Mitigating panel attrition with synthetic data

Statistical disclosure control for linked panel survey and register data

Maarten Koomen

Master thesis in Applied Data Science and Measurement

Supervisor: Jörg Drechsler   Examinator: Stefan Bender

Link to Github repository with R code.

15 12, 2022

Contents

1 Introduction

Attrition can be a significant problem for panel studies. Selectivity in who chooses to participate in a panel survey diminishes the representativeness of the survey data and, by extension, the general validity of statistical inferences. By combining panel survey and register data it is generally possible to fill in some of the gaps created by the explicit or implicit (e.g. nonresponse) refusal to participate in panel waves. Publishing such linked data without explicit consent would be a serious infringement on the respondents' privacy and a violation of most conventional data protection regulations. However, generating synthetic data based on linked data might offer an opportunity to make such sensitive information accessible for research, without infringing on the respondents' privacy.

In this paper, I will assess the feasibility of generating synthetic data from the Swiss education panel survey (TREE) linked with registry spell data on educational enrollment (LABB) from the Swiss Federal Statistics Office (FSO). Merging these two data sets combines the feature richness of survey data with the (theoretical) full coverage of the target population in the registry data. The goal is to combine these two data sources and generate synthetic data which are analytically comparable to the observed data without an unreasonable increase in the likelihood of disclosing any sensitive information where these data to be publicly disseminated.

I will start by giving a short overview of the history and practical applications of data synthesis as a method of statistical disclosure control for the dissemination of micro-data. In Section 2, I will discuss some metrics commonly used to measure the data utility and disclosure risk of synthetic data, and select some of them to fine-tune and assess the quality of the synthetic data. In Section 3, I will briefly describe the data collection process and main characteristics of both sets of observed data; the TREE panel survey and the LABB registry data. Finally, I will compare different synthesis configurations and use the data utility and disclosure risk metrics discussed in Section 2 to judge the performance of the final synthesis model.

2 Statistical disclosure control with synthetic data

Releasing micro-data to the public carries with it a, usually unknown, risk of information disclosure. Such a risk could entail that an attacker, i.e. someone intended on disclosing information, can determine that (i), a specific individual was in the original sample or (ii), that specific information can with high probability be linked to individuals. Increasingly inexpensive computing power has led to a greater demand by researchers for direct access to micro-data. This has created a need for statistical agencies and other data producers to find solutions to data dissemination that allow researchers to access sensitive micro-data whilst limiting the potential for privacy breaches. Not all such solutions focus on releasing micro-data to the public. Data centers can also set up on-site or online infrastructures where researchers are allowed to access the sensitive micro-data. In on-site infrastructures, researchers are typically allowed to analyze the data in a controlled environment and take with them only the aggregated results after they are checked for potential disclosure issues. This method is flexible but can be rather burdensome on both the data center and the researchers. The data center needs to provide isolated machines for the researchers and manually check any outputs they wish to take with them. In addition, the researchers have to physically travel to the data center, making any later discovered mistakes or additional need for analysis fairly costly. Online infrastructures give easier access to researchers and are potentially less costly for institutions. However, controlling the level of access and a priori defining how detailed output can be is challenging as it is difficult to assess how they affect the risk of disclosure.

An alternative to these solutions is releasing scientific-use-files that have been modified in some way by the data disseminator. Methods of modification depend on the type of data but they have broadly focused on information reduction, where information that poses a disclosure risk is suppressed in some way (e.g. top coding, rounding, and global re-coding), and data perturbation, where values are generally altered to improve data confidentiality (e.g. value swapping, noise addition, and data synthesis). A downside of many of these methods is that they often require specific knowledge of the modification process to properly analyze the altered data. Synthetic data as a method of statistical disclosure control offer a promising alternative because the data can be analysed by using relatively straightforward methods.

2.1 Synthetic data: synopsis of methodology and practical applications

The idea of data synthesis as an approach to statistical disclosure control was first proposed by Rubin (1993) and Little (1993). These early contributions focused on using multiple imputation techniques to generate multiple synthetic records for non-observed units or replacing sensitive records with synthetic ones. In the computer science literature, the idea can be traced back to Liew et al. (1985) who discuss using a very similar approach as a method of data distortion. After these initial proposals, it took another ten years before the methodology was fully formalized.

The general idea behind generating synthetic data is that models are fitted to the original, observed data, then, random draws from the fitted models are used to replace the original data. Broadly speaking, there are two approaches; full- and partial data synthesis. As the name suggests, in fully synthetic data, all records and values are fully synthetic. In partial synthetic data, only some sensitive records and/or values are replaced with synthetic data, meaning some original records or values remain in the synthetic data. In theory, fully synthetic data should offer better privacy protection because they contain no (directly) observed information. The flip-side is that, compared to partial synthesis, the analytic validity of fully synthetic data depends more heavily on the specification of the synthesis models.

Raghunathan et al. (2003) and Reiter (2003) formulated the rules for achieving valid inferences from fully-, and partially synthetic data, respectively. These rules are a variant of the combining rules for multiple imputation techniques for nonresponse, differing slightly for full- and partial synthesis. Raab et al. (2016) further expanded these combining rules by adding a method to achieve valid inferences in cases where only one synthetic data set is generated. This can be helpful in cases where the data synthesis is computationally intensive (for example when synthesizing data for large samples of observed data) or if data publishers fear that releasing multiple synthetic data sets generates an unacceptable additive risk of disclosure.

In the social sciences, early practical adoptions of synthetic data have focused primarily on fitting parametric models. Over the years, these methods have been expanded by including modelling approaches based on machine learning (ML) algorithms and models that can account for the complex sampling structures of modern survey data. ML approaches are especially promising where higher order relationships between a large number of variables need to be modeled in the synthesis process. For example, parametric models that have dozens of categorical predictors might not converge and suffer from issues of multicollinearity or perfect prediction. ML approaches can be helpful in these cases as they are not affected by these problems and offer an automated way of modelling any relevant higher order relationships in the original data. A suit of different ML techniques have been tested as synthesizers over the years; Classification and Regression Trees (CART) (Reiter 2005), Random Forest (Caiola and Reiter 2010), Support Vector Machines (Drechsler 2010), and Generative Adversarial Networks (GANs) (Choi et al. 2017; Park et al. 2018). Comparing some of these ML methods, Drechsler and Reiter (2011) and Little et al. (2021) show that CART models generally outperform some of the newer ML algorithms when used as a data synthesizer.

In addition to selecting a parametric or ML data synthesizer, another choice in the data synthesis process is the use of either sequential or joint modelling. Joint modelling aims to directly specify the joint distribution of the original data and generate synthetic data by drawing values from this joint distribution (see Schafer 1997 for a detailed discussion on joint modelling for the multivariate normal- and log-linear model). Valid synthetic data can be easily generated if the joint distribution of the original data is correctly specified. However, given the complexity of real-world data, it is often difficult to correctly identify the joint distribution. This is especially true if the data consist of both continuous and categorical variables, a common characteristic of social science micro-data. An alternative to joint modelling is synthesizing variables in sequence where each variable is synthesized by using as predictors only those variables that have already been synthesized previously, plus, in the case of partial synthesis, any variables that remain unchanged in the final data set. The assumption of sequential modelling is that the underlying joint distribution of a particular set of variables can be represented by the product of their conditional univariate distributions. This approach is very flexible since it allows each variable to be modeled separately, given the option of using either parametric or ML methods.

The earliest real-world application of synthetic data is from 1997, when the U.S. Federal Reserve Board decided to synthesize certain information at high risk of disclosure in the Survey of Consumer Finances (Kennickell 1997). An early example of the usefulness of synthetic data in the disclosure control of linked data is given by Abowd and Woodcock (2001), who generate a synthetic data based on data from the French National Institute of Statistics and Economic Studies (INSEE). Their goal was primarily to reduce the disclosure risks when linking data from several different official registers. To date, the most complex and extensive linked synthetic data has been released by the U.S. Census Bureau (Abowd et al. (2006); Benedetto et al. (2018)). This data contains synthesized records based on linked data from the Survey of Income Program Participation (SIPP), the Social Security Administration, and the Internal Revenue Service. This longitudinal data contains over 600 variables, almost all of which have been synthesized. In Europe, several data centers have started to produce similar statistical products. The German Institute for Employment Research (IAB) released a partial synthetic data in 2011 that was based on one wave of its Establishment Panel (Drechsler 2012). The Scottish Longitudinal Study (SLS) has used a tailor-made synthetic data approach to grant access to census data linked with sensitive records from health and death registers (Nowok et al. (2017)). In 2015, Eurostat published a synthetic version of the EU Statistics on Income and Living Conditions (EU-SILC) (de Wolf 2015). The synthetic EU-SILC data are not designed to lead to valid inferences, instead, they facilitate a way for researchers to access sensitive micro-data and prepare their analysis while they request full access to the original data.

2.2 Assessing the quality of synthetic data

Generating and publishing synthetic data is only useful if both the data disseminator and the research community can be equally convinced of its utility as a proxy for the original data, and of the effectiveness of the synthesis process in reducing the risk of disclosure. These two characteristics of synthetic data are logically opposed, i.e. synthetic data that can in no way be discriminated from its original base would offer zero additional protection against re-identification of individuals or the disclosure of sensitive information. A broad range of formal tests for both the utility and disclosure risk of synthetic data have been proposed. In the following two subsections, I will introduce some of the most prevalent of these metrics and select three that I will use in the remainder of the paper.

2.2.1 Measuring data utility

Data utility measures can roughly be grouped into three broad categories. First, synthetic data is usually checked on its general consistency and distribution. These checks, which can be labelled as fit-for-purpose measurements, are aimed towards comparing marginal or conditional distributions and checking whether values in the synthetic records are in themselves plausible (e.g. no negative number of children) and conditionally plausible (e.g. no unemployment with full-time job). These checks are generally not designed to illustrate the utility of the data; synthetic data with a one-to-one marginal distribution to the original data might still lead to invalid inferences in more complex multivariate analyses. However, they are a first check to see if there are underlying issues with the synthesis models. For this paper, I will limit these checks to comparing the marginal distributions of the original and synthetic variables and checking for implausible values in the synthetic data.

