Putting synthetic people in place: creating synthetic data for spatial analysis at the individual level

(QCumber-EnvHealth project: WP3 Accessible health data)

Beata Nowok and Chris Dibben

Administrative Data Research Centre - Scotland, University of Edinburgh, School of GeoSciences, Drummond Street, Edinburgh EH8 9XP, United Kingdom, {beata.nowok, chris.dibben}@ed.ac.uk

Abstract. Improved availability of microdata with detailed geographic information, combined with appropriate methods of spatial analysis, enables better understanding of the relationships between people and their physical and social environments. It also makes it possible to evaluate the effects of policy interventions at the individual or any higher level. Access to microdata for small geographic areas is especially problematic and therefore very restricted due to privacy concerns. In this paper we present an approach to share such data by producing, using the R package *synthpop*, their completely synthetic versions. They are generated from probability distributions and as such contain artificial units only. Synthetic data substantially limit the risk of identification while maintaining most of the research value of the data. There is, however, no warranty that all statistical features are preserved.

1 Introduction

The great potential of microdata for policy and scientific purposes is very often unfulfilled due to highly restricted access to personal information provided under promises of confidentiality. This is particularly true of detailed geographic data, since precise information about spatial location can be perfectly identifying. Very often, therefore, geographic data are aggregated before release and much spatial modelling is constrained in spatial resolution and limited to ecological regressions with related issues such as the modifiable unit problem. Access to individual level data with detailed information on geographic location would enhance research and our knowledge about the impact of the physical and environmental context of human activities and a successful balance between required precision and the confidentiality protection needs to be found.

Generating and releasing synthetic data is one of the most promising approaches for disseminating individual level data without compromising confidentiality (Burgette and Reiter, 2013; Drechsler and Hu, 2015; Machanavajjhala et al., 2008; Quick et al., 2015; Paiva et al., 2014; Sakshaug and Raghunathan, 2014; Wang and Reiter, 2012; Yu et al., 2017). In most applications partial synthesis strategy is adopted, when real individuals are assigned synthetic locations. Geographic information is

often represented by latitude and longitude but methods for aggregated spatial areas have been also proposed.

In this article we present a synthetic data approach to release data on individuals for small geographic areas in order to enable spatial analysis at less aggregate level. We outline methods for producing spatial synthetic data, suggest a metric for evaluating their utility and identifying synthesised data with the best features for a particular use case. We use the *synthpop* package (Nowok et al., 2016) for R (R Core Team, 2017) and generate both partially and completely synthetic data using classification and regression trees (CART) models (Breiman et al., 1984).

2 Data

In this study we use a linked data set (24 variables) that includes maternity data (43,265 records) for mothers living in the city of Glasgow council area between 2009-2014 and information on various characteristics of their place of residence at more detailed spatial scale. The availability of geographic information is summarised in Table 1. The mother's data zone of residence is the finest level of geography available in a safe haven environment but environmental variables were linked at a postcode level. Table 2 presents attributes according to the level of aggregation. A complete list of all variables with names that are used for plotting can be found in the Appendix A.

Table 1. Geography of the city of Glasgow council area.

Geographic identifier	# of areas	Availability
Intermediate zone 2001	133	Available in a safe haven
Data zone 2001	693	Available in a safe haven
Postcode	14,049	Not available; some attributes can be linked at postcode level and made available in a safe haven

Information on mothers, their pregnancies and babies comes from the 'SMR02 - Maternity Inpatient and Day Case' dataset collected by the NHS Scotland and from the National Records of Scotland's data on births. Mothers are described by their demographic, economic and behavioural characteristics and variables related to their pregnancies (see Table 2). Data on air pollutants concentration were provided by the Cambridge Environmental Research Consultants (CERC) and data on green spaces and tobacco retailers by the Centre for Research on Environment, Society and Health (CRESH).

Table 2. Description of attributes.

