Proceedings of Statistics Canada Symposium 2016
Growth in Statistical Information: Challenges and Benefits

Sampling Procedures for Assessing Accuracy of Record Linkage

Paul A. Smith, Shelley Gammon, Sarah Cummins, Christos Chatzoglou, and Dick Heasman[1]

Abstract

The use of administrative datasets as a data source in official statistics has become much more common as there is a drive for more outputs to be produced more efficiently. Many outputs rely on linkage between two or more datasets, and this is often undertaken in a number of phases with different methods and rules. In these situations we would like to be able to assess the quality of the linkage, and this involves some re-assessment of both links and non-links. In this paper we discuss sampling approaches to obtain estimates of false negatives and false positives with reasonable control of both accuracy of estimates and cost. Approaches to stratification of links (non-links) to sample are evaluated using information from the 2011 England and Wales population census.

Key Words: Inverse Sampling, Controlled Sampling, Linkage Error.

1. Introduction

Record linkage (matching) is becoming an increasingly important part of official statistics with the increased pressure to use administrative datasets as a component of their production. It is therefore important to be able to assess the quality of matching processes. There is wide agreement (eg Ferrante & Boyd 2012, Randall *et al*. 2013, Vatsalan *et al*. 2014) that the key measures are the precision of the matching process, expressed as $P = \frac{TP}{L} = \frac{TP}{TP + FP}$ where TP is the number of correct links ('true positives'), and FP is the number of false positives among L links made in the matching procedure; and recall, expressed as $R = \frac{TP}{TP + FN}$ where FN is the number of missed links ('false negatives'). Where an overall measure is needed, the f-measure $f = 2\frac{PR}{P + R}$ (the harmonic mean of precision and recall) quantifies the trade-off between the two measures.

In order to estimate these measures in a reasonable way for datasets that contain potentially many matches, we need estimates for TP, FP and FN. If we aim for 'gold standard' evaluations from detailed investigation (clerical resolution) of possible links, a sampling procedure is needed to define suitable numbers of links and non-links to investigate, in order to make estimates of precision and recall satisfactorily accurate. In this paper we set out how a stratified sampling procedure can be used for this process, and the inputs and data sources that are needed to tune the sampling.

A full linkage process typically involves multiple stages (here labelled by h as we will consider stratifications based on them) – for example, first an exact matching process, then various rule-based procedures, and then probabilistic matching methods, which may be repeated with a range of different blocking strategies (Herzog *et al*., section 12.1). In this paper we use the matching of the 2011 England and Wales Census and Census Coverage Survey (CCS) as an illustrative example; in this case the clerical match outcome from the Census processing is treated as the "true" match status, so it is known. We then evaluate an automated matching process, first matching together using seven matchkeys (a deterministic phase) and a logistic regression (a probabilistic phase). We assume that the precision, p,

[1]Paul A. Smith, S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ, UK (p.a.smith@soton.ac.uk); Shelley Gammon, Sarah Cummins, Christos Chatzoglou, and Dick Heasman, Office for National Statistics, Segensworth Road, Fareham, PO15 5RR, UK

varies according to the matching method (Winkler 2004 indicates that the variation in matches from different blocking passes may be substantial, for example).

2. Stratification

Stratified sampling is most effective when the characteristics of interest are homogeneous within strata, and heterogeneous between strata (Cochran 1977, section 5.7). Therefore, in order to choose an efficient stratification, we will look for variables that have good explanatory power for the characteristic to be investigated, and use those preferentially to define the strata.

We use the presumption that a high-quality matching strategy contains a number of stages, and that values of precision vary by the stage at which the match was declared (which therefore should be recorded). The population for false negatives is all the pairs which were not identified as a match in any stage. Therefore they are not distinguished by the matching stage, which is not used in estimating recall. We would like both to take advantage of the differences between stages when designing the sampling for precision and to have estimates by stage to provide feedback on the usefulness of the stages. We therefore expect that stratification by the method or blocking procedure that determined the match is useful. It is possible that characteristics of the record pairs (other than the linkage method) also affect the probability of a true match, and this sort of information can also be used in stratifying for estimating recall.

