Catalogue no. 11-522-x
ISSN: 1709-8211

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Generating smart deep files: the example of synthesizing hierarchical data

by Héloïse Gauvin

Release date: October 29, 2021

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2021
Adopting Data Science in Official Statistics to Meet Society's Emerging Needs

Generating smart deep files: the example of synthesizing hierarchical data

Héloïse Gauvin[1]

Abstract

The Government of Canada's Directive on Open Government aims to ensure that Canadians have greater access to government data and information. One solution for open data is smart synthetic files, which retain as much analytical value as possible and take into account confidentiality issues that arise from collecting personal information.

In recent years, Statistics Canada has acquired a recognized expertise in producing synthetic data files of high analytical value. In a current project, Statistics Canada is tackling a new challenge to synthesize a database and preserve hierarchical structures in the form of families, where records are linked and share common traits that must be maintained. These challenges are also encountered when synthesizing structured data such as business data.

This paper presents the challenges and solutions for building synthetic data with such hierarchical structures. Application of this strategy will be illustrated with the development of a synthetic database that supports the development of retirement income policies. This database includes over 20 variables and 8 million records structured into approximately 4 million family units. We will present how family structures have been preserved, discuss the practical and technical challenges inherent in developing such a large and complex database, present the risk and utility of the data, and propose avenues for future research.

Keywords: synthetic data of high analytical value; family structures; modern data access solution.

1. Introduction

1.1 Synthetic data

Since the term "synthetic data" is used quite loosely, it is important to define what synthetic data refer to for this project. Synthetic data are data generated knowingly so as to preserve the analytical value of an original dataset while protecting confidentiality. Analytical value means that the same statistical conclusions will be obtained from the synthetic data or the original file when analyses are done where these analyses are not known in advance. The underlying goal of synthetic data is to have more useful data more readily available to the public than other current data access options, while protecting confidentiality. The challenge in developing synthetic data lies in the ability to preserve the content of the file in terms of correlations and interactions between variables, whether or not these relationships have been identified beforehand.

1.1.1 Data synthesis process

The approach we have chosen for data synthesis is based on an important premise: the original dataset is only one of countless possible datasets, all of which are considered to be independent occurrences of the same statistical distribution. Note that other approaches exist and that this paper focuses on the approach implemented at Statistics Canada thus far (Sallier 2020). This is a two-step process: 1) identify the statistical distribution governing the original data and 2) use the resulting model to randomly generate a new set of data points, i.e., synthetic data.

[1] Héloïse Gauvin, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (heloise.gauvin@statcan.gc.ca).
Disclaimer: The content of this paper reflects the author's position, not necessarily that of Statistics Canada.

1.2 Project presentation and objectives

Statistics Canada is currently developing a new Canadian microsimulation model of retirement income in partnership with Employment and Social Development Canada and HEC Montréal (École des hautes études commerciales de Montréal). The purpose of this project is to develop an open simulation model that researchers can run on their own computer using an open, non-confidential database. The original database comprises a portion of the 2016 Census long-form questionnaire. The microsimulation model can be used to analyze Canada Pension Plan (CPP) policy changes. Our goal is to create a synthetic version of the database for this project. This synthetic dataset will have high analytical value and will be disseminated outside a secure environment.

2. Creating synthetic data with a family structure

2.1 General data synthesis approach

The data synthesis approach adopted for this project is the complete conditional specification (CCS) (Drechsler 2011). This method has been used in the past in such projects at Statistics Canada (Sallier 2020). The CCS assumes that the analytical value of the original file is derived from the joint distribution of the data. To preserve the information, the CCS decomposes the multidimensional joint distribution into a series of one-dimensional conditional distributions. If a dataset with p variables is considered, this approach results in the following formula:

$$(1)\quad f_{X_1,X_2,\ldots,X_p} = f_{X_1} \times f_{X_2|X_1} \times \ldots \times\ f_{X_p|X_1,X_2,\ldots,\ X_{p-1}}$$

Essentially, instead of trying to explain all the relationships in the dataset at once, the synthesis captures them sequentially, one variable at a time. In practice, the method follows these steps:

1. Take a simple random sample of $x_{1,obs}$ and define it as $x_{1,syn}$

2. Adjust a model $f\left(x_{2,obs}\middle|x_{1,obs}\right)$ and draw $x_{2,syn}$ from $f\left(x_{2,syn}\middle|x_{1,syn}\right)$

