[1] Statistics Canada Statistique Canada [2]

SFU CO-OPERATIVE EDUCATION

Barindervir Thind

Co-op Student

Job Title: Methodologist[3]

Supervisor: Shuai Zhang

Major/Degree: B.Sc. Honors - Statistics

Work Term: Winter 2017 - Work Term 1

Business Register & Record Linkage

[1] *Here is the logo of the agency.*

[2] Disclaimer: The contents of this report reflect the views of the author and not necessarily the official views or opinions of Statistics Canada.

[3] I am not a permanent employee of Statistics Canada.

Table of Contents

1 Objective

The Business Register (BR) is a vast and versatile database that takes in information from various sources (such as the administrative data, surveys, profilers, etc.) and centralizes it for the purposes of facilitating the frame creating process for Statistics Canada's survey programs as well as supporting calculation of national statistics. This centralization allows for a multitude of advantages such as consistency throughout estimates within the agency, an ease of cross-validation, and a verified pool of information from which to infer. However, due to the sheer mass of such a data set[4], there are bound to be complications. My research entailed improving the BR's quality by contributing to the record linkage methodology to help identify potential duplicates. In addition, I had the benefit of working with geographical variables such as of latitude and longitude[5].

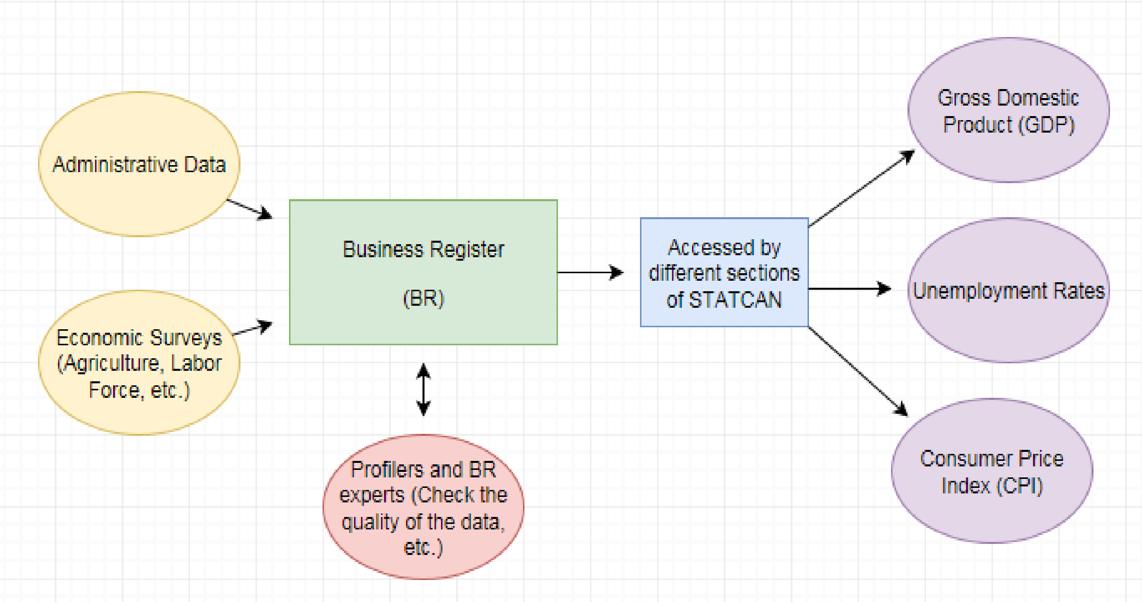

Administrative Data
Economic Surveys (Agriculture, Labor Force, etc.)
Business Register (BR)
Profilers and BR experts (Check the quality of the data, etc.)
Accessed by different sections of STATCAN
Gross Domestic Product (GDP)
Unemployment Rates
Consumer Price Index (CPI)

Figure 1.1: A map of how the BR operates and the various inputs/outputs involved

[4] There are approximately 4 million active businesses and over 100 variables

[5] These variables were first added to the BR in January 2017

2 Preliminaries and Notes

This was my first work term and I couldn't have asked for a better situation. Working at Statistics Canada has been a rewarding, enriching, and enjoyable experience that I would recommend to any student or graduate. My supervisors - Shuai Zhang and Javier Oyarzun - have been helpful, kind, and courteous while also pushing me to come up with better and more concrete solutions for the problems I faced throughout the work term. They are also responsible for a lot of the progress I made in this research.

For this work-end project, I decided to do a technical report. This made the most sense because of the type of work I did and the amount of information I felt I needed to relay so that the reviewer could get an adequate understanding of the work.

Table 2.1 below contains a list of widely used terms and their corresponding definitions. More detailed explanations/definitions are presented in the appropriate sections.

Term	*Definition*
Business Register (BR)	**Mandate**: Collect, compile, and maintain a full repository of businesses in Canada, in order to provide a complete, unduplicated, and up-to-date frame for economic programs **Target Coverage:** All businesses engaged in producing goods and/or services in Canada
Administrative Data	Administrative data refers to data acquired by the Administrative Data Division. This division acquires tax data from the CRA (Canada Revenue Agency) and other data sources that are required for many Statistics Canada programs.
Score	A score is a value assigned to a matched pair which is then compartmentalized and used to determine the validity of that association.
Threshold	A "cutoff" set of values based on the score of a match that allows us to determine whether a linkage association will be considered to be valid or not.

Geocoding	This is the process of assigning geographic identifiers (codes or x, y coordinates) to map features and data records. The resulting geocodes permit data to be linked geographically to a place on Earth.
Rare/Common-Words	A set of words weighted differently when being taken into account with respect to the score function because of how frequently (or not) they are used in business names (e.g. farms in "Smith's farms" for agriculture).
Block-face (BF)	A block-face is one side of a street between two consecutive features intersecting that street. The features can be other streets or boundaries of standard geographic areas.
Dissemination Block (DB)	A dissemination block is an area bounded on all sides by roads and/or boundaries of standard geographic areas. The dissemination block is the smallest geographic area for which population and dwelling counts are disseminated. Dissemination blocks cover all the territory of Canada.
Dissemination Area (DA)	A dissemination area is a small, relatively stable geographic unit composed of one or more adjacent dissemination blocks. It is the smallest standard geographic area for which all census data are disseminated. Das cover all the territory of Canada
North American Industry Classification System (NAICS)	NAICS is a comprehensive system encompassing all economic activities. It has a hierarchical structure. At the highest level, it divides the economy into 20 sectors. At lower levels, it further distinguishes the different economic activities in which businesses are engaged.
Match	An initially produced data frame (with matched observations) from which links are determined.
Link	A match is called a true link if it is indeed the case that the businesses are the same. A possible link is one that we are not sure of and otherwise, the match is a false link.
Street Dissemination Block	Street linked to some number of dissemination

(STDB)	blocks.
Postal Code Dissemination Block (PCDB)	Postal code linked to some number of dissemination blocks.

