Applications of record linkage to population statistics in the UK

Dick Heasman, Mark Baillie[1], James Danielis, Paula McLeod, Meghan Elkin
Office for National Statistics, Fareham, Hampshire, PO15 5RR, United Kingdom
dick.heasman@ons.gsi.gov.uk

Abstract: Population statistics in the UK are benchmarked by a decennial Census, the latest one having been held in March 2011. The design of the process to match the Census to the Census Coverage Survey (CCS) is briefly described. It is possible that traditional census-taking will not have the same role in the future of UK population statistics, the creation of a population spine by integrating administrative data sources being one of the alternative options under investigation. Some of the challenges involved in linking such data sources are outlined. Significant questions include the sensitivity of matching procedures to failures of the conditional independence assumption and how to handle missing values. The use of simulated data to research these and other issues in record linkage is presented.

Keywords: record linkage, administrative data, population statistics, missing values, conditional independence

1. Introduction

Population statistics in the UK are benchmarked by a decennial Census, the latest one having been held in March 2011. To adjust for undercount the Office for National Statistics (ONS), and its associated statistical offices for the devolved administrations of Scotland and Northern Ireland, hold a Census Coverage Survey (CCS) in order to make a final estimate of key population domains using the Dual System Estimator. This process is extremely sensitive to the accuracy of the match between Census and CCS. Section 2 reports briefly on how ONS is performing this matching.

It is possible that traditional census-taking will not have the same role in the future of UK population statistics. ONS has established the Beyond 2011 Programme, which aims to investigate: the feasibility of improving population statistics in the UK by making use of integrated data sources to replace or complement existing approaches; and whether alternative data sources can provide the priority statistics on the characteristics of small populations, typically provided by a Census.

The creation of a population spine by integrating administrative data sources is one of the options being investigated by the Beyond 2011 Programme. However, the data sources available are not designed with population statistics in mind, nor do any of them aim to achieve complete population coverage. The challenges involved in linking the data sources are therefore expected to be considerable, and to include factors such as

[1] Formerly of the Office for National Statistics

variations in definitions and missing values in data fields, as well as poorly recorded data and unaligned dates of data capture.

Section 3 briefly describes the Beyond 2011 administrative data sources option, early plans for carrying out the record linkage and the problems in terms of data quality that ONS anticipates having to address. Section 4 discusses some of the methodological research work being carried out to provide evidence on the best way to proceed with the linkage. The design of simulated administrative data sources and ongoing research into the sensitivity of matching procedures to failures of the conditional independence assumption, the handling of missing data and optimal matching methods are outlined.

2. Matching in the Context of 2011 Census Coverage Assessment

2.1 The 2011 Census

The constituent nations of the United Kingdom (UK) are England, Wales, Scotland and Northern Ireland. ONS is responsible for the conduct of the 2011 Census in England and Wales. Devolved administrations are responsible for the Census in Scotland and Northern Ireland. All three offices work closely together to deliver a Census for the whole of the UK, which is currently estimated to have 62.3 million usual residents.

England and Wales comprise 376 Local Authorities (LAs). ONS is confident that the two key targets for coverage have been hit, namely that at least 94% of the population have responded to the Census, and that there has been a response rate of over 80% in every LA. In addition, it is estimated that less than 10% of LAs have below a 90% response rate and that Inner London Boroughs (a subset of LAs that had particular coverage problems in 2001) had 5 to 15 percentage points higher response than in 2001.

2.2 The 2011 Census Coverage Survey

Despite these encouraging outcomes to the Census Field Operation, a Census coverage and adjustment strategy is necessary to meet the requirements that the national population estimate is within +/-0.2% of the truth, and that all LA population estimates should be within +/-0.3%, both with a 95% confidence interval.

To achieve this adjustment ONS conducted, as in 2001, the Census Coverage Survey (CCS). It is a survey covering all households and individuals in households in about one per cent of all postcodes (a UK postcode typically consists of between 15 and 25 dwellings). The sampling strategy ensures a sufficient presence in the survey for each LA, but also uses stratification by 'Hard To Count' categories, ensuring a greater sample in those areas where Census response is expected to be lowest. ONS is confident that the CCS in 2011 achieved a 90% response rate, a significant achievement for a voluntary survey in a time of falling response rates to ONS social surveys.

