguistics, several research teams thus attempted to automate the matching process. Two software packages were specifically designed to classify South Asian names: Nam Pahchan (Macfarlane et al. 2007) and SANGRA (Nitsch et al. 2009). Both, however, concentrate on classifying ethnicity and output religion only as a by-

product; moreover, neither algorithm's source code is available in the public domain. South Asia is also covered by the onomap project, which identified empirical name clusters rather than using pre-defined categories (Mateos, Longley, and O'Sullivan 2011); but onomap, too, concentrates on ethnicity rather than religion, and the avoidance of pre-defined categories – theoretically convincing as it may be – limits comparisons and integration with existing data sources (particularly with government records).

A final shortcoming of all existing techniques, automated or not, is their focus on providing merely one 'best bet' for a name bearer's likely religion, without taking multiple possibilities or varying degrees of likelihood into account. In contrast, the algorithm introduced here approaches name matching as a probabilistic exercise throughout; it outputs not only the most plausible categorization, but all potential alternatives, as well as a certainty index that allows for flexible accuracy thresholds. The algorithm also specifically concentrates on religion, is capable of processing input in a range of Indian scripts as well as Latin transliteration and integrates fuzzy soundex matching technology to alleviate the impact of spelling variants as well as, to an extent, misspellings. Finally, the algorithm and a sample dataset (covering the religious demography of India's largest state, Uttar Pradesh) are available under open licenses (Susewind 2013).

REFERENCE LIST

The accuracy of name classification algorithms such as the one proposed here relies heavily on the quality of the reference list against which an individual's name parts are matched. The reference list used here has been crawled from the website indiachildnames.com, a database

which links roughly 23,000 given names to gender and the religious categories Hindu, Muslim, Sikh, Christian, Jain, Parsi and Buddhist. It also lists 4,200 surnames, but apart from 'Christian' and 'Muslim' names, these are classified by regional origin rather than religion. Such regional names were thus subsequently fed into the matrimonial website vivaah.com; if the name was found in adverts from a specific religious community at least twice as often as in those from the second most frequent community, it was categorized accordingly.

Importantly, the resulting reference list does not indicate the empirical frequency of names in different communities, merely whether they are thought to exist in a community at all. It is thus neither possible to estimate the convergence of the reference list with empirical name distributions, nor does the probabilistic logic depend on this information. Moreover, thirteen per cent of all names were listed with more than one religious connotation, with 'Jain' and 'Sikh' names being particularly ambiguous: they overlap with 'Hindu' names by 52% and 26% respectively; the algorithm might thus be less successful in distinguishing between Hindu, Sikh, and Jain populations. Geographically, the reference list covers all of South Asia, even though the algorithm arguably works best in the North Indian, Pakistani, Nepali or Bangladeshi context, where names tend to consist of a sequence of given names potentially prefixed by honorific titles and followed by family or caste names as well as designations of regional origin (Haroon 1984: 23). South Indian and Sri Lankan conventions in contrast often prioritise village names, which are not specific to any religion and are also frequently abbreviated, as are (upper) caste names – a practice which significantly reduces the textual material available for classification, and consequently reduces accuracy.

Finally, the reference list comes in Latin script, but was transliterated into a range of Indian

scripts frequently used on public name lists using Google's transliterate API. The algorithm can thus process input in Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, Telugu, as well as with limited functionality in Marathi, Nepali, Sanskrit, Sinhalese and Urdu. Moreover, it incorporates the fuzzy IndicSoundex algorithm developed by Thottingal (2009), which matches based on pronunciation rather than spelling and thus consolidates various spelling alternatives and ameliorates the impact of typographical mistakes (so that, for instance, Chowdhury and Chaudry would be considered the same name). While both transliteration and fuzzy soundex matching introduce an unknown error margin, the evaluation below indicates that the same is outweighed by the positive effects of increased coverage.

CLASSIFICATION ALGORITHM

To achieve the actual classification, the algorithm first tries to distinguish surnames from (male or female) given names, and subsequently counts how often each of those can be found in the respective subset of the reference list, noting the community suggested by each match. Should no matches be found, these restrictions are incrementally lifted by looking for presumably male names among female ones or for given names among surnames as well. Since each match is conducted once according to spelling and once according to pronunciation, this first step results in a record similar to Table 1, which lists all matches for this paper's example, the fictional 'Mohammad Ram Lal Yadav', a person of unknown gender.

