Creation of synthetic microdata in 2021 Census Transformation Programme (proof of concept)

Robert Rendell[*] & Cal Ghee[**]

[*] Office for National Statistics, Robert.Rendell@ons.gov.uk
[**] Office for National Statistics, Cal.Ghee@ons.gov.uk

Abstract: The Census Transformation Programme of the Office for National Statistics in England and Wales is investigating the use of synthetic data for a number of potential uses. One application is the creation of a household microdata sample. Due to concerns about preserving the confidentiality of individual respondents, we have not been able to provide a 2011 Census household microdata sample accessible outside secure research environments, with sufficient utility for users, using standard disclosure control techniques. The method being tested to create a household microdata file punches holes in a microdata sample, and uses the edit and imputation process from 2011 Census (using CANCEIS) to fill in the holes. This attempts to preserve the relationships between variables within households and individuals, but introduces sufficient uncertainty to mitigate disclosure risk. Methods are being investigated to test utility and risk in the resultant data.

This presentation will demonstrate issues we have to overcome, and the methods we are investigating to come up with a solution to provide useful, non-disclosive, microdata for a wider range of users, and how we are measuring risk and utility. Given these issues, this is just a proof of concept at this stage, to test whether the approach is feasible.

1 Introduction

Access to high quality microdata samples provides academic researchers and government departments the ability to conduct bespoke analysis which would not be possible using standard census tabular outputs. The dissemination of such microdata samples is dependent on the data contained complying with Principle 5 of UK Statistics Authority Code of Practice (CoP): Confidentiality. One key element of this principle is to ‘Ensure that official statistics do not reveal the identity of an individual or organisation’. Traditional disclosure control techniques implemented to protect the confidentiality of respondents include top coding, categorising continuous variables, coarsening categorical variables and even dropping variables completely. The effect of utilising such data coarsening techniques is a reduced risk of personal information being disseminated, but the increased security is made possible to the detriment of data utility.

In addition to the UK statistics Authority CoP, the Office for National Statistics and its employees have a legal obligation to protect personal information held by the office (Statistics and Registration Act 2007 Section 39).

Alongside the legal ramifications of personal information being disclosed, the impact on data quality can also be substantial if respondents feel that the information they provide will not be kept confidential. Respondents who don't have confidence that the information they supply is held securely may be less likely in the future to provide accurate information, or may even refuse to take part entirely, resulting in a lower response rate, which in turn could impact on data utility or the applicability of any inferences made from it.

One novel method of disclosure control which has been developing in recent years is the creation of synthetic, or partially synthetic data. Rubin (1987) originally proposed the concept of creating synthetic data using multiple imputation. Multiple imputation is the process of removing observed values from variables within a dataset and imputing them with synthetic data. Little (1993) built upon this concept by suggesting a closely related approach by creating partially synthetic data in which only sensitive information is imputed. There are several possible methods of creating multiply imputed synthetic data (explored further in section 2) but this study aims to test the feasibility of producing a synthetic microdata sample using CANCEIS (Canadian Census Edit and Imputation Software). This paper will explore the possibility of producing a partially synthetic detailed microdata sample from the 2011 Census for England and Wales database, with the potential to be widely available to international and commercial users which is currently not possible using conventional disclosure control methods and access restrictions.

Disclosure risk analysis is out of scope of this paper as in the interest of objectivity, risk analysis will be conducted independently of this study. Risk analysis will take the form of both a traditional unique analysis and an empirical intruder testing and will be reported separately.

2 Method

2.1 Data Preparation

For this proof of concept project, a sample area was chosen that was deemed broadly representative of the overall characteristics of England and Wales. From the complete 2011 Census database for this area, a 5% systematic random sample was drawn at the household level. This 5% sample represents our microdata sample. From this, some values for some variables were randomly selected and removed. Precise details of which variables and records were imputed cannot be disclosed at this point in order to maintain uncertainty of the identity of respondents. This being said, it can be stated that all records underwent some level of imputation; resulting in a dataset in which no record contained in the synthetic dataset matches the original observed values exactly.

