



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
www.statcan.gc.ca



Project for the Regional Advancement of Statistics in the Caribbean - PRASC

Funded by the
Government
of Canada

Canada



Statistics
Canada

Statistique
Canada

Canada



Statistics Canada
www.statcan.gc.ca

Project for the Regional Advancement of Statistics in the Caribbean - PRASC

Component: Business Survey Infrastructure

Funded by the
Government
of Canada

Canada



Record Linkage for Business Surveys

Mike Hidioglou and Wesley Yung
Business Survey Methods Division

January 19, 2016
Port of Spain, Trinidad and Tobago

Outline

- Introduction
- Probabilistic Record Linkage
- Deterministic Record Linkage
- Pre-Match Preparation
- Business Survey Record Linkage
- Summary

Introduction

- Record linkage is the process of matching records in two files which represent identical persons, objects or events
 - Can be used to find duplicates in a single file by matching it to itself
- In statistical offices, this exercise is becoming more and more common due to an increased use of administrative data
 - Business Register / Frame creation
 - Replacement of collected survey data
 - Big Data?

Introduction

- There are two main types of record linkage
 - Statistical matching
 - Exact matching
- Statistical matching attempts to link files that may have few units in common
 - More interested in relationships between variables on the files
 - Suppose variables X and Y are on one file and X and Z on a second file and we are interested in the relationships between X, Y and Z
 - Used more for analysis purposes

Introduction

- Exact matching
 - Links records that represent the same person, object or event
 - Assumes that the same unit is on both files
 - Matching is done by some unique key or based on one or more variables
- Within exact matching, there are two approaches
 - Probabilistic
 - Deterministic

Probabilistic Record Linkage

- Used when a unique identifier is not available
- Estimates the likelihood that two records are referring to the same entity
- Theory based on Fellegi-Sunter (JASA 1969)
 - Formalized linkage practices and specified properties of optimal linkage rules
- Link records for which some variables are not in complete agreement

Probabilistic Record Linkage

- Based on measures of ‘similarity’, pairs of records are assigned a status of match, possible match or rejected
- The assignment is based on ‘linkage scores’ and decision rules
- Matched records are accepted as linked and possible matches are manually reviewed
 - Cost/quality trade-off

Probabilistic Record Linkage

- Measures of similarity are based on string comparators
 - How many characters are the same/different?
 - How many changes are required to make the two strings equal?
- Variables can be assigned different 'weights' to reflect their discriminatory power
 - Full name is more discriminate than first initial and last name
- More in depth discussion is out-of-scope for the workshop

Deterministic Record Linkage

- Determines whether record pairs agree or disagree on a set of identifiers
 - Outcome is 'agree' or 'disagree'
 - No notion of likelihood of agreement
- Simplest case is direct matching
 - Linking of records using a unique identifier or key
 - For example, a business number, a BR identifier, social security number, etc.
- Unfortunately, not all sources share a unique identifier

Deterministic Record Linkage

- Alternate method is Hierarchical Exact Matching
 - An iterative or stepwise approach
 - Multiple passes with different matching criteria
 - Usually starts with the most stringent matching criteria and moves to less stringent ones
 - When determining the most stringent criteria, must consider
 - Power – Ability to find true pairs
 - Specificity – Ability to reject false pairs

Deterministic Record Linkage

- Hierarchical exact matching – Example

<u>Pass 1</u>	<u>Pass 2</u>	<u>Pass 3</u>	<u>Pass 4</u>
First name	First name	First name	First initial
Middle name			
Last name	Last name	Last name	Last name
Street number and name	Street number and name	Street name	Street name
Postal code	Postal code	Postal code	Postal code
Gender	Gender	Gender	Gender

Deterministic Record Linkage

- Once all passes have been done, it is important to review the links
 - All identified links should be stored in an accumulated links file
 - Multiple links with the same record should be reviewed
 - Could confirm the link or be in conflict
 - If the percentage of conflicts from a pass is high, may not want to use the results

Pre-match Preparation

- Given that variables must match, it is important to properly prepare the files
- All linkage variables should be standardized across files
 - Same case: Upper or lower case
 - Same format: 01SEP2013
 - Same content: No punctuation, no digits in text field, etc.
 - Same length
 - Similar names: Charles to Chuck, Richard to Dick, etc.
 - Remove prefixes and suffixes: Dr. John Smith Jr. to John Smith

Pre-match Preparation

- Remove accents: Éric to Eric
- Normalize abbreviation and compound words
 - P.R.A.S.C to PRASC
 - Baby sitter to Babysitter
- Algorithms exist to account for misspelling
 - Strings converted to phonetic codes (ex. Soundex, New York State Immigration Information System)

Pre-match Preparation

- Keep only those fields needed for linkage
 - Reduces execution time and space required
- Check values for accuracy
 - Feb. 14, 11962 to Feb. 14, 1962
- Delete unimportant words
 - Inn of Roses to Inn Roses

Pre-match Preparation

- Identifiers can be parsed into separate pieces of information
 - Names can be parsed into first, middle and last name
 - Dates of birth can be parsed into day, month and year
 - Addresses can be parsed into street, city province and postal code
 - This may allow enough partial matches to provide sufficient evidence that the records being compared are the same
 - Ex. Address is the same but full name on one file includes middle initial and not middle name

Business Survey Record Linkage

- Becoming more and more necessary due to trend of using more alternate sources of data
- Many administrative files
 - Lack common identifiers
 - Have poor quality common identifiers
 - Do not have standard formats
 - Contain typographical errors
 - Are very large in size

Business Survey Record Linkage

- Standardization of business names
 - Convert all letters to upper case
 - Remove all accents
 - Replace certain common strings
 - A/C to Air conditioning
 - Standardize geographical names (Provinces)
 - Remove spaces between words
 - Fast Food to Fastfood
 - Remove hyphens
 - Co-op to Coop
 - Drop trivial words/letter (LTD, Co., 's', etc.)