Level of attributes	Description
Individual level (mother/pregnancy/ baby)	Age, marital status, mother's height, parity (whether first pregnancy), income, smoking status during pregnancy, year of delivery, baby's weight, baby's sex, indicator of low birthweight ($<$ 2,500g), gestational age, mode of delivery, indicators of a preterm birth ($<$ 32 weeks, $\geq$ 32 and $<$ 36 weeks)
Data zone level	Scottish Index of Multiple Deprivation (SIMD) rank, 2006
Postcode level	% greenspace within various buffers (100m, 250m, 500m), density of tobacco retailers within 800m, pollutant (NO_2, PM_{10}, $PM_{2.5}$) concentrations averaged over duration of pregnancy

3 Synthesising methods

Data are synthesised using sequential modelling based on CART models as implemented in the *synthpop* package for R. The sequential approach is adopted in many implementations for generating synthetic data and it consists in modelling each variable by all variables that are earlier in the synthesising order. It is preferred to joint modelling not only because of the ease of implementation but also because of their flexibility to apply methods that take into account structural features of the data such as logical constraints or missing data patterns. The basic idea of CART models is to recursively split a data set into groups with increasingly homogeneous outcome. The splits are specified as yes-no questions referring to the predictor space. The values in each final group approximate the conditional distribution of the predicted variable for units with predictors meeting the criteria that define that group. The synthetic values are generated by sampling from an appropriate group. CART models were suggested for generation of synthetic data by Reiter (2005) and they have been successfully used in many applications, also to model point-referenced geographic area (Wang and Reiter, 2012; Yu et al., 2017).

We produce partially and completely synthetic data sets. In the former case, only individual level variables are synthesised (13 variables) and data zone and postcode level variables are used as predictors in synthesising models (see Table A for details on variables). In the latter case, individual, data zone and postcode level variables are all synthesised. The synthesising order can be consulted in the Appendix A. The results presented in the next section refer, however, only to partially synthetic data sets.

In order to identify a strategy that is most effective in replicating spatial and non-spatial relationships we consider three approaches with different synthesis strategies that vary in their geographic stratification. For each approach we generate ten synthetic data sets ($m = 10$). The approaches are summarised in Table 3. In Approach 1 we do not use any stratification and intermediate zone and data zone variables

are synthesised at the very end. This approach led to some computational problems and two strategies were implemented to overcome them. In the first one, the intermediate zone variable is excluded from the predictors of data zone. In the second one, synthetic data zones are obtained by bootstrap sampling from the original data zones within each intermediate zone. In Section 4 only results from the former method are reported. Approach 2 involves stratification by intermediate zone, which means that synthesis is carried out in each intermediate zone separately and different synthesising models estimates can be obtained in different strata. An average number of observations in a stratum is around 300. Synthetic data zone values are generated within stratum at the very end of the synthesising process. Here and in Approach 1 synthetic geographic indicators are generated as the last ones, because categorical variables with a large number of unique values cause computational issues when used as predictors. In Approach 3 synthesis is stratified by data zone with an average number of births equal to 60.

Table 3. Synthesis strategies.

Approach no.	Synthesis strategy
Approach 1	Overall synthesis with intermediate zone and data zone synthesised at the very end*
Approach 2	Synthesis stratified by intermediate zone; data zone synthesised at the very end
Approach 3	Synthesis stratified by data zone

Notes: *To overcome memory problems two strategies were implemented: a) intermediate zone excluded from predictors of data zone, b) bootstrap sampling from the original data zones within each intermediate zone

In all the approaches data zone is the finest level of geographic information. To obtain synthetic data for finer geography, e.g. post code level, synthetic mothers can be distributed randomly within data zones based on pollution level.

