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

# Comparing Survey Data to Administrative Sources: Immigration, Labour, and Demographic data from the Longitudinal and International Study of Adults

James Hemeon, Meghan Fulford[1]

## Abstract

Administrative data, depending on its source and original purpose, can be considered a more reliable source of information than survey-collected data. It does not require a respondent to be present and understand question wording, and it is not limited by the respondent's ability to recall events retrospectively. This paper compares selected survey data, such as demographic variables, from the Longitudinal and International Study of Adults (LISA) to various administrative sources for which LISA has linkage agreements in place. The agreement between data sources, and some factors that might affect it, are analyzed for various aspects of the survey.

Key Words: Administrative data, longitudinal, survey, record linkage, data linkage

## 1. Introduction

Government agencies collect and retain data for administrative purposes, such as taxation of the population or its businesses, or regulation of the flow of goods and people across borders. With the growing demand for statistical data, and the response burden and costs associated with survey collection, administrative sources have become increasingly sought after as sources of data to be used for statistical purposes.

The linkage of survey data to administrative data sources is a means to collect data on sensitive topics that a person may not feel comfortable disclosing to an interviewer, information which may be difficult for a person to recall precisely, and/or information that would result in a very lengthy interview if conducted by survey. However, when using administrative data for statistical purposes, the original purpose, scope, variable definitions, and limitations must be understood.

The Longitudinal and International Study of Adults (LISA) is a Canadian longitudinal social survey run by Statistics Canada that collects a wide variety of socioeconomic data. The LISA also contains data from multiple administrative data sources for each year onward from its 2011 inception, as well as historically, with administrative data from as early as 1980, and up to 2013, for some respondents. Due to the variety and amount of information, there are a few variables for which the LISA has both a survey and an administrative source of data. In cases where there are discrepancies between sources, the LISA does not make data corrections under the assumption that one source is correct.

[1] James Hemeon, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6, James.Hemeon@Canada.ca. Meghan Fulford, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6, Meghan.Fulford@Canada.ca.

The present paper will compare three socioeconomic variables that exist both in the LISA survey data and the LISA-linked administrative data: year of immigration (landing), industry of employment, and year of marriage. It will measure the coherence between sources, and seek to explain some of the differences, in order to better understand their strengths and limitations, and how they might complement each other.

## 2. Data source

The LISA includes data on respondent demographics, family and household composition, caregiving, education and training, skills, health, income and wealth, retirement, and labour market participation. The sample was selected from a population frame constructed of dwellings from Canada's ten provinces that responded to the 2011 Census of Population, and excludes regular members of the Canadian Forces, individuals living in institutions, and individuals living on reserves and other Aboriginal settlements.

The LISA currently has two survey databases, known as LISA 2012 and LISA 2014. LISA 2012 contains 23,926 respondents, aged 15 and older. 10,299 (43%) were interviewed in 2011, and 13,627 (57%) in 2012.

The LISA 2014 contains 19,178 respondents, aged 15 and older, including 16,895 people who had been interviewed in LISA 2012, 1,582 people who were part of the LISA 2012 sample and had not responded, and 701 people who joined the household of a LISA 2012 respondent between LISA 2012 and LISA 2014. LISA 2014 interviews took place in 2014.

In total, LISA has at least one survey response for 26,209 respondents aged 15 and older.

## 3. Linkage

The LISA has data linkages to the following administrative databases: the T1 Family File (T1FF), the statement and summary of compensation paid by employers (T4 file), the Business Register (BR), the Pension Plans in Canada (PPIC) file, and the Longitudinal Immigration Database (IMDB).

To link to these databases, the LISA is linked to the T1 personal income tax master files, using deterministic and probabilistic linkage methods, in order to obtain a Social Insurance Number (SIN). In LISA 2012, SIN's were linked using the 2011 and 2010 T1 master files, containing data for all taxfilers in Canada for those respective years. In LISA 2014, SIN's were linked using the 2013 and 2012 T1 master files. The LISA linked to a SIN (and administrative data) for 24,763 (94.5%) of its 26,209 total respondents, including 23,081 (96.5%) of LISA 2012 respondents, and 18,209 (95.0%) of LISA 2014 respondents. The linkage rate of previous waves improves over time, as SIN linkage is attempted for existing and new respondents.

