

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Remaniement : processus d'identification des partenariats T1

par Shaundon Holmstrom

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Remaniement : processus d'identification des partenariats T1

Shaundon Holmstrom¹

Résumé

Au Canada, le formulaire d'impôt T1 est utilisé pour la déclaration du revenu personnel, qu'il ait été gagné en tant que personne employée ou travailleur autonome. Le revenu provenant d'un travail autonome, ou « revenu d'entreprise T1 », est déclaré par des entreprises détenues par une seule personne ou en partenariat. Un partenariat T1 comprend deux entités juridiques ou plus qui font une déclaration conjointe pour une entreprise partagée. Les données d'entreprises T1 sont reçues sous forme de déclarations individuelles; ce qui signifie que les données pour les partenariats sont reçues séparément pour chaque partenaire. Le couplage d'enregistrements interne dans la base de données T1 sur les entreprises est effectué pour identifier les partenariats et prévenir le surdénombrement dans la population finale des entreprises T1. Ce nouveau processus d'identification des partenariats T1 tire parti de nouveaux algorithmes, comme le groupement numérique DBSCAN par appariement flou, pour déterminer les couplages internes. La théorie des graphes est utilisée pour bâtir la liste des partenariats à partir des paires de rangées relevées dans le processus de couplage.

Mots clés : Couplage d'enregistrements; DBSCAN; théorie des graphes; données administratives.

1. Introduction

1.1 Données T1

Le formulaire d'impôt T1 est utilisé pour la déclaration du revenu personnel, y compris celui gagné par des travailleurs autonomes. Ces derniers sont considérés comme des entreprises non constituées en société et produisent leur déclaration de revenus au moyen de formulaires et d'annexes d'impôt sur le revenu des particuliers. La location d'une propriété, la vente de produits d'artisanat dans un magasin personnel, et la vente de fruits et légumes cultivés dans une terre agricole appartenant à des particuliers sont quelques exemples d'activités réalisées par des entreprises T1.

Bien que les activités commerciales des entreprises T1 ne soient pas exclusives aux entités non constituées en société, de nombreux propriétaires d'entreprises choisissent de demeurer non constitués en société pour diverses raisons, comme une plus grande souplesse sur le plan des affaires, une réduction du fardeau administratif et une réduction des frais de déclaration.

1.2 Organisation des entreprises T1

Les entreprises T1 sont organisées en entreprises individuelles et en partenariats. Une entreprise individuelle consiste en une seule déclaration d'un seul propriétaire d'entreprise. Les entreprises T1 peuvent produire des déclarations en tant qu'entreprises détenue en partenariat, c'est-à-dire une entreprise détenue par au moins deux partenaires qui déclarent une entreprise unique, partagée et non constituée en société.

Bien qu'il existe des partenariats entre d'autres entités juridiques (p. ex. autres partenariats, entreprises constituées en société, organismes de bienfaisance), elles sont considérées comme hors du champ pour le processus relatif aux partenariats T1.

¹ Shaundon Holmstrom, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, shaundon.holmstrom@statcan.gc.ca

1.3 Déclaration de revenus des entreprises détenues en partenariat

Les partenaires au sein d'une entreprise détenue en partenariat ont pour directive de soumettre une déclaration contenant le total des activités financières de l'entreprise. Par conséquent, un partenaire ne doit pas soumettre ce qu'il a dépensé individuellement en dépenses de services publics, mais plutôt le coût total de toutes les dépenses de services publics de l'entreprise au cours d'une période de déclaration donnée. En l'absence de tout autre renseignement, les états financiers des partenaires devraient être semblables.

Les renseignements sur l'entreprise, comme le nom de l'entreprise lorsqu'il y a lieu, l'adresse de l'entreprise, le code postal de l'entreprise, et le code du Système de classification des industries de l'Amérique du Nord (SCIAN), devraient correspondre, ou être très semblables entre les déclarations individuelles. Les partenariats peuvent également enregistrer un numéro d'entreprise; ce qui fournirait une clé unique pour identifier l'entreprise.

Les partenaires peuvent également déclarer leur part en pourcentage des intérêts dans l'entreprise, ainsi que le nom d'autres partenaires au sein de l'entreprise détenue en partenariat.