Table 2.1: Definition table for commonly used words in this report

3 Methodology

Record Linkage is a task of finding records that refer to the same entity across different data sources. A methodology of record linkage is applied here in order to find potential duplicated records within the BR. This section is broken into 5 sub-sections: Initial matches, Rare Words, Score Function (and the applications of Zipf's law in this context), threshold methodology, and some comments on comparing previous methods of finding potential duplicates versus the methods in the report (geodistance vs. non-geodistance). In each sub-section, I highlight, in detail, the reasoning and logic behind my approach and what particular assumptions I made. I also make some short notes on potential shortcomings.

3.1 Initial Matches and Geodistance

Since the dataset is so large, reasonable measures must be taken so that the scope of the linkage being attempted is manageable. I borrowed code written by Javier and modified so that it incorporated the new geographical co-ordinate variables.

In the initial step, I separated out the data by province and whether it was an active business or not[6]. The data was also separated by NAICS2 to allow for an even smoother process when running the code. Finally, a last group of subsets is made which divides the observations on how accurate the geographical information is. That is, they are split on BF, PCDB, STDB, and DA. The reason that I opted to take this approach is because it allows us to make a better determination on how accurate our potential matches are. For example, if we have 2 observation linked on BF, then that is more likely to be a "true" link than if those observations had matched with the accuracy of the geographical information being DA.[7]

[6] This idea was borrowed from Javier Oyarzun (Senior Methodologist, Statistics Canada)

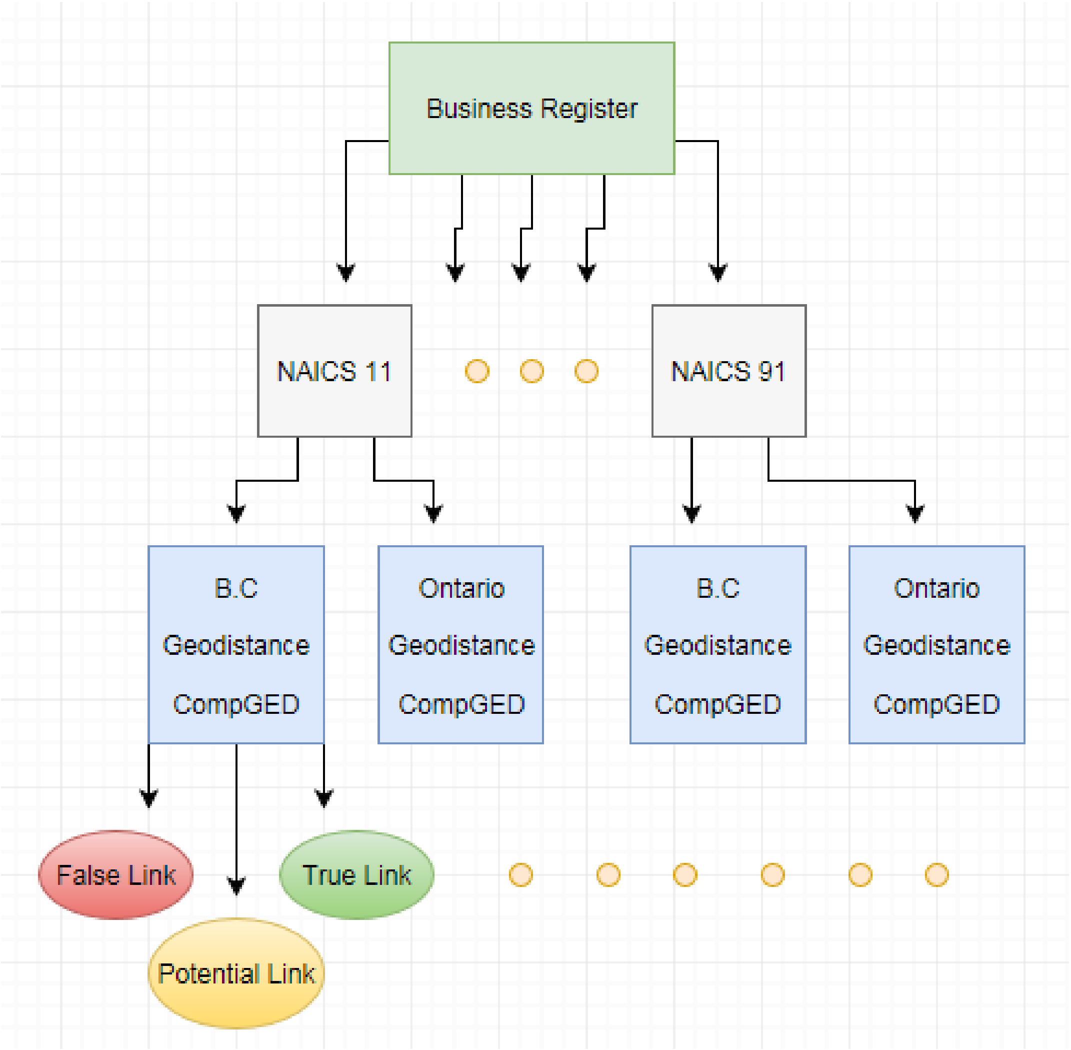


Figure 3.1: An overview of the whole process.

