Continuity Measures for Statistics Canada's Longitudinal Business Database

Mike Sirois[1], Alain Therrien[2], Costa Glikofridis[2] and Javier Oyarzun[1]
[1]Statistics Canada, Statistical Integration Methods Division, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6
[2]Statistics Canada, Data Integration Infrastructure Division, Jean-Talon Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

Abstract
The Business Register (BR) at Statistics Canada provides monthly snapshots of the Canadian economy with a list of active businesses and auxiliary data from various administrative sources. Economic and classification variables are used in survey programs to create frames and collect relevant information from respondents to fulfill business statistical programs requirements. Analysts are responsible for understanding the changes in their population from one period to the next by tracking specific business transformation events like mergers, acquisitions and splits. The Longitudinal Business Database (LBD) currently being developed will reduce this burden by providing the users of the BR with more information on businesses longitudinal changes. The LBD will offer tables of events with effective dates, longitudinal identifiers with graphical tools to visualize continuity and other useful data which will be used to increase the amount of business demography statistics produced. This presentation will focus on the continuity rules like labour tracking and metadata comparison as well as the combined score function used to establish links between businesses over time.

Keywords: Longitudinal Business Database, Business Demography Statistics, Labour Tracking, Score Function;

1. Introduction

The BR at Statistics Canada is an inventory of all businesses operating in Canada. It provides monthly snapshots of the Canadian economy to assist the researchers in producing business related statistics. Nearly two hundred business surveys use this information in various ways to support their activities, mainly for establishing a survey frame, sampling, collecting and processing data, and producing estimates. The BR has a direct impact on the efficiency of the business survey process, the reliability of data produced by business statistics programs and the coherence of the national accounting system. As of 2020, the BR contains around 7 million active businesses with 120 variables ranging from business name to size of business (revenue, employment, asset and more), geography (street address, geolocation) and industrial classification. The BR includes the following legal types of business: incorporated businesses (public and private), sole proprietorships, governments (municipal, provincial, and federal), trusts and special funds, and partnerships. Information about charities and non-profit organizations is also available on the BR.

The BR is constantly evolving. Current projects include adding more administrative data from various Government of Canada departments, integrating other registers such as the Building and Address Registers, and providing longitudinal information for the users. The later project is called the *Longitudinal Business Database* (LBD) and is the subject of this paper.

2. LBD

The LBD is an extensive project over many years to develop a platform which will provide continuous and historically-updated frame records that will add a longitudinal component to the BR. All businesses operating in Canada will have a Statistics Canada *Longitudinal Identifier* on top of their regular Statistical Identifier created on the BR. It will also provide users with business events (Birth, Death, Merger and more) and effective date changes for characteristics covered. These dates will show when change occurs like a new address or a new main industrial activity. The LBD will then allow to:

- Support the compilation of historically accurate statistics and data from non-survey sources including business counts
- Foster the development of new and innovative business demographic statistics
- Support researchers doing record linkage
- Better understand changes in the Canadian economy over time through continuity studies

2.1 Continuity

The Canadian economy is dynamic. Thousands of new businesses appear on the market each months and others become inactive due to bankruptcy, retirement of owner and other reasons. This constant movement has an impact on the quality of survey frames as well as total business counts and other business demographic statistics produced. Survey analysts are responsible in determining which businesses are truly new and which cases show a sign of continuity, i.e. a business replacing a previous one. They spend a great deal of time analyzing their frames, reading newspaper and searching the Web to detect events like split-offs, mergers and acquisitions which would explain changes to their frame over time. This helps in preventing overcoverage and undercoverage of their population.

One example of continuity would be a restaurant changing ownership and operating name but that would still operate as a restaurant – at the same location – and with many employees in common. They would be two distinct businesses legally and the BR might publish administrative data for both at the same time until the first business becomes officially inactive. Establishing a continuity link between these two businesses would prevent double-counting and overestimation in the statistics. This situation will be observed in the LBD by seeing two different business identifiers for the two businesses associated to the same longitudinal identifier.

The challenge with continuity is to develop a methodology establishing links between these millions of businesses in a fully automated way with low risk of creating false links. First, there is a need to identify which units are new (entrants) and which ones are leaving the economy (exits). These lists are the base for matching businesses. Then different variables are compared in order to confirm the relationships between the in-coming and out-going businesses. These comparisons are called continuity measures and are described in detail next. Continuity links have been produced on an historical yearly basis for this study. They will eventually be produced on a current monthly basis.