4 Evaluation of synthetic data utility

We evaluate synthetic data sets using goodness-of-fit statistic for two- and three-way cross-tabulations that include geographic variable (data zone or intermediate zone) as one of the dimensions. The basic idea is that the overall resemblance between two data sets can be assessed by comparing sufficient number of multivariate tabulations. We use the statistic proposed by Voas and Williamson (2001):

$$VW = \sum_{i=1}^{k} \frac{(O_i - S_i)^2}{(O_i + S_i)/2} \tag{1}$$

where i is a cell indicator in tables with a total number of cells equal to k. S_i and O_i are counts in a table based on synthetic and observed data respectively. This

statistic avoids the problem of zero cells in one of the tables. Where observed and synthesised cells are both zero, they do not contribute to the sum. If the synthesising model is correct this measure should have chi-square distribution for large samples. A low value of the statistic equates to a good fit.

Figure 1 and Figure 2 shows Voas-Williamson statistic (average over ten synthesis) for two- and three-way cross-tabulations with data zone and intermediate zone respectively for Approach 1 (*none*), Approach 2 (*intzone2001*) and Approach 3 (*datazone2001*). Results are ordered by increasing value of the statistic for Approach 3. We can see that overall synthesis produces the worst fit for both levels of geography and therefore in further analysis we will focus on the other two approaches only. As we could expect, synthesis stratified by data zone gives, except for a few exceptions, better results at data zone level than synthesis stratified by intermediate zone (see Figure 1). Note, however, that it does not translate into a better fit at higher geographical level. Some relationships between variables that are present at intermediate zone level are not captured when estimated at data zone level (see Figure 2). The most problematic cross-tabulations include indicator of a preterm birth and indicator of low birth weight.

Figure 3 and Figure 4 show two-way cross-tabulations results for each of the ten synthetic data sets separately in order to assess variability between them, since in some instances only one synthetic data set is released. From the figures, it is clear that there are differences between synthesised data sets generated using the same approach and we can choose the data set that preserves best the relationships we are interested in. A general strategy would be to select a data set that on average has the lowest Voas-Williamson ratio and performs well for critical dependencies.

Figure 5 shows box plots for percentage difference in proportion of smokers in intermediate zones between real and synthetic data sets synthesised using Approach 2 (*intzone2001*) and Approach 3 (*datazone2001*). Two copies for each strategy are presented. Again, we can see that utility of the synthetic data varies between different copies of the data synthesised following the same strategy. In addition, there are some intermediate zones with significant discrepancies between real and synthetic data, which requires further investigation if percentage of smokers is of our interest.