Next, the initial matches were made. Using SAS's *proc sql*, the BR was matched with itself and all observations such that the geographical distance between them was less than 0.1 kilometers, had the same province, and had a compGED[8] of less than 500[9]. In order to move on to the next step, the following definitions are required:

[7] Now, clearly there will be some links between say BF and PCDB which will not be captured if we just matched BF vs. BF. Fortunately, the way I have written the SAS code, it was easily possible to make all these different connections (BF vs. BF, BF vs. PCDB, ..., DA vs. DA)

Definition 3.1.1 The values of α_i take on 1 if the i^{th} word of name A and the i^{th} word of name B are the same. Otherwise, it takes on the value of 0.

Definition 3.1.2 The *initial score* value of a match is defined as the sum of all corresponding words between two names such that they are assigned a value of 1 if they are the same and 0 otherwise. That is:

$$S=\sum_{i=1}^{n} \alpha_i$$

Where S is the initial score, α_i is whether the i^{th} word is the same for both names, and n refers to the number of words examined in each name (for this report, 10 was chosen as that encompasses the length of nearly all the names in this dataset).

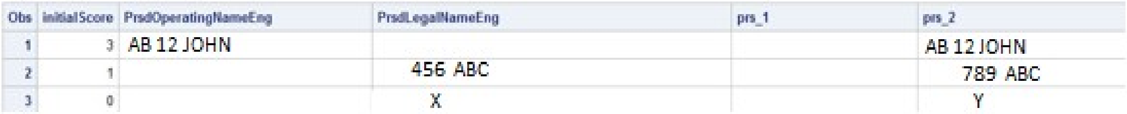
As seemingly an arbitrary choice, I decided to look at the parsed representations of the names (of the businesses)[10]. Then, Definition 3.1.1 was applied to all six combinations of names. In other words, the parsed operating name and the parsed legal name of the first observation was compared with the parsed operating and legal name of the associated observation. This resulted in binary values for every combination of words as defined above. Then, applying Definition 3.1.2, we get

[8] A SAS function which calculates how close to each other two strings are. At a score of 100, you have some very minimal difference such that changing one letter in one of the string would get you to the other string. As this scores increases, the further apart the strings are. The formula used is:

[9] These numbers are adjustable. The score assigned here has an inverse relationship with the number of missed links. That is, a lower threshold based off the compGED could imply a higher amount of missed potential links

[10] The operating and legal names are the names off the businesses which pertain to from where they operate and from where their business is done. The parsed versions of these names is a standardized version such that special characters, etc. are removed.

some integer as an initial score. We now have a raw measure of what is a likely to be a true link and what is not. Figure 3.2 below is an example of what the results would look like.



Obs	initialScore	PrsdOperatingNameEng	PrsdLegalNameEng	prs_1	prs_2
1	3	AB 12 JOHN			AB 12 JOHN
2	1		456 ABC		789 ABC
3	0		X		Y

F

igure 3.2: Initial Score Example [Example Data][11]

3.2 Rare Words

In the previous section, we established a basic score value from which we can determine whether a match is a link or not. However, if we were to stick with that value convention, we are making some problematic assumptions. Consider the following two matches:

1. On A Car - **matched with** - On A Truck
2. Kuzuls - **matched with** - Kuzuls

If we were to use Definition 3.1.2, we would get a value of 2 for the first match and a value of 1 for the second match. But, by examining it manually, we can see that the reason the first match has a higher score is because of words like “on” and “a”. And, even more abhorrent, we see that the second match has a score of 1 based on a word that is hardly ever seen - particularly as the title of a business. Now the implicit assumption becomes obvious: we are claiming that, if we use the above values to make our link decisions, all the words in these names are used in equal frequency. Clearly, this is not true.

In order to deal with such a situation and improve the scoring model, I decided to first, look at how frequently each word comes up and then, looking at those frequencies, I weight the initial scores accordingly[12]. Using SAS’s *proc freq*, the tables in Figure 3.3 were obtained. This seemed the most reasonable path to go down for such a problem as it would, theoretically, take care of the problem with those most commonly used words while also emphasizing the value of the less

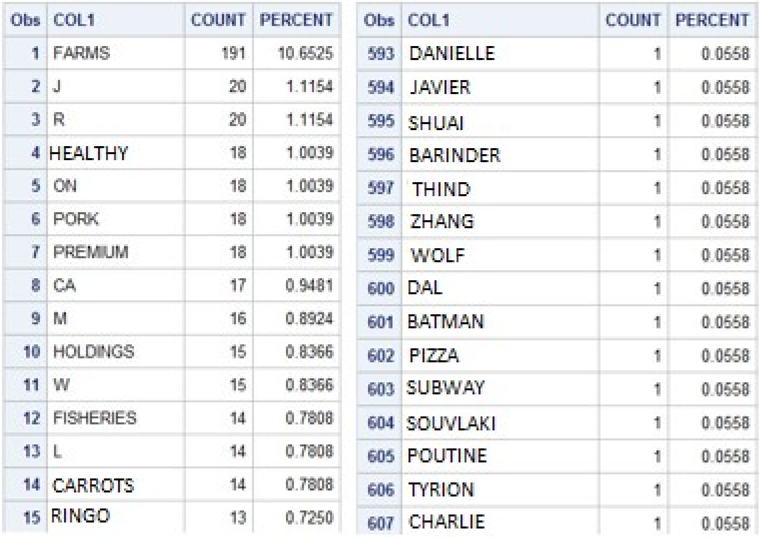
[11] This report will also be going to my school and so, this data is made up.

[12] See: Score Functions in section 3.3

frequently used or “rare”[13] words. In the next section, I highlight some of the different approaches someone could take to resolve this problem.

3.3 Score Function

While identifying the scenario of “rare words” allows us to nullify an assumption, it also opens the door to another problem: What is a reasonable weight? As a cautionary precursor, it is important to note that there is no single, correct answer to such a question. And, in fact, it is entirely possible that some particular[14] results could be attained if a specific weighting method is pursued. Hence, it is crucial to always have some kind of cross-validation to minimize bias. With that said, there are methods of weighting that are clearly superior to others.



