Population synthesis with regression trees

Kirill Müller[1] and Gunnar Flötteröd

Introduction

The scope of our work is the generation of a synthetic population for an urban study region. A synthetic population is a *statistically consistent* person-by-person representation of a real population, or a large sample thereof. In most cases, obtaining detailed survey data for the full population or for large samples is infeasible. Thus, the concrete challenge we are facing is to create a full synthetic population from a much smaller sample of the real population.

From a conceptual point of view, the generation of a synthetic population consists of estimating a population model and then simulating it to obtain the desired sampling fraction. This is true even for the procedures that expand a weighted sample to the desired size by duplicating observations: The estimation step consists of estimating the weights, and the simulation step is simply the expansion using these weights. Many procedures for generating a synthetic population have been presented in the recent past (e.g., Pritchard and Miller (2012), Ye et al. (2011), Auld and Mohammadian (2010), Guo and Bhat (2007)); many of these models are based on the work of Beckman et al. (1996), i.e., the simulation step is the expansion of weights, while the estimation step is often described in a procedural fashion. A notable exception is the recent work of Farooq et al. (2013) where conditional probabilities for all attributes are estimated and the simulation is performed by applying Gibbs sampling. Frazier and Alfons (2012) use multinomial logit models for generating a synthetic population by sequential imputation; this is somewhat similar to the approach discussed in this paper.

Assuming a given representative population sample $\mathbf{X}^{*} = (X_1^{*}, X_2^{*}, \quad X_{\ddot{n}}^{*})$, the joint distribution can be decomposed as follows:

$$P(\mathbf{X}) = P(X_1) \cdot \quad P(X_2 \mid X_1) \quad P(X_n \mid X_1, \ldots X_{n-1}).$$

By approximating each conditional distribution with a regression tree (CART, see Breiman et al. (1984)), an approximation of the full joint distribution can be obtained. A regression tree is a machine learning approach to classifying data through a sequence of partition steps (hence the notion of a "tree"). While CART models can be estimated also on continuous variables, the dependent variable is always categorical, and hence all attributes must be categorical. This does not pose a substantial restriction in practice, since any continuous variable can be made categorical through discretization (and in many cases even continuous data is available only in categories).

The CART-based sequence of regressions offers some advantages:

- CART trees can operate on missing values, which hence do not have to be treated in

[1] Corresponding author: kirill.mueller@ivt.baug.ethz.ch

advance and whose presence will not affect the estimation. (Missing values are more the rule than the exception in transport-related data.)

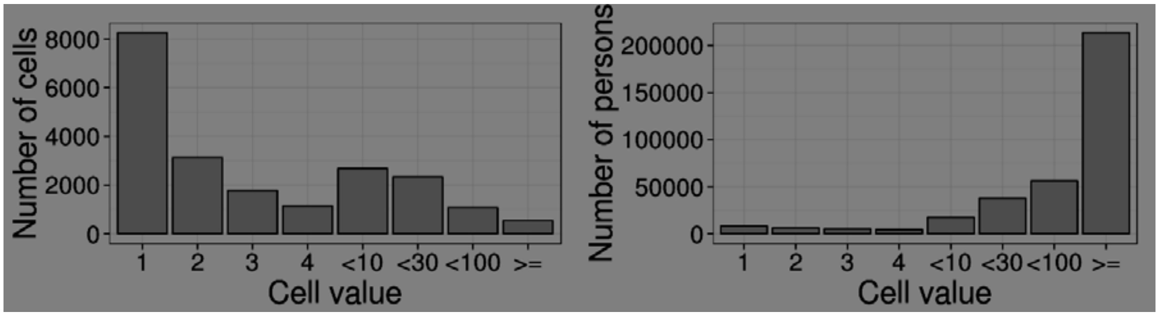
- The model is able to assign a nonzero probability to combinations of attributes that are not observed but still likely. This is interesting especially when considering datasets with many attributes where, by the curse of dimensionality, the sample can cover only a tiny subset of the possible combinations of attributes.
- The model can be set up and learned with very little effort, an open-source implementation of CART is available for the R platform for statistical computing. Simulation is straightforward. Furthermore, the likelihood for each combination of attributes can be computed exactly.

