Evaluating Record Linkage Software for Agricultural Surveys

Michael E. Bellow[1], Kara Daniel[1], Mark Gorsak[1], Andreea L. Erciulescu[1, 2]

[1]National Agricultural Statistics Service, 1400 Independence Ave., SW, Washington, DC 20250
[2]National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, NC 27709

Abstract
Reducing duplication and matching records lacking unique identifiers are common practices associated with the construction and maintenance of a list sampling frame. The U.S. Department of Agriculture's National Agricultural Statistics Service (NASS) employs a record linkage system built using AutoStan and AutoMatch (originally developed by MatchWare Technologies) to maintain its list frame of farm operators and agribusinesses. The overall process consists of four main steps: 1) reformatting, 2) standardizing (AutoStan), 3) matching (AutoMatch) and 4) review. Because the current match engine is no longer supported and becoming increasingly obsolete, NASS has recently begun to explore alternative software options, such as Statistics Canada's G-Link package. In this paper, we describe the results of a preliminary study comparing G-Link with AutoMatch using list frame data from a national survey of organic farmers and discuss issues associated with upgrading the agency's record linkage system.

Key Words: Record Linkage, List Sampling Frame, Duplication, AutoMatch/AutoStan, G-Link

1. Introduction

1.1 Record Linkage at NASS

The U.S. Department of Agriculture's National Agricultural Statistics Service (NASS) is responsible for conducting the Census of Agriculture (COA) and additional agricultural surveys. The samples for these surveys are drawn from NASS's list sampling frame (LSF) of known and potential farms. In order to ensure estimation accuracy, it is important that this frame be as complete and as free of duplication as possible. An important tool in maintaining and updating the LSF is *record linkage*, a technique that uses computer algorithms to link records not having unique identifiers (which would make the matching trivial).

NASS keeps its LSF as current as possible by obtaining new lists of individuals and operations likely to be involved in agriculture. The lists come from various sources, such as other USDA agencies, state departments of agriculture, state property assessor lists and agricultural membership lists. Analysts apply record linkage techniques to match outside source lists to the LSF and then use the matched records to update information on existing LSF records. In addition, they mark records from outside source lists that are not found on the list frame as potential farms and then check whether those records satisfy the agency's farm definition and should be included in the population for sample surveys and the COA.

NASS also uses record linkage on a regular basis to identify and remove duplication within the LSF. Other applications include routine LSF maintenance (such as updating phone numbers or identifying deceased operators) and overlapping records residing on one frame to another. NASS uses a database tool known as ELMO (Enhanced List Maintenance Operations) to build and maintain its list frame (Bailey and Apodaca, 2015).

In 1994, NASS selected *AutoMatch/AutoStan* as its record linkage software system after an evaluation of six candidate products to replace the previous system (known as RECLSS) which had been deemed inadequate for future list frame maintenance needs (Day, 1995). At the time, responsibility for the COA had recently been transferred from the U.S. Bureau of the Census to NASS, and the latter needed to be prepared for the additional data processing challenges presented by this new assignment.

In conducting the COA for the final time in 1992, the Census Bureau had used a computerized record linkage model based on the probabilistic *Felligi-Sunter theory* (Felligi and Sunter, 1969) for the first time and also introduced effective algorithms for dealing with typographical errors. When compared with the previous COA in 1987 (where ad hoc methods for parsing names and addresses had been used), these measures resulted in a reduction of the number of clerical staff hours by about 50 percent and the proportion of duplicate records on the final list frame from ten to two percent (Herzog et al., 2010).

Matt Jaro (a former Census Bureau employee) and his team at MatchWare Technologies developed the original AutoMatch and AutoStan programs (Jaro, 1999). AutoStan was designed to standardize incoming records and break variables into their component parts, while AutoMatch employs the Felligi-Sunter methodology to link records within or between lists and to classify them as either definite matches, possible matches or non-matches based on comparing probability ratios with pre-selected threshold values. In separate empirical studies conducted using data from North Carolina (Day, 1996) and Ohio (Broadbent, 1996), respectively, AutoMatch was found to perform substantially better than RECLSS.

Since 2005, IBM Corporation has owned sales and marketing rights to AutoMatch/AutoStan as part of its quality stage program. While support costs have escalated and significant updates have been made to the system, NASS has been using the same version since 2001 and thus no longer pays for support or receives upgrades. Although this version of AutoMatch/AutoStan still meets its record linkage needs, for the long term NASS would like to implement an available product that is more efficient, uses state of the art technology and is fully supported. The first potential replacement software system to be considered is Statistics Canada's *G-Link,* which NASS obtained free of charge. The primary focus of this paper is a pilot study comparing G-Link with AutoMatch.

