Linking Canadian Holders of U.S. Patents to Statistics Canada's Business Register 2000 to 2011

Paul Holness

Presented at the 2016 Methodology Symposium, Palais de Congrès, Gatineau, Quebec.

Wednesday, March 23, 2016.

Outline

1. Background
2. Objective
3. Record Linkage Framework
4. Data and Methods
5. Overall Match Results
6. Evaluation of Overall Match Quality
7. Limitations

Background

- Linkage of United States Patent and Trademark Office (USPTO) data to Statistics Canada's Business Register (BR)
- Integrate business micro-level data on patent frequency, patent class with firm characteristics such as employment, revenues, assets and liabilities
- Study period includes the years 2000 to 2011
- Provides a rich retrospective panel to support empirical studies on innovation and technical change in Canada

Objectives

- Seek an innovative and cost-effective solution to produce reliable data on the use of patents by Canadian enterprises
  - Harvest and re-use the distance calculations and the labeled corpus from an unsupervised approach to inform the linkage and gain efficiencies in the supervised approach
  - Integrate coding and classification modules in a single application
  - Implements statistical quality assurance methods, diagnostic measures and visualization techniques to evaluate linkage quality
  - Document a proof of concept that could be integrated into Statistics Canada's generalized systems to extend the tools available to users and help develop more robust linkage systems