3. Continuity measures

Continuity is determined by comparing different attributes between businesses in order to measure their similarities. These quantified comparisons or continuity measures produce links, sometimes with a quality indicator based on the level of similarities. The results are later used in a score function that establishes the final set of continuity links for a given pair of consecutive years (reference years). The five continuity measures retained so far for the LBD are described in this section.

In theory, the attributes to be compared should be the production factors of a business. A production factor is any good or service used to produce output such as employment, machines and equipment, land, buildings and more (UNECE, 2019). In practice, administrative data available on the BR and the LBD are used as proxies. Reference years 2013 up to 2018 have been evaluated for this study.

3.1 Labour Tracking

Businesses in Canada are required to provide the Canada Revenue Agency with tax forms about their employees and other employment related data. Researchers at Statistics Canada can access a table with over 600 million records showing the history of all businesses with employment data by year. Businesses can be connected throughout time based on the estimated number of employees in common between them in a process referred to as *Labour Tracking*. There are a few business longitudinal studies involving labour tracking in the literature. At Statistics Canada, the *Longitudinal Employment Analysis Program* established in the 1980's followed by the *National Accounts Longitudinal Microdata File* (NALMF) as a successor have benefited from a long experience with methodology for finding links between businesses through employment data in order to measure employer dynamics. It served has a strong basis for this project.

There are two main challenges regarding labour tracking. The first is timeliness, as the administrative data is available a few months after the reference years. Also, the data is produced on an annual basis while continuity links will be produced on a monthly basis. There are a few special cases to consider like businesses submitting employment tax forms even after becoming inactive. For the purpose of this project, labour tracking was run over a period of three years to make sure all potential links between businesses are caught, taking into account administrative delays for learning the status of a business. As an example, the years 2014, 2015 and 2016 would be used when measuring continuity for the reference years 2014-2015.

Labour tracking is run four times using different combinations of years to produce potential continuity links. For the previous example where the reference years were 2014-2015, labour tracking would be run on the pairs 2014-2015, 2014-2016, 2015-2016 as well as 2015 matched to itself as continuity might happen within a year.

Entrants are businesses providing employment data only for the second year while *Exits* are those with data only in the first year. *Incumbents* are businesses with data for both years. Those definitions are based on employment tax reports. However, the BR definition for entrants and exits (see section *3.2 Address comparison*) is used for continuity. This is one reason why incumbents are also used in the labour tracking process.

All businesses are confronted and pairs are kept only if the ratio of employees in common over the total number of employees is greater than a pre-specified threshold for both

businesses. These thresholds (see *Table 1*) based on the business size were derived from NALMF (Rollin, 2013):

Table 1: Labour Tracking Link Acceptance Thresholds

	Size of the business (number of employees)					
	250 or more	**51 to 249**	**16 to 50**	**8 to 15**	**6 to 7**	**5**
Proportion of shared employees required	25%	30%	50%	50% if target is an *Exit\Entrant* 60% if target is an *Incumbent*	70%	100%
Examples	75/300	18/60	10/20	5/10 *Exit\Entrant*; 6/10 *Incumbent*	5/6	5/5

The links kept from the four runs are merged together. The final step retains links with businesses appearing under only one link. The cases of businesses involved with multiple links require further study and this will be considered in a future phase of the study. A pair of reference years will detect on average around 13,000 links based on labour tracking, 3,000 when excluding incumbents.

3.2 Address comparison

The BR produces a list of active businesses every month. New businesses appearing on the BR (entrants) for a given month of the second reference year are matched to all businesses which stopped appearing in the monthly snapshot of the BR (exits). For the exits, a period going from twelve months before and after the entrants' month, thus a window of 24 months, are considered to cover for administrative data treatment delays. Two methods were selected to measure the similarities between addresses.

The first uses the street address variable made of the civic number and the street name. Such addresses are spelled in many ways like *555 Park Ave W* and *555 Park Avenue West* could designate the same physical location. The BR includes a standardized version of the address which solves this potential comparison issue. The streets are compared by computing a generic edit distance between the two strings based on the type of differences. A missing character in one string will add more to the distance than a character swap for example. Pair of businesses with a distance greater than zero were eventually rejected as this could have a significant impact for this continuity measure, for example *19 Dirt Road* could be kilometers away from *91 Dirt Road* in a rural area. Thus, the value *Strong* is given to the address measure as a quality indicator when the distance calculated is null which means the street address is identical between the two businesses.