A second approach, often classified as global or general utility metrics, aims to test the utility of synthetic data by making general, formal, comparisons between the synthetic and original data. Many global measures utilize some distance metric to compare the synthetic and original data, such as the Kullback-Lieber divergence (Karr et al. 2006) or the Hellinger distance (Gomatam and Karr 2003). Another technique that has gained in popularity is based on the literature of propensity score matching (Rosenbaum and Rubin 1983). In this approach, the synthetic and original data are stacked into a single file, then, the probability for each row being a synthetic row is calculated. The closer the synthetic data are to the original data, the harder it would be for the propensity model to distinguish between synthetic and original records. Several metrics can be used to evaluate the difference between propensity scores, such as the Kolmogorov-Smirnov distance (Bowen et al. 2021) or the propensity score mean squared error ($pMSE$) (Woo et al. 2009; Snoke et al. 2017). The $pMSE$ has become a particular popular global utility measure in recent years, it is defined as:

$$pMSE = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2 \tag{1}$$

Where $\hat{p_i}$ is the estimated propensity for record i in the stacked synthetic and original data ($N = n_{org} + n_{syn}$) to be part of n_{syn}, with $c = n_{syn}/N$. Smaller $pMSE$ values indicate higher analytic validity for the synthetic data as it is harder for a propensity model to distinguish between observed and synthetic records in the stacked data set. A downside to this approach is that the $pMSE$ generally increases with the number of predictors in the synthesis model. To overcome this, Snoke et al. (2017) define a standardized $pMSE$ by subtracting the $pMSE$ with its expected value under a null model (a correctly specified synthesis) divided by its standard deviation. The null $pMSE$ is distributed as a multiple of a chi-squared distribution with $(k-1)$ degrees of freedom with the expected value and standard deviation defined as:

$$E[pMSE] = (k-1)(\frac{n_{org}}{N})^2(\frac{n_{syn}}{N})/N = (k-1)(1-c)^2c/N \tag{2}$$

$$StDev(pMSE) = \sqrt{2(k-1)}(\frac{n_{org}}{N})^2(\frac{n_{syn}}{N})/N = \sqrt{2(k-1)}(1-c)^2c/N \tag{3}$$

As a global utility measure, the standardized propensity score mean squared error (S_pMSE) offers a clear interpretation and straightforward comparison between different synthesis models. For the remainder of the paper, I will use the S_pMSE to fine-tune the synthesis parameters in Section 4. One shortcoming of the propensity score measure is that it is dependent on the model specified to calculate the propensity scores. A common traditional approach is to fit a Logit model that uses all variables in the data as estimators. I will additionally look at CART models to calculate the propensity scores, both as robustness check and because CART models can often be more efficient in automatically detecting relevant interactions and high order terms.

The advantage of global utility measures is that ther is no need to know a priori how synthetic data will be analyzed. However, there is no guarantee that synthetic data with a high global utility are in fact suitable for any one specific analysis. A third and last class of measurements therefore focuses on outcome-specific utility metrics, where specific analyses on the synthetic and original data are run and compared in parallel. Point estimates (means, regression coefficients) obtained from the synthetic and original data can be plotted against each other. If the synthetic data have a high outcome utility, these point estimates should cluster around a diagonal with a gradient of one. However, this approach does not account for any differences caused by the inherent uncertainty of estimation. In some situations, deviating point estimates obtained from the synthetic and original data could simply be an artifact of a large sampling error. A widely used metric that can take this estimation uncertainty into account is based on the confidence intervals (CI) overlaps of point estimates (Karr et al. 2006). The CI overlap J is defined as the average relative overlap for point estimate k as:

$$J_k = \frac{1}{2}[\frac{U_{over,k} - L_{over,k}}{U_{org,k} - L_{org,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}] \tag{4}$$

Where U_{org}, L_{org} and U_{syn}, L_{syn} are the upper- and lower bounds of the CI on the original and the synthetic data, respectively, and U_{over} and L_{over} are the upper- and lower bounds of the overlapping sections of the CI estimates on the original and synthetic. In Section 4.3, I will use this method to asses the specific utility of the generated synthetic data sets.

2.2.2 Measuring disclosure risk

There is a general distinction between measuring the disclosure risk of full- or partial synthetic data. With partial synthesis, at least some records or values remain unchanged and there is usually a one-to-one relation between the synthetic and original data. With full synthesis, this one-to-one relation does not exist and the original and synthetic data can be of different sizes. However, even though no original information remains in a fully synthetic data set, this does not mean that it cannot be used to learn about and disclose information contained in the original data. For example, Stadler et al. (2022) show how prior knowledge about the true values of some target records can be used in combination with information on the synthesis process to determine whether specific units are contained in the original data. Another approach proposed by Taub and Elliot (2019) uses a measure that directly matches cases in the original and synthetic data based on a defined set of key variables, and then calculates the disclosure risk of specific information in target variables. Specifically, the authors calculate a *Targeted Correct Attribution Probability* (TCAP) by finding records in the synthetic data for which a specific combination of *key* variables can uniquely identify a set of *target* variables, and then calculating the probability that those records can disclose the true value of the *target* variables in the original data. Let d_o be the original data and K_o and T_o vectors for the key and target variables so that:

$$d_o = \{K_o, T_o\} \tag{5}$$

With the synthetic data d_s being:

$$d_s = \{K_s, T_s\} \tag{6}$$

First, the *Within Equivalence Class Attribution Probability* (WEAP) is calculated:

$$WEAP_{s,j} = Pr(T_{s,j}|K_{s,j}) = \frac{\sum_{i=1}^{n}[T_{s,i} = T_{s,j}, K_{s,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{s,i} = K_{s,j}]} \tag{7}$$

The WEAP score for record j is the probability of the target variables conditional on the key variables. The square brackets are Iverson brackets (1 if condition is true, 0 otherwise) and n is the number of records in the synthetic data. The aim is to retain records with a high WEAP score since those synthetic records would be most helpful in trying to guess the true values of the target variables in the original data. For rows with a high WEAP score (for example WEAP=1), the TCAP for record j in the synthetic data is given by:

$$TCAP_{s,j} = Pr(T_{s,j}|K_{s,j})_o = \frac{\sum_{i=1}^{n}[T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j}]}{\sum_{i=1}^{n}[K_{o,i} = K_{s,j}]} \tag{8}$$

For synthetic records j that have no counterpart in the original data d_o matched on the key variables, the denominator in Equation 8 would be zero and the TCAP would be undefined. For records that do match, the TCAP score lies somewhere between 0 and 1, where a score of 0 would mean there is no disclosure risk for these records and a value of 1 would indicate that, for records with a high WEAP score, there is a considerable risk to disclose the true values of the target variables if the synthetic data were to be made public. The TCAP measure seems highly appropriate for assessing the added disclosure risk of linked data and I will use it in Section 4.4 to assess the disclosure risk of the synthesized data.

3 Observed data

The main goal of this paper is to specify a synthesis model that can be used to generate scientific-use-files from data that would otherwise be too sensitive to be published. Specifically, I am interested in linking survey data from the Swiss education panel (TREE) with register data on educational enrollment (LABB) from the Swiss Federal Statistical Office. The main idea is that this approach could facilitate the creation of a data set with the full TREE baseline sample of +20'000 participants who can be followed throughout their educational careers. This merger would reap the benefits of having data that is rich in features (TREE) and data that has a high degree of coverage (LABB) of the target population. In the next two subsections, I will briefly describe some important design features of both data sources.

3.1 The TREE and LABB data

The Transitions from Education to Employment (TREE) panel is a multi-cohort, multidisciplinary longitudinal large scale survey providing high-quality data on educational and occupational pathways in Switzerland. The target population are school leavers who are first surveyed at the end of compulsory school at the age of approximately 15 to 16 (see Hupka-Brunner et al. (2021) for more details). For the purpose of this paper, I will focus on the second TREE cohort who left compulsory school and were first surveyed in 2016. The baseline survey wave for the second TREE cohort contains detailed student and parental background characteristics and measurements of cognitive skills. The TREE survey was part of the Assessment of the Attainment of Educational Standards (AES), a national monitoring scheme designed to capture student skills in mathematics at the end of lower-secondary education in Switzerland. In 2016, the AES was designed as a compulsory, cross-sectional in-school assessment, carried out under the responsibility of the Swiss Conference of Cantonal Ministers of Education (EDK/CDIP). Because it was compulsory, the response rate of the baseline survey was close to a hundred percent. The TREE survey gathered explicit consent from the AES respondents to link their data with any further collected longitudinal information (with an overwhelming majority (95%) giving their approval). Subsequent TREE survey waves continue to collect data on education and labor market trajectories, contextualized by a rich set of complementary information. An downside of these subsequent waves, at least in terms of responses, is that participation is voluntary and response rates for the first two baseline surveys were around 85 and 75 percent, respectively.

The *Längsschnittanalysen im Bildungsbereich* (LABB) data are micro-data on education trajectories. Started in 2014, the LABB combines data from several FSO registers, primarily from several education monitors (*Statistik der Lernenden* (SdL), *Statistiken der Abschlüsse* (SBA), *Statistik der beruflichen Grundbildung* (SBG), and the *Schweizerisches Hochschulinformationssystem* (SHIS)), in addition to data from the population register (*Statistik der Bevölkerung und der Haushalte* (STATPOP)) and the Swiss micro-census (*Strukturerhebung* (SE)). Similar to the TREE data, the LABB micro-data can be used to analyze transitions into post-compulsory, upper-secondary education. Unlike the TREE data, however, the LABB micro-data is limited to educational programs that last at least one full-time semester. It does not contain information on some intermediate educational options in the post-compulsory transition period like the Motivation Semester or the Au-pair or Foreign Language Stays (for more detail see: FSO (2021, 2022)). For all remaining formal educational programs, the LABB data provide population-level data that does not suffer from nonresponse biases that usually plague voluntary-based surveys like TREE.

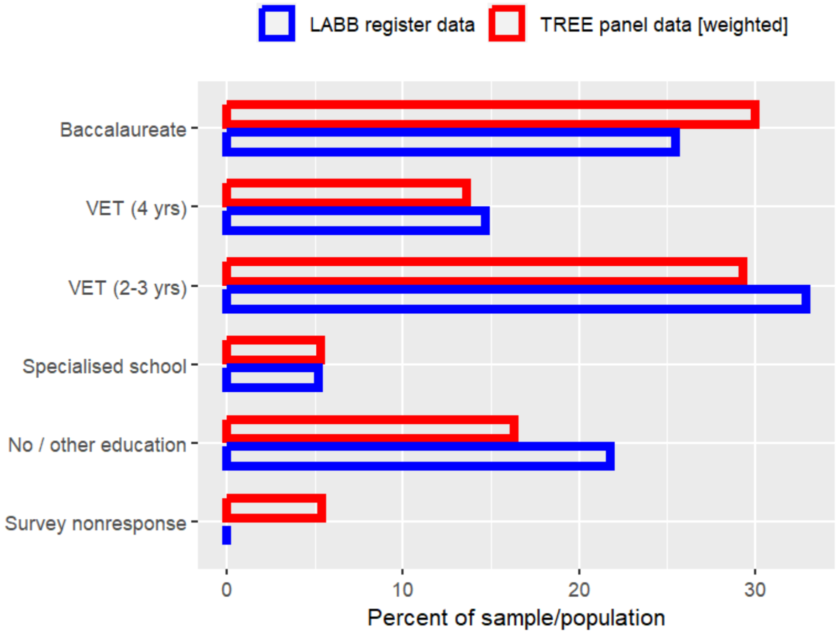

Figure 1: 2017 Enrollment of 2016 school-leaving cohort.

To illustrate how nonresponse impacts survey statistics in the TREE panel, Figure 1 shows the population and survey response percentages of educational enrollment in 2017 in the LABB register and TREE survey data, respectively. The TREE percentages are adjusted by their sampling design weights to represent population averages. In 2017, the first TREE survey wave had around five percent nonresponse. Among respondents, students that enroll into a baccalaureate-type of upper-secondary education (high-school, college, etc.) are over-represented. The LABB figures represent the true population enrollment distribution. In comparing both data sources, respondents in the TREE panel are biased towards students in higher academic tracks and under-represented by students in vocational- or other educational tracks. The LABB register data can be used to understand this bias and, ideally, to fill in some of the gaps in the TREE data. However, it would be unethical to published linked data that ignores the explicit or implicit refusal to panel participation. The aim of this paper is therefore to see if it would be feasible, both from a data validity and privacy standpoint, to create a set of synthetic data that could be used as a publicly available scientific-use-file.

