The use of synthetic data to support researchers

March 2020

By Robert Beatty

Northern Ireland
Statistics and Research Agency
Gníomhaireacht Thuaisceart Éireann
um Staitisticí agus Taighde

Should statistical offices produce synthetic data to enable researchers to conduct modelling work otherwise conducted in safe settings?

Introduction

1.While statistical outputs (typically tables) meet the vast majority of Census users' requirements, some users require access to individual level data (henceforth termed 'microdata'), usually to enable the development of statistical models. The disclosure risks in providing such access are obvious. The standard approaches are to provide ready access to low-detail microdata (teaching files) or to controlled access to detailed microdata data (secure settings). The former (teaching files) has limited analytical potential, while the latter (secure settings) can be cumbersome to users. An alternative approach is the generation of synthetic data that mimics (and retains the statistical information within) real data but being 'not real' (apparently) removes / reduces the disclosure risk. This short paper discusses synthetic data further, and makes recommendations in a local context to NISRA.

Respecting confidentiality – an over-riding priority

2.Respecting the confidentiality of respondents to statistical inquiries is a key duty of all national statistical offices. It is covered in the Code of Practice for Statistics[1] (Section T6.1 of version 2.0, February 2018) and is implicit in the Data Protection Act 2018[2] (first principle that processing must be lawful and fair, and sixth principle that data must be processed in a secure manner). The Census is a special case in that the population are required by law (The Census Act (Northern Ireland) 1969[3]) to provide responses [Section 3 (1) e]; this is balanced by the Act explicitly stating that Census returns must be kept "secret" [Section 6]. Thus it is a legal requirement that Census returns must be kept confidential.

The tension between the desire publish information while protecting confidentiality

3.All official statistics offices encounter a natural tension brought about by two intuitive objectives, to exploit the information in data to the maximum extent and – simultaneously - respect the confidentiality of data subjects. This is reflected in NISRA's recent 2021 Census Proposals paper[4] (April 2019). Section 1.6 of the Proposals paper describes NISRA's key strategic objectives for the Census, of which the first two are;

To provide high quality, value for money, fit-for-purpose statistics that meet user needs, and which are consistent, comparable and accessible across the UK; and

To protect, and be seen to protect, confidential Census personal information

4.It stands to reason that as progressively more information is published (usually in the form of statistical abstracts from the data), the greater the risk of disclosure of some data that can be associated with a data subject.

The user need for access to microdata

5.One simple consequence of the desire to protect respondent confidentiality is that most Census outputs take the form of statistics ('abstracts' in the language of the Census Act). This restriction of outputs to (mostly) tables leads to a subsequent restriction on the analyses that users can perform. For example, the true relationships between variables may be visible only at the level of individual subjects and disguised at aggregate level. Access to microdata helps researchers avoid the ecological fallacy[5] and consider the potential impact of the Modifiable Areal Unit Problem[6].

6.To meet this demand NISRA, in common with other official statistical offices, does allow access to microdata in two ways – readily available safe data (teaching files) and controlled access to detailed data (safe settings)[7].

Teaching files

7.The first option is to make the data very safe and publish the microdata; this has led to the 'microdata teaching file' which consists of Census records for a sample of the population – internationally this is typically a 1% sample. The protection of respondent confidentiality is ensured through the data set having restricted content for each individual record, for example age is reported as one of a small number age-bands (8 in Northern Ireland) and, in particular, no information is provided on the location of the Census record within Northern Ireland.

8.The microdata teaching file is described by NISRA as a taster for what is available. It is worth noting that even this 'light' dataset could contain a record that could be identified with a single person, and Census Offices can take further steps to prevent this – for example, a check for 'population uniques' could be performed and such records are removed from the microdata sample.

The development of safe settings

9.In common with other Census Offices, NISRA acknowledges that the teaching file will not meet user need for detailed microdata upon which to develop sophisticated statistical models. This is the same tension as described earlier; microdata that would enable such models would be inherently disclosive. In line with other statistical offices, NISRA has developed a secure environment termed a secure data-laboratory. Users wishing to analyse such microdata must conduct their analyses within the NISRA data laboratory. There are further security factors involved including; the user must go through an approval process, likewise the project must be approved, the laboratory contains no output media, all statistical outputs are vetted by NISRA staff prior to release to the user.

