Abstract

In this paper, we provide a method to generate synthetic population at various administrative levels for a country like India. This synthetic population is created using machine learning and statistical methods applied to survey data such as Census of India 2011, IHDS-II, NSS-68th round, GPW etc. The synthetic population defines individuals in the population with characteristics such as age, gender, height, weight, home and work location, household structure, preexisting health conditions, socio-economical status, and employment. We used the proposed method to generate the synthetic population for various districts of India. We also compare this synthetic population with source data using various metrics. The experiment results show that the synthetic data can realistically simulate the population for various districts of India.

INTRODUCTION

The Census Organisation in India has been publishing the tabulated results of Census since the initiation of modern Census in 1872 in various book forms. As the information given by any respondent is treated as confidential, this micro-data is not publicly available. However Agent-based models, such as infectious disease modelling for Covid-19 require individual level information to model and simulate the behavior of the system’s constituent units (the agents) and their interactions. Though, within the 12th Five Year Plan (201217) Indian Census set-up 18 Workstations country-wide for research on micro-data from Census, it lacked the flexibility which researchers need in using such huge datasets. Thus it is crucial to have the ability to model population data and generate synthetic population at various administrative levels like, the Country, State, District, Sub-District, Town, Village or Ward in Town as the case may be.

In this paper, we use a variety of data sources to generate a population of individuals and households with demographic attributes that are statistically identical to real data. This population is generated using a hybrid of statistical methods and machine learning algorithms that are flexible enough to generate data at various administrative levels, ranging from small communities to states. The primary sources of data for these algorithms include the Census of India (Office of the Registrar General and Census Commissioner of India, India, 2011), the India Human Development Survey (IHDS) (Desai et al., 2018), the National Sample Survey (NSS) (NSSO, 2018), and the Gridded Population of the World (GPW) (CIESIN - Columbia University, 2016).

While the synthetic population should faithfully reproduce demographic statistics, it must also incorporate other realistic network structures, such as those appropriate to households and workplaces. Because different kinds of data respond well to different techniques, a hybrid process is used to scale up these datasets. First, the data is cleaned to remove obvious inconsistencies. Next, a customized hybrid of Iterative Proportional Fitting (IPF) (Beckman et al., 1996; Deming & Stephan, 1940), Iterative Proportional Updating (IPU) (Ye et al., 2009), and a specialized variant of a neural network, called Conditional-Tabular Generative Adversarial Network (CTGAN) (Xu et al., 2019), is used to generate new data.

Briefly, Iterative Proportional Fitting finds a joint distribution that matches the marginals, while trying to stay as close to the sample distribution as possible. Iterative Proportional Updating is a heuristic iterative approach which can simultaneously match or fit to multiple distributions (constraints). Finally, Conditional-Tabular Generative Adversarial Networks is a method to model the tabular data distribution and sample rows from the distribution. A Generative Adversarial Network (GAN) (Goodfellow et al., 2014) uses two “competing” neural networks, the generator and the discriminator. The generator creates realistic samples with
the goal that the discriminator should be unable to differentiate between a real sample and a generated sample. In this zero-sum game, capabilities of both the networks are enhanced iteratively. Critically, our techniques are designed to work seamlessly across data-scarce and data-rich areas; even if a particular area has error-prone or missing data, a synthetic population can still be generated, albeit of a lower quality.

**THE POPULATION GENERATION PROCESS**

We use IPF to generate a base population, using census data for the demographics and the IHDS survey dataset for personal and household attributes. The base population thus consists of individual data and household data. We assign each household to an administrative unit within a district.

We also experimented with CTGAN to generate a base population. The major advantage of IPU over CTGAN is that IPU is capable of matching individual level and household level characteristics of an individual while making sure that members of the household have a realistic age and gender joint distribution.

To assign job labels to individuals, the relevant data from the IHDS dataset is used. For the time-being, we classify individuals below the age of 18 as students, but could easily relax this assumption. A subset of the population is also assumed to be home-bound. This subset consists of unemployed individuals, homemakers, infants and children under the age of 3 and elderly people over the retirement age. We use data from the NSS survey to determine the percentage of adult males and females in a city who are home-bound. A random independent sample is drawn from a Bernoulli distribution with this gender-based marginal value as a parameter in order to decide if an individual will be home-bound or not.

Each student in the population is assigned a school. Similarly, each working individual is assigned a workplace based upon their job label. We generate a synthetic latitude and longitude pair for each home, school and workplace in our dataset using GADM grid population density data (J. Hijmans, 2018). We select a subset of grid points that lie within a given geographical boundary and sample grid-points with replacement grid points from the subset, weighing each point by the population density in the associated grid. We add independent random noise drawn from a uniform distribution to the latitudes and longitudes, rejecting those samples which fall outside the given geographical boundary. We follow this process to generate synthetic geolocation data for households, schools and workplaces.