3.2 Variable selection: who accesses secondary education in Switzerland?

The TREE panel scientific-use-files contain hundreds of variables. Synthesizing all of these is firmly beyond the scope of this paper. As a feasibility study, I will limit myself to synthesizing a handful of variables that model the transition from lower-secondary to post-compulsory education in Switzerland. In the following sections I will therefore briefly introduce some relevant characteristics of the Swiss educational system and prior research findings into the determinants of educational outcomes.

The Swiss educational landscape is federalized with each canton having a high level of autonomy in setting its own educational policies. The Swiss system is characterized by early tracking with relative high levels of differentiation and stratification. Tracking generally starts in lower-secondary education around age 12, in one of five distinctive tracks, with three differentiated by their academic requirements (basic, intermediate, or advanced academic requirements), plus two tracks with integrated or no tracking on academic ability (SKBF 2018). Placing guidelines for lower-secondary tracks vary between cantons but they are most commonly based on prior scholastic performance measured by grades (Neuenschwander et al. 2012).

The main differentiation at the upper-secondary level is between vocational education and training (VET) and general education, mainly typified by baccalaureate type schools that grant direct access to university. VET tracks can be completed with an additional vocational baccalaureate that also grants access to tertiary education at universities of applied science. For both VET and general upper-secondary education, admission to the more academically demanding tracks is mainly based on prior school track and performance measured by grades (Buchmann et al. 2016). In French- and Italian-speaking cantons there generally is a higher share of youths that enroll for general (i.e. non-VET) upper-secondary education compared to the German-speaking regions (SKBF 2018). Scientific research into determinants of upper-secondary track placement have generally confirmed the importance of grades and early tracking (Baeriswyl et al. 2006; Neuenschwander and Malti 2009; Beck 2015). Compared to boys, girls are found to be slightly over-represented in tracks with higher academic requirements, mostly because of better performances measured by grades, but also because they have higher achievement motivation (Glauser 2015). Having a migration background increases the likelihood of attending a track with lower academic requirements, although there is considerable variation between different migrant origin groups (Beck 2015; Glauser 2015). Another common research finding is that there is a persistent influence of social background (e.g. parental socio-economic status or educational attainment) on upper-secondary track placement (Imdorf 2005; Neuenschwander and Malti 2009;

Hupka-Brunner et al. 2010; Falter and Wendelspiess 2011; Falter 2012; Combet 2013; Glauser 2015).

Drawing on these research findings, I select 15 variables for data synthesis from the linked TREE & LABB data. Table 1 lists those variables. First, I include personal background information on gender, language region, and immigration status. In the unweighted TREE data there is a slight bias towards female participation (54.5% females versus 45.5% males). The language region is coded as binary with the Italian language region being grouped together with French-speaking regions because of their low case counts (69.8% and 30.2% for German- and French/Italian language regions, respectively). Immigration status is a categorical variable with three levels; native Swiss (72.7%), second-generation immigrants (18.7%), and first-generation immigrants (9.3%). Second, I include information on the socio-economic background of respondents in the form of their parents' occupation, educational attainment, reading interest, educational aspirations (for the respondent), family affluence, and family wealth. Parental occupation is measured by the highest parental ISEI-08 code (Ganzeboom 2010). Parental education is measured by a variable with two categories; parents with secondary education or lower (62.3%), and parents with at least one completed tertiary education (40.7%). Reading interest is measured by a composite variable based on the reading interest of both parents (Sacchi and Krebs-Oesch 2021). The parental aspirations for the respondents' educational careers is measured by a categorical variable with four levels; tertiary (34.5%), vocational education (50.6%), compulsory school (1.4%), and no opinion (13.5%). Family affluence and wealth are both composite variables scaled on household possessions and spending patterns (for more details see Kunter and Weiss (2002); Hartley and Currie (2016); Hobza et al. (2017)). On early tracking, ability, and performance, I include the attended type of lower-secondary school, average school grade in the regional language, maths, and science classes, and the score on the compulsory standardized maths test (AES) taken in 2016. Lower-secondary school type is measured by the level of academic requirements for the attended school with four categories; high requirements (29.1%), advanced requirements (39.1%), basic or low requirements (29.7%), and no or alternative differentiation based on skill level (2.1%). The average school grade is the mean of grades in maths, natural science courses, and marks in the language of AES test (e.g. mostly equivalent to primary regional language, either German, French, or Italian). The maths test score is a weighted likelihood estimate based on the individual AES maths test items (for details on test design and item scaling see Angelone and Keller (2019)). Lastly, respondents' idealistic educational aspiration and their embodied cultural capital are added. The idealistic educational aspiration is measured as a categorical variable with three levels; compulsory education (0.5%), upper-secondary (42.5%), or tertiary (57%). Embodied cultural capital is a composite variable scaled on questions on behavioral and verbal skills (for details see Hupka-Brunner (2016); Sacchi and Krebs-Oesch (2021)).

Table 1: Selected synthesis variables

#	Variable	Measured in	Description	Values
1	sex	2016	Gender of respondent	1=Female, 2=Male
2	langreg	2016	Language region	1=German, 2=French/Italian
3	wlem	2016	Maths test: weighted likelihood estimates (WLE)	*Continuous*
4	marks	2016	Mean school marks (test-language/maths/science)	*Continuous*
5	ls_req	2016	Lower-secondary school requirements	1=High, 2=Advanced, 3=Basic/Low, 4=No/non-assignable
6	hisei08	2016	Highest parental occupational code (ISEI 08)	*Continuous*
7	pareduc	2016	Parents' highest educational attainment	0=Secondary or less, 1=Tertiary
8	immig	2016	Immigration status	1=Native, 2=Second generation, 3=First generation
9	wealth	2016	Household possessions: family wealth	*Continuous*
10	joyreadp	2016	Parental reading interest	*Continuous*
11	fas	2016	Family affluence scale	*Continuous*
12	inccap	2016	Embodied cultural capital	*Continuous*
13	aspmf	2016	Parents' educational aspirations	1=Tertiary, 2=VET, 3=Compulsory school, 4=No opinion
14	aspideal	2016	Student's idealistic educational aspirations	0=Compulsory school, 1=Upper-secondary, 2=Tertiary
15	us_enroll	2017	Upper-secondary: education enrollment	1=No/other education, 2=Bridge programme, 3=Upper-sec VET, 4=Upper-sec VET (Bac), 5=Upper-sec Academic Bac.

Besides forming a theoretically cohesive set, the variables in Table 1 are also selected because they do not contain any dependencies that might be difficult to reproduce in the data synthesis process. In principle, for these specific variables, there is no configuration of values which is inherently impossible. This will make the data synthesis process much more straightforward because there is no need to explicitly model dependencies between sets of variables. For example, if age and marital status would have been included, the data synthesis should take into account that no person under 18 could be legally married under Swiss law. In addition, missing values that are present in the observed data, will not be replaced before data synthesis, meaning that they will effectively count as extra categories and will be present in the synthetic data sets. This approach is useful if the goal of the synthetic data is to give researchers easy access to data that preserves the structure of the observed data as closely as possible, including information on missing patterns present in the original data.

4 Data synthesis

To generate the linked TREE and LABB synthetic data, I will focus on fitting CART and parametric models, ignoring some of newer ML algorithms. Further, I will only explore sequential synthesis, i.e. synthesizing each variable in sequence. An alternative to this approach would be to model the joined distribution of the entire data, something that is often prohibitive with real-world survey data, even with only a handful of variables. In the next two sections, I will explore the synthesis sequencing and the modelling of each variable individually to find parameters that generate synthetic data with high utility. For the data synthesis and most model evaluations, I will be using the R package *synthpop* (Nowok et al. 2016). This package was created as part of the SYLLS (Synthetic Data Estimation for UK Longitudinal Studies) project. The goal was to create a toolkit that can be used to quickly create bespoke synthetic data from sensitive micro-data that fits the particular needs of individual researchers. The package offers a range of tools that are useful for both the creation and the analysis of synthetic data. It should be noted that the initial idea behind the *synthpop* package is that the synthetic data generated are meant to be used as test data for researchers to explore and test their models on. Code developed on synthetic data should ultimately be run on the original data to verify results. The *synthpop* package is flexible, supporting the use of a range of parametric models and ML algorithms. In addition, the *synthpop* package can handle missing data in categorical and continuous variables. For missing values in categorical variables, this process is relatively straightforward since they are simply regarded as separate categories in the imputation process. For continuous variables, this process is a little more involved. First, an auxiliary categorical variable is created containing all parallel missing categories. Values in this auxiliary variable are synthesized by either a polytomous or CART model. The remaining, non-missing, values are fitted separately and then both these variables are used in the final synthesis and as a predictor for the synthesis of remaining variables.

4.1 Sequencing

A first step in finding a well-performing synthesis model for our observed data is to consider the order in which our variables are synthesized, which can have a significant impact on the utility of the synthetic data. A brute-force method for finding the optimal sequence of data synthesis is computationally prohibitive for data sets with even a handful of variables. For n variables a total of $n!$ synthesis sequences have to be compared. According to Drechsler and Reiter (2011), it can be computationally beneficial to place categorical variables at the end of the synthesis sequence, especially if they have many categories. Sub-setting numerical and categorical variables in the synthesis sequencing can significantly reduce the search grid for a brute-force approach. In addition, variables that follow a normal or binomial distribution would be easier to synthesize earlier in the sequence because, even with few predictors in the synthesis model, fairly accurate synthesis can be achieved by simply taking random draws from the normal or binomial distribution.

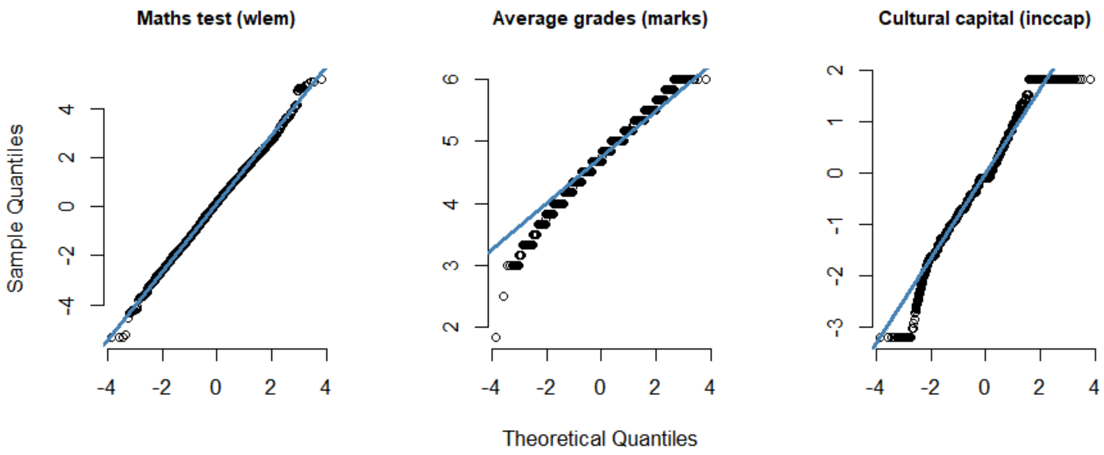

Figure 2: Normal Q-Q Plot.

