CIQSS | QICSS
CENTRE INTERUNIVERSITAIRE QUÉBÉCOIS DE STATISTIQUES SOCIALES | QUEBEC INTER-UNIVERSITY CENTRE FOR SOCIAL STATISTICS

GESS SSSG Groupe d'étude en statistique sociale Social Statistics Study Group

Le potentiel des données administratives et couplées pour la recherche en sciences sociales

Un survol en bref

Xavier St-Denis, INRS

Webinaire

Centre interuniversitaire québécois de statistiques sociales (CIQSS)

2 novembre 2023

INRS Institut national de la recherche scientifique

Objectifs et plan de la présentation

1. Principales caractéristiques et avantages des données administratives
2. Survol des données administratives canadiennes les plus utilisées avec exemples
3. Limites et opportunités de développement de sources de données administratives

Institut national de la recherche scientifique

Données administratives: définition et contexte

- **Données numériques « trouvées »** et traitées pour un usage secondaire plutôt que « collectées »

- **Utilisation principale autre que la recherche** par l'organisation productrice ou dépositaire des données

- Le contexte canadien:

  - **Relation/interface entre les chercheurs et les des dépositaires** et producteurs des données assuré par organismes statistiques fédéraux (StatCan) et provinciaux (ISQ) assurent l'interface dans une grande partie des cas (relation avec ministères, etc.)

  - **Éthique en recherche et consentement des individus** « observés » est gouverné par des cadres légaux au fédéral et au provincial

Institut national de la recherche scientifique

Principales sources de données administratives au Canada

- Historiquement: données de l'état civil et de paroisses (naissances, décès, mariages)

- Données fiscales
  - **Revenu personnel:** Fichier familles T1 (T1FF) et ses fichiers dérivés (DAL et extraits pour couplages), autres relevés
  - Données d'**entreprises**: BEAM, CEEDD, FDLMO, etc.

- Données de programmes et données ministérielles
  - **Immigration**: BDIM
  - **Éducation**: SIEP, données en éducation primaire, secondaire et post-secondaire au Québec et autres provinces
  - **Santé et services sociaux**: DAD/NACRS, RAMQ et autres données provinciales (opioïdes, etc.)
  - **Assurance-emploi** (PVAE)**, pensions** (RPC) **et autres programmes** de soutien au revenu
  - **Système judiciaire**

- Couplages de données
  - Avec enquêtes et recensements
  - Avec données de programme ou données institutionnelles, gouvernementales ou non
  - Liste des couplages approuvés: https://www.statcan.gc.ca/en/record/summ

- Données accessibles aux chercheurs: https://www.statcan.gc.ca/en/microdata/data-centres/data

INRS Institut national de la recherche scientifique

Les données administratives: éléments de base

Caractéristiques	Avantages	A	R	E
Données numériques « trouvées » préexistantes chez dépositaires	• Ne dépend pas de la participation des répondants • Moins d'enjeu de sous-dénombrement dû à la non-réponse* [*mais: dépend de la couverture « primaire »] • Peu d'attrition lorsque longitudinal • Coût moindre vs collecte de données équivalentes	+	±	-
Production selon fréquence régulière souvent annuelle (« haute vélocité »?)	• Analyses de phénomènes contemporains et priorités avec données récentes* [*mais: délais du cycle de production]	+	±	?
Volume élevé d'observations similaire à données massives (*big data*)	• Inférence statistique plus précise • Données sur minorités, petites unités géo, etc.	+	+	±
Mesures provenant d'enregistrements plutôt que réponses auto-rapportées	• Haute précision et moindre risque d'erreur de mesure* • Uniformité d'une période à l'autre*	+	-	-
Identifiants personnels uniques ou information équivalente	• Analyses longitudinales • Couplages entre sources de données	+	±	?