The demand for microdata in users' own environments

10. The procedures around the secure data laboratories are complicated, but this merely reflects the sensitive nature of the confidential data held in the laboratory. In particular, access is typically restricted to a specific physical location which researchers must access. Not surprisingly, this had led to demand (from some users) for access to microdata that the users can analyse 'locally' on their own machines. Whilst users can be required to sign confidentiality guarantees and so forth, releasing microdata outside the data laboratory is, in practical terms, making the microdata publicly available. Notwithstanding this, a number of secure data facilities across the UK and Ireland (including official statistics offices) are now providing 'secure remote access' as a practical alternative to physical presence in the secure data facility.

11. The tension between the Statistics Offices (confidentiality is paramount) and the researchers desire for ready access is obvious. This has led to the work investigating the development of 'synthetic' data, which reflects the properties of the original 'real' data to a degree that is sufficient to enable valid analyses but (somehow) protects subject confidentiality.

12. Having outlined the rationale for the production of synthetic data, this short paper now discusses the potential for synthetic data to be developed and used in a NISRA context.

A classification of synthetic data

13. Firstly, a definition of synthetic data from the US Census Bureau (taken from the ONS Working Paper[8]) –

"Synthetic data are microdata records created to improve data utility while preventing disclosure of confidential respondent information. Synthetic data are created by statistically modelling original data and then using those models to generate new data values that reproduce the original data's statistical properties. Users are unable to identify the information of the entities that provided the original data."

14. The Office for National Statistics (ONS) has usefully produced a classification (or spectrum) of synthetic data[8], reproduced below.

Table 1 - A classification of synthetic data (Source – ONS[8])

Spectrum level	Synthetic classification	Description
1.Structural	Synthetic	Preserves data types and formats. No analytic value. No disclosure risk.
2.Valid	Synthetic	As structural, but removes implausible records (e.g. no employed infants). No analytic value. No disclosure risk.
3.Univariate Plausible	Synthetically augmented	Replicate univariate distributions. No associations preserved. Minimal analytic value. Non-zero but minimal disclosure risk.
4.Multivariate plausible	Synthetically augmented	As for plausible. Replicate multivariate distribution 'loosely' at high geographic area. High disclosure risk.
5.Multivariate detailed	Synthetically augmented	As for multivariate plausible. Replicate multivariate distributions at small geographies. Some analytic value. Very high disclosure risk.
6.Replica	Synthetically augmented	As for multivariate detailed. Could be used in place of real data, extremely high disclosure risk, likely to be available only in secure research facility.

Note – While the author of this paper has paraphrased and shortened the original spectrum from the ONS working paper, the descriptions use only words from the ONS text.

15. At this point, it's worth noting the view (articulated within this classification, final row and column above) that "replica" synthetic data should be available only in a secure research facility.

How are synthetic data produced?

16. In very basic terms, most methods to produce synthetic data generate variables sequentially, the first variable follows its observed marginal distribution with each successive variable synthesised through a distribution that is conditioned upon all the previous variables.

17. An elementary example is shown below; it usefully outlines the method but also the inherent danger with regard to confidentiality. Consider the following 'true' data, describing a small population in terms of age (4 age bands) and gender.

Table 2 - True data

True data	Age band A	Age band B	Age band C	Age band D	All ages
Male	10	20	30	40	100
Female	80	60	40	20	200
Persons	90	80	70	60	300

18. Although trivial, note that the full underlying data call be expressed as 300 cases, where the first ten are (Male, Age A), the next 20 are (Male, Age B) and so forth. That is, the underlying data can be reconstructed exactly.

19. Going back to the ONS spectrum, structural synthetic data will have 300 records where each record is of the form (Gender, Age Band). If a random process was employed there would be approximately 37 or 38 in each cell of the table. The limitation is obvious – see below; even the marginal age and gender distributions are not reproduced for this simple example. [With just age x sex, no invalid records are possible, hence Table 3 is also 'valid'.]