Once a SIN is identified, it is used to link to the T1FF, T4, and IMDB files. The T4 data is then used to link to the BR and the PPIC. T1FF data is available from 1982 to 2013. T4, BR, and PPIC data are available from 2000 to 2013. The IMDB data is a longitudinal database containing data from 1980 to 2012.

The samples used for this paper vary depending on the analysis. LISA 2012 and LISA 2014 are used to compare year of immigration (comparing to the IMDB), LISA 2012 is used to compare industry of employment (comparing to the 2011 and 2012 T4/BR), and LISA 2014 is used to compare year of current marriage (comparing to the 1982-2013 T1FF).

# 4. Results

## 4.1 Year of immigration: LISA vs. IMDB

The IMDB is a database which combines linked immigration data with taxation records (primarily from the T1FF). The database contains those who obtained immigrant status since 1980, and up to 2012 (at the time of the LISA 2014 linkage), and who filed at least one tax return since 1982.

Of the 26,209 respondents in the combined sample of LISA 2012 and LISA 2014, 4,864 (18.6%) reported being a landed immigrant and provided a year of immigration. Of these, a total of 3,276 (67.4%) were successfully linked to the IMDB file. 1,258 (25.9%) respondents were unable to be linked to the IMDB, but had reported a year of immigration prior to 1980 or post-2012, and were therefore out of the scope for the IMDB. The remaining 330 (6.8%) were unable to be linked, and reported a landing year between 1980 and 2012. Of those who reported immigrating between 1980 and 2012, 90.8% were linked to the IMDB.

All LISA respondents were linked to IMDB, regardless of the immigrant status they reported in the survey. The LISA IMDB file contains 3,446 respondents. 3,278 (95.1%) of the linked respondents reported a landing year during the interview. The remaining 170 (4.9%) did not report being a landed immigrant in the survey.

Some reasons why some respondents appear to be a landed immigrant in one data source but not the other could include response errors in survey collection, or processing errors in data linkage. Furthermore, a respondent with a year of immigration prior to 1980 or post-2012, or who has not filed at least one tax return since 1982, would not be on the IMDB.

After considering the differences between data sources, the coherence will now be examined. The majority (77.9%) of people who reported a LISA landing year between 1980 and 2012 and who were linked to the IMDB reported a year that is equal to their IMDB landing year, and 88.9% reported a year within 1 year of their IMDB landing year (see table 4.1.1).

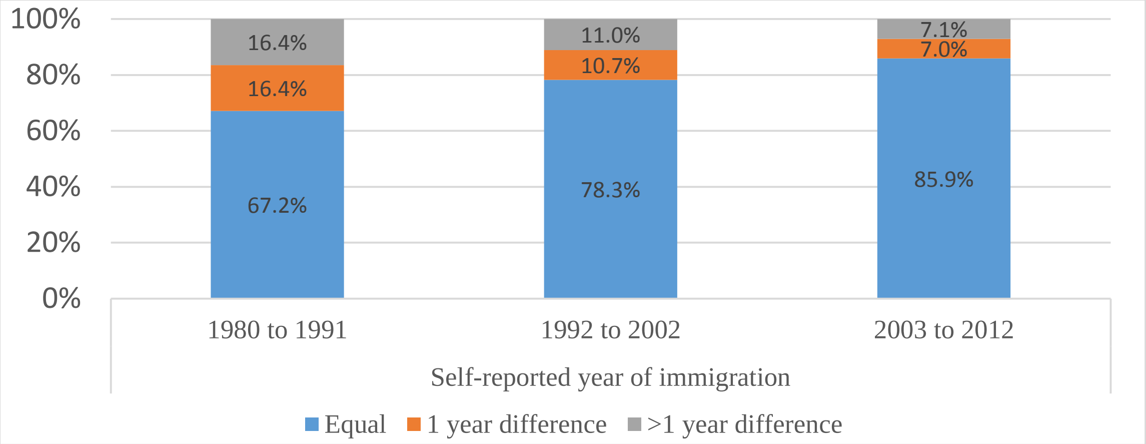
**Table 4.1.1: Coherence between LISA year of immigration and IMDB year of immigration**

| Coherence | % | Cumulative % |
|---|---:|---:|
| Equal | 77.9% | 77.9% |
| 1 year difference | 11.0% | 88.9% |
| 2-5 years difference | 9.6% | 98.4% |
| >5 years difference | 1.6% | 100.0% |

**Source:** Statistics Canada, Longitudinal and International Study of Adults, 2012 & 2014; Statistics Canada, Longitudinal Immigration Database, 1980-2012.

