Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team

Mark Elliot

October 2014

1. Background

I have been asked by the SYLLS project, University of Edinburgh to provide an analysis of the disclosure risk posed by the release of synthetic data produced by the methodology developed as part of the SYLLS project and implemented in the *synthpop* package for R. The general approach here is to investigate what a would-be intruder might be able learn about an individual from a synthetic data set.

It is commonly assumed that synthetic data is by definition non-disclosive. However, this position is based on a slightly naïve understanding about what synthetic data is. 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. This of course would not be practicable for synthetic data nor is it the aim.[1]

However more realistically it is true that the one is aiming to reproduce as many the statistical properties of the original data as possible. And, by induction, the better one is able to do that the closer one must be getting to the original data and therefore the issue of whether the data are disclosive or not is an empirical question.

1.1 The data situation

It is possible that the synthpop package could be used to make some extracts from the LSs available as public use data files. This is not what is planned in the first instance since synthesised LS extracts will only be released to accredited researchers who have signed an undertaking not to make them available to anyone else. This "cautious approach first" is commendable. However, in assessing the risk here we are looking at the open data situation.

A slight complication in the project design was that in order to properly test the disclosiveness would require me to have access to the original data and that was not possible within the resource and time constraints. However, the primary idea was to test the disclosiveness of SYLLS as a system rather than a particular synthetic dataset and the timing of the project enabled us to take advantage of some recent work that I carried out for ONS.

This was a penetration test of the same Living Costs and Food survey (LCF) dataset that was used in this study. That study had established that if the LCF was released as open data then there it would be possible to identify people within it with a high degree of certainty. This then is a fairly interesting test of the non-disclosiveness of

[1] Methods such those implemented in synthpop could not produce synthetic data from a fully saturated model as there would be no residuals to sample from and a model with hundreds of thousands of parameters would be needed. This is in any case moot as nobody would actually want to use such a model for producing synthetic data. The point is that synthetic data sits on a disclosure risk continuum which runs from completely random data to the original raw data. Disclosure controlled data sits at a point on that continuum and so does synthetic data.

the synthetic SYLLS method. If synthetic versions of the LCF were non-disclosive when tested similarly then that would be evidence they could be regarded as safe to release as open.

The practical upshot was that a synthetic version of the 2010, 2011 and 2012 LCF survey datasets were generated and passed to me.

2. Methodology

Assessing disclosure risk of synthetic data cannot be done using orthodox disclosure risk measurement tools. All of these rely on the structure of the data and test for the risk of reidentification. Most pragmatically assume that the data is completely non-divergent (that it corresponds exactly to the world). With synthetic data that assumption is demonstrably false (actually it is for non-synthetic data too but that is another issue). However, the critical point here is that it is not meaningful to think about synthetic data in terms of reidentification.

Reidentification only is disclosive in so far that it leads to accurate attribution. It is the attribution of a piece of information to a particular individual that constitutes disclosure. Therefore if disclosure is meaningful in the context of synthetic data it is through direct and accurate attribution. In order to test for this I constructed a new method with a working title of "Empirical differential privacy".

The basic principle of differential privacy is whether to make more accurate inferences about individual X from a dataset that contains X than the same dataset without X. One can apply that principal in a slightly stronger way using the procedure shown in Box 1.

Box 1: By record empirical differential privacy procedure assuming a categorical target variable

1) Obtain two consecutive years of the same survey dataset.
2) Generate synthetic versions of those data sets.
3) Take a record r at random from the original dataset and using a predefined key (K).
    a. Match r back onto the original dataset (O)
    b. Match r against the synthetic version of O (S)
    c. Match r against the synthetic dataset for the other year (S')
4) Select a target variable T
5) Each of 3 a-c will produce a match set (a set of records which match r). The proportion of each match set with the correct value on T is the probability of an accurate attribution (PAA) value for that match set.
6) Repeat 4 and 5 several times with different targets
7) Repeat 3-6 with different records until the mean PAA values stabilise.

If the mean PAA values for S are indistinguishable from S′ then S is differentially private with respect of K. One can also compare the values for S with the values for a baseline estimate simply based on draw from the univariate distribution for T.

Now, in fact, one can generalise this procedure across the whole of O as shown in box 2.

Box 2. General empirical differential privacy procedure assuming a categorical target variable

1) For each record in O record their multivariate class membership for both K and K+T. The equivalence class for K+T divided by the equivalence class for K for a given record will be the PAA for that record.
2) Repeat 1 for each record in S. Impute the PAA values from S into O against the corresponding records (matched on K+T).
3) Repeat 2 for S′

A slight complicating factor is that not every combination of K or K+T within O will be represented in S or S′. The imputed PAA values for these will be recorded as 0. This seems reasonable as the probability of accurately attributing from a non-match is de facto zero.