The second method is based on geolocation with longitude and latitude coordinates. A new distance is calculated by measuring the physical coordinates' gap. A set of coordinates might include dozens of businesses in a dense city area. The geolocation cannot be trusted as much as the street address for that reason. Thus, the value *Weak* is given to the address measure quality indicator when the geolocation distance is null. The *Strong* indicator is

kept when a pair of businesses has both *Strong* and *Weak* attributes, i.e. both street and geolocation are the same for the two businesses.

The sole address continuity measure is not sufficient to validate continuity between two businesses as many special cases may arise. One example would be a restaurant and a clothing store in a shopping mall which would have the same street value but are certainly not linked based on their production factors. Other avenues involving addresses have been explored but the real challenge here is the heavy computer processing. Businesses were compared within the same province and postal code area in order to get results within hours instead of days. This could have led to a few links being missed.

Around 145,000 strong links are found for any given pair of reference years and one million weak links.

3.3 Ownership

The ownership continuity measure consists in matching the owners of the businesses between the same pool of exits and entrants as defined previously with the address measure. All exits and entrants are matched together as potential link. For each pair, it is a straight comparison of the owner's identifiers so it must be a perfect match in order to be flagged as an ownership link. This is simple when there is only one owner for each of the two businesses.

The BR is made of many informative tables from various sources of administrative data. Tables with complete list of owners with their share information as provided by the Canada Revenue Agency which collect tax forms are accessible. There is a lot of pre-processing involved with complex economic rules to calculate the percentage share of an owner and special cases like partnerships can be a challenge. However, the rule for continuity is simple once the input file is ready. The total shares from owners in common between the two businesses must cover at least 50% of all shares for each business from the pair so the link is kept as potential continuity for the rest of the continuity process.

A pair of reference years will detect on average approximately 15,000 ownership links for incorporated businesses and about 115,000 for non-incorporated ones.

3.4 NAICS

The North American Industry Classification System (NAICS) has been developed by the statistical agencies of Canada, Mexico and the United States. The six-digit NAICS code on the BR defines the main economic activity of a business. There are over 900 different NAICS codes and some are specific to a country like the 5 cannabis industries in Canada (Statistics Canada, 2020). The more NAICS digits used to define an industrial group, the more precise is the production factor of a business (see *Table 2*).

Table 2: NAICS Hierarchical Example

Breakdown of an industrial group (NAICS code)	
11	Agriculture, forestry, fishing and hunting
111	Crop production
1111	Oilseed and grain farming
11111	Soybean farming
111110	Soybean farming

NAICS are compared for all links found in one of the previously described continuity measures: labour tracking, address and ownership. The quality indicator for the NAICS continuity measure is derived from the number of matching NAICS digits between the two businesses, starting from the left side of the NAICS code. It is *Strong* if the number of matching digits is 5 or 6 and it is *Weak* if it is 3 or 4. Otherwise, the NAICS code is not considered as a factor for further validation of the link as continuity.

3.5 Phone numbers

Two businesses sharing the same phone numbers show a potential relationship for continuity. Businesses will sometimes provide the same phone number to Government's administration for many of their operating entities or franchises. They have one person, like an accountant or someone at the head office, in charge of filling forms or respond to survey questionnaires. Phone numbers cannot be used as the sole indicator of continuity for that reason as they are not necessarily linked to the physical location of a business.

Phone numbers are available from various administrative sources on the BR. The *Legal* and *Operating* phone numbers have been compared for businesses in the links found with labour tracking, address and ownership continuity measures. A match is defined when at least one phone number (legal or operating) is identical between the two businesses.

4. Score function

Each continuity measure, presented in section 3, produces a set of potential continuity links between businesses for two contiguous years. The quality indicator for each of these links or the linkage success varies by type of measure as explained previously. The address comparison has produced strong and weak links for example, and the phone number measure simply produced a successful or failed match. The links obtained from all continuity measures are merged into one file so that all quality indicators for a given link are found on one row. Linkage is a major challenge as using various administrative sources led to multiple variation of business identifiers. The next step is a score function which converts all these quality indicators into numbers which sum gives a global score for each link. The score is used to assess the presence or absence of continuity.

4.1 Continuity measure scores

For most continuity measures, the score is '1' if the measure's quality indicator is impactful and '0' otherwise. Labour tracking is a much more reliable measure as determined by the NALMF experience and the analysis from this study so the score is set at '2' when there is a match. The *Table 3* summarizes the calculation of each continuity measure score:

Table 3: Individual Score Rules by Continuity Measure

Continuity measure	Quality indicator	Individual score
Labour tracking	Link	2
	No link	0
Address	Strong	1
	Weak or No link	0
Ownership	Link	1
	No Link	0
NAICS	Strong or Weak	1
	3 or less digits matching	0
Phone	At least one identical phone number	1
	Any digit difference for both phone numbers	0

Concatenating all five individual scores as one vector variable (e.g. 21001 or 01110) proved to be useful for research purposes. Producing frequencies of this vector was used to detect special patterns and frequent combinations in continuity measure success. Some of these vector values led to manual investigation which later allowed to better understand the links between the businesses and further improve the decision tree (see *Figure 1*).

