Statistics Canada
www.statcan.gc.ca

Couplage des détenteurs canadiens de brevets américains au Registre des entreprises de Statistique Canada 2000 à 2011

Paul Holness

Présenté au Symposium international de 2016 sur les questions de méthodologie,
Palais de Congrès, Gatineau, Québec

Le mercredi 23 mars 2016

Aperçu

1. Contexte
2. Objectif
3. Cadre de couplage des enregistrements
4. Données et méthodes
5. Résultats globaux de l’appariement
6. Évaluation de la qualité globale de l’appariement
7. Limitations

Contexte

- Couplage des données du United States Patent and Trademark Office (USPTO) au Registre des entreprises (RE) de Statistique Canada
- Intégration des microdonnées des entreprises sur la fréquence et la catégorie des brevets aux caractéristiques des entreprises telles que l'emploi, les revenus, les actifs et les passifs
- Étude couvrant la période de 2000 à 2011
- Panel rétrospectif riche à l'appui des études empiriques sur l'innovation et les progrès techniques au Canada

Objectifs

- Recherche d'une solution novatrice et rentable qui permettrait de produire des données fiables sur l'utilisation des brevets par les entreprises canadiennes
  - Collecte et réutilisation non supervisées des calculs des distances et du corpus étiqueté afin d'éclairer le couplage et de réaliser des gains d'efficience dans le cadre d'une approche supervisée
  - Intégration des modules de codage et de classification dans une seule application
  - Mise en œuvre de méthodes statistiques d'assurance de la qualité, de mesures diagnostiques et de techniques de visualisation afin d'évaluer la qualité des couplages
  - Documentation d'une validation de principe qui pourrait être intégrée aux systèmes généralisés de Statistique Canada afin de développer davantage les outils mis à la disposition des utilisateurs et d'aider à mettre au point des systèmes de couplage plus robustes

