CREATING A SYNTHETIC VERSION OF A LONGITUDINAL AND STRUCTURED FILE: CHALLENGES AND LESSONS

Héloïse Gauvin[1]

ABSTRACT

In recent years, generating synthetic data has increasingly been viewed by statistical agencies as a means of disseminating useful statistical information while fulfilling their obligations to protect the personal data they have been entrusted with. In 2018, Statistics Canada was one of the first national statistical agencies to release smart synthetic data files, where there is a goal of maintaining the analytical value of the original data. Since then, through various initiatives, its expertise has kept growing, establishing the Agency as a leader in the field. This presentation will focus on the production of more elaborate smart synthetic dataset than those previously produced. Intended to safely run open-source micro-simulations for the new Canadian retirement income model (PASSAGES), a massive smart synthetic file including both longitudinal and family components was created. We will describe how the synthesis process had to be enhanced to handle the challenges of capturing temporal correlation and life events.

KEY WORDS: Synthetic Database, Historical and Hierarchical Data, Microsimulation model.

RÉSUMÉ

Ces dernières années, la production de données synthétiques est de plus en plus considérée par les organismes statistiques comme un moyen de diffuser des informations statistiques utiles tout en remplissant leurs obligations de protection des données personnelles qui leur sont confiées. En 2018, Statistique Canada a été l'un des premiers organismes statistiques nationaux à publier des fichiers de données synthétiques « smart », dans le but de maintenir la valeur analytique des données originales. Depuis, grâce à diverses initiatives, son expertise n'a cessé de croître, faisant de l'Agence un leader dans le domaine. Cette présentation se concentrera sur la création d'un jeu de données synthétiques « smart » plus élaboré que ceux produits précédemment. Destiné à exécuter en toute sécurité des microsimulations à code source ouvert pour le nouveau modèle canadien des revenus de retraite (PASSAGES), un énorme fichier synthétique « smart » comprenant à la fois des composantes longitudinales et familiales a été créé. Nous décrirons comment le processus de synthèse a dû être amélioré pour relever les défis liés à la capture des corrélations temporelles et des événements de la vie.

MOTS CLÉS : Base de données synthétiques; données historiques et hiérarchiques; modèle de microsimulations.

1. INTRODUCTION

To promote openness, transparency and accountability in the Government of Canada, the National Action Plan on Open Government is established every two years. The 2018-2020 plan (Government of Canada, 2018) sought to "help Canadians understand the data and models used to design and study government programs" by ensuring that "microsimulations models, including underlying datasets, are made publicly available to help to explain how the government uses these models to design programs and to estimate their impacts".

Recently, Statistics Canada disseminated an open-source dynamic socio-economic microsimulation population model, called PASSAGES (Statistics Canada, 2024). PASSAGES supports policy analysis and research relating to Canadian retirement income system outcomes at the individual and family levels. This work was a collaborative effort between Employment and Social Development Canada, Statistics Canada and the Retirement and Savings Institute at HEC Montréal. For this project, being open translated directly into having an open-source model as well as an open database. To run the microsimulation model two components are required: a starting population and various modules to make the status of its individuals evolve over time.

[1] Statistics Canada, Ottawa, Canada, heloise.gauvin@statcan.gc.ca

To represent the starting population, a large confidential database was constructed involving different sources such as the long-form questionnaire from the 2016 Canadian Census of Population, multiple historical annual tax files and several pension benefits administrative files. Since that database contained sensitive information that was not to be disclosed, an open synthetic database was developed allowing researchers to run the open-source microsimulation model outside the secure environment of Statistics Canada.

The microsimulation model includes modules for demographic trends (fertility, migration, union, education, mortality), economic trends (employment, earnings) and pension benefits rules (retirement, post-retirement, survivor, disability, death, children). The synthetic data were developed to provide the confidentiality protection needed for a safe general release while boasting high fidelity to the original data to have significant analytical value. The open model combined with the open synthetic data allows users to run simulations and analyze the results as if these had come directly from the confidential database. The synthetic dataset produced is also interesting in its own right since the census data is one of the most used datasets in the Canadian network of research data centres (Cranswick and Hotton, 2022).

The complexity of creating a synthetic database of this magnitude should not be underestimated. It was the first time Statistics Canada released such a file that included a hierarchical component as well as a longitudinal component. The following section describes the multiple strategies used to create the synthetic version of the database. Section 3 describes the verifications implemented to check the fidelity of the synthetic data at the different stages of its creation. Section 4 presents how disclosure risk was managed and assessed.