In this paper we investigate this approach to show its advantages and potential shortcomings. We report here excerpts from a more comprehensive study that would be presented in full length at the conference.

Results

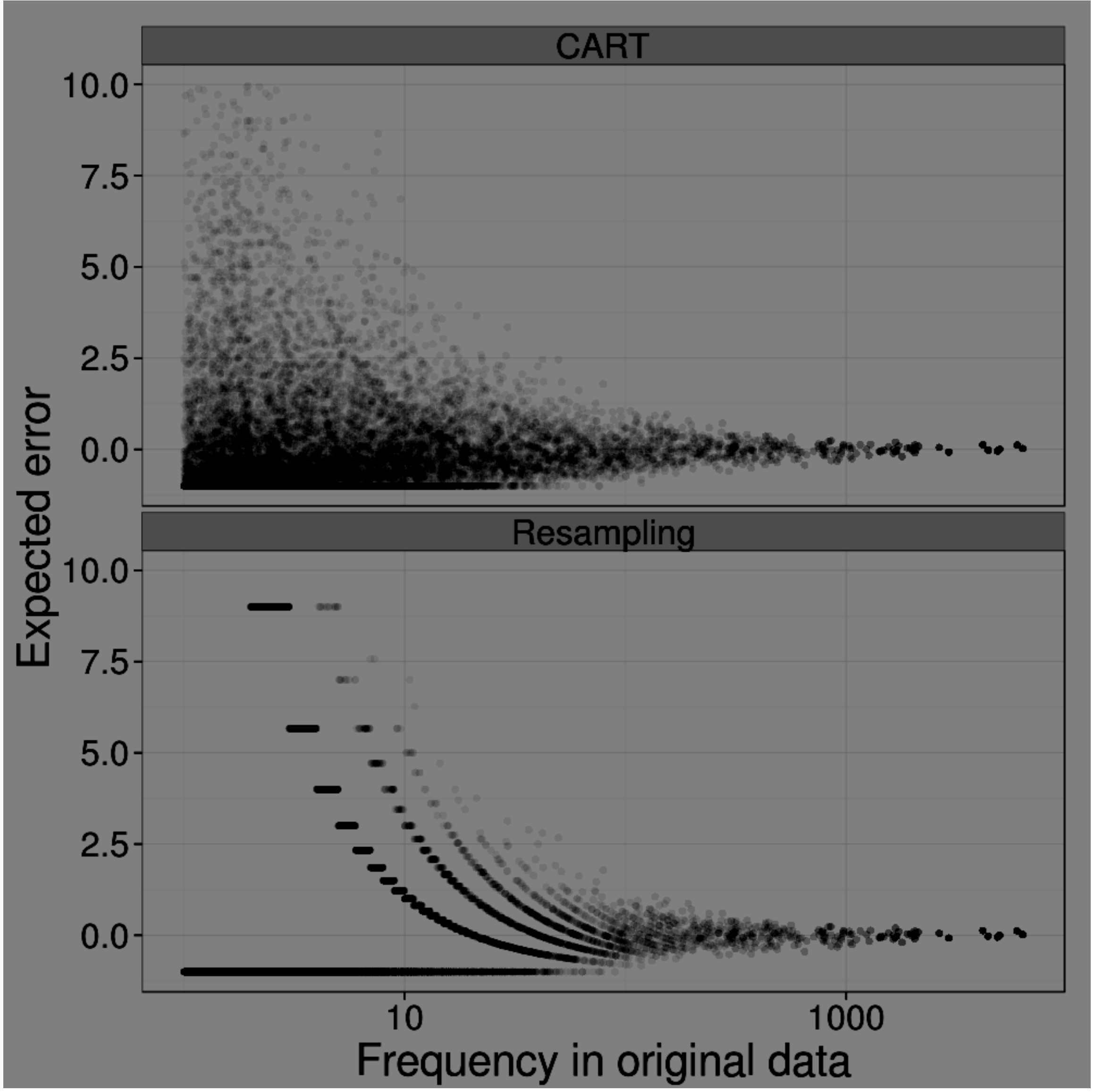
The Public Use Sample of the Swiss census contains categorical data on 5 % of the persons surveyed in the year 2000 census. It has been analyzed by Müller and Axhausen (2012). Almost all attributes contain a (sometimes substantial) fraction of missing values for purposes of anonymization. For simplicity, we use a version where missing values have been imputed. Furthermore, only 9 attributes are considered: Sex, Age, Nationality, Language, Marital status, Education, Work status, Workload, and Employment status.