Generic Record Linkage Framework

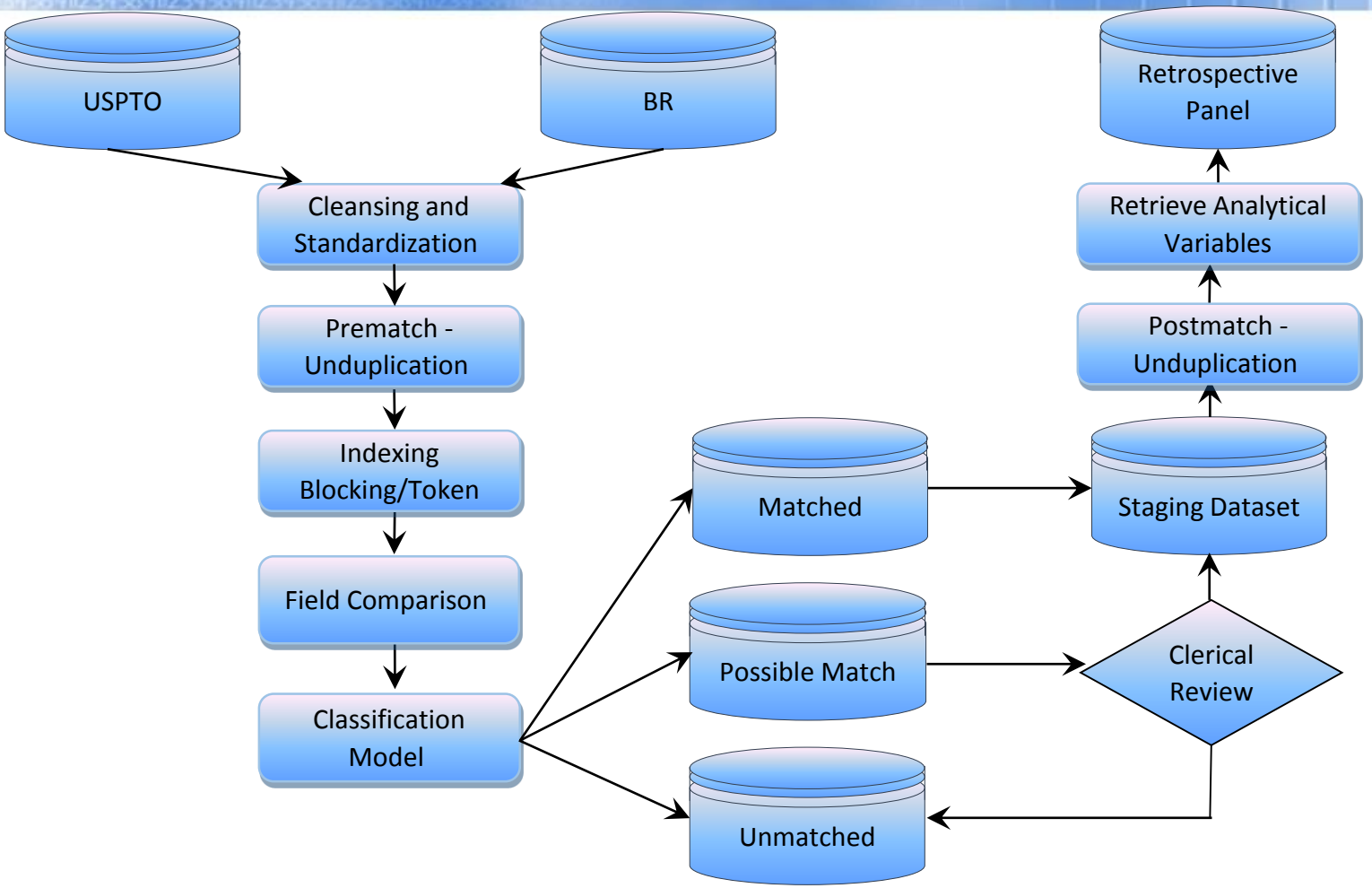


Figure 1: General Record Linkage Framework for USPTO and BR, 2000 to 2011

Data

- United States Patent and Trademark Office (USPTO) dataset of 41,619 of Canadian entities that received U.S. patents between 2000 and 2011

- The 41,619 patents are distributed among 14,162 distinct patent holders: 8,844 individuals (62.5%) and 5,318 institutions* (37.5%).

- Statistics Canada's Business Register had approximately 2.4 million statistical enterprises

*Institutions include businesses, post-secondary institutions of higher learning and government agencies.

Matching Fields

Primary Matching Fields

USPTO Fields	BR Fields	Description
Patent Year	Reference Year + 1, Jan. ed.	Reference Period
Assignee name	Legal/Operating Names	Enterprise Name
Province	Province	Geographic Jurisdiction
Secondary Matching Fields		
Vendor_DMKX	BR_Vendor_DMK	Phonetic fingerprint coded name
Clean_NameX	BR_Clean_Name	Name; no punctuation
Std_NameX	BR_Std_Name	Name; no stopwords such as inc. co.
Company NumberX	BR_Company Number	Incorporation Certification Number
K1X, K2X, K3X	K1, K2, K3	First, Second and Third word in name
CityX	BR_City	City; no punctuation
Postal CodeX	BR_PostalCode	Postal Code; no punctuation

Methods

- Unsupervised learning
  - Grouping data instances that are similar (near) to each other in one class or cluster and those instances that are very different (far away) from each other into different classes without prior knowledge of the relationships between the attributes
  - Examples include blocking or clustering records based on distance functions i.e. Generalized Edit Distance (GED)
- Supervised learning
  - A two-stage approach where an initial process is used to discover patterns that relate data attributes with match class
  - This a priori information is then used to predict the values of the target attribute in future data instances

Unsupervised Classification

- Deterministic linkage of USPTO/BR datasets
- Approximate string matching used to cluster the unlabeled datasets by selected matching fields
- Decomposed matching fields into tokens (words) and phonetically encoded strings
- Compared the values of USPTO/BR attributes
- Used match results to create comparison vectors to label USPTO/BR candidate pairs on an ordinal scale from zero (perfect match) to nine (unmatched)

