PRASC

Project for the Regional Advancement of Statistics in the Caribbean

Projet régional pour l'avancement de la statistique dans les Caraïbes

Funded by the Government of Canada

Canada

Record Linkage

As Presented to the Central Statistical Office of Trinidad & Tobago

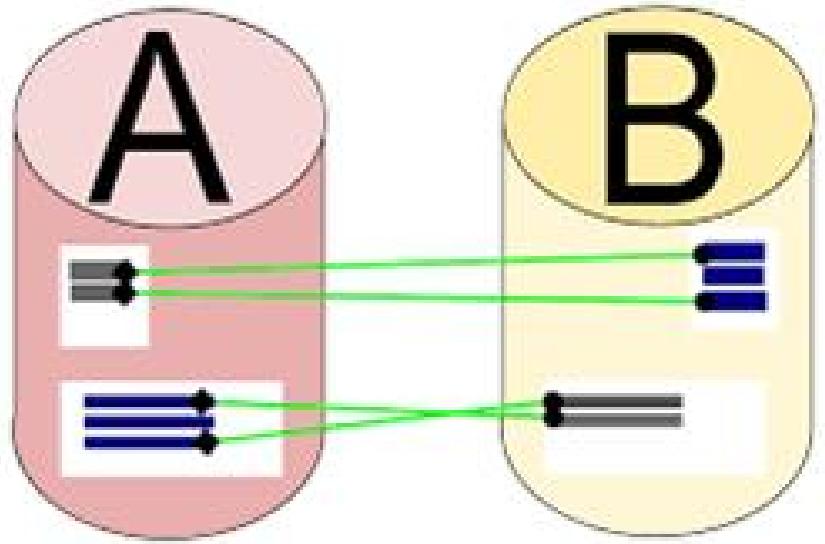
Project for the Regional Advancement of Statistics in the Caribbean (PRASC)

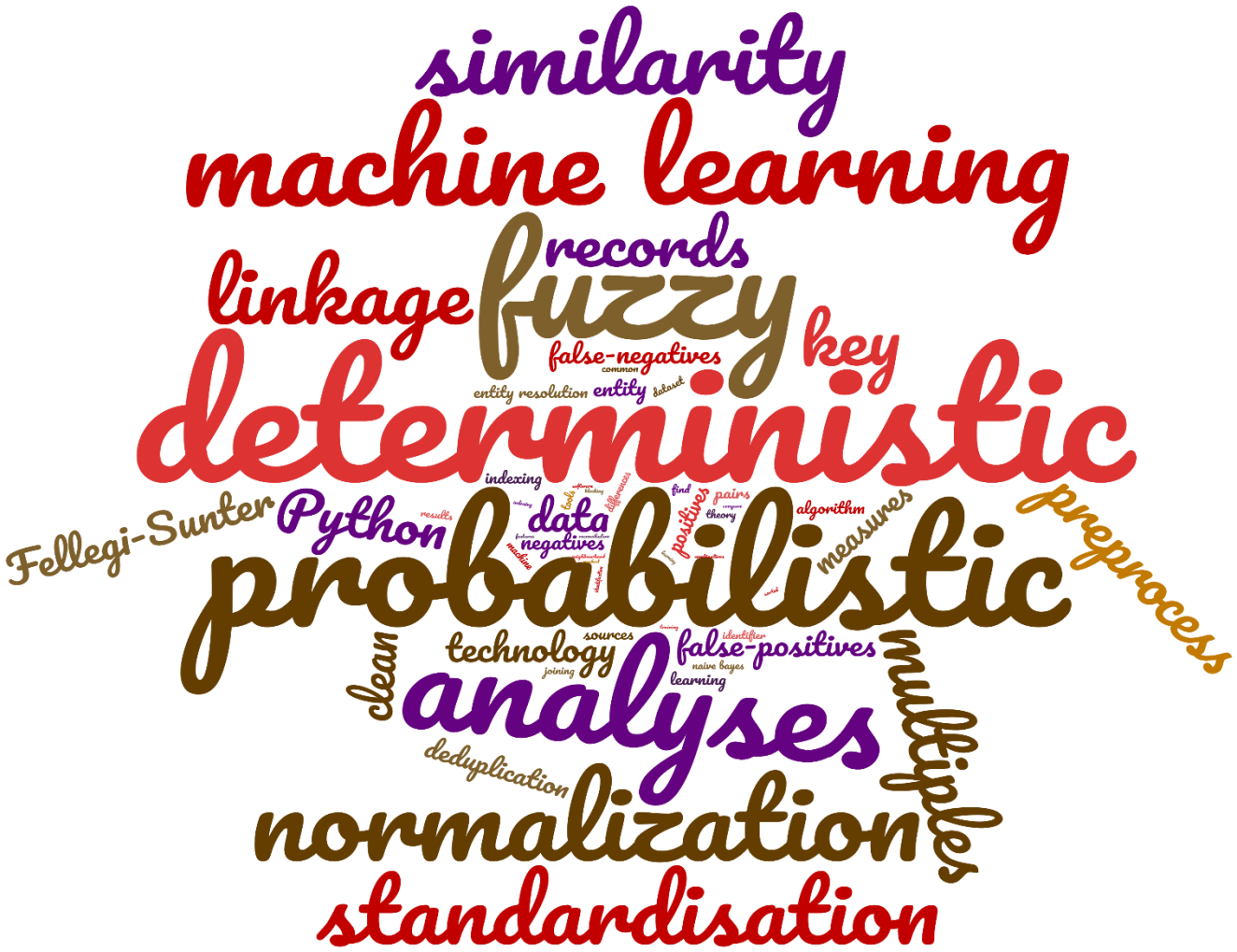
Gaétan St-Louis and Jeff Mondoux
Statistics Canada
March 2020
Port of Spain, Trinidad