Légende: **A** Données administratives; **R** Recensement canadien; **E** Enquêtes traditionnelles

Principales variables dans les données fiscales (T1FF/DAL)

- Identifiant personnel unique
- Variables sociodémographiques de base (âge et sexe)
- Revenu
- Structure familiale
- Lieu de résidence
- Complété avec données d'autres sources
  (couplages avec enquêtes ou données administratives d'autres programmes)

INRS Institut national de la recherche scientifique

Le **revenu** dans les données fiscales canadiennes

- Éléments de base:
  - Revenu personnel rapporté dans le formulaire de déclaration T1 et relevés fiscaux associés
  - Information agrégée au niveau annuel
  - Niveaux d'analyse: individuel et familial

- Types de sources de revenu
  - Revenu d'emploi (T4, travail autonome, pourboires et commissions, etc.)
  - Revenu de marché (investissements, gains en capital, etc.)
  - Taxes et transferts gouvernementaux de toute sorte (allocations familiales, pensions, aide sociale, PCU)

- Identification indirecte:
  - À travers source de revenu (santé, statut d'emploi, etc.)
  - Exercice: identification des chômeurs

INRS Institut national de la recherche scientifique

Extrait de l'information sur le revenu dans un formulaire T1

Étape 2 – Revenu total

En tant que résident du Canada, vous devez déclarer vos revenus de toutes les sources canadiennes et étrangères.

Revenus d'emploi (case 14 de tous les feuillets T4)			10100	1
Revenu exonéré d'impôt versé aux volontaires des services d'urgence (lisez la ligne 10100 du guide)	10105			
Commissions incluses à la ligne 10100 (case 42 de tous les feuillets T4)	10120			
Cotisations à un régime d'assurance-salaire (lisez la ligne 10100 du guide)	10130			
Autres revenus d'emploi (lisez la ligne 10400 du guide)			10400 +	2
Pension de sécurité de la vieillesse (PSV) (case 18 du feuillet T4A(OAS))			11300 +	3
Prestations du RPC ou du RRQ (case 20 du feuillet T4A(P))			11400 +	4
Prestations d'invalidité incluses à la ligne 11400 (case 16 du feuillet T4A(P))	11410			
Autres pensions et pensions de retraite (lisez la ligne 11500 du guide et la ligne 31400 de la déclaration)			11500 +	5
Choix du montant de pension fractionné (remplissez le formulaire T1032)			11600 +	6
Prestation universelle pour la garde d'enfants (PUGE) (consultez le feuillet RC62)			11700 +	7
Montant de la PUGE désigné à une personne à charge	11701			
Prestations d'assurance-emploi (AE) et autres prestations (case 14 du feuillet T4E)			11900 +	8
Prestations de maternité et parentales de l'AE et prestations du régime provincial d'assurance parentale (RPAP)	11905			
Montant imposable des dividendes de sociétés canadiennes imposables (utilisez la feuille de travail fédérale) :				
Montant des dividendes (**déterminés** et **autres que déterminés**)			12000 +	9
Montant des dividendes (**autres que déterminés**)	12010			
Intérêts et autres revenus de placements (utilisez la feuille de travail fédérale)			12100 +	10
Revenus nets de société de personnes (commanditaires ou associés passifs seulement)			12200 +	11
Revenus d'un régime enregistré d'épargne-invalidité (REEI) (case 131 du feuillet T4A)			12500 +	12
Revenus de location (consultez le guide T4036) Bruts 12599		Nets	12600 +	13
Gains en capital imposables (remplissez l'annexe 3)			12700 +	14

Source: https://www.canada.ca/content/dam/cra-arc/formspubs/pbg/5005-r/5005-r-22f.pdf

INRS Institut national de la recherche scientifique

Extrait de l'information sur le revenu dans un formulaire T1

Revenus d'un travail indépendant (consultez le guide T4002) :										
Revenus d'entreprise	Bruts	13499		Nets	13500			20		
Revenus de profession libérale	Bruts	13699		Nets	13700	+		21		
Revenus de commissions	Bruts	13899		Nets	13900	+		22		
Revenus d'agriculture	Bruts	14099		Nets	14100	+		23		
Revenus de pêche	Bruts	14299		Nets	14300	+		24		
Additionnez les lignes 20 à 24.			Revenus nets d'un travail indépendant			=		▶	+	25
Ligne 19 plus ligne 25									=	26
Indemnités pour accidents du travail (case 10 du feuillet T5007)					14400			27		
Prestations d'assistance sociale					14500	+		28		
Versement net des suppléments fédéraux (case 21 du feuillet T4A(OAS))					14600	+		29		
Additionnez les lignes 27 à 29 (lisez la ligne 25000 à l'étape 4).					14700	=		▶	+	30
Ligne 26 plus ligne 30							**Revenu total**	15000	=	31