Business Survey Record Linkage

■ Examples

ORIGINAL NAME1	STANDARDIZED NAME1	MODIFICATIONS
MACKENZIE TRANSPORT LIMITED	MACKENZIE TRANSPORT	1) Drop "LIMITED"
JONES BROS. DEVELOPMENT CO. LTD.	JONES BROS DEVELOPMENT	1) Trimming of "." from "BROS." 2) Drop "CO." 3) Drop "LTD."
CONSTRUCTIONS SHERBROOKE LTÉE	CONSTRUCTION SHERBROOKE	1) Drop "S" from "CONSTRUCTIONS" 2) Drop "LTÉE"
DVM BUSINESS INTERIORS INC	DVM INTERIOR	1) Drop "BUSINESS" 2) Drop "S" from "INTERIORS" 3) Drop "INC"
ROYAL REFINING CANADA LTD	ROYAL REFINING CA	1) Change "CANADA" to "CA" 2) Drop "LTD"
T & T ENGINEERING	T ENGINEERING	1) Drop repeated letter "T" 2) Drop "&"
CENTURY SPAS (WINNIPEG) LTD	CENTURY SPAS WINNIPEG	1) Remove brackets 2) Remove "LTD"

Business Survey Record Linkage

- Once files are cleaned up, direct or hierarchical exacting matching can be performed
- If files are large, a blocking approach can be used to reduce time required
 - Blocking is when records are first matched on a good quality variable to reduce the number of potential pairs
 - For example, the province

Business Survey Record Linkage

- Consider the following example

Master File

EntID	Legal Name	Province
1	ABC Farms	AB
2	Al's Diner	AB
3	Wheel Factory	MB
4	Pizza To Go	MB

Birth File

EntID	Legal Name	Province
11	Kiddy Central	AB
12	Fred's Lighting and Decor	AB
13	JJ Technologies	MB
14	Wheel Factory	MB

Business Survey Record Linkage

- Without blocking, there are 16 possible pairs

Comparison Count	Master			Birth		
	EntID	Legal Name	Province	EntID	Legal Name	Province
1	1	ABC Farms	AB	11	Kiddy Central	AB
2	1	ABC Farms	AB	12	Fred's Lighting and Decor	AB
3	1	ABC Farms	AB	13	JJ Technologies	MB
4	1	ABC Farms	AB	14	Wheel Factory	MB
5	2	Al's Diner	AB	11	Kiddy Central	AB
6	2	Al's Diner	AB	12	Fred's Lighting and Decor	AB
7	2	Al's Diner	AB	13	JJ Technologies	MB
8	2	Al's Diner	AB	14	Wheel Factory	MB
9	3	Wheel Factory	MB	11	Kiddy Central	AB
10	3	Wheel Factory	MB	12	Fred's Lighting and Decor	AB
11	3	Wheel Factory	MB	13	JJ Technologies	MB
12	3	Wheel Factory	MB	14	Wheel Factory	MB
13	4	Pizza To Go	MB	11	Kiddy Central	AB
14	4	Pizza To Go	MB	12	Fred's Lighting and Decor	AB
15	4	Pizza To Go	MB	13	JJ Technologies	MB
16	4	Pizza To Go	MB	14	Wheel Factory	MB

Business Survey Record Linkage

- With blocking, there are 8 possible pairs

Alberta Comparisons

Comparison Count	Master			Birth		
	EntID	Legal Name	Province	EntID	Legal Name	Province
1	1	ABC Farms	AB	11	Kiddy Central	AB
2	1	ABC Farms	AB	12	Fred's Lighting and Decor	AB
3	2	Al's Diner	AB	11	Kiddy Central	AB
4	2	Al's Diner	AB	12	Fred's Lighting and Decor	AB

Manitoba Comparisons

Comparison Count	Master			Birth		
	EntID	Legal Name	Province	EntID	Legal Name	Province
5	3	Wheel Factory	MB	13	JJ Technologies	MB
6	3	Wheel Factory	MB	14	Wheel Factory	MB
7	4	Pizza To Go	MB	13	JJ Technologies	MB
8	4	Pizza To Go	MB	14	Wheel Factory	MB

Summary

- Knowledge of record linkage becoming more important
 - Increase use of data from different sources (data integration)
- Probabilistic record linkage a complex undertaking but free software does exist
 - FRIL, Link King, Link Plus, etc.
- Deterministic record linkage less powerful but easier to perform
 - Can be done manually in Excel

Summary

- Key to any linkage is in pre-linkage preparation
 - Standardization, parsing, etc.
- Business survey record linkage mostly performed on business name
- Consider blocking on additional variables
 - For example, province or industry



You can contact the PRASC team at:

prasc@statcan.gc.ca

or

statcan.prasc-prasc.statcan@canada.ca

Canada