An important part of NASS's initial implementation of AutoMatch/AutoStan was the development of user interfaces using *PowerBuilder*, a tool developed by Sybase for building object-oriented client/server applications. A front-end program simplifies parameter preparation and also has functionality of template parameters for routine matches. When NASS converts to a new record linkage system, the front ends will likely no longer be used as most of their functionality is specific to AutoMatch/AutoStan. There is also an interface for reviewing linked records which interacts with the LSF. This program allows reviewers to identify records believed to be matches and perform actions within the resolution system that will result in updates to existing list records and identification of new potential farm records to be added to the frame. Inasmuch as substantial resources have been invested in developing this resolution system, hopefully it can be incorporated into NASS's future record linkage system as part of the review process.

1.2 List Frame Updating Procedure

The first step in the general procedure for matching a list of potential agricultural operators to NASS's LSF is to place the list in a standard format. Analysts read the incoming list into SAS and insert variables into basic fields with formatting consistent with the frame. For example, names on an incoming list are commonly in 'surname on the left' format (e.g., Smith, Bob). These names are converted to signature format as they will eventually be stored on the list frame (i.e., Bob Smith).

Another example involves address, city, state and zip code as lists will often contain two or more of such variables in the same field. The reformatting process splits these variables into distinct fields as required by the list frame. Reviewers examine city, state and zip code entries to ensure that they meet postal standards and phone numbers to check their validity. The reformatted dataset is output as a fixed field text file and then input into AutoStan, which further parses each string into its component parts. The program standardizes all component parts and places them into identifiable fields. For example, the name Bob Smith Jr would be formatted with given name = ’Robert’, surname = ’Smith’ and suffix = ’JR’.

Following standardization, the list is ready for matching. The AutoMatch program matches the standardized outside source file to a similar file extracted from the LSF and also looks for duplication within the outside source records. The user runs a series of passes with different blocking and matching variables to bring records together (multiple passes are used to overcome problems with missing and incorrect data as well as situations where data are maintained differently). After the AutoMatch processing is done, a SAS program combines linked records from all passes into conglomerate link groups for a singular review. Outside source records may be linked to different list frame records in different passes and those classified as non-matches in a given pass are not removed from consideration in upcoming passes. The conglomerate final link group contains the outside source record and each of the linked list frame records. The SAS program also brings in any additional list frame records related to the same operation. A common occurrence is for many individuals to be associated with a single agricultural operation and it is also possible for one individual to be involved in multiple operations. Such related agricultural records have identification numbers on the LSF that group them together and are used to ensure that they will appear in the same final link group for review.

Each conglomerate link group is classified as either a match, possible match or non-match group. The conglomerate link groups are populated to the NASS resolution system for review. If any individual linkage between records within a conglomerate link group is classified as a possible match, the entire group is classified as a possible match group. Analysts review all possible matches, a portion of the non-matches, matches to list frame records coded as non-farms and link groups where an outside source record is linked to multiple known farms. Duplication within the LSF frame can often be identified and corrected through this review. Non-matches are reviewed if key elements (e.g., an address) are missing. The NASS resolution system interacts with the LSF and was designed to make the review process as simple as possible. Reviewers are trained to make decisions that will keep the frame current and accurate. Quality control programs are run for each project to ensure that the review is consistent with the training.

1.3 Required and Preferred Features

The following are some key required and desirable features of a future record linkage system at NASS:

1. *UNIX Compatibility*
   NASS has set up a secure area on a UNIX platform to protect Personally Identifiable Information and Federal Tax Information (IRS) data. The new record linkage system must be able to run on that platform.

2. *SAS Compatibility*
   Since SAS is the main software package used by NASS for its survey processing applications, the new system should be capable of reading and writing SAS datasets.

3. *Standardization*
   NASS will need a software tool (replacing AutoStan) for converting names and addresses into standard form and cleansing. This program could either come with the matching engine or be obtained separately.

4. *Interactivity/Dynamic Analysis*
   NASS would like to have the ability of performing record linkage interactively and assessing the effect of user selected input parameters so that adjustments can be made if necessary.