Initially we have no information on which to assess which variables are important for stratification. However, once a quality evaluation has taken place, we have a sample of record pairs with 'true' outcomes which can be analysed to provide guidance for further studies. From an evaluation we therefore need data containing the record pair, including the variables of interest for stratification (eg location, demographic characteristics, paradata variables (for example time between extraction/collection dates of the information in the records to be matched)). We also need the outcome of the quality evaluation – whether the match was assessed to be correct or incorrect. We use this data to develop a logistic regression model to identify the best predictors of the true/false match indicator.

To estimate FN (which we do by estimating $Q = FN/(FN + TN)$ and multiplying by the number of non-matches) we look amongst non-matched pairs of records and examine in the quality assessment whether they should have been matched or not. Although all case pairs passed all stages without being linked, the probability of a correct match from the final, probabilistic matching stage is likely to be an important predictor. The number of non-matched pairs is, in general, very large, and a great many of these pairs are true non-matches (TN). In order to make an assessment practical, we will generally assume that a large number of these pairs (eg male to female matches) have such low probability of being a false non-match (ie a true match) that they can be ignored. Therefore we will have a stratum which is never considered for sampling (a cut-off sample, Haziza et al. 2010, Smith 2013 section 5.3.4). For non-matched pairs excluding this cut-off stratum, we can construct a logistic regression model for the 'true' outcome in a similar way.

2.1 Predictors of false positive errors in using the Census-CCS linkage

To investigate the factors associated with false positive errors, a fully automated approach was first used to redo the matching for the 2011 Census and CCS using deterministic and probabilistic linkage (blocked by postcode). The resulting linked record pairs were then analysed using the original 2011 Census to CCS linkage as a gold standard matched dataset to indicate the 'true' match status.

A total of 619,458 links were made, with 0.22% being false positives (incorrect links). Building a model from the full data would swamp the false links, so 500 true positive record pairs and 500 false positive record pairs were sampled (by srs) as training data for the model. Samples were taken within linkage method strata, proportional to the size of the strata, to ensure that all linkage methods were represented

A logistic regression model for true match status (false positive = 0, true positive = 1) was built from the available explanatory variables:

- Census hard to count index (1-5) (Hopper, 2011)
- Sex (1, 2, or missing)

- Age group (0-17, 18-24, 25-39, 40-64, 65+, missing)
- Whether in London (1) or not London (0)
- Ethnicity (White, Asian/Asian British, Black/Black British, mixed, other, missing)
- Match pass - exact (1), rule-based (2- 7), probabilistic (8)

The variables sex, age group and ethnicity were checked for conflicts between the matched pair of records, excluding missing. Since the ethnicity variable was coded differently in Census and CCS, the codes were matched up as closely as possible. There were 3.84% record pair conflicts for ethnicity, 0.29% conflicts for sex and 0.70% conflicts for age. This was considered unlikely to affect the model.

A stepwise modelling procedure was used on eight different samples of 1000 records. Match pass was highly significant ($p<0.0001$) in all samples and the hard to count index was highly significant ($p<0.0001$) in six out of eight samples, indicating that these two variables are strong predictors of false positives. Not all categories in match pass and hard to count index were significant and so analysis of false positive rates and expert judgement were used to simplify the categories: Hard to count was recoded to (1,2) and (3,4,5) and match pass to (1), (2-7) and (8), which corresponds to exact, deterministic and probabilistic matching.

Modelling proceeded using the Bayes Information Criterion (BIC; or Schwarz criterion) to choose between models, which provides strong protection against overfitting. The simplified (recoded) model had a far lower BIC than the original model, indicating an improved model fit. For the final model, the area under the ROC curve was 0.8634, indicating that the model is a good predictor of false positives.

