Studying Indian politics with large-scale data: Indian Election data 1961–today

This is a post-print of the following article:

Jensenius, Francesca R. and Gilles Verniers (2017). *Studies in Indian Politics,* 5(2), 269–275.

Francesca R. Jensenius[1]
Gilles Verniers[2]

There has been a data revolution in the study of Indian politics in recent years. Indian politics has always been a fascinating field, in comparative perspective and in its own right. Accounts of fieldwork in India have challenged and informed discussions about democracy in the developing world for decades. Since the 1990s, the National Election Studies conducted by Lokniti, CSDS, have enabled informed quantitative analyses of political patterns across the country. More recently, technological innovations and the increasing availability of online data sources have opened the door to new ways of studying Indian politics quantitatively. Scholars can now create and merge a wide range of large-scale datasets, making it possible to establish new empirical trends, fact-check commonly held political narratives, and test hypotheses developed in other democratic contexts.

In this research note, we first describe some techniques and tools used for creating and merging large-scale datasets. Next, we introduce two datasets we have developed: constituency-level datasets of Indian State Elections and General Elections from 1961 until today.[3] We describe the process of creating these datasets, the efforts involved in cleaning the data, and how the data can be utilized. In conclusion, we offer some reflections on the limitations of over-relying on quantitative data in research on Indian politics. We hope to get more scholars and practitioners interested in using the publicly available datasets developed by ourselves and others, and to inspire students and scholars to invest in the quantitative skills needed to develop new quantitative datasets.

Data sources for the study of Indian politics

India has some extraordinary data sources for studying politics and society. The decennial censuses provide large amounts of information about villages and towns, which can be aggregated to higher administrative levels. Several large surveys, including the Indian Sample Surveys and the Indian National Election Studies, are internationally acclaimed for their innovative design and sizable samples. Most of India's government agencies also collect impressive amounts of information about their work. Of particular importance for political scientists is that the Election Commission of India

1 Senior Research Fellow, Norwegian Institute of International Affairs (NUPI), Oslo, Norway.

2 Assistant Professor of Political Science and Co-Director, Trivedi Centre for Political Data, Ashoka University, Rajiv Gandhi Education City, Sonipat, Haryana.

3 We wish to acknowledge Dr. Sudheendra Hangal's contribution towards restructuring the dataset, and wish to thank research assistants at UC Berkeley as well as the research team at TCPD, Ashoka University, who helped with the tedious job of cleaning these data.

publishes detailed information about all elections in India, and, since the early 2000s, on the socio-economic backgrounds of election candidates.[4]

However, scholars find it difficult to use some of these data sources for statistical analysis. Many sources are hard to come by, as they have been recorded and kept by single offices at the state – or district – level. Census data prior to 1991 were published in book format, not in soft copy. Historical election results are published as PDFs. Furthermore, public data in India are often available in inconsistent or unpractical formats, making it cumbersome (or near-impossible) to merge them or work with them. The sheer scale of the data, and the time and effort required to digitize data manually from handwritten notes, books, PDFs and scanned documents – or simply to access data sources – has deterred many scholars from working with quantitative data, or has made them work with small sub-samples of data.[5]

This is now changing. With the spread of the Internet, increased storage capacity, and a focus on transparency and data sharing, more and more Indian government agencies and private organizations are putting large amounts of data online. These data are generally intended to inform Indian citizens about the work and activities of their elected politicians and bureaucracy – but they also provide a goldmine of new information for students of Indian politics and society.

However, much of this data is not readily usable or immediately inter-compatible. The data must be entered, reformatted, cleaned, and streamlined so that they can 'speak' with one another. Here are six tools particularly useful for digitizing and merging large-scale data:

1. **OCR** (Optical Character Recognition) makes it possible to digitize data from books or images. OCR software recognizes letters and numbers and creates a digital version of them. Recognition is not always perfect, especially with older text sources and poorly scanned documents, so further manual cleaning and verification may be required, but using OCR software can massively reduce the time and cost of entering large amounts of data. This has been used recently to digitize sub-district level data from the 1931 and 1971 Indian Censuses.[6] Many OCR software products are available, but few of them can handle many pages of data simultaneously. For large data materials, it can make sense to outsource the OCR job to companies that specialize in using such technologies.