Obs	COL1	COUNT	PERCENT
1	FARMS	191	10.6525
2	J	20	1.1154
3	R	20	1.1154
4	HEALTHY	18	1.0039
5	ON	18	1.0039
6	PORK	18	1.0039
7	PREMIUM	18	1.0039
8	CA	17	0.9481
9	M	16	0.8924
10	HOLDINGS	15	0.8366
11	W	15	0.8366
12	FISHERIES	14	0.7808
13	L	14	0.7808
14	CARROTS	14	0.7808
15	RINGO	13	0.7250

Obs	COL1	COUNT	PERCENT
593	DANIELLE	1	0.0558
594	JAVIER	1	0.0558
595	SHUAI	1	0.0558
596	BARINDER	1	0.0558
597	THIND	1	0.0558
598	ZHANG	1	0.0558
599	WOLF	1	0.0558
600	DAL	1	0.0558
601	BATMAN	1	0.0558
602	PIZZA	1	0.0558
603	SUBWAY	1	0.0558
604	SOUVLAKI	1	0.0558
605	POUTINE	1	0.0558
606	TYRION	1	0.0558
607	CHARLIE	1	0.0558

Figure 3.3: The frequency at which the words in NAICS 11 (agriculture) appear. The left table displays the most frequent words, whereas the right graph displays the

[13] Words like “Kozuls” above are something that might be classified as a rare word because of how infrequently it would be used in any nomenclature conventions. For a more strict threshold, you can define it to be a word that is used less than an *x* number of times

[14] desirable

least frequent. Note: these results come from a subset of NAICS 11[15]. COL1 here is a variable which refers to individual words used in the NAICS. [Example Data]

An initial avenue to look into more for weighting methodology might be to see the amount of times a unique word appears in the list of words (for e.g. a list like Figure 3.3). The word “Farms” occupies far more space in that set than all the other words and there is a significant drop off afterwards. This can lead to the following weight/score method:

$$S_A = \sum_{i=1}^{n} \alpha_i \cdot (1 - \frac{p_i}{100})$$

Where S_A is the first score function, α_i is whether the i^{th} word is the same for both names, and p_i is the percentage of the frequency that the word is used.

The percent is the relative frequency at which a particular word appears in the parsed operating and legal names of a NAICS[16]. This method is very simple in its approach and its simplicity makes it attractive. However, issues arise under minimal scrutiny of this method. For example, just looking at the word “Ringo” and the word “Tyrion”, which are apart in frequency by 12 words, we notice that the weight has little to no influence on the resulting score. In fact, the weights for “Ringo” and “Tyrion” would be 0.99275 and 0.99944 respectively. When multiplied by the α_i values (of 1), this results in minimal difference[17]. This can be problematic for multiple reasons, including being susceptible to rounding errors and having razor thin thresholds that can fall victim to margin of errors. Another problem with this method is a variation of one we have already encountered earlier in the paper. Since the difference is so negligible between two different values (some of which of a great magnitude of difference in terms of percent), this does not fully resolve the problematic assumption made earlier. One way to get around such a problem then, is to make the percent as an input into some function that exacerbates the difference in effect. An appropriate function is the exponential one and it gives way to this second weighting scheme:

[15] This is made-up data to use as an example

[16] So, ”Farms” would be 10.6525 percent

[17] ≈ 0.007 in this case

$$S_B = \sum_{i=1}^{n} \alpha_i \cdot (\frac{1}{e^{\psi \cdot \frac{p_i}{100}}})$$

Where S_B is the second suggested score function, α_i is whether the i^{th} word is the same for both names, p_i is the percentage of the frequency that the word is used, and ψ is some constant scalar that is treated as a weight.

The ψ is some constant scalar. The reason that the exponential function is used is because it allows, even with ψ set to 1, there to be some increasingly (or decreasingly) significant effect on the weight value depending on what value you allow percent to take on. Clearly, this effect becomes more impactful the more often a word is used. A methodology for calculating ψ is presented below.

3.3.1 Zipf's Law and its Application

Identifying the value of ψ can be difficult because of the subjectivity involved. However, there are ways to decrease this uncertainty. One way to go about finding this value is to use intuition and minimize the error in intuition through repeated trials.

Before getting into the specifics of the procedure, some knowledge of Zipf's law is required. In short, Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table[18]. That is to say, the word that occurs the second most occurs half as many time as the word that occurs the most. This is an observable phenomenon that has been shown to be true a nearly countless number of times. In the English language, the most used word is "the" and it makes up about 6.68%[19] of all words used. The next most common word is "be" (which is about 3.84% of all words used). This pattern continues and results in the following discrete Pareto distribution shown in Figure 3.4.

[18] This is the Wikipedia definition

[19] From an online resource (wordfrequency.com) that collects data on word usage