2.2 Predictors of false negative errors using the Census-CCS linkage

The non-linked record pairs from the automated analysis and gold standard outcomes from the original 2011 Census to CCS linkage (see 2.1) were used as inputs to a model for false negatives.

Just over 4 million record pairs were rejected at the probabilistic matching stage, 16,000 (0.40%) of which were true matches (false negatives). Building a model from the full data would swamp the false negatives, so 500 true negative record pairs and 500 false negative record pairs were sampled (by srs) as training data for the model. Samples were taken within groups of probabilistic matching scores, proportional to the size of the strata, to ensure that a range of probabilities were represented.

A logistic regression model for true non-match status (true negative = 0, false negative = 1) was built from the available explanatory variables:

- Census hard to count index (1-5)
- Match probability from probabilistic stage of matching (0-0.1, 0.1-0.2, … , 0.9-1)
- Whether in London (1) or not London (0)
- Sex in Census (1, 2 or 0 for missing)
- Age group in Census (0-17, 18-24, 25-39, 40-64, 65+, missing)
- Ethnicity in Census (White, Asian/Asian British, Black/Black British, mixed, other, missing)

Census variables for age group, sex and ethnicity were considered in the model; however these are highly likely to conflict with CCS values, especially for true negatives. Conflicts were high for sex (47%), age group (42%) and ethnicity (36%). This should be noted before interpreting results.

A stepwise modelling procedure was used on eight different samples of 1000 records. Probability group ($p<0.0001$) and age group ($p<0.05$) were significant in all samples and sex ($p<0.05$) was significant in seven out of eight samples, indicating that these three variables are strong predictors of false negatives. Not all categories for these variables were significant and so analysis of false negative rates as well as expert judgement was used to simplify the categories. The BIC (Schwarz criterion) was again used to assess model fit, to protect against overfitting.

The false negative rate for probability group was largely affected by the placement of the threshold score in the automated matching phase (in this case at 0.5): record pairs just under the threshold had a much higher false negative rate than those above (except for those with a very high score). The categories 0-0.1 (1), 0.1-0.3 (2), 0.3-0.5 (3), 0.5-0.7 (4), 0.7-0.9 (5), 0.9-1 (6) produced a good model fit. Probability score groupings should take account of the probabilistic method and match designation threshold used. For age group, the category for missing had the

highest false negative rate as well as the oldest age group (65+), so the categories were grouped as follows: Missing (0), 0-64 (1), 65+ (2). Similarly, for the sex variable, missing also had the highest false negative rate whilst males and females had similar rates, so missing was retained as a category and male and female were grouped. Both grouped variables improved the model fit. The final model with the further grouped probability score, age group and sex had an area under the ROC curve of 0.8785, indicating that these factors are good predictors of false negatives.

3. Sample size determination and allocation

Once we have determined an appropriate stratification, from an analysis of, or by analogy with, past data, or by assumption, we then need to specify the target quality for the estimates of precision and recall, and use them to determine the required overall sample sizes and how they should be allocated between the strata. In the exposition below we focus on precision, the simplest case.

The estimates of precision and recall are estimates of proportions, *p* and *r* respectively, and in many instances these proportions are expected to be quite close to 1. There is therefore a risk that confidence intervals will have their upper bound above 1. Also when *p* becomes small, a variance constraint is insufficient to allow tests for differences between meaningful values of *p*. To control these risks, achieving a target coefficient of variation is much better than achieving a target variance, particularly when there is no prior information to give an initial estimate for *p*.