4.2 Global score

The score of each continuity measure is summed up for each link to form a global combined score. Scores range from 0 to 6. A score of zero happens when all individual scores are equal to zero, meaning all quality indicators were weak or null for the link. For example, a link could have been found only through the address comparison which produced a weak indicator and thus, it obtained a score of 0. For the reference years 2014-2015, only three businesses had a perfect score of 6: many common employees, same address, same owner(s), same industry and same phone number.

4.3 Continuity events

A decision tree (see *Figure 1*) is used to confirm if a link is considered as a continuity link or not. First, the links are split into three groups based on their global score. A score of 3 or higher is considered excellent and form the group 1. A score of 2 may require further investigation and is classified in group 2. A score of 0 or 1 is poor and is already rejected as a continuity event.

A manual look at the links brought up an important factor to consider. Links involving businesses not appearing in other links, called *simple* links, tend to have a higher probability of being true. A link is called *complex* when at least one of the two businesses is part of other links. Complex links need to be further studied as they can be special cases like multiple similar businesses at the same address, e.g. restaurants in a food court or a split-off.

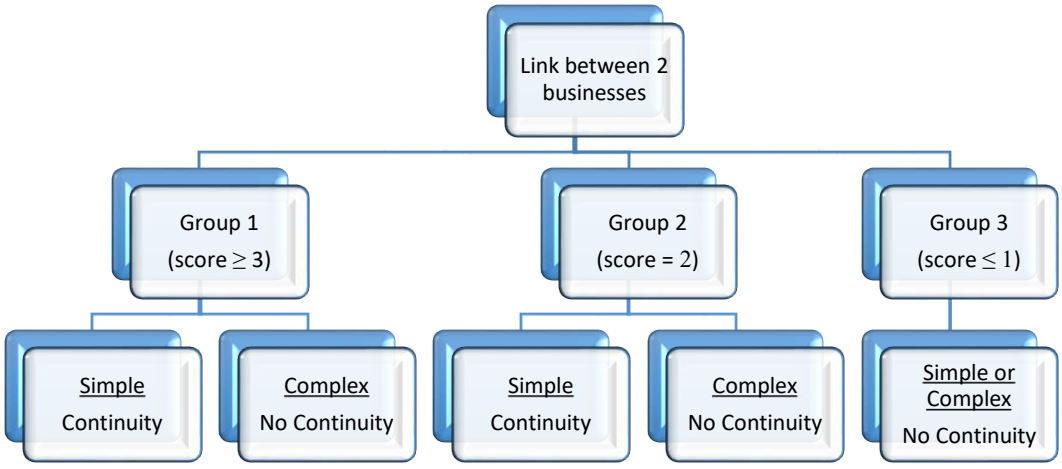

Figure 1: Decision tree for validating continuity based on complexity and global score

Around 120,000 links have a combined score of 2 or higher for the reference years 2014-2015. A total of 44,564 of them are simple links from group 1 and group 2 and are considered continuity links. Further investigation is undergoing for group 2, simple links to potentially identify situations leading to false continuity links, for example links based solely on the labour tracking measure. The goal is to find combinations of continuity measures detecting true continuity links. The *Table 4* shows how each continuity measure contributed to the continuity links for 2014-2015.

Table 4: Continuity Measures Contribution for 2014-2015

	44,565 Continuity links in total	
Continuity measures	**Number of links**	**Percentage %**
Labour tracking	7,696	17.3
Address	7,330	16.4
Ownership	32,291	72.5
NAICS	39,136	87.8
Phone number	4,295	9.6

The contribution of labour tracking for the number of continuity links is similar to the results from the NALMF program at Statistics Canada. The owners of a business and the industrial groups are the top contributing production factors for determining continuity. More analysis is required at this point for better understanding the role of each of these continuity measures in this process.

5. Next steps

The continuity studies for the LBD is still ongoing. The main goal is to provide continuity links and longitudinal identifier relationships to data user in a timely manner using an automated method with the lowest risks possible of being false. A closer look at the complex links is required to see if it is possible to identify any pattern like an economic event (split-off, merger), a specific type of businesses (partnerships, consolidations) or any

pattern in the profile of these businesses which could lead to additional branches in the decision tree (see *Figure 1*) and more positive continuity events.