[INSERT TABLE 1 ABOUT HERE]

In a second step, a certainty index for each name part / community combine is calculated and further multiplied with the percentage of unambiguous names in the respective subsets of the reference list, assuming that increased overlap in a whole class of names or matching procedure renders this entire class or procedure less reliable. If, for instance, surnames matched by spelling tend to be less ambiguous than female given names matched by pronunciation, the 'quality factor' applied to them would be higher, too. The certainty index for the suggestion that name part X signifies community Y is calculated as follows. Let spelling (X) and pronunciation (X) be the frequency of this name part across all religious communities on the reference list, matched according to spelling and pronunciation, and let spelling (X,Y) and pronunciation (X,Y) be the frequency with which spelling or pronunciation matches indicated community Y:

$$index(namepart_X \in community_Y) = \left(1 - \frac{spelling_X - spelling_{XY}}{spelling_X} \times \frac{pronunciation_X - pronunciation_{XY}}{pronunciation_X}\right) \times quality_{spelling} \times quality_{pronunciation}$$

The outcome of this second step for 'Mohammad Ram Lal Yadav' is listed in Table 2 (with all quality factors assumed to be 1 for the sake of simplicity). Importantly, the certainty index is not a straightforward probability, nor does it necessarily reflect the likelihood of a given name part / community combine in empirical name distributions. The index does, however, vary between 0 and 1 for each name part/community combine, and rises if a classification seems more trustworthy (and, in this sense, 'likely').

[INSERT TABLE 2 ABOUT HERE]

In a final step, the certainty indices of individual name parts are further aggregated to arrive at

an overall best bet and list of alternatives. This second step uses the following formula, with name parts reflecting the absolute count of name parts (in the example, this would be 4) and entries reflecting the total number of matches in the reference list with either procedure (in the example, this would be 34):

$$index(name \in community_Y) = 1 - \left(\prod_{x=1}^{nameparts} \frac{entries - index(namepart_X \in community_Y)}{entries} \right)$$

Based on this second aggregation, the algorithm guesses that 'Mohammad Ram Lal Yadav' seems to be Hindu with a certainty index of .33, which is 16 index points more likely than him or her being Parsi, 19 index points more likely than him or her being Muslim, and 30 index points more likely than him or her being Christian. As mentioned before, these certainty indices are not to be confused with straightforward probabilities, but nonetheless indicate rank. How accurately this classification reflects the religious identity of 'Mohammad Ram Lal Yadav', that is the category legally assumed for matters of personal law or for purposes such as the decennial census, is evaluated in the next section.

EVALUATION OF ACCURACY

Accuracy of name matching algorithms can principally be assessed from two angles: with respect to the algorithm's internal working – how often and how clearly does it succeed in providing a classification at all and does its probabilistic logic seem convincing? – and to an external 'gold standard' – how well does this classification reflect the real world, especially the self-categorization of individuals involved? As to the first question, the breadth of the reference list and the relatively low ambiguity of its classifications ensure that the algorithm

succeeds in classifying most names, albeit with varying certainty. Still, the extent of errors remains inestimable since the reference list does not reflect empirical name distributions. More important, therefore, is the evaluation against an external 'gold standard'.

How well the algorithm reflects such a standard depends on both the specific list of names to be classified and on the religious categories of interest. The following remarks are limited to one exemplary context drawn from my own field of research; they only evaluate how well the algorithm categorizes Muslim vs. non-Muslim names in Uttar Pradesh, India's largest state. This example seems appropriate for three reasons. Firstly, Muslims are the state's poorest religious minority, so that understanding social exclusion better is arguably most urgent in their case (Sachar et al. 2006). Secondly, Muslim names in North India are also linguistically more distinct than those of other religious communities for historical reasons, making name matching a particularly promising technology (Christian names are similarly distinct, but frequently adopted as 'modern' or 'Western' names across the religious spectrum and thus no reliable indicator of religious demography). Finally, Muslims were chosen for the pragmatic reason that an independent test corpus of Muslim and non-Muslim names from Uttar Pradesh was readily available.