2. SYNTHESIZING A HISTORICAL AND HIERARCHICAL DATABASE

Because of the complexity of the database from a risk and utility perspective, the construction of the synthetic version was divided into 5 phases: the pre-treatment of the information, the synthesis of the baseline units (families), the addition of information about their life trajectories, the synthesis of historical income information as well as pension benefits and finally, some adjustments to the database.

2.1 Pre-Synthesis: Making the Data Less Sensitive

To further protect the confidentiality of the respondents on the database and to simplify the synthesis process, various standard anonymization techniques were first applied to the original data (Duncan *et al.*, 2011). Even if synthesis proceeds globally and not on a record-by-record basis, a synthetic record may in the end look similar to one of the original records and give the appearance of disclosure. Therefore, the number of details contained in the data and the risk of re-identification had to be managed. All modifications were applied to the original database before the synthesis was undertaken. Similar to Bonnéry *et al.* (2019) and past experiences (Sallier, 2020), the original data were simplified and transformed to create a slightly less complex and less sensitive dataset from which the synthetic data were derived.

From a geographic perspective, the original data provided information on all thirteen Canadian provinces and territories. However, because pension allocations are the same throughout the country except for the province of Quebec, the geography dimensions for this project are Quebec (Qc), the rest of Canada (ROC) and outside of the country. This greatly reduced any identification risks in the file. Other categorical information such as the marital status, the level of education achieved and the school attendance information were aggregated. As for continuous monetary variables to be synthesized, amounts were top-coded and bottom-coded to remove self-disclosing outliers. Furthermore, values were rounded.

2.2 Synthesis: The Family and Baseline Information Layers

All synthesized variables were obtained using the Fully Conditional Specification (FCS) approach (Drechsler, 2011). This means that variables are generated iteratively, one at a time, with each variable being generated based on all of the previously generated information. The argument is that the overall joint distribution is preserved since the joint distribution of the data is a product of conditional distributions that are simpler to model. However, the entire dataset was not generated in one simple sequence but in different waves to limit the computation needs. For instance, families were first generated separately based on their province of residence in 2016 (Qc/ROC) and their type. Then, historical earnings and pension benefits were generated at the individual level and linked by conditioning on the relevant information.

For all synthetic data generated, the implementation was performed by using the R package 'synthpop' (Nowok *et al.*, 2016). Most variables were synthesized using a Classification and Regression Trees (CART) model as this method was proven to provide results with good utility and low disclosure risks (Drechsler and Reiter, 2011, El Kababji *et al.*, 2023) and has been used successfully with other data synthesis projects at Statistics Canada (Sallier, 2020).

The desired final synthetic product is a microdata file where each synthetic record represents an artificial individual in the Census long form file. However, records from the database are not independent from one another since family units were maintained in the original data. Therefore, instead of generating synthetic individuals directly, family structures were synthesized first and then the people within those structures were synthesized in a second phase. The database variable 'Family Type' broadly follows the Census 'Household' definition. Exceptionally, children over the age of 25 would form their own family unit on the data file, whereas on the Census they would be considered part of the same household.

The first phase created different types of family structures (single individuals, couples with or without children, and lone-parent families) with characteristics such as age of the family head and the family size. In the case of couples, their age difference as well as their composition (same sex or not) were created.

The second phase of the synthesis took place at the individual level and synthesized demographic information (sex, marital status), immigration, mobility, work status and income as of the year 2016. To maintain a correlation among family members, variables were generated sequentially alternating between the first and second member of the couple based on the information generated beforehand. This was done because it was observed in the original data that the couples' characteristics were strongly correlated. For example, immigrants are more likely to form a couple with another immigrant; similarly, a person with a university degree is more likely to form a couple with another person with the same type of education profile. Also, in the case of families including children, information for their members was generated by conditioning on the parents' information while making sure logical constraints were respected when applicable (e.g., many variables were 'Not available' for people under 15 years old). The important lesson here is that generating synthetic data relies on a complete understanding of the data rather than simply allowing the data synthesis to run. Determining the order as well as how the variables fit together are essential to put all the pieces of the puzzle together and obtain meaningful synthetic data.