Now, we can obtain summary statistics for the PAA's for O, S and S′, most particularly the means and standard errors. If the mean PAA for S is not significantly greater than the mean PAA for S′ then S is empirically differentially private with respect of K and T.

With a continuous variable as the target, the procedure – shown in box 3 - is slightly different.

Box 3. General empirical differential privacy procedure assuming a continuous target variable

1) For each record in O record their multivariate class membership for both K.
2) Calculate the mean M of T for each equivalent class.
3) Calculate the absolute residual between M and T.
4) Repeat 1 for each record in S. Impute the residual values from S into O.
5) Repeat 2 for S′

Comparison of the mean residual sizes rather than mean PAA is then the risk assessment. In these cases non-matches are replaced by the means of the original files.

2.1 A note about the scenario

The particular statistic that the above method generates (the PAA) has an unusual relationship with the issue of response knowledge. The critical point is that one could make a correct inference even if the match is incorrect – indeed even if the target individual population unit is not actually represented in the file. Therefore, there is not a direct mapping of the statistic onto the level of response knowledge. In future research I will explore this mapping of response knowledge and inference however this relationship is pre-theoretical at present.

The LCF data was selected for this study, because it would allow comparison with the results of the pen test later, we generically assumed that the linkage is carried out by an intruder who has access some information about a household that s/he knows is in the original dataset. Because of the aforesaid lack of mapping this is actually a weaker assumption than in the original pen test and actually as we will see this comparison is not that illuminating in any case.

2.2 Key selection

In disclosure control processes key selection is usually an important issue. In this case because we have previously carried out pen tests against these data we have good empirical data to guide us. Critical keys were the geographical indicators: government office region and output area classifier, tenure and dwelling type. Secondary keys were based on the age and sex structure of the household.

Multiple combinations of keys were experimented with. In fact, the key selection does not in this case affect the overall pattern of the results. Four keys are used in the main analysis reported here

Key 1: GOR, Output area classifier,

Key 2: GOR, Output area classifier, tenure.

Key 3: GOR, Output area classifier, tenure, dwelling type.

Key 4: GOR, Output area classifier, tenure, dwelling type, Internet in hh.

The important point is to examine the extent to which increments of the in the keys size impact on the ability to make inferences.

3. Analysis

The PAA values for four (hypothetical) categorical targets with all four keys are shown in Tables 1-4. The patterns here are replicated with other key and target combinations.

Table 1: Mean Probability of accurate attribution(PAA) of number of cars using two keys against the LCF and three synthetic files

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	0.45	0.60	0.73	0.78	0.11
2010 synth	0.35	0.32	0.23	0.20	-0.05
2011 synth	0.37	0.33	0.26	0.22	-0.05
2012 synth	0.33	0.31	0.23	0.20	-0.04
baseline	0.35	0.35	0.35	0.35	
2011 synth-baseline[2]	0.02	-0.02	-0.09	-0.13	
DP residual[3]	0.03	0.02	0.03	0.02	

Table 2: Mean Probability of accurate attribution(PAA) of age of the oldest member of the household using two keys against the LCF and three synthetic files

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	0.12	0.34	0.53	0.62	0.17
2010 synth	0.02	0.02	0.02	0.01	0.00
2011 synth	0.02	0.02	0.02	0.02	0.00
2012 synth	0.02	0.02	0.02	0.01	0.00
baseline	0.02	0.02	0.02	0.02	
2011 synth-	0.00	0.00	0.00	0.00	
DP residual	0.00	0.00	0.00	0.01	

Table 3: Mean Probability of accurate attribution(PAA) of number of workers in the household using two keys against the LCF and three synthetic files

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	0.38	0.60	0.73	0.78	0.13
2010 synth	0.29	0.30	0.22	0.18	-0.04
2011 synth	0.29	0.31	0.24	0.20	-0.03
2012 synth	0.28	0.30	0.21	0.18	-0.03
baseline	0.30	0.30	0.30	0.30	
2011 synth-	-0.01	0.01	-0.06	-0.10	
DP residual	0.00	0.01	0.03	0.02	

[2] This is the difference between the synth 2011 value and the baseline which is generated by drawing randomly from the univariate frequency distribution for the target.

[3] This is the mean difference between the value for the 2011 synth file and the the other two synth files.