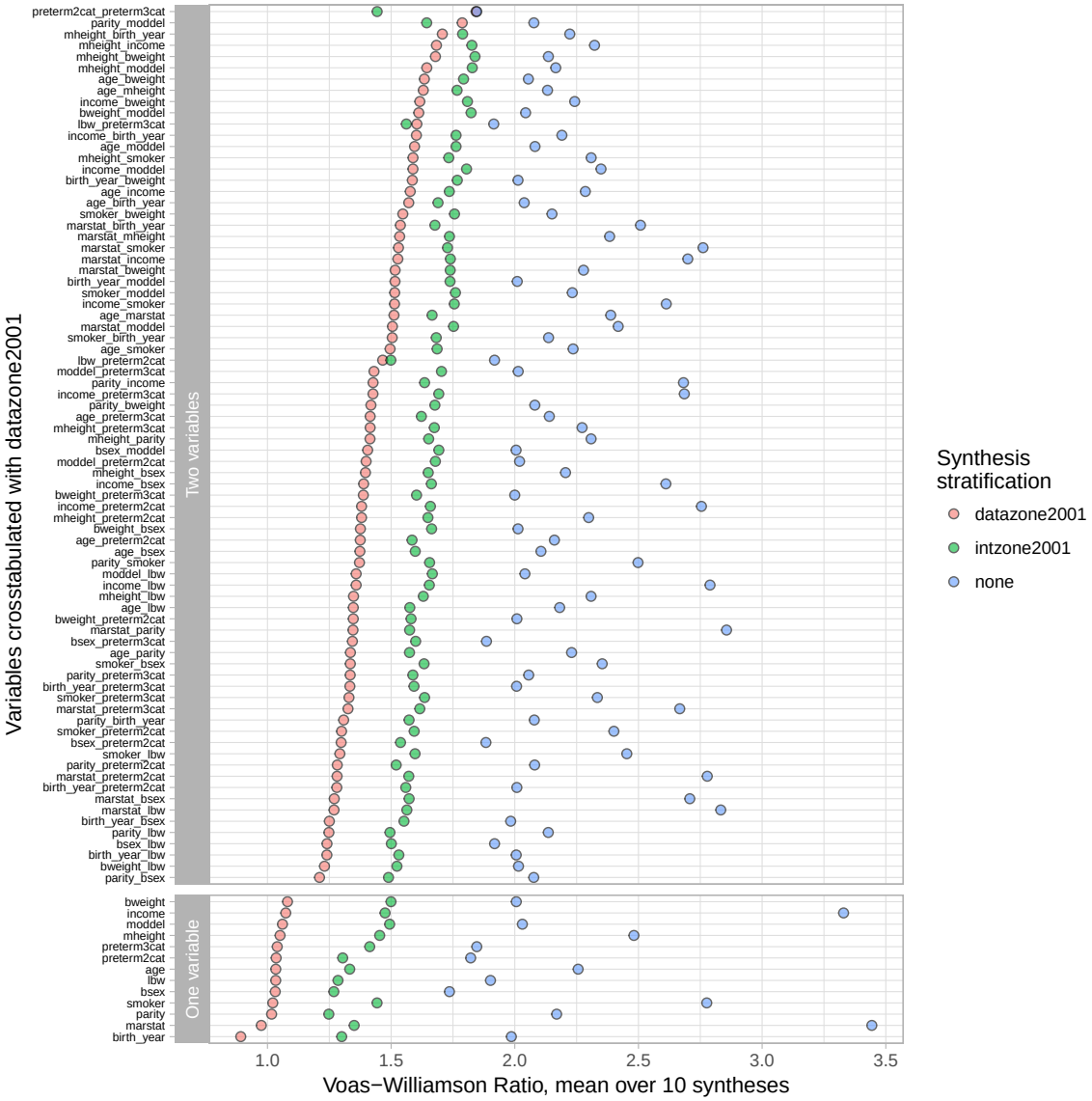

Figure 1. Voas-Williamson ratio for two- and three-way cross-tabulations with data zone; mean over ten syntheses.