All numeric variables in the observed data are composite variables that are based on sets of underlying categorical variables. In general, these variables are approximately normally distributed with some degree of variability. The variables *wlem*, *marks*, and *inccap* have distributions that are closest to normal. Figure 2 shows the normal quantile plots for these variables. Of these three, *wlem* is closest to having a normal distribution since the observed values lay close to the blue diagonal line that represent the expected values under normality. The variables *marks* and *inccap* are only quasi-normal in that their distributions have fat tails or that they are clustered around certain values. I group these three variables together under n_{norm}. The remaining numeric variables; *hisei08*, *wealth*, *joyreadp*, *fas* follow distributions that depart even furhter from a normal distribution with more skewness or clustering. I group these separately under n_{num}. The variables *sex*, *langreg*, and *pareduc* follow relatively symmetric binomial distributions and are grouped together under n_{bin}. Lastly, I create two separate groups for categorical variables with three levels (n_{cat3}) and one for categorical variables with four or more levels (n_{cat4}). This reduces the sequencing search space from $n!$ to $\prod(n_{norm}!, n_{bin}!, n_{num}!, n_{cat3}!, n_{cat4}!)$, or from 1.3 trillion combinations to $\prod(3!, 3!, 4!, 2!, 3!)$, a more manageable 10,368 combinations of synthesis visit sequences.

To find optimal synthesis sequences, I calculate and compare the standardized propensity mean-squared error (S_pMSE). For each synthesis sequence, I construct two synthetic data sets with *synthpop*, one using CART and one using parametric models for data synthesis. For each variable, the parametric synthesis models use all other, previously synthesized variables as predictors with no interactions or higher order terms. To compare the validity of the synthesis, I calculate the S_pMSE using both CART and Logit propensity score models. Figure 3 plots the S_pMSE for each synthesis sequence. The axes show the S_pMSE calculated with CART propensity score models for both

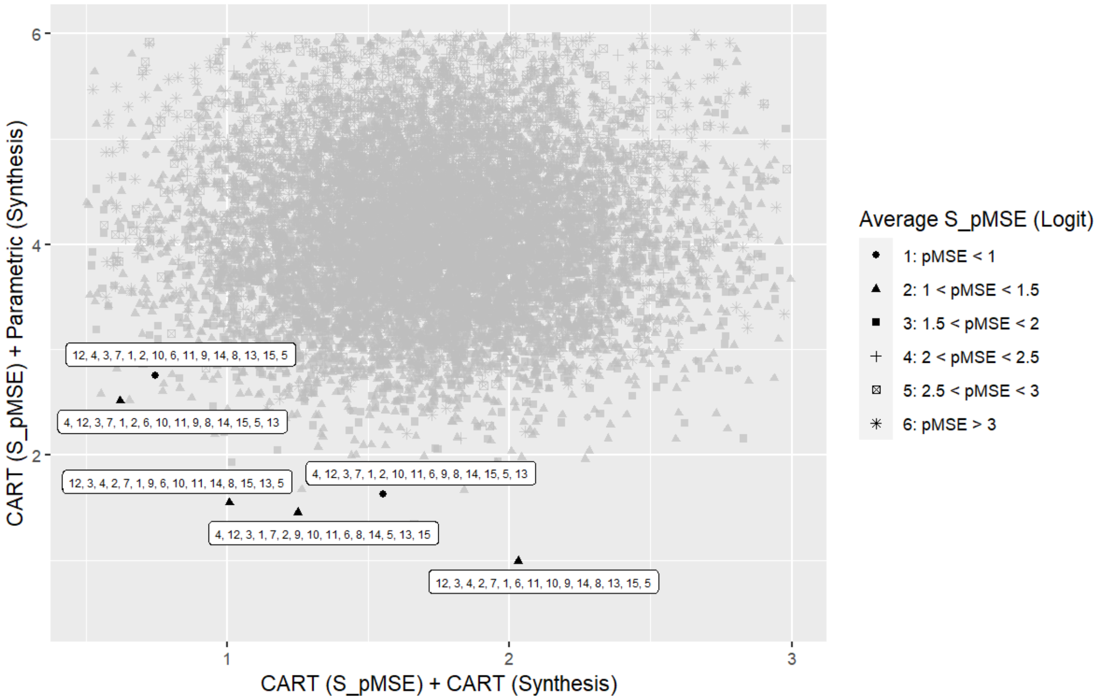

Figure 3: Propensity mean-squared error per synthesis visit sequence.

sets of synthetic data, one generated by CART models, and one by parametric models. To make sure that the low S_pMSE scores calculated by the CART propensity score models are robust, I also calculate the average S_pMSE by using a Logit propensity score model. The six points highlighted in Figure 3 have an average S_pMSE smaller than 1.5 (as calculated by both the CART and Logit propensity score models). I will use these six synthesis sequences to further fine-tune the synthesis models for each individual variable in the following section. The numbers in these sequences correspond to the order of the variable list in Table 1.

4.2 Modelling

For each of the six best variable visit sequences, Figure 4 plots the S_pMSE calculated with CART models per variable for both the CART and parametric data syntheses. Highlighted are sequences and models where the S_pMSE is below 0.8.

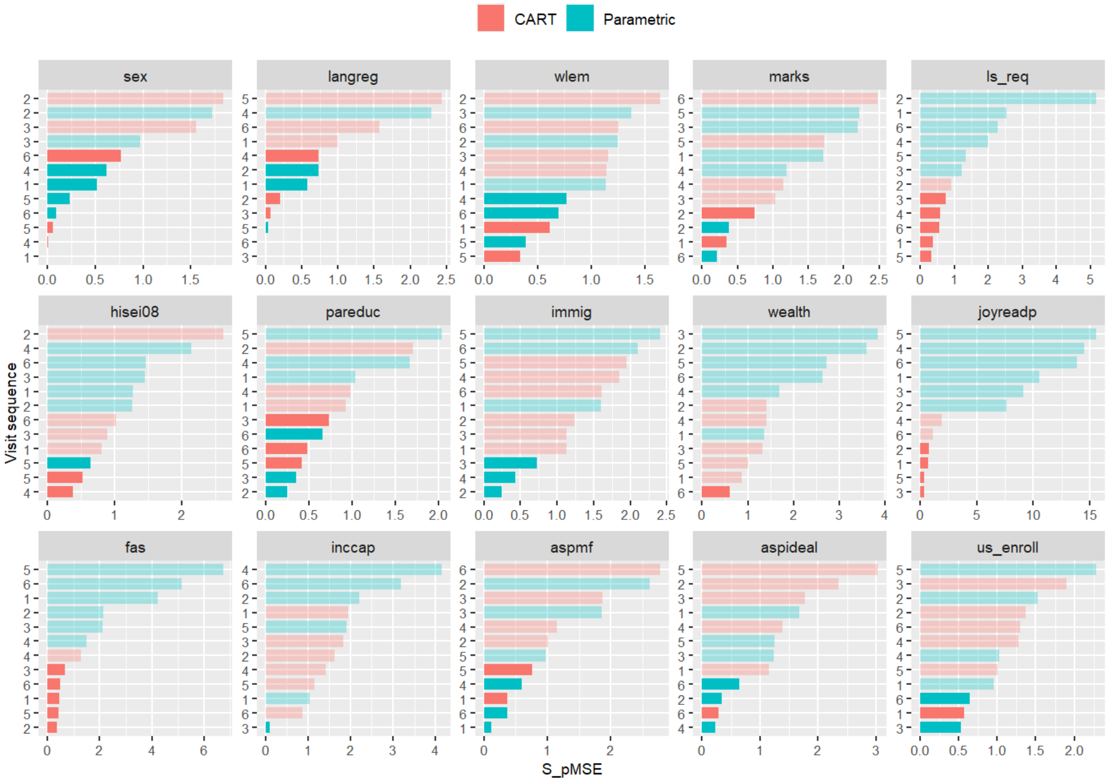

Figure 4: Variable utility with CART or Parametric synthesis for 6 best performing visit sequences

Each panel in Figure 4 contains the utility in S_pMSE for one of the 15 variables. The vertical axis list 12 synthesis runs (the best 6 sequences for both CART- and Parametric synthesis) and orders them by their S_pMSE to get a better idea if some variables consistently perform better with either CART or Parametric data synthesis. Most of the variables show a mixed picture where, depending on the sequencing, either CART or parametric modeling leads to the lowest S_pMSE. However, there are some variables for which it is relatively clear that CART synthesis outperforms parametric modeling. It should be noted that this comparison is only between CART and a relatively straightforward parametric model that includes all remaining variables as main effects without any interactions or higher order polynomials. The parametric modelling could be improved to better model the data generating process or, for continuous variables, some data transformation can be performed to better approximate normality and improve the fit of the linear synthesis models. For example, the variable *joyreadp* (parental reading interest) has a left-skewed distribution that is relatively poorly modeled by fitting a linear regression model. Figure 5 shows a density plot of the regression residuals for the parametric (linear) model for *joyreadp* that would hypothetically be used as a synthesizer. The left-skewness of the residuals indicates that the linear model is not a great fit for the observed data.

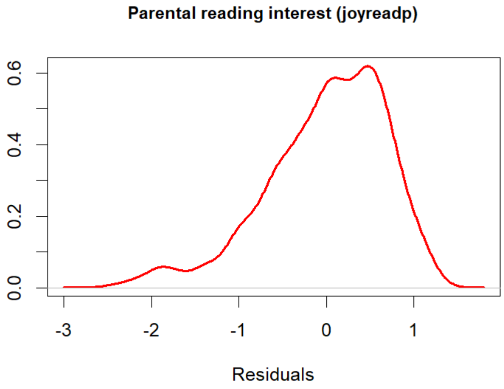

Figure 5: Regression residuals density plot.

To check whether some data transformation or more complex model could improve the fit of a linear model, I separately perform a square-root transformation on *joyreadp* and run a linear model that includes a range of interaction terms as well as second and third order polynomials for all continuous variables. The square-root transformation of *joyreadp* does improve the fit of the linear synthesis model, the residuals are smaller but the distribution of the residuals still exhibit the same skewness seen in Figure 5 (see Figure 11 in Appendix). In terms of expanding the linear model, adding interaction and higher order terms does not improve the model fit in terms of regression residuals, it only marginally increases the explained variance from 0.13 to 0.15 (McFadden adjusted R-squared). Even with these marginal improvements, it seems reasonable to only consider CART models for *joyreadp* since this would likely lead to synthetic data with better utility without the need for more complex modelling or data transformations.