5. *Automatic Weight Computation*
   NASS would also like to have the capability of computing outcome and frequency weights automatically (with minimal user input).

6. *Deterministic and Probabilistic Matching*
   NASS wants records with certain matching variables (for example *Employer Identification Number*) brought together for review even if they have little else in common. The agency has always used Felligi-Sunter methodology for its record linkage applications, and the new system should employ either that or an alternative probabilistic technique.

Other factors that NASS will consider when selecting a new record linkage system include cost, support, features, continual development, statistical defensibility and ease of setting up projects. Regarding the last item, a desirable feature would be the ability to create templates for similar projects. For example, in preparing for the June Area Survey (an annual area frame based sample survey), NASS has only about a week to complete its overlap processing during which area frame extracts are matched to the list frame in 49 states (Cotter et al., 2010). The new system should be capable of processing large files (100 million or more records) and handling regional differences (e.g., varying address formats). The resolution software must be able to interact with ELMO in a single review as opposed to having to review multiple passes performed by the matching engine.

In the remainder of this paper, a software system called *G-Link* is described and compared with AutoMatch based on available features and matching accuracy. Additional candidate record linkage software packages may be considered in the future.

2. G-Link Record Linkage System

2.1 Description

G-Link is an iterative record linkage matching engine developed by Statistics Canada. Like AutoMatch, G-Link uses probabilistic Felligi-Sunter methodology. This system is an upgraded version of an earlier Statistics Canada product called the *Generalized Record Linkage Software (GRLS)* (Nuyens, 1993). NASS considered GRLS as a replacement for RECLSS in 1994, but ultimately rejected it due to the system's reliance on Oracle databases (which made it inappropriate for use with a Sybase database such as ELMO). However, G-Link uses SAS instead of Oracle as its database and thus is compatible with ELMO. Most of the core logic for G-Link has been in place since the mid 1990s, with subsequent changes to the system primarily involving the user interface. G-Link does not come with a companion standardization package.

The system is comprised of two components: 1) the client (interface, business logic and some metadata) implemented in C# and 2) the server implemented in SAS. The client portion must run on a Windows-capable platform and the machine running the client must also have Microsoft's .NET Framework 4.5.1 or higher. The server component handles all of the heavy data processing and can run either on a local PC, a Windows server or a UNIX machine. The linkage process is carried out as a sequence of distinct phases, where each phase involves choosing values for system parameters (for example rules and the criteria for comparing attributes), examining their effect and making any necessary adjustments to the values before moving on to the next phase. An earlier phase can be rerun with new adjustments that were suggested by later phases. G-Link does not physically alter the file or files that it is linking, meaning that the same file can be involved in several two-file linkages at the same time.

G-Link performs record linkage in three basic stages: 1) *search*, 2) *decide*, and 3) *group*. In the *search* stage, the user 1) loads the input (SAS) data sets, 2) creates a set of potential linked pairs by specifying a blocking condition using a subset of fields in the input files (e.g. surname or postal code), 3) creates a random sample of non-linked pairs that will be used later to calculate non-linked outcome weights, 4) builds compound (multiple criteria) matching rules, 5) performs a field-by-field comparison of records identified as potential pairs to generate outcomes based on the rules, 6) outputs record pairs with associated rule outcomes and odds ratios and 7) classifies pairs as either definite matches, possible matches or non-matches based on comparing odds ratios to user specified threshold values. Following step 1, the user can examine analysis tables to identify properties of the data (e.g., percentage of missing records per field) which may be useful in selecting blocking variables.

In the *decide* stage, the user can 1) adjust the odds ratio for existing linked pairs (without having to compare the input data again), 2) change threshold values, 3) revise outcome probabilities and automatically apply them to record pairs, 4) calculate value-specific odds ratios and apply them to record pairs and 5) calculate and apply frequency probabilities based on field values. After adjusting odds ratios for rule outcomes and frequency probabilities, the user reclassifies the pairs so that those with modified odds ratios falling below the lower threshold are now considered non-matches. Examination of various samples of record pairs as well as tabular and graphical displays available in G-Link can assist the user in adjusting the threshold values and selecting the best ones for a specific record linkage project.

The *group* stage involves grouping records according to the status of links between them. G-Link arranges records joined by definite or possible links into *weak link groups,* which can be very large. Within the weak groups, the program further divides records joined by definite links into *strong link groups* that contain the best links. G-Link identifies conflicts where a record on one of the input files is linked to several records on the other file. The user has the option of either allowing automatic conflict resolution (called *mapping*) or doing it manually via on-screen updating of group contents. A combined approach that first applies mapping and then allows the user to examine the results and perform some manual rearrangement is also possible.