Source: https://www.canada.ca/content/dam/cra-arc/formspubs/pbg/5005-r/5005-r-22f.pdf

Institut national de la recherche scientifique

Transposition de l'information dans la base de Données administratives longitudinales (DAL) de StatCan

Revenu total – Définition de StatCan (XTIRC)

(1982 à présent)

Définition : Le revenu total (TIRC), qui figure à la ligne 150 du formulaire d'impôt T1, représente la somme du revenu d'un déclarant pour les besoins de l'Agence du revenu du Canada. La DSR a apporté certaines modifications à cette variable afin d'obtenir sa propre définition du revenu total (XTIRC). Celle-ci comprend le revenu du déclarant provenant de sources imposables et non imposables. Cette définition a été changée au cours des années afin de refléter les modifications apportées au formulaire d'impôt, aux crédits d'impôt remboursables et aux calculs du revenu. La relation entre la définition de l'Agence du revenu du Canada et celle de la DSR est la suivante (voir la section 14, tableau 4, pour une liste complète des variables) :

XTIRC = TIRC - {rajustements des dividendes} - {gains en capital} + {crédits d'impôt remboursables} + {autre revenu non imposable}

Pour une comptabilité complète des variables particulières utilisées pour définir le XTIRC pour des années particulières, et les différences entre le XTIRC et le TIRC, veuillez consulter la section 11 de ce dictionnaire de données.

Dérivée de : traitement du fichier T1FF

DAL : XTIRC I, F, P, K

Source: https://www150.statcan.gc.ca/n1/fr/pub/12-585-x/12-585-x2021001-fra.pdf?st=zXTYZu9U

INRS Institut national de la recherche scientifique

Transposition de l'information dans la base de Données administratives longitudinales (DAL) de StatCan

Prestation canadienne d'urgence (CV19CERB_)

(2020)

Définition : La Prestation canadienne d'urgence (PCU) a fourni un soutien financier aux employés et aux travailleurs indépendants canadiens qui étaient touchés directement par la COVID-19. S'ils étaient éligibles, les individus pouvaient recevoir 2 000 $ pour une période de quatre semaines (l'équivalent de 500 $ par semaine) pour un maximum de 7 périodes (28 semaines). Le PCU est un avantage imposable.

À compter du 27 septembre 2020, les personnes qui étaient toujours dans l'incapacité de travailler ont été transférées vers un programme simplifié d'assurance-emploi (AE) ou vers la Prestation canadienne de la relance économique (PRC).

Dérivée de : traitement du fichier T1FF

DAL : CV19CERB_ I, F, P

Source: https://www150.statcan.gc.ca/n1/fr/pub/12-585-x/12-585-x2022001-fra.pdf?st=VBauy78T

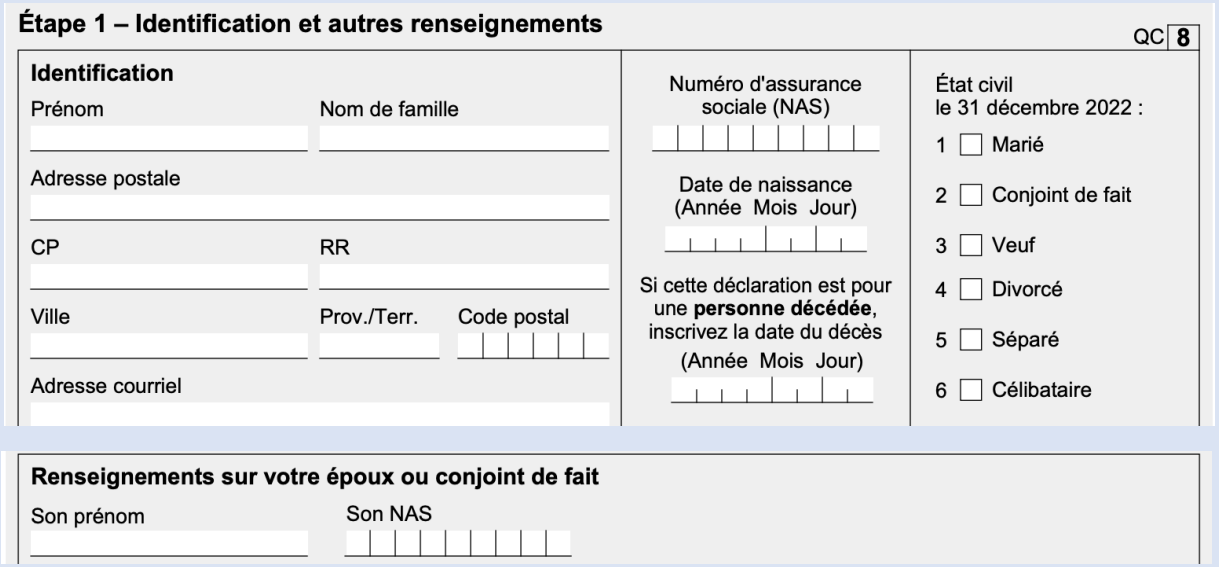
INRS Institut national de la recherche scientifique