2.2 Imputation

After sampling and removing values from some variables, the missing values were imputed using CANCEIS, an imputation engine designed and created by Statistics Canada.

2.2.1 CANCEIS

A traditional method of multiple imputation is to use a classification and regression tree (CART) method (Reiter, 2005). CART models are a flexible tool for estimating the conditional distribution of a univaraiate outcome given multivariate predictors (Breiman et al. 1984). Although the CART methodology is applicable to multiple imputation, it does have its limitations. The time taken to specify and create an imputation model that accounts for all relationships and interactions between variables and create an imputation that produces plausible values can be quite substantial. In addition, (Reiter, 2005) identifies that it is less straightforward to implement the CART approach when data are missing for multiple variables, and suggests that one approach to overcome this obstacle is to impute from chains of single variable trees conditional on previous imputations. This approach has potential to introduce error as if a relationship is omitted/incorrectly specified in the imputation model, using that variable to inform the imputation of further variables would result in inaccuracies across all subsequent imputations. Utilising CANCEIS has the potential to avoid such scenarios.

CANCEIS is the imputation software that was used for edit and imputation for the 2011 Census of England and Wales. In order to ensure the software is suitable for such a task it was extensively tested (Rogers and Wagstaff, 2005). One aspect of this suitability testing involved removing some observed values and then imputing them using CANCEIS and assessing the level of agreement between the observed and the imputed data, essentially creating a partially synthetic dataset. It was concluded that overall, the analysis provided strong evidence that CANCEIS recovered the marginal and joint distributions extremely well.

CANCEIS uses a nearest-neighbour methodology to identify donor records at the household level. Potential donors are selected from complete households that are as similar as possible to the household to be imputed. By utilising decision logic tables (DLT's) and user defined data dictionaries, it is possible to ensure that the imputed value is plausible (e.g. a 5 year old will not be given a marital status of 'married') and the marginal and joint distributions of the data are maintained. All edit rules and DLT's have been previously specified and created as part of the processing stage of the 2011 Census, therefore removing the time consuming and resource intensive requirement of creating a new imputation model.

2.3 Utility Analysis

Utility can be thought of in this context as how accurately the synthetic data matches the original observed data. It can be measured with narrow or global measures. Narrow measures are specific to individual variables, assessing how closely the distributions of specific variables match before and after synthesis whereas, global measures look at the overall distributions for all variables within the dataset.

Narrow measures of utility used in this study are a comparison of the marginal distributions for specific variables from the synthetic and original datasets. Kappa statistics (Fleiss, 1975) are used to provide an index of agreement, since they take account of the agreement between the true and imputed values which may have occurred by chance. If there is complete agreement between the true and imputed values then K = 1, if the observed agreement is greater than chance then K > 0 whilst if the agreement is equal to chance then K = 0 . Landis and Koch (1977) provide some arbitrary, but useful, benchmarks for Kappa. These are displayed in Table 1.

Table 1 Evaluation of observed Kappa values (Landis and Koch, 1977)

K	**Strength of agreement**
0.00	Poor
0.01 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

Analysis using narrow measures of utility identifies the movement between categories within variables. The limitation of this utility measure is that it only takes into account specific variables and can result in a dataset that is very accurate for some variables but other variables may be neglected.

One solution to this issue is to use a global measure of utility. The global measure of utility used for this project is the propensity score measure. First described by (Rosenbaum and Rubin, 1983), the propensity score is the probability of being assigned treatment, given covariate values x. Treatment assignment and covariates are conditionally independent given the propensity score. Therefore, when two large groups have the same distributions of propensity scores, the groups should have similar distributions of covariates.

Woo et al. (2009) describe a method whereby propensity scores can be utilised to assess the utility of a synthetic dataset. In this approach, the original and synthetic datasets are merged (by stacking) and creating a dummy variable T equal to ‘1’ for all records from the original dataset and equal to ‘0’ for all records from the synthetic dataset. Propensity scores can be estimated via a logistic regression of the T variable on functions of all variables x within the dataset. When the odds ratios are close to 1 (even odds) for all records in the original and synthetic data, the chances are that the distributions of the variables are similar. Not only is this approach a good indicator of utility, it is also a very useful diagnostic tool for identifying deficiencies in the synthetic data creation, whereby variables with significant coefficients in the logistic regression have different distributions in the original and synthetic data.