G-Link can perform record linkage in a single pass in which a rule is created with multiple outcomes (e.g., complete agreement, NYSIIS agreement and typo agreement). The program determines the best possible outcome for each pair in one pass with the user having the option of creating a set of matching rules that are executed simultaneously (consistent with the methodology used to calculate the weights). However, processing can also be done in multiple passes.

The final output of G-Link is a SAS dataset containing the group information. This table can then be joined to the appropriate records from the two input files to enable further processing depending on the application.

2.2 Feature Comparison with AutoMatch/AutoStan

Table 1 indicates whether or not each of the required or desirable features of a NASS record linkage system listed in Section 2 is available with G-Link and AutoMatch/AutoStan, respectively.

Table 1. Software Products Feature Comparison

Feature	G-Link	AutoMatch/AutoStan
UNIX Compatibility	x	x
SAS Compatibility	x	x
Standardization		x
Interactivity and Dynamic Analysis	x	
Automatic Weight Computation	x	x*
Deterministic and Probabilistic Matching	x	x

* - semi-automated

3. Pilot Software Comparison Study

3.1 Plan

The immediate goal of the current record linkage research efforts at NASS is to assess the viability of G-Link as the future matching engine for the agency's list frame building and maintenance activities. A preliminary study done with non-NASS data verified the feasibility of the G-Link software that NASS obtained from Statistics Canada (Erciulescu, 2015). The next step was to carry out a larger study comparing G-Link with the version of AutoMatch currently used by NASS for operational record linkage with NASS data. This study represents a preliminary comparison of the two software products in two states based on a set of matching accuracy metrics, and the results should not be considered definitive.

The study was conducted using a version of G-Link installed on a laptop PC. The input data came from NASS's 2015 Certified Organic Survey (COS), a census of all operations having certified organic production. Data for the COS are collected directly from producers who participate voluntarily and on a confidential basis. NASS builds the population for this survey using a number of outside source lists, including the USDA Agricultural Marketing Service's (AMS) National Organic Program (NOP) list which was used for this pilot study.

Pennsylvania and Iowa (which ranked third and tenth, respectively in the U.S. in terms of organic sales according to the 2014 COS) were selected as the two test states for the study. NASS's Frames Maintenance Group (FMG) had already used AutoMatch to update its existing ELMO data on organic farms in those two states by linking the LSF with monthly lists provided by AMS as part of its INTEGRITY database of certified organic operations. Output files from the AutoMatch matching (prior to resolution) were available and used for the comparison study. G-Link was then run on the same two data sets and the results compared with those from AutoMatch. AutoStan was used to standardize names and addresses for both AutoMatch and G-Link.

The operational matching with AutoMatch involved multiple passes but (based on advice from Statistics Canada) single passes were run with G-Link and thus a pass-by-pass comparison between the two products was not possible. To ensure a valid comparison, the blocking variables for G-Link were a subset of those used previously with AutoMatch and the same matching variables were used. We conducted four separate trial runs for G-Link in both states, each with a different combination of certain input parameters. Although operator and operation phone numbers had been used as blocking variables in the AutoMatch processing, they were not used with G-Link due to the high percentage of missing values for these variables in the data sets.

Table 2 shows the general categories of blocking and matching variables used with each trial of either AutoMatch or G-Link, including whether or not *frequency weights* (which adjust for the fact that some values of an identifier are more common than others) were calculated and used in the classification. The labels GL-1 through GL-4 refer to the four sets of input parameters used in the G-Link trials for both states.

Table 2. Input Parameters for AutoMatch and G-Link Test Runs (Both States)

Category	**Variable Name**	**Trial Run**				
		AutoMatch	GL-1	GL-2	GL-3	GL-4
Blocking Variables	Operator Name	x	x	x	x	x
	Residence Address	x	x	x	x	x
	Operator Phone	x				
	Operation Name	x	x	x		
	Operation Address	x	x	x		
	Operation Phone	x				
Matching Variables	Operator Name	x	x	x	x	x
	Residence Address	x	x	x		x
Frequency Weights	Operator's First Name			x		x