1.4 Défi

Les déclarations T1 sont reçues individuellement de chaque propriétaire d'entreprise. Pour les entreprises individuelles, le dépôt est traité comme une déclaration complète d'entreprise. Toutefois, pour les entreprises détenues en partenariat, la déclaration est reçue sous forme de multiples déclarations individuelles, soit une par propriétaire de l'entreprise. La législation relative au T1 n'exige pas que les déclarants soumettent une clé unique pour identifier leur entreprise; cela a donc une incidence sur l'identification du partenariat.

L'une des principales composantes du traitement des données d'entreprises T1 consiste à créer une population finale d'entreprises T1 à des fins statistiques. Pour ce faire, un couplage d'enregistrements interne est nécessaire afin de créer une seule déclaration pour une entreprise détenue en partenariat à partir des états individuels déposés par chaque partenaire. Le défaut d'identifier tous les partenaires se traduirait par un surdénombrement de la population d'entreprises T1 : des doublons du partenariat seraient présents dans la population d'entreprises. Le fait d'identifier de façon erronée une entreprise détenue par une seule personne en tant qu'entreprise détenue en partenariat se traduirait par un sous-dénombrement de la population d'entreprises : les renseignements d'une ou de plusieurs entreprises individuelles ne seraient pas inclus dans la population d'entreprises T1.

1.5 Raison du remaniement

Le processus original d'identification des partenariats T1 est utilisé depuis plus de 15 ans. Depuis son introduction, les besoins des utilisateurs de données ont beaucoup évolué, tout comme les technologies auxquelles nous avons accès, de sorte qu'un remaniement du système de traitement s'impose.

La transition vers un système entièrement nouveau a permis l'intégration de méthodes contemporaines de science des données, de l'apprentissage automatique et de structures de données complexes faisant partie intégrante de la nouvelle méthodologie.

2. Couplage des enregistrements T1

2.1 Étapes du processus de couplage

Le nouveau processus de couplage comporte trois principales étapes :

1. regroupement des résultats des colonnes financières;
2. création d'un ensemble de liens par paire dans chaque regroupement créé en 1.)
 - a. utilisation de méthodes de comparaison de chaînes pour déterminer les colonnes de correspondance entre les paires de rangées;

- b. conservation des paires de rangées qui satisfont au critère d'appariement minimal;
3. création d'un graphique à l'aide de l'ensemble de paires de rangées conservé en 2.)

2.2 Regroupement financier

Les entreprises détenues en partenariat devant, idéalement, fournir des valeurs similaires, voire identiques, dans les champs financiers. L'ensemble des liens potentiels se fait par regroupement des champs financiers déclarés.

Le processus relatif aux partenariats utilise un sous-ensemble de champs financiers fréquemment déclarés et dont l'écart est faible. Les champs ont été choisis à l'aide d'une combinaison d'expertise spécialisée, ainsi qu'en déterminant les champs financiers présentant la plus forte fréquence d'écart nulle entre les partenariats identifiés dans le cadre de l'ancien processus relatif aux partenariats (p. ex. total des revenus, total des dépenses, achats/coûts du matériel, téléphone et services publics). Les champs financiers propres à la personne ou présentant une forte probabilité de déclaration au niveau personnel plutôt qu'au niveau de l'entreprise ont été exclus du processus (p. ex. frais de déplacement, revenu après impôt).