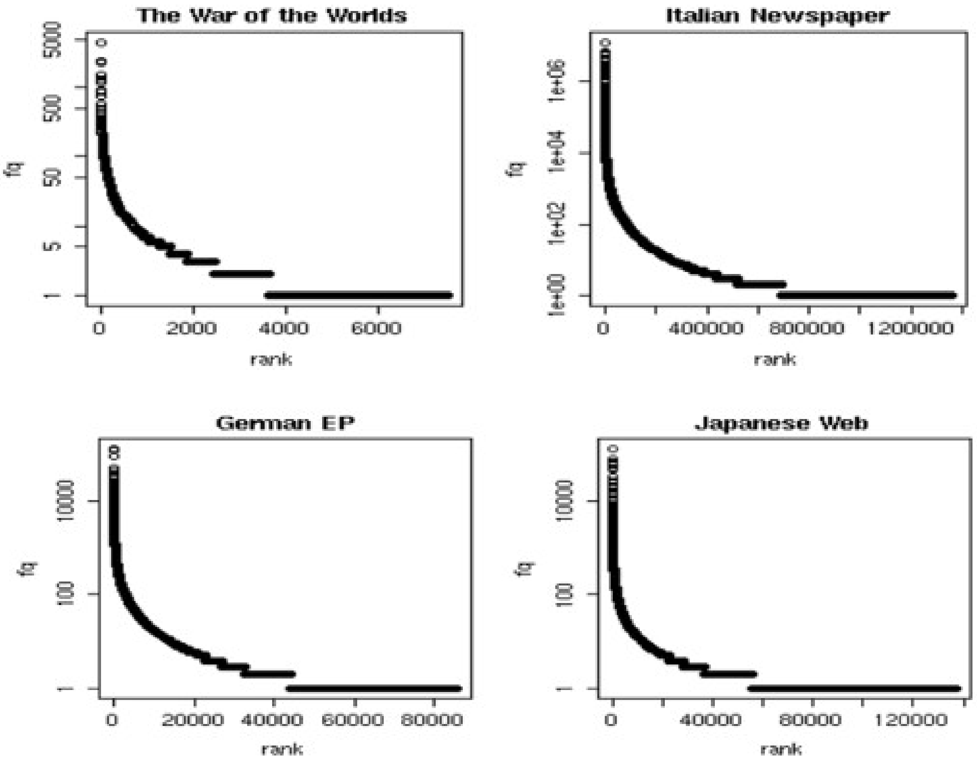


Figure 3.4: The distributions of the frequency of words for four different literary works[20]. A couple of examples of how the NAICS distributions of word frequencies is available in the Appendix.

Now, assuming that this Law holds for the English language (it is reasonable to assume it does), we can say that there is a mapping from the subset of our NAICS to the set of all English words. This gives us a point of reference and allows us to make good, intuitive guesses as to how much we should weight a particular word. As an example to illustrate this, we can consider the word “farms” from NAICS 11.

[20] This was taken from Javier Oyarzun’s presentation of NAICS classification of words

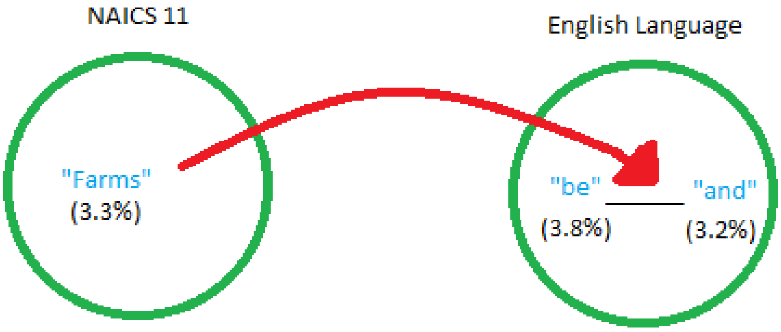


Figure 3.5: Mapping from NAICS 11 to English Language usage

This word makes up about 3.3% of all words in its NAICS. If we superimpose the distribution of NAICS 11 over the distribution of all English words, we can begin to see what value “farms” should take in terms of weight. In Figure 3.6, we can see that, assuming for the sake of argument that “be” and “and” make up 3.8% and 3.2% respectively of the frequency of all words used in the English language, “farms” will land somewhere in the space between those two. This gives rise to a more intuitive thinking process when deciding what the weight should be. A more clear representation of this is given in the following figure:

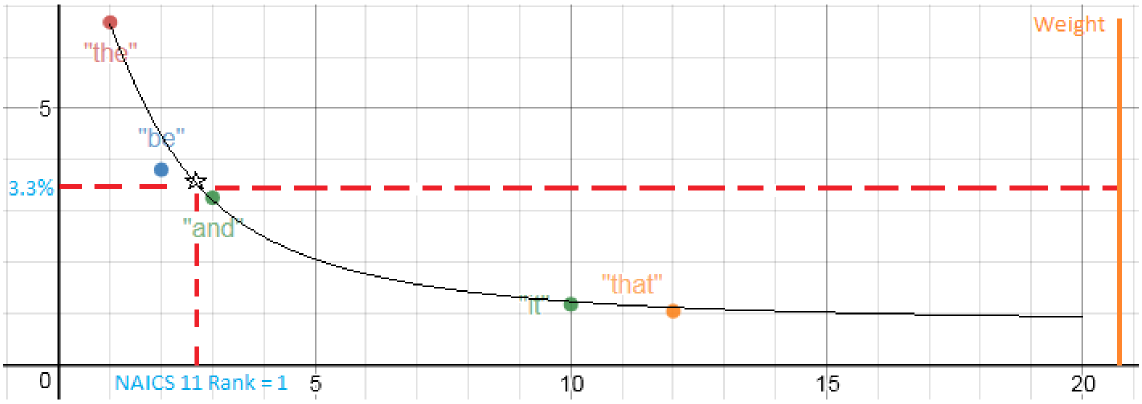


Figure 3.6: A superimposition of “farms” (the star) from the NAICS 11 word frequency/rank distribution onto to the frequency/rank distribution of all words[21].

[21] The x-axis here is the rank and the y-axis is a percentage of the frequency rather than the frequency itself

In this figure, we see that the word “farms” in NAICS 11 is equivalent to a position in the space between “be” and “and”. We would (presumably) never think that two people have said the same thing if we observe that they both used the word “be” or “and” and hence, you can then conclude that you should never think that two businesses are a link purely based on the fact that they matched on the word “farms”. However, it can still be difficult to decide what the weight should be. Let’s assume that you decide, from this new comparative information, that “farms” should be weighted then at 0.1[22]. This is enough for us to estimate a value of *ψ*[23]. However, to add an element of cardinality (the user can get a relative gauge as opposed to a single guess), we can pick multiple words from our subset (in this example, NAICS 11), map them onto the set of all English words, and make similar comparisons as above.

This will result in a set of *ψ’s*. Then, the following will give us an estimate for $\hat{\psi}$:

$$\hat{\psi} = \frac{\sum_{i=1}^{n} \psi_i}{n}$$

Where *n* is the number of words superimposed onto the distribution of all words. There are other ways to get an estimate for $\hat{\psi}$ such as by looking at the median. However, with this measure, we get an unbiased expectation of what the user believes the words they chose have in value when compared to words we use every day.

3.4 Thresholds/Links

The threshold is probably the least objective aspect of any record linkage project. It is difficult to say with certainty that below or above some particular score that you will get all true or false links. This problem, when combined with the size of the BR, will almost always result in false links being true or links being missed. It can by chance, just so happen to be that, even though a match scored very high, it was all coincidental. Therefore, the guidelines I have written here should be taken very cautiously and the reader is welcome to make any reasonable modifications.