3.2 Metrics

The *weighted Kappa coefficient* (Cohen, 1960) is a measure of association between two separate classifiers (or raters) working with the same set of items. If the two classifiers are labelled *A* and *B*, then this metric is defined in general terms as:

$$\kappa_W = 1 - \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}p_{ij}}{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}e_{ij}}$$

where:

c = number of categories
w_{ij} = weight associated with category *i* for classifier *A* and category *j* for classifier *B*
p_{ij} = observed proportion of items classified to category *i* by *A* and category *j* by *B*
e_{ij} = expected proportion of items classified to category *i* by *A* and category *j* by *B*

The computation weights just defined should not be confused with the outcome and frequency weights used in matching. In this specific application where *A* refers to AutoMatch, *B* to G-Link, category *1* to non-matches, *2* to possible matches and *3* to definite matches, the computation weights can be assigned as follows:

$w_{ii} = 0$ ($i = 1, 2, 3$)
$w_{i,\ i+1} = 1$ ($i = 1, 2$)
$w_{i,\ i-1} = 1$ ($i = 2, 3$)
$w_{1,3} = w_{3,1} = 2$

This choice of weights reflects the fact that classifying a non-match as a definite match (or vice versa) is a more severe error than any misclassification involving possible matches (which are reviewed manually). The expected proportions (under the assumption that the two classifiers are independent) can be estimated as the product of the two overall proportions:

$$e_{ij} = p_{i.}p_{.j}$$

where:

$p_{i.}$ = observed proportion of records classified to category *i* by AutoMatch ($i = 1, 2, 3$)
$p_{\cdot j}$ = observed proportion of records classified to category *j* by G-Link ($j = 1, 2, 3$)

The weighted Kappa coefficient then reduces to:

$$\kappa_W = 1 - \frac{p_{12} + p_{21} + 2(p_{13} + p_{31})}{p_{1.}p_{.2} + p_{2.}p_{.1} + 2\,(p_{1.}p_{.3} + p_{3.}p_{.1})}$$

3.3 Results

Table 3 compares the four G-Link trial runs in both states based on κ_W and the following two metrics:

p_{13} = proportion of AutoMatch definite matches classified as non-matches by G-Link
p_{23} = proportion of AutoMatch possible matches classified as non-matches by G-Link

Clearly, low values of p_{13} and p_{23} and high values of κ_W are desirable. Note that p_{13} was identical for the four test runs in both states while p_{23} was only slightly affected by whether or not the operation name and address were used as blocking variables or frequency weights were applied.

For this reason, the focus here is on the test runs that resulted in the highest values of κ_w in each state (i.e., GL-3). While $\kappa_w = 0.218$ for both GL-3 and GL-4 in Pennsylvania, GL-3 was chosen because it did not require the computation of frequency weights.

Table 4 shows a cross-tabulation of classified outside source (AMS) records for both states corresponding to the trial runs labelled GL-3, showing for the three AutoMatch categories (definite, possible and non-matches) the number and percent of outside source records that were classified in each of those classes by the combination of G-Link and the post-match processing in SAS.

Table 3. Comparison of G-Link Test Runs

State	**Run Label**	**Metric**		
		p_{13}	p_{23}	κ_w
Iowa	GL-1	0.018	0.108	0.306
	GL-2	0.018	0.111	0.257
	GL-3	0.018	0.108	0.307
	GL-4	0.018	0.108	0.255
Pennsylvania	GL-1	0.019	0.159	0.192
	GL-2	0.019	0.159	0.185
	GL-3	0.019	0.151	0.218
	GL-4	0.019	0.151	0.218

Table 4. Cross-Tabulation of Classified Outside Source Records

State	**AutoMatch Category**	**G-Link Category**			
		Definite Match	Possible Match	Non-Match	Total
Iowa	Definite Match	330 (67%)	151 (31%)	9 (2%)	490
	Possible Match	120 (38%)	162 (51%)	34 (11%)	316
	Non-Match	0 (0%)	0 (0%)	12 (100%)	12
	Total	*450*	*313*	*55*	*818*
Pennsylvania	Definite Match	225 (47%)	243 (51%)	9 (2%)	477
	Possible Match	101 (28%)	203 (57%)	54 (15%)	358
	Non-Match	0 (0%)	0 (0%)	17 (100%)	17
	Total	*326*	*446*	*80*	*852*

The most critical cases are AutoMatch definite matches that were categorized as non-matches by G-Link since, in general, very few non-matches are input into the resolution system for clerical review (note that there were nine such cases in each state). The AutoMatch definite matches that G-Link classified as possible matches (34 in Iowa and 54 in Pennsylvania) are of lesser concern since in actual practice they would be put through the resolution process.