Delivering insight through data, for a better Canada

Presentation Outline

- Overview of Record Linkage
- Types of Matching
- Pre-Processing of Linkage Variables
- Deterministic Matching with Example
- Ongoing BR Update Processes
- Conclusions




similarity
machine learning
records
linkage
fuzzy
key
false-negatives
common
entity resolution
entity
robust
deterministic
indexing
Fellegi-Sunter
Python
data
find
pairs
positives
theory
algorithm
preprocess
negatives
measures
probabilistic
sources
identifier
technology
false-positives
clean
naive bayes
learning
analyses
multiples
deduplication
normalization
standardisation

Overview of Record Linkage

- The process of matching files together to:
  - Identify records in each file that represent the same person, business, object or event
  - Identify duplicates within a file by matching it to itself (deduplication)
  - Produce datasets for analysis, production of statistics, initialization or maintenance of Statistical Business Register (SBR)
- Essential when using administrative data from various sources
  - Lack of unique common identifiers
  - Poor quality of unique common identifiers
  - Lack of standardized formats
  - Contain typographical errors
  - Large volume of records

Overview of Record Linkage (cont.)

- In National Statistics Offices (NSO), this exercise is essential due to the increased use of multiple administrative data sources

    - Business Register / Frame creation
    - Replacement of collected survey data
    - Coherency and data confrontation analysis

Overview of Record Linkage (cont.)

- Links records that represent the same entity (person, object, event, etc.)
- Assumes that the same entity is on both files
- Matching is done using one or more common variables

- Each file should contain a unique identifier
  - Permits the entity within future files (from the same sources) to be quickly and easily linked
  - For example: registration number, tax identification number, etc..
  - Administrative sources often assign their own unique identifier for their particular program

    National Insurance (National Insurance Reg. No./Employer Reg. No.)
    Value Added Tax (VAT registration No.)
    Annual Income Tax (BIR File No./PAYE)
    Business Registration (Certificate No.)
    Business Licence

Overview of Record Linkage (cont.)

- The relationships of the unique identifiers from each source are stored in an accumulated links file for future use
  - Often called a concordance file

Example:

STATID	BIRID	NIBID	VATID
S001	T123	E234	V346
S002	T124	E244	V456
S003		E255	V366

Overview of Record Linkage (cont.)

- Data matching has three primary approaches
  - Deterministic
  - Probabilistic
  - Machine Learning

- All matches should be reviewed in some manner to minimize false matches (false positives)
  - Of particular interest ➔ Multiple links to the same record

Types of Matching, Deterministic Record Linkage

Used to match the same or similar values from different sources using a series of defined rules (test for equality)

- ☑ Quick to run (low computational)
- ☑ No special software required
- ☑ Powerful if a number of matching variables are available
- ☑ Easy to understand and interpret
- ☑ Usually results in very high quality matches

- ☒ Can be more manually intensive to develop and implement business rules
- ☒ Lacks robustness/flexibility
- ☒ Depending on data, can yield low match rates

Types of Matching, Deterministic Record Linkage (cont.)