Although the CCS fieldwork operation is kept strictly independent of that of the Census, the CCS questionnaire is designed specifically to facilitate the matching of its results to the Census. When this matching has been completed, the true population count for the key population groups (quinary age band by sex by LA) are estimated using the Dual System Estimator. While this paper is not the place to discuss its workings, it is worth noting that the estimates are sensitive to errors in matching, due to the small size of the

survey sample relative to the population. Sensitivity is even greater in populations with low response in either the Census or the survey.

2.3 Possible Census overcount

As well as the risk of undercount through Census non-response, there is also, for a variety of reasons, the risk that duplicate Census returns will be made for the same individuals. ONS is running a matching exercise to detect duplicates in samples of the Census database, in order to estimate the size of any overcount.

2.4 Matching the 2011 CCS and Census

As noted above, the impact of matching errors, whether false positives or negatives, is unusually high for this particular matching exercise. To prevent such errors from occurring as far as is practicable, a large number of clerical matchers is used. They are deployed at various stages of the procedure: resolving all multiple matches arising from the automated matching; quality checking samples of the matches made by the automated matching (every LA will have a sample checked); resolving the pairs allocated to the clerical review region by the automated matching; and searching the Census database for residual units from the CCS that are not matched.

Clerical matching is aided by the clerical matchers having access, where necessary, to the images of Census returns. Thus all sorts of contextual clues can be taken into account. In total, the clerical matching will be a full-time job for at least 23 people for 9 to 12 months. Not all of this effort will be directed towards matching the CCS and the Census: some will also be expended, for instance, on resolving the clerical review region in the detection of duplicates.

The matching has an overarching hierarchical structure: households are firstly matched, followed by individuals within households, then individuals within the set of unmatched households. Finally, residual units still not matched are submitted for clerical review.

The automated matching has been written in SAS and designed in a modular format using macros for flexibility and ease of development. It is not the intention of this paper to go through each macro or to give an exhaustive account of how it works. Instead, the key choices, from a methodological point of view, made in the design of the matching are now briefly listed and explained.

- The matching exercise is carried out on pre-imputed data, e.g. dates of birth are not donated to records where this information is missing.
- Since name variables are strong personal identifiers, any records with null or blank values in these fields are excluded from the automatic matching and clerical resolution. They are made available for matching in the final clerical review.
- Data are cleaned by removing any white-space characters, transforming all strings to upper case and the removal of common tokens such as titles. Standardisation is carried out by the use of look-up lists for common abbreviations and acronyms in variables such as address and ensuring a standard entry for telephone numbers.
- Data on date of birth are collected as three separate variables and all are used in the matching of individuals.
- In each phase, exact matching (i.e. perfect agreement on all fields) is used before the residues are matched using probabilistic matching.

- The probabilistic method used is that of Fellegi and Sunter. Macros have also been prepared to carry out the method of Copas and Hilton (used in 2001) but it is unlikely that these will be needed.
- The parameters of the probabilistic matching are not determined by using training data. The main reason for this is a practical one: data is processed and comes in to the matching exercise on an LA by LA basis. It will take several months before the data from the last LA to be processed are received. If training data were to be used they would have to come from the 'early' LAs, but would not necessarily be typical of data for the rest of the country.
- Choice of matching variables may also vary from one LA to another. The matching system has the functionality to evaluate the discriminative power of each variable on a per LA basis. If the diagnostics indicate that the default matching variables provide low levels of matches for one LA, a new set of matching and blocking variables may be selected and the matching process rerun.
- Throughout the probabilistic matching process, conditional independence is assumed. Although the matching variables used will not in fact be fully statistically independent, they are chosen to remove redundancy as far as possible.
- Search space reduction is achieved through blocking. Consideration is given to the level of independence between blocking and matching variables. Multiple blocking passes are used to guard against the blocking strategy creating false negatives.
- Macros have been written for many different string comparators. Matchers can experiment with using a different choice if from the diagnostics the default does not appear to be working well. The default is bigrams, based on the assumption that most textual errors will result from the difficulty of reading Census returns using optical character recognition technology. Each string comparator must be input with a threshold, as the output is a decision on whether the strings agree or disagree.
- If the value of a variable is missing from either dataset, it is not entered for comparison and the variable makes no contribution to the total matching score for the record pair.
- Parameter estimation is achieved using the EM algorithm. Testing has found that starting values of 0.51 for m-probabilities and 0.49 for u-probabilities work perfectly well.
- The clerical review regions are initially set to be wide, but are likely to be narrowed after results of clerical reviews are fed back from the first few LAs.