2.3 Donor-Imputation Approach: The Life Trajectories Layer

The PASSAGES database contains an impressive amount of longitudinal information, and part of it describes historical trajectories of unions, kids and the place of residence. This information has complex longitudinal correlations that must be maintained for the synthetic data to be useful. Following Bocci and Beaumont (2009) strongly correlated variables were attributed simultaneously using a donor-imputation approach.

Based on specific criteria (the region of residence, family type, marital status, age and year of immigration) a pool of real families was created for each cross-sectional synthetic family. One family within the pool (referred to as a donor) was selected at random for each synthetic family (referred to as a recipient). For synthetic families that had no real family in their pool, criteria were gradually relaxed until all synthetic families had at least one family in their pool. Over 50% of families with or without kids were matched perfectly (meaning on all criteria) to a donor family, while that was the case for over 95% of single individuals. Historical trajectories of unions, kids, province of residence and other variables from 1966 to 2015 were assigned to the recipients using the donor values. Note that for non-perfect matches, post-edits were applied to maintain the age, sex and year of immigration from the synthetic database.

2.4 Synthesis: The Earnings and Pension Benefits Layer

The synthesis of longitudinal earning variables was performed following a conditional sequence, inspired by Kinney *et al.* (2011). Since the structure of the families as of 2016 was synthesized in the first phase, the values from 2015 to 1966 were created in a reverse order. The model was conditioning on the earnings for the next three years except for 2013, 2014 and 2015 for which there were fewer years available. The sequences are shown below for the synthesis of the variables imeam ('Wages, salaries and commissions') and imnetse ('Net-self-employment earnings') for year 2012 to year 1966:

$$f\left(imearn^{(year)} \middle| \begin{matrix} imearn^{(year+3.2.1)}, imnetse^{(year+3.2.1)}, \\ imspearn^{(year+3.2.1)}, imspnetse^{(year+3.2.1)} \\ life.trajectory.vars^{(year)}, 2016.vars \end{matrix} \right) \quad (1)$$

and

$$f\left(imnetse^{(year)} \middle| \begin{matrix} imearn^{(year+3.2.1.0)}, imnetse^{(year+3.2.1)}, \\ imspearn^{(year+3.2.1)}, imspnetse^{(year+3.2.1)} \\ life.trajectory.vars^{(year)}, 2016.vars \end{matrix} \right) \quad (2)$$

where:

- $imearn^{(year+3.2.1)} = imearn^{(year+3)}$, $imearn^{(year+2)}$, $imearn^{(year+1)}$;
- *imspearn* is the spouse's wages, salaries and commissions variable;
- *imspnetse* is the spouse's net-self-employment earnings variable;
- *life.trajectory.vars* are the variables relating to the province of residence, the union status and the count of kids under 7 years old, all for concerning the current year being modelled; and
- 2016.*vars* are demographic variables (age, sex, education and immigration).

In addition to the retirement pension, the Canada Pension Plan (CPP) includes different benefits for disabled contributors and their children, the legal partner of a deceased contributor and their children. CPP benefits can have more than one pay pattern if the benefit amount changes for any reason (such as a reoccurring disability or survivor benefits starting) other than the annual inflationary adjustment. People receiving pension benefits could have more than one pay pattern. To facilitate the synthesis process relating to retirement and disability benefits, and to avoid producing unique trends, only the first pay pattern was kept. For retirement and disability benefits, a start time was synthesized as well as a start amount (rounded and capped at the yearly maximum) using a CART model. For CPP orphans' benefit and disabled contributors' child's benefit, a start and end time were synthesized using a CART model.

Finally, note that for Quebec residents, due to the lack of administrative data, no pension information was synthesized prior to 2015. However, some pension information was synthesized as part of the baseline information layer and those were preserved for Quebec residents to make up for the shortfall.

2.5 Final Step: Database Adjustments

The synthetic dataset was created before the original database took its final form. Therefore, at the beginning of the project, the database containing the 2016 long-form census only represented a snapshot in time, that of census day May 10, 2016. However, many of the key variables in the microsimulation model, such as earnings, use calendar-year measurement periods. Consequently, the decision was made to create a starting population database that represents the population as of December 31, 2015. This decision was taken after the synthetic database was already created. To adjust the synthetic database from May 2016 back to December 2015, those born early in 2016 were removed. If there was a change of spouse during this period, then the affected families were also restructured. Other variables were adjusted accordingly, such as union status and place of residence while some, such as the level of education and the population living in institutions would remain the same between December 2015 and Census Day 2016. In Canada, pension may be subject to sp litting after a separation or a divorce. To account for past relationships and to compensate for any spouses lost in the process of shifting the database, some records were created using a combination of multiple methods of imputation. Those records do not have full historical information and they represent 4.8% of all individuals on the database.

