A Hybrid Framework for Generating A Country-scale Synthetic Population

Anonymous authors
Paper under double-blind review

Abstract

Population censuses are vital to public policy decision-making. They provide insight into human resources, demography, culture, and economic structure at local, regional, and national levels. However, such surveys are very expensive (especially for low and middle income countries with high populations, such as India), and may also raise privacy concerns, depending upon the kinds of data collected.

We introduce a novel hybrid framework which can combine data from multiple real-world surveys (with different, partially overlapping sets of attributes) to produce a real-scale synthetic population of humans. Critically, our population maintains family structures comprising individuals with demographic, socioeconomic, health, and geolocation attributes: this means that our "fake" people live in realistic locations, have realistic families, etc. Such data can be used for a variety of purposes: we explore one such use case, Agent-based modelling of infectious disease in India.

We use both machine learning and statistical metrics to gauge the quality of our synthetic population. Our experimental results show that synthetic data can realistically simulate the population for various administrative units of India, producing real-scale, detailed data at the desired level of zoom – from cities, to districts, to states, eventually combining to form a country-scale synthetic population.

1 Introduction

Advancements in computing have made it feasible to train deep neural networks with more than a 100 billion parameters and build complex models for real world scenarios. For example, BERT Devlin et al. (2018), released in 2019, has around 350 million parameters while GPT-3 Brown et al. (2020), released in 2021, has more than 175 billion parameters. These large models require relevant and sufficient data for training. Furthermore, financial institutions, government agencies, think tanks, etc. are using techniques like agent-based modelling(ABM) Bonabeau (2002) to simulate increasingly complex scenarios for decision making. Real world constraints such as high costs of acquisition, privacy laws, etc. might prevent the collection of extensive datasets needed for model training and Agent-based modelling. Hence, the quantity and quality of available datasets is a bottleneck.

Population censuses and various other surveys e.g., for health, economics, education etc. help inform public policy decision making. Periodic censuses and surveys provide aggregate statistics on variables describing the demography, socioeconomic distribution and health status of the constituents at various levels of zoom. They also help decision makers examine the effects of previous public policy decisions on the evolution of these statistics. Surveys that provide data at an individual level instead of marginals are of special interest as, once released, they allow for superior analysis of correlations between different attributes. However, data collection for censuses and surveys is a time consuming and expensive task, especially for low and middle income countries. For example, in India, a country with a population of around 1.35 billion, collecting population data is a gigantic operation. The latest Census in India would require nearly three million "enumerators" and "supervisors" (Office of the Registrar General, India, 2019), resulting in huge time and resource investments. Further, national censuses do not often release full scale individual data for public access due to privacy reasons which limits the potential for analysis.

Privacy preserving synthetic populations can be used as an alternative to real individual level datasets. To increase the accuracy of downstream tasks, it is essential that the marginal and joint distributions of the various attributes are similar across the original and the synthetic population. Further, to address privacy concerns, it is necessary to ensure that the synthetic dataset does not leak any personally identifiable information about the individuals who were originally surveyed. Some analysis tasks might require a dataset with a specific combination of attributes for which a survey has not been conducted. To address this problem, two or more datasets with different, partially overlapping sets of attributes can be combined to generate a synthetic population with the specific set of attributes required for the downstream task. For example, SynC Wan et al. (2019) is framework for generating individual level synthetic data and can combine multiple datasets to generate a synthetic population with more attributes than any of the underlying real datasets.

Additionally, with the advent of graph neural networks and the growing complexity of Agent-based models, it is necessary to produce synthetic datasets where individuals are parts of networks and the linkage is based on shared attributes like a household, a workplace, a school, common direction of travel, etc. Preexisting frameworks for synthetic population generation are limited in this regard.

To focus our discussion, we will be looking at the problem with the use case of modelling disease spread (via Agent-based modelling) throughout this paper.

In this context, we make the following contributions:

- A novel hybrid framework which combines various state of the art statistical and machine learning models to generate a real-scale synthetic population with network and geolocation data.
- A combination of metrics to help verify the synthetic population
- To promote transparency and reproducibility, our code and datasets are available at this link: `https://anonymous.4open.science/r/synthpop-C67C/`

The rest of the paper is organised as follows. Section 2 is about related work to generate synthetic population. In Section 3, we present our approach in detail. The experiments and evaluation are presented in Section 4. We discuss future work and conclude in Section 5.

2 Related Work

In the task of modeling and generating synthetic populations, there are various challenges in modelling the tabular data. For example, a table might have mixed data types with discrete and continuous columns. The columns in the table (especially those with continuous values) might not follow, say, the Gaussian distribution. Further, a column might have multi-modal distributions. The data may also be highly imbalanced resulting in insufficient training opportunities for minor classes. Some ABMs require the synthetic population to be realistically geolocated, which, to best of our knowledge, has not been attempted, especially not for India.

Though, various methods have been suggested to tackle these issues, they do not provide a solution to the issues specified. For example, Bayesian networks Koller & Friedman (2009) combine low dimensional distributions to approximate the full-dimensional distribution of a data set, and are a simple but powerful example of a graphical model. But, Bayesian methods can not model tabular data effectively when both continuous and discrete columns are present. Deep Generative models are generally more sophisticated but in practice can not outperform Bayesian methods as the tabular data may contain Non-Gaussian continuous columns and/or imbalanced discrete columns. Some methods have tried to mitigate these issues. PrivBayes Zhang et al. (2014) can model the table with discrete variables but all continuous variables need to be discretised. Additionally, noise injection to preserve privacy, may result in low-quality synthetic data. MedGAN Choi et al. (2017) has been used to generate health records but it does not support mixed data types. It also can not handle Non-Gaussian, Multi-modal, imbalanced categorical columns, lack of training data and missing data. Although, CTGAN Xu et al. (2019) can work with tables with mixed data types, high dimensionality, imbalanced categorical data etc., it can not handle lack of training data and missing values. Also, it can not model realistic family structures. Iterative Proportional Updating (IPU) Ye et al. (2009) can work with limited data and model realistic family structures, but it does not scale well with high dimensional data. To tackle these issues, we propose our framework which uses a hybrid 1 of

IPU and CTGAN to take advantage of the best from these two models. Our framework allows for combining data from various sources and generates a synthetic population with required columns.

3 Our Framework

A synthetic population is a limited individual-level representation of the actual population. However, not all the attributes of an individual are included (for example, hair colour or shoe-size might be irrelevant for modelling disease spread, while co-morbidities like diabetes would be included). As such, a synthetic population does not aim to perfectly mimic reality – this would be impossible. Instead, it attempts to sufficiently match various statistical measures observed in the real population.

The components of any pipeline used for generating a synthetic population would largely depend on the variables that are required for the downstream task. As stated earlier, our pipeline as shown in diagram 1 is built with a focus on synthetic population data for district level infectious disease modelling but we believe that our methods can be used to generate synthetic populations which can be useful for a diverse set of tasks.

Our objective is to generate a real scale synthetic population for an entire country with distributions and joint distributions of various variables that match those of the actual population. This is necessary to ensure that the results of the downstream tasks are reliable.

For the chosen downstream task, the data should capture various heuristics like the size of a family; the joint distribution of age and sex of individuals within a family; administrative units like cities, districts, states; geographical distribution of households, workplaces, schools within an administrative unit; the number of people who are associated with a workplace or school; and various other individual level statistical correlations that can be observed in any population. Iterative Proportional Updating and Conditional Tabular GAN were used to generate families and individual, and their related metadata. Further, reject sampling was used to sample geo-location points within a city's boundaries which followed the per unit area population density distribution. A method for random assignment of external locations like workplaces, schools and public places was used where the probability of selecting any external location was inversely proportional to the distance between a given individuals household and the external location Ministry of Education (2019).

3.1 Base Data Generation : IPU and CTGAN

IPU Ye et al. (2009) is a sampling method which ensures that the marginal distributions for both household and individuals in the synthetic population are in line with the marginals distributions of the real population. Hence, sampling with IPU helps in generating a population in which households have a realistic distribution of the number of family members and their age and gender. CTGAN Xu et al. (2019) can be used for adding individual level attributes for which marginals are not available. The CTGAN model can be conditioned on attributes which are present in the base population generated by IPU which would allow for a reasonable joint distribution of various individual level attributes in the synthetic population. Multiple CTGAN models can also be used to join survey datasets which have an overlapping set of attributes. This allows for expansion of attributes as per the needs of the downstream task. More importantly, this can synthesize populations with a list of attributes for which no survey has been conducted.

3.2 Population Density Distribution Sampling

Given that the downstream task is to analyse the spread of infectious diseases in a given region, it is necessary to have accurate geographical distribution of population density and workplace density. To this end, grid population density data Gaughan et al. (2013) is used. A row in the grid population density data consists of latitude (X), longitude (Y) and the number of individuals (Z) which live within a square with edge length S and with it's centre at the given latitude and longitude. For the data set used in the current pipeline, the length of the side of square $S = 0.3\ arctan$ which translates to a length of about $1\ km$ on the equator. The sampling method also needs a polygon for the boundaries of the region for which data is being generated.

Let n be the number of latitude longitude pairs that need to be sampled within this region.

A filtered subset of the population density dataset is obtained by only retaining those rows for which the latitude X and longitude Y are within the boundary polygon of the region. We sample with replacement k latitude longitude pairs from this filtered subset and use the Z value as the weight for this sampling. For a given row in the sample, we then sample points within the squares corresponding to the latitude longitude pair by adding independent random uniform noise $A, B \sim U(-S/2, S/2)$ to the latitude X and longitude Y respectively. We reject those latitude and longitude pairs that are not within the polygon. Given, this method uses reject sampling, we initially need to sample more points than required and hence we need to use a suitable value for $k > n$ and then chose at random from the pairs which were not rejected.

We use this method to assign geolocation to houses, workplaces, schools and public places. Given we do not have access to workplace distribution data, we assume that it follows the population density distribution as well. However, this assumption is not valid in cities where people live in outer sub urban areas and work closer to the city center. With the appropriate dataset, we can substitute the population density data with workplace density data and get a distribution similar to the ground reality with the sampling method described earlier.

3.3 External Location Assignment

Since infectious disease modelling often includes analysis based on contact network graphs, individuals in this synthetic population would need to be associated with external locations like workplaces, schools and public places which they might periodically visit. This assignment is based on the assumption that adults work at workplaces, children go to schools and people are more likely to visit public places which are closer to their homes. We use L2 distance and not a geodesic distance metric for this computation since the curvature of the earth within a city is negligible. Given an individual with home latitude X_h and longitude Y_h and a list of possible external locations with latitude longitude pairs (X_{E_1}, Y_{E_1}), ..., (X_{E_k}, Y_{E_k}), we calculate the L_2 distance between the home of the individual and the external location. Then we choose a function $f(x)$ which is strictly decreasing in the positive real domain. Then we weigh the probability of the individual being assigned the external location by the value $f(D((X_h, Y_h), (X_{E_k}, Y_{E_k})))$ where D is the L_2 distance in 2 dimensions.

4 Experiments and Evaluation

4.1 Datasets and Data Preparation

As described in *Our Framework* section, we use various datasets as input to our framework. Here we describe these datasets and the methods to clean these datasets in order to make them suitable for our framework. We used sample survey data from India Human Development Survey-II Desai et al. (2018), marginals from Census, 2011 Office of the Census Commissioner of India (2011a), Data for employment from NSS National Sample Survey Office,NSSO (2012), and population density from GADM grid population density dataset J. Hijmans (2018).

4.2 Generating Synthetic Population

A synthetic population for Agent-based disease modelling should have demographic, socioeconomic, health related and geographic variables for each individual in the population. We will go over the steps for generating a synthetic population for the city of Mumbai, located in the state of Maharashtra in India. We acquire marginals for individual and household attributes. For individuals the attributes are age, sex [1], religion and caste. For households, the only attributed is household size i.e. the number of members in the household. We also use a subset of IHDS 2 dataset obtained by filtering for individuals and households which are situated in the state of Maharashtra. Instead of filtering for individuals and households in the city of Mumbai, we use entire dataset for Maharashtra because the Mumbai dataset is very small. We then use IPU which use the marginals and the microdata to generate households of varying sizes. Each household has a certain number of individual family members, each with their own age, sex, religion and caste. There may be some shared attributes between family members like common religion

[1] We used the same terminology as used in the Census of India

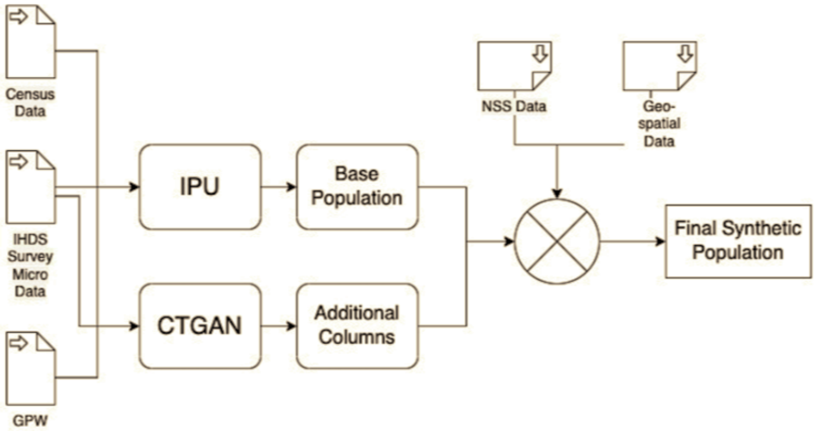


Figure 1: Our Framework : Pre-processed data is used for generating data with various models

and caste. Further the distribution of age and sex of members within a family of a given size resembles that of a typical family of the given size that one would expect to see in the real world. We then sample latitude longitude pairs using the method described in Section 3.2 for each household. The same household latitude and longitude is attached to each member of the household.

We then add individual level attributes like weight, height, comorbidities and job description. For weight, height and comorbidities, we use a CTGAN model conditioned on age and sex. The CTGAN model is, once again, trained on the subset of IHDS 2 dataset containing individuals situated within the state of Maharashtra. For individuals below the age of 3, ’Homebound’ is assigned as the job description. For individuals above the age of 3 and below the age of 18, ’Student’ is assigned as the job description. For individuals, above job description is sampled with replacement from the list of job descriptions of the individuals in the IHDS 2 subset. The weight assigned to each possible job description is linearly proportional to the distribution observed in the IHDS 2 subset for Maharashtra.

We then generate synthetic workplaces, schools and public places for our synthetic individuals to go to. The number of workplaces, schools and public places are parameters which need to be adjusted on the basis of the district under consideration. Each workplace has three attributes, workplace type, latitude and longitude. The workplace type attribute is generated by sampling with replacement from the list of job descriptions of the individuals in the IHDS 2 subset. Latitude and longitude pair are once again sampled with method described in Section 3.2. Schools are obtained by filtering for subset with workplaces which have workplace type set to ’Teacher’. We then assign schools to students and workplaces to individuals who are not homebound or students. Assigning schools in straightforward as we use the method from Section 3.3. While assigning workplaces we need to consider the fact that not all individuals at a given workplace might have the same job description. To address this issue, we first sample from a Bernoulli distribution with a sensible parameter. If the Bernoulli random variable is 1, we sample a workplace using the method from Section 3.3 from a subset of workplaces for which the workplace type attribute is the same as the job description attribute of the individual under consideration. If the Bernoulli random variable is 0, we sample a workplace from a subset of workplaces for which the workplace type attribute is not the same as the job description attribute of the individual under consideration.

We have used the above steps to generate synthetic populations for multiple districts of India. This synthetic population has been used in a downstream task of infectious disease modelling, under another framework called BharatSim, described in Section 4.4. Here, we describe our synthetic population for the combined districts of Mumbai and Mumbai Suburban. According to Census of India, 2011 Office of the Census Commissioner of India (2011b), the district of Mumbai had a total population of 3,085,411 and the district of Mumbai Suburb had a total population

Table 1: Statistical Metrics

Test	Average Score
CSTest	0.99
KSTest	0.94

Table 2: ML efficacy tests

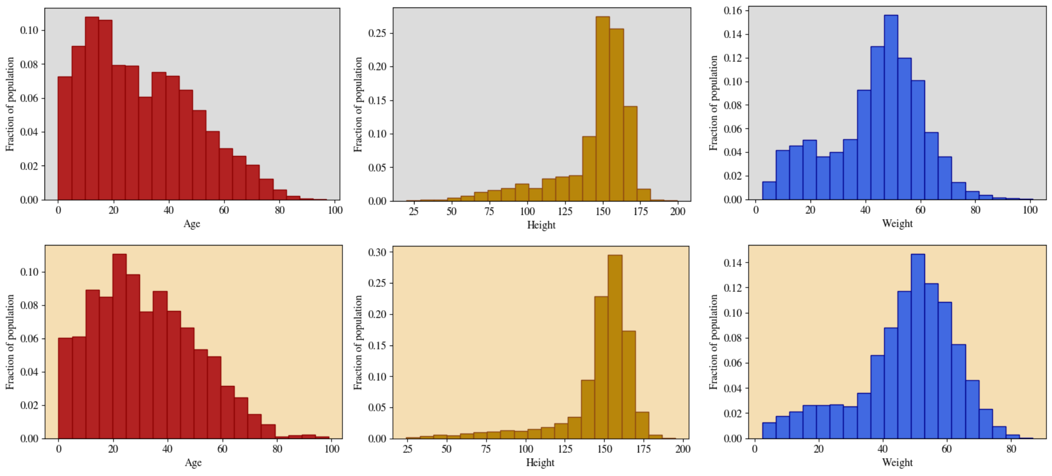
Test	Source Population	Synthetic Population
Linear Regression for Weight	0.72	0.69
MLP Regressor for Weight	0.78	0.76
Linear Regression for Height	0.69	0.65
MLP Regressor for Height	0.71	0.72

of 9,356,962. We used our framework to generate the entire population for these two districts synthetically. A sample of our synthetic population is shown in *the Appendix* section.

4.3 Evaluating data generated by our framework

Though, there are no standard techniques for evaluating the quality of synthetic population data, various papers have tried different methods. For example, Wan et al. (2019) uses ML classification tasks to check the usefulness of synthetic population augmented data. Yue et al. (2018) also uses classification tasks and regression tasks to validate the synthetic population. We evaluate the synthetic population based on the following three criteria:

- Distribution of individual features in synthetic population match those in sample surveys, i.e., histogram should look similar
- Joint distribution of selected features in synthetic population match those in sample surveys, i.e., scatter plot should look similar
- Relationship between features and selected target variable in synthetic population match the same in sample survey

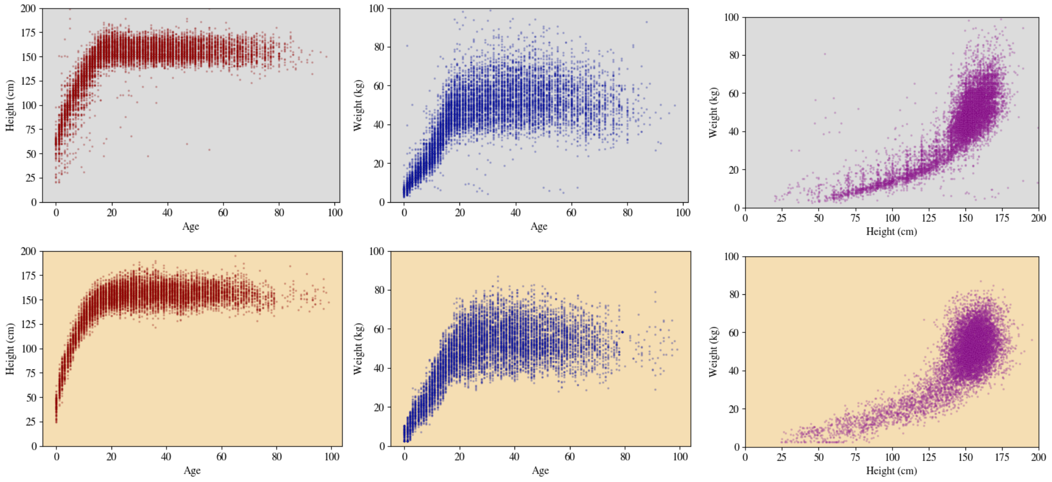


(a) Population age distribution (b) Population height distribution (c) Population weight distribution

Figure 2: Histogram for marginal distributions of age, height and weight: Comparing source population (top) with synthetic population for the combined data for the districts of Mumbai and Mumbai Suburban, India

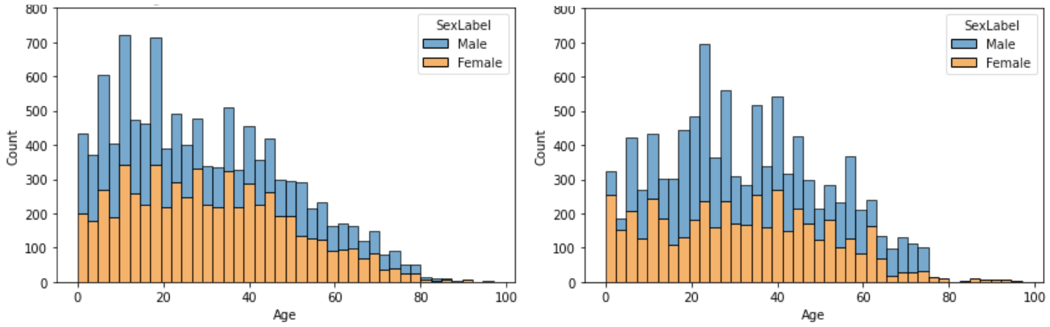
Figure 2 helps compare the marginal distributions of basic attributes age, height and weight across the real and the synthetic population for the city of Mumbai. The marginal distribution have similar shapes even though the synthetic population is real scale. Figure 3 juxtaposes the scatter plots of basic attributes like age, height and weight for real and synthetic populations. The scatter plots depict the joint distributions which are roughly similar across the real and the synthetic population. The joint distribution of age and gender can be seen in figure 4. Figure 5 helps visualize the geographic distribution of synthetic households, workplaces and schools in the combined district of Mumbai and Mumbai Suburban.

Table 1 shows results for two statistical metrics, the two-sample Kolmogorov–Smirnov test (KSTest) and the Chi-Squared test (CSTest). The statistical tests are executed on all the compatible columns (so, categorical or boolean columns for CSTest and numerical columns for KSTest) in the real and synthetic population, and the average scores are reported in the Table 1. Another set of tests we conducted to evaluate if our synthetic population can replace the real population to solve a machine learning problem. For this experiment, we use two ML models, Linear Regression and MLP Regressor to train on the synthetic data and then test on the real data. We compare this result with the one obtained when trying to make the prediction using real data as input. The comparative results are shown in Table 2.



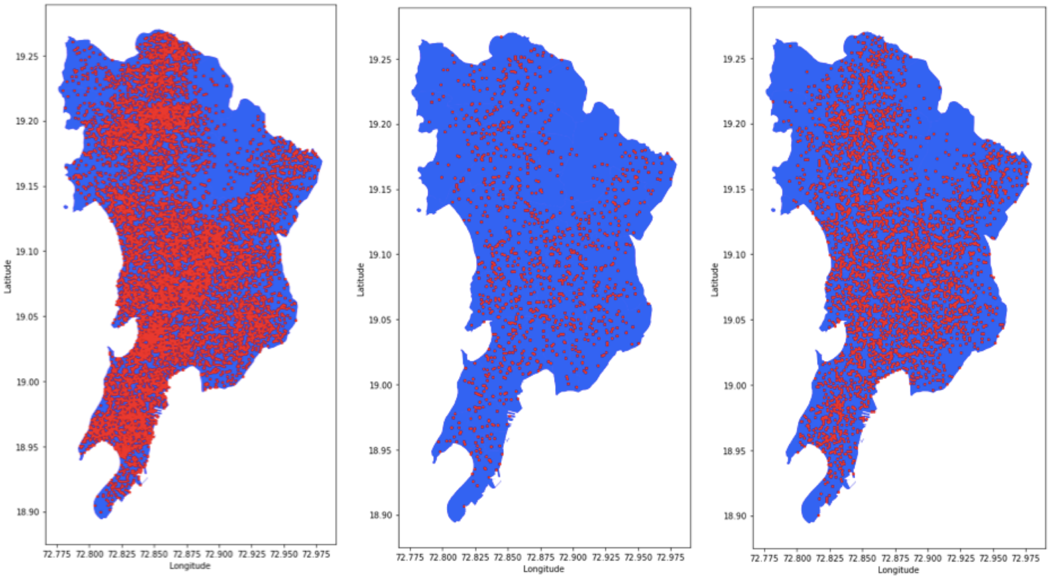
(a) Age Height joint distribution (b) Age Weight joint distribution (c) Height Weight joint distribution

Figure 3: Scatter plots for joint distributions of Age, Height and Weight: Comparing source population (top) with synthetic population for the combined data for the districts of Mumbai and Mumbai Suburban, India



(a) (b)

Figure 4: Histogram for marginal distributions of age-wise sex distribution: Comparing source population (left) with the combined synthetic population for the districts of Mumbai and Mumbai Suburban, India



(a) Households (b) Schools (c) Workplaces

Figure 5: Geographical distribution of households, schools and workplaces for the combined synthetic population for the districts of Mumbai and Mumbai Suburban, India

4.4 Our Results on BharatSim

BharatSim, an Agent-based model, is capable of allowing for complex social dynamics. Real-world systems involve interactions between individuals with different attributes (age, weight, etc.) and geographies. These interactions lead to emergent phenomena, while events like a pandemic affect individuals according to their attributes. Agent-based modeling accounts for individual differences and allows us to simulate scenarios of varying complexity. These simulations can guide policy level interventions (eg. lockdowns).

The synthetic population gives us information on the homes and workplaces of all agents in the population, and the individual's movements in a day are charted by defining a "schedule" specific to that individual, which decides how individuals move back and forth between their homes and workplaces or schools during a day. The infection spreads stochastically between individuals who share a location at any given time, at rates determined by the infection parameters. In this model, we have constructed a simple SIR compartmental structure describing a hypothetical disease: Susceptible (S) individuals can be infected at a rate proportional to the fraction of infected (I) individuals they are in contact with. The constant of proportionality, which determines how infectious the disease is, is age-stratified. Once an agent is infected, they can infect other individuals until they recover. The time to recovery is assumed to be drawn from an exponential distribution with a mean of 7 days. For a simulated lockdown, individual agents' schedules are modified so that they remain at home. All agents are required to follow this rule, except those that are designated by the synthetic population as essential-workers, and those who have a low probability of adherence to policy-level interventions. Whether or not an individual falls into the latter category is decided by comparing a random coin-toss to their "Adherence to interventions" value. The reduction in the number of people in different network locations thus reduces contacts between individuals and restricts the spread of the disease. In our simulations, we have considered two counterfactual scenarios, one in which a lockdown is imposed when the number of active cases is 1% and 2% of the total population, respectively. As can be seen in Figure 6, the lockdown flattens the curve, keeping the peak in active cases low. Additionally, starting the lockdown earlier is much more effective in curtailing the spread of the disease. The results are averaged over 20 runs, with the individual runs in the background.

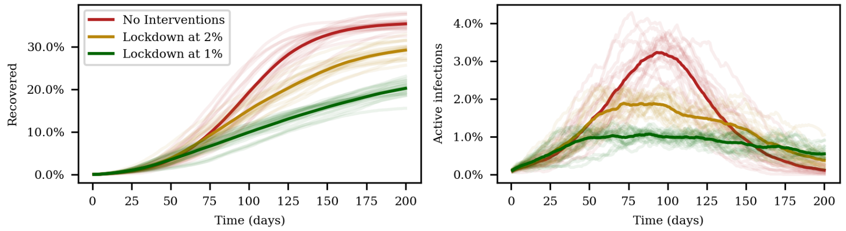


Figure 6: No. of recovered individuals (left) and active infections over time

5 Conclusion and Future Work

In this paper, we propose a novel hybrid framework for generating synthetic populations at various scales ranging from a few thousand to a billion individuals. We also suggest a combination of metrics to verify the generated synthetic population.

Future work would include the exploration of methods to generate a synthetic population without the need for sample survey data, as well as the modelling of complex features such as economic variables. Another target would be a realistic intra-city geographical distribution of individuals depending on their demographic attributes.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL http://arxiv.org/abs/2005.14165.

Edward Choi, Siddharth Biswal, Bradley A. Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. volume 68, 2017. URL http://dblp.uni-trier.de/db/conf/mlhc/mlhc2017.html#ChoiBMDSS17.

Sonalde Desai, Reeve Vanneman, 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL http://arxiv.org/abs/1810.04805.

Andrea E Gaughan, Forrest R Stevens, Catherine Linard, Peng Jia, and Andrew J Tatem. High resolution population distribution maps for southeast asia in 2010 and 2015. *PloS one*, 8(2): e55882, 2013.

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

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2009. URL https://books.google.co.in/books?id=7dzpHCHzNQ4C.

GoI Ministry of Education. Steps taken by government to provide education to poor student, Jul 2019. URL https://pib.gov.in/PressReleasePage.aspx?PRID=1578389.

National Sample Survey Office,NSSO. NSS 68th Round, 2012. URL http://www.icssrdataservice.in/datarepository/index.php/catalog/91.

Office of the Census Commissioner of India. Census Tables, 2011a. URL https://censusindia.gov.in/census.website/data/census-tables.

Office of the Census Commissioner of India. Districts of Maharashtra, 2011b. URL https://www.census2011.co.in/census/state/districtlist/maharashtra.html.

Office of the Registrar General, India. Centre of India, 2021 - Circular No. 6, 2019. URL https://censusindia.gov.in/nada/index.php/catalog/40515/download/44147/ORGI_circular006_2021.pdf.

Colin Wan, Zheng Li, Alicia Guo, and Yue Zhao. Sync: A unified framework for generating synthetic population with gaussian copula, 2019. URL https://arxiv.org/abs/1904.07998.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

Xin Ye, Karthik Konduri, Ram Pendyala, Bhargava Sana, and Paul Waddell. Methodology to match distributions of both household and person attributes in generation of synthetic populations. 01 2009.

Yang Yue, Ying Li, Kexin Yi, and Zhonghai Wu. Synthetic data approach for classification and regression. 2018. URL http://dblp.uni-trier.de/db/conf/asap/asap2018.html#YueLYW18.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: private data release via bayesian networks. 2014. URL http://dblp.uni-trier.de/db/conf/sigmod/sigmod2014.html#ZhangCPSX14.

A Appendix

A.1 Sample Population

A sample of the synthetic population. The columns are split to fit in the page.

	Age	Height	Weight	PSUID	M_Fever	M_Cough	M_Diarrhea
0	77	147.41	49.06	23	0	0	0
1	77	146.96	49.55	15	0	0	0
2	65	166.68	54.36	22	0	0	0
3	67	167.50	55.72	14	0	0	0
4	68	159.10	60.12	10	0	0	0
5	78	147.05	50.28	4	0	0	0
6	78	145.84	49.08	7	0	0	0
7	76	147.51	49.21	7	0	0	0
8	65	154.97	53.50	22	0	0	0
9	69	151.37	51.58	2	0	0	0

	M_Cataract	M_TB	M_High_BP	M_Heart_disease	M_Diabetes	M_Leprosy
0	0	0	0	0	0	0
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0

	M_Cancer	M_Asthma	M_Polio	M_Paralysis	M_Epilepsy	SexLabel
0	0	0	0	0	0	Female
1	0	0	0	0	0	Female
2	0	0	0	0	0	Male
3	0	0	0	0	0	Male
4	0	0	0	0	0	Male
5	0	0	0	0	0	Female
6	0	0	0	0	0	Female
7	0	0	0	0	0	Female
8	0	0	0	0	0	Male
9	0	0	0	0	0	Male

	StateLabel	Religion	Caste	District	JobLabel	JobID
0	Maharashtra	Hindu	SC	Pune	Construction	95
1	Maharashtra	Hindu	SC	Pune	Electrical	85
2	Maharashtra	Hindu	other	Pune	Clerical supe	30
3	Maharashtra	Hindu	other	Pune	Homebound	0
4	Maharashtra	Hindu	other	Pune	Construction	95
5	Maharashtra	Hindu	SC	Pune	Homebound	0
6	Maharashtra	Hindu	SC	Pune	Homebound	0
7	Maharashtra	Hindu	SC	Pune	Clerical nec	35
8	Maharashtra	Hindu	other	Pune	Construction	95
9	Maharashtra	Hindu	other	Pune	Ag labour	63

	essential_worker	PublicTransport_Jobs	AdminUnitName
0	0	1	Ambegaon Dattanagar-Katraj Gaothan
1	1	1	Upper Super Indira Nagar
2	0	1	Vadgaon Dhayari-Vadgaon Budruk
3	0	1	Janata Vasahat-Dattawadi
4	0	1	Tadiwala Road-Sasoon Hospital
5	0	0	Erandwana-Happy Colony
6	0	1	Savitribai Fule University-Wakadewadi
7	0	1	Deccan Gymkhana-Model Colony
8	0	1	Sahakar Nagar-Padmavati
9	0	1	Salisbury Park-Maharshi Nagar

	AdminUnitLatitude	AdminUnitLongitude	H_Lat	H_Lon	HHID
0	18.44944	73.84900	18.44485	73.84942	99800000013
1	18.46199	73.86911	18.47127	73.87378	99800000027
2	18.46598	73.82463	18.47746	73.82425	99800000031
3	18.49020	73.83970	18.49854	73.84571	99800000046
4	18.52606	73.87160	18.52639	73.87114	99800000051
5	18.49625	73.81471	18.50547	73.82750	99800000055
6	18.55236	73.82666	18.55333	73.82133	99800000058
7	18.51845	73.83391	18.53135	73.83434	99800000059
8	18.47564	73.85440	18.47696	73.85511	99800000060
9	18.49162	73.86610	18.50113	73.87002	99800000071

	Agent_ID	WorkPlaceID	W_Lat	W_Lon	school_id	school_lat	school_long	public_place_id	public_place_lat	public_place_long	Adherence_to_Intervention
0	99800000000	2001000001258	18.52211	73.85910	0	NaN	NaN	3001000000170	18.45862	73.85568	1.0
1	99800000001	2001000010579	18.57208	73.77733	0	NaN	NaN	3001000001400	18.46185	73.88527	1.0
2	99800000002	2001000009601	18.48865	73.80390	0	NaN	NaN	3001000001329	18.52939	73.84523	1.0
3	99800000003	0	NaN	NaN	0	NaN	NaN	3001000000193	18.50061	73.86271	1.0
4	99800000004	2001000003138	18.49784	73.86794	0	NaN	NaN	3001000000444	18.49296	73.92796	1.0
5	99800000005	0	NaN	NaN	0	NaN	NaN	3001000000546	18.51937	73.85556	1.0
6	99800000006	0	NaN	NaN	0	NaN	NaN	3001000000975	18.54279	73.81889	1.0
7	99800000007	2001000006776	18.51390	73.86056	0	NaN	NaN	3001000000733	18.53214	73.83162	1.0
8	99800000008	2001000000836	18.46321	73.86382	0	NaN	NaN	3001000000125	18.47872	73.87102	1.0
9	99800000009	2001000010221	18.51484	73.84733	0	NaN	NaN	3001000001375	18.51651	73.92492	1.0