[22] This value is measured in comparison 1 as established in Definition 3.1.1

[23] The calculations are shown in section 4.1

A logical way to go about creating a threshold is to consider what percentage of the matches are false links when given some score. Then, you can decide whether or not that is the number you want to go with or whether you want to adjust this score value to meet your needs. The following algorithm should be applied when determining these values:

1) Pick an error rate (whether that 1%, 5%, 10%, etc.)

2) Randomly (but reasonably) pick any score value that is contained within the set of all score values that your dataset takes (take into account any new information)[24]

3) Subset the data such that the score value is greater than or equal to your picked score value from the previous step

4) Use SAS's *proc surveyselect* (or any other appropriate method) to sample (with the size depending on the size of the population) from the previous step's subset

5) Manually scan over the sample and determine your error rate with the following formula:

$$\varepsilon = \frac{\textit{Numnber of False Matches}}{\textit{Total Number of Duplicates}}$$

[24] In particular, the information obtained from step 6 of this algorithm

Where ε is the error rate.

6) Do the following based on your result in 5)

   a. If the error rate is greater than the one you wanted, repeat steps 2 to 5 but pick a score value higher than the one picked previously

   b. If the error rate is lesser than the one you wanted, repeat steps 2 to 5 but pick a score value lower than the one picked previously

   c. If the error rate is as you wanted, then you are done!

7) – Optional – Resampling might be a good idea for confirmation

In the results[25] section, there will be a table provided that lists the threshold score values that correspond with some common set of error rates (a conservative and an aggressive one).

3.5 Geocoding vs. No Geocoding

Another important element of this project was to look at the benefits of having the newly added quantified geographical variables, i.e., the longitude and the latitude. One way to make this comparison is to examine how many false links are made when the above threshold criteria is used while the code for matching via geography is excluded. In particular, I looked at the final results of both approaches (without geographical co-ordinates and with them) and considered only the

[25] Section 4.2

matches deemed to be a link. Then, from these subsets, I took out random samples. Finally, from these samples, I manually inspected them to see how many false links are produced. The results are presented in section 4.3.

4 Results

This section of the report is broken into 3 sections: the calculations for the constant *ψ,* the number of potential duplicates/potential links I got for the thresholds set at aggressive[26] and conservative[27] levels (for both the matches at the block-face level only and matches at the block-face – PCDB level), and a comparison/discussion of the results I got using the geographical coordinates and this new scoring system vs. what was presented before.

4.1 The ψ Calculation

In this section, I will just provide one calculation for ψ as an example. The user results will obviously differ but the underlying procedure should be the same. In this example, we will use the word “farms” from NAICS 11. Assume that “farms” makes up 3.3% of all words in this NAICS. Looking back at Figure 3.6, we see that this is equivalent to having it be between “be” and “and” in terms of everyday use. Seeing as how these words are extremely common in use, it might even be plausible to say that “farms” should have little to no weight on the score. For simplicity’s sake, we can say that “farms” should have a weight of 0.1. Then, the following calculations will get us our value of $\hat{\psi}$ (and, on repetitions, values for the rest of the constants[28]):

[26] An error rate of 10% to 20%

[27] An error rate of 0% to 6%

[28] For n = 5 in this example; the last 4 do not have their calculations shown to avoid redundancy. The n = 5 in this case refers to the number of words taken from the NAICS 11 dataset and super-imposed onto the distribution of all the words (See: Figure 3.6). One of those words is “farms” and that calculation is shown above to get the first value for ψ. The next 4 are “Carrots”, “Beets”, “Celery”, and “Wolfs”. Score claims were made for these words and the same calculations were done.

$$\frac{1}{e^{\psi_1(\frac{percent}{100})}} = ScoreClaim$$

$$\frac{1}{e^{\psi_1(\frac{3.3}{100})}} = 0.1$$

$$\psi_1 \cdot 0.033 = ln(10)$$

$$\psi_1 \approx 69.78$$

Repeating these calculations for 4 other words while adjusting previous score claims, we get the following results:

$$\psi_2 \approx 68.49$$

$$\psi_3 \approx 72.36$$

$$\psi_4 \approx 71.82$$

$$\psi_5 \approx 65.81$$

This results in an estimate of 69.65 for the value of $\hat{\psi}$ [29]. This was the value used to obtain the threshold results in the next section. The results may vary in terms of the score on follow-ups depending on the discretion of the user.

4.2 Thresholds/Links Results

First, a table for the total number of matches found for block-face only:

NAICS	# of Potential Links
11	4321
21	35
22	93
23	6269
31	152
32	73
33	201

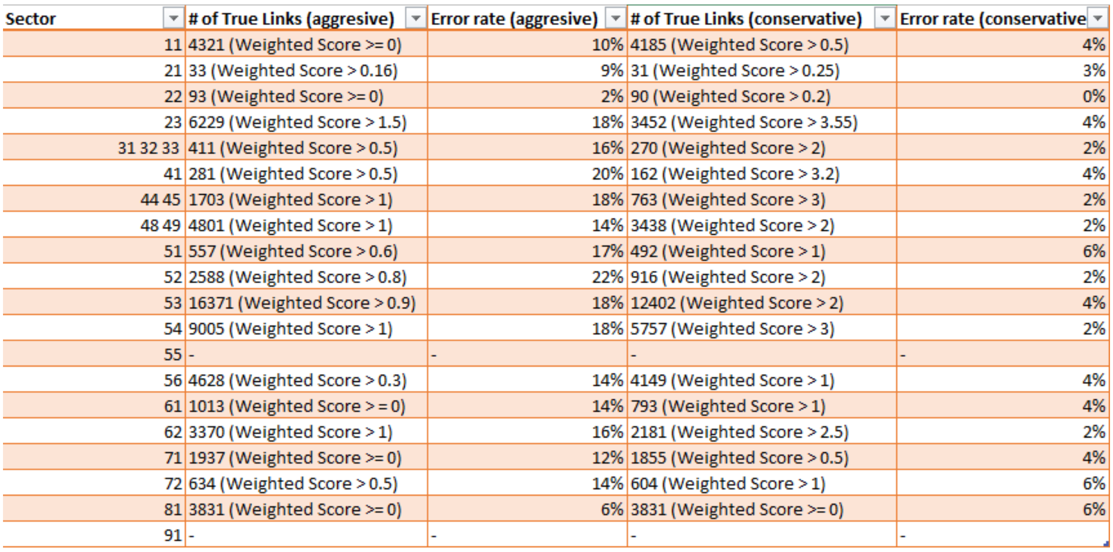
[29] This was found using the formula:

Where *n* is the number of words superimposed onto the distribution of all words.