This test corpus consists of the names of 10249 Haj pilgrims from Uttar Pradesh in 2012, the names of 2305 undergraduates admitted to Lucknow University through the scheduled caste quota in the same year and the names of the 6752 candidates who contested Uttar Pradesh's state elections since independence. Since non-Muslims are legally barred from the Haj, and Muslims legally barred from claiming scheduled caste benefits, the first two lists are mutually exclusive, while the third was manually categorized by Gilles Verniers (Jaffrelot and Verniers

2012). In addition to a Muslim / non-Muslim classification, the first two lists contain both a person's own as well as their father's (or male relative's) names, while the third provides names and gender. Since the test corpus consists of roughly the same number of Muslim and non-Muslim names, the latter were randomly duplicated until the overall ratio reflected the religious demography of Uttar Pradesh; this does not affect sensitivity and specificity, but renders more meaningful positive and negative predictive values.

Accepting this test corpus as an external 'gold standard', the algorithm demonstrated a sensitivity (i.e. a rate of 'true' Muslims classified correctly) of 96% and a specificity (i.e. a rate of 'true' non-Muslims classified correctly) of 99%. The algorithm's positive predictive value (i.e. the rate of 'true' Muslims among all those identified as Muslims) stood at 95%, its negative predictive value (i.e. the rate of 'true' non-Muslims among all those so identified) at 99%. Overall, 5% of names could not be classified and were discarded. Compared to other name matching technologies reviewed by Mateos (2007), the algorithm proved to be very capable of identifying Muslim population shares in this test corpus, which likely reflects both the distinctness of empirical name pattern among Muslims and non-Muslims in Uttar Pradesh and intrinsic advantages of the algorithm over earlier technologies, such as the fuzzy soundex logic.

One could potentially refine the outcome further by excluding unclear matches that either fall below an absolute threshold in the certainty index or whose certainty indices do not differ enough from the second most likely categorization. Such thresholds will always be a trade-off between coverage and accuracy; in this specific example, they did not markedly improve accuracy, but significantly reduced coverage, so that even those classifications for which the

best bet seemed only marginally more likely than the second-best alternative were taken into account. In other contexts and/or to differentiate between other religious categories, flexible thresholds might prove useful and might constitute a major advantage of this algorithm over earlier ones.

In order to reasonably assume a similarly high accuracy in categorizing a wider population, it is important to consider how the names in the test corpus might differ from those in this wider population. Those Muslims going on the Haj are for instance arguably older, as are candidates in state elections – while university students are younger. Students admitted under the SC quota also tend to come from lower economic strata – a bias arguably offset by an elite bias in the list of candidates. Finally, the kind of names found in the test corpus might differ in clarity from the average 'Muslim' and 'non-Muslim' name. One could argue that those going on the Haj tend to come from more pious backgrounds and might thus carry less ambiguous names; they might also hail from the richer sections of society, which among North Indian Muslims often means an *Ashraf* origin discernible in names drawn from Arabic or Persian heritage. With the SC admission lists, one could argue the opposite: individuals on these lists tend to be poorer and less educated; their names might thus on average be less Sanskrit, Arabic, or Persian in origin than that of an 'average' non-Muslim. In the absence of empirical studies on naming practices in North India, it is hard to conclusively assess these potential distortions, but even with a generous margin of error, the algorithm's accuracy would remain pretty high.

ETHICAL IMPLICATIONS

The mere fact that a new methodology is technically feasible need not imply that it is

normatively unproblematic. This begins with the underlying notion of a stable communal identity (Mateos 2011). Names are usually given by parents, and need not reflect much more than that: the experience and (inter-)subjective meaning of 'being Muslim' (or Hindu, Christian, Parsi, Sikh, Buddhist or Jain) goes far beyond the respective legal category. This problem is not specific to the methodology suggested here, however, but is similarly problematic in the context of other datasets, such as the Indian Census. While it remains worth emphasizing that there can never exist a wholly accurate method of coding religion based on names because neither religious names nor religious categories exist as objective, bounded entities, religious categories do exist as statistical approximations. By explicitly taking a probabilistic stance, this algorithm tries to circumvent some of the ethical problems scholars have rightly identified in endeavours such as the ongoing Indian Caste Census. Often, the resulting data, simplified as their categories may be, should be better than the alternative: having no data at all or only insufficiently disaggregated sets.