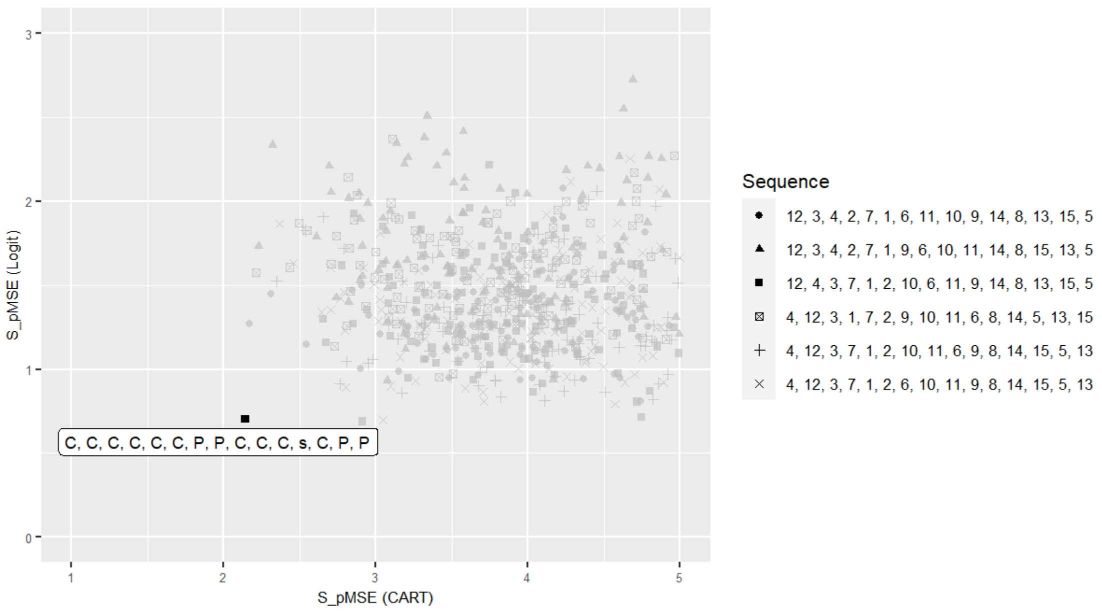

Figure 6: CART / parametric modeling per visit sequence.

For the six visit sequences, I test which combination of modelling will lead to the smallest S_pMSE. For the variables *sex, ls_req, hisei08, wealth, joyreadp,* and *fas* I will only consider CART modelling, for the variables *langreg, wlem, marks, pareduc,* *inccap, aspmf,* and *us_enroll* I will test both CART and parametric models, and for *immig,* and *aspideal* I will look at parametric models exclusively. Since only half of the 15 variables can vary between models, a total of $\prod_{n=1}^{7}(2!)$ or 128 different combinations per visit sequence have to be tested. Figure 6 shows the best performing modelling and variable visit sequences, with capital letters "C" and "P" standing for "CART" and "parametric", respectively, and lower case letter "s" standing for "sample" (i.e. the first variable in the synthesis sequence). Again, the numbers in the sequences correspond to the position of the variables in the data set (i.e. the variable numbers in Table 1). I will use these modelling parameters to run the final data synthesis:

Table 2: Synthesis sequence and modelling

Synthesis order	Variable name	Variable number	Variable description	Synthesis model
1	inccap	12	Embodied cultural capital	sample
2	marks	4	Mean school marks (test-language/maths/science)	CART
3	wlem	3	Maths test: weighted likelihood estimates (WLE)	CART
4	pareduc	7	Parents' highest educational attainment	parametric
5	langreg	2	Language region	CART
6	sex	1	Gender of respondent	CART
7	wealth	9	Household possessions: family wealth	CART
8	fas	11	Family affluence scale	CART
9	joyreadp	10	Parental reading interest	CART
10	hisei08	6	Highest parental occupational code (ISEI 08)	CART
11	aspideal	14	Student's idealistic educational aspirations	parametric
12	immig	8	Immigration status	parametric
13	us_enroll	15	Upper-secondary: education enrollment	parametric
14	ls_req	5	Lower-secondary school requirements	CART
15	aspmf	13	Parents' educational aspirations	CART

Only variables *pareduc*, *immig*, *aspideal*, and *us_enroll* will be synthesized by fitting parametric models. Of those, *immig*, *aspideal*, and *us_enroll* are categorical variables which will be modeled by fitting multinomial logit models. Variable *pareduc* would normally be dichotomous, however, since missing values are treated as separate categories in the synthesis process, this variable will also be treated as categorical and be synthesized by fitting a multinomial logit model. The variable *pareduc* is the fourth synthesized variable and the synthesis model only has three explanatory variables (*inccap*, *marks*, and *wlem*). The variables *immig*, *aspideal*, and *us_enroll* are all synthesized later in the sequence (12th, 11th, and 13th place, respectively) and contain all prior synthesized variables as predictors.

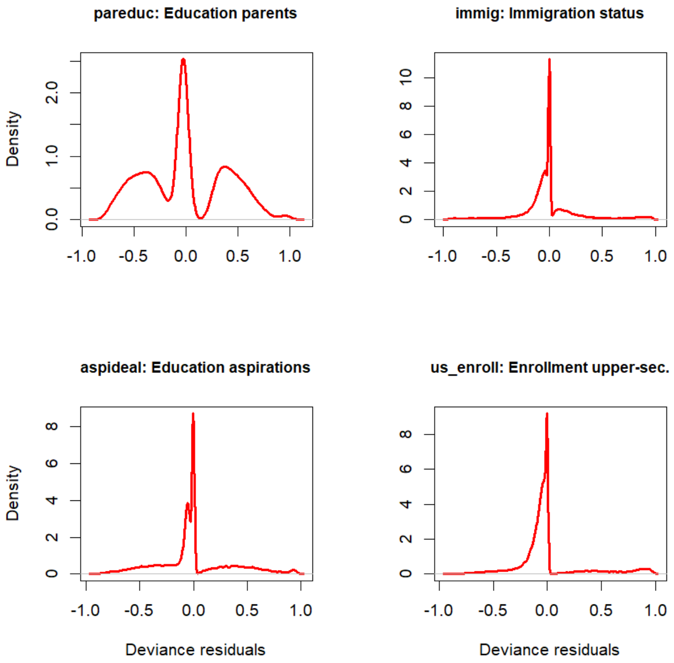

Figure 7: Deviance residuals for multinominal logit synthesis models.

To see how well these models describe the data, Figure 7 shows the density plots for the deviance regression residuals for *pareduc*, *immig*, *aspideal*, and *us_enroll*. Deviance residuals are a useful diagnostic tool in situations where model-fitting is achieved by maximum likelihood estimation (Pierce and Schafer 1986). The residuals are calculated by comparing the fitted values in the specified model versus a fully-saturated model that perfectly describes the observed data by including a separate parameter for each observation. The deviance residuals measure how much the estimated probabilities of the specified model differ from the estimated probabilities in the fully-saturated model. The regression models for *immig*, *aspideal*, and *us_enroll* seem to fit the data quite well since the residuals mostly spike around zero. The residuals for *pareduc* show that there are some more discrepancies between the observed and estimated values in the regression model but the residuals are generally still small.

4.3 Data utility

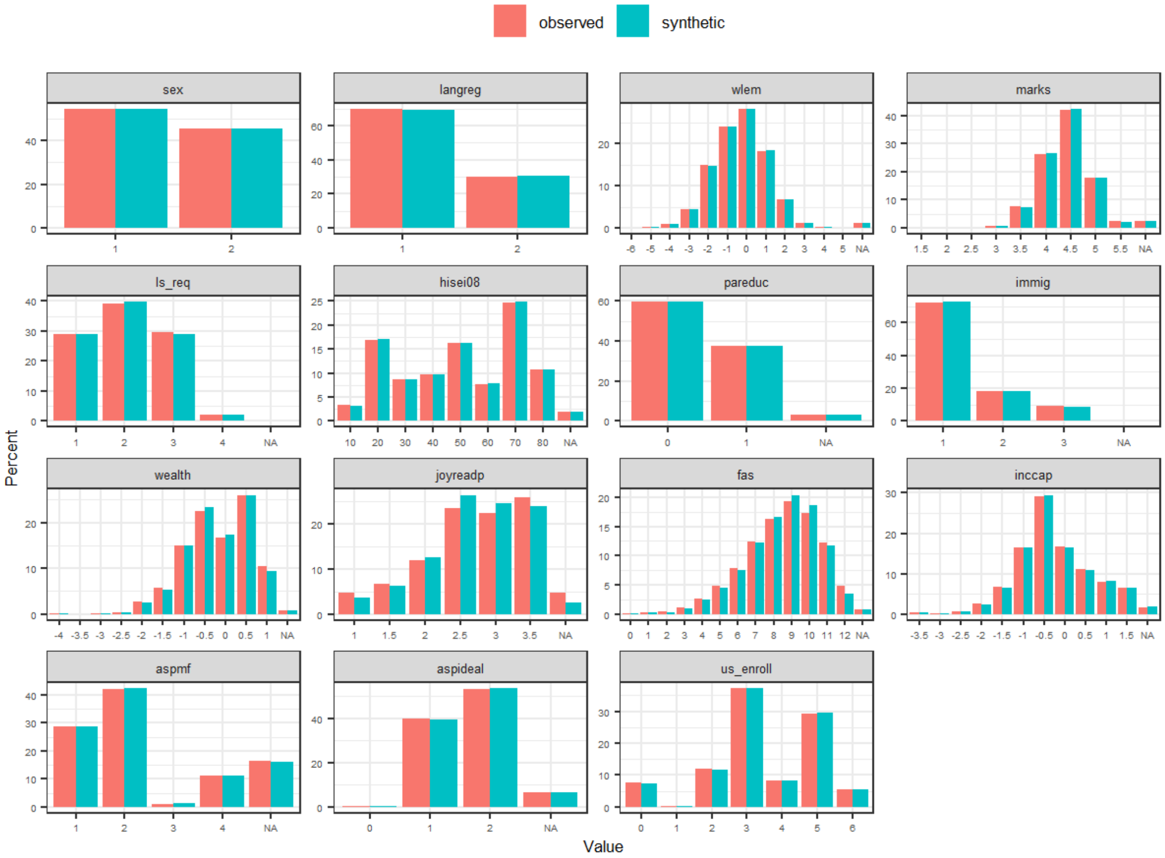

Figure 8: Marginal distributions for synthetic and observed data.

With the sequence and modelling parameters set in section 4.2, I generate five synthetic data sets using *synthpop* (Nowok et al. 2016). As mentioned previously, missing values in the observed data are retained in the synthetic data. Figure 8 plots the marginal distributions of each of the 15 variables in the synthetic and the observed data. The percentages for the synthetic data are based on the pooled sets of generated synthetic data. In the univariate comparison, the synthetic data generally follow the observed data fairly closely. Some more notable differences exist between the synthetic and observed data for the variables *joyreadp* and *fas*. The data synthesis for these variables could likely be improved by adding additional information to the data or tweaking the data and modelling. Both the marginal distributions for *joyreadp* and *fas* are somewhat skewed. Exploring other data transformation techniques or trying out other modelling approaches might improve the utility of the synthetic data, however, doing so is beyond the scope of this paper.

To look at the multivariate validity of the data, I run a set of regression models on the synthetic- and the observed data and compare the overlap of the confidence intervals (CI). For each variable, the regression model contains all remaining variables as predictors without any interactions or higher order terms. The variables *sex*, *langreg*, and *pareduc* are estimated by fitting a Logit model, the variables *wlem*, *marks*, *hisei08*, *wealth*, *joyreadp*, *fas*, and *inccap* by fitting a linear model, and the variables *ls_req*, *immig*, *aspmf*, *aspideal*, and *us_enroll* by fitting a multinomial logit model.