Table 3 - Synthetic data Structural and Valid (level 2 as defined in ONS classification)

Synthetic	Age band A	Age band B	Age band C	Age band D	All ages
Male	37	38	37	38	150
Female	38	37	38	37	150
Persons	75	75	75	75	300

20. Now consider moving to plausible data, where the marginal distributions are maintained. Fitting the age distribution first, the original marginal age distribution is necessarily returned (the persons row). For the interior cells, the overall Male : Female marginal distribution (1:2) is applied to each age band giving the table below. The marginal totals are reproduced, but there are errors at the cell level because the age by gender relationship has not been considered. Considered multivariately, a plausible relationship is shown (that of independence) but the true relationship between age and sex is not reproduced.

Table 4 - Plausible synthetically augmented (level 3/4 as defined in ONS classification)

Plausible	Age band A	Age band B	Age band C	Age band D	All ages
Males	30	27	23	20	100
Females	60	53	47	40	200
Persons	90	80	70	60	300

21. Now consider creating further "synthetic augmented" data that takes account of relationships between the variables. Fitting the age distribution first, the original marginal age distribution is necessarily returned (the persons row). For the interior cells, the specific Male : Female ratio is applied to each age band separately giving the table below. It is immediately evident that the original data have been reconstructed. (Note the levels 5-6 in the ONS Spectrum differ in the level of geographical detail. This simple example operates only at a single geographic level, hence it goes straight to level 6 on the ONS Spectrum.)

Table 5 – Replica synthetic augmented (level 6 as defined in ONS classification)

Synthetic augmented	Age band A	Age band B	Age band C	Age band D	All ages
Male	10	20	30	40	100
Female	80	60	40	20	200
Persons	90	80	70	60	300

22. It is stressed that this is of course an elementary example. The exact reconstruction of the original data has come about because of the simplicity of the raw data, in particular the limitation of the data to a two-way table. Mathematically, this is similar to acknowledging that a straight line in two dimensions can be defined by two points.

23. However, the principle involved can be extrapolated. If the data are based on n classification variables, a fully saturated n-way table will fully describe the data - which this method will recover. This is why high-dimensional tables are generally not produced by Census Offices, except perhaps at high levels of geography (such as Northern Ireland in a NISRA context).

24.A few further comments –

These high dimension fully saturated tables are equivalent to data-cubes, for which it is the size of the smallest cells (by count) that is important from a disclosure perspective.

The full reproduction of the original data depends on the table being fully saturated.

The inclusion of non-discrete variables will protect against full reproduction.

25. It is exaggerating to make a point, but while synthetic augmented data are obviously 'not real' this does stop them being disclosive. The elementary example (above) demonstrates that – in terms of protection against disclosure - synthetic data that fall towards the bottom end of the ONS spectrum might be considered as similar to de-identified data. Thus, there are no geographic or name identifiers, but otherwise the data reflect real data. As Census Offices are well aware, simple de-identification provides little protection against disclosure.

[**26.**The dangers of simple de-identification are demonstrated in the well-known example of William Geld, Governor of Massachusetts. In the late 1990s, in the interest of the public good, the state made available a hospital record dataset, de-identified but including zipcode (US equivalent of postcode) and date-of-birth. An IT student matched the hospital dataset against publically available voter-registration data, and posted the Governor his personal medical record.[9]]

How good are synthetic data in practice?

27. While it's not the only approach, the Synthetic Data Estimation for UK Longitudinal Studies (SYLLS) team has developed routines for the production of synthetic data within R (called Synthpop[10]) using a methodology that seems to be widely used[11].

28. The SYLLS group has no access to the full Scotland Census, but have created a synthetic (synthetic augmented in ONS terminology) data set based on a publically available Polish survey containing 35 variables of various types across 5,000 individuals. The paper derives a model that predicts "likely intention to work abroad" using a set of predictor variables covering sex, age, educational attainment (3 variables) and income. The paper concludes that "The fact that the results from synthetic data can have a similar pattern to the results from the real data is encouraging for further developments of synthetic data tools." (page 23, very end of Section 4)

29. The SYLLS paper states (Section 5, Concluding remarks, opening sentence), "In this paper we presented the basic functionality of the R package synthpop for generating synthetic versions of microdata *containing confidential information* so that they *are safe* to be released to users for exploratory analysis." The italic emphasis is by this author; if the italicised phrases are removed, the paper justifies the quote. In the view of this author, the SYLLS paper demonstrates the ability of synthpop to produce synthetic data that can be the basis of serious modelling – but the paper has not examined the disclosure risk inherent in the synthetic microdata.