Below an analysis of a cross-tabulation over all attributes is shown. This cross-tabulation is a 9-dimensional table where the cells correspond to unique combinations of characteristics, and their count represents the number of persons. While there are fairly many unique or rare *combinations of categories* (i.e., cells with a low value in the figure at the left), the main share of *observations* is obtained from cells with large values (figure at the right). This effect is mitigated to some extent when using more attributes.

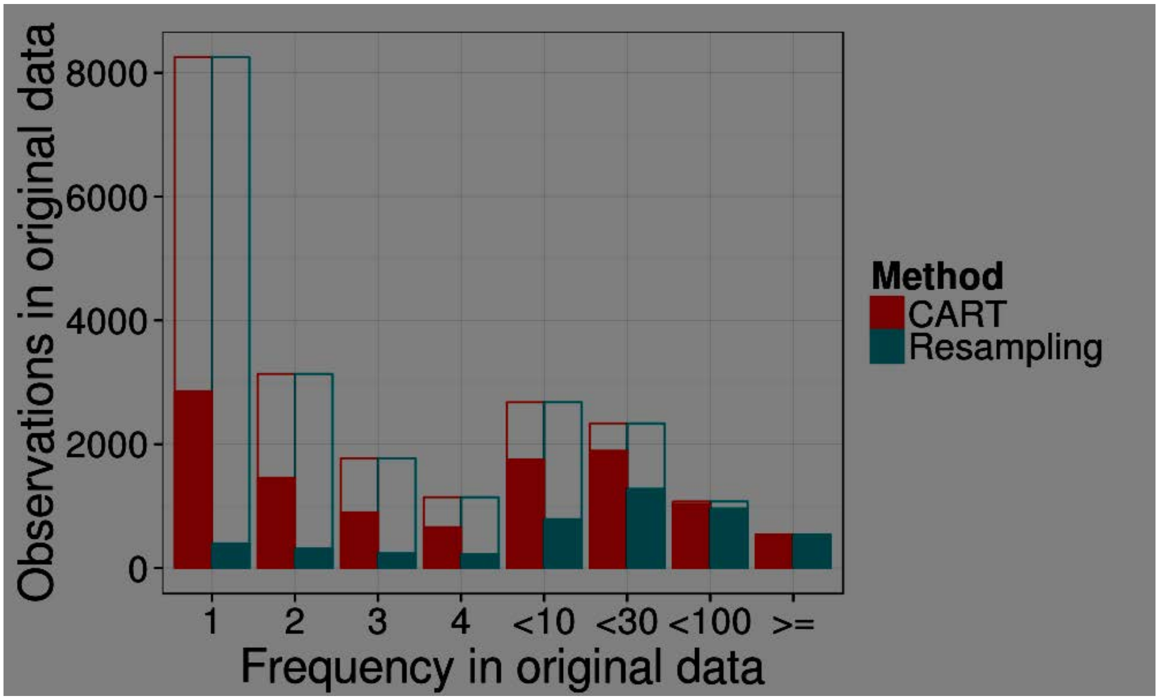


The model performance is assessed by drawing a 5% sample from the data and analyzing the likelihood of generating observations from the full dataset using the model. A single experiment is analyzed in some detail and compared to a naive resampling model.