3. Adjust a model $f\left(x_{3,obs}\middle|x_{1,obs},x_{2,obs}\right)$ and draw $x_{3,syn}$ from $f\left(x_{3,syn}\middle|x_{1,syn},x_{2,syn}\right)$

4. And so on, until $f\left(x_{p,syn}\middle|x_{1,syn},x_{2,syn},\ldots,x_{p-1,syn}\right)$,

   where$x_{i,obs}$ denotes the i^e observed variable and $x_{i,syn}$ denotes the i^e synthetic variable.

This approach is available in the R synthpop library (Nowok et al. 2016), which is the one used for the project. The models chosen are a mixture of parametric models and decision trees, depending on the variable to model. The order in which variables are synthesized is important since it affects the quality of the synthesis. Note that it is preferable to synthesize the variables of interest at the end. In addition, to minimize the risk of re-identification and facilitate the creation of their synthetic version, the data are preprocessed (reduction of the information present at different levels, see section 4.2).

2.2 Hierarchical data synthesis process

Until now, data synthesis has been a rather simple process, where observations (in this case, individuals) were first analyzed and then equivalent but synthetic observations were generated. However, in this project, the structure of the observations (the fact that not all observations are independent of each other) plays an important role for the simulations. Retirement income is influenced by having had children (child-rearing clause) or having a spouse (separation of income in the event of divorce, for example). The question is therefore how to preserve the correlations between the variables knowing that the individuals themselves are related (presence of couples and families)?

To do this, the problem is divided into four types of family structures: 1) individuals outside the census family (single individuals); 2) couples without children; 3) couples with children; and 4) lone-parent families. The synthesis process is twofold: first, the family structures (first synthesis) are generated, then the synthesis process for each type of structure (second synthesis) is applied. In other words, we divide the data, generate the structures (create families) and then we fill in these structures (create persons).

Specifically, for the first phase, the synthesis focuses on five variables, namely 1) type of family structure, 2) family size, 3) the age of the first member of the couple, 4) the age difference and 5) the type of couple (opposite sex or same sex). Note that the last two variables are present only if a couple is present. In addition, family size is always 1 for non-census families and 2 for childless couples.

The second phase depends largely on the type of structure to complete. For non-census family persons, the first phase provides the age of the individual. All other characteristics of the individual must be generated in the second phase. The variables to synthesize are grouped under different categories and the variables of the different categories are synthesized in the following order: demographic, immigration and mobility, education, work, income, and pension plan benefit variables.

2.2.1 Couples without children

The order of data synthesis is the same for the other types of structures, with a few variations. For couples, the first phase provides information on the age of the first member of the couple, the age difference between the two members of the couple and the type of couple. Therefore, the age of the second member of the couple is subtracted from the age of the first member and the age difference between them. Then, the type of couple identifies the sex of each member. Once these variables have been deduced, we continue with the other variables, alternating between the first and second member of the couple, to always consider the information generated beforehand. We do this since we observed in the original data that the couples are not formed randomly. For example, immigrants are more likely to form a couple with another immigrant; similarly, the data show that it is more common for a person with a university degree to form a couple with another person with the same type of profile. In other words, birds of a feather flock together, and for the synthetic data to reflect this, information on the entire couple is used in the synthesis. Tangibly, let’s take the example of the year of immigration of the second member of the couple. We will adjust a conditional model on this person’s demographic information, on the couple’s demographic information, on the first person’s modelled migration variables, on the second person’s migration variables, and lastly on the same variable (year of immigration, if applicable) of the first person.

2.2.2 Couples with children and lone-parent families

For couples with children and lone-parent families, the procedure is similar. In both cases, we generate the parents first. Couples with children are treated the same as couples without children. For lone-parent families, the parent is treated as a person outside the census family. Then we move on to children for whom all or part of the information is generated by age. We start by creating children by family size, then generate their age and sex. Then, we generate the other variables by conditioning on the parents and respecting the constraints, when applicable. For example, some variables are absent (N/A) when the children are under 15 years. Finally, when all family structures are completed, they are combined into a single dataset.