3 Results

3.1 Utility

3.1.1 Narrow Measures

For the purposes of this study, narrow measures refer to comparisons of the marginal distributions between the synthetic and original datasets. Figure 1 shows a histogram of the marginal distributions of the original and synthetic for the health variable. The histogram shows a strong level of agreement, supported by the Kappa scores (reported in section 3.1.2), between original and synthetic values.

Fig 1 Marginal distributions for health

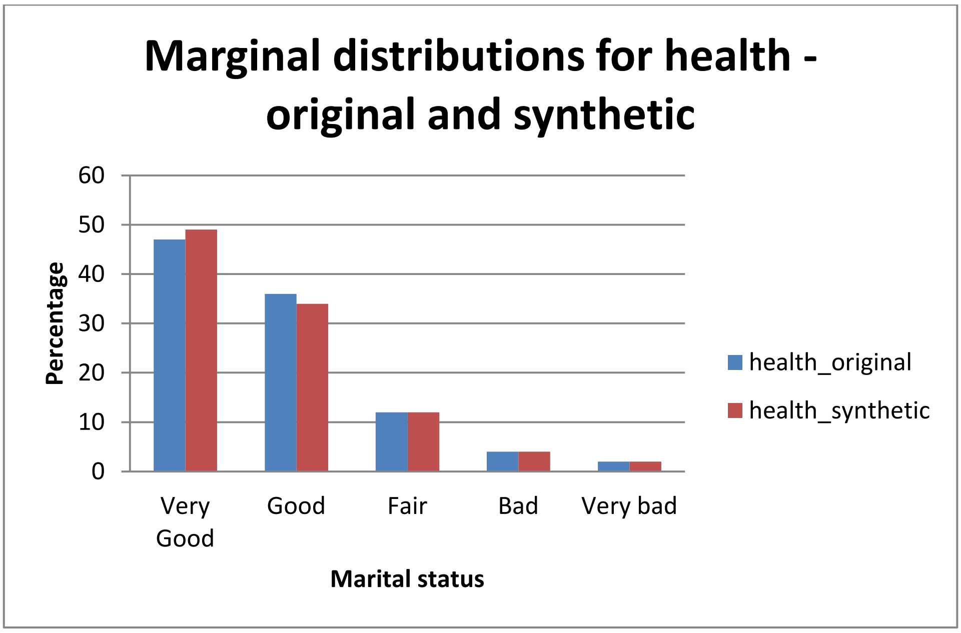

Figure 2 shows a histogram of the marginal distributions of the original and synthetic for the marital status variable.