The disclosure risk associated with synthetic data

30. The elementary example above (paras 17 to 21 above) and the SYLLS paper show that - beginning with discrete categorical data in saturated or near-saturated high dimension tables – synthetic data can be generated that closely resemble the original data.

31. While the above SYLLS paper largely focusses on the statistical quality of the synthetic data, and doesn't really address the disclosure risk, it's only fair to note that SYLSS has commissioned work on the disclosure risk from Mark Elliot (University of Manchester). Elliot's paper[12] notes that (Section 1) "At the extreme, it is in principle possible to specify the data generation process for the synthetic data so that it would completely replicate the original data; a fully saturated model would do this". Elliot goes on to point out that, in a Census context, this would require a model with thousands of parameters and "nobody would want to use such a model for producing synthetic data".

32. As with the first SYLLS paper, Elliot does not have full access to the Census, and his disclosure research is based on an analysis of the 2011 Living Costs and Food Survey (LCF), previously published as Open Data. Elliot concludes that "the disclosure risk present in a synthetic version of the 2011 LCF is very small". He rightly is cautious about extrapolating this to other surveys "particularly if it is to be used in a strong decision making context for example to release synthetic samples as open data. However, the results here are compelling and suggest that open synthetic datasets ought to be technically possible."

33. This author accepts Elliot's analysis, but – in a Statistical Office context - would query his view that "nobody would want to use such a (fully saturated) model for producing synthetic data". This will be followed up in the concluding discussion.

The disclosure risk with small population groups

34. Consider a table on the whole population when all cells contain large numbers. Nothing further can be inferred about any single person, whether the table is based on real or synthetic people.

35. Conversely, consider a small population group. International migrants from ethnic minorities are mostly fairly recent arrivals to Northern Ireland, and consequently there are relatively few older ethnic-minority migrants living in Northern Ireland. Consider a detailed high-dimension table including age and ethnic group based on synthetic data, but where the synthetic data are generated using a 'multivariate detailed' or 'replica' method. The number of older ethnic-minority migrants will be small in the table, and probably similar in number to the (Census) 'truth'. Some of the counts may well be 'ones'. Further, the characteristics of such people in the synthetic data will be very similar to that of the real population. The fact that the synthetic data are not real does mean that, in a strictly literal sense, the synthetic data cannot 'belong' to a member of the population. But in an example like this, the risk of disclosure must be high, and a hypothetical argument put forward by a Statistics Office that 'the disclosure risk is low simply because data are synthetic' wouldn't suffice (in this author's view).

36. The main aim of synthetic data is to reflect the underlying relationships in the real data. In broad terms, the quality of the synthetic data (with respect to this aim) will increase as further variables are introduced into the process of producing the synthetic data. An inevitable consequence of this increased quality is that individual synthetic records will resemble real records to a greater degree, with an associated increase in the likelihood of (apparent) disclosure. (It is for this reason that in 2011 Census Office used targeted record swapping.) It is noted that Elliot (end of section 4) suggests that the perturbative nature of generating synthetic data means that "unusual records" (for example, those linked with small population groups) are affected to a proportionately greater extent, providing additional protection to such records.

Synthetic data – discussion and conclusion

37. As shown in the elementary example above (paragraphs 17-22) and the SYLSS paper, from a starting point of tabular data modelling has the capability of generating synthetic data that reflects the statistical properties of the original data. A consequence of this is that individual synthetic records will (virtually) replicate individual real records, leading to a high risk of apparent disclosure.

38. The ONS methodology paper states that, within the spectrum of synthetic data, 'multivariate detailed' synthetic data has a 'very high' disclosure risk while 'replica' data has an 'extremely high' disclosure risk and should accessed only in a secure setting.

39. The initial SYLLS paper[11] demonstrates the ability of synthpop to generate high quality synthetic data, and states that "In this paper we presented the basic functionality of the R package synthpop for generating synthetic versions of microdata *containing confidential information* so that they *are safe* to be released to users for exploratory analysis." (this author's italics).