2. **Web scraping** involves harvesting or extracting information from websites. With applications like *SiteSucker*, online information can be automatically downloaded, without requiring programming skills. However, standardized applications cannot deal with all websites. For complex websites, for example those with drop-down menus that must be completed to access data, some programming might be needed. Programming languages like *Python* can generate inputs for drop-down menus – making it possible to download large amounts of information very rapidly.

3. **PDF conversion:** Even if they can be downloaded from the Internet, many data sources are locked inside PDFs. Some PDFs can readily be converted into tables or text files using programs like *Adobe Acrobat Pro*, or online PDF converters, but those usually only allow for the conversion of one page or one PDF file at a time. To convert several PDFs simultaneously, it is better to employ

4 See Vaishnav (2017) for more information about these data.

5 See Jensenius (2014) for reflections on how to collect quantitative data through fieldwork.

6 The data were copied from books digitized using OCR. For further information about the 1931 data, see Suryanarayan (2016); for the 1971 data, see Bhavnani and Jensenius (2015).

some form of programming. For instance, it is simple to convert large numbers of PDFs using the utility *pdftohtml*. Explanations of how to do this can be found online.

4. **Parsing:** The end-product of downloading online data or converting a PDF is usually a messy file that must be reordered and cleaned in order to look like a normal dataset. The term for reordering the data is *parsing,* which means breaking data into smaller, more manageable chunks. In the election data described below, the PDF for each election was converted to html. We then extracted from those htmls whatever information we wanted. For example, we searched for the term 'constituency' to identify where new constituencies started, the term 'total voters' to find out where in the file we had voting data, and so on.

5. **GIS mapping** – Geographic Information System (computer-generated map files) – is another rapidly developing technology. GIS maps are usually organized in points or polygons, and can represent administrative or electoral boundaries, towns, or geographic phenomena in map form. Such maps are extremely useful for illustrating patterns like election results, but are also excellent for merging datasets, as GIS programs allow users to lay one map on top of another – for instance, to show which electoral constituency overlaps with an administrative sub-district. Industry-standard GIS programs are prohibitively expensive, but *QGIS* is a free, user-friendly alternative. One can also work with GIS data and maps in statistical programs like *Stata* and *R.*[7]

6. **Fuzzy name matching** is another useful tool for merging data. Traditionally, data are merged by using a unique identifier across two datasets. For example, in Indian census data, the 16-digit census code is a unique identifier for each geographical unit in the data. However, data often do not have a unique identifier because they were not collected with the intention of being merged with other data sources. They may have a name in common, for example a town name, but since such names are often transliterated from local languages they end up being spelled differently across datasets. One way of dealing with this problem is to create merging scripts identifying the names in each datasets that have as many shared characters as possible, or that start with the same letter. Algorithms can then identify matching names with varying degrees of differentiation. This makes it possible to find perfect matches, most-probable matches, or at least reduce the manual work to choosing between a few alternatives.[8]

In addition to these techniques, considerable thinking and cleaning often go into creating a dataset. We show this below, describing how we created two datasets covering all Indian elections.

Creating datasets on Indian elections

After every general and state election in India, the Election Commission of India (ECI) compiles constituency-wise reports of the results, and the data are made available online in PDF documents. Each of these PDFs were downloaded, converted from PDFs to html files, and then parsed using the statistical computing environment R (R Core Team 2013). The result was a state assembly constituency-level dataset of political variables for all Indian state elections held between 1961 and

7 There are also many R packages developed for GIS visualization, such as *maps* and *Shiny*.

8 Bhavnani (2009) and Ashed and Novosad (2017) have made publicly available tools they have developed to 'fuzzy match' South Asian names. Raphael Susewind (2015) has an algorithm allowing for the probabilistic inference of religious community from names. The Trivedi Centre for Political Data (TCPD) has also developed an algorithm for matching names and generating automatic unique IDs (beta version at http://tcpd.ashoka.edu.in:8080/surf/).