The difference in landing year between sources may be the result of recall bias. While respondents may have difficulty accurately recalling events from long ago, the administrative data does not have this limitation. When comparing landing year from the LISA and the IMDB, a larger coherence is present between sources for those who reported immigrating more recently (see table 4.1.2). Of those who immigrated between 2003 and 2012, 85.9% reported a year that is equal to their IMDB year. Going back in time, the percentage that reported a year equal to their IMDB year decreases. When landing year is between 1980 and 1991, the percentage reporting a year equal to their IMDB year decreases to 67.2%.

**Figure 4.1.2: Comparison between sources: LISA year of immigration and IMDB year of immigration**



**Source:** Statistics Canada, Longitudinal and International Study of Adults, 2012 & 2014; Statistics Canada, Longitudinal Immigration Database, 1980-2012.

This suggests that the administrative source is not subject to recall bias that might exist for respondents who, in their interview, tried to recall an event that occurred as early as 1980.

## 4.2 Industry of employment: LISA 2012 vs. Business Register (2011-2012)

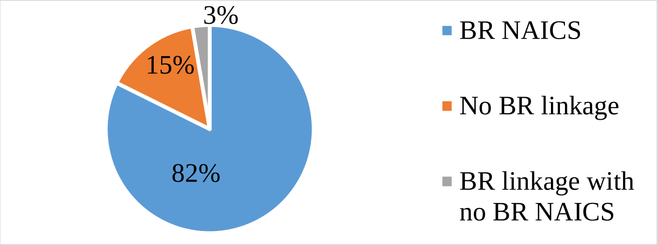
Of the 23,926 LISA 2012 respondents, 19,582 (81.8%) have a North American Industry Classification System (NAICS) code assigned from LISA data. Of these, 15,168 (77.5%) also have a NAICS from the Business Register (BR), linked from their T4 of highest earnings for the tax year corresponding to their LISA 2012 year of interview (2011 or 2012). 504 respondents with a LISA NAICS (2.6%) were linked to a BR observation with a blank NAICS, which can occur if the business associated to the respondent's T4 did not provide a sufficient description for a NAICS to be assigned. 3,936 respondents with a LISA NAICS (20.1%) were not linked to the BR. However, one must consider that, in LISA 2012, the industry of employment was collected from respondents who reported working in the 5 years prior to their interview, and respondents who had not worked during their year of interview are less likely to obtain a T4 for that year.

A total of 4,344 respondents (18.1%) have no NAICS code in LISA. 429 (9.9%) respondents with no LISA NAICS have a BR NAICS, 345 (80.4%) of which indicated not having worked within the 12 months prior to their interview (341 of which not having worked in the 5 years prior to their interview). There is therefore a discrepancy between sources: the respondents indicated not having worked; however, they were linked to a T4.

The remaining 3,915 respondents with no LISA NAICS (90.1%) also have no BR NAICS. Of these, 3,862 (98.6%) indicated not having worked within the 12 months prior to their interview (3,834 of which not having worked in the 5 years prior to their interview). This is consistent, as those who did not work are less likely to receive a T4.

If considering only the 18,142 respondents who reported working at a paid job within the 12 months prior to their interview, 18,005 (99.2%) were assigned a LISA NAICS. Of those with a LISA NAICS, 14,822 (82.3%) also had a BR NAICS, 2,699 (14.9%) had no BR linkage, and 484 (2.7%) were linked to a BR observation with no NAICS (see Figure 4.2.1). Of the 14,906 respondents with a BR NAICS, 14,822 (99.4%) also had a LISA NAICS.