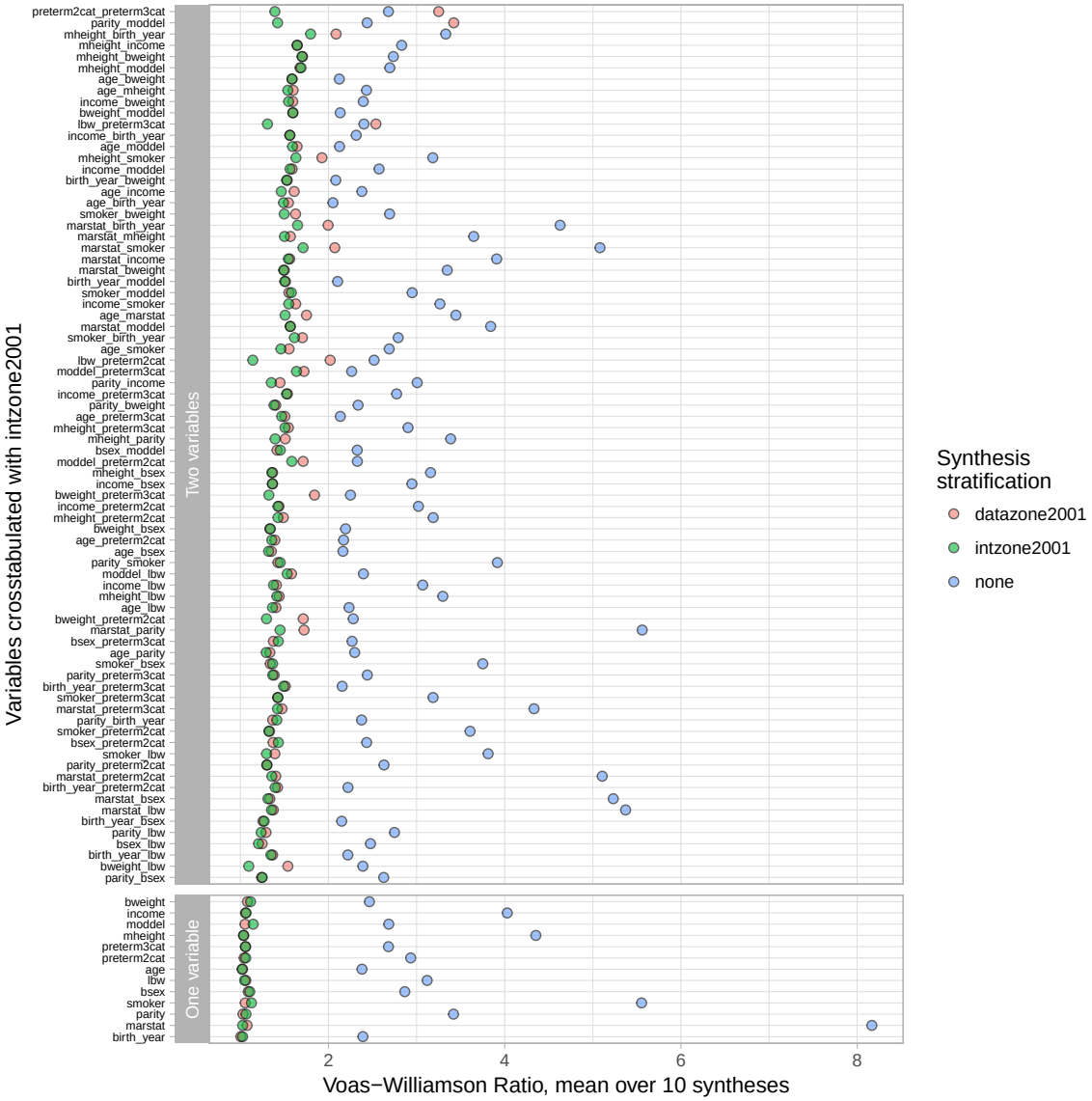

Figure 2. Voas-Williamson ratio for two- and three-way cross-tabulations with intermediate zone; mean over ten syntheses.