Revenu: Exemples de projet

- Trajectoires de revenu
  - Entrées et sortie de la pauvreté (DAL)
  - Trajectoires de revenu des immigrants (BDIM)
  - Entrée sur le marché du travail des diplômés de différents programmes (PLEMT)
  - Trajectoires de revenu de minorités sexuelles
- Inégalités, revenu pré/post-taxe et influence de certains programmes et transferts sociaux
- Analyse de faisabilité de déclaration fiscale automatique

Institut national de la recherche scientifique

Extrait de l'information sur la famille dans un formulaire T1



Étape 1 – Identification et autres renseignements QC 8

Identification

Prénom | Nom de famille

Adresse postale

CP | RR

Ville | Prov./Terr. | Code postal

Adresse courriel

Numéro d'assurance sociale (NAS)

Date de naissance (Année Mois Jour)

Si cette déclaration est pour une **personne décédée**, inscrivez la date du décès (Année Mois Jour)

État civil le 31 décembre 2022 :

1 ☐ Marié
2 ☐ Conjoint de fait
3 ☐ Veuf
4 ☐ Divorcé
5 ☐ Séparé
6 ☐ Célibataire

Renseignements sur votre époux ou conjoint de fait

Son prénom | Son NAS

Source: https://www.canada.ca/content/dam/cra-arc/formspubs/pbg/5005-r/5005-r-22f.pdf

Institut national de la recherche scientifique

Structure familiale et données fiscales: Éléments de base

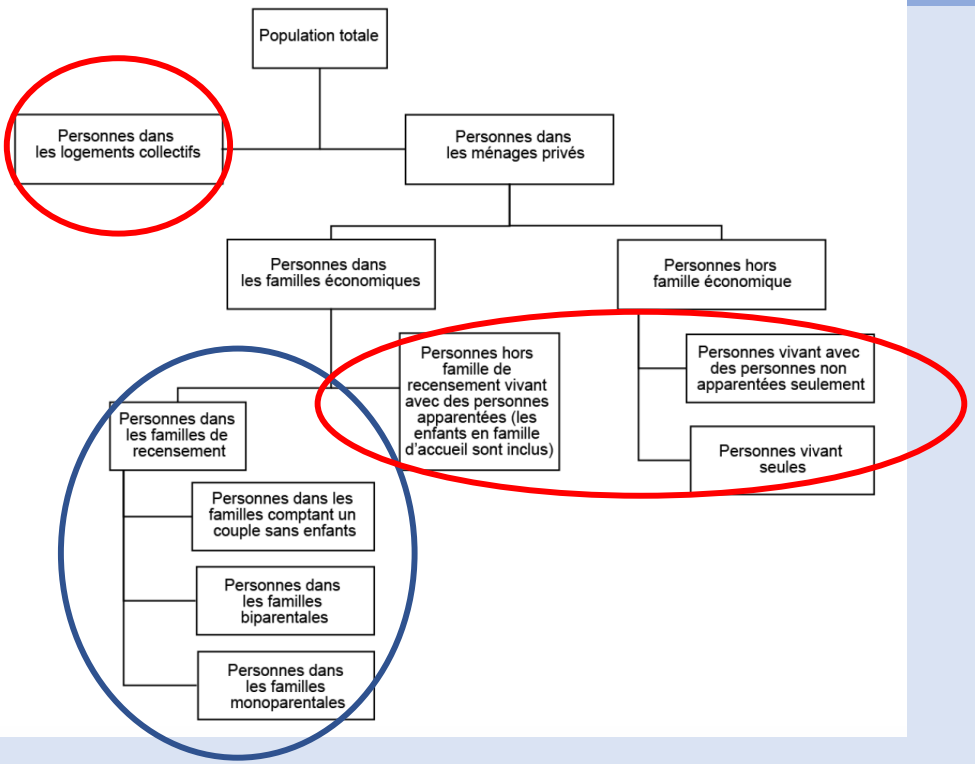
- Principales variables
  1. FIN (identifiant familial transversal)
  2. ID de l'époux.se
  3. Structure de la famille (couples avec/sans enfants, personne seule, et famille monoparentale)
  4. Statut de l'individu dans la famille (conjoint.e, enfant, personne seule)
  5. Statut matrimonial auto-rapporté
  6. Nombre d'enfants selon l'âge (incluant les non-déclarants)
  7. Informations sur enfants provenant de relevés complémentaires (programme d'allocation familiale)