More problematically, however, names have not only been used to discriminate (literally, distinguish) people, but also to discriminate *against* them, for instance on the housing market. More dramatically, rioters involved in major bouts of Hindu-Muslim violence are known to have used name lists to identify their targets. Unfortunately, the ethical implications of this and similar 'big data' methodologies vis-a-vis such absue are far from straightforward and the process of formulating ethical guidelines is still very much in its infancy (Boyd and Crawford 2012: 672). On the one hand, all name lists mentioned in the introduction – electoral rolls, admission lists, property records or BPL lists – are in the public domain, as is knowledge about religious connotations of names. On the other hand, the algorithm makes this existing data and knowledge significantly more accessible by automating the matching process,

effectively generating data otherwise considered too sensitive to be published, and doing so based on name lists not originally intended for this purpose. While accessibility is often valuable and desirable – not least to better understand religion-based discrimination and inequality – the new analytical possibilities this opens up have to be weighed against the risks that might go with them.

These risks differ with context and scale. In the Indian context, for instance, discrimination as well as violence against religious groups tends to be orchestrated on the local level; although parts of the state machinery have often been implicated for its slow and indecisive interventions, it is hard to see systematic, sustained, and large-scale state-led exclusion on the basis of religion. On the local level, however, the algorithm arguably makes the least difference: if rioters wish to identify members of a particular community in a few urban neighbourhoods, they can easily match the electoral register by hand (and have done so in the past); likewise, if property owners wish not to rent to members of a particular community, they do not need a statistical tool to identify those community members. On the other hand, the algorithm plays out its computational strengths on larger scales of application, and it is on these scales that more fine-grained analyses of religion-based social exclusion might ultimately allow the state to better combat the same. Whether such political action will materialize is beyond the scope of this paper, but without appropriate data, good policy design, implementation, and evaluation would certainly remain difficult.

Overall, the ethical benefits from employing name matching technology thus arguably outweigh the risks in the South Asian context – even though a final ethical assessment will have to be made in the context of specific research applications.

POTENTIAL APPLICATIONS AND SAMPLE DATASET

This paper began by noting the scarcity of data on religious demography in South Asia, and in India in particular, and suggested to exploit the connotations of names as a probabilistic workaround. I presented a computer algorithm that automates this process and evaluated its accuracy with respect to one particularly relevant example: the identification of Muslim names in North India. The algorithm is not, however, limited to this task. In which development blocks of Pakistani Punjab are Sikhs under-represented in the BPL lists? Where did Yadavs in Bihar *not* vote for the Samajwadi Party? How strong is religion-based residential segregation in rural vs. urban areas of Bangladesh? These are just some of the questions that could potentially be explored with more fine-grained demographic data. The algorithm presented here allows to generate such data.

As a proof of concept and to allow colleagues to experiment with this approach, I recently compiled a comprehensive dataset on religion and politics in Uttar Pradesh (Susewind 2013). The dataset combines election results, GIS data, and Census records with an estimate of religious demography derived from the algorithm introduced in this paper. The smallest unit of measurement is that of a polling booth, which serves roughly a 300 m radius in urban areas or a small village in rural ones. Like the algorithm itself, this dataset is available under an open share-alike license and continuously updated.

Names are important signifiers of social categorization, which in turn build the basis for many sociological studies. As long as official data on religion remain inaccessible for political or

practical reasons in South Asia, inferring religious demography from the colouring of names might thus remain the only feasible way of making this basis more fine-grained and reliable. This paper has shown how such an approach can remain attentive to degrees of probability while nonetheless rendering fairly accurate approximations.

REFERENCES

Basant, Rakesh and Abulsaleh Shariff. 2010. "The state of Muslims in India: An overview." Pp. 1-23 in *Handbook of Muslims in India*, edited by R. Basant and A. Shariff. Oxford: Oxford University Press.