Cadre générique de couplage des enregistrements

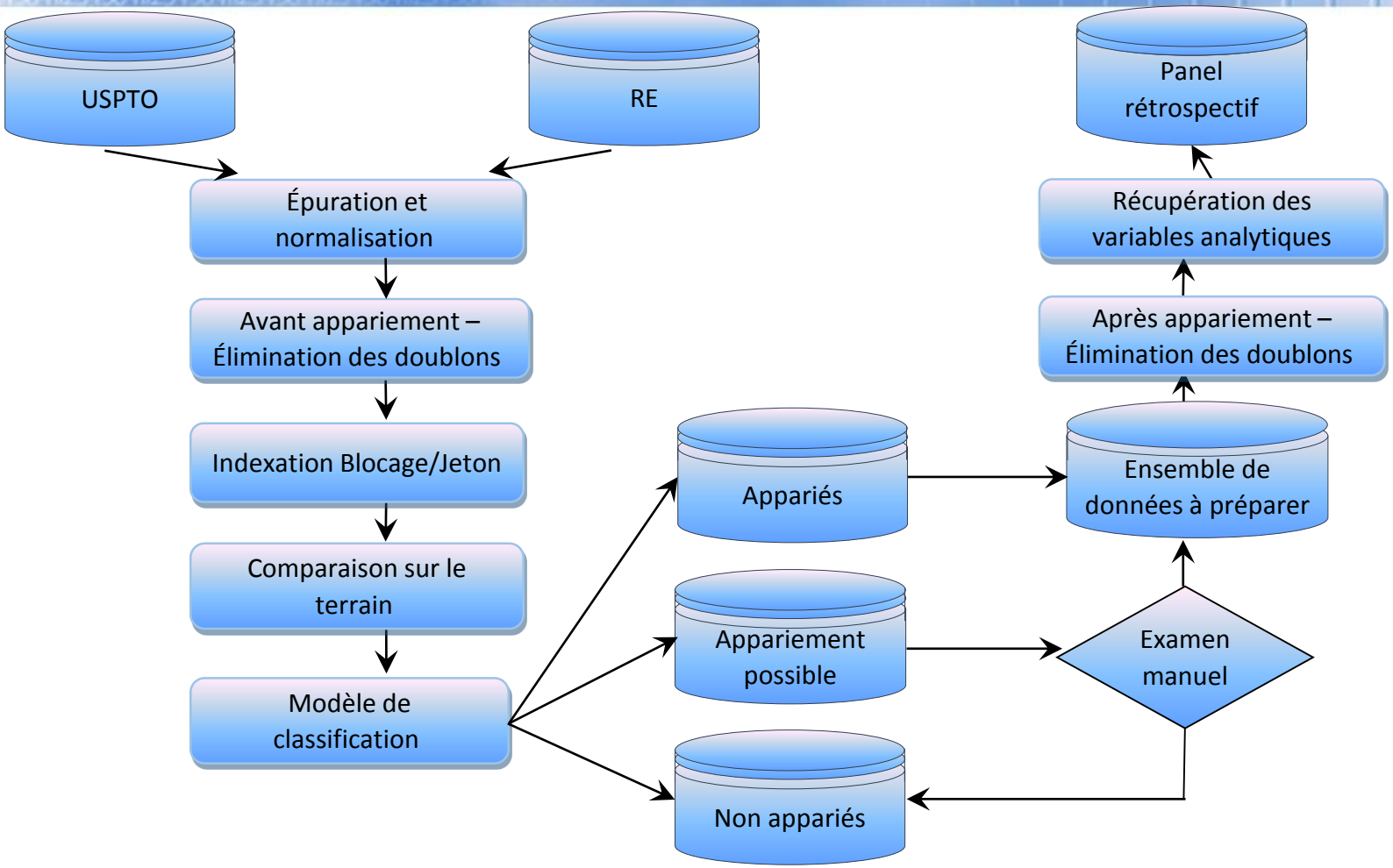


Figure 1 : Cadre générique de couplage des enregistrements pour l'USPTO et le RE, 2000 à 2011

Données

- Ensemble de données du United States Patent and Trademark Office (USPTO) sur 41 619 brevets américains accordés à des entités canadiennes entre 2000 et 2011

- Les 41 619 brevets sont répartis entre 14 162 détenteurs distincts, soit 8 844 personnes (62,5 %) et 5 318 institutions* (37,5 %).

- Le Registre des entreprises de Statistique Canada incluait environ 2,4 millions d'entreprises statistiques.

*Les institutions englobent les entreprises, les établissements d'enseignement postsecondaire et les organismes gouvernementaux.

.

Champs d'appariement

Champs d'appariement primaires

Champs de l'USPTO	Champs du RE	Description
Année d'octroi du brevet	Année de référence + 1, édition de janvier	Période de référence
Nom du cessionnaire	Dénomination sociale/ Nom commercial	Nom de l'entreprise
Province	Province	Territoire géographique

Champs d'appariement secondaires

Vendor_DMKX	BR_Vendor_DMK	Nom codé phonétiquement
Clean_NameX	BR_Clean_Name	Nom; sans punctuation
Std_NameX	BR_Std_Name	Nom; sans mot vide comme inc. ou co.
Company NumberX	BR_Company Number	Numéro de certificat de constitution
K1X, K2X, K3X	K1, K2, K3	Premier, deuxième et troisième mots du nom
CityX	BR_City	Ville; sans punctuation
Postal CodeX	BR_PostalCode	Code postal; sans punctuation

Méthodes

- Apprentissage non supervisé
  - Regroupement des instances de données semblables (proches) dans une catégorie ou une grappe et des instances très différentes (éloignées) dans des catégories différentes sans connaissance préalable des relations entre les attributs
  - Les exemples comprennent le blocage ou le regroupement en grappes des enregistrements fondé sur les fonctions de distance (distance d'édition généralisée)
- Apprentissage supervisé
  - Approche en deux étapes où un processus initial est utilisé pour repérer les tendances qui mettent les attributs des données en relation avec la catégorie d'appariement
  - Cette information a priori est ensuite utilisée pour prédire les valeurs de l'attribut cible dans les futures instances de données