Table 4: Mean Probability of accurate attribution(PAA) of economic position of household reference person using two keys against the LCF and three synthetic files					
File	**Key 1**	**Key 2**	**Key 3**	**Key 4**	**mean cumulative key impact**
2011 original	0.35	0.59	0.72	0.78	0.14
2010 synth	0.31	0.31	0.23	0.20	-0.04
2011 synth	0.31	0.34	0.26	0.23	-0.03
2012 synth	0.31	0.31	0.24	0.21	-0.03
baseline	0.26	0.26	0.26	0.26	
2011 synth-	0.05	0.08	0.00	-0.03	
DP residual	0.00	0.03	0.03	0.03	

There are four main findings:

1) In all cases matches against the synthetic files produce much lower PAAs than matches against the original file.
2) There is a small difference in the PAAs for matches against the 2011 synthetic file compared to the other two synthetic files. This represents an uplift of 5- 10% across different keys.
3) The effect of adding variables onto the key (indicated by the final column of the tables) is to increase the PAA for the original data and decrease it for the synthetic data.
4) The 2011 synthetic data is no better than the baseline of drawing randomly from the univariate distribution.

One of the assumptions used for the analysis above is non-matches are in effect a PAA of zero. However, an intruder would very likely deal with a non-match on the primary key by aborting that particular attempt. So a better approach might be to treat a non-match on the key as undefined. The relevant statistic is the PAA conditional on a match. Tables 5 through 8 give these data.

Table 5: Mean Probability of accurate attribution (PAA) of number of cars using two keys against the LCF and three synthetic files given that a match has occurred.					
File	**Key 1**	**Key 2**	**Key 3**	**Key 4**	**mean cumulative key impact**
2011 original	0.45	0.60	0.73	0.78	0.11
2010 synth	0.42	0.46	0.45	0.49	0.02
2011 synth	0.39	0.43	0.45	0.46	0.02
2012 synth	0.35	0.40	0.40	0.42	0.02
baseline	0.35	0.35	0.35	0.35	
2011 synth-	0.03	0.07	0.09	0.10	
DP residual	0.01	0.00	0.02	0.00	

Table 6: Mean Probability of accurate attribution (PAA) of age of the oldest member of the household using two keys against the LCF and three synthetic files given that a match has occurred.

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	0.45	0.60	0.73	0.78	0.11
2010 synth	0.02	0.03	0.04	0.02	0.00
2011 synth	0.02	0.03	0.03	0.04	0.01
2012 synth	0.02	0.03	0.03	0.02	0.00
baseline	0.02	0.02	0.02	0.02	
2011 synth-	0.00	0.01	0.01	0.02	
DP residual	0.00	0.00	0.00	0.02	

Table 7: Mean Probability of accurate attribution (PAA) of number of workers in the household using two keys against the LCF and three synthetic files given that a match has occurred.

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	0.38	0.60	0.73	0.78	0.13
2010 synth	0.35	0.43	0.43	0.44	0.03
2011 synth	0.30	0.40	0.42	0.42	0.04
2012 synth	0.30	0.38	0.37	0.38	0.03
baseline	0.30	0.30	0.30	0.30	
synth2011-baseline	0.00	0.10	0.11	0.11	
DP residual	-0.02	-0.01	0.02	0.01	

Table 8: Mean Probability of accurate attribution (PAA) of economic position of household reference person using two keys against the LCF and three synthetic files given that a match has occurred.

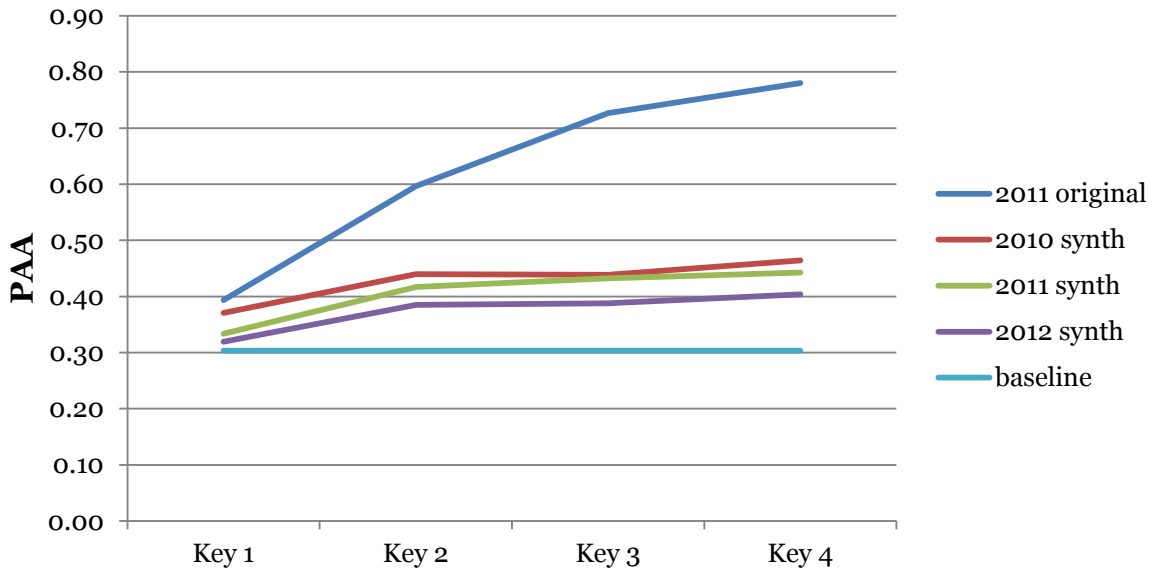
File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	0.35	0.59	0.72	0.78	0.14
2010 synth	0.37	0.45	0.45	0.49	0.04
2011 synth	0.33	0.44	0.45	0.48	0.05
2012 synth	0.33	0.40	0.42	0.44	0.04
baseline	0.26	0.26	0.26	0.26	
2011 synth-	0.06	0.17	0.18	0.21	
DP residual	-0.02	0.02	0.02	0.01	