Match Results, Unsupervised

Chart 1

Match rates by individual patent assignees, 2000 to 2011

Percentage

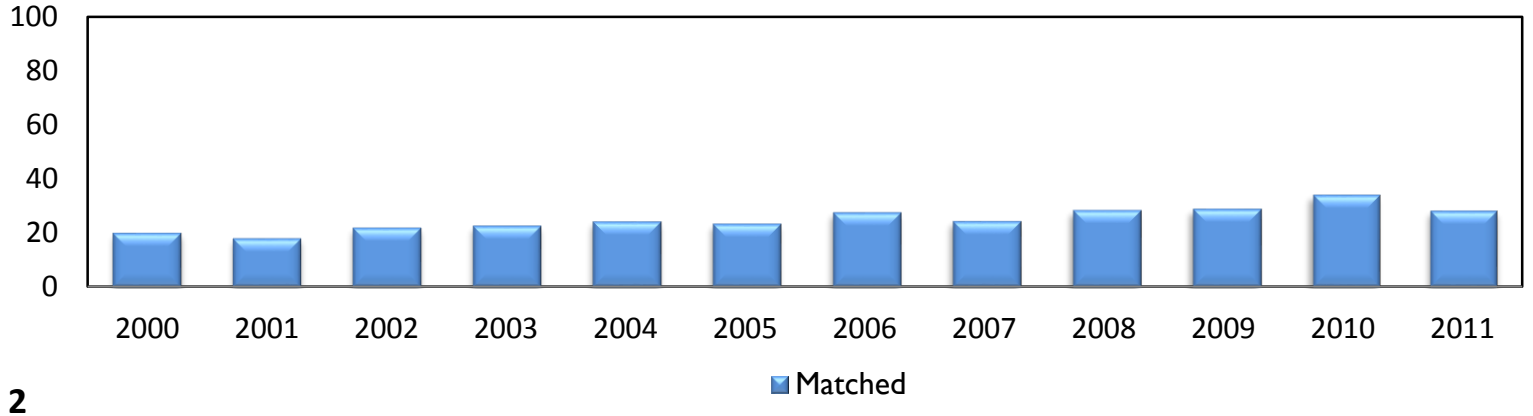
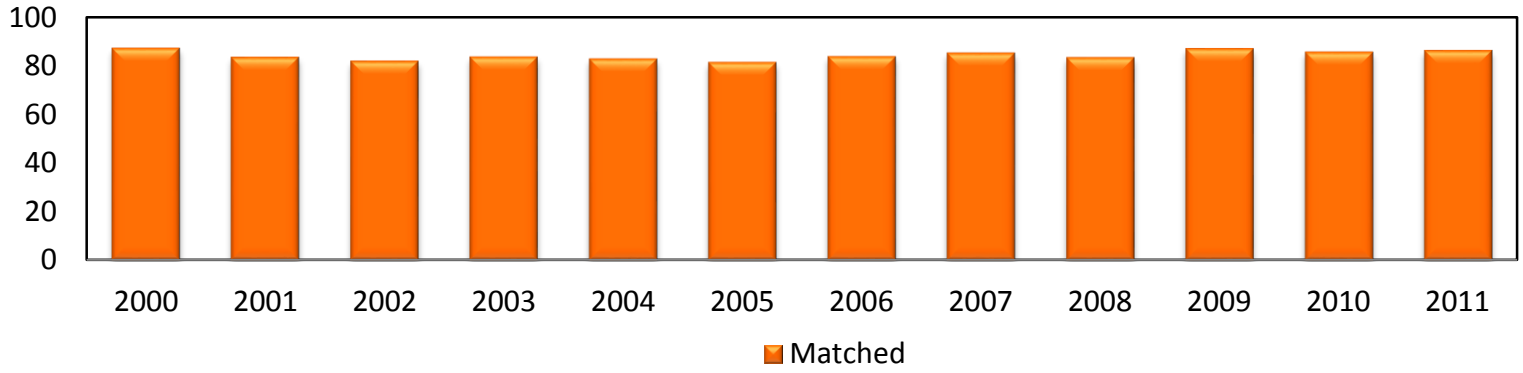


Chart 2

Match rates by institutional patent assignees, 2000 to 2011

Percentage



Matching, Supervised Classification

- Harvest distance metrics and labels from unsupervised process and use them as parameters along with other attributes to train a multinomial regression model to infer match classes
- Partition data frame into two disjoint datasets: Training (75.0%) and Testing (25.0%) datasets
- Evaluate model results using chi-square significance test as statistical evidence as whether there is a relationship between the log odds of the match score and the combination of Generalized Edit Distance (GED) measures
- Create a Cartesian product** of the unmatched USPTO/BR records with weighted mean GED values ≤ threshold value of 7
- Score the Cartesian product to predict the probable match class

**Cartesian Product is of A × B is the set of all ordered pairs (a, b) where a ∈ A and b ∈ B.

Model and Feature Selection

Multinomial Logistic Regression Model

Specified Model

$$\log[p(c \le j)] = \alpha_j + \beta_1 (GEDName_j) + \beta_2 (GEDCity_j) + \beta_3 (GEDPostalCode_j) + \beta_4 (Length(USPTOClean\_Name)_j) + \beta_5 (Length(BRClean\_Name)_j) + \beta_6 (Length(USPTOClean\_Name)_j * Length(BRClean\_Name)_j) + e_j \quad (2)$$

e_j Random error terms

where c = the match class with an ordinal range from (0 to 9) and the subscript j denotes the institution

Model Diagnostics

Table 1. Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	9	1	-24.8477	236.2	0.0111	0.9162
Intercept	8	1	-5.2236	0.2557	417.4869	<.0001
Intercept	6	1	-3.8817	0.2344	274.3457	<.0001
Intercept	5	1	-3.8328	0.2338	268.7112	<.0001
Intercept	3	1	-2.4193	0.2226	118.1610	<.0001
Intercept	2	1	-0.4872	0.2170	5.0408	0.0248
Intercept	1	1	-0.4769	0.2170	4.8304	0.0280
RelScoreCompName		1	0.6595	0.0234	795.3086	<.0001
RelScoreCompCity		1	0.0384	0.00193	395.5514	<.0001
LengthCleanNameX		1	-0.1322	0.0142	86.0800	<.0001
LengthCleanName		1	0.0865	0.0131	43.8193	<.0001
LengthCle*LengthClea		1	0.00138	0.000322	18.3293	<.0001