The database is used in the microsimulation model as a starting population and is meant to reflect the entire Canadian population. However, individuals living in collective dwellings (1.95% of the population in 2016) were not on the long-form census which was used here. Collective dwelling refers to a dwelling of a commercial, institutional or communal nature, and it includes nursing homes, hospitals and staff residences. Since collective dwellings are included in the short-form census population, the short-form census was used to (a) provide the number of individuals in collective dwellings to be added to the model's starting population, and (b) identify appropriate donors for the imputation of the characteristics of individuals living in collective dwellings. The process of creating additional records was done by matching individuals in collective dwellings on the short-form census to similar individuals on the database (based on their marital status, age, province, income, sex, pension benefits) and simply cloning these database records to represent those in collective dwellings.

Family-level weights were developed for the confidential database to ensure that any samples drawn could represent the actual population as of December 31, 2015. Some preliminary results showed some slight differences in population proportions for different groups of interest (difference in age and sex groups) for the confidential database and the synthetic database. Therefore, separate calibration weights were obtained for the synthetic database.

3. DATA FIDELITY

3.1 Analytical Assessment Throughout the Synthesis Process

At each step of the synthesis process, tables and graphs were prepared (Gauvin, 2021) to ensure that the synthetic data represented well the actual dataset and to adjust the synthesis process when needed. For all categorical variables, frequency counts were compared between the original and synthetic datasets. When relevant, those comparisons were conducted across the whole 50-year period (1966-2015). For continuous variables, descriptive statistics (minimum, maximum, mean, median and standard deviation) were compared between the original dataset and the synthetic versions, and also on an annual basis when possible. For income variables, in addition to comparing the distributions of values, the annual proportion of individuals having non-zero income, the annual proportion of individuals becoming employed or unemployed, the number of transitions from employed to non-employed (and vice versa) and the proportion of lifetime unemployment were compared. Overall, data from both versions were considered sufficiently close.

3.2 Database Comparison

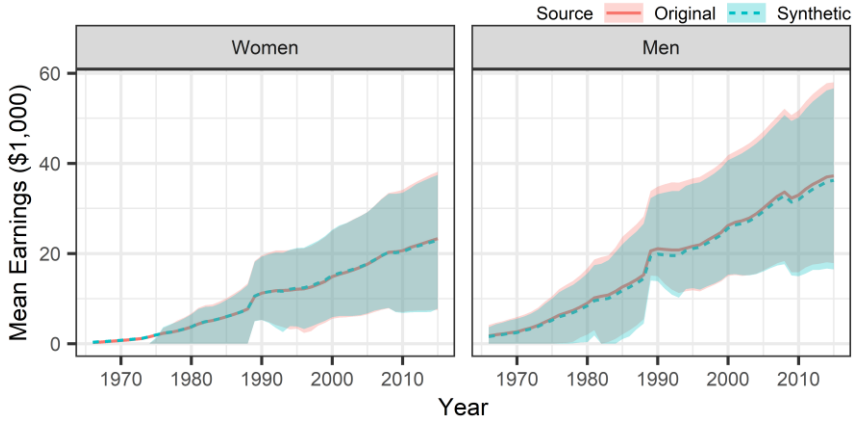
The synthetic database includes more records (see Table 1) than the original database. The reasons for this are twofold. First, the synthetic database was generated while the original database was being created. For example, after the creation of family units for 2016, records were removed from the original database for various reasons such as too much information missing from tax files. Since this process could not be easily applied to the synthetic database, no such records were removed. The second was that the synthetic database needed twice as many records for the database time adjustment to be performed as the original version. Indeed, in the original database certain individuals could be linked to one another over time, whereas this was not feasible for the synthetic version since family links were limited to a single time point (in 2016). An adjusted set of calibration weights was calculated for the synthetic database to take these differences into account. When the two databases are compared with their family-level weights, the population totals are very close and both reflect the population counts on December 31, 2015. Proportions across family types are also very close. The biggest difference comes from the synthetic database having 1.1% fewer family units that are couples with children.