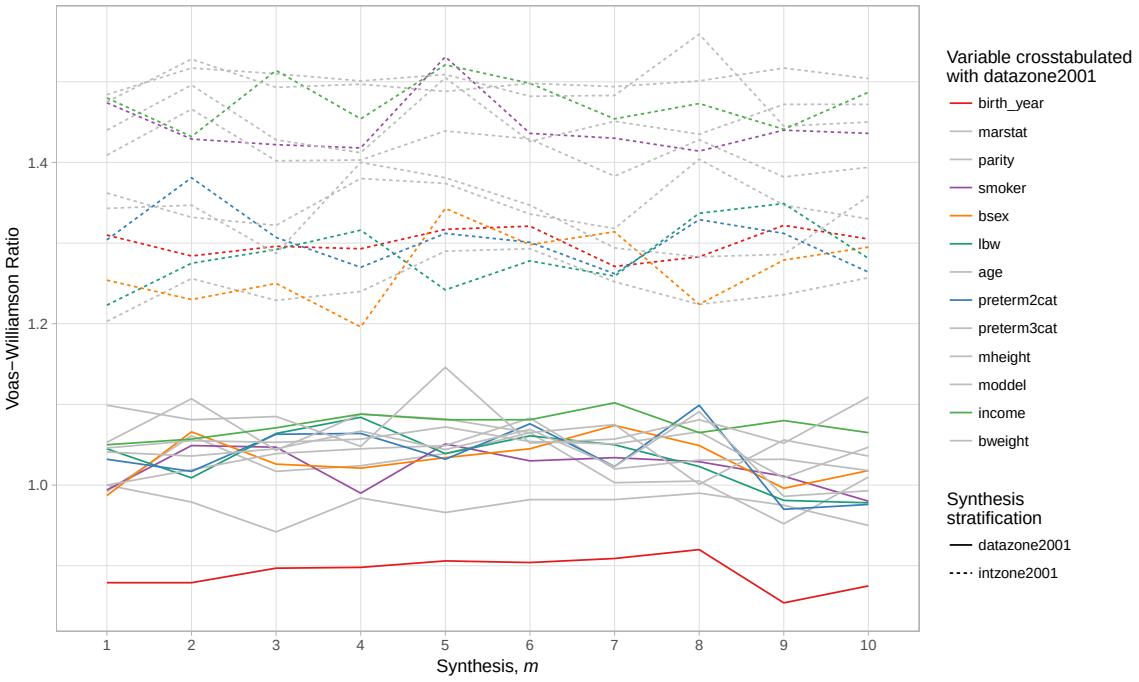

Figure 3. Voas-Williamson ratio for two-way cross-tabulations with data zone for ten synthetic data sets.

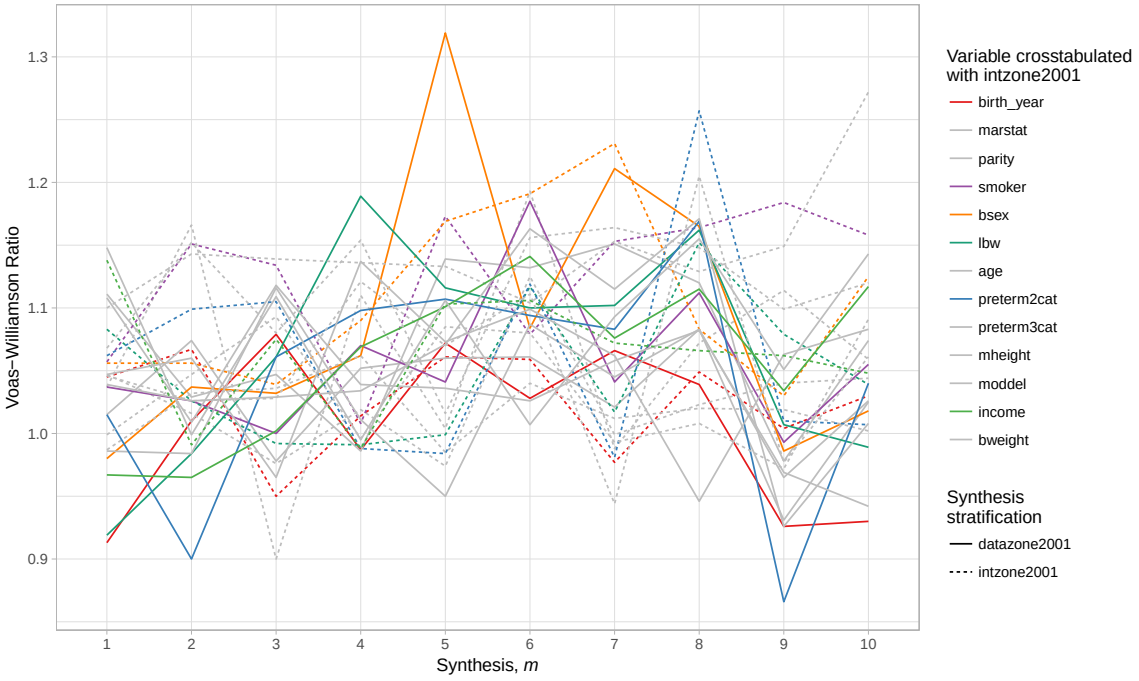

Figure 4. Voas-Williamson ratio for two-way cross-tabulations with intermediate zone for ten synthetic data sets.

