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

An Overview of Business Record Linkage at Statistics Canada: How to link the “Unlinkable”

Javier Oyarzun and Laura Wile[1]

Abstract

Statistics Canada’s mandate includes producing statistical data to shed light on current business issues. The linking of business records is an important aspect of the development, production, evaluation and analysis of these statistical data. As record linkage can intrude on one’s privacy, Statistics Canada uses it only when the public good is clear and outweighs the intrusion. Record linkage is experiencing a revival triggered by a greater use of administrative data in many statistical programs. There are many challenges to business record linkage. For example, many administrative files not have common identifiers, information is recorded is in non-standardized formats, information contains typographical errors, administrative data files are usually large in size, and finally the evaluation of multiple record pairings makes absolute comparison impractical and sometimes impossible.

Due to the importance and challenges associated with record linkage, Statistics Canada has been developing a record linkage standard to help users optimize their business record linkage process. For example, this process includes building on a record linkage blocking strategy that reduces the amount of record-pairs to compare and match, making use of Statistics Canada’s internal software to conduct deterministic and probabilistic matching, and creating standard business name and address fields on Statistics Canada’s Business Register. This article gives an overview of the business record linkage methodology and looks at various economic projects which use record linkage at Statistics Canada, these include projects in the National Accounts, International Trade, Agriculture and the Business Register.

Keywords: Record Linkage, Deterministic matching, Business Register, Partnership, Standardization.

1. Introduction

Statistics Canada’s mandate includes producing statistical data to shed light on current business issues. The linking of business records is an important aspect of the development, production, evaluation and analysis of these statistical data. Statistics Canada has been linking business records for many years, often performing direct linkage using the business number (BN). Statistics Canada also conducts extensive record linkage of data collected on individuals. Data collected on individuals can be less volatile in comparison to business data: individuals generally keep the same name, the same gender, and they age every year. The changing nature of businesses can lead to additional complexities when conducting record linkage: businesses can change their name, close down, reopen, and go through mergers and acquisitions. A recent increase in the number of business data record linkage projects at Statistics Canada has been observed and can be partially attributed to the fact that more statistical programs are using administrative data sources (which can lack BNs) and more of Statistics Canada’s existing programs are required to link to the Agency’s centralized Business Register (BR). Due to the importance and challenges with business data record linkage, Statistics Canada is working to develop a *generalized* record linkage strategy to help users conduct business data record linkage. This article discusses a proposed business data record linkage methodology in detail. Section 2 provides examples of business data record linkage projects conducted at Statistics Canada. Section 3 introduces some of the challenges one may encounter when trying to link business records. Section 4 describes the proposed business data record linkage methodology. The last section of this paper discusses plans for future work.

[1] Javier Oyarzun, Statistics Canada, Business Survey Methods Division, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (javier.oyarzun@canada.ca);
Laura Wile, Statistics Canada, Business Survey Methods Division, R.H. Coats Building, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (laura.wile@canada.ca).

2. Record Linkage and its Applications at Statistics Canada

This section will briefly discuss projects that use business data record linkage at Statistics Canada from which ideas are being drawn for establishing a generalized record linkage strategy.

There are a number of Statistics Canada's divisions that link administrative files without BNs to the BR.

- For example, data for exported goods are obtained from the Canada Border Services Agency and United States Bureau Census as well as other sources. Declarations are used in record linkage to create and maintain Statistics Canada's Exporter Register. Due to the lack of BNs on the Exporter Register, International Accounts and Trade Division links this register to the BR using business names and business addresses to obtain demographic information (see Auger (2015) or Byrd (2016) for details). As well, patent data files obtained from the Canadian Intellectual Property Office, the United States Patent and Trademark Office and the European Patent Office are also linked to the BR.

- In 2012, in an effort to reduce the number of frames maintained by Statistics Canada, the Farm Register (FR) was migrated to the BR. Since the FR did not include the BN for all farms, extensive record linkage work was conducted to reconcile both registers (Dongmo Jiongo *et al*., 2013). This reconciliation of the two registers revealed a discrepancy between the number of businesses identified for the agricultural sector of the BR and the number of farms included on the FR. This discrepancy can be mostly explained by unidentified partnerships (see Section 2.2 for more details about business partnerships) or misclassified farms. On-going record linkages have been conducted to identify potential duplicate records using business name, telephone number, business address, and administrative (tax) data (Gutoskie, 2016).