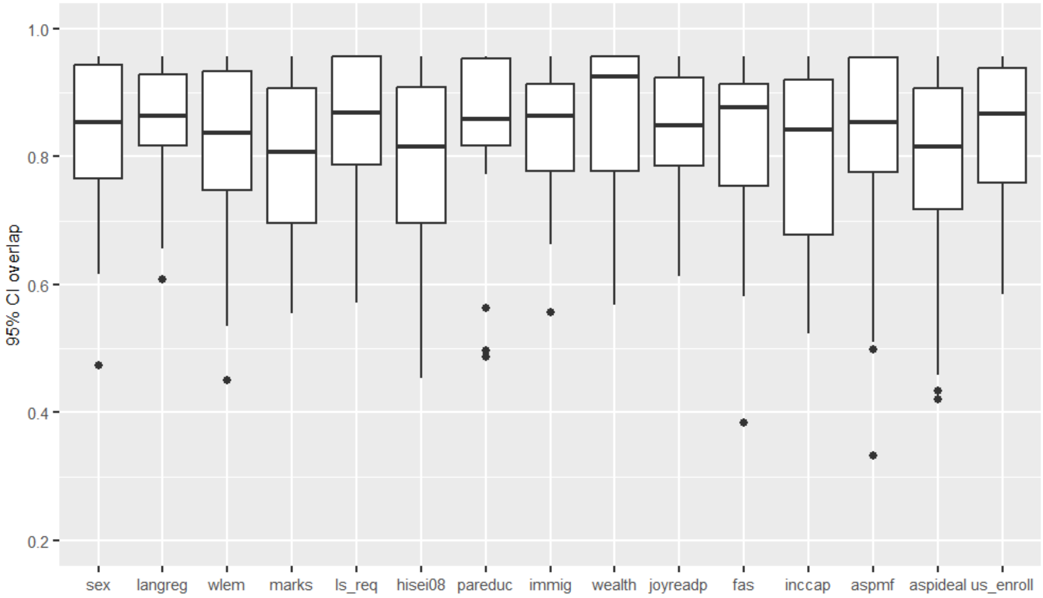

Figure 9: Box-plot for confidence interval overlaps per outcome variable

Figure 9 plots the distribution of the CI overlaps for all the regression coefficients in each of the 15 regression models. The median CI overlap generally lies around or above the 75 percent line. However, for some variables a significant number of coefficients overlap relatively poorly. Specifically, the tail ends of the CI-overlap distributions of *hisei08* and *aspideal* are rather long, indicating that, for these models, a larger share of the point estimates' CI fitted on the synthetic data deviate from those fitted on the original data. In some cases, this lack of overlap is partially driven by point estimates with a high *p*-value in both the regressions on the original and synthetic data (see Appendix for detailed regression outputs and CI comparison for *hisei08*). A further step that could be taken is fine-tuning the comparison for each variable by dropping these estimators from the regression models to check whether or not this improves the CI overlaps.

4.4 Disclosure risk

The appropriate measure of disclosure risk of disseminating synthetic data based on linked data from the TREE panel survey and the LABB register is identifying the added risk of disclosing information about the true value of educational outcomes, in particular for those who explicitly or implicitly refused to disclose this information in the TREE panel survey. For the calculation of the *Targeted Correct Attribution Probability* (TCAP), the target variable is therefore defined as the upper-secondary enrollment in 2017. As key variables, it seems reasonable to pick demographic and background information that might be available in other publicly available data sources. Therefore, I define vectors K and T as:

$$K = \{sex, langreg, ls\_req, pareduc, immig\} \tag{9}$$

$$T = \{us\_enroll\} \tag{10}$$

Figure 10 plots the percentage of rows in each of the five synthetic data sets that have a WEAP score of one (i.e. synthetic records for which the combination fo key variables can uniquely identify the target variables), plus, for those records, the mean TCAP score.

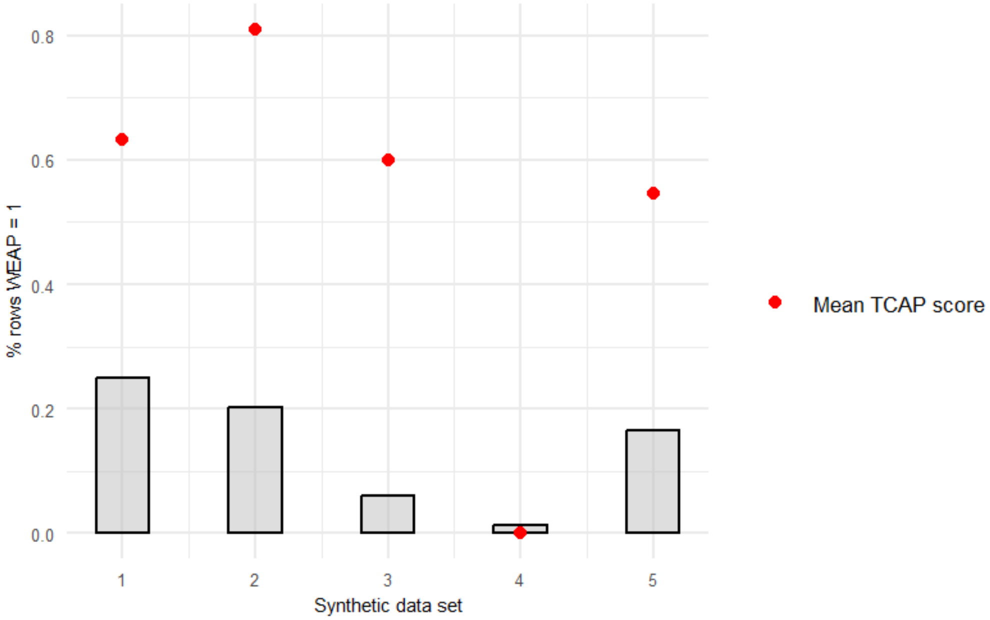

Figure 10: TCAP score per generated synthetic data set.

Figure 10 shows that the risk of disclosure fluctuates between the synthetic data sets. The share of rows in each synthetic data with a WEAP score of one (records for which the combination of key variables can uniquely identify the target variables) fluctuates between 0.25 and 0.01 percent. In data set four, only 0.01 percent of synthetic records have a WEAP score of one and the risk of disclosure, measured by TCAP, is zero. For the other data sets, the mean TCAP score fluctuates between 0.81 and 0.55. This means that for the rows with a WEAP score of one, not all of them pose an equal risk of disclosure with regard to the true value of the target variable in the original data. The closer the TCAP is to one, the higher the risk associated with publishing that particular data set. However, it is not fully clear how to interpret this risk. Even if rows can be identified in the synthetic data that have a low variability in the target variable given some unique configurations of the key variables, without prior knowledge, it would not be unclear to an attacker if this low variability in the target variable, given the key variables, is also present in the original data. However, if the disclosure of the target variable is deemed especially sensitive, to reduce the risk to an absolute minimum, only synthetic data sets with a TCAP score below a particular threshold could be published (e.g. with a max TCAP would be set at 0.6 only synthetic data sets 4 and 5 would be published). Limiting the amount of synthetic data to be released in this way would diminish the utility of the data since they would be more susceptible to misspecification of the underlying synthesis models. Making a trade-off between utility and disclosure risk ultimately depends on the use-case of the synthetic data and the sensitivity of original data. In the case of the linked TREE and LABB data, the true values of educational outcomes for those who explicit or implicit refuse to participate in the TREE panel survey should be deemed extremely sensitive. Even if the true risk of disclosure is difficult to determine, I think it would be prudent to reduce even the perceived risk of disclosure by publishing only synthetic data that do not contain any records with a TCAP score above 0.5.

5 Conclusion

The aim of this paper is to illustrate the feasibility of generating synthetic data based on linked panel survey and register data. The main motivation for this exercise is the knowledge that panel attrition can lead to biases in survey statistics. Ideally, panel survey data could be supplemented with data taken from official registries, wherever available. However, with participation in panel surveys usually comes the explicit consent to have this information (anonymously) published. Nonresponse, in that way, can be seen as an explicit or implicit refusal for the disclosure of such information, even in anonymous form. Therefore, even if information from registries could be used to complement panel survey data, it would neither be ethical nor legal, in most cases, to publish this data without additional layers of disclosure control.

In this paper, I have shown that such an approach is feasible. Generating synthetic data based on linked panel survey and register data can produce statistical products that have a reasonable level of utility without adding much risk of disclosure if they were to be disseminated. Ideally, such a product could be used in a similar way to the synthetic versions of the EU-SILC published by Eurostat. If publicly available, researchers could develop their analyses on the synthetic version of the linked TREE and LABB data and check their results on the original data in a more controlled setting.

The synthesis models used in this paper are fairly straightforward and with more model fine-tuning I believe that the utility of a final statistical product could be further improved. In addition, I have selected the synthesis variables to best fit the modelling of educational outcomes with an overall straightforward data structure. The selection of variables and the complexity of the data could be expended to facilitate a wider range of analyses. Here it is important to note that I have approached the data synthesis process by using brute-force tactics in setting the variable synthesis sequence and the modelling choices for each individual variable. An upside to this process is that it is easy to implement and, as I have shown, it can lead to synthetic data that have good utility. However, a major downside to this approach is that it quickly becomes computationally prohibitive when more synthetic variables are added. Even with only a handful of variables, I had to make some informed decisions on how to compare the synthesis sequences and the modelling of each variable. Still, with increasing computing power, such automated approaches will likely be more feasible in the future, making it easier and quicker for data producers to make and disseminate synthetic statistical products of sensitive micro-data.

References

Abowd, John, Martha Stinson, and Gary Benedetto. 2006. "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project." https://hdl.handle.net/1813/43929.

Abowd, John, Simon Woodcock, Benoit Dostie, Sam Hawala, Janet Heslop, Paul Massell, Carol Murphree, et al. 2001. "Disclosure Limitation in Longitudinal Linked Data." *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. https://www.researchgate.net/publication/267956705_Disclosure_Limitation_in_Longitudinal_Linked_Data.

Angelone, Domenico, and Florian Keller. 2019. "Überprüfung Des Erreichens Der Grundkompetenzen (ÜGK) Im Fach Mathematik Im 11. Schuljahr. Technische Dokumentation Zur Testentwicklung Und Skalierung." Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK). https://www.uegk-schweiz.ch/wp-content/uploads/2019/05/ÜGK2016_Technischer-Bericht_ADB.pdf.

Baeriswyl, Franz, Christian Wandeler, Ulrich Trautwein, and Katrin Oswald. 2006. "Leistungstest, Offenheit von Bildungsgängen Und Obligatorische Beratung Der Eltern." *Zeitschrift Für Erziehungswissenschaft* 9: 373–92. https://doi.org/10.1007/s11618-006-0056-6.

Beck, Michael. 2015. *Bildungserfolg von Migranten - Der Beitrag von Rational-Choice-Theorien Bei Der Erklärung von Migrationsbedingten Bildungsungleichheiten in Bern Und Zürich*. https://doi.org/10.18747/PHSG-coll3/id/308.

Benedetto, G., J. C. Stanley, and E. Totty. 2018. "The Creation and Use of the Sipp Synthetic Beta V7. 0." *United States Census Bureau*. https://www.census.gov/library/working-papers/2018/adrm/SIPP-Synthetic-Beta.html.