3. Linking Administrative Data for the Beyond 2011 Project

3.1 Background

The Beyond 2011 Programme is investigating three major options: administrative data based options, census options and survey options. Within the first of these, the record level model is a possible candidate. Section 3.2 outlines the major administrative data sets available to ONS, section 3.3 outlines the record level model and section 3.4 discusses quality issues connected to the data sets available.

3.2 Administrative sources

The two administrative data sources that have the potential to include the national population at all ages are the Patient Register Database (PRD), which holds details on individuals registered with doctors, and the 'Customer Information System' (CIS), which holds records of people who have been a client or customer of the tax or benefit administrations, i.e. tax payers and benefit and pensions claimants.

Some other administrative sources provide good coverage of certain sub-populations which may be hard to count on these broad coverage sources. These include:

- The Migrant Worker Scan, which provides information on international migrants to the UK who have registered for and been allocated a National Insurance number.
- The School Census which collects data on state school pupils in England. It has good coverage of children aged 5-15 in England and collects a broad range of demographic information.
- HESA Student Data which records students registered at Higher Education institutions who are following a course that will lead to a qualification.
- Birth, marriage and death records and the electoral roll.

3.3 The record level model

Under this model the broad coverage administrative sources, the CIS and PRD, would be linked together at the individual level to produce an initial population spine. This might then be linked to the other datasets to attempt to improve coverage in hard to count groups. Address information in the spine might also be linked to a register of addresses. Rules would need to be developed on what combination of appearances in the data sets would qualify a record to appear in the population spine.

Linking the sources would lead to a degree of over and under estimation of the sizes of the key population groups. A coverage assessment (which is likely to take the form of a survey) would therefore be needed to assess the accuracy of the initial population count. Individuals in the coverage survey would be matched to records arising from the original data linkage process. This would enable ONS to estimate the extent of over and under count in different domains and to estimate weights which could be applied to adjust the initial population estimates to correct for this.

3.4 Quality of the data sources

The primary reason for collecting the data sources described above was not to enhance demographic statistics. Consequently they contain some features likely to present ONS with a challenge when it comes to link them. An example of this appeared in an exercise where PRD was matched to School Census data for 5 to 15 year olds. Exact and unique matches by sex, date of birth and postcode alone[2] outnumbered the exact matches made when first and last name were added to the matching variables. This was despite the fact that excluding the name fields led to a reduced number of unique matches as matches involving multiple birth siblings could not be distinguished as unique. Only when allowance was made for typographical errors and spelling variations in the name matching using the careful application of a string comparator did the inclusion of name data lead to a better matching result.

[2] Sampling and clerical checking revealed very few of these matches to be false links.

For legal reasons, ONS has yet to gain access to record-level CIS data. The issues that may occur with this source are therefore not fully explored, but there is no particular reason to expect that it will be of better quality than others. In one respect, it is expected to be worse, in that a record will only be updated after some contact between the tax or benefit authorities and the individual. For some, these episodes may be years apart.

In all the data sources but to varying degrees, ONS expects to find missing values in some variables. For record linkage purposes, missing data is still of some use and clearly of more use than imputed data. ONS is therefore requesting its suppliers not to impute data into these sources. Despite this, with computer collection making it increasingly hard to miss data, cases are 'forced' through, for instance where benefit claimants are given the postcode of the Job Centre, or members of some ethnic groups all have recorded date of birth as 1 January. To guard against this type of risk, ONS will need to look at histograms of matching variables and consider whether the values recorded at a spike are genuine. Where it appears more likely that they are not, it will be better to recode these values as missing.

4. Research on Record Linkage Methods

Within ONS there is a Methodology Directorate (MD) to provide the technical foundation for the production and analysis of official statistics, one part of which is currently providing methodological support for the planning of the Beyond 2011 record level model. This work is currently concentrating on the treatment of missing values, the advantages and risks involved with making the conditional independence assumption, evaluating possible different matching methods (e.g. deterministic, probabilistic), and evaluating software packages that might be used to carry out the linkage for the model.