Table 1 – The number of families of each type and the number of individuals in each family type for the original and synthetic database and their weighted totals

Family types	Number of families		Number of individuals	
	Original	Synthetic	Original	Synthetic
Single	1,836,589	2,095,699	1,836,589	2,095,699
Couple without children	1,018,142	1,174,032	2,102,082	2,543,448
Couple with children	871,119	927,521	3,433,545	3,686,193
Lone-parent family	289,268	304,772	762,726	805,872
Total	4,015,118	4,502,024	8,134,942	9,131,212
Weighted total	17,799,133	17,979,279	35,848,798	35,848,800

Apart from the number of records and their family types, there are many variables to compare. For income, the Fig. 1 presents the average value for earnings over the 1966-2015 period by sex for both original and synthetic databases. On the synthetic database any value of earnings below $3,500 was set to $0. The $3,500 value represents the annual amount on which CPP contributions are not required to be made. The same was assumed for the original database and this adjustment was made before the databases were compared. The comparison suggested that the synthetic mean values are on average 3.0% lower than the original mean values. The relative difference for the mean tends to be slightly larger for men (4.6%) than for women (-1.8%) but the range of relative difference per year is wider for women (1.6% to -19.7%). The data distributions are shown in Fig. 1 where the interval is delimited by the median and third quartile. The median was picked

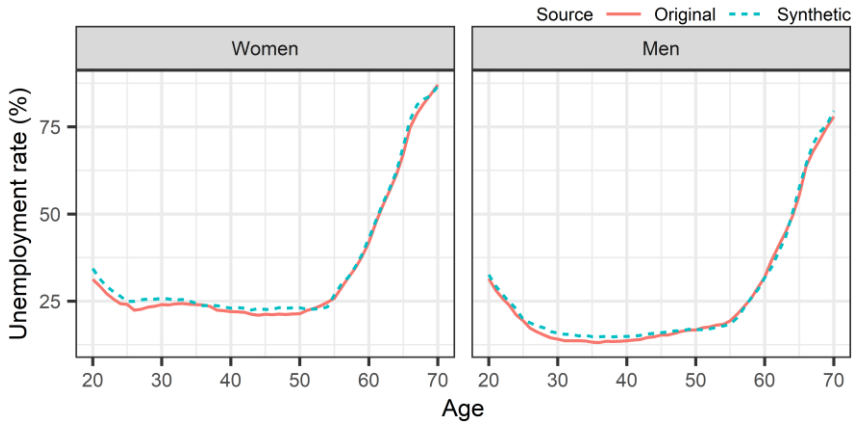
Figure 1 – Plot of the annual values for the 'Wages, salaries, and commissions' variable. The line represents the mean, and the ribbon presents the 50th and 75th percentile's range.



over the first quartile since the first quartile was always $0. Those intervals present similar trends. Note that the sudden raise in 1989 is explained by the fact that for Quebec residents there were no tax data available prior to that year.

Also relating to earnings, the unemployment rates by age from 20 to 70 years old in 2015 were compared (Fig. 2). Here the unemployment rate is defined by the proportion of individuals for which both the earning and self-earning variables are below $3,500. Compared on an age basis, the synthetic unemployment rates are on average 4.3% higher than the original rates. The relative difference for the rate by sex are the same but the range of relative differences per year is slightly wider for men (-13.6% to 5.5%) than women (-11.8% to 5.4%). Overall, the two databases present some small differences. However, they are not a concern because the goal here was to confirm the data from the synthetic database was close enough to the original database so that microsimulation results would be similar (see below). In short, the comparison was pointing towards a successful replication.

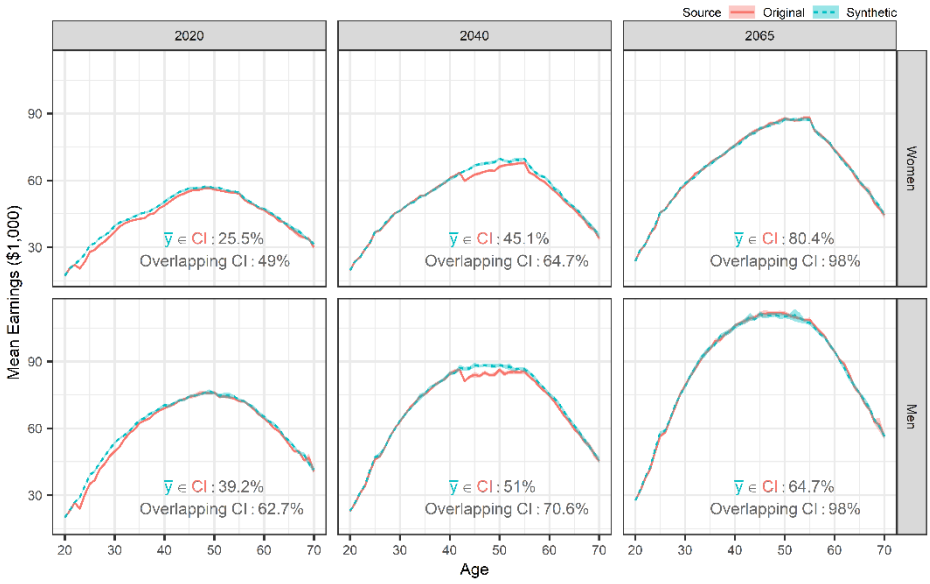
Figure 2 – Plot of the unemployment rate by sex and age for 2015.