Bowen, Claire, Fang Liu, and Bingyue Su. 2021. "Differentially Private Data Release via Statistical Election to Partition Sequentially: Statistical Election to Partition Sequentially." *METRON* 79 (March). https://doi.org/10.1007/s40300-021-00201-0.

Buchmann, Marlis, Irene Kriesi, Maarten Koomen, Christian Imdorf, and Ariane Basler. 2016. "Differentiation in Secondary Education and Inequality in Educational Opportunies: The Case of Switzerland." In *Models of Secondary Education and Social Inequality – an International Comparison*, edited by H.-P. Blossfeld, S. Buchholz, J. Skopek, and M. Triventi, 111–28. Edward Elgar Publishing.

Caiola, Gregory, and Jerome P. Reiter. 2010. "Random Forests for Generating Partially Synthetic, Categorical Data." *Trans. Data Privacy* 3 (1): 27–42. https://dl.acm.org/doi/10.5555/1747335.1747337.

Choi, Edward, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. 2017. "Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks." arXiv. https://doi.org/10.48550/ARXIV.1703.06490.

Combet, Benita. 2013. "Zum Einfluss von Primären Und Sekundären Effekten Der Sozialen Herkunft Beim Zweiten Schulischen Übergang in Der Schweiz. Ein Vergleich Unterschiedlicher Dekompositions- Und Operationalisierungsmethoden." *Schweizerische Zeitschrift Für Bildungswissenschaften* 35: 447–71. https://doi.org/10.25656/01:10303.

Drechsler, Jörg. 2010. "Using Support Vector Machines for Generating Synthetic

Datasets." In *Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, 148–61. PSD'10. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-15838-4_14.

———. 2012. "New Data Dissemination Approaches in Old Europe - Synthetic Datasets for a German Establishment Survey." *Journal of Applied Statistics* 39: 243–65. https://doi.org/10.1080/02664763.2011.584523.

Drechsler, Jörg, and Jerome P. Reiter. 2011. "An Empirical Evaluation of Easily Implemented, Nonparametric Methods for Generating Synthetic Datasets." *Computational Statistics & Data Analysis* 55 (12): 3232–43. https://doi.org/10.1016/j.csda.2011.06.006.

Falter, Jean Marc. 2012. "Parental Background, Upper Secondary Transitions and Schooling Inequality in Switzerland." *Swiss Journal of Sociology* 38: 201–22. https://doi.org/10.5169/seals-815118.

Falter, Jean Marc, and Florian Wendelspiess Chávez Juárez. 2011. "Does Tracking Shape the Intergenerational Transmission of Educational Attainment? Evidence from Switzerland." *Labor: Human Capital eJournal* 10: 1–15. https://doi.org/10.2139/ssrn.1938281.

FSO. 2021. *Übergänge Und Verläufe in Der Obligatorischen Schule*. 16804389. Neuchâtel: BFS; Bundesamt für Statistik (BFS); Bundesamt für Statistik (BFS). https://dam-api.bfs.admin.ch/hub/api/dam/assets/16804389/master.

———. 2022. "Längsschnittanalysen Im Bildungsbereich (LABB)." Swiss Federal Statistical Office. https://www.bfs.admin.ch/bfs/de/home/statistiken/bildung-wissenschaft/erhebungen/labb.html.

Ganzeboom, Harry. 2010. "A New International Socio-Economic Index (ISEI) of Occupational Status for the International Standard Classification of Occupation 2008 (ISCO-08) Constructed with Data from the ISSP 2002–2007." *Annual Conference of International Social Survey Programme*. http://www.harryganzeboom.nl/Pdf.

Glauser, David. 2015. *Berufsausbildung Oder Allgemeinbildung: Soziale Ungleichheiten Beim Übergang in Die Sekundarstufe II in Der Schweiz*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-09096-8.

Gomatam, S., and Alan Karr. 2003. "Distortion Measures for Categorical Data Swapping. Technical Report." https://www.niss.org/sites/default/files/technicalreports/tr131.pdf.

Hartley, Levin, J. E. K., and C. Currie. 2016. "A New Version of the HBSC Family Affluence Scale - FAS III: Scottish Qualitative Findings from the International FAS Development Study." *Child Indicators Research* 9: 233–45. https://doi.org/10.1007/s12187-015-9325-3.

Hobza V, Bucksch J, Hamrik Z, and De Clercq B. 2017. "The Family Affluence Scale as an Indicator for Socioeconomic Status: Validation on Regional Income Differences in the Czech Republic." *Int J Environ Res Public Health* 14. https://doi.org/10.3390/ijerph14121540.

Hupka-Brunner, Sandra, Ben Jann, Maarten Koomen, Dominique Krebs-Oesch, Thomas Meyer, Barbara Müller, Christina von Rotz, Stefan Sacchi, and Barbara Wilhelmi. 2021. "Tree2 Study Design." University of Bern: TREE. https://boris.unibe.ch/152018/16/Hupka-Brunner_etal_2021_TREE2_Study_Design.pdf.

Hupka-Brunner, Sandra, Stefan Sacchi, and Barbara Stalder. 2010. “Social Origin and Access to Upper Secondary Education in Switzerland: A Comparison of Company-Based Apprenticeship and Exclusively School-Based Programmes.” *Schweizerische Zeitschrift Fur Soziologie/Revue Suisse de Sociologie/Swiss Journal of Sociology* 36: 11–31. https://doi.org/10.5167/uzh-43185.

Imdorf, Christian. 2005. *Schulqualifikation Und Berufsfindung. Wie Geschlecht Und Nationale Herkunft Den Übergang in Die Berufsbildung Strukturieren.* VS Verlag für Sozialwissenschaften Wiesbaden. https://doi.org/10.1007/978-3-322-93537-3.

Karr, Alan, C. N. Kohnen, Anna Oganian, J. P. Reiter, and A. P. Sanil. 2006. “A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality.” *The American Statistician* 60 (February): 224–32. https://doi.org/10.1198/000313006X124640.

Kennickell, Arthur. 1997. “Multiple Chapter Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances.” *Survey of Consumer Finances*. https://www.federalreserve.gov/Pubs/Oss/oss2/papers/expers95.pdf.

Kunter, Schümer Gundel, Mareike, and Manfred Weiss. 2002. “PISA 2000: Dokumentation Der Erhebungsinstrumente.” Heenemann GmbH, Berlin. https://www.iqb.hu-berlin.de/fdz/studies/PISA-2000/pisa2000_SH.pdf.

Liew, Chong K., Uinam J. Choi, and Chung J. Liew. 1985. “A Data Distortion by Probability Distribution.” *ACM Trans. Database Syst.* 10 (3): 395–411. https://doi.org/10.1145/3979.4017.

Little, Claire, Mark Elliot, Richard Allmendinger, and Sahel Shariati Samani. 2021. “Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study.” arXiv. https://doi.org/10.48550/ARXIV.2112.01925.

Little, Roderick J. A. 1993. “Statistical Analysis of Masked Data.” *Journal of Official Statistics* 9 (2). https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

Neuenschwander, Markus, Michelle Gerber, Nicole Frank, and Benno Rottermann. 2012. *Schule Und Beruf. Wege in Die Erwerbstätigkeit.* https://doi.org/10.1007/978-3-531-94156-1.

Neuenschwander, Markus, and Tina Malti. 2009. “Selektionsprozesse Beim Übergang in Die Sekundarstufe i Und II.” *Zeitschrift Für Erziehungswissenschaft* 12: 216–32. https://doi.org/10.1007/s11618-2009-0074-2.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. “Synthpop: Bespoke Creation of Synthetic Data in r.” *Journal of Statistical Software* 74 (11): 1–26. https://doi.org/10.18637/jss.v074.i11.

Nowok, Beata, Gillian Raab, and Chris Dibben. 2017. “Providing Bespoke Synthetic Data for the UK Longitudinal Studies and Other Sensitive Data with the Synthpop Package for r.” *Statistical Journal of the IAOS* 33 (January): 1–12. https://doi.org/10.3233/SJI-150153.

Park, Noseong, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. “Data Synthesis Based on Generative Adversarial Networks.” *Proceedings of the VLDB Endowment* 11 (10): 1071–83. https://doi.org/10.14778/3231751.3231757.

Pierce, Donald A., and Daniel W. Schafer. 1986. “Residuals in Generalized Linear

Models." *Journal of the American Statistical Association* 81: 977–86. https://doi.org/doi.org/10.2307/2289071.

Raab, Gillian M., Beata Nowok, and Chris Dibben. 2016. "Practical Data Synthesis for Large Samples." *Journal of Privacy and Confidentiality* 7 (3): 67–97. https://doi.org/10.29012/jpc.v7i3.407.

Raghunathan, Trivellore, Jerome P. Reiter, and Donald Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16. http://www2.stat.duke.edu/~jerry/Papers/jos03.pdf.

Reiter, Jerome P. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology* 29 (2): 181–88. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6785-eng.pdf?st=lCiB-Wcb.

———. 2005. "Using CART to Generate Partially Synthetic, Public Use Microdata." *Journal of Official Statistics* 21. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-cart-to-generate-partially-synthetic-public-use-microdata.pdf.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (April): 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rubin, Donald B. 1993. "Discussion: Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.

Sacchi, Stefan, and Dominique Krebs-Oesch. 2021. "Scaling Methodology and Scale Reporting in the Tree2 Panel Survey. Documentation of Scales Implemented in the Baseline Survey (2016)." TREE, University of Bern. https://doi.org/10.48350/152055.

Sandra Hupka-Brunner, Thomas Meyer, Ben Jann. 2016. "Erläuterungen Zum Kontextfragebogen Der ÜGK 2016: Allgemeiner Teil." https://www.uegk-schweiz.ch/wp-content/uploads/2019/05/UEGK_2016_CQ_Erlaeuterungen_Allgemein.pdf.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Chapman; Hall/CRC.

SKBF. 2018. "Swiss Education Report 2018." Swiss Coordination Centre for Research in Education. https://www.skbf-csre.ch/fileadmin/files/pdf/bildungsberichte/2018/Swiss_Education_Report_2018.pdf.

Snoke, Joshua, Gillian Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2017. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181. https://doi.org/10.1111/rssa.12358.

Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso. 2022. "Synthetic Data – Anonymisation Groundhog Day." arXiv. https://doi.org/10.48550/ARXIV.2011.07018.

Taub, Jennifer, and Mark Elliot. 2019. "The Synthetic Data Challenge." UNECE: conference of european statisticians. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Synthethic_Data_Challenge_Elliot_AD.pdf.

Wolf, Peter-Paul de. 2015. "Public Use Files of Eu-Silc and Eu-Lfs Data." *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality: Helsinki, Finland,* 1–10. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_21_Session_5_-_Netherlands___de_Wolf_.pdf.

Woo, Mi-Ja, Jerome Reiter, Anna Oganian, and Alan Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1. https://doi.org/10.29012/jpc.v1i1.568.