R-Square	0.7219	Max-rescaled R-Square	0.7602

Source: USPTO, Author's calculations

Model Outcomes

- The primary outcome was the response match class, with ordinal values ranging from zero to nine
- The model classified the USPTO/BR candidate pairs according to the strength of the relationship between the model covariates and the response class
- The diagnostic measures that follow, show that the model effectively and reliably related the logits to the response class

Evaluating Model Results

Table 2. Metrics for logistic classification model

		True Condition: Matched	True Condition: Unmatched	
Inferred Condition	Matched	T^+ 171	F^+ 68	PPV= $(T^+)/(T^+ + F^+)$ 71.5%
	Unmatched	F^- 24	T^- 538	NPV= $(T^-)/(T^- + F^-)$ 95.7%
		MMR =$(F^-)/(T^+ + F^-)$ (Type I error) Sensitivity=1-MMR 87.6%	FMR=$(F^+)/((F^+)+(T^-))$ (Type II error) Specificity=1-FMR 88.8%	

where: True positive (T^+), true negative (T^-), false positive (F^+), false negative (F^-), positive predictive value (PPV), negative Predictive Value = (NPV), missed match rate (MMR), false match rate (FMR)

Overall Match Results

Table 3. Distribution of matched institutions by matching variables

Response match class	Matching variables	Frequency	Percentage	Cumulative frequency	Cumulative percentage
0	Clean_Name, City, Province	2340	44.00	2340	44.00
1	Clean_Name, City	8	0.15	2348	44.15
2	Clean_Name, Province	1207	22.7	3555	66.85
3	Std_Name, City, Province, RelScoreCompName < 10	458	8.61	4013	75.46
4	Clean_Name, PostalCode	0	0	0	0
5	Company Number, IncorporationJurisdiction	11	0.21	4024	75.67
6	Vendor_DMK, First word, Second word, Third word, City, Province RelScoreCompName < 10	262	4.93	4286	80.59
7	Multinomial logit followed by clerical review	195	3.67	4481	84.26
8	Clerical review (manual)	24	0.45	4505	84.71
9	Unmatched	813	15.29	5318	100.00

Source: USPTO, Author's calculations

Overall Evaluation of Match Quality

Table 4. Metrics for Overall Match results

		True Condition		
		Matched	Unmatched	
Inferred Condition	Matched	T^+ 340	F^+ 0	PPV= $(T^+)/(T^+ + F^+)$ 100.0%
	Unmatched	F^- 30	T^- 31	NPV= $(T^-)/(T^- + F^-)$ 50.8%
		MMR =$(F^-)/(T^+ + F^-)$ (Type I error) Sensitivity=1-MMR 91.9%	FMR=(F+)/((F+)+(T-)) (Type II error) Specificity=1-FMR 100.0%	

where: True positive (T^+), true negative (T^-), false positive (F^+), false negative (F^-), positive predictive value (PPV), negative Predictive Value = (NPV), missed match rate (MMR), false match rate (FMR)

Limitations

- Key assumptions:
  - The Gold Standard: Where the clerical review process manually labeled USPTO/BR candidate pairs are 100.0% correct.
  - The selected sample used in the training dataset and in the evaluation dataset is sufficiently representative of the overall USPTO dataset.
- Violations of these assumptions could lead to potential bias in estimates regarding the impact of the patent on the assignees that matched entities in the BR.

Potential for Bias

- The manual review process was conducted on a top-down basis to maximize the coverage of patents and optimize the use of available resources
- The results of the chi-square test (see below) of independence confirmed the presence of bias
- The number of patents held by the assignee influences the match outcome of the USPTO/BR candidate pairs Therefore, estimates generated from the resulting dataset are potentially subject to selection bias

Table 5. Statistics for Table of NoOfPatents by MatchGroup

Statistic	DF	Value	Prob
Chi-Square	98	133.7798	0.0095

Contact Information

Paul Holness

International Cooperation and Corporate Statistical Methods Division

Statistics Canada

Office: (613) 864-0176

Mobile: (613) 866-0367

Paul.Holness@canada.ca