3.3 Comparing Results from Microsimulation Runs

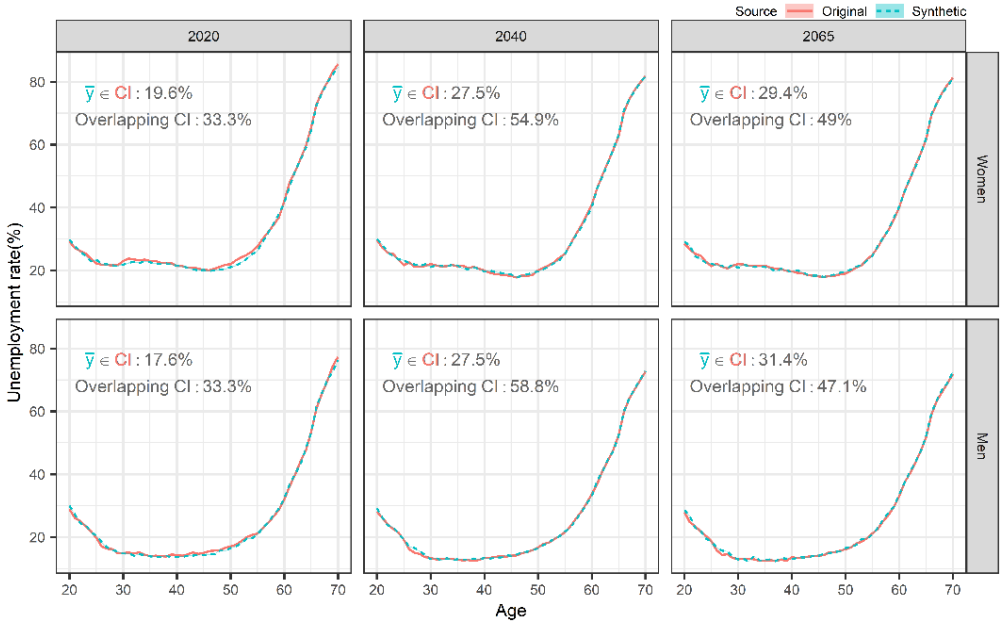
Another way to validate the quality of the synthetic data is to compare results obtained from the microsimulation model it was developed for. Running the PASSAGES microsimulation model on the full database requires time and mobilizes a lot of computing power, which was not available for this study. Therefore, for comparison purposes here, 10 random samples of 300,000 individuals were drawn, and each sample was used twice for a total of 20 microsimulation runs.

Figure 3 – Plot of the 'Wages, salaries, and commissions' variable by age, by sex and for 3 times forward in time (2020, 2040 and 2065). The line represents the mean, and the ribbon presents the 95% confidence interval around that mean.



In Fig. 3 the mean earning values are shown by sex and age for 3 different years forward in time (2020, 2040 and 2065) i.e., 5-year, 25-year and 50-year simulations as well as confidence intervals (CI) over those mean computed from the 20 microsimulation runs. In the long run, results from either the original database or the synthetic version seem to converge toward very similar distributions. The proportion of 95% CI that overlap increases over time. Part of the explanation for that comes from the CI around the means that are getting wider as there is more variability from microsimulation results in the long term. When analyzing results from microsimulations one must consider that having more samples and bigger samples would reduce the variability from the microsimulations as well as doing more replications would reduce the Monte Carlo variability. In Fig. 3 the proportion of original CI including the synthetic mean is also reported and is increasing over time. Overall, results show a relatively close fit between projections using either database. Note that there is a noticeable gap in 2020 and 2040 between the synthetic and original data. There is no current explanation for this dissimilarity. It could be due to differences in the databases or the result of the microsimulation model. Note that the PASSAGES microsimulation model itself is still being validated and improvements could be made in future versions.