- Méthode:
  - **Conjoints légaux et les conjoints de fait:** reliés à partir du numéro d'assurance sociale (NAS) de leur conjoint inscrit sur le formulaire d'impôt ou par un appariement effectué en fonction du nom, de l'adresse, du sexe et de l'état matrimonial.
  - **Enfants déclarants:** identifiés à partir d'un algorithme semblable et de fichiers complémentaires.
  - **Enfants non-déclarants:** Avant 1993, identifiés à partir des renseignements sur la déclaration de revenus de leurs parents et autre données (allocations familiales). Depuis 1993, identifié à travers le programme de prestations fiscales pour enfants.

INRS Institut national de la recherche scientifique

Classification des types de familles de Statistique Canada

Appartenance aux familles économiques et appartenance aux familles de recensement, Recensement de la population de 2021



Source: https://www12.statcan.gc.ca/census-recensement/2021/ref/98-500/002/98-500-x2021002-fra.pdf

Institut national de la recherche scientifique

Variables sur la structure familiale dans les données fiscales

Code de description individuelle – sexe, âge, imputé, état matrimonial (INDFL)

(1982 à présent)

Définition : La description du particulier est un code numérique attribué aux personnes d'une même catégorie descriptive.

Voici une liste des codes et de leur description :

1 : Homme, adulte, déclarant, marié ou en union libre;
2 : Homme, adulte, non déclarant (personne imputée), marié ou en union libre;
3 : Femme, adulte, déclarante, mariée ou en union libre;
4 : Femme, adulte, non déclarante (personne imputée), mariée ou en union libre;
5 : Enfant déclarant;
6 : Enfant non déclarant (imputé); (disponible seulement de 1993 à présent);
7 : Adulte, déclarant, parent seul;
8 : Personne hors famille, déclarante.

Si une personne meurt au cours d'une année donnée, son statut avant son décès est défini par cette variable.

Il n'y a aucune restriction sur l'âge des enfants. Un enfant est défini comme toute personne célibataire qui vit avec un ou deux parents. Par exemple, un enfant de 50 ans peut demeurer avec un parent âgé de 70 ans. Cette famille serait classifiée comme une famille monoparentale.

Dérivée de : traitement de la banque DAL

DAL : INDFL I, F, P, K

Source: https://www150.statcan.gc.ca/n1/fr/pub/12-585-x/12-585-x2022001-fra.pdf?st=VBauy78T

INRS Institut national de la recherche scientifique

Variables sur la structure familiale dans les données fiscales

Flag - Same sex couple (SSFLG)

(2000 to present)

Definition: Starting in 2000, a same sex couple could report on the tax form that they are a common-law family.

