Are generative models for synthetic populations enough?

Extended abstract

R. Daniel Jonsson

2024-01-30

In this paper we discuss how both established methods and new machine learning approaches to the process of creating synthetic populations solve part of the problem very well but fails to address a crucial policy relevant part. We briefly outline the state of the art and touch on some recent research of our own to establish what existing methods do well. In the extended abstract we then argue that these methods is less help than we might think when it comes to analysing counterfactual och future scenarios. A forthcoming paper adds examples from computer simulations to establish this. This conference contribution can be viewed as a call for future research into the intersection of population synthesis, land-use and transport interaction, and planning decision support.

Introduction

The process of population synthetization is relatively common in practice, maybe more so than shows in the scientific literature. The increasing reliance on agent based simulation in transport policy analysis is one reason that many metropolitan areas are creating them. But even before agent based simulation was common it used in the United States because of the way census data was provided to research. Micro-data was and still is available (in the Public Use Microdata Sample, or PUMS) as a sample drawn from a city or county, while smaller census tracts or traffic analysis zones (TAZ) are reported as aggregates of a smaller set of variables, such as number of households, age distribution, etc. Population synthesis was one way of combining the information from these sources into something useful on the finest geograpical level.

Synthetic populations can be of use in countries even if registry data exists on the household or individual level. Data protection and privacy concerns play a role, such that even if we could use registry data we would not want to in planning. Registry data comes with a lot of reasonable restrictions, which in turn would restrict the use of transport models only to those who can get access to the data. That could pose a democratic problem if no one can check the results because of data protection. Furthermore, in many cases, we are not particularly interested in analysing only the present population. We often want to be able to analyse what would happen to some other, possible future, population. If we are to have any hope of being able to generate plausible future scenarios, we might as well use that method to generate a present population as well, since that also validates the method.

The next section will review four approaches to generate synthetic populations and set the foundation for the discussion and conclusions that follow. The review is necessarily short, see Fabrice Yaméogo et al. (2021) for a recent overview of different methods. We mostly use the same categorisation in this review. As the review shows, although new methods can help fix some issues, this is a mostly solved problem. In this paper we will argue that, although we have many methods that we can use to generate a present day population, we have not progressed quite as far when it comes to generating alternative populations that we can use to reason about counterfactuals and futures.

Methods

The problem

Every model and transport planning project is more or less unique in terms of what data is available, and in what variables the model requires. For the purposes of this paper we do not need to consider the full range of possibilities. Let's instead set up a simple representative problem with a few of the most common. For the purpose of describing the algorithms this is without loss of generality.

Imagine that we have a set of observed individuals representative of the population in the region we want to generate the population for (Table 1). We also have a description of each zone, but that does not necessarily contain all of the variables that we have in the sample.

Table 1: A sample of individuals.

Age	Income	Employment	Children	Car
23	3	Student	No	No
44	8	Fulltime	Yes	Yes
...	...	...	...	...
72	5	Retired	No	Yes

For each zone this becomes a 2D matrix with income group on one axis and employment status on the other. The task is to arrive at a population that as much as possible looks like the one in the sample, while obeying the constraints set by the zone data. Another way of thinking about it is that we want to find a multidimensional distribution of the four variables in the sample that simultaneously fits the distribution in the sample and the marginal distributions given by the zone constraints.

Table 2: Aggregate land use data giving information about zone marginals for two variables.

(a) Income data.

Zone	Income	Number
1	1	23
1	2	12
...	...	...
945	10	44

(b) Employment data.

Zone	Employment	Number
1	Student	5
1	Fulltime	34
...	...	...
945	Retired	6

Synthetic reconstruction

Some variation of synthetic reconstruction (SR) has long been the standard method for creating synthetic populations (e.g. Beckman et al., 1996; Müller, 2017). The general algorithm consists of two stages. The first is to find values for the elements in our 2D matrix that fit the marginal distributions, often by applying an iterative method where the elements of the matrix are adjusted to fit with one marginal distribution at the time until it eventually converges. [1]

The implementation details vary in different applications but the mathematics usually work out to something equivalent to finding the distribution that maximises the entropy (or minimises the information) given the marginal constraints. The second step is to use the elements of the matrix as weights on the corresponding observations in the sample. The matrix will provide the joint probability of someone being full time employed and in income group 7, which can then be used to draw a set of observations from the sample to put in that zone.

Combinatorial optimisation on population sample

Combinatorial optimisation (CO) takes a more brute force approach to the problem. In its simplest form we draw a random observation from the sample (Table 1), and assign it to a zone. If adding the observation improves the fit with the marginal constraints, we keep it. If not, we can check whether we can improve the fit by swapping out an existing observation. If neither improves the fit we move on and draw another observation and start over. The fit is defined as some distance between the zone marginals from Table 2 and what we get if we aggregate over the observations we have already assigned to the zone. Z scores have been shown to work well for this purpose (Huang and Williamson, 2001; Ryan et al., 2009).

This is a computationally intensive method, and the literature tends to think that it is only useful for smaller populations. Our own work shows, however, that it can readily be used for populations of millions. A key is that it is very easy to parallelize the process since each zone is independent from every other zone. It is possibly less efficent than

[1] The method, Iterative Proportional Fitting (IPF) by Deming and Stephan (1940) can be used for other similar tasks like matrix balancing too.

Generative machine learning

Generative machine learning learns an approximation of the multidimensional distribution that describes the households. After training we can draw new households from that distribution. There is, however, no guarantee that the households in a zone will respect aggregate constraints on income or age distribution if we draw from this distribution. Borysov et al. (2019) showed that Variational Autoencoders (VAE) performs better than other common methods when the number of dimensions in the sample is large. They also overcome the zero-cell problem that both SR and CO suffer from. That is, if some rare combination of variables is missing from the sample, it will be missing from the synthetic population too. SR and CO will, in other words, tend to inflate omissions that occur because it is a random sample and not the full population.

Meanwhile, research in Generative Adversarial Networks (GAN) has shown that they can work well for things that machine learning has had trouble with such as tabular (Xu et al., 2019) and missing data (Bernal, 2021; Lee et al., 2019). Xu et al. (2019) also develop a conditional GAN that can generate synthetic data conditional on one or more variables. In other work (Rastogi et al., 2024) we have investigated a combination of them in a deep generative model that learns the multidimensional distribution of a sample of individuals from a travel survey with missing data.

Combinatorial optimisation with generative ML

As noted by Fabrice Yaméogo et al. (2021), whether they are based on GAN, variational encoders, or some bayesian technique, generative models are not particularly well suited to respect zone constraints. The reason is that they are trained on the sample (Table 1), which does not contain information about the zones. The solution is simple. We can combine the generative model with either SR or CO in a straightforward way. In our research we combine the deep generative model from Rastogi et al. (2024) with combinatorial optimisation. It is a good fit because the generative model is a straightforward and easily implemented drop-in replacement. Instead of drawing candidate observations from the sample we can draw a generated synthetic observation instead.

Discussion

Increasing computing power and advances in generative machine learning have led to renewed interest in the generation of synthetic populations. The properties of households and individuals can be viewed as variables drawn from a multidimensional distribution and generative machine learning techniques are well suited to solve that particular problem.

What is much less discussed in the literature, however, is what to do if we want to create a counterfactual or alternative future population. The naïve approach is to change the zone marginals (e.g. to accomodate a population change to some future year) and then use the same sample [2] to fill in the population. But that comes with two related problems.

The first is that it is not straightforward to create new zone marginals. Lets say that we want to analyse a future where we have made some non-marginal change to housing policy and the tax code. It could easily mean that a completely different composition of households have the means to live in a zone than today, and that the present income distribution in the zone is a poor guide. We also have to be careful to not adjust marginals to implausible combinations. If we had a marginal distribution over single- and multi-family housing and changed that for some future year, we would have to take that into account when we adjust the income distribution as well.

The second problem is that we use the correlations in the sample (Table 1) to add information from it to each zone. The households or agents we add from the sample according to the joint probabilities that the marginals imply will bring with them the variables that we have no aggregate information about. So, in our example we would get an estimate of car ownership and children to the extent that they correlate with income and employment. But policies may well change those correlations. But, this problem seems easier to address because we can always break out variables we think will be affected by policy into a separate model. It is not uncommon to do that for car ownership, for instance.

These problems could possibly be solved by introducing a full land-use and transport interaction (LUTI) model (see Acheampong and Silva (2015) for a comprehensive review). But that comes with a host of other problems such as data requirements and building in many assumptions about how the transport and land-use system works. Having experience developing both LUTI models (Börjesson et al., 2014; Jonsson, 2008) and population synthetization tools it is our conclusion that we need to find a more balanced approach. Population synthetization builds in too much of our current patterns, while LUTI models builds in too many unverifiable assumptions.

Conclusion

Methods for generating a synthetic population that we can use in planning and agent based simulation of the transport system are well established and widely used. Recent advances in machine learning help overcome some known issues such as the zero-cell problem. In this paper we argue that while interesting and worth continuing development, they do not help solve the crucial question of what to do when we do not want to assume that the distribution of households will stay the same in the future.

Consider this a call for more research into how we can reason about future changes. We want the reasoning to be grounded in data and theory while avoiding locking in present patterns by assuming that things stay the same.

[2] For this discussion it does not matter if we draw from the sample directly or use a generative model

References

Acheampong R A, Silva E, 2015, "Land use–transport interaction modeling: A review of the literature and future research directions" *Journal of Transport and Land Use* **8**(3)

Beckman R J, Baggerly K A, McKay M D, 1996, "Creating synthetic baseline populations" *Transportation Research Part A: Policy and Practice* **30**(6) 415–429

Bernal E A, 2021, "Training Deep Generative Models in Highly Incomplete Data Scenarios with Prior Regularization", in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (IEEE, Nashville, TN, USA), pp 2631–2641

Börjesson M, Jonsson R D, Lundberg M, 2014, "An ex-post CBA for the Stockholm Metro" *Transportation Research Part A: Policy and Practice* **70** 135–148

Borysov S S, Rich J, Pereira F C, 2019, "How to generate micro-agents? A deep generative modeling approach to population synthesis" *Transportation Research Part C: Emerging Technologies* **106** 73–97

Deming W E, Stephan F F, 1940, "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

Fabrice Yaméogo B, Gastineau P, Hankach P, Vandanjon P-O, 2021, "Comparing Methods for Generating a Two-Layered Synthetic Population" *Transportation Research Record* **2675**(1) 136–147

Huang Z, Williamson P, 2001, "A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata", Department of Geography, University of Liverpool

Jonsson R D, 2008, "Analysing sustainability in a land-use and transport system" *Journal of Transport Geography* **16**(1) 28–41

Lee D, Kim J, Moon W-J, Ye J C, 2019, "CollaGAN : Collaborative GAN for Missing Image Data Imputation" *arXiv:1901.09764 [cs, stat]*, https://arxiv.org/abs/1901.09764

Müller K, 2017 *A generalized approach to population synthesis*, ETH Zurich

Rastogi T, Jonsson D, Karlström A, 2024, "Population Synthesis using Deep Generative Models"

Ryan J, Maoh H, Kanaroglou P, 2009, "Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms" *Geographical Analysis* **41**(2) 181–203

Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K, 2019, “Modeling Tabular data using Conditional GAN”, https://arxiv.org/abs/1907.00503