today, plus a parliamentary constituency-level dataset of the national elections during the same period.[9]

The portion of the reports included in our datasets is the 'Detailed Results' at the end of each report, with the name, party, sex, and number of votes for each candidate running for election, as well as information about each constituency, with the size of the electorate and the total number of voters and whether a constituency was reserved for SCs or STs. We also included data about the gender distribution of the electorate, for the years and states for which such data are available. The two datasets are now available at website of the Trivedi Centre for Political Data (http://lokdhaba.ashoka.edu.in/LokDhaba-Shiny/) and will be updated as new elections take place. The dataset is accompanied by detailed documentation of how the original ECI data were cleaned and what variables are included, but here is a summary.

Cleaning the data
Irregularities in the formatting made it difficult to re-create the data perfectly using only programming. These were some of the most important steps taken to clean the data:[10]

- checking that the 'Number of candidates' listed in the report corresponded with the actual number of candidates in the data
- checking that there were more electors than voters in the data
- checking that the votes for all candidates added up to the 'Total votes' in the report
- checking that all percentages of votes added up 100
- harmonizing the spellings of constituency names and candidate names to make it easier to identify the names over time (including standardizing titles such as Doctor, Advocate or Engineer in the candidates' names).
- correcting the labelling of candidates as 'Male'/'Female' in case of obvious error
- correcting the reservation status of constituencies in case of obvious error.

There are some remaining mistakes in the data due to gaps or errors in the original PDF files. Also, the names of constituencies and candidates are often spelt differently due to differences in spelling across the reports, and some problems may have escaped our attention. We will continue seeking to make the data as clean possible.

The problem with 1950s data
On the ECI website there are PDFs for elections in the 1950s. We downloaded and cleaned these files, but have not included them in the dataset. The reason is that in the 1950s India's reserved seats were organized as double-member constituencies (and some triple-member constituencies). That means that in some constituencies both a general-category candidate and an SC or ST candidate were elected. The election reports list all candidates, but we cannot tell who were general-category candidates and who were reserved-category candidates. The result is electoral turnouts higher than 100% and other illogical metrics that make it hard to use these data. Starting from 1961, all politicians were elected in single-member constituencies.

---

9 This work was originally initiated in order to study the long-term effects of electoral quotas in India (see Jensenius 2017), and was later continued as a collaboration between Jensenius and Verniers at TCPD. At TCPD we would particularly like to acknowledge the work of Sudheendra Hangal, Chinmay Narayan and Rajkamal Singh.

10 We are deeply grateful to research assistants at UC Berkeley and at Ashoka University who helped with the tedious job of cleaning these data.

Creating new variables
The election data are useful as such, but we have added several variables relevant for discussions of politics:

- Turnout: 100*voters/electorate
- Margin: Vote share of a candidate minus the vote share of the next candidate in the position order (this includes the *Margin of Victory*; the difference between the vote share of the winner and the runner up)
- Effective Number of Candidates: a measure of how fractionalized the vote was, using the formula developed by Laakso and Taagepera (1979)
- Administrative districts (for elections after 2007)
- Parliamentary constituencies that state assemblies are part of
- Sub-regions in large states

Using the data
The state assembly election data and the general election data are available at the Trivedi Centre for Political Data's website.[11] Here, data can be downloaded for statistical analyses or can be summarized and visualized through the **Lok Dhaba interface**, a web tool built in R that enables users to download the raw data or to build visualizations of the data as charts or maps.

Merging the data with other datasets
In the election data, single constituencies can be uniquely identified using the state code and the constituency code. The state codes and district codes follow the Indian Census of 2011. The constituency codes follow the delimitation used at the time of each election. Since the constituency codes change from one delimitation to the next, users should take care when creating lagged variables or merging the data over time, to ensure that the constituencies that are used are actually the same.

The election data can easily be linked to any dataset that has constituency identifiers, such as the National Election Studies (see Heath, Verniers and Kumar, 2015), or the Indian Legislators Dataset.[12] In the latter case, information about candidates and legislators can be merged with information about their electoral performance and information on their constituencies.

The state assembly data can also be linked to administrative districts by aggregating variables, since state assembly constituencies fit neatly into administrative districts. Linking the election data to lower-level administrative units, such as sub-districts or villages, can be done manually or by using GIS maps.[13]

11 See http://lokdhaba.ashoka.edu.in/LokDhaba-Shiny/. The general election data also form part of an international database of parliamentary elections maintained by the Centre for Political Studies, Institute for Social Research, University of Michigan: http://www.electiondataarchive.org/

12 The Indian Legislators Dataset is a database containing information about the sociological profile of elected representatives in the Lok Sabha and in the State Assemblies, such as sex, age, caste, religion, occupation and other biographical information about MPs and MLAs across time. It is hosted by the Trivedi Centre for Political Data and is still a work in progress.

13 See Bhavnani and Jensenius (2015) for a discussion on how to link census data and pre-delimitation state assembly constituencies.

Conclusions

Integrating constituency-level data with other forms of data opens several research avenues that we hope will shape the empirical research on Indian politics. The data described in this research note allow us to examine associations between socio-economic or demographic data and political phenomena like turnout, party performance and voter preferences – issues often studied mainly through case studies (see, e.g., Banerjee, 2014, p. 20). They also enable us to explore patterns at many levels of observation – assembly constituency, parliamentary constituency, district, state, and the all-India level.

New technologies may increase efficacy and measurement reliability, but they are not substitutes for more traditional forms of dealing with data. They can help in handling large-scale datasets that are unapproachable through manual data-entry methods. But they always require manual verification and data cleaning. Looking at raw data remains an important way to learn what these data have to say and what their limitations are. Manual verification makes it possible to spot recurrent issues, unusual values, or outliers that may require more qualitative investigation. It is also essential to keep in mind that the value of any dataset fundamentally comes down to the quality of each data-point. If measures are poorly conceptualized, operationalized, or measured, simply having more does not help much.

Even high-quality data may not always hold the answers to the research questions at stake. Many questions are best approached with observations drawn from historical sources, interviews, or experiments. In our view, political science research should ideally draw on both qualitative and quantitative data. Large-scale quantitative data can serve as a starting point for a more qualitative inquiry or a testing ground for hypotheses derived from such work. Either way, by making our datasets publicly available, we hope to provide a large audience of students and scholars with basic empirical information for their own research.

Sources

Asher, Sam, and Paul Novosad (2017). Politics and local economic growth: Evidence from India. *American Economic Journal: Applied Economics*, *9*(1), 229–273.

Banerjee, M. (2014). *Why India votes?* New Delhi: Routledge.

Bhavnani, R. R. (2009). Do electoral quotas work after they are withdrawn? Evidence from a natural experiment in India. *American Political Science Review*, *103*(1), 23–35.

Bhavnani, R. and F.R. Jensenius (2015). Socioeconomic profile of India's old electoral constituencies, 1971–2001. In M. Sanjeer Alam and K. C. Sivaramakrishnan (Eds), *Fixing electoral boundaries in India: Laws, processes, outcomes and implication for political representation*. M. Sanjeer Alam and K. C. Sivaramakrishnan New Delhi: Oxford University Press.

Heath, O., Verniers, G. & Kumar, S. (2015). Do Muslim voters prefer Muslim candidates? Co-religiosity and voting behaviour in India. *Electoral Studies* 38, 10–18.

Jensenius, F.R. (2014). The fieldwork of quantitative data collection. *PS: Political Science & Politics*, *47*(2), 402–404.

Jensenius, F.R. (2017). *Social justice through inclusion: Consequences of electoral quotas in India.* New York: Oxford University Press.

Laakso, M. & Taagepera, R. (1979). 'Effective' number of parties: a measure with application to West Europe. *Comparative Political Studies 12*(1), 3–27.

R Core Team (2013). *R: A language and environment for statistical computing*. http://www.R-project.org/.

Suryanarayan, P. (2016). *Hollowing out the state: Status inequality and scale capacity in colonial India*. Ph.D. thesis, Columbia University, New York.

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

Vaishnav, M. (2017). *When crime pays: Money and muscle in Indian politics*. New Haven, CT: Yale University Press.