Fig 2 Marginal distributions for marital status

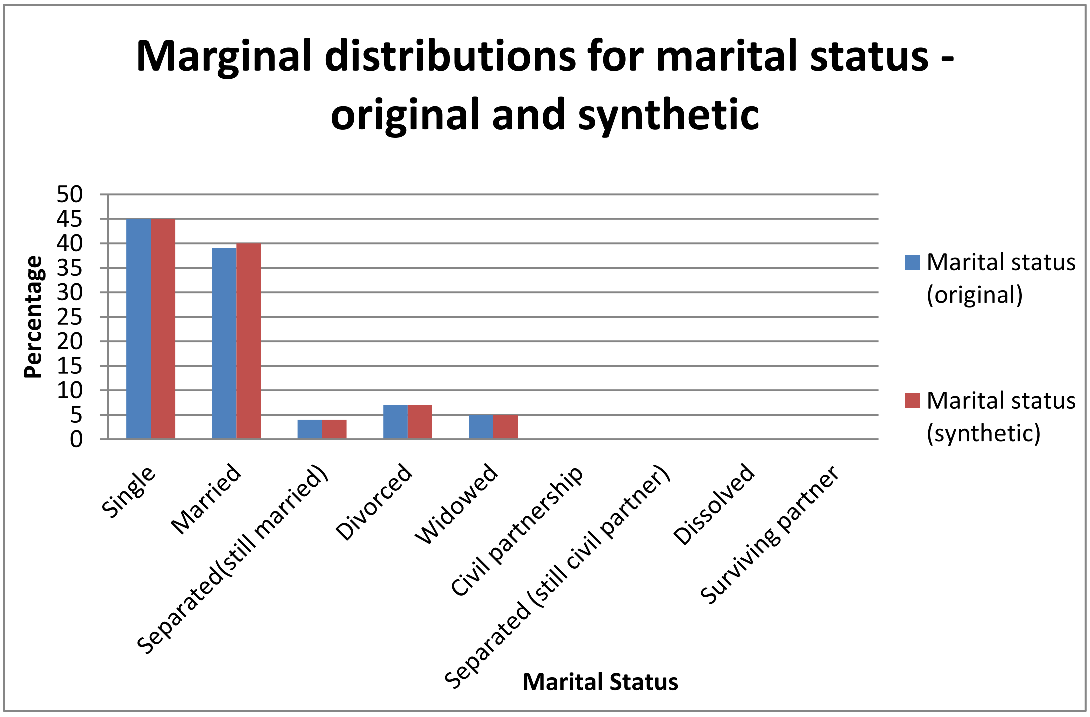

Figure 2 shows a very high level of agreement between the original and synthetic values.

3.1.2 Kappa scores

Table 2 shows the Kappa scores for several key variables within the dataset. The majority of variables fall within the ‘substantial’ and ‘almost perfect’ categories of agreement described by Landis and Koch (1977). Sex and Age have the lowest levels of agreement.

Table 2 Kappa index of agreement between the distributions of the original and synthetic datasets

Variable	Kappa Score (K)
Health	0.77
Marital Status	0.83
Sex	0.69
Age	0.68
Carer	0.79
Student	0.86
Year last Worked	0.92

3.2 Global Measures

Figure 3 shows odds ratios of variables from the merged (stacked) original/synthetic dataset. The results show that the vast majority of the variables included fall very close to 1, suggesting similar distributions between the original and synthetic datasets. The only variable with a significant difference in the distributions is the student variable.

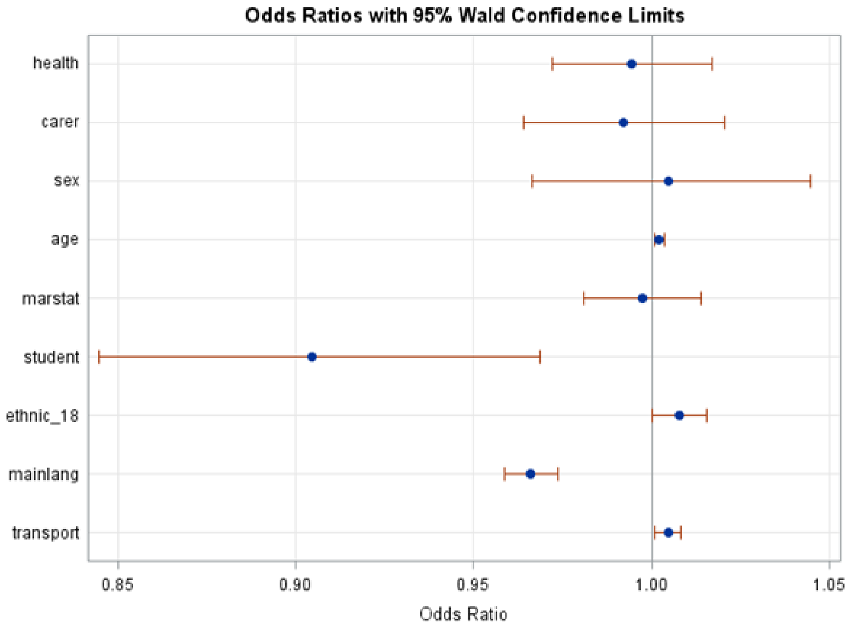

Fig 3 Odds ratios with 95% Wald Confidence Limits

4 Discussion/Conclusion

Overall, the distributions of variables in the synthetic data closely match the original data, suggesting a high level of utility. The maintenance of the marginal distributions seen in fig 1 and fig 2 are indicative of an accurate imputation model. The odds ratios displayed in fig 3 all fall close to 1 which suggests an equal probability of a variable being ‘assigned’ to the synthetic or original dataset. The ‘Student’ variable is the only variable which has odds falling closer to 0.9 than to 1, which isn’t a substantial variation, but when compared to how other variables are clustered closer to the even odds it does stand out. Further investigation is required to identify the exact cause of this anomaly, but the kappa scores for this variable are still very reassuring, falling in the ‘almost perfect’ category of agreement with K = 0.86.