Table 5 shows the number and percent of G-Link definite and possible matched AMS records that did not appear in the same link group as the 'true' matched LSF record based on the AutoMatch classification, post-match processing and resolution. Such outside source records have no chance of being matched to the correct record, assuming that the operational processing (which includes manual reviews) matched them correctly. Note that the overall percentages were roughly the same in both states.

Table 5. Statistics on Outside Source Records Misgrouped by G-Link

G-Link Match Type	**Iowa**		**Pennsylvania**	
	Number Misgrouped	Percent	Number Misgrouped	Percent
Definite	23	5.1	21	6.5
Possible	24	7.8	28	6.5
All	*47*	*6.2*	*49*	*6.5*

4. Conclusion

NASS makes extensive use of record linkage for various applications, including updating its LSF via outside source lists. The current record linkage system, which uses AutoMatch/AutoStan, is becoming outdated and will eventually need to be replaced. G-Link (which was obtained free of charge from Statistics Canada) is currently being evaluated as a potential replacement system.

Results of a pilot study comparing G-Link with AutoMatch using Certified Organic Survey data from Iowa and Pennsylvania were promising as the percentages of AutoMatch matches (definite or possible) classified as non-matches by G-Link were by no means unacceptably high. There were also significant reductions in processing time due to performing all of the matching in a single pass (with G-Link) compared with the multiple passes routinely used by NASS in its operational record linkage with AutoMatch.

G-Link is currently being installed on a UNIX platform at NASS for further evaluation in an environment that closely emulates the operational record linkage. NASS plans to explore standardization tools available from outside providers as possible replacements for AutoStan and may also evaluate additional matching products such as *LinkSolv* and *LinkageWiz* (Dusetzina et al., 2014).

5. References

Bailey J. and Apodaca M. (2015), "Compilation and Maintenance of the Master Frames used in the United States Agricultural Estimation Program and Census of Agriculture," 60th World Statistics Congress ISI 2015, Rio de Janeiro, Brazil.

Broadbent, K. (1996), *Record Linkage III: Experience Using AUTOMATCH in a State Office Setting*, Research Report No. STB-96-02, National Agricultural Statistics Service.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales". Educational and Psychological Measurement. **20** (1): 37–46.

Cotter, J., Davies, C., Nealon, J. and Roberts, R. (2010) "Area Frame Design for Agricultural Surveys", in (R. Benedetti, M. Bee, G. Espa and F. Piersimoni, eds.), *Agricultural Survey Methods,* John Wiley & Sons, Ltd, Chichester, UK, Ch. 11.

Day, C. (1995), *Record Linkage I: Evaluation of Commercially Available Record Linkage Software for Use in NASS,* Research Report No. STB-95-02, National Agricultural Statistics Service.

Day, C. (1996), *Record Linkage II: Experience using AUTOMATCH for Record Linkage in NASS*, Research Report No. STB-96-02, National Agricultural Statistics Service.

Dusetzina, S.B., Tyree, S., Meyer, A.M. et al. (2014), *Linking Data for Health Services Research: A Framework and Instructional Guide (Internet),* Rockville, MD: Agency for Healthcare Research and Quality (US): 2014 Sept. 4, "An Overview of Record Linkage Methods". Available from http://www.ncbi.nlm.nih.gov/books/NBK253312/.

Erciulescu, A.L. (2015), *Record Linkage Project – G-Link Test 1*, NASS internal document.

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

Herzog, T.N., Scheuren, F. and Winkler, W.E. (2010), "Record Linkage", in (D.W. Scott, Y. Said and E. Wegman, eds.), *Wiley Interdisciplinary Reviews: Computational Statistics*, New York, N.Y.: Wiley, 2 (5), September/October, 533-543.

Jaro, M. (1999), "Matchware Product Overview", in *Record Linkage Techniques – 1997, Proceedings of an International Workshop and Exposition*, National Academy Press, Washington, DC.

Nuyens, C. (1993), "Generalized Record Linkage at Statistics Canada". *Proceedings of the International Conference on Establishment Surveys*, Buffalo, NY, 926-930.

Statistics Canada (2011), *G-Link Concepts Guide*, Report No. 2011-03-01, Systems Engineering Division.