Derived from: T1FF processing

LAD: SSFLG I character

Source: https://www150.statcan.gc.ca/n1/fr/pub/12-585-x/12-585-x2022001-fra.pdf?st=VBauy78T

Institut national de la recherche scientifique

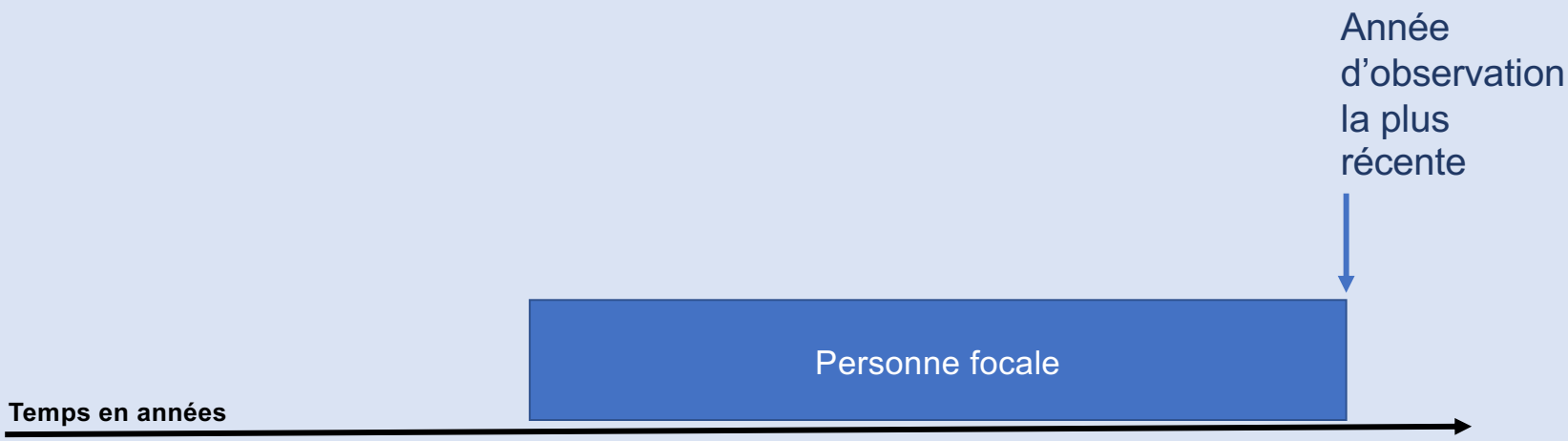
Diversité des structures familiales dans les données fiscales

- Concept de famille: la **famille de recensement**
  - Conjoint ou conjointe
    - Couples de même sexe
  - Enfants si présents
    - Et donc fratrie
    - Y compris enfants adultes sans enfant et célibataire)
  - N'inclut pas: colocs, grands-parents, frère/sœur si forme une famille de Recensement séparée, etc.

- **Identification indirecte** à travers les données longitudinales
  - **Parents** de déclarants adultes formant leur propre famille de recensement (cohortes de naissance 1965 et + environ)
  - Autres personnes dans le **réseau familial élargi**? → Innovation nécessaire
  - **Exercice**: Comment identifier ….
    - Un grand-parent?
    - Un.e oncle/tante?
    - Un.e ex-conjoint.e?