The kappa scores identified a ‘substantial’ or ‘almost perfect’ level of agreement for all variables synthesised. The variables with the lowest level of agreement were sex and age. This is potentially due to the high weighting attributed to these variables during imputation in CANCEIS. Sex and age are identified as good predictors of other characteristics and therefore contribute highly in the imputation. By removing these values, the variance in imputation increases and the accuracy of the imputation for these variables could be impaired. Sex and age are also considered highly disclosive at the household level, so the lower level of agreement in these variables increases uncertainty, which is desirable from a disclosure control perspective.

One obstacle synthetic data needs to overcome is the perception some researchers, government departments and other users have of using data that has been synthesised. Some individuals and organisations find the concept of using synthetic data difficult, and insist on using only real data. It is however important to consider that many of the traditional disclosure control methodologies have a substantial impact on data utility (Lane, 2007 and Winkler, 2007) and any comparisons to synthetic data should therefore be made using datasets to which other disclosure control techniques have been applied rather than the original data. The similarities in the original and synthetic data’s marginal distributions and propensity score analysis should go some way to instilling confidence in the utility of a synthetic microdata sample produced using this methodology.

The benefit of using CANCEIS as the imputation engine cannot be overstated. An obstacle facing the production of widely available synthetic data is that agencies argue that ‘developing synthetic data is labour intensive, takes too long and requires experts that are familiar with the data, have a strong knowledge of Bayesian statistics and excellent modelling skills to generate synthetic data with a high level of data utility' (Drechsler, 2011). By using CANCEIS, which has already been specified, tested and includes strict edit rules to ensure plausibility of outputs, the time consuming task of building a specific imputation model can be averted, and the results of this study suggest that the synthetic data produced have a high level of utility.

This work is still in the early 'proof of concept' stage and disclosure risk is yet to be tested. The accuracy of the imputation suggests a high level of utility, but whether this sufficiently protects the identity of respondents is yet to be seen. The benefit with synthetic data is the doubt placed within the mind of an ill intentioned user. The fact that all records contained within the sample have undergone some level of imputation will go some way to discouraging a potential intruder, but whether this is sufficient needs to be assessed independently using both traditional disclosure risk techniques and an empirical intruder test.

Overall, the results of this study suggest using CANCEIS is a time and cost effective method of generating synthetic data. From a user's perspective, the synthetic data can be considered to have a high level of utility and suggests that using CANCEIS as an imputation engine for creating synthetic microdata is feasible. The next step is to carry out disclosure risk assessment and then iterating towards the most appropriate risk and utility balance.

References

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation.* New York.

Lane, J. I. (2007). Optimizing the use of microdata: An overview of the issues. *Journal of Official Statistics* 23, 299-317.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* 21, 441-462.

Rogers, S. and Wagstaff, H. (2005). Application of CANCEIS to 2001 Census data. *Office for National Statistics Internal Report*

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41-55.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*, *462-468.*

Statistics and Registration Service Act (2007)
http://www.legislation.gov.uk/ukpga/2007/18/section/39

UK Statistics Authority Code of Practice for Official Statistics (2009)
https://www.statisticsauthority.gov.uk/wp-content/uploads/2015/12/images-codeofpracticeforofficialstatisticsjanuary2009_tcm97-25306.pdf

Winkler, W. E. (2007b). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Tech. rep., Statistical Research Division, U.S. Bureau of the Census.

Woo, M., Reiter, J. P., Oganian, A., Karr, A. F. (2009) "Global measures of data utility for microdata masked for disclosure limitation", *Journal of Privacy and Confidentiality*, 01 (01), 111-124.