40.A second SYLLS paper (Elliot[12]) focusses on the disclosure risk associated with synthetic data. It acknowledges that synthetic data based on a saturated (or near saturated) model essentially reproduces the original data with the attendant disclosure risk. But Elliot suggests that saturated models are not needed to produce sufficiently high quality synthetic data, and that such synthetic data (from non-saturated models) are safe.

41. This is the crux of the matter and worth further consideration.

42. Suppose a data set contains 20 variables each with 6 levels. A fully saturated table contains 6^{20} cells (approximately 4 x 10^{15} or 4 followed by 15 zeros), which is obviously beyond any realistic computing capacity. The simplest model would retain the 20 marginal distributions, requiring 20 x 6 = 120 control cells. The next step up would retain all two-factor interactions (variable A by variable B); there are 190 such interactions (20*19/2), each requiring 36 control cells, so 190 * 36 = 6,840 control cells (in addition to the original 120). Considering three-factor interactions (variable A by variable B by variable C); there are 1140 such interactions (20*19*18/(3*2)), each requiring 216 control cells, that is 246,240 additional cells. The number of control cells explodes exponentially with each additional level of interaction is included. The data requirements for these early stages are summarised below in table 6.

Table 6 - Control cells required as model becomes more complex

20 Variables, each at 6 levels

Model complexity	Additional Control cells	Total Control cells
Marginal distributions	120	120
Plus 2 factor interactions	6,840	6,960
Plus 3 factor interactions	246,240	253,200
Plus 4 factor interactions	6,279,210	6,532,320
And so forth, until		
Fully saturated		About 4 x 10^{15}

43.A researcher conducting an initial exploratory analysis might initially be content with synthetic data that only go as far as two-factor interactions, and be prepared to accept that the synthetic data cannot examine how factor C varies within any cell of (factor A by factor B). In order to meet this request, the Statistics Office needs synthetic data based on 6,960 control cells, which is probably feasible.

44. However, after conducting this exploratory analysis, the researcher decides that he needs to focus on just 4 variables, and wants to explore all possible factor interactions. This would require the Statistics Office to generate a complete new synthetic dataset containing all possible interactions involving the selected 4 variables in order to generate the researcher's required dataset.

45. Elliot's view is correct if each researcher request is conducted separately, with its own unique synthetic data set generated. However, there is a lot of work involved in generating a synthetic dataset (that is, a multivariate detailed synthetic dataset), and – more importantly – the Statistical Office will want to generate a single synthetic database and run all queries off that database. The synthetic data that the Statistics Office would need to generate to pre-empt all possible researcher requests would require going some distance down table 6 above. Doing so has two problems, firstly the sheer computing power required, and the disclosure risk increases rapidly increases.

46. In summary, if the desire is to create synthetic data that replicate the statistical properties of the original data, all variables of interest must be included in the modelling – by definition this leads to synthetic records that are similar to the real records with the associated disclosure risk. Removing variables (and / or interactions) from the modelling process (or using a random process for some variables) would reduce the disclosure risk but at the same time invalidate any modelling.

47. Putting this another way, models derived from the synthetic data must reflect relationships within the synthetic data - but the relationships within the synthetic data can reflect only those relationships used in generating the synthetic data. It's all a bit circular.

48. Finally, section 1.1 final paragraph of the initial SYLSS paper states –"These test data should resemble the actual data as closely as possible, but would never be used in any final analyses." This is a laudable objective, but this author would suggest that there is merit in the synthetic data having clear, well-advertised shortcomings especially if it reduces the disclosure risk. Especially if the researcher accepts that their final models can only be determined on the underlying data in the secure setting.

49. In conclusion, the disclosure risk with synthetic data that is fit for purpose (with respect to generating valid models) is correlated with the quality of the synthetic data. This author would not advise NISRA to put resource into attempting to create synthetic data that can both be released and sufficient for sophisticated modelling. This author would advise NISRA to invest further in the safe setting secure data laboratory, perhaps exploring further the potential for safe ways of providing secure remote access.

Addendums

Using a sample of raw data to produce a large-scale synthetic data set

50. It has been suggested that a small sample of real records (say 500 records) could be used as the base for the production of synthetic data, and the models could be used to produce synthetic data with an apparently large sample size, perhaps 100,000 or more. In a technical sense this is possible, but there are at least two shortcomings.