To assign an individual a school, we sample from the list of schools within that geographical boundary, weighing each school by the inverse of the euclidean distance between it and the individual’s home. This weighting factor increases the probability of assigning an individual a school that is closer to their home (Ministry of Education, GoI, 2019). We follow a similar method to assign workplaces to adults. Additionally, based on every individual’s job label, a workplace is assigned at random from a suitable subset of allowed workplaces.

A number of additional attributes are included in our synthetic population, including whether an individual uses public transport or whether an individual is an essential worker. These values are assigned using the individual’s job label.

**POPULATION VERIFICATION METRICS**

To compare and verify the generated synthetic population with the survey data, we used several methods. We used the Bhattacharya distance to quantify the similarity of the joint age-height and age-weight distributions. In addition, apart from comparing the two populations visually as seen in Figure 1 and Figure 2, we have also used a number of other metrics such as statistical likelihood techniques (CS-test, KS-test). We also visualise the geographical spread of the households, schools and the workplaces in the population as in Figure 3. The visual comparison shows the synthetic population resembles the real population.
CONCLUSION

Lack of data due to access and privacy issues results in poor model design. To tackle this issue, we propose a hybrid method to generate a synthetic population. We also provide a combination of metrics to verify the generated data. In ongoing work, we are generating the synthetic population for the entire country of India. As future work, we want to explore the possibility of modelling more nuanced and complex features for the synthetic population.

Acknowledgements

The authors are grateful for support from the Mphasis F1 Foundation and the Bill and Melinda Gates Foundation, Grant No: R/BMG/PHY/GMN/20.

References

Centre for Research in Micro Census Data, 2011. URL https://www.isical.ac.in/~library/census.php.

Beckman, R. J., Baggerly, K. A., and McKay, M. D. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice, 30(0965-8564):415–429, 1996.

Bonabeau, E. Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences, 99(3):7280–7287, 2002.

CIESIN - Columbia University. Gridded population of the world, (GPWv4): Population density, 2016.

Deming, W. E. and Stephan, F. F. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. The Annals of Mathematical Statistics, 11(4):427–444, 1940.

Desai, S., Vanneman, R., and National Council of Applied Economic Research. India Human Development Survey-II (IHDS-II), 2011-12. Inter-university Consortium for Political and Social Research, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, pp. 2672–2680, Montreal, Quebec, Canada, 2014.

J. Hijmans, R. Database of Global Administrative Areas, 2018. URL https://gadm.org/data.html.

Ministry of Education, GoI. Steps taken by Government to Provide Education to Poor Students, 07 2019. URL https://pib.gov.in/PressReleasePage.aspx?PRID=1578389.

NSSO. India - Household Consumer Expenditure, NSS 68th Round . Technical report, 2018. URL http://microdata.gov.in/nada43/index.php/catalog/126.

Office of the Registrar General and Census Commissioner of India, India. CENSUS TABLES, 2011.
Generating Synthetic Population

URL https://censusindia.gov.in/census.website/data/census-tables.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional gan. In Advances in Neural Information Processing Systems, 2019.

Ye, X., Konduri, K., Pendyala, R., Sana, B., and Waddell, P. Methodology to match distributions of both household and person attributes in generation of synthetic populations. 01 2009.
### A. Sample Synthetic Population

A sample of ten individuals in the synthetic population is reproduced below. The columns are split to fit in the page.

| Age | SexLabel | Height | Weight | HHID   | H_Lat | H_Lon | District | AdminUnitName |
|-----|----------|--------|--------|--------|--------|--------|----------|--------------|
| 56  | Male     | 165.970| 44.381 | 5.190e+10 | 19.024 | 72.911 | Mumbai   | M/E          |
| 10  | Male     | 138.150| 34.388 | 5.190e+10 | 19.074 | 72.832 | Mumbai   | H/W          |
| 3   | Female   | 65.460 | 7.205  | 1.038e+11 | 19.258 | 72.867 | Mumbai   | R/N          |
| 63  | Male     | 170.400| 59.962 | 1.038e+11 | 19.046 | 72.933 | Mumbai   | M/E          |
| 37  | Female   | 143.530| 60.381 | 5.190e+10 | 19.189 | 72.837 | Mumbai   | P/N          |
| 46  | Male     | 163    | 63.064 | 1.038e+11 | 19.071 | 72.889 | Mumbai   | L            |
| 35  | Female   | 165.660| 70.111 | 1.038e+11 | 19.013 | 72.886 | Mumbai   | M/W          |
| 29  | Male     | 152.520| 59.198 | 5.190e+10 | 19.131 | 72.898 | Mumbai   | S            |
| 74  | Female   | 169.560| 37.874 | 5.190e+10 | 18.952 | 72.797 | Mumbai   | D            |