4.1 Missing values

A search of the literature reveals that under standard probabilistic methods, missing values can be catered for by leaving out any variable that contains missing values from the computation of the total weight for the record pair in which the missingness occurs. This approach, recommended in Hand & Yu, 2001[3], is implemented in matching the 2011 Census to the CCS and Tromp *et al*, 2008. Under the standard Fellegi-Sunter model of record linkage, which respectively assigns a positive or negative individual weight to a variable depending on whether there is agreement or disagreement, this is equivalent to assigning a zero weight to a variable where a missing value occurs in the pair.

Other methods can also cater for missing values. The software FRIL, for instance, employs user-allocated and automatically tuned weights for matching variables. Each of these weights is multiplied by a 'score' representing the agreement between the values in the two data sets, the score being a number in the interval [0,1]. Where missing values occur, the score can be set by the user. The default missing value score is 0.5.

[3]Hand & Yu is not a paper about record linkage *per se,* but a paper on supervised classification methods. Record linkage is a special case of such methods, which seeks to classify record pairs into two or three classes. The authors claim that this approach to missing values is consistent with making the conditional independence assumption, but not consistent with accounting for dependencies.

4.2 The Conditional Independence Assumption (CIA)

The CIA is that the joint distributions of the agreement statuses of the match variables, both conditional on a record pair being a match and on it being a non-match, are independent. Failures of the CIA among matches can arise, for instance, when one of the sources is collected using optical character recognition on a handwritten return. Since respondents with poor handwriting can give rise to errors in a number of matching variables, agreement/disagreement with another source on these variables is positively correlated.

Failures among non-matches can be more pronounced. In the UK, fashions in the naming of babies change considerably over time. Hence if two different individuals agree on year of birth, the probability that they also agree on first name is increased. Clearly, if sex is used as a matching variable, first name has dependency on it.

The classical record linkage model uses the CIA, as also do methods such as those employed by FRIL. MD has conducted a review of the literature on the use of the CIA in matching, with the particular aim of discovering evidence from empirical studies on the sensitivity of the accuracy of the matching outcome[4] to failures in the CIA.

Hand & Yu discuss the independence model in the field of supervised classification methods. They find that this approach, despite being based on a model that is clearly unrealistic in most cases, has a long and successful history. They argue that the main reasons for this are: the simplicity of the model means that it requires fewer parameters to be estimated than alternative methods, resulting in a lower variance for the estimates; that the model may not give accurate probability estimates but these are not needed for classification as all that needs to be preserved is rank order; and that in real problems, variables typically undergo a selection process before being combined to yield a classification, resulting in a tendency towards using only weakly correlated variables.

The review found some papers reporting empirical findings on the performance of the independence model in record linkage. These are marked ^ in the references section. Several of these focus on designing methods to model the dependencies, and it is therefore not surprising that they have found that accounting for dependencies, where they exist, results in improved matching. Some (e.g. Schürle, 2003, using street name, postcode and district in the Berlin telephone directory; Tromp *et al* using child's expected data of birth and child's actual date of birth in matching perinatal data) use highly correlated variables in coming to these conclusions. Exceptionally Sharp (2011) investigates the use of only moderately correlated variables and finds the independence model yields a better performance than one that accounts for all dependencies.

The review found that the conclusion of Winkler (1999) is still valid; matching quality is improved using dependence methods but it has not been demonstrated that accounting for dependencies is assured to yield appropriately good quality matching in actual record linkage software on a day to day basis.

[4] By 'matching outcome' we mean the classification of the pairs into matches and non-matches. At the moment, MD is less interested in how sensitive the underlying statistical model is to failures in the CIA.

Tromp *et al* make an interesting point about dependencies between variables in the non-matches. They note that these usually arise from a latent factor, in their case the timing of the pregnancy. Matching administrative data to provide a population spine will encounter similar, if less severe, situations. For instance, where ethnic minorities exist in the population and choose names from a different name pool to the majority population, dependency between first name and last name will arise.