41	292
44	758
45	1982
48	13926
49	655
51	587
52	10074
53	17457
54	10562
55	0
56	5150
61	1013
62	3926
71	1937
72	655
81	3831
91	0
Total:	**83949**

Table 4.1: Initial matches for each NAICS BF vs. BF (Total = 83949)

The following results pertain to all matches made at the block-face level. The thresholds were created individually for each NAICS following the procedure from section 3.4.



Sector	# of True Links (aggresive)	Error rate (aggresive)	# of True Links (conservative)	Error rate (conservative
11	4321 (Weighted Score >= 0)	10%	4185 (Weighted Score > 0.5)	4%
21	33 (Weighted Score > 0.16)	9%	31 (Weighted Score > 0.25)	3%
22	93 (Weighted Score >= 0)	2%	90 (Weighted Score > 0.2)	0%
23	6229 (Weighted Score > 1.5)	18%	3452 (Weighted Score > 3.55)	4%
31 32 33	411 (Weighted Score > 0.5)	16%	270 (Weighted Score > 2)	2%
41	281 (Weighted Score > 0.5)	20%	162 (Weighted Score > 3.2)	4%
44 45	1703 (Weighted Score > 1)	18%	763 (Weighted Score > 3)	2%
48 49	4801 (Weighted Score > 1)	14%	3438 (Weighted Score > 2)	2%
51	557 (Weighted Score > 0.6)	17%	492 (Weighted Score > 1)	6%
52	2588 (Weighted Score > 0.8)	22%	916 (Weighted Score > 2)	2%
53	16371 (Weighted Score > 0.9)	18%	12402 (Weighted Score > 2)	4%
54	9005 (Weighted Score > 1)	18%	5757 (Weighted Score > 3)	2%
55	-	-	-	-
56	4628 (Weighted Score > 0.3)	14%	4149 (Weighted Score > 1)	4%
61	1013 (Weighted Score > = 0)	14%	793 (Weighted Score > 1)	4%
62	3370 (Weighted Score > 1)	16%	2181 (Weighted Score > 2.5)	2%
71	1937 (Weighted Score >= 0)	12%	1855 (Weighted Score > 0.5)	4%
72	634 (Weighted Score > 0.5)	14%	604 (Weighted Score > 1)	6%
81	3831 (Weighted Score >= 0)	6%	3831 (Weighted Score >= 0)	6%
91	-	-	-	-

Table 4.2: Thresholds for each NAICS along with the relevant error rates. The total number of matches: Aggressive – 61806, Conservative – 47226

The total number of empty cells means that either it was too difficult to determine whether the links were false or not (because of the nature of the NAICS) or there weren't enough observations in the particular NAICS to make any kind of determination.

There are actually less matches initially found here when compared with the results that were obtained previously using a different methodology. Here are two reasons as to why this might be:

1) I had the added constraint of matching on the geographical co-ordinates whereas the original attempt left out this information (it wasn't available). This significantly impacts what constitutes a match as SAS will now not recognize that two things potentially might be a link based on name only.

2) My matches are only based on block-face coordinate points. This accounts for another chunk of the disparity. It might make sense to continue down this path if I do not finish in time in generalizing the code. For e.g., matching on block-face vs. dissemination block.

Finally, here are the matches made at the BF vs. PCDB and BF vs. STDB level. Using postal codes (Table 4.3), there are way less matches made than if no postal codes were used (those results will be available in the appendix).

NAICS	# of Potential Links BF vs. PCDB	# of Potential Links BF vs. STDB
11	2816	**408**
21	7	**0**
22	26	**6**
23	1028	**187**
31	2	**0**
32	2	**0**
33	24	**5**
41	23	**2**
44	452	**12**

45	1839	**45**
48	555	**183**
49	43	**6**
51	284	**8**
52	792	**83**
53	8124	**587**
54	1336	**149**
55	0	**0**
56	2938	**89**
61	129	**21**
62	498	**45**
71	127	**14**
72	112	**10**
81	706	**101**
91	0	**0**
Total:	**21863**	**1961**
Overall Total:	23824	

Table 4.3: Number of matches for BF vs. PCDB (left) & BF vs. STDB (right) [compGED = 500]

In the appendix, there will also be results for when compGED = 100 and, as mentioned before, results for when no postal code constraints were added.

4.3 Geocoding vs. Non-Geocoding

When I avoided the geographical coordinates as a constraint for the initial matches, I got 12 false links from 62 matches. I used a compGED of 100 for these results. An example of a false link made under these circumstances is a match between "Kevin Farms" and "Kevin Hamilton". If the constraint of the GPS had been there, this false link would not have been made.

In contrast, taking a subset using the GPS constraint, I observed 0 false links from 19 matches in NAICS 11. Clearly, there is a significant decrease in false links made, at least in this context. These results were found with using a compGED of 100. In another sample, using compGED of 500[30] (with the GPS constraint), there was 7 potentially false links made from 62 matches. So, even with strings that are further

[30] A larger compGED implies that the strings are further apart

apart, it seems that just having the GPS constraint significantly decreases the error rate that you would get – without applying any thresholds.

5 Conclusions and Future Considerations

The problem of finding an appropriate score function seems to be a difficult one. It is often challenging to gauge where thresholds should be placed and how to value the likeness of names. Through this project, there has been the introduction of solving the score problem by weighing the impact of the words in the names via an application of Zipf's law. On top of that, the threshold problem was somewhat remedied by an algorithm which gives a measuring stick to the user as to where they would want the cut-offs to be. Lastly, the potential for geographical co-ordinates to have an impact on future work is huge. Even in their initial implementation in this project, they decreased error rates extensively before a threshold of score is ever even applied.