Classification non supervisée

- Couplage déterministe des ensembles de données de l'USPTO et du RE
- Utilisation de l'appariement approximatif de chaînes pour regrouper les ensembles de données non étiquetées en fonction de champs d'appariement choisis
- Décomposition des champs d'appariement en jetons (mots) et en chaînes codées phonétiquement
- Comparaison des valeurs des attributs de l'USPTO et du RE
- Utilisation des résultats d'appariement pour créer des vecteurs de comparaison afin de coter les paires candidates de l'USPTO et du RE sur une échelle ordinale de zéro (appariement parfait) à neuf (non appariées)

Résultats d'appariement, non-supervisé

Chart 1

Match rates by individual patent assignees, 2000 to 2011

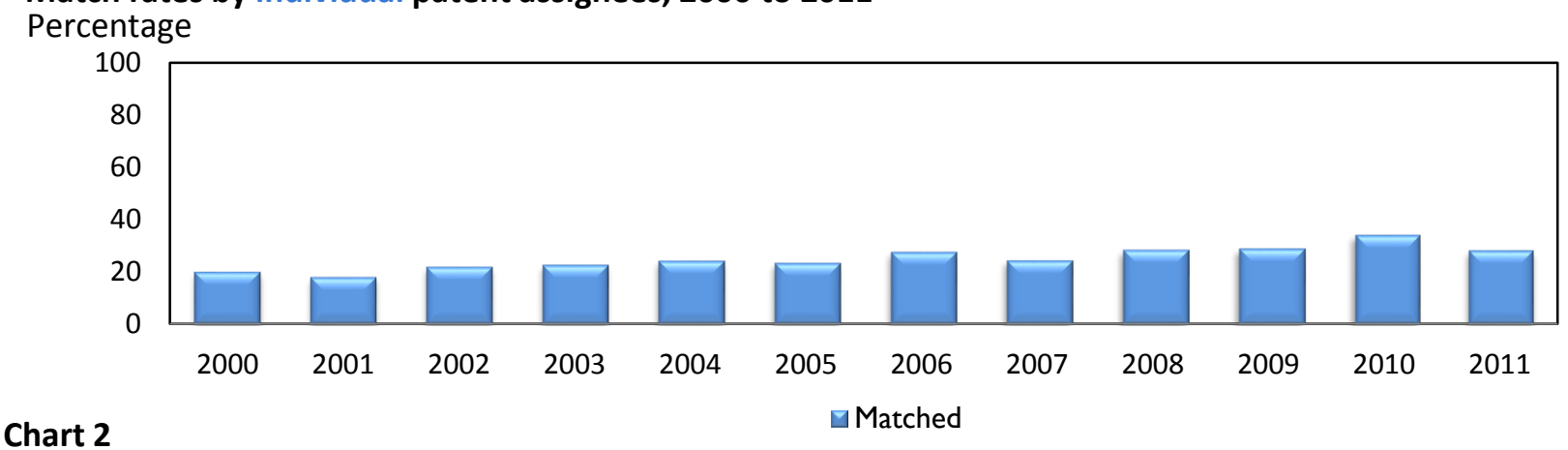
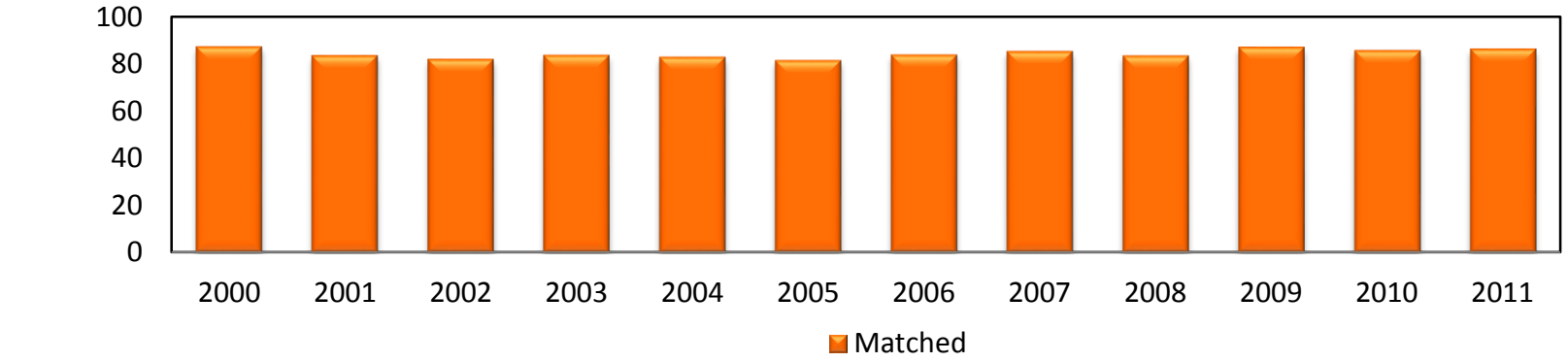