| AdminUnitLatitude | AdminUnitLongitude | Religion | Caste | JobLabel | JobID | WorkPlaceID |
|-------------------|--------------------|----------|-------|----------|-------|-------------|
| 19.056            | 72.922             | Hindu    | other | Carpenters | 81   | 2.001e+12   |
| 19.056            | 72.835             | Hindu    | other | Student   | 199  | 0           |
| 19.120            | 72.852             | Hindu    | other | Homebound | 0    | 0           |
| 19.056            | 72.922             | Hindu    | other | Homebound | 0    | 0           |
| 19.188            | 72.842             | Hindu    | other | Homebound | 0    | 0           |
| 19.070            | 72.879             | buddhist | other | Labour nec | 99   | 2.001e+12   |
| 19.084            | 72.906             | Hindu    | SC    | Construction | 95   | 2.001e+12   |
| 19.061            | 72.899             | Hindu    | other | Student   | 199  | 0           |
| 19.139            | 72.930             | Hindu    | other | Construction | 95   | 2.001e+12   |
| 18.963            | 72.813             | Hindu    | other | Construction | 95   | 2.001e+12   |

| W_Lat | W_Lon | essential_worker | Adherence_to_Intervention | PublicTransport_Jobs | school_id |
|-------|-------|------------------|---------------------------|----------------------|-----------|
| 19.036| 72.867| 0                | 0.900                     | 1                    | 0         |
| nan   | nan   | 0                | 0.800                     | 1                    | 2.001e+12 |
| nan   | nan   | 0                | 1                        | 1                    | 0         |
| nan   | nan   | 0                | 1                        | 1                    | 0         |
| nan   | nan   | 0                | 0                        | 1                    | 0         |
| 19.071| 72.890| 0                | 0.900                     | 1                    | 0         |
| 19.192| 72.865| 0                | 0                        | 1                    | 0         |
| nan   | nan   | 0                | 1                        | 1                    | 2.001e+12 |
| 19.155| 72.883| 0                | 0.200                     | 1                    | 0         |
| 19.024| 72.850| 0                | 1                        | 1                    | 0         |

| school_lat | school_long | public_place_id | public_place_lat | public_place_long | Agent_ID | PSUID |
|------------|-------------|-----------------|------------------|-------------------|----------|-------|
| nan        | nan         | 3.001e+12       | 19.024           | 72.910            | 5.191e+10| 1     |
| 19.177     | 72.867      | 3.001e+12       | 19.179           | 72.828            | 5.191e+10| 20    |
| nan        | nan         | 3.001e+12       | 19.098           | 72.844            | 5.191e+10| 9     |
| nan        | nan         | 3.001e+12       | 19.148           | 72.842            | 5.190e+10| 14    |
| nan        | nan         | 3.001e+12       | 19.189           | 72.830            | 5.190e+10| 17    |
| nan        | nan         | 3.001e+12       | 19.092           | 72.889            | 5.191e+10| 5     |
| nan        | nan         | 3.001e+12       | 19.059           | 72.895            | 5.190e+10| 8     |
| 18.962     | 72.838      | 3.001e+12       | 19.223           | 72.866            | 5.191e+10| 22    |
| nan        | nan         | 3.001e+12       | 19.129           | 72.913            | 5.191e+10| 13    |
| nan        | nan         | 3.001e+12       | 19.046           | 72.851            | 5.190e+10| 1     |
| M_Fever | M_Diarrhea | M_Cataract | M_Heart_disease | M_Diabetes | M_Leprosy | M_Cancer |
|---------|------------|------------|-----------------|------------|-----------|----------|
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |
| 0       | 0          | 0          | 0               | 0          | 0         | 0        |

| M_Asthma | M_Paralysis | M_Epilepsy |
|----------|-------------|------------|
| 0        | 0           | 0          |
| 0        | 0           | 0          |
| 0        | 0           | 0          |
| 0        | 0           | 0          |
| 0        | 0           | 0          |
| 0        | 0           | 0          |
| 0        | 0           | 0          |
| 0        | 0           | 1          |
| 0        | 0           | 1          |
| 0        | 0           | 1          |
| 0        | 0           | 0          |