In order to be incorporated, a business must obtain a BN from the Canada Revenue Agency (CRA). There are a number of reasons why the "same" business could have multiple BNs on the BR. For example, the same enterprise can have a number of different BNs for the purposes of reporting financial data or the CRA may assign a new BN to the same enterprise if it undergoes a change, such as a change in ownership (Rollin, 2013). As well, for a given unincorporated business that does not require a BN, multiple entries for the "same" business can appear on the BR if the enterprise was not correctly identified in a partnership structure.

- The Economic Analysis Division (EAD) at Statistics Canada maintains a longitudinal database of Canadian enterprises called the National Accounts Longitudinal Microdata File. To track a business longitudinally for the purpose of labour analysis, it is not possible to rely solely on BNs because an enterprise can change BN without any change in its employees, which is considered as a false enterprise birth and a false enterprise death. To obtain information about the entry and exit of an enterprise, labour tracking is currently conducted (Rollin, 2013).The opportunity exists to explore other possible linkage variables for the purpose of other types of analysis, which can rely on the business name, business address, and industrial classification (North American Industry Classification System). The work conducted on this project can help in the improvement of the BR's identification process of predecessors and successors businesses.

- The administrative data on the BR is updated on a regular basis using administrative files from CRA through the process of record linkage. As well, record linkage can be used to conduct a frame quality assessment in order to identify and resolve duplication, overcoverage or undercoverage (Oyarzun and Wile, 2016). Recently, studies have been undertaken to improve the identification of partnerships on the BR. A partnership exists when two or more individuals are partners in the same unincorporated business. For tax purposes, each partner must report the same financial information, but with his or her share of the partnership. As explained previously, partnerships of unincorporated businesses without a BN can be difficult to identify. Failing to identify partnerships can lead to overcoverage which can result in overestimation for variables such as revenue. Therefore, partnerships need to be identified, namely using record linkage. An example of this record linkage application will be explored more in detail in Section 4.

3. Challenges with Business Record Linkage

When Fellegi and Sunter (1969) formalized probabilistic record linkage, most researchers did not have access to the administrative and computerized data that is available today. The recent drive towards the use of administrative data has led to a revival of record linkage and many new challenges. The challenges addressed in this paper fall into four categories: (i) administrative data processing, (ii) business names and addresses characteristics, (iii) standardization strategy, and (iv) identification of true matches.

3.1 Administrative Data Processing

The use of administrative files at Statistics Canada has greatly increased in recent years. Administrative data can have several uses - for example, it can be used for frame creation, sampling, imputation, estimation, and analysis. However, these data can often lack common or unique identifiers and these identifiers can sometimes be inaccurate. As well, administrative data files can contain information recorded in non-standardized formats or with typographical errors. In addition, administrative datasets are occasionally very large, making evaluation of every potential pairs often impossible when conducting record linkage. Therefore, additional pre-processing of administrative data files may need to be conducted.

3.2 Business Names and Addresses Characteristics

Business names and business addresses are often available when conducting record linkage. Extensive work has been conducted at Statistics Canada on household surveys to establish matching rules for names and addresses of individuals and households. However, these rules may not to be appropriate for businesses. Business names can be less distinct than names of people and their lengths can vary more than names of people (for example, a business name can be an acronym, a single word, or contain ten words or more). Business names also tend to contain common words such as “farm” or “holdings”. Addresses pose many challenges as well - for example, the same business can have various addresses that can lead to several matches or many business can have the same address which can lead false matches (for example, businesses located in the same office tower). As is the case for business names, addresses included on the BR are currently not standardized.

3.3 Standardization Strategy

Name standardization (parsing) is a crucial process to make variables more comparable. Different groups at Statistics Canada have their own strategies for standardizing business names and business addresses, each of which have different objectives and quality. Some basic rules for standardization are to convert letters to upper case, to put words in alphabetical order, to remove French accents, to drop trivial words (for example, to drop “company” and “limited”), to convert to standardize spelling (for example, province names could be standardized to use two character representation). More complex rules for standardizing business names and business addresses are required than these. However, it is necessary to evaluate how aggressive the standardization should be as this can influence the resulting links identified.

3.4 Identification of True Matches

The final step when trying to link records together is to use a methodology that will create pairs using either a deterministic[2] or probabilistic[3] approach. At Statistics Canada, a software called MixMatch (Lachance, 2014) conducts deterministic record linkage and a generalized system called G-Link conducts probabilistic record linkage. For both approaches, the user needs to carefully consider which linkage strategy should be used. For example, when using MixMatch, the user must determine which string comparators and parameters should be used. When using G-Link, the user must determine linkage weights that are used for each linkage field in order to decide if two individuals match.