51. The first, and main issue, is the quality of the synthetic data. As outlined earlier, for the synthetic data to have value requires the use of a saturated or near-saturated model for the production of the synthetic data. Even with a modest number of variables, each with a limited number of levels, the number of cells in the high level table rises exponentially. For example 6 variables each at 4 levels has 4^6 or 256 cells. With only 500 people, the average cell count is just 2 leading to unstable modelling in the production of the synthetic data. The synthetic data could be produced, and would yield models, but the quality of these models is informed not by the 100,000 cases in the synthetic data set but the underlying quality of the 500 cases used in the creation of the synthetic data.

52. In practice, there are likely to be more than 6 variables at 4 levels within the data set, with the underlying model requiring probably thousands of cells. A random sample of (say) 500 cases would result in the number of model cells exceeding the number of observed cases, which would not enable meaningful modelling. Extrapolating the sample cases (essentially through replication) – to thousands of cases - would not improve the modelling capability of the data.

Facilitating researchers

53. This paper has argued that synthetic data, of sufficient quality that models could be based on it, should not be released outside safe settings, negating the work required to produce the synthetic data in the first place. Because of the disclosure risks in high-quality synthetic data, model building must be restricted to the safe setting.

54. How far could NISRA go in the ONS synthetic data classification to facilitate researchers, through the provision of synthetic data that would enable researchers to, for example, check computer code before coming into the safe setting? The "Plausible, synthetically augmented" (level 3 in the ONS spectrum) would have some value for researchers. This would be generated per Table 4 in the earlier simple example; thus -

55. For each variable in turn, the marginal (univariate) distribution would be derived. This need not a theoretical distribution, merely the percentage distribution across the values of the variable – for example for variable 1, X_1% of cases have value 1, X_2% of cases have value 2, and so forth. The synthetic data will mimic this for synthetic variable 1, with X_1% of cases have value 1, X_2% of cases have value 2, and so forth. The marginal (univariate) distribution of variable 2 is then derived in the same way. This distribution is applied, in turn within each value of variable 1. This is repeated for each variable in turn. In general, the overall marginal distribution of each variable is retained, and an assumption of independence between all variables is assumed.

56.Analysis of the resulting synthetic data would yield the correct marginal univariate distribution for each variable, giving some reassurance to the researcher. But, the correlation matrix will be the Identity matrix, implying zero correlations and no explanatory models will be possible.

57. It is recommended that validity checks are not performed. Thus the synthetic data may contain, for example, employed 5 year olds. This will further emphasise the synthetic nature of the data.

Robert Beatty

March 2020

References

[1] UK Statistics Authority - Code of Practice for Statistics (2018) https://www.statisticsauthority.gov.uk/code-of-practice/

[2] Data Protection Act 2018 http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[3] Census Act (Northern Ireland) 1969 https://www.nisra.gov.uk/statistics/2011-census/background/legislation

[4] 2021 Census Proposals for Northern Ireland https://www.nisra.gov.uk/statistics/planning/2021-census-proposals-document

[5] The ecological fallacy, sample definition - https://support.esri.com/en/other-resources/gis-dictionary/term/3dc42628-48b6-426d-a5dd-0d06c8d9c5f7

[6] The Modifiable Areal Unit Problem (MAUP) https://en.wikipedia.org/wiki/Modifiable_areal_unit_problem

[7] Microdata section of NISRA Census website https://www.nisra.gov.uk/statistics/2011-census/results/specialist-products

[8] ONS Methodology Working Paper 16 – Synthetic Data pilot Bates et al (2019) https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot

[9] Data anonymization, see for example - https://aircloak.com/history-of-data-anonymization/

[10] Synthpop package https://synthpop.org.uk/about-synthpop.html

[11] Synthpop: Bespoke Creation of Synthetic Data in R (Nowok B, Raab G and Dibben C, 2016) https://www.jstatsoft.org/article/view/v074i11

[12] Final Report on Disclosure Risk Associated with Synthetic Data Mark Elliot (SYLLS) 2014 http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2015-02%20-Report%20on%20disclosure%20risk%20analysis%20of%20synthpop%20synthetic%20versions%20of%20LCF_%20final.pdf