3.1 Neyman allocation

Cochran (1977, section 5.12) gives an expression for the sample size in stratified sampling of proportions under presumed optimal allocation using a target variance. It requires initial estimates of the stratum proportions $\tilde{p}_h$. The variance can be replaced by $c^2\tilde{p}^2$ where *c* is the target cv and $\tilde{p}$ is an initial estimate of the overall precision, and this yields an expression for the required sample size for fixed cv,

$$n_0 = \frac{\left(\sum_h W_h \sqrt{\tilde{p}_h(1-\tilde{p}_h)}\right)^2}{c^2\tilde{p}^2}, \qquad n = \frac{n_0}{1+\frac{1}{Nc^2\tilde{p}^2}\sum_h W_h\tilde{p}_h(1-\tilde{p}_h)}$$

where n_0 provides a first approximation if the finite population correction is negligible and *n* takes the fpc into account. This procedure is, however, dependent on the initial estimates $\tilde{p}_h$, which should be realistic. In order to have better control over the cv, however, we would like a method which is not so dependent on the assumed $\tilde{p}_h$.

3.2 Inverse sampling

Haldane (1945) introduced a technique which has become known as inverse sampling, which approximately controls the coefficient of variation using a sequential design which continues until *m* events have occurred – that is, for estimating precision, until *FP* = *m*. In this case the final sample size is a random variable, which is more tricky for resource planning purposes, but allows the accumulation of information in a sequential approach.

We can deduce the required *m* for fixed cv constraints in a single stratum using Haldane's result that $\hat{v}(\hat{p}) \approx \frac{\hat{p}^2(1-\hat{p})}{m-2}$, with the approximation good for small values of $\hat{p}$, the estimated probability of a FP. Solving for *m* gives $m = \frac{(1-\hat{p})}{c^2} + 2$, where *c* is the target cv, and again with the approximation holding for small values of $\hat{p}$. Therefore *m* is approximately constant and independent of $\hat{p}$ over ranges that are important for assessing matching accuracy. Cochran (1977, p77) uses a different approximation but its solution is very close to the approximation derived from Haldane's variance estimator.

It is therefore straightforward to use inverse sampling in a large population, sorted randomly, with varying probabilities of observing a "success" to obtain an unbiased estimate for the average probability in the population. In our case we would like additionally to have information on the probabilities within each stratum (as quality information on the linkage strategy), and might hope to achieve a result with similar accuracy using a smaller sample size if we were able to use inverse sampling within strata. Approximately constant m does not however translate to constant sample size, and extremely large expected sample sizes arise for small values of $\hat{p}$. But for inverse sampling it is not important to have accurate (or even any) estimates of $\hat{p}$ to determine the sampling – it is sufficient to proceed to accumulate 'successes' towards the target total r.

3.3 Stratified Inverse Sampling

It is possible to write down an expression for the cv of a stratified inverse sample using Finney's (1949) unbiased estimator for the variance of an inverse sample. No analytical solution to minimise $\sum_h n_h$ to achieve a fixed cv in this setup has yet been forthcoming. Instead we take a heuristic numerical approach. The basic algorithm (with some comments on its implementation) is:

Step	Comments
1. Allocate a minimum value for m_h in each stratum.	Minimum can be set to improve accuracy for small cost when successes are relatively common as in the generalised sampling plans of Kim & Nachlas (1984).
2. Evaluate cv using the sample sizes from step 1.	cv estimated on the fly using the results from the first cases.
3. If cv target not met, add 1 to the target m_h for each stratum in turn, leaving the others unchanged, and recalculate the expected cv (cv achieved if next 'success' occurs after $1/\hat{p}_h$ trials). Retain the sample with the extra unit with the lowest overall cv.	Need *expected* cv as this is all that is available sequentially. The simulation may get 'stuck' where one variance dominates. The frequency of this can be drastically reduced (but not eliminated) by limiting the number of consecutive increases to m_h in the same stratum (10 is used as a maximum in many simulations here); the stratum with the second best effect on the overall cv is chosen instead.
4. Repeat step 3, until cv target achieved or all strata reach maximum.	Sample size can be capped above to control costs with an associated variance penalty, as in the generalised sampling plans of Kim & Nachlas (1984).

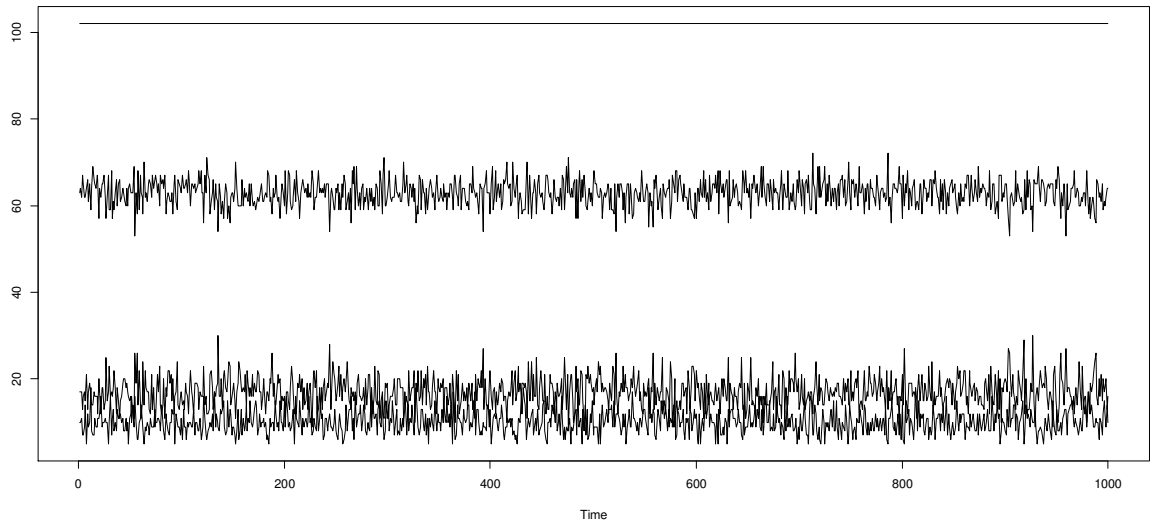
Table 3.3-1
Three-stratum small example. $\tilde{p}_h$ assumed known.

stratum	N (millions)	P
1	5	0.01
2	3	0.03
3	2	0.20

Consider the small example in Table 3.3-1. (A value of $p = 0.2$ in the final stratum is rather large for the assumption of 'p small' in inverse sampling.) Applying the above algorithm 1000 times on series generated from Table 3.3-1 gives outcomes as shown in Fig. 3.3-1. The top line is the constant $m = 102$ required for a 0.1 cv in inverse

sampling. In some sense we are more interested in the overall sample size, and summing m_h in the three-stratum approach gives $m \approx 90$. So at least in m there is a saving from using stratification. However, considering how these m translate into realised sample sizes n, there is essentially no difference – for single sampling $n = 1884$ on average, and for 3 strata $n = 1905$.

Figure 3.3-1
Numbers of successes required for cv = 0.1 treating whole population at once (top line), and using stratified inverse sampling algorithm (lower three lines, one per stratum).



4. Application to 9-stratum artificial example

Extending to a more realistic (but still artificial) 9-stratum example (Table 4-1) gives the outcomes in Fig.4-1. The Neyman allocation, done with the correct probabilities and therefore the best that could possibly be achieved (but unlikely to be approached in practice) looks very similar in quality to inverse sampling. This is corroborated by the boxplots of the cv's in Fig. 4.1. Although the range of cv's under Neyman allocation is wider, this is partly artefactual, because the stopping criterion for the inverse sampling is when the required cv is achieved, which reduces the variation in the cv's. It is also notable that the stratified inverse sampling seems to produce an estimate of the overall probability with a small downward bias. This result seems counterintuitive, as there is an unbiased estimator in each stratum and the weights are known, but it is repeated across the simulations.

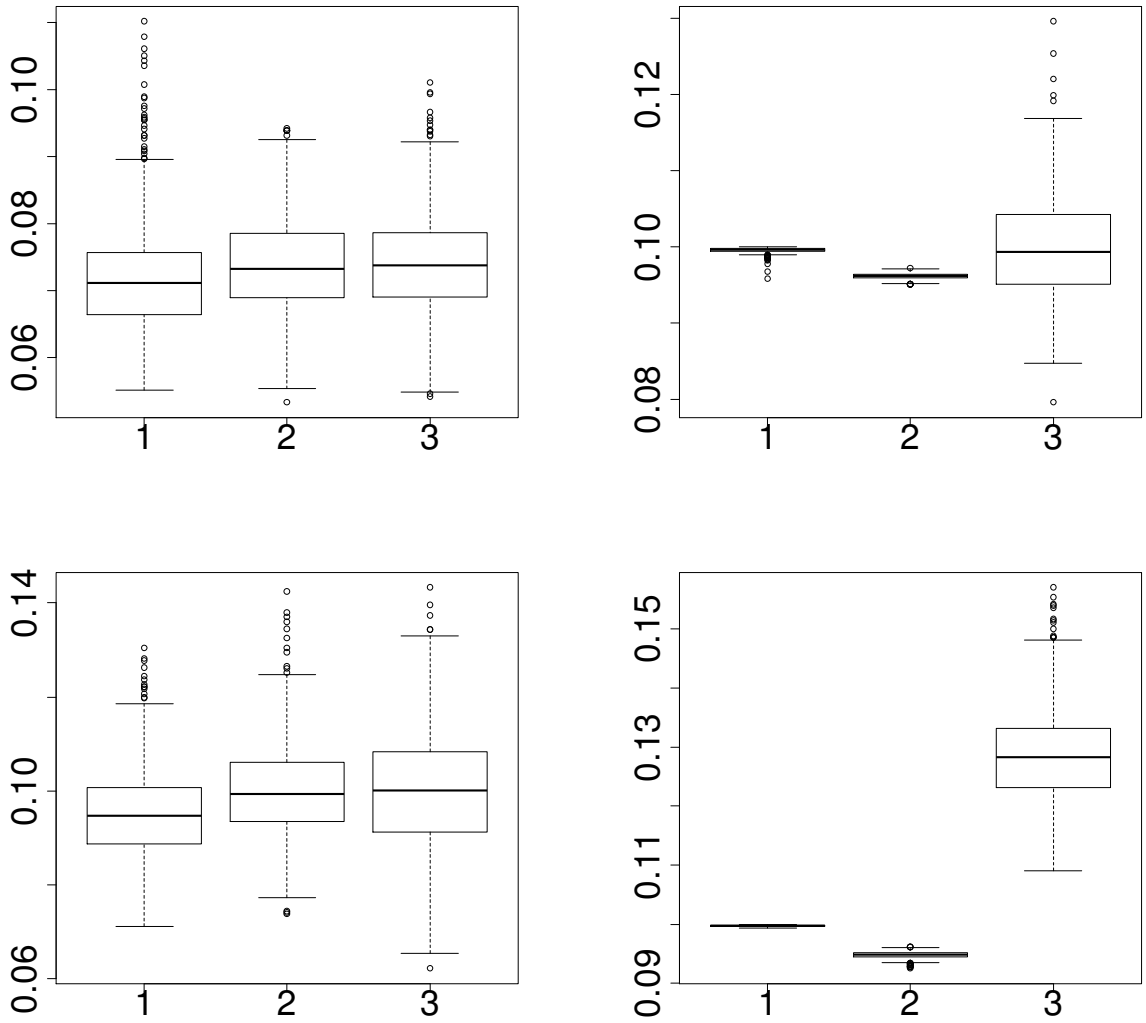
Table 4-1
Nine-stratum example.

stratum	*N* (of pairs, million)	Pr(*FP*) used in Neyman allocation	true Pr(*FP*)	
			set A	set B
1	20	0.005	0.005	0.1
2	20	0.01	0.01	0.1
3	16	0.05	0.05	0.1
4	10	0.06	0.06	0.1
5	10	0.07	0.07	0.1
6	90	0.1	0.1	0.1
7	7	0.2	0.2	0.1
8	5	0.3	0.3	0.1
9	3	0.4	0.4	0.1

Comparing the overall sample sizes, however, the sample sizes derived from Neyman allocation to meet a target cv are generally smaller than the inverse sampling total sample size for the same cv. In this case, where the cost of clerical resolution of links is quite high, a general strategy of inverse sampling could lead to high costs.