Examples:

- Matching by common unique identifiers assigned by administrative sources
- Exact matching on a data item (or combination of data items) that are expected to produce an accurate match (example: Business Name)
- Matching on transformed or normalized data allowing similar data items to be linked which in combination with other data items produce an accurate match
- Using the concordance file to reuse matches found in the past

Types of Matching, Probabilistic Record Linkage

Used to match similar values from different sources using a series of calculated weights for each data item based on its estimated ability to correctly identify a match or a non-match (test for similarity)

- ☑ Minimal preprocessing of matching variables
- ☑ Offers flexibility and robustness (partial/varying degree of matching)
- ☑ Can increase match rates over deterministic matching
- ☑ Valuable method in the absence of a common identifier and reduced quality of data

- ☒ Special software is required
- ☒ Increased computational requirements vs. deterministic matching
- ☒ Increased runtime, can be slow for larger datasets*
- ☒ Match results harder to interpret
- ☒ Potential to increase noise in the match result set
- ☒ Must include some review of matches

Types of Matching, Probabilistic Record Linkage (cont.)

- Uses a measure of similarity of each matching variable
  - Edit based – Generalize Edit Distance, Levenshtein, Jaro/Jaro-Winkler
  - Token based – Cosine similarity, Jaccard Coefficient, etc
  - Other – Locality Sensitive Hashing (LSH), MinHash, etc.

Example use case: ‘Andrew Martin Smith’ vs ‘Andy Martin Smith’

given some measure of similarity: 90% similar*

Example tools:

- Fuzzy Match add-in for Microsoft Excel (Free)
- G-Link, developed by Statistics Canada
- Python: Record Linkage Toolkit, Scikit-learn, Jellyfish (open source)

Types of Matching, Machine Learning

Extension of probabilistic linkage theory that uses a mathematical model to determine likely matches between data sets.

- ☑ Can be very powerful at identifying difficult to make matches
- ☑ Can be beneficial when data and volume warrants

- ☒ Often requires adequate training data or a 'truth' file to achieve quality results
- ☒ Transparency of match results (black box)
- ☒ Specialized software required
- ☒ Expertise required to understand and successfully apply models
- ☒ Computational and resource intensive

*Not typically used by National Statistics Offices for frame creation

Pre-processing of Linkage Variables

- It is important to properly assess and prepare each file for linkage
  - All linkage variables should be standardized/normalized across files
  - Keep only the fields needed for linkage
    - Reduces execution time and space required

Pre-processing of Linkage Variables (cont.)

- Analyse & Prepare data
  - Identify Linkage Variables
    - Business names: Legal Name, Trade Name
    - Addresses
    - Telephone numbers
    - Etc.
  - Formatting Files
    - Standardize values
    - Rename Variables to Identify Source

Pre-processing of Linkage Variables (cont.)

- Standardizing business names
  - Convert to upper case
  - Standardize common strings
    - A/C ➔ AIR CONDITIONING
    - ST-, ST., SAINT-, SAINT. ➔ ST
    - **Found by reviewing values**
  - Remove punctuation and spaces
    - MOM's FAST FOOD ➔ MOMSFASTFOOD
    - CO-OP ➔ COOP

Pre-processing of Linkage Variables (cont.)

- Standardizing business names (cont.)
  - Drop trivial words/letters
    - LTD, LIMITED
    - CO, COMPANY
    - INC, INCORPORATED
    - AND, OF, THE

- Standardize addresses by data field
  - Address – number, thoroughfare name and type
  - District/village
  - City

- Standardize geographical names
  - District/village names
  - Thoroughfare types

Pre-processing of Linkage Variables (cont.)

- Algorithms exist for standardizing or creating alternate representations for some fields

  - Strings:
    - Direct Match Key (DMK) (useful for business names)
    - Soundex (phonetic)
    - MatchRatingCodex (phonetic)
    - Metaphone (phonetic, targeted at person names)
    - NYSIIS (phonetic, targeted at person names)

    ….
  - Others:
    - Postal
    - Hashing

    ….

Jeff’s Bakery Limited → JFBKR

Jeff’s Bakery Limited → J121

Jeff’s Bakery Limited → JF’MTD

Sean Shawn Steven Stephan → SN PHN

Sean Shawn Steven Stephan → SAN SAN STAFAN STAFAN

Deterministic Matching

- An iterative approach
- Multiple passes with different matching criteria
- Start with the most stringent criteria and move to less stringent ones
- When determining the most stringent criteria, consider
  - The ability to find true pairs
  - The ability to reject false pairs

Deterministic Matching (cont.)