The use of additional measures like comparing the phone numbers in this study could help to take a better decision for continuity in the presence of null, incoherent or misclassified values for some of the other measures. Continuity measures being scarce in the literature, it is important to keep exploring multiple sources of administrative data and consult with various groups including economists, survey analysts, BR specialists and methodologists. Improving record linkage methods will also help in measuring as accurately as possible the distance between two business attributes.

It would be interesting to develop new coding tools using machine learning techniques for example, to impute data in some administrative sources. Finally, the first iteration of the continuity rules has been established at the legal entity level. Looking at a lower level of complex businesses like the location could help discover more relationships between businesses.

The methodology has been applied for years 2013 to 2018 so far and the exploration will go as far back as the year 2000 in order to have access to a maximum of historical continuity links as possible.

6. Conclusion

The LBD development is a massive project spread over many years and involving multiple groups including analysts, computer specialists and methodologists. The international recommendations, like those provided by the (UNECE, 2019) serve as a foundation for many key components of the LBD. The LBD expands on continuity recommendations with additional measures for validation. Continuity data will allow data analysts to better understand movement amongst businesses in the Canadian economy. Continuity links will help producing more accurate business demography statistics, in particular the counts of new businesses (births) and those leaving the economy (deaths). Longitudinal identifiers, tables of events (including continuity) and retroactive effective dates are some of the components that will assists the users of the BR in understanding changes in their respective populations. The continuity measures might be extended to include more comparison variables (production factors) until the release of the LBD in a near future. The score function and decision tree used to determine continuity between two businesses will evolve as new sources of administrative data are explored, machine learning techniques like *fast*Text for comparing names are integrated, and the labour tracking measure is expanded to include more links than the 1-1 matches. The continuity study also proved to be useful in the identification of duplicates and this aspect will also be explored. As a final note, the continuity study has been a fruitful experience not only for better understanding the longitudinal movement of businesses, but also to discover more potential use of administrative data for record linkage purposes and frame improvement.

Acknowledgements

The authors would like to thank the following contributors who helped make this project possible: Julie Trépanier, Director of the Data Integration Infrastructure Division overlooking the LBD project, Michelle Simard, Director of the Statistical Integration Methods Division at Statistics Canada.

References

Australian Bureau of Statistics. 2020. Microdata: Business Longitudinal Analysis Data Environment, BLADE. https://www.abs.gov.au/ausstats/abs@.nsf/mf/8178.0 (accessed June 7, 2020).

Baldwin, J.R., R. Dupuy, and W. Penner. 1992. Development of longitudinal panel data from business registers: the Canadian Experience. In *Statistical Journal of the United Nations,* 9, 289-303.

Canada Revenue Agency. 2017. The Business Number and Your Canada Revenue Agency Program Accounts. In *Pamphlet RC2*. http://www.farrow.com/documents/rc2-12e.pdf (accessed June 1, 2020).

European Commission. 2010. Continuity rules for the enterprise. In *Business Register Recommendations Manual*. Luxembourg. 128-134.

Geurts, K., M. Ramioul, and P. Vets. 2010. Employee flows to study firm and employment dynamics. In *Proceedings of Statistics Canada Symposium 2010*. Ottawa, Canada. 325-330.

Oyarzun, J. and L. Wile. 2016. An Overview of Business Record Linkage at Statistics Canada: How to Link the Unlinkable. In *Proceedings of Statistics Canada Symposium 2016*. Ottawa, Canada. 1-7.

Robertson, Kenneth, Huff, Larry, Mikkelson, Gordon, Pivetz, Timothy and Winkler, Alice. 1999. Improvements in Record Linkage Processes for the Bureau of Labor Statistics' Business Establishment List. In *Record Linkage Techniques -- 1997: Proceedings of an International Workshop and Exposition*. Washington, DC: The National Academies Press, 1999. doi: 10.17226/6491.

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

Statistics Canada. 2020. North American Industry Classification System (NAICS) Canada 2017 Version 3.0. (last updated 2018, August 29). https://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=1181553 (accessed June 7, 2020).

United Nations Economic Commission for Europe (UNECE). 2019. Definitions and key concepts of business demography statistics In *Guidelines on the use of statistical business registers for business demography and entrepreneurship statistics*. 9-29. http://www.unece.org/fileadmin/DAM/stats/publications/2018/ECECESSTAT20185.pdf (accessed June 7, 2020).