Chart 2

Match rates by institutional patent assignees, 2000 to 2011



Appariement, classification non supervisée

- Collecte des mesures de distance et des étiquettes du processus non supervisé à utiliser comme paramètres avec d'autres attributs pour apprendre à un modèle de régression multinomiale à déduire les catégories d'appariement
- Partition du cadre de données en deux ensembles de données disjoints : Formation (75,0 %) et Évaluation (25,0 %)
- Évaluation des résultats du modèle au moyen du test d'hypothèse du chi-deux comme preuve statistique d'une relation entre le logarithme du rapport de cotes du score d'appariement et la combinaison des mesures de distance d'édition généralisée (DEG)
- Création d'un produit cartésien** des enregistrements non appariés de l'USPTO et du RE dont les valeurs DEG moyennes pondérées sont inférieures ou égales à la valeur seuil de 7
- Évaluation de la capacité du produit cartésien à prédire la catégorie d'appariement probable

**Le produit cartésien de $A \times B$ est l'ensemble de paires ordonnées (a, b) où $a \in A$ et $b \in B$.

Sélection des modèles et des caractéristiques

Modèle de régression logistique multinomiale

Modèle spécifié

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

e_j *Termes d'erreur aléatoire*

où c est la catégorie d'appariement dont l'échelle ordinale s'étend de 0 à 9 et l'indice j dénote l'institution

Diagnostics du modèle

Tableau 1. Analyse des estimations du maximum de vraisemblance

Paramètre		DF	Estimation	Erreur type	Wald Chi-carré	Pr > Chi-carré
Ordonnée à l'origine	9	1	-24,8477	236,2	0,0111	0,9162
Ordonnée à l'origine	8	1	-5,2236	0,2557	417,4869	<0,0001
Ordonnée à l'origine	6	1	-3,8817	0,2344	274,3457	<0,0001
Ordonnée à l'origine	5	1	-3,8328	0,2338	268,7112	<0,0001
Ordonnée à l'origine	3	1	-2,4193	0,2226	118,1610	<0,0001
Ordonnée à l'origine	2	1	-0,4872	0,2170	5,0408	0,0248
Ordonnée à l'origine	1	1	-0,4769	0,2170	4,8304	0,0280
RelScoreCompName		1	0,6595	0,0234	795,3086	<0,0001
RelScoreCompCity		1	0,0384	0,00193	395,5514	<0,0001
LengthCleanNameX		1	-0,1322	0,0142	86,0800	<0,0001
LengthCleanName		1	0,0865	0,0131	43,8193	<0,0001
LengthCle*LengthClea		1	0,00138	0,000322	18,3293	<0,0001

R-carré	0,7219	R-carré remis à l'échelle max.	0,7602

Source : USPTO, calculs de l'auteur

Résultats du modèle

- Le résultat principal était la catégorie d'appariement des réponses, les valeurs ordinales allant de zéro à neuf.
- Le modèle classait les paires candidates de l'USPTO et du RE selon la force de la relation entre les covariables du modèle et la catégorie de réponse.
- Les mesures diagnostiques qui suivent montrent que le modèle a établi des liens entre les logits et la catégorie de réponse de manière fiable et efficace.