- Typically starts with:
  - Raw values then moves towards preprocessed values
    - The raw values can have some minor preprocessing such as the removal of accents and case conversions
  - Deterministic matches then moves towards less precise matches using transformed data
    - Allows deterministic matching to quickly find the most likely links
    - Smaller files of remaining unlinked records are entered into the rules requiring more computing intensive transformations

  - The art of record linkage is to find the best order to execute the various possible combinations of variables and approaches

Deterministic Matching (cont.)

- After each pass ➔ important to review the links
  - If the first 100 links are valid, consider accepting all links from the pass
  - If the first 100 links are invalid, consider rejecting all links from the criteria

- All true links are saved to the concordance file
- All unlinked records enter the next pass of the strategy

- Final Check: Multiple links with the same record should be reviewed
  - Determine which are likely true links
  - If more than one link appears valid, determine how to resolve

Deterministic Matching (cont.)

Pass 1	Pass 2	Pass 3	Pass 4
Legal Name	Legal Name		
Operating Name		Operating Name	Operating Name
Street Name	Street Name	Street Name	
City	City	City	City
Telephone Number	Telephone Number	Telephone Number	Telephone Number

Legal name = Enterprise
Operating name = Establishment

Deterministic Matching

Example

Deterministic Matching Example

BIR File

BIRID	Legal Name	Trade Name	Address	Tel#	Revenue
T001	123 Bahamas	Goulet and St. Louis Ltd	123 Main Street	123-234-0202	1,000,000
T002	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-234-1111	500,000
T003	Bahamas Hotel	Beach Front Hotel	5 Beach Street	123-234-2525	5,000,000
T004	1202 Bahamas	Island Steak House	10 George Street	123-234-1010	2,000,000
T005		Cycle Repair	15 Cloud Street	123-234-2345	10,000

National Insurance File

NIBID	Legal Name	Trade Name	Address	Tel#	Emp.
E010	123 Bahamas	Goulet & St-Louis Limited	123 Main	123-234-0202	100
E020	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-235-1111	50
E030	Bahamas Hotel	Beach Front Hotel	5 Beach Street	123-234-2525	75
E040	1202 Bahamas	Island Steak House	10 George Street	123-234-1010	20

Deterministic Matching Example (Cont.)

BIR File

BIRID	Legal Name	Trade Name	Address	Tel#	Revenue
T001	123 Bahamas	Goulet and St-Louis Ltd	123 Main Street	123-234-0202	1,000,000
T002	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-234-1111	500,000
T003	Bahamas Hotel	Beach Front Hotel	5 Beach Street	123-234-2525	5,000,000
T004	1202 Bahamas	Island Steak House	10 George Street	123-234-1010	2,000,000
T005		Cycle Repair	15 Cloud Street	123-234-2345	10,000

National Insurance File

NIBID	Legal Name	Trade Name	Address	Tel#	Emp.
E010	123 Bahamas	Goulet & St-Louis Limited	123 Main	123-234-0202	100
E020	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-235-1111	50
E030	Bahamas Hotel	Beach Front Hotel	5 Beach Street	123-234-2525	75
E040	1202 Bahamas	Island Steak House	10 George Street	123-234-1010	20

Matched Result File #1

BIRID	NIBID	Legal Name	Trade Name	Address	Tel#	Revenue	Emp.
T003	E030	Bahamas Hotel	Beach Front Hotel	5 Beach Street	123-234-2525	5,000,000	75
T004	E040	1202 Bahamas	Island Steak House	10 George Street	123-234-1010	2,000,000	20

Deterministic Matching Example (Cont.)

BIR File

BIRID	Legal Name	Trade Name	Address	Tel#	Revenue
T001	123 Bahamas	Goulet and St.Louis Ltd	123 Main Street	123-234-0202	1,000,000
T002	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-234-1111	500,000
T005		Cycle Repair	15 Cloud Street	123-234-2345	10,000

National Insurance File

NIBID	Legal Name	Trade Name	Address	Tel#	Emp.
E010	123 Bahamas	Goulet & St-Louis Limited	123 Main	123-234-0202	100
E020	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-235-1111	50

Pre-processed BIR File

BIRID	Legal Name	Trade Name	Address	Tel#	Revenue
T001	123BAHAMAS	GOULETSTLOUIS	123 Main	123-234-0202	1,000,000
T002	FRESHAIR	FRESHAIR	22 Cool	123-234-1111	500,000
T005		CYCLEREPAIR	15 Cloud	123-234-2345	10,000

Pre-processed National Insurance File

NIBID	Legal Name	Trade Name	Address	Tel#	Emp.
E010	123BAHAMAS	GOULETSTLOUIS	123 Main	123-234-0202	100
E020	FRESHAIR	FRESHAIR	22 Cool	123-235-1111	50

Deterministic Matching Example (Cont.)