Bhalotra, Sonia, Irma Clots-Figueras and Lakshmi Iyer. N.d. Politician identity and religious conflict in India. Available from http://www.isid.ac.in/~pu/seminar/20_04_2012_Paper1.pdf.

Boyd, Danah and Kate Crawford. 2012. "Critical question for Big Data." *Information, Communication & Society* 15(5): 662-679

Field, Erica, Matthew Levinson, Rohini Pande and Sujata Visaria. 2008. "Segregation, rent control, and riots: The economics of religious conflict in an Indian city." *The American Economic Review* 98(2): 505-510

Galonnier, Juliette. 2012. "Aligarh: Sir Syed Nagar and Shah Jamal, contrasted tales of a 'Muslim' city." Pp. 129-158 in *Muslims in Indian cities: Trajectories of marginalisation*, edited by L. Gayer and C. Jaffrelot. London: Hurst.

Gayer, Laurent and Christophe Jaffrelot. 2012. "Introduction: Muslims of the Indian city. From centrality to marginality." Pp. 1-22 in *Muslims in Indian cities: Trajectories of marginalisation*, edited by L. Gayer and C. Jaffrelot. London: Hurst.

Jaffrelot, Christophe and Gilles Verniers. 2012. "Castes, communities and parties in Uttar Pradesh." *Economic & Political Weekly* 47(32): 89-93.

Jaffrelot, Christophe and Sanjay Kumar, ed. 2009. *Rise of the plebeians? The changing face of the Indian legislative assemblies*. New Delhi: Routledge.

Haroon, Mohammed. 1984. *Cataloguing of Indian Muslim names*. Delhi: Indian Bibliographies Bureau.

Macfarlane, Gary J., Mark Lunt, Benedict Palmer, Cara Afzal, Alan J. Silman and Aneez Esmail. 2007. "Determining aspects of ethnicity amongst persons of South Asian origin: The use of a surname-classification programme (Nam Pehchan)." *Public Health* 121(3): 231–236.

Mateos, Pablo. 2007. "A review of name-based ethnicity classification methods and their potential in population studies." *Population, Space and Place* 13(4): 243–263.

Mateos, Pablo. 2011. "Uncertain segregation: The challenge of defining and measuring ethnicity in segregation studies." *Built Environment* 37(2): 226-238.

Mateos, Pablo, Paul A. Longley and David O'Sullivan. 2011. "Ethnicity and population structure in personal naming networks." *PloS one* 6(9): 1–12.

Nitsch, D., L. Kadalayil, P. Mangtani, R. Steenkamp, D. Ansell, C. Tomson, I. Dos Santos Silva and P. Roderick. 2009. "Validation and utility of a computerized South Asian names and group recognition algorithm in ascertaining South Asian ethnicity in the national renal registry." *QJM* 102(12), 865–72.

Sachar, Rajindar, Saiyid Hamid, T. K. Oommen, M. A. Basith, Rakesh Basant, Akhtar Majeed and Abusaleh Shariff. 2006. Social, Economic and Educational Status of the Muslim Community of India. New Delhi: Government of India.

Susewind, Raphael. 2013. Data on religion and politics in Uttar Pradesh. Available from http://data.raphael-susewind.de under an Open Database License (ODbL)

Thottingal, Santhosh. 2009. Swathanthra Indian Language Computing Project (Indic

Soundex). Available from http://smc.org.in/silpa/Soundex under a GNU Affero General Public License

Table 1. Number of matches in the reference list, broken down by religious connotation

	Mohammed	Ram	Lal	Yadav
by spelling	Muslim: 1x	Hindu: 1x Parsi: 1x	Hindu: 1x Parsi: 1x	Hindu: 9x
by pronunciation	Muslim: 3x	Hindu: 3x Parsi: 1x Christian: 1x	Hindu: 2x Parsi: 1x	Hindu: 9x

Table 2. Certainty indices for each name part/community combine

Mohammed	Ram	Lal	Yadav
Muslim: .100	Hindu: .80 Parsi: .60 Christian: .20	Hindu: .83 Parsi: .67	Hindu: .100