4.3 Synthetic data

MD has created synthetic data sets for use in record linkage research, training and software evaluation. These were first used for the on-the-job training course provided by the Data Integration ESSnet in January 2011. The data sets created are a 'truth' data set and three others meant to simulate the 2011 Census and contemporaneous versions of the PRD and CIS. The truth data set contains approximately 25,000 records of individuals. The other three data sets are large subsets (90 or 95 per cent) of the truth set with errors, some of which are correlated, introduced into the matching variables. The errors include replacing values by blanks, as missing values are known to be a problem.

The truth data set is built up in layers. Initially a district is chosen, then postcodes within it. Next, street and streets numbers are allocated from a set of street names. Now each address is populated with a household, with the number of persons being randomly generated. In most cases all members of the household are allocated the same last name. Finally first names and dates of birth are allocated to the persons, with controls over the year of birth to make the population profile and structure of households realistic.

4.4 Deterministic methods (rule-based matching)

ONS has a long tradition of employing deterministic methods of matching, most notably in updating the Longitudinal Study (an anonymised research database based on a sample of persons, having one of four specified birthdays, extracted from the Census and updated using birth and death records) from the latest Census. MD prefers the phrase used by the Relais developers: rule-based matching.

A rule-based match consists of a number of sub-rules. A sub-rule states a condition that the record pair must satisfy in order to be classified as a match, and may consist of conditions that must be checked at once; these conditions are separated by an "AND" operator. The different sub-rules are separated by an "OR" operator[5]. Schemes for rule-based matches can be very complicated, and to aid understanding are often illustrated by a flow diagram.

MD uses the term 'score-based matching' to draw a contrast with rule-based matching. This term is meant to cover any method whereby a weight is allocated to each matching variable; the weight is multiplied by a factor on the interval [0,1] which represents agreement status on that variable for the record pair to provide a score for the variable; and the variable scores are summed to provide a total score for the record pair.

At first sight, there seems to be an obvious principle: for any rule-based match there exists a score-based match that will make the same matches, and if it also makes other matches, these will be as good in quality as those made by the lowest quality sub-rule.

[5] If desired, the sub-rules could be structured in such a way that they are mutually exclusive.

But actually this only applies under the CIA. Indeed, it is possible to construct a highly artificial example with dependent variables as a counter-example.

MD has undertaken a piece of empirical work to discover whether the principle applies in realistic data sets. Using the PRD and CIS sets in the synthetic data a rule-based match was devised for matching them, and twice refined to improve its performance in the light of results. The 'true' match status can be readily checked in the synthetic data, and performance is measured simply by the number of true matches made minus the number of false matches made.

A score-based match was then derived from the refined rule-based match, allocating weights to the matching variables and setting a threshold in such a way that all the matches made by the rule-based match must be made by the score-based match. Table 1 shows that the extra matches made by the score-based match include more true than false links and that therefore overall performance is improved. While carrying out this work it became apparent that a score-based match with a higher threshold performed better still: although it did not make all the matches made by the rule-based match it made both more true matches and fewer false matches. This is also shown in table 1.

Table 1: *Performance of rule-based and score-based matching options in matching synthetic PRD and CIS data sets*[1]

	Rule-based match	***Score-based match*[2] *threshold = 5***	***Score-based match threshold = 6***
True matches made	21,298	21,704	21,508
False matches made	145	524	112
Difference[3]	21,153	21,180	21,396

[1]Total number of true matches = 22,860
[2]Formulated to make all the matches made by the rule-based match
[3]Simple measure of performance of the matching method

To complete this research work the data sets were made more of a challenge for score-based matching by increasing the dependence between some variables. In matches, this was done by making errors or blanks occur simultaneously in some records for day, month and year of birth. In non-matches, the population was divided into three different artificial ethnicities, distinguished by having distinct sets of names to choose from for first and last name. The first names in the majority ethnicity were further subdivided into three sets and assigned respectively to three different age cohorts of the population. This has given ONS a new and perhaps more useful set of data for record linkage research. The above exercise will now be repeated on the new PRD and CIS data sets.

4.5 Score-based and probabilistic matching

The Fellegi-Sunter method for record linkage is one type of score-based matching, where a weight is the logarithm of the ratio of the m-probability to the u-probability, and the multiplying factors are simply 1 for agreement and 0 for disagreement. Nadeau *et al.*, 2006, describe probabilistic matching in a loose sense as allowing for alternative ways other than probabilities to determine the weight for each variable, and MD's use of the term score-based matching has the same intention.