Appendix

Generalized linear model for variable *hisei08* on original data.

```
Call:
glm(formula = hisei08 ~ ., family = gaussian, data = obs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-60.819  -12.132    1.039   12.162   55.734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.0187     5.1446   4.669 3.10e-06 ***
sex2          0.9367     0.4529   2.068 0.038648 *
langreg2     -1.7009     0.4974  -3.420 0.000631 ***
wlem          0.5299     0.2312   2.292 0.021964 *
marks        -0.3578     0.5398  -0.663 0.507492
ls_req2      -2.2816     0.6914  -3.300 0.000972 ***
ls_req3      -5.5829     0.9070  -6.155 7.98e-10 ***
ls_req4       1.3636     1.5956   0.855 0.392814
pareduc1     11.9279     0.4938  24.157  < 2e-16 ***
immig2       -9.1583     0.6309 -14.517  < 2e-16 ***
immig3       -5.8423     0.8264  -7.069 1.73e-12 ***
wealth        0.2798     0.4824   0.580 0.561958
joyreadp      3.9356     0.3076  12.793  < 2e-16 ***
fas           1.9469     0.1698  11.463  < 2e-16 ***
inccap        0.5346     0.2529   2.114 0.034554 *
aspmf2       -2.9496     0.6158  -4.789 1.71e-06 ***
aspmf3       -4.3244     1.9233  -2.248 0.024589 *
aspmf4       -0.8090     0.7193  -1.125 0.260760
aspideal1     4.0009     4.1085   0.974 0.330192
aspideal2     3.5839     4.1143   0.871 0.383737
us_enroll1    8.9223     6.0303   1.480 0.139039
us_enroll2   -0.3368     1.0533  -0.320 0.749157
us_enroll3   -2.2897     0.9109  -2.514 0.011968 *
us_enroll4   -1.2571     1.1586  -1.085 0.277984
us_enroll5    1.5713     1.0703   1.468 0.142106
us_enroll6    0.8716     1.3000   0.670 0.502581
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 282.3691)

    Null deviance: 2661200  on 6079  degrees of freedom
Residual deviance: 1709463  on 6054  degrees of freedom
  (2349 Beobachtungen als fehlend gelöscht)
AIC: 51593

Number of Fisher Scoring iterations: 2
```

Generalized linear model for variable ***hisei08*** on pooled synthetic data

```
Fit to synthetic data set with 5 syntheses. Inference to coefficients
and standard errors that would be obtained from the original data.

Call:
glm.synds(formula = hisei08 ~ ., family = gaussian, data = syn)

Combined estimates:
            xpct(Beta) xpct(se.Beta)  xpct(z) Pr(>|xpct(z)|)    
(Intercept)  24.749433      5.074619   4.8771      1.077e-06 ***
sex2          0.680505      0.459192   1.4820      0.1383509    
langreg2     -1.307207      0.501571  -2.6062      0.0091546 ** 
wlem          0.010260      0.232488   0.0441      0.9648005    
marks         0.543446      0.539884   1.0066      0.3141281    
ls_req2      -2.319335      0.704070  -3.2942      0.0009871 ***
ls_req3      -6.742166      0.916393  -7.3573      1.877e-13 ***
ls_req4       0.145613      1.594177   0.0913      0.9272218    
pareduc1     12.142418      0.498175  24.3738      < 2.2e-16 ***
immig2       -8.194026      0.626958 -13.0695      < 2.2e-16 ***
immig3       -5.183242      0.837989  -6.1853      6.197e-10 ***
wealth        0.784478      0.477918   1.6414      0.1007041    
joyreadp      4.103512      0.329903  12.4386      < 2.2e-16 ***
fas           1.611479      0.166359   9.6868      < 2.2e-16 ***
inccap        0.470870      0.253523   1.8573      0.0632676 .  
aspmf2       -1.941913      0.619990  -3.1322      0.0017352 ** 
aspmf3       -2.496155      1.925041  -1.2967      0.1947425    
aspmf4       -0.045726      0.725882  -0.0630      0.9497718    
aspideal1     0.860004      4.083044   0.2106      0.8331774    
aspideal2     0.515830      4.088373   0.1262      0.8995973    
us_enroll1    1.365748      6.192035   0.2206      0.8254309    
us_enroll2   -0.539672      1.080918  -0.4993      0.6175875    
us_enroll3   -2.541597      0.929568  -2.7342      0.0062538 ** 
us_enroll4   -0.810955      1.179060  -0.6878      0.4915800    
us_enroll5    1.884180      1.088751   1.7306      0.0835250 .  
us_enroll6    1.221016      1.315848   0.9279      0.3534432    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CI overlap between generalized linear models for variable ***hisei08***

```
Call used to fit models to the data:
glm.synds(formula = hisei08 ~ ., family = gaussian, data = syn)

Differences between results based on synthetic and observed data:
              Synthetic   Observed        Diff Std. coef diff CI overlap
(Intercept) 24.74943336 24.0186758  0.73075759     0.14204470  0.9564355
sex2         0.68050491  0.9367489 -0.25624394    -0.56579385  0.8640292
langreg2    -1.30720681 -1.7008788  0.39367200     0.79145899  0.8089685
wlem         0.01025981  0.5298729 -0.51961314    -2.24720895  0.4537755
marks        0.54344621 -0.3577558  0.90120203     1.66959082  0.5947103
ls_req2     -2.31933547 -2.2815535 -0.03778196    -0.05464876  0.9564355
ls_req3     -6.74216637 -5.5829057 -1.15926066    -1.27807257  0.6902381
ls_req4      0.14561346  1.3636130 -1.21799956    -0.76332897  0.8158320
pareduc1    12.14241756 11.9279000  0.21451761     0.43445706  0.8960745
immig2      -8.19402595 -9.1582850  0.96425909     1.52844228  0.6291496
immig3      -5.18324166 -5.8423404  0.65909870     0.79752989  0.8074872
wealth       0.78447818  0.2797705  0.50470767     1.04626394  0.7467978
joyreadp     4.10351187  3.9355783  0.16793352     0.54588473  0.8688869
fas          1.61147853  1.9469350 -0.33545650    -1.97503481  0.5201841
inccap       0.47086951  0.5346372 -0.06376765    -0.25214512  0.9405573
aspmf2      -1.94191333 -2.9495547  1.00764133     1.63619124  0.6028596
aspmf3      -2.49615533 -4.3243559  1.82820053     0.95053874  0.7701541
aspmf4      -0.04572576 -0.8089986  0.76327284     1.06113305  0.7431698
aspideal1    0.86000417  4.0008919 -3.14088774    -0.76448305  0.8155504
aspideal2    0.51583003  3.5839178 -3.06808772    -0.74571871  0.8201288
us_enroll1   1.36574808  8.9223000 -7.55655195    -1.25309446  0.6963326
us_enroll2  -0.53967241 -0.3368125 -0.20285988    -0.19259260  0.9550877
us_enroll3  -2.54159729 -2.2897315 -0.25186574    -0.27651714  0.9346107
us_enroll4  -0.81095530 -1.2570623  0.44610699     0.38503010  0.9081343
us_enroll5   1.88418026  1.5713412  0.31283902     0.29230147  0.9307595
us_enroll6   1.22101604  0.8716206  0.34939541     0.26876451  0.9365023

Measures for 5 syntheses and 26 coefficients
Mean confidence interval overlap:  0.7947251
Mean absolute std. coef diff:  0.8430104

Mahalanobis distance ratio for lack-of-fit (target 1.0): 4.24
Lack-of-fit test: 110.2365; p-value 0 for test that synthesis model is
compatible with a chi-squared test with 26 degrees of freedom.

Confidence interval plot:
```

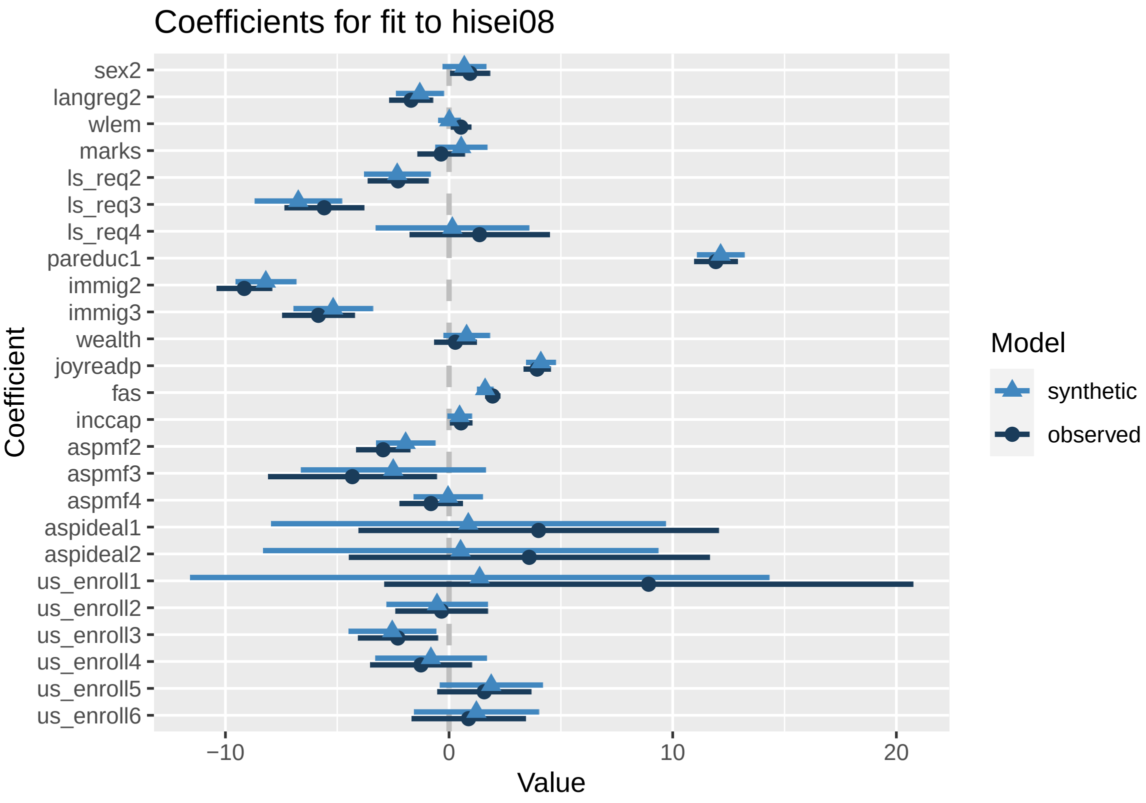

Figure 11: Regression residuals for linear model fit *joyreadp* after square-root transformation.

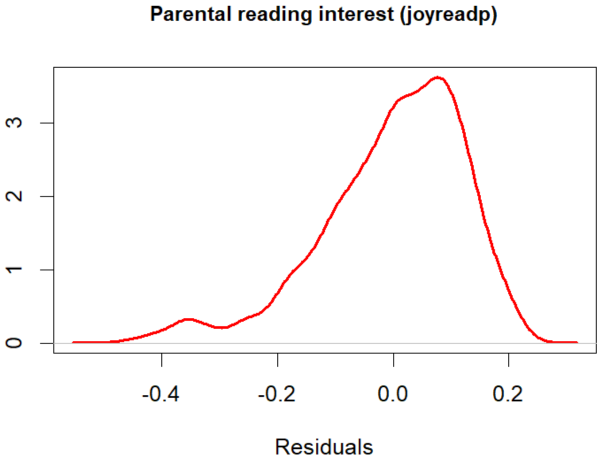

Figure 11: Regression residuals density plot.