**Figure 4.2.1 Proportion of LISA 2012 respondents who worked in the 12 months prior to their interview, with a LISA NAICS, with/without a BR NAICS**



**Source:** Statistics Canada, Longitudinal and International Study of Adults, 2012; Statistics Canada, Business Register, 2011-2012

The NAICS code derived from LISA survey data is 4-digits, and the NAICS code from the BR is 6-digits, which provides additional detail on industry of employment. To compare the sources, the coherence was measured at the 4-digit level for the 14,822 respondents who had both a LISA and a BR NAICS. When comparing the two sources, 43% had the same NAICS at the 4-digit level, 57% at the 3-digit level, and 64% at the highest (2-digit) level (see Table 4.2.2). This can likely be explained by the differences between data sources and coding: the LISA NAICS is coded at Statistics Canada from information provided by the respondent, using the same process used by other Statistics Canada surveys. The BR NAICS, on the other hand, is assigned by the Business Register Sub-Division at Statistics Canada, using multiple data sources, including the description provided by the business upon registration with the Canada Revenue Agency, and the tax filings of the business (as of 2011). The NAICS is also updated by Business Register users (business profilers and business survey subject matter specialists) at Statistics Canada, as well as from feedback obtained from business surveys. In addition to these differences, a LISA respondent (employee) may have described their employer differently than the business owner's description (provided upon incorporation of the business). Furthermore, businesses sometimes change their main activity (NAICS) over time, and the timing of the linkages may not account for these changes.

Aside from the data sources and coding, there are minor conceptual differences between sources. The LISA NAICS was coded from a respondent's self-description of their current job; however, if a respondent worked at more than one job, the job in which they worked the most hours was used to code a NAICS. If a respondent worked at more than one job but worked the same amount of hours in all jobs, the job with the highest earnings was used. Therefore, some respondents may have potentially had a LISA NAICS coded for a job in which they worked the most hours, but another job was the highest earning. For these cases, comparing LISA to the T4 of highest earnings would not be comparing NAICS codings of the same job.

**Table 4.2.2. Coherence between 4-digit LISA NAICS and BR NAICS**

| **Coherence** | **%** | **Cumulative %** |
|---|---|---|
| 4-digit | 43% | 43% |
| 3-digit | 14% | 57% |
| 2-digit | 7% | 64% |
| No match | 36% | 100% |

**Source:** Statistics Canada, Longitudinal and International Study of Adults, 2012; Statistics Canada, Business Register, 2011-2012

## 4.3 Year of marriage: LISA 2014 vs. T1FF (1982-2013)

For respondents who reported being 'Married' or 'Separated' at the time of their interview, the LISA collects the year in which a respondent's current marriage began. The year of current marriage can also be derived using the marital status variable on the T1FF, from 1982 to 2011. By merging the marital status from each T1FF year into one dataset, a marital status history can be created. For this analysis, those who reported being 'Married' in the LISA were used.

When comparing LISA current marriage data to the processed T1FF data (which will be referred to as 'Raw') for 6,371 LISA 2014 respondents who reported being 'Married' post-1982 in the LISA data, 63.9% had the same marriage year derived from T1FF data. For the 3,484 LISA 2014 respondents who reported being married in 1982 or prior in the LISA data, 60.2% had a marriage year of '1982 or prior' derived from the T1FF data. The T1FF data series begins in 1982; therefore, if a respondent is married as of their 1982 tax filing, they may have been married in 1982 or any year prior.

One consideration is that the T1FF data is being used differently than its intended purpose, with tax data being used to derive marital status. Some limitations of the marital status variable in the T1FF data were noted which suggest that, with some reasonable "clean-up" to the T1FF marriage status history, more coherence could be found with the LISA data. First, there were years in which data was missing, and there were marital status values that were unlikely or impossible. For example, there were one year anomalies where people who were 'Married' became 'Common-law' or 'Separated' for one year (which is possible, but unlikely), became 'Single, never married' (which is impossible) for one year, or had a year of missing T1FF data, then returned to being 'Married'. These values were changed to the value that preceded/followed them (see 'Step 1' in Table 4.3.1 and Table 4.3.2).