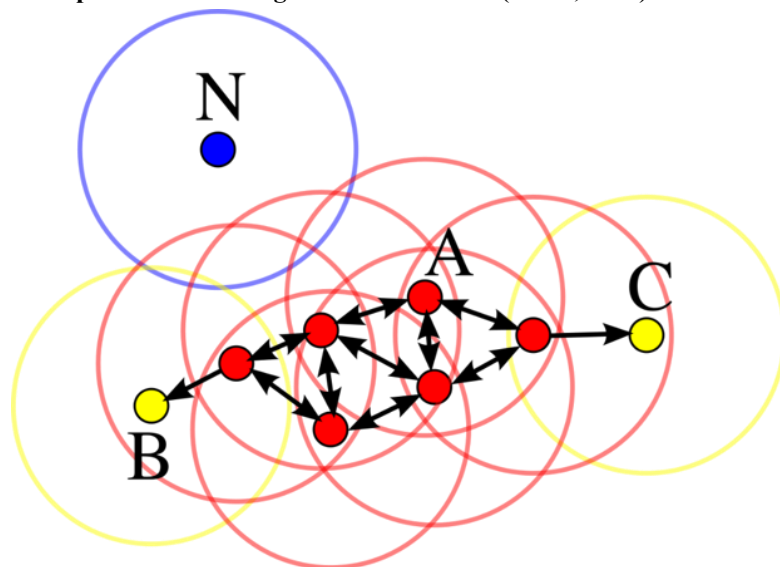
2.2.1 Algorithme de groupement : DBSCAN

DBSCAN est un algorithme de regroupement numérique conçu pour trouver des regroupements dans un grand ensemble de données bruyantes (Ester et coll., 1996). L'algorithme DBSCAN calcule une mesure de distance entre deux points et choisit de regrouper ces points s'ils se trouvent à l'intérieur d'un Epsilon (ϵ) de borne d'erreur donné l'un par rapport à l'autre. Les points qui ne se trouvent pas à ϵ d'un autre point ne sont pas regroupés.

La méthode DBSCAN peut créer des regroupements de taille et de forme arbitraires et n'exige pas que l'utilisateur précise au préalable le nombre de regroupements (Ester et coll., 1996). Les seuls paramètres requis fournis par l'utilisateur sont la fonction de distance et la distance autorisée entre deux points (ϵ).

En raison de la nature des données T1, il est impossible de savoir au préalable exactement le nombre d'entreprises détenues en partenariat au cours d'une année donnée. Le fait de ne pas avoir à déterminer au préalable le nombre de regroupements, et la capacité de DBSCAN à gérer les bruits qui ne devraient pas être liés sont deux propriétés très attrayantes pour régler ce problème. Les points de bruit sont présumés être des entreprises individuelles, en l'absence de tout autre critère.

Figure 2.2.1-1
Exemple illustré de l'algorithme DBSCAN (Chire, 2011)



A (rouge) : points centraux; connectés à deux autres points ou plus.

B, C (jaune) : points périphériques; connectés à un point central.
N : points de bruit; situés au-delà de ε d'un autre point.

2.3 Comparaison de paires de rangées dans un regroupement

Une fois un ensemble de rangées attribué à un même regroupement, toutes les comparaisons par paires sont effectuées à l'intérieur de ce regroupement. Les scores de similitude des comparaisons de chaînes sont produits pour le nom et l'adresse de l'entreprise. D'autres champs, comme le numéro d'entreprise, le code SCIAN et le nom du partenaire, sont comparés pour obtenir une correspondance exacte.

2.3.1 Comparaisons entre les colonnes

Des coquilles et différents formats de déclaration figurent dans les champs d'adresse et de nom de l'entreprise. Les scores de similitude des chaînes sont utilisés pour tenir compte des incohérences déclarées. On a déterminé que d'autres champs étaient trop courts ou trop précis pour qu'il soit nécessaire de tenir compte des coquilles ou d'autres erreurs, et seules des correspondances exactes ont été permises pour les champs autres que le nom ou l'adresse.

La similarité de Jaro-Winkler a été choisie comme mesure de la distance pour les chaînes, puisqu'elle traite les longueurs de chaîne ne correspondant pas et a fait ses preuves dans le couplage d'enregistrements (Winkler, 1990). Son rendement sur le plan des données T1 a été évalué et jugé acceptable.

2.3.2 Critères d'appariement minimaux

Une fois que toutes les colonnes d'appariement et de non-appariement déterminées, un ensemble de combinaisons de colonnes d'appariement est choisi comme critère d'appariement minimal. Nous conservons des paires de rangées qui répondent à au moins une des combinaisons minimales de colonnes d'appariement.

Par exemple, un appariement est considéré comme adéquat si l'analyse du regroupement, le nom de l'entreprise et le code postal de l'entreprise correspondent entre deux rangées.

2.4 Représentation graphique des paires de rangées conservées

Un graphique est une structure de données composée de nœuds et d'arêtes qui les relie. Aux fins du présent projet, les nœuds du graphique représentent des rangées de l'ensemble de données, et les arêtes sont l'ensemble de paires de rangées acceptées.

La liste restante de paires de rangées satisfaisant aux critères d'appariement minimal est transmise à une bibliothèque de graphiques (Hagburg et coll., 2008). Un graphique est créé à partir des paires de rangées, et des sous-graphiques sont extraits. Chaque ensemble de nœuds joints dans le produit correspond à un partenariat finalisé.