Pre-processed BIR File

BIRID	Legal Name	Trade Name	Address	Tel#	Revenue
T001	123BAHAMAS	GOULETSTLOUIS	123 Main	123-234-0202	1,000,000
T002	FRESHAIR	FRESHAIR	22 Cool	123-234-1111	500,000
T005		CYCLEREPAIR	15 Cloud	123-234-2345	10,000

Pre-processed National Insurance File

NIBID	Legal Name	Trade Name	Address	Tel#	Emp.
E010	123BAHAMAS	GOULETSTLOUIS	123 Main	123-234-0202	100
E020	FRESHAIR	FRESHAIR	22 Cool	123-235-1111	50

Matched Result File #2

BIRID	NIBID	Legal Name	Trade Name	Address	Tel#	Revenue	Emp.
T001	E020	123 Bahamas	Goulet and St. Louis Ltd	123 Main Street	123-234-0202	1,000,000	1,000,000

Deterministic Matching Example (Cont.)

Pre-processed BIR File

BIRID	Legal Name	Trade Name	Address	Tel#	Revenue
T002	FRESHAIR	FRESHAIR	22 Cool	123-234-1111	500,000
T005		CYCLEREPAIR	15 Cloud	123-234-2345	10,000

Pre-processed National Insurance File

NIBID	Legal Name	Trade Name	Address	Tel#	Emp.
E020	FRESHAIR	FRESHAIR	22 Cool	123-235-1111	50

Matched Result File #3

BIRID	NIBID	Legal Name	Trade Name	Address	Tel#	Revenue	Emp.
T001	E020	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-234-1111	500,000	50

Deterministic Matching Example (Cont.) - Final result

BIRID	NIBID	Legal Name	Trade Name	Address	Tel#	Revenue	Emp.
T001	E010	123 Bahamas	Goulet and St-Louis Ltd	123 Main Street	123-234-0202	1,000,000	100
T002	E030	Fresh Air Inc.	Fresh Air Inc.	22 Cool Street	123-234-1111	500,000	50
T003	E040	Bahamas Hotel	Beach Front Hotel	5 Beach Street	123-234-2525	5,000,000	75
T004	E050	1202 Bahamas	Island Steak House	10 George Street	123-234-1010	2,000,000	20
T005			Cycle Repair	15 Cloud Street	123-234-2345	10,000	

Ongoing BR Update Processes Using Administrative Data

- Develop linkage process or procedures
  - Possible linkage variables: legal & trade names, addresses, telephone numbers, etc..
  - Final output: Concordance table containing all the unique administrative business IDs and the BR unique business ID

- Identify key variables to be updated or created on the BR
  - Tombstone (legal name, trade name, address, telephone #, etc..)
  - Size variables (revenue, employees etc..)
  - Classification variables (geography, industry, others)

Ongoing BR Update Processes Using Administrative Data (cont.)

- Develop business status rules or procedures (birth, active, inactive and ceased)
- Create “update date” and “source” variables for maintenance purposes

Resources

- https://uwaterloo.ca/networks-lab/blog/post/pre-processing-recordlinkage
  - University of Waterloo: complete tutorial of record linkage work flow using Python Record Linkage Toolkit. Great tutorial that summarizes most of the concepts of record linkage and is a great introduction to this useful library

- **An Overview of Selected International Business Record Linkage Programs**
  - https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2016001-eng.htm
  - Statistics Canada report summarizing record linkage activities including challenges and best practices for a number or NSOs

- Guidelines for Hierarchical Record Linkage (electronic handout, to be provided)
  - Provides a summary of record linkage concepts and as well as example implementation of a DMK
- Excel fuzzy match…

Conclusions

- Record linkage or data integration processes bring many advantages to a statistical organization
  - Reduce the need to contact businesses
  - Reduce costs
  - Replace existing survey modules
  - Develop new data series just based on BR data
  - Help in creating a complete BR
    - Overall population
    - Variable values
- However, need to:
  - Deploy initial effort to develop the processes and procedures to link records for the same businesses from various data sources
  - Apply the procedures at pre-determined frequency to ensure new records get linked and to attempt to link previously unlinked records.

You can contact the PRASC team at:

statcan.prasc-prasc.statcan@canada.ca