Évaluation des résultats du modèle

Tableau 2. Mesures pour le modèle de classification logistique

		Condition réelle		
		Apparié	Non apparié	
Condition inférée	Apparié	V^+ 171	F^+ 68	VPP= $(V^+)/(V^+ + F^+)$ 71,5 %
	Non apparié	F^- 24	V^- 538	VPN= $(V^-)/(V^- + F^-)$ 95,7 %
		TAM =$(F^-)/(V^+ + F^-)$ (Erreur de type I) Sensibilité=1-TAM 87,6 %	TFA=$(F^+)/((F^+)+(V^-))$ (Erreur de type II) Spécificité=1-TFA 88,8 %	

où : vrai positif (V^+), vrai négatif (V^-), faux positif (F^+), faux négatif (F^-), valeur prédictive positive (VPP), valeur prédictive négative = (VPN), taux d'appariements manqués (TAM), taux de faux appariements (TFA)

Résultats d'appariement globaux

Tableau 3. Répartition des institutions appariées selon les variables d'appariement

Catégorie d'appariement des réponses	Variables d'appariement	Fréquence	%	Fréquence cumulative	% cumulatif
0	Clean_Name, ville, province	2340	44,00	2340	44,00
1	Clean_Name, ville	8	0,15	2348	44,15
2	Clean_Name, province	1207	22,7	3555	66,85
3	Std_Name, ville, province, RelScoreCompName < 10	458	8,61	4013	75,46
4	Clean_Name, PostalCode	0	0	0	0
5	Numéro de société, IncorporationJurisdiction	11	0,21	4024	75,67
6	Vendor_DMK, premier mot, deuxième mot, troisième mot, ville, province RelScoreCompName < 10	262	4,93	4286	80,59
7	Logit multinomial suivi par un examen manuel	195	3,67	4481	84,26
8	Examen manuel	24	0,45	4505	84,71
9	Non apparié	813	15,29	5318	100,00

Source : USPTO, calculs de l'auteur

Évaluation globale de la qualité de l'appariement

Tableau 4. Mesures pour les résultats d'appariement globaux

		Condition réelle		
		Apparié	Non apparié	
Condition inférée	Apparié	V^+ 340	F^+ 0	VPP= $(V^+)/(V^+ + F^+)$ 100,0%
	Non apparié	F^- 30	V^- 31	VPN= $(V^-)/(V^- + F^-)$ 50,8 %
		TAM =$(F^-)/(V^+ + F^-)$ (Erreur de type I) Sensibilité=1-TAM 91,9 %	TFA=(F+)/((F+)+(T-)) (Erreur de type II) Spécificité=1-TFA 100,0 %	

où : vrai positif (V^+), vrai négatif (V^-), faux positif (F^+), faux négatif (F^-), valeur prédictive positive (VPP), valeur prédictive négative = (VPN), taux d'appariements manqués (TAM), taux de faux appariements (TFA)

Limitations

- Hypothèses clés :
  - La norme de référence : Lorsque les paires candidates de l'USPTO et du RE évaluées durant le processus d'examen manuel sont correctes à 100,0 %.
  - L'échantillon sélectionné utilisé dans les ensembles de données de formation et d'évaluation est suffisamment représentatif de l'ensemble de données global de l'USPTO.
- La violation de ces hypothèses pourrait introduire un biais potentiel dans les estimations concernant l'impact du brevet sur les cessionnaires appariés à des entités du RE.

Biais potentiel

- L'examen manuel a été effectué selon la méthode descendante afin de maximiser la couverture des brevets et d'optimiser l'utilisation des ressources disponibles.
- Les résultats du test chi carré (voir ci-dessous) de l'indépendance ont confirmé la présence d'un biais.
- Le nombre de brevets détenus par le cessionnaire influence le résultats de l'appariement des paires candidates de l'USPTO et du RE. En conséquence, les estimations générées à partir de l'ensemble de données résultant pourraient être sujettes à un biais de sélection.

Tableau 5. Statistiques pour le tableau du nombre de brevets par groupe d'appariement

Statistique	DF	Valeur	Probabilité
Chi carré	98	133,7798	0,0095

Coordonnées

Paul Holness

Division de la coopération internationale et des méthodes statistiques institutionnelles

Statistique Canada

Bureau : 613-864-0176

Mobile : 613-866-0367

Paul.Holness@canada.ca