There were also respondents with multiple years of missing T1FF data. For example, a respondent's T1FF data series may begin in 1996, where they reported being 'Married' from then onward. Because data prior to that was missing, it cannot be assumed that this was the year of marriage. To allow for a more meaningful comparison to LISA data, we did the following steps. First, the distribution of current marriage age was analyzed using existing LISA and T1FF data, and the mode (age 23) was chosen. The missing data, when preceding a tax filing where the respondent was 'Married', was therefore imputed retrospectively to a respondent age of 23 (see 'Step 2' in Table 4.3.1 and Table 4.3.2).

Lastly, values of 'Common-law', 'Single, never married' were changed to 'Married' if they immediately followed a year of a person being 'Married' (see 'Step 3' in Table 4.3.1 and Table 4.3.2). For example, if a person became 'Single, never married' for several years immediately following a year in which they reported being 'Married', these 'Single, never married' values were changed to 'Married'.

For self-reported marriages post-1982 (6,230 respondents), Steps 1 to 3 increased the coherence between LISA and T1FF year of current marriage from 62.5% ('Raw') having the same year in both sources, to 68.0%, 69.1%, and 69.5%, respectively (see Table 4.3.1). For self-reported marriages in 1982 or prior (3,450 respondents), Steps 1 to 3 increased the ability to predict, using T1FF data, if the marriage occurred in 1982 or prior from 57.3% ('Raw') to 70.4%, 90.4%, and 91.0%, respectively (see Table 4.3.2). 175 respondents who reported being married did not have tax data, and therefore could not be analyzed.

**Table 4.3.1. Coherence between LISA and T1FF-derived marriage year, for those who self-reported a marriage post-1982**

| T1FF | Raw | Step 1 | Step 2 | Step 3 |
|---|---|---|---|---|
| Equal | 62.5% | 68.0% | 69.1% | 69.5% |
| 1 year difference | 8.3% | 8.9% | 9.4% | 9.5% |
| 2 year difference | 3.6% | 3.2% | 4.3% | 4.3% |
| 3-5 year difference | 6.5% | 5.2% | 6.0% | 6.0% |
| 6-10 year difference | 7.5% | 6.0% | 4.5% | 4.3% |
| 10+ year difference | 11.2% | 8.2% | 3.5% | 3.3% |
| T1FF 1982 or prior | 0.4% | 0.5% | 3.2% | 3.2% |

**Source:** Statistics Canada, Longitudinal and International Study of Adults, 2014; Statistics Canada, T1 Family File, 1982-2013

**Table 4.3.2. Coherence between LISA and T1FF-derived marriage year, for those who self-reported a marriage in 1982 or prior**

| T1FF | Raw | Step 1 | Step 2 | Step 3 |
|---|---|---|---|---|
| 1982 or prior | 57.3% | 70.4% | 90.4% | 91.0% |
| T1FF post-1982 | 42.7% | 29.6% | 9.6% | 9.0% |

**Source:** Statistics Canada, Longitudinal and International Study of Adults, 2014; Statistics Canada, T1 Family File, 1982-2013

## 5. Conclusion

It is uncommon for survey programs to have a survey and an administrative source of the same data. The present paper is the first to do such an analysis on LISA data. The analysis in this paper shows that information obtained from administrative records is usually, but not always, the same as what is given by survey respondents.

The LISA has survey data for years which do not exist in the linked administrative sources, and has concepts and definitions that can be controlled by a survey subject matter team. Administrative data, by nature, was not collected for statistical purposes, and careful attention must be paid to ensure that the concepts are well understood before replacing one source with the other.

The year of immigration, when comparing survey to administrative data sources, was the same for many respondents. For this comparison, the concepts were the same, although the years of coverage were limited for the administrative source. For industry of occupation (NAICS), the survey and administrative sources were the same in some cases; however, many had a different NAICS in each source, which may have been due to differences in concepts and coding. For year of current marriage, the survey and administrative sources were quite different, as the administrative data was used to derive a year of marriage from a series of tax data. However, when making some assumptions to fill in missing years of data, as well as to “clean up” impossible or unlikely values, the coherence was greatly improved.

When comparable concepts of immigration year were analyzed between sources, the coherence between survey data and administrative data is greater for more recently occurring events. This suggests that, provided the concepts and quality of the administrative data remain constant over time, administrative data may be a more accurate data source for events that occurred long ago.

By comparing the two sources, their strengths can be leveraged, and the sources may complement each other to provide more complete data than would exist in either source by itself.