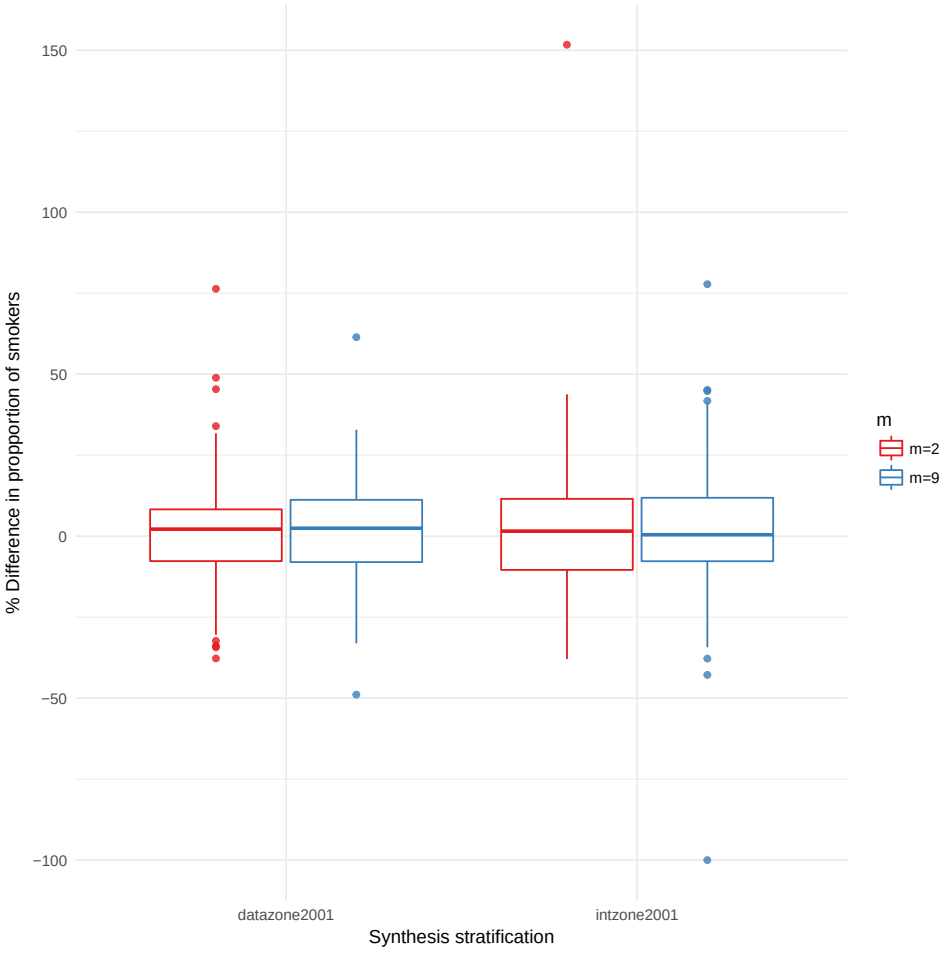

Figure 5. Box plots for % difference in proportion of smokers in intermediate zones between real and synthetic data sets synthesised using two different stratification strategies; two copies for each strategy ($m = 2$ and $m = 9$ denotes here 2nd and 9th synthesised data set).

5 Conclusions

Synthetic data that replicates the structure and statistical properties of the original data set without using original content offer a way to enhance the use of confidential microdata, also with detailed geographic identifiers. They can be used to conduct reliable inference in many analyses but in some cases they will give approximate results at best. Further research is needed to investigate their usefulness when advanced analytical methods are applied for data analysis. In addition, procedures for disseminating such data have to be developed concurrently to exploit fully the potential of synthetic data. Synthetic data generated in our example offer, when released, possibility to estimate individual level spatial models or to investigate impact of different policy scenarios at individual level, which is otherwise impossible outside safe haven. The choice of a specific synthetic data set for release from all the versions created depends on the intended level of geographical analysis and the desirable risk-utility profile. Nonetheless, a spatially stratified synthesis is a preferred approach over an overall synthesis.