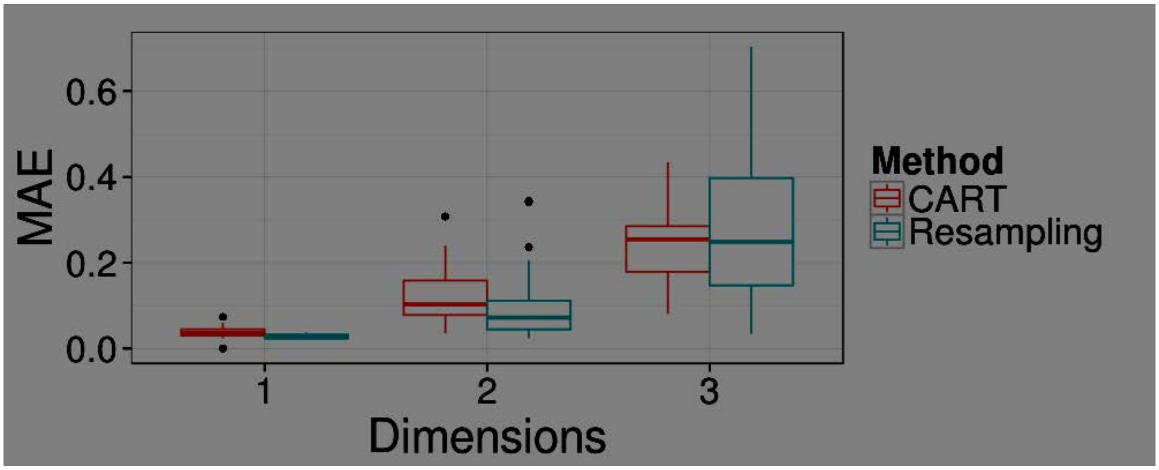
The next plot shows expected relative error against original frequency for each original observation. For both CART and resampling, this error can be substantial for rare observations, in some cases (see the concentration at -1) the models are is even incapable of replicating the input. This plot already shows the interpolating features of the regression-based model, compared to the resampling model which either heavily overestimates or just fails to replicate. The expected error was too large to fit on the plot for some of the observations.



Below a histogram of original observations is shown. The share of observations that can be generated by the model is filled. Many unique or rare observations cannot be generated at all, while almost all frequent observations can be reproduced. Again, this effect seems to be much stronger with the resampling model.



Finally, marginal distributions between observed and simulated data are compared. The next plot shows distributions of mean absolute error of all marginal distributions with one to three dimensions. It is defined as the mean of the relative error of differences in the cell values of the corresponding cross-classification tables. On average, both models perform comparably well; however the variance of the error is much larger – the error exceeds 1.0 for some attributes.



References

Auld, J. and A. K. Mohammadian (2010) Efficient methodol synthetic populations with multiple control levels, *Transportation Research Record*, **2183**, 19–28.

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

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen (1984) *Classification and regression trees*, CRC press.

Farooq, B., M. Bierlaire, R. Hurtubia and G. Flötteröd (2013) Simulation based synthesis of population.

Frazier, T. and A. Alfons (2012) Generating a close-to-reality synthetic population of ghana, *Open Access publications from Katholieke Universiteit Leuven*, Katholieke Universiteit Leuven.

Guo, J. Y. and C. R. Bhat (2007) Population synthesis for microsimulating travel behavior, *Transportation Research Record*, **2014** (12) 92–101.

Müller, K. and K. W. Axhausen (2011) Hierarchical IPF: Generating a synthetic population for Switzerland, paper presented at the *51st Congress of the European Regional Science Association*, Barcelona, September 2011.

Müller, K. and K. W. Axhausen (2012) Preparing the Swiss Public-Use Sample for generating a synthetic population of Switzerland, paper presented at the *12th Swiss Transport Research Conference*, Ascona, May 2012.

Pritchard, D. R. and E. J. Miller (2012) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously, *Transportation*, **39** (3) 685–704, May 2012.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana and P. A. Waddell (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.