3. Exemple fictif détaillant le processus relatif aux entreprises détenues en partenariat

3.1 Regroupement financier

Supposons que nous avons une population des six (6) déclarations T1 suivantes. Au cours de la première étape, nous ne tenons compte d'aucun autre renseignement sur l'entreprise et nous nous concentrons uniquement sur le regroupement au moyen des rapports financiers. Nous exécutons DBSCAN avec un Epsilon (ε) de borne d'erreur adapté au domaine de données. À titre d'illustration, $\varepsilon = 200$ est utilisé dans l'exemple suivant.

Tableau 3.1-1**Exemple de données : DBSCAN sur les données financières T1**

Code ID	Nom du déclarant	Revenus totaux	Dépenses totales	Coût des services publics	Revenus de location	Regroupement DBSCAN
132_1	John Smith	50 000	1 200	0	0	1
132_2	John Smith	50 100	1 150	0	0	1
865_1	Alice Jones	50 100	1 150	0	0	1
611_1	Eliza Rochefort	25 000	1 500	900	25 000	2
376_1	Emilie Rochefort	25 000	1 500	900	25 000	2
978_3	Marie Dupoint	30 000	800	750	30 000	-1

3.2 Appariement des paires dans un regroupement

Les colonnes d'identification de l'entreprise sont comparées grâce aux regroupements créés par DBSCAN, dont le profil est illustré au tableau 3.2-1.

Intuitivement, il est facile pour un être humain de déterminer que la deuxième déclaration de John Smith (132_2) et celle d'Alice Jones (865_1) concernent une entreprise détenue en partenariat. Ils ont déclaré la même adresse, mais leur présentation est telle qu'une machine aura de la difficulté à établir des correspondances strictement par comparaison de chaînes. Les différents noms d'entreprise et numéros d'entreprise peuvent être attribués à des coquilles. Les noms des partenaires sont également une indication claire d'un partenariat.

Eliza et Emilie Rochefort sont également des candidates évidentes pour une entreprise détenue en partenariat. Elles ont déclaré la même adresse, une part de partenaire de 50 % et le même code d'industrie.

Tableau 3.2-1**Exemple de données : Colonnes de renseignements sur l'entreprise**

Code ID	Nom du déclarant	Regroupement DBSCAN	Numéro d'entreprise	Nom de l'entreprise	Adresse de l'entreprise	Code d'industrie	Pourcentage des partenaires	Nom du ou des partenaires
132_1	John Smith	1			Rue 456	333		
132_2	John Smith	1	885	Gold Star Staffing	#3, rue 467,	334	50	Jones
865_1	Alice Jones	1	8885	Gold Star Staff	Unité 3, rue 467 N.-O.	334		John Smith
611_1	Eliza Rochefort	2			123, Sésame	531	50	
376_1	Emilie Rochefort	2			123, rue Sésame	531	50	
978_3	Marie Dupoint	-1			Rue 1-476	531		

Dans chaque regroupement DBSCAN, la série complète de comparaisons par paires est effectuée sur chacune de ces colonnes. Le regroupement 1 fera trois comparaisons par paire et le regroupement 2, une seule comparaison par paire. Marie Dupoint (978_3) n'a reçu aucun regroupement (point de bruit) et ne sera pas incluse à l'avenir.

Tableau 3.2-2

Exemple de données : Comparaisons entre colonnes de paires de rangées

Code ID 1	Code ID 2	Regroupement DBSCAN	Correspondance du numéro d'entreprise	Correspondance du nom de l'entreprise	Correspondance de l'adresse de l'entreprise	Correspondance du code d'industrie	Somme des pourcentages de partenaire = 100	Correspondance du nom du partenaire	Correspondance acceptée?
132_1	132_2	1	0	0	0,59	0	0	0	Non
132_1	865_1	1	0	0	0,63	0	0	1	Non
132_2	865_1	1	0	0,97	0,61	1	0	1	Oui
611_1	376_1	2	0	0	0,89	1	1	0	Oui

La paire (132_1, 132_2) est immédiatement ignorée, car une personne ne peut pas être un partenaire pour elle-même.

La paire (132_1, 865_1) n'est pas non plus acceptée. Même si Alice Jones a déclaré que John Smith était partenaire et que John Smith a soumis deux déclarations au sein du regroupement, aucun autre critère d'appariement satisfait ne suggère que l'appariement est de piètre qualité. L'appariement n'est pas accepté.