3. Results

3.1 Overview of the data

It is important to keep in mind that the original data here represent a sample of the Canadian population, that some records have been removed, and that the data have been pre-processed. Since we decided to model family units first, both datasets (original and synthetic) have the same number of *families* (4,236,024), but the number of *individuals* is slightly different (8,651,677 for the original data and 8,655,913 for the synthetic data). When we look at the distribution of family units by type of family structure, we see that the distribution is similar between the synthetic and original data. The same is true when looking at the distribution of the number of individuals in each type of family structure (Table 3.1-1).

Table 3.1-1
Number of individuals in each family structure for original and synthetic data

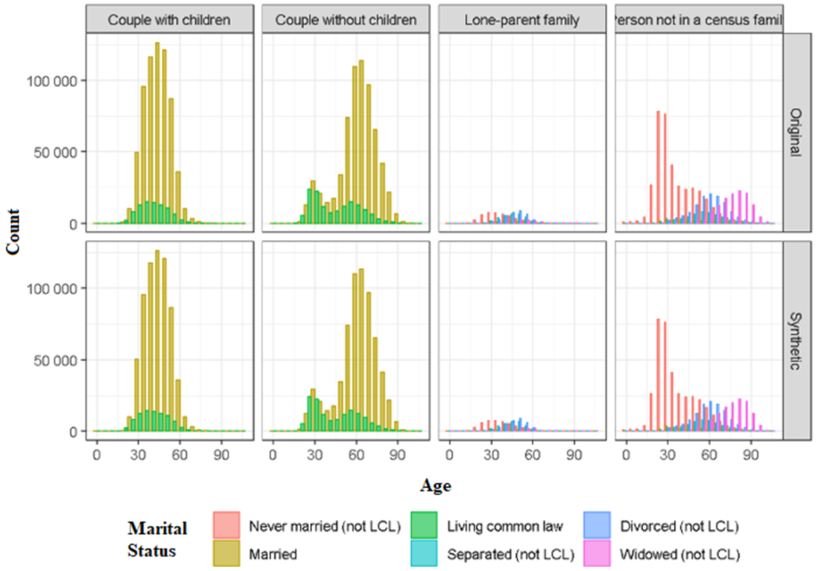
Individuals in each family structure	Data	
	Original	Synthetic
Person not in census family	1,854,855	1,852,585
Couple without children	2,198,240	2,201,314
Couple with children	3,734,342	3,734,804
Lone-parent family	864,240	867,210
Total	**8,651,677**	**8,655,913**

3.2 Synthetic variables

For the synthetic file to be considered of high analytical value, it must be verified that the similarities with the original data go beyond the general file structure. Specifically, it must be ensured that the analyses on the original and synthetic data lead to almost identical statistical conclusions. In this case, we know that the data will be used to conduct simulations, but it is difficult to know exactly which analyses will be conducted and their conclusions, and therefore which correlations should be preserved. We will be able to compare the results when the microsimulation model is ready. In the meantime, we conducted a first assessment of the analytical value by ensuring that the univariate, bivariate and multivariate distributions were in agreement between the original and synthetic data.

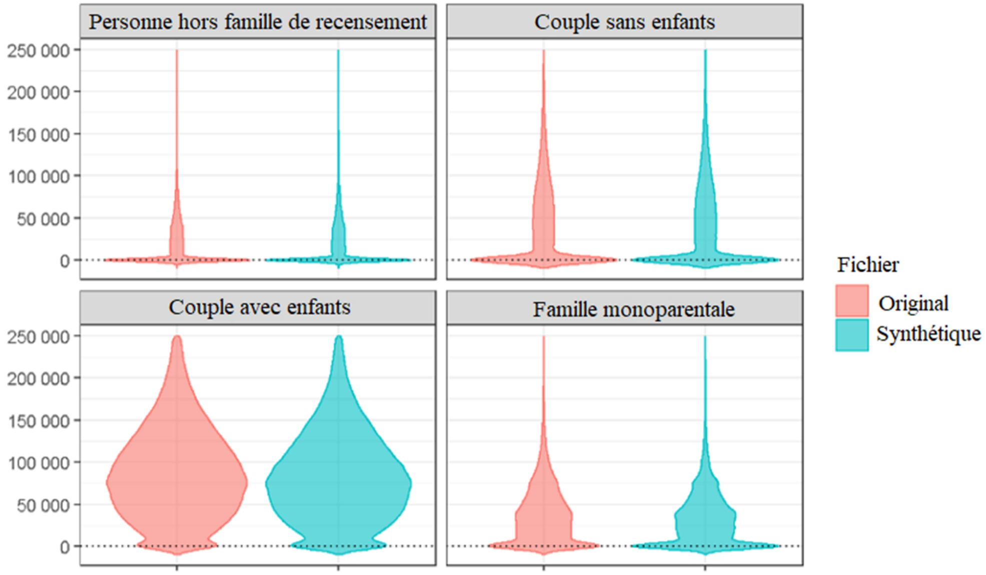
Two figures illustrate this approach. First, Figure 3.2-1 presents histograms of ages for each type of family structure by marital status. Note that the distributions from the synthetic data appear identical to those from the original data. Both figures present results for one region only.