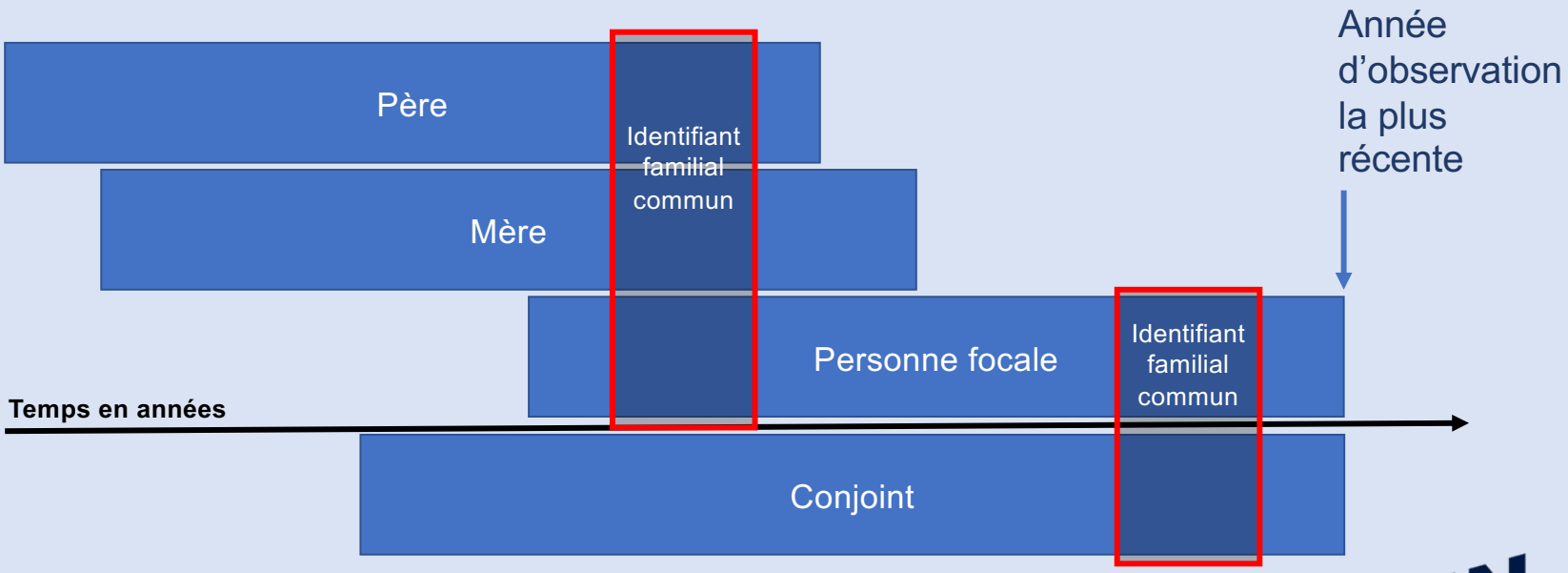
INRS Institut national de la recherche scientifique

Diversité des structures familiales dans les données fiscales



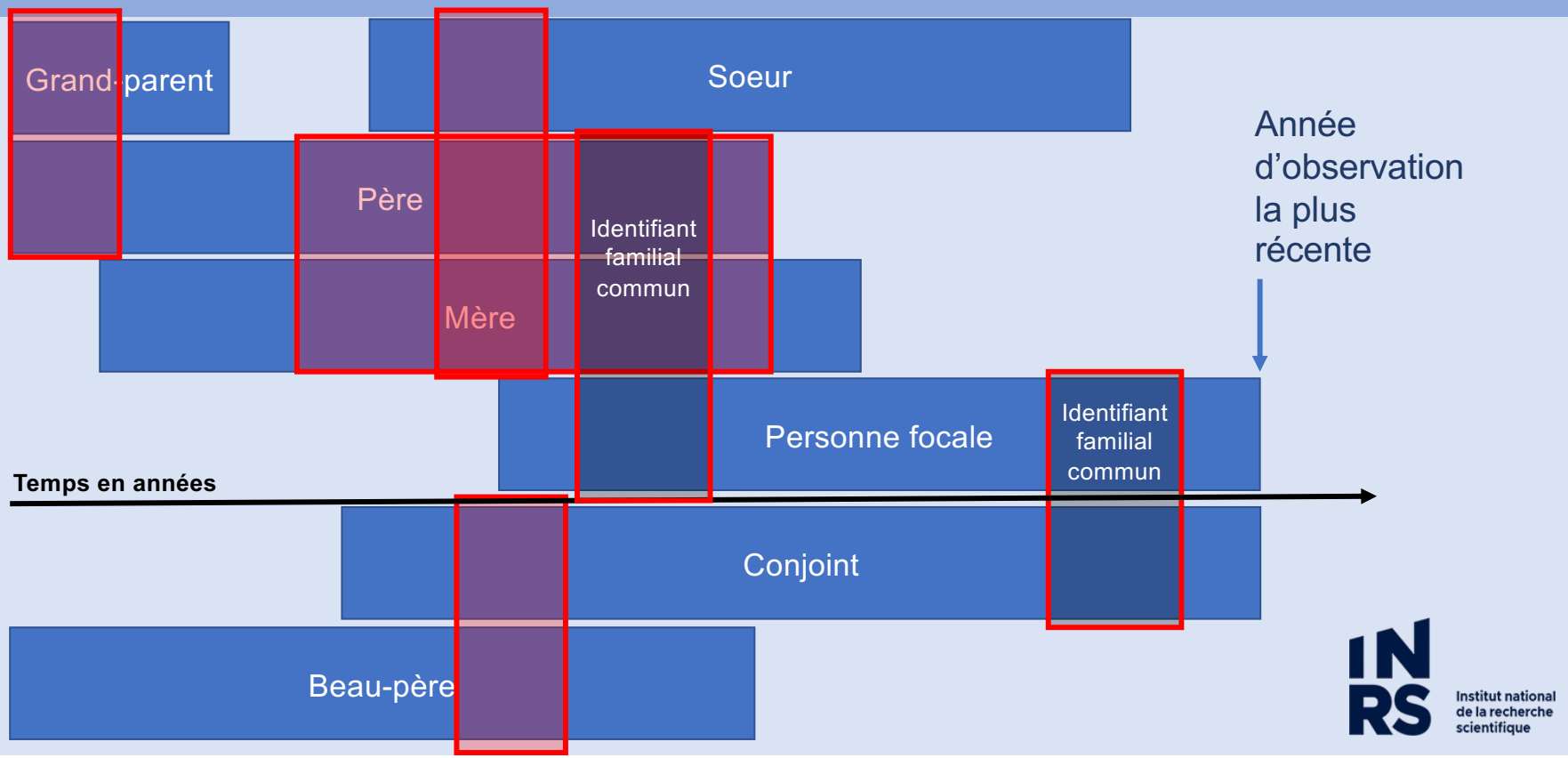
Institut national de la recherche scientifique