Figure 4 – Plot of the unemployment rate by age, by sex and for 3 times forward in time (2020, 2040 and 2065). The line represents the mean, and the ribbon presents the 95% confidence interval around that mean.



Next, the unemployment rates were compared in Fig. 4. The rates are very similar for both databases. Here the proportion of original CI including the synthetic mean is also increasing over time, but the increase is less steady. The rate of CI overlapping in this case is even decreasing in 2065 compared to 2040. For the rate of unemployment, CIs are very narrow and their sizes even decrease over time. Despite some gaps, these two comparisons are showing reasonably close results.

The comparisons made here are limited in scope in contrast with the amount of information available. Further comparisons would show more results, and a more refined analysis for specific groups important to the CPP should be carried out. For example, the education level of individuals, their immigrant profiles or the type of pension benefits received could be considered. Investigating alternative scenarios involving parameters which are different from the default one (where among other things, inflation is null) would also be interesting. Indeed, one of the important features of the microsimulation model is the fact that simulation parameters are customizable.

4. DISCLOSURE RISK

4.1 Risk Management

It is well understood that the synthetic data can pose a certain disclosure risk for those in the original database. That risk can be understood in terms of the sensitivity of the information or the harm that it would cause an individual as well as the probability of correctly connecting a synthetic record with an individual in the population. It is important to note that the standard risk measures would assume that a match between databases would indicate a disclosure risk. However, in the case of a well-managed synthesis process, these would only be perceived risks. A careful evaluation of the risk management strategy was conducted.

Prior to looking at the disclosure risks from the synthetic database, some points are worth being highlighted about the confidential database (the original one). First, note that this database underwent many adjustments to improve its representativeness of the total Canadian population. Second, a certain number of imputations were made to complete records and make them consistent because information was not always compatible across data sources or coherent over time. For instance, for the historical dataset used in the donor imputation approach about 80% of those records had some imputation. These inconsistencies could also be the results of linkage errors. In summary, what is being called here ‘original data’ is not entirely what one would see in the various data sources that were used.

Regarding the sensitivity of the data, before the synthesis took place each variable underwent a review to evaluate its content, the necessity to keep it intact and the ways of altering it when needed. All outliers were identified at the beginning, and all unique patterns were examined with the intent of making them non-unique. For example, families with a person having more than 4 unions in their lifetime were excluded from the donor pool when life trajectories were selected, and income values were top-coded, bottom-coded and rounded. Lastly, the information on the original database underwent some aggregation which was also translated on the synthetic version.

Hence, once the data were further anonymized, the synthesis was undertaken. For most variables the modeling was done through CART. In the implementation of CART from the ‘synthpop’ package, synthetic values are randomly drawn from the members of a terminal node (Nowok *et al.*, 2016). To minimize the risk of replicating real records, a minimum number of observations in the terminal nodes was specified in most models. Different thresholds were tried, and the number was chosen so that the balance between synthetic data quality and risk disclosure was achieved.

4.2 Risk Assessment

Despite all the measures taken to lower the risk of disclosure the synthetic data may present; it is still possible for some synthetic records to be identical or very close to the original ones. Therefore, there was an evaluation made to quantify how frequently this was happening. Initially the synthetic database did not include as many variables as it does now, so our attention was focused on unique records, which are seen as riskier. With 12 variables from the cross-sectional layer, the rate of unique records was only about 3.8% for both original and synthetic databases. From those unique records in the original database, fewer than 1 out of 5 (19.7%) are also unique in the synthetic database. The more information was considered, the more unique the records became, but the probability of reproducing them also decreased. Overall, the unique matching rate was negligible. It is also worth mentioning that a unique record in the original dataset does not necessarily translate into a

unique in the population. Recall that the original dataset is based on the long-form census questionnaire, which is a subset of the population. This further reduced the identification risk.