Figure 3.2-1
Age distribution by type of family structure and marital status, original and synthetic data



Since age and marital status are among the first variables to model, we chose to also present results for a variable that occurs later in the synthesis process, namely income (the sum of employment earnings of the parents in the family). Figure 3.2-2 presents family income by family type for the original and synthetic data. Again, the original and synthetic distributions match. Note that the scale on the income axis is clipped and shows only values from $0 to $250,000.

Figure 3.2-2
Distribution of family incomes by type of family structure, original and synthetic data



Lastly, note that the distributions of the variables within the couples matched were also checked. As mentioned above, couples do not appear to be formed randomly in real life, and the synthetic data must therefore reflect this fact.

4. Discussion

4.1 Usefulness

The results (comparison of the original and synthetic data using univariate, bivariate analyses, descriptive statistics such as minimum, maximum, mean, median, standard deviation, etc.) show that the analytical value seems to be well preserved. The usefulness specifically for this project will be confirmed when we can compare the results of the microsimulations performed on the original and synthetic data. The usefulness of the synthetic data may even go beyond this project, depending on the quality of the data. In other words, if the synthesis has been carried out correctly, the analysis potential of the synthetic data will be unlimited.

4.2 Privacy risk management

Before discussing the risk that synthetic data might pose, it should be noted that some conventional approaches to preserving data confidentiality have already been put forward. First, we aggregated several variables (e.g., school attendance was reduced from nine [9] categories to three [3]), we reduced the geographic information available, and the monetary variables were rounded and limited to the extremes. In addition, because the data are from the long-form census, they are derived from a systematic sample of one in four private Canadian dwellings. Therefore, a single record in the original dataset is not necessarily unique in the population. This means that replication of original records is not necessarily problematic. Consideration should be given to their uniqueness in the original population as well as the information presented to measure the alleged risk.

4.3 Privacy risk assessment

In an attempt to assess the risk posed by the synthetic data, we performed an exact match on all variables. This matching shows that almost 70% of the synthetic data are new (i.e., records that are absent from the original data). The other part (the remaining 30%) consists of records that are identical to the original data. Of these, 10% are unique records on the original database, equalling 3% of the total records. Lastly, of these records, one third are from people outside of the census family (≈1%), the rest are part of a family or couple. This is important to stress. Looking at the analysis of families or couples, we have come to the conclusion that none of these is replicated in its entirety. In other words, a record may look the same on an individual level but when we consider the rest of the family or the other member of the couple, it is not identical to the original. It remains to be assessed whether this fraction of identically reproduced records (≈1%) that are rare on the original data can cause confidentiality concerns.

4.4 Next steps

The privacy risk analysis is not complete. We need to determine the risk of re-identification and how to quantify the uniqueness of records. Specifically, we need to assess the risk associated with the ability to infer information if some is already available, and whether certain combinations of variables lead to a higher risk. We will also focus on the perceived risk. We want to minimize it and assess different strategies to do so. Lastly, when the assessment is complete, we will release the first version of the synthetic database and address additional challenges such as synthesizing historical data.

References

Drechsler, J. 2011. *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer.

Nowok, B., Raab, G. M., and Dibben, C. 2016. *synthpop: Bespoke Creation of Synthetic Data in R*. Journal of Statistical Software, 74:1–26. [https://www.jstatsoft.org/article/view/v074i11]

Sallier, K. 2020. *Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis,* Statistical Journal of the IAOS, Vol. 36, No. 4, pp. 1059–1066.