While there are still some problems with this methodology, a lot of them stem from practicality. Often, because of processing times, there is a lag between running code and actually getting results. It can also be fairly time consuming to go through samples to figure out error rates. In the future, it might be worth looking into figuring out ways to minimize this processing time. Some routes to consider could be an alternative way of matching how scores are given or looking at methods in which the initial matches take less time to make.

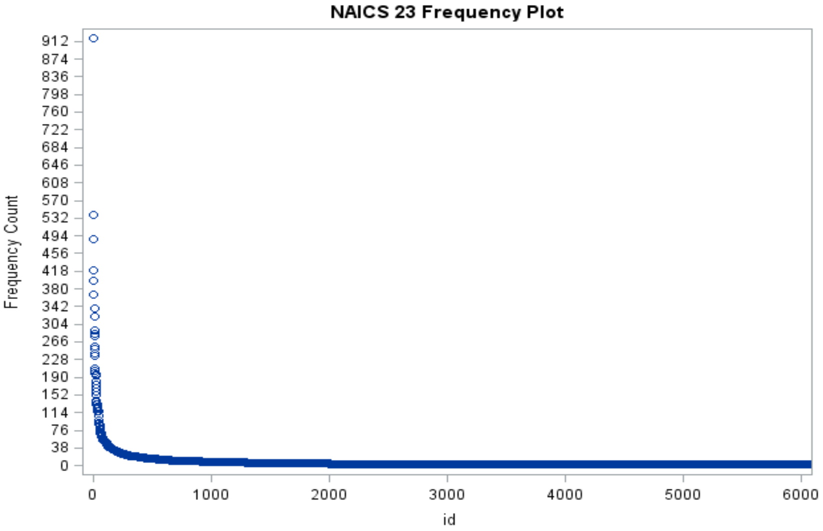
6 Acknowledgements

Thank you again to Shuai Zhang and Javier Oyarzun for all the help they provided me throughout the work term while also showing me just some of the great things that Statistics Canada has to offer. Also, thank you to all my co-workers who made my first co-op placement as enjoyable as it was!

7 References

- Rainsford, Chris, Lifang Gu, Rohan Baxter, and Deanne Vickers. *Record Linkage: Current Practice and Future Directions* (n.d.): n. pag. Web.
- "Word Frequency Data." *Word Frequency: Based on 450 Million Word COCA Corpus*. N.p., n.d. Web. 23 Mar. 2017.
- "Free Flowchart Maker and Diagrams Online." *RSS*. N.p., n.d. Web. 21 Mar. 2017
- "Zipf's Law." *Wikipedia*. Wikimedia Foundation, 21 Mar. 2017. Web. 23 Mar. 2017.ss

8 Appendix



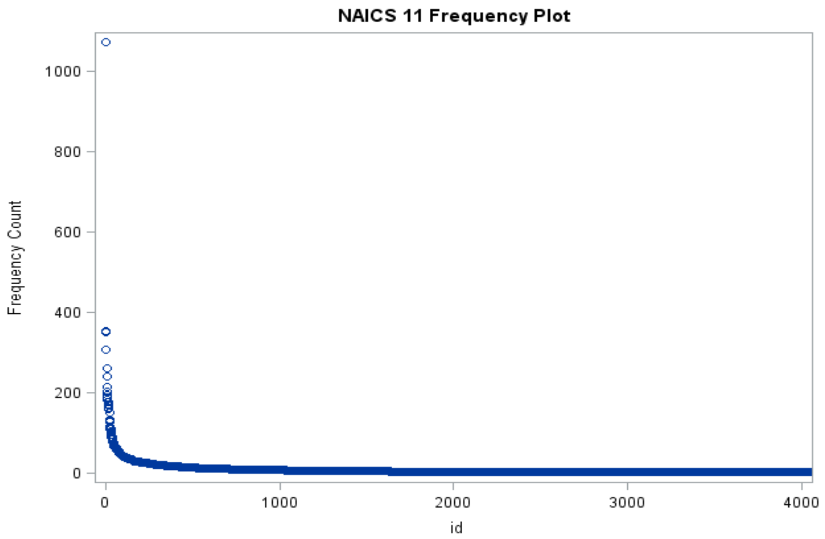


Figure 7.1: Graphs for NAICS 23 and NAICS 11 that show the manifestation of Zipf's law (id is rank).

NAICS	# of Potential Links BF vs. PCDB	# of Potential Links BF vs. STDB
11	492	**58**
21	5	**0**
22	8	**4**
23	282	**17**
31	1	**0**
32	0	**0**
33	8	**2**
41	7	**0**
44	38	**2**
45	105	**15**
48	227	**47**
49	29	**2**
51	21	**2**
52	528	**52**
53	1036	**63**

54	390	**35**
55	0	**0**
56	113	**15**
61	71	**11**
62	174	**12**
71	39	**3**
72	64	**5**
81	191	**17**
91	0	**0**
Total:	**3829**	**362**
Overall Total:		4191

Table 7.2: Number of matches for BF vs. PCDB (left) & BF vs. STDB (right) [compGED = 100]

NAICS	# of Potential Links BF vs. PCDB (no postal code constraint)
11	31725
21	30
22	4289
23	18653
31	65
32	30
33	145
41	1007
44	4603
45	40193
48	58622
49	1261
51	1595
52	80344
53	187916 (Stopped)
54	

55	
56	
61	
62	
71	
72	
81	
91	
Total:	**430478**

Table 7.3: Matches made BF vs. PCDB using no postal code constraint[31]

Notes:

- The data must be subset on the business code and the province – otherwise SAS seems to run out of resources
- Processing time for block-face vs. block-face at compGED 500: 1586 minutes (Approximately 26 hours)
- Processing time for block-face vs. PCDB at compGED 100: < 1 hour (using postal code)
- Processing time for block-face vs. STDB at compGED 100: < 1 hour (using postal code)
- The word count is significantly impacted by the figure/table descriptions, the definition table in the beginning, and the references. It is actually closer to 3000.
- It makes sense to run block-face with PCDB even if it takes a long time initially. There ends up being more matches and then the scoring system can be applied to these results.

[31] Stopped due to the processing time (ran for over 3 days)