MD has recently started another piece of research work to compare matching under the Fellegi-Sunter model with the alternative strategies that still use the concept of weights, which are grouped under the description of the 'allocated weights method'. Part of this will be an information-gathering exercise. Box 1 lists the main *perceived* advantages and disadvantages of the two methods. For example, the Fellegi-Sunter model, or the softwares currently developed for implementing it, appears to be restricted to handling only binary levels of variable agreement, thus discarding useful information, while the alternative methods appear to be based on no model at all. The research will aim to find out to what extent these perceptions are true, how important the advantages and disadvantages are in practice, and whether there are work-arounds for the disadvantages.

Box 1: *Comparison of perceived advantages and disadvantages of two different types of score-based match*

	Fellegi-Sunter method	***Allocated weights method***
Advantages	1. Gives a framework for estimating the parameters 2. Parameters can be tuned using the EM algorithm 3. Model can be used for error estimation 4. Can cater for missing values when CIA is made	1. Only half the number of parameters need be estimated, one for each variable instead of two 2. Parameters can be tuned (e.g. in the FRIL software) 3. Can cater for missing values in a way that allows user flexibility 4. Can flexibility cater for partial agreements
Disadvantages	1. Double the number of parameters to estimate in the model makes for more uncertainty 2. Restricted to binary agreement values	1. Initial parameter estimation is at best expert opinion and at worst guesswork 2. Lacks an underlying statistical model

Finally, research is planned to see if the two methods converge under certain conditions. A software that uses the Fellegi-Sunter method and the FRIL software could both be set to match the synthetic PRD and CIS data sets: both would use the same set of matching variables; both would use only binary agreement values; both would use the same blocking strategy; and both would use the EM algorithm to tune their parameters. Both methods would do the same job of classification into matches or non-matches if they imposed the same rank order on the pairs. If they were different, the distribution of the true matches on the ordinal scale could be compared to determine if either method had a superior performance to the other in this experiment.

References

Fellegi, I. P. & Sunter, A. B. (1969) A Theory for Record Linkage, *Journal of the American Statistical Association*, 64(328), 1183-1210.

Hand, D. J. & Yu, K. (2001) Idiot's Bayes: Not So Stupid After All?, *International Statistical Review*, 69(3), 385-398.

Jurczyk, P. (2009) FRIL: Fine-grained Integration and Record Linkage Tool V3.2: Tutorial. Available at http://fril.sourceforge.net/. Copyright: Emory University, Math&CS Department, 2009.

Nadeau, C., Beaudet, M. P. & Marion, J. (2006) Deterministic and Probabilistic Record Linkage, *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health.* Available at:

http://www.statcan.gc.ca/pub/11-522-x/2006001/article/10404-eng.pdf.
Office for National Statistics (2010) papers on the Census to CCS matching are available from the presenter on application by email.
Office for National Statistics (2011) Census Roadshows – September 2011. Available on request from census.customerservices@ons.gsi.gov.uk.
Office for National Statistics (2011) Beyond 2011: Administrative data sources and low-level aggregate models for producing population estimates *(presented at the Annual Conference of the British Society for Population Studies 2011).*
^Schürle, J. (2003) A method for consideration of conditional dependencies in the Fellegi and Sunter model of record linkage, *Statistical Papers*, 46, 433-449.
^Sharp, S. (2011) The Conditional Independence Assumption in Probabilistic Record Linkage Methods *(presented at the sixteenth GSS Methodology Symposium).* Edinburgh: National Records Scotland.
^Thibaudeau, Y. (1989) Fitting Log-Linear Models in Computer Matching, *Proceedings on the Section on Statistical Computing, American Statistical Association*, 283-288.
^Thibaudeau, Y. (1993) The Discrimination Power of Dependency Structures in Record Linkage, *Survey Methodology*, 19(1), 31-38.
^Tromp, M., Méray, N., Ravelli, A. C., Reitsma, J. B. & Bonsel, G. J. (2008) Ignoring Dependency between Linking Variables and Its Impact on the Outcome of Probabilistic Record Linkage Studies, *Journal of the American Medical Informatics Association*, 15, 654-660.
^Winkler, W. E. (1989) Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage, *Survey Methodology*, 15(1), 101-117.