La paire (132_2, 865_1) indique une excellente correspondance entre le nom de l'entreprise et les partenaires déclarés. Un code d'industrie correspondant est également indicatif. L'appariement est accepté.

La paire (611_1, 376_1) présente une correspondance adéquate entre l'adresse de l'entreprise et le pourcentage des partenaires qui atteint bien 100 %. L'appariement est accepté.

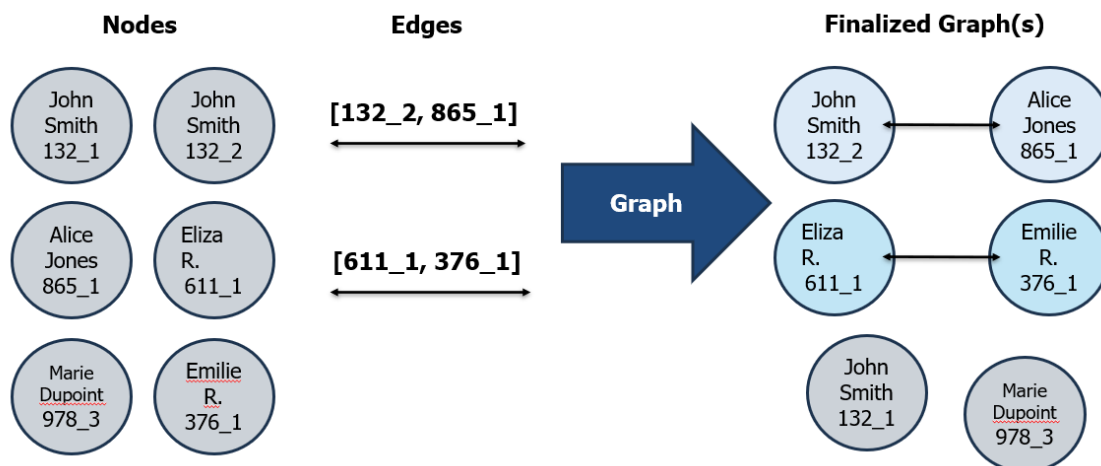
3.3 Représentation graphique des paires de rangées acceptées

La théorie des graphes permet la construction de l'ensemble final de partenaires à partir des paires de rangées relevées en (3.2). La saisie des paires de rangées dans le logiciel de graphique crée le graphique où chaque sous-graphique de points connectés produit représente un partenariat finalisé.

Dans l'exemple fourni, les paires de rangées de correspondances acceptées sont utilisées comme arêtes dans le graphique. Le produit identifie deux partenariats de deux points chacun : John Smith (132_2) et Alice Jones (865_1), Eliza Rochefort (611_1) et Emilie Rochefort (376_1). Deux entreprises individuelles, John Smith (132_1) et Marie Dupont (978_3), sont également identifiées.

Figure 3.3-1

Aperçu du processus de représentation graphique



4. Conclusion

La méthodologie ajustée pour le processus d'identification des entreprises détenues en partenariat a produit des résultats prometteurs. Les industries dont les modèles de déclaration sont complexes présentent encore des difficultés en termes de méthodologie, au moment de rédiger cet article, mais d'autres ajustements apportés à l'application de DBSCAN et le fractionnement du processus par industrie entraînent des améliorations au processus. Le remaniement signifie également que nous disposons d'une plateforme plus sophistiquée que le logiciel que nous avons l'intention de remplacer, avec laquelle nous pouvons élaborer de futures améliorations.

Bibliographie

- Chire, (2011), "DBSCAN-Illustration," *Wikimedia Commons*. Extrait le 15 avril, 2025 de <https://commons.wikimedia.org/w/index.php?title=File:DBSCAN-Illustration.svg&oldid=872952835>.
- Ester, M., H.-P. Kriegel, J. Sander et X. Xu. (1996), « A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise », *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, p. 226-231.
- Hagberg, A.A., D.A. Schult et P.J. Swart. (2008), « Exploring network structure, dynamics, and function using NetworkX », *Proceedings of the 7th Python in Science Conference (SciPy, 2008)*, p. 11-16.
- Jaro, M.A. (1989), « Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida ». *Journal of the American Statistical Association*, 84(406), p. 414-420.
- Winkler, W.E. (1990), « String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage », *Evaluative Report*, Washington DC, États-Unis d'Amérique : U.S. Bureau of the Census Stat. Research Div.