Diversité des structures familiales dans les données fiscales



Institut national de la recherche scientifique

Diversité des structures familiales dans les données fiscales



Institut national de la recherche scientifique

Structure familiale: Exemples de projets

- Évolution de la prévalence des divorces post-2008
- Études sur la mobilité sociale et la transmission intergénérationnelle du revenu
- Inégalités socioéconomique au sein de personnes ayant déjà formé des couples de même sexe et homogamie selon le type de couple (même sexe ou non)
- Au-delà des couples et des paires parent-enfant: identification des grands-parents, frères et sœurs, ex-conjoints, beaux-parents, etc.
  → Proximité résidentielle

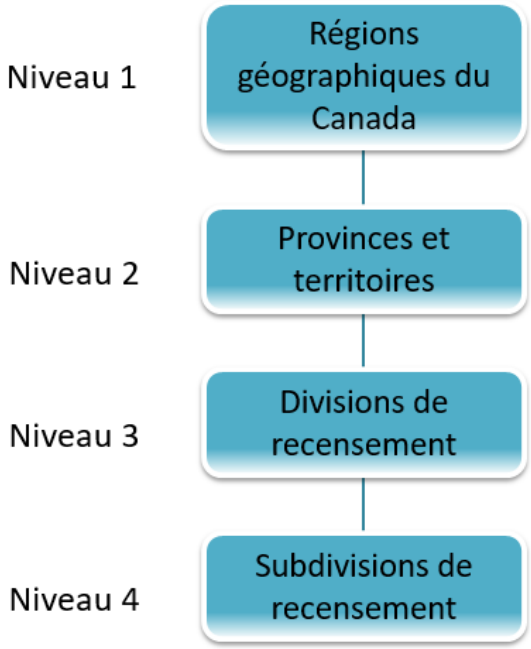
INRS Institut national de la recherche scientifique

Lieu de **résidence** et données fiscales: Éléments de base

- Identifiants géographiques:
  - Province de résidence le 31 décembre de l'année fiscale concernée
  - Adresse de correspondance lors de la soumission de la déclaration
  - Information auxiliaire: Adresse (province) de l'employeur dans relevé T4
- Code postal et codes correspondants de la *Classification géographique type (CGT)*
- Enