Targeting Indian Census 2021: Systematic Approaches for Reliable and Efficient Data Collection

Amandeep Rathee*

Academics and Research - Data Science | Machine Learning and AI, UpGrad

Abstract

India is the second largest country in population and seventh largest in size. The Indian government conducts decennial census [1] which aims to keep a record of its demographics and other information important for the betterment of the country. The projected population for the next census is 1.4 billion [2] . The census data is used by the Indian government to provide better facilities in areas such as agriculture, infrastructure, education, healthcare, transportation, information technology and several other domains. However, as far as technology is concerned, a lot has changed since the last census held in 2011. India has experienced a boom in big data technologies and data science in general. And these technologies will only grow in the upcoming years. The world is becoming more data-driven, and so is India. Therefore, if the country has to make use of the next census data which will be collected in 2021, it needs to focus and improve the current practices involved in the data collection. The last census which was held in 2011 cost the government a whopping 2200 crore INR (USD 310 million) and several months to collect such humongous data. In this paper, I propose various techniques for the data collection personnel - researchers, teachers and lecturers in government institutions and other government employees who are involved in collection of the data during the census - to collect the data in the next census in a more efficient and reliable manner such that the data science practitioners don't face the data quality issues while analyzing it. The current technology such as cloud storage, fast and efficient computers, IoT devices, along with some basic knowledge of stats and data storage will help the data collection personnel to collect it in a more efficient and reliable way. The ultimate goal of this proposal is to make sure that the data science community doesn't face any problems while working on the census data which could eventually be used for the betterment of the country.

*E-mail: arathee2@gmail.com

1 Introduction

Data collection tasks like for census data requires intensive use of human resource. The data has to be collected across the country which has different states, varying color, creed, caste, religion and other factors. There is a huge diversity in the mindset of people along with their tradition and languages. If the data is collected across one region, there can be multiple sources of error and the collection personnel need to deal with the diversity of the country and make sure that it doesn't impede the data collection process. It becomes really difficult to collect data in such a diverse country with a uniform set of practices. Most of the people who collect data are either not trained properly or don't pay attention to issues that might prove detrimental to the researchers that use the census data. When data collection agents, who are commonly government employees such as lecturers or school teachers, go on the field to collect data, they face certain issues such as unreliable data collection equipment, skeptical mindset of people to share information, collecting incorrect information, dealing with issues pertinent to a particular region, etc.

2 Issues With Current Practices

There are various issues with the current data collection practices adopted by the government. These are listed below.

2.1 Physical Forms

The data is collected on physical application forms. There are three forms which need to be filled. The problem with physical forms, in this day and age, is that these forms are really slow to process and are unreliable to store information. The data collected in the census takes about 2 months. However, this data takes 4-6 months just to send to the data collection center for collection. Moreover, data in physical application forms gets destroyed or lost which leads to a wastage of time and money for the government.

2.2 Incorrect Information

People are really skeptical to share information asked in the census data. Information such as income source and income amount are the ones that people are most reluctant to share. People either don't share this kind of information or they provide incorrect information. While analyzing the data, this false information leads to misleading insights.

Another problem is when people are not present at their homes. Data is collected at the household level, that is, the government agent visits each household and asks for the information about the family from the head of the family or the person who is present at home. If a family is not at home, the family's neighbors are inquired to gather the data about them which most of the times turns out inaccurate in one or the other aspects.

2.3 Duplicate Information

Data is collected at the household level, that is, the government agent visits each household and asks for the information about the family from the head of the family or the person who is present at home. This type of information collection practice can lead to duplicate data. Consider this case - a family which is living in a city. A member of the family, being a student, has moved to another city to study. While collecting the data, the family head will provide the information of the student as well along with the other members of the family present at home. At the same time, the student can also provide information about him/her at the place where he/she is currently residing. The data related to the student is collected two times. And there is no kind of identification required while providing the information about an individual which leaves no room for identification of duplicate information.

2.4 Type of Data

The type of data that's collected needs to be updated. Currently, only demographic data is collected. If the government wants to use the census data for the growth of the country, it needs to ask for other types of data which can help in the betterment of the country. Moreover, the data is collected doesn't proper data types. For example, dates are collected as strings, not as a date object. The state and the district while recording the address are also collected as strings, not as categorical data.

3 Proposed Solution

The issues listed above can be addressed if we follow the practices mentioned in the text that follows.

3.1 Digital Collection of Data

India has already started working to become a digitally sound nation by introducing the 'Digital India' campaign [3]. To collect the data efficiently, the census needs to adopt the usage of electronic devices such as smart tablets or laptops. The most important advantage of this technique is that it will reduce the error and the effort both. Once the data is collected digitally, we need not put the efforts in converting the handwritten data on the application forms to digital form - a process which is currently used and takes weeks to complete. In addition, for converting handwritten data to digital form, the government uses the state-of-the-art technology OCR machines [4] but even these are not 100% accurate, and they have a substantial chance of error. In addition, collecting the data digitally will drive the whole data collection process paperless. Since almost everybody has a smartphone, the government can either design an app which can be used by the data collection personnel on their smartphones/tablets, or they can provide them with tablets specifically designed to collect census data. This adaptation single-handedly will enable the process to become fast, error-free and eco-friendly.

3.2 Usage of Cloud Storage

Currently, the data collected in the forms are sent to data collection centers. The process takes days to reach the data center. If the data is collected digitally, it becomes easy to store is on the cloud, a private cloud, of course. As of now, the data is uploaded after scanning the forms. But the forms are handled by human beings when they go from one place to other. This can lead to a data breach. Data security is a very important issue and it the data gets into the wrong hands, it may even pose a threat to the national security. Storing the data on the cloud will mitigate this problem to a great extent. So, the data must be securely stored to the cloud without anyone handling or tampering with it in the intermediate stages. Moreover, digital collection of data will enable to store the data on a secure cloud server via a secure internet connection. Obviously, cloud storage also has its own problems and threats such as someone hacking in the cloud and accessing the data. But data is much harder to access on a secure server than when multiple personnel all over the country are involved with the data directly.

3.3 Using Unique Identification

In the issue section, I had mentioned the problem of duplicate data. A person's information can be recorded multiple times at multiple places by his/her family members. Duplicate data is not only misleading, in the sense that it gives false insights about a demography, but also more costly to store. To eradicate this problem, the government can use the national identity of an Indian resident - the Aadhaar number [5]. The Aadhaar can be used to identify duplicate information about a person. Even if information about a person is recorded at multiple places and times, the Aadhaar number of the person can be checked before entering the details. If an entry related to a particular Aadhaar number already exists in the database, that means the information is already recorded. Therefore, there would be no need to record his/her information again. It will lead to a more reliable data with no duplicate information. Moreover, this practice would save a tremendous amount of resources

The use of Aadhaar has another very important advantage. Generally, while providing information about income, people are reluctant because of various reasons. Some people want to save tax by giving false income figures while some people just don't want to share their income with someone who they don't know. In India, the Aadhaar number is linked to each individuals PAN card [6] - a dedicated identity used to collect tax from an individual. The PAN is linked to a person's bank account. Therefore, the Aadhaar identity can be used to find out the real income of the person. At the least, it can be used to check if the bank statements corroborates the stipulated income figure provided by an individual. In case he/she has received an income significantly more than the stipulated income, the system can tell that it's a false information provided in order to save tax. Therefore, an individual can be asked to provide the correct information. Or at the very least, there can be a flag that can be raised which tells that the provided income information is false. It will help the data scientists to understand the data better and it will be at their discretion whether to use or discard that data which would eventually lead to use a much more useful

utilization of the data.

3.4 Standardized Data Collection

When entering data into the digital form, one has to make sure that it's standardized. The data needs to be uniform across the country as far as the format in which the data is collected. For example, the government can mandate that all the dates should be collected in dd/mm/yyyy format. So, across the domain, the data entered should be in one format, be it manually entering the date, or selecting date from the app. One could imagine this practice would not at all be difficult if the data is collected digitally. All the data types and their standard ways to collect them are listed in the Table 1.

3.5 Handle Missing Values

Missing values is a lot of pain for any data analyst. To this day, there are no effective ways to compensate for missing data. There are various techniques such as:

1. Imputing mean/median for numeric data and mode for categorical data
2. Using machine learning algorithms to predict missing data
3. Removing data points or features with missing data

All the above-listed techniques are just workarounds and each of them lead to a significant bias. None of these techniques is a reliable way to handle missing values. Each technique leads to its own bias. There can be two solutions to fix the problem of missing values while collecting the data:

1. While collecting the data, the data collection agent should make sure that no information is left unfilled.
2. Sometimes, a person is unwilling to share their information because of privacy reasons. In that case, there should be a separate option to fill that particular field with a predefined token such as 'NA' or 'Choose not to answer'. These situations can be predetermined and special tokens can be arranged for different situations. For example, for someone whose age is unknown, an 'NA' value would be the most appropriate to enter rather than filling random number like '0' or '9999', or a '-' or white space.

Handling missing values properly will lead to better quality data. Working with the data will become easier for someone to analyze it.

4 Conclusion

If India wants to become data-driven, the data collection practice should be efficient and reliable. There is no data collection exercise that's bigger than the decennial census data

TABLE 1
Efficient ways to collect data to ensure data quality

Data Type	Ideal Way to Collect Data
Date	Record in a uniform format such as dd/mm/yyyy throughout the country.
Text	The recording device should automatically remove unwanted characters such as the common prefixes or suffixes such as Mr or Miss, leading/trailing/multiple spaces between words, etc.A built-in spell corrector could help identify misspellings in text such as profession type.
Address	The address is collected as a string and includes a separate pin code (zip code). The address is a text data. However, the pin code is numeric. A pin code should be used to fill the major part of the address. Each place has a pin code in India. A pin code narrows down the choices for filling the complete address since a major part of the address such as district, city and state could be filled by the pin code automatically, in case the data is being collected digitally.
Numbers	Make sure that the range of the particular attribute or feature is kept in mind while recording data. For example, the age of an individual can possibly range from 1 - 100 years (not considering outliers). If someone enters negative age, that should not be excepted. An age of over 400 years also doesn't make sense. Similarly, data such as income figure should be exact. Suppose if a person has an income of INR 40,000 then the number "40000" must be entered rather than entering any value number such as 40k, or 40 thousand or 40.
Categories	Most of the data that's collected is categorical by nature. But as per current practices, it's filled as text. There must be predefined categories for fields such as gender, occupation type, the field of service, marital status, etc. Categories will help to record the data without any mistake since the person entering the data would need to choose from a list of predetermined options. Hence, it leaves no scope for a wrong entry as far as spelling mistakes are concerned. The data will be in a standard form after collection and there would be no need of manual labour identifying and correcting these issues.

collection. The government can use the census data to improve its administration. It needs to foresee the kind of problems that data science community will likely face once the data is collected. It also needs to learn from its past mistakes that have been encountered during the census. And last but not the least, it should make use of the latest technology to collect, store and process the data. This paper tries to present a list of measures that can be taken to progress towards this veritable endeavor of efficient and reliable data collection. The aim of this research is to edify the lecturer and teaching community about the practices in data collection that they need to follow in the next census.

References

[1] Census of India. URL http://censusindia.gov.in.

[2] Indian Population Statistics. URL https://www.statista.com/statistics/263766/total-population-of-india/.

[3] Digital India - Home. URL https://www.digitizeindia.gov.in/.

[4] Census of India - FAQs. URL http://censusindia.gov.in/2011-FAQ/FAQ-Public.html.

[5] Aadhaar - Home Page. URL https://uidai.gov.in/your-aadhaar/about-aadhaar.html.

[6] Indian Supreme Court orders for the Aadhaar and PAN card linkage. URL https://uidai.gov.in/images/news/Supreme_Courts_Order_in_WP_247_277_304_of_201716062017.pdf.

Author biography

Amandeep Rathee Content Strategist - Data Science | Machine Learning & AI, UpGrad