Figure 4-1

Estimates of $\hat{p}$ (left) and estimated cv's (right) from 1 – stratified inverse sampling, 2 – inverse sampling on the whole sample (unstratified), 3 – stratified sampling with Neyman allocation. The upper plots use Neyman allocation derived from the correct probabilities (set A in Table 4-1), the lower plots use the same allocation, but the true probabilities are constant at 0.1 (set B).



5. Discussion

Unexpectedly, stratified inverse sampling appears not to offer much advantage over ordinary inverse sampling (but neither to have a penalty). Its main advantage is that allows inverse sampling to be employed to provide indicators on the quality of matches coming from each stage of the matching process. The sample size requirement for inverse sampling is in general larger than that for a Neyman allocation for the same target cv. The cv's from the Neyman allocation are generally more variable, as the heuristic numerical procedure for stratified inverse sampling stops when (and only when) the target is reached.

This suggests a strategy for sampling to assess the quality of linkage consisting of:

1. If reasonable estimates of $\tilde{p}_h$ are available, use them in a Neyman allocation in a stratified design. This will give the smallest sample size with reasonable chance of achieving the required cv.
2. If these are not available, and only an overall estimate is required, use inverse sampling on randomly sorted data.
3. If separate estimates in the strata are desirable, follow the algorithm for stratified inverse sampling.

There may be further options where inverse sampling is used to obtain rough estimates of $\tilde{p}_h$, which are then plugged in to the Neyman allocation, and this may provide better control of overall sample size when the $\tilde{p}_h$ are initially unknown. This awaits further research.

References

Cochran, W.G. (1977), *Sampling techniques*. New York: Wiley.

Ferrante, A. & J. Boyd (2012), "A Transparent and Transportable Methodology for Evaluating Data Linkage Software", *Journal of Biomedical Informatics*, 45, pp. 165-172.

Haldane, J.B.S. (1945), "On a Method of Estimating Frequencies", *Biometrika*, 33, pp. 222-225.

Haziza, D., G. Chauvet, and J.C. Deville (2010), "A Note on Sampling and Estimation in the Presence of Cut-off Sampling", *Australian and New Zealand Journal of Statistics*, 52, pp. 303-319.

Herzog, T. N., F. J..Scheuren, and W. E. Winkler (2007), *Data quality and record linkage techniques*. New York: Springer Science & Business Media.

Hopper, N.A. (2011), "Predicting patterns of household non-response in the 2011 Census", *Survey Methodology Bulletin*, 69, pp. 9-22.

Kim, S., and J.A. Nachlas (1984), "Estimation in Bernoulli Trials Under a Generalized Sampling Plan", *Technometrics*, 26, pp. 379-387.

Randall, S.M., A.M. Ferrante, J.H. Boyd, and J.B. Semmens (2013), "The Effect of Data Cleaning on Record Linkage Quality", *BMC Medical Informatics and Decision Making*, 13:64, pp. 1-10.

Smith, P. (2013), "Sampling and Estimation for Business Surveys", in G. Snijkers et al. *Designing and conducting business surveys*. Hoboken, New Jersey: Wiley, pp. 165-218.

Vatsalan, D., P. Christen, C. O'Keefe, and V.S. Verykios (2014), "An Evaluation Framework for Privacy-Preserving Record Linkage", *Journal of Privacy and Confidentiality*, 6, pp. 35-75.

Winkler, W. E. (2004), "Approximate string comparator search strategies for very large administrative lists", *Proceedings of the Section on Survey Research Methods*, 2004, pp. 4595-4602.