There are four main findings:

1) With one small exception in Table 8, matches against the synthetic files produce lower PAAs than matches against the original file. Notably the exception was actually on the 2010 file rather than the 2011 one.
2) The differences in the PAAs for matches against the 2011 synthetic file compared to the other two synthetic files are much smaller than in Tables 1-4 and sometimes in the opposite direction. The mean nett uplift is only about 1% across different keys and targets.
3) The effect of adding variables onto the key (indicated by the final column of the tables) is to increase the PAA for the original data and to slightly increase it for the synthetic data.
4) All of the synthetic data is better than the baseline of drawing randomly from the univariate distribution.

It is worth reflecting on these results in more detail as superficially they present a somewhat more concerning picture than the original set. One point is that for small keys and course targets the ability to make accurate inferences is pretty much the same for the synthetic data and the real data. The exception here is Table 6 where the target is age, where the PAAs for the synthetic data files don't get off the floor. Arguably, age should have been treated as a continuous variable. If we ignore that table for the current discussion, then figure 1 shows the mean effects of key size across the four keys.

Figure 1: Mean PAA scores over 3 categorical targets across four keys of increasing size.



This does illuminate the results well. The original 2011 data shows the expected increase in PAA as key size increases. The synthetic data shows modest increases in

PAAs as key size increases in key size and by the key4 all are noticeably better than the baseline. However, the particular interesting thing to note is that it is the 2010 file that is consistently the highest performer rather than the 2011 file.

An additional point to consider here is the hit rate. This is shown table 9. Unsurprisingly, the % of records on the original for which there is at least one corresponding record in the synthetic data file decreases markedly as the size of the key increases. By the time you get to key 4 the hits rates have fallen below 50%. This provides some further context for the PAAs above; the hit rate is dropping much faster than the PAA is climbing. The rates for the 2011 and 2012 synthetic files are virtually indistinguishable. In a reversal of the result in Figure 1 the 2010 synthetic file is the worst performer.

Table 9: Hit rate for primary key matches from the original file onto the synthetic file.

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	100%	100%	100%	100%	0%
2010 synth	84%	69%	51%	41%	-14%
2011 synth	95%	77%	58%	48%	-16%
2012 synth	94%	78%	57%	48%	-15%
DP residual	6%	3%	3%	4%	

Table 10 shows the results for the one continuous target outcome using the alternative method shown in box 3 in section 1. There are four main findings:

1) The matches against the synthetic files produce much higher residuals than matches against the original files.
2) There are small differences in the residuals for matches against the 2011 synthetic file compared to the other two synthetic files. This represents a reduction of between 2-6% across different keys.
3) The effect of adding variables onto the key is to decrease the residual markedly for the original data and on average to increase it for the synthetic data.
4) The 2011 synthetic data is only slightly better than the baseline of assuming the mean income.

Table 10: Residual sizes for estimated weekly income using two keys against the LCF and three synthetic files

File	Key 1	Key 2	Key 3	Key 4	mean cumulative key impact
2011 original	246.72	246.72	174.84	147.00	-33.24
2010 synth	401.29	378.60	396.71	397.72	-1.19
2011 synth	369.69	368.35	384.99	380.96	3.76
2012 synth	376.36	374.08	389.79	387.61	3.75
baseline	385.36	385.36	385.36	385.36	
2011 synth-	-15.67	-17.01	-.37	-4.40	

DP residual	19.13	7.99	8.26	11.71	

These findings are consistent with those for the PAA shown in Tables 1-9.