6 Acknowledgements

The authors would like to thank the NHS National Services Scotland and QCumber-EnvHealth project partners for providing the data for this study.

References

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

Burgette, L. F. and J. P. Reiter (2013). Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian Anal. 8*(2), 453–478.

Drechsler, J. and J. Hu (2015). Generating synthetic geocoding information for public release. *Joint UNECE/Eurostat work session on statistical data confidentiality*, Helsinki, Finland, 5-7 October 2015.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, pp. 277–286.

Nowok, B., G. M. Raab, and C. Dibben (2016). synthpop : Bespoke creation of synthetic data in R. *Journal of Statistical Software 74*, 1–26. Available at `https://www.jstatsoft.org/article/view/v074i11`.

Paiva, T., A. Chakraborty, J. Reiter, and A. Gelfand (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine 33*(11), 1928–1945.

Quick, H., S. H. Holan, C. K. Wikle, and J. P. Reiter (2015). Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics 14*, 439 – 451.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reiter, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics 21*(3), 441–462.

Sakshaug, J. W. and T. E. Raghunathan (2014). Generating synthetic microdata to estimate small area statistics in the American Community Survey. *Statistics in Transition 15*(3), 341–368.

Voas, D. and P. Williamson (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling 5*(2), 177–200.

Wang, H. and J. P. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. *Ann. Appl. Stat. 6*(1), 229–252.

Yu, M., J. P. Reiter, L. Zhu, B. Liu, K. A. Cronin, and E. J. R. Feuer (2017). Protecting confidentiality in cancer registry data with geographic identifiers. *American Journal of Epidemiology 186*(1), 83–91.

A Appendix - A list of variables and synthesising order

Synthesising order (excluding *datazone2001* and *intzone2001*): *simd_inc_rank_2006*, *tobdens800*, *pc100_total*, *pc250_total*, *pc500_total*, *no2_mean9*, *pm10_mean9*, *pm25_mean9*, *age*, *marstat*, *mheight*, *parity*, *income*, *smoker*, *birth_year*, *bweight*, *bsex*, *gest*, *moddel*, *lbw*, *preterm2cat*, *preterm3cat*

No.	Variable name	Description
1	*age*	Age
2	*marstat*	Marital status
3	*mheight*	Mother's height
4	*parity*	Parity (whether first pregnancy)
5	*income*	Income
6	*smoker*	Smoking status during pregnancy
7	*birth_year*	Year of delivery
8	*bweight*	Baby's weight
9	*bsex*	Baby's sex
10	*lbw*	Indicator of low birthweight ($<$ 2,500g)
11	*gest*	Gestational age
12	*moddel*	Mode of delivery
13	*preterm2cat*	Indicator of a preterm birth ($<$ 32 weeks)
14	*preterm3cat*	Indicator of a preterm birth ($<$ 32 weeks, $\geq$ 32 and $<$ 36 weeks)
15	*simd_inc_rank_2006*	Scottish Index of Multiple Deprivation (SIMD) rank, 2006
16	*pc100_total*	% greenspace within 100m buffer
17	*pc250_total*	% greenspace within 250m buffer
18	*pc500_total*	% greenspace within 500m buffer
19	*tobdens800*	Density of tobacco retailers within 800m
20	*no2_mean9*	NO_2 concentration averaged over duration of pregnancy
21	*pm10_mean9*	PM_{10} concentration averaged over duration of pregnancy
22	*pm25_mean9*	$PM_{2.5}$ concentration averaged over duration of pregnancy
23	*datazone2001*	Data zone
24	*intzone2001*	Intermediate zone