There were several limitations in the risk analysis. There were too many variables to conduct a full matching exercise. Using part of the longitudinal information i.e., only years 2015, 2014 and 2013, and only the variables relating to 1) wages and salaries and 2) net self-employment income, 3) same information for the spouse (if available), 4) same for separated spouse (if one), 5) union status, 6) place of residence and 7) spouse age plus the cross-sectional information and familial information, 74,648 unique original records were matched uniquely to a synthetic record (0.9%). It is worth mentioning that 97% of those records had missing information or $0 for values such as wages and salaries or net self-employment income. In other words, those matches were not informative. Considering more years will decrease the number of matches; however, records with a lot of missing information (like children under 15 years old) or zero values for monetary variables will still potentially match across databases.

This whole exercise tends to indicate that the identification risk associated with the synthetic data file is negligible and would consist mostly of a perceived risk. Ultimately, any attributes taken from a linkage made with a set of variables would be considered either random or incorrect.

5. CONCLUDING REMARKS

Different lessons were learned throughout the completion of that project. Combining different techniques to generate synthetic data and re-using information were key elements in the development of this complex database involving life trajectories. Having a good understanding of the data was also essential to put all the pieces of the puzzle together and generate meaningful data. Knowing also how the data would be used and looking at specific analytic results allowed to assess the data utility of the synthetic data. Finally, synthesizing an evolving database posed its own set of challenges. Different adjustments were required because the original database changed along the way. Those changes were not directly applicable to the synthetic version as both versions do not share one-on-one link between records. Therefore, some additional steps and techniques were required to maximize the synthetic data fidelity.

Analytical comparison of both datasets is still ongoing. Exploration of results from the PASSAGES model, looking at smaller population groups and potential improvements to the synthetic database are on the radar as next steps. The comparison exercise confirmed, up to now, that the synthesis reproduced trends and characteristics consistent with the original data. At least results that were expected from the models that were created. Hopefully the data fidelity goes beyond the content that was deliberately preserved. Eventual feedback from users could help to establish that or guide further improvements to the synthetic database.

Throughout this project, Statistics Canada continued to develop its expertise in the field of synthetic data production. This paper described how Statistics Canada's most complex synthetic dataset so far was created, as well as how its analytical utility and its disclosure risk were assessed. The creation of synthetic data allows for the dissemination of useful statistical information while fulfilling an obligation to protect personal information. Here the synthetic data supports the model, allows users to run microsimulations and analyze the results as if they were looking at the results coming from the confidential database.

ACKNOWLEDGMENTS

The author would like to thank Steven Thomas, Claude Girard and Peter Wright for all their comments on this paper. The author would also like to thank the PASSAGES team from the Social Analysis and Modelling Division at Statistics Canada, in particular Jennifer Jones for answering numerous questions and Mahbubur Rahman for running the microsimulations.

REFERENCES

Bocci, C. & Beaumont, J.F. (2009). Synthetic Data Creation for the Cross National Equivalent File. In: *Proceedings of Statistics Canada Symposium 2009 Longitudinal Surveys: from Design to Analysis*, Statistics Canada, Ottawa.

Bonnéry, D. et al. (2019). The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-Level Multi-Agency Longitudinal Data. *Journal of research on educational effectiveness* **12**, 616-647.

Cranswick, K. and Hotton, T. (2022). Research Data Centres Program : Manager's Report Spring 2022. 37. Internal Document. Data Access Division, Statistics Canada.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*. 1st ed. New York: Springer.

Drechsler, J. & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*. **55**, 3232-3243.

Duncan, G. T., Elliot, M. & Salazar-González, J. (2011). *Statistical Confidentiality*. New York: Springer.

El Kababji, S. et al. (2023). Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. *JCO Clinical Cancer Informatics*, **7**, e2300116.

Gauvin, H. (2021). Generating smart deep files: the example of synthesizing hierarchical data. In: *Proceedings of Statistics Canada Symposium 2021 Adopting Data Science in Official Statistics to Meet Society's Emerging Needs*, Statistics Canada, Ottawa.

Government of Canada (2018). *Canada's 2018-2020 National Action Plan on Open Government | Open Government - Government of Canada*, https://open.canada.ca/en/content/canadas-2018-2020-national-action-plan-open-government, last accessed 2024/04/15

Kinney, S. K. et al. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review*, **79**, 362-384.

Nowok, B., Raab, G. M. & Dibben, C. (2016). synthpop : Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*. **74**, 1-26.

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* **36**, 1059-1066.

Statistics Canada. (2024). *The Daily — New retirement income microsimulation model now available*, https://www150.statcan.gc.ca/n1/daily-quotidien/240423/dq240423c-eng.htm, last accessed 2024/05/13