As a final empirical check I also drew on the pen test data. Although this was originally a focus of the basis of this consultancy, the methodological development that are described above make this a less critical test than was imagined at the outset. The issue being that the pen test was essentially focused on identification not attribution.

Nevertheless, by focusing on the successful unique matches from the pen test it is possible to assess whether *a posteriori* risky records are any more problematic than the file means. There were thirteen such records in the pen test. The keys used varied between the target individuals as they depended on the specific data collected for that individual; here we use the same keys as were used in the pen test. As these keys sometimes included the categorical targets used in tables 1-9 we focus on the income estimates (i.e. equivalent to table 10). This is shown below.

Table 11: Residual sizes for estimated weekly income using two keys against the LCF and three synthetic files

File	hit rate	residuals	mean of table 10	Difference
2011 original	100%	0.00		
2010 synth	46%	342.35	393.58	51.23
2011 synth	46%	426.04	376.00	-50.04
2012 synth	62%	439.09	381.96	-57.13
baseline		385.36	385.36	
synth2011- baseline				
DP residual				

Comparing the residuals sizes with the means of Table 10, we can see that these are a little worse than the file means. Although one should not read too much into the figures as they are based on a small number of cases. So the errors on the figures will be quite wide. However, there is certainly no evidence that these *a posteriori* risky records present any greater risk than the file means for the synthetic data.

4. General discussion

The above results indicate that for the synthetic data produced for this trial are close to differentially private.

The uplift of the 2011 synthetic file of the other two synthetic files probably represents the likely fact that the 2011 synthetic file is structurally more similar to the original file than the other two rather than any actual increase in disclosure risk because of the presence of the case in the 2011 file. Once we adopt the - perhaps

more realistic - approach of treating non-matches as undefined then the effect is in any case very small and indeed in many instances the synthetic version of the 2010 files produce higher PAAs than that for the 2011 file.

The divergent key addition effect that appears to present when we code non-matches on the primary key as yielding a PAA of zero is interesting and warrants further investigation. However, the effect does not hold if we treat non matches on the primary key as undefined. In some ways, this slightly disappointing as the effect would be extremely protective. However, it would be surprising if the increasing the key size did not increase the capacity to make inferences as this would suggest that the synthetic data was of little analytical value. The cumulative impact of key size is still small relative to the original file.

The comparison of the baseline data with the 2011 synthetic dataset is particularly telling. The synthetic data is only marginally more accurate than simply assuming that each person earns the mean income.

In the interim report there were two other pieces of work that looked like they may be useful additions.

Firstly, it was thought that it might be useful in order to clarify whether the uplift for the 2011 synthetic files was due to the shared information about the record in question analyses could be run on the original files for 2010 and 2012. The logic being if the same differences emerge then we can say that the 2011 synthetic file is differentially private. However, since the uplift virtually disappeared for the analyses reported in tables 4-8 that appears unnecessary now. In any case it was simply an exercise in crossing as it is unlikely that the uplift - such as it is – is being caused compositionally at the record level it is much more likely that it is a side effect of the differences between the three original samples. Some more detailed experiments with synthetic versions of the same data set with single records removed or not would be a much more consistent way of assessing the basic proposition and that is one piece of possible future research work.

Another possible piece of work identified in the interim report was *"for the PAA calculations it may be better to assume a draw from a random distribution rather than 0 for non-match."* On reflection this does not produced a meaningful statistic. An intruder when faced with a non-match is likely to abort or try a different (probably smaller) key. We have looked at the impact of the abort strategy in Tables 4 to 8. The different key strategy would require some complex modelling which is beyond the scope of this consultancy. However, given the distribution of information of the PAAs and hits across Tables 4-9, it is not difficult to estimate what the effect of such a strategy would be: a small decrease in the PAAs but an increase in the number of hits.

One final point before concluding, it is an interesting feature of the results that there is a general tendency across the data for hit rate and PAA to be inversely related.

This is highly suggestive that it is more unusual records that are failing to hit. This makes sense if you think about the synthetic data production as a perturbative mechanism. If so then this is again a significant protective factor as the unusual records are the risky ones in disclosure risk terms. Further research – outside the scope of this consultancy - would be required to confirm this but these results are certainly indicative.

Conclusions

Overall, the pattern of results described in this report indicates that the disclosure risk present in synthetic version of the 2011 LCF dataset is very small. Note that this should not be taken as a general conclusion about all synthetic data nor even all synthetic data produced by the SYLLS method. Further work would be necessary to generalise this conclusion, 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.