[2] *Deterministic record linkage* produces links based on one or many identifiers that match among the available data sets.

[3] *Probabilistic record linkage* takes a wider range of potential identifiers, computing a weight for each identifier based on its ability to correctly identify a match or a non-match, and then using these weights to calculate the probability that two given records refer to a matching entity.

[4] *Blocking*, like stratification, serves to functionally subset a large dataset into a smaller dataset of individuals with at least one common characteristic.

For both tools, the user needs to consider the use of blocking[4] to reduce processing time. Finally, users will have to create a methodology to assign a quality assessment score to the pairs identified. Sometimes, a manual revision to evaluate potential links is required. This last step can be time consuming and requires subject-matter expertise.

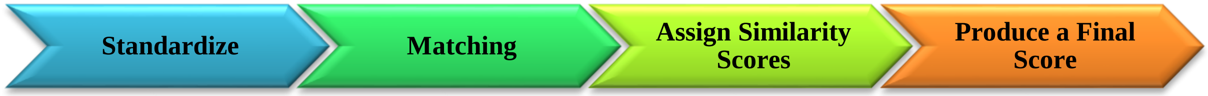
4. Proposed Business Record Linkage Methodology

Using ideas from business record linkage projects conducted at Statistics Canada, the Business Survey Methods Division has proposed a business record linkage methodology. In theory, probabilistic record linkage is the appropriate approach for linking businesses. In practice, this method is often computationally challenging for large data sets (both in terms of observations and variables), such as the BR. A Cartesian product, which is needed when taking a probabilistic approach, simply cannot be handled by today's computers. In contrast, the deterministic record linkage approach can be more feasible and fit the user's needs.

The proposed methodology presented in this paper has the ability to generate reproducible links, and considers the record linkage variables independently. Many deterministic record linkage techniques use a hierarchical series of matching criteria, stopping when a link has been identified – for example, criterion 1: units must match based on business name and business address; if not criterion 2: units must match based on business name; if not criterion 3: units must match based on business address.

When the new proposed methodology is applied independently to (a) business names, (b) business addresses and (c) administrative data, similar but independent steps are conducted (Figure 4-1).

Figure 4-1
Proposed record linkage methodology steps



4.1 Standardization

Standardization of business names is conducted using a Statistics Canada generalized system called G-Code. A parsing strategy is provided to G-Code, created by combining multiple strategies used by different programs at Statistics Canada. Standardization of addresses is conducted using a generalized system called PCODE. PCODE is another Statistics Canada generalized system that takes a Canadian address, standardizes it according to Canada Post Corporation guidelines and generates a postal code for the address. Pre-processing of administrative data at Statistics Canada is conducted by the Administrative Data Division (ADD). ADD conducts editing, imputation, outlier detection, and partnership identification after receiving data from the CRA.

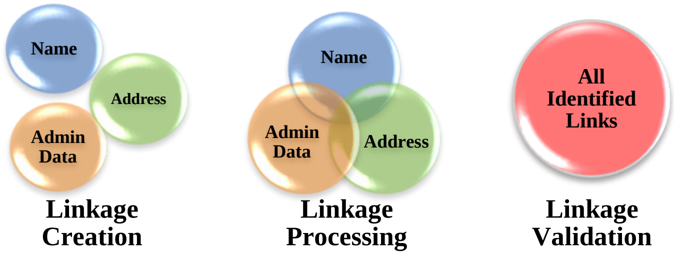
4.2 Matching

Business record links are identified by performing record linkage on a business name, address, a combination of both or any other appropriate variable. For the purpose of this paper, the authors link records using business name, business address and administrative data as three independent processes. First, to identify links using business names, MixMatch is used to perform a deterministic record linkage (Lachance, 2014). Secondly, geographical proximity links are created using business addresses. Lastly, the links using administrative data are identified with a difference measure between common variables. In the example of partnership detection, records that have many revenue fields in common are more likely to be in a partnership than records that only have one field in common.

4.3 Assign Similarity Scores

Following the completion of each matching step, the three sets of links (i.e., business name links, business address links and administrative data links) are merged (see Figure 4.3-1). Three similarity scores are assigned to each link, one for each of the three variables used to identify links.

Figure 4.3-1
Combination of business name links, address links and administrative data links



4.3.1 Business Name Score

For identifying duplicates, a business name score is derived from the Generalized Edit Distance (GED) score that summarizes the degree of differences between two text strings. GED is a generalization of Levenshtein edit distance, which is a measure of dissimilarity between two strings. This edit distance measures the number of deletions, insertions, or replacements of single characters that are required to transform string *a* into another string *b*. For example, the string “Balloons” compared to the string “Baollon” would receive a penalty, translating to a higher score (S_{NAME}), because of the swap of the first “o” and the missing “s”. Mathematically, the Levenshtein distance between two strings *a* and *b* of length $|a|$ and $|b|$, respectively, can be calculated as follows:

$$S_{NAME} = lev_{a,b}(|a|, |b|)$$

where

$$lev_{a,b}(i,j) = \begin{Bmatrix} \max(i,j) & if \min(i,j) = 0 \\ min\begin{Bmatrix} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{Bmatrix} & otherwise \end{Bmatrix}$$

where $i = 1, \dots, |a|$ and $j = 1, \dots, |b|$, and $1_{a_i \neq b_j}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise, and $lev_{a,b}(i,j)$ is the distance between the first *i* characters of *a* and the first *j* characters of *b*. Note that the first element of the *min* argument corresponds to a deletion, the second to an insertion and the third element represents a match or mismatch. Therefore, a lower score means that the two strings are more similar. In this example, “Balloons” and “Baollon” would get a GED score (S_{NAME}) of 120 (20 for the “o” swap and 100 for the “s” insertion).

4.3.2 Business Address Score

For business address linkage, a distance measure can be based on the physical distance between the two linked records. This can be done by evaluating each record’s Global Positioning System (GPS) coordinates. The formula to measure the distance ($S_{ADDRESS}$) between two GPS coordinates (L1 and L2) is the following:

$$S_{ADDRESS} = R * C = 6{,}371 * C$$

where *R* is the earth’s radius (6,371 kilometers in average), *C* is a function of the latitude (ϕ) and longitude (λ) of L1 and L2:

$$C = 2 * atan2(\sqrt{a}, \sqrt{1-a})$$

a is defined as

$$a = \sin^2\left(\frac{L_{1,\phi} - L_{2,\phi}}{2}\right) + cos(L_{1,\phi}) * cos(L_{2,\phi}) * \sin^2\left(\frac{L_{1,\lambda} - L_{2,\lambda}}{2}\right)$$

For example, if two addresses are located in the same building, the difference between their latitude and longitude would equal 0, making the parameters *a* and *C* equal 0. Therefore, the distance score ($S_{ADDRESS}$) between the two addresses would be 0.

4.3.3 Administrative Data Score

To assign a score to links identified using administrative data, in this example financial data, a distance measure is assigned to each link that considers the size of the tax data values and the number of fields (represented as *i* in the equation below) in common between the two units (x_1 and x_2) :

$$S_{ADMIN} = \frac{1}{Log\left(\sum_i (x_{1,i} + x_{2,i})\right)} * \log(1 + \sum_i \left(x_{1,i} - x_{2,i}\right)^2)$$

For example, a husband and wife declaring the same financial declaration would obtain a difference score (S_{ADMIN}) of 0 because the sum of $\left(x_{1,i} - x_{2,i}\right)$ would be null.

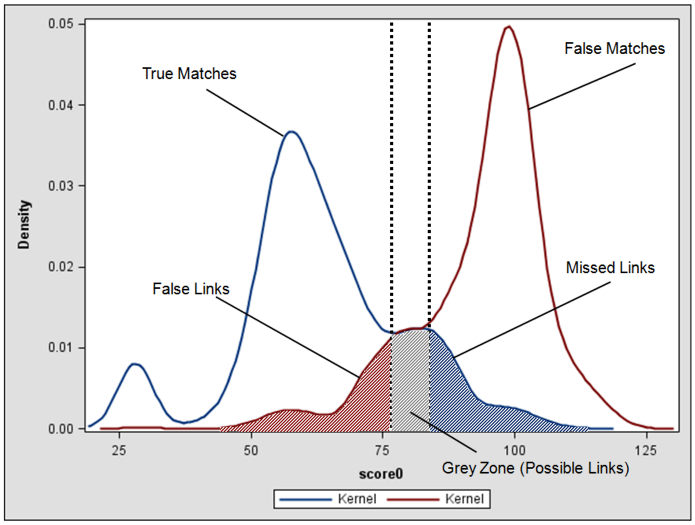
4.4 Produce a Final Score

Each link is assigned a global (or final) score based on the three independent scores (name, address and administrative) to distinguish the stronger links from the weaker ones. For example, a link with a low score for business name, business address and financial data (administrative data) is more likely to be a true match than a link with a large score. A global score can then be derived on the three scores presented in Section 4.3 and calculated as:

$$\mathrm{G_L} = \mathrm{W}_1 S_{NAME,L} + \mathrm{W}_2 S_{ADDRESS,L} + \mathrm{W}_3 S_{ADMIN.L}$$

where S_{NAME} is the business name similarity score, $S_{ADDRESS}$ is the business address similarity and S_{ADMIN} is the administrative similarity data score. Each of these scores (name, address and administrative data) needs to be weighted with values ($\mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3$) to contribute in the same magnitude to the final global score, $\mathrm{G_L}$, where *L* represents a link. The links with the lowest global scores should be automatically accepted (under a pre-defined threshold), or prioritized (sorted) and analysed by the subject matter analyst and rejected if necessary. Methodologists, subject-matter experts, or a computerized system can determine the boundaries/threshold that should be applied in order to distinguish "true matches" from "false matches". However, the use of a grey zone in between the upper threshold for identifying "false matches" and the lower threshold for identifying "true matches" should be used to identify cases that are possible "true matches" but that may also be "false matches". The cases in this grey zone should then be reviewed by subject matter experts to determine if they are correctly identified as "true matches" or "false matches". As shown in Figure 4.4-1, the methodology provides the user with the option to change the threshold to either (a) obtain more "true matches" (by consequence more "false matches"), (b) obtain less "true matches" (by consequence less "false matches").

Figure 4.4-1
Score for true matches and false matches



5. Future Work

The creation of a generalized business data record linkage strategy is very much in development. Progress will be made in consultation with Statistics Canada's technical and steering committees as well as with groups currently conducting business data record linkage. In the short term, work will be conducted to add standardized business names and standardized addresses to the BR to assist users in their linkages. Additionally, the processes used to standardize business names and business addresses will need to be made available to users in order to standardize datasets other than the BR. Work will continue on establishing a generalized strategy for conducting record linkage and for assigning scores to identify links. Additionally, a strategy will be developed to evaluate linkage quality. The generalized strategy will be used, for instance, (1) to create a more complete business continuity (predecessor/successor) table of enterprises for longitudinal analyses; (2) to improve BR's partnership identification process; and (3) to assist Statistics Canada's projects currently conducting linkage of business records. The ultimate objective is to offer business analysts a template that allows the identification of links with an associated quality indicator.

Acknowledgements

We would like to thank the following contributors who helped make this project possible: Sylvie Auger, Susie Fortier, Kimberly Fyfe, Josh Gutoskie, Paul Hunsberger, Leon Jang, Rob Kozak, Martin Lachance, Pierre Lavallée, Danielle Lebrasseur, William Liu, Chris Mohl, Martin Montreuil, Sylvain Poirier, Alexander Reicker, Anne-Marie Rollin, Abdelnasser Saïdi, Ling Su, Anthony Yeung, and Wesley Yung.

References

Auger, S. (2015), "Exporter Register", unpublished document, Ottawa, Canada: Statistics Canada.

Byrd, C. (2016), "Exporter Register Process Overview", unpublished document, Ottawa, Canada: Statistics Canada.

Dongmo Jiongo, V., Émond, N. and J. Lynch, "The Migration of Agricultural Surveys to Statistics Canada's Business Register", *Proceedings of Statistics Canada Symposium 2013*, pp. 294 – 298.

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

Gutoskie, J. (2016), "Use of G-CODE and MixMatch for 2016 CEAG Frame", unpublished document, Ottawa, Canada: Statistics Canada.

Lachance, M. (2014), "Useful functionalities for record linkage", *Proceedings of Statistics Canada Symposium 2014*.

Mayda, M. (2015). "Address Matching by the Address Register Team", unpublished document, Ottawa, Canada: Statistics Canada.

Oyarzun, J., Su, L. and D. Lebrasseur (2015). "An Overview of Record Linkage Applications in BSMD", Presented at Statistics Canada's Business Survey Methods Division Technical Committee, February 27, 2015.

Oyarzun, J. and L. Wile, (2016), "Business Register: Agriculture Project", internal document, Ottawa, Canada: Statistics Canada.

Rollin, A.-M. (2013), "Developing a Longitudinal Structure for the National Accounts Longitudinal Microdata File (NALMF)", *Proceedings of Statistics Canada Symposium 2013*, pp. 306-311.

Saïdi, A. (2014). “Overview of Record Linkage at Statistics Canada”, Technical Reported Presented at Statistics Canada’s Advisory Committee on Statistical Methods, May 5-6, 2014.

Statistics Canada (2014). “User Guide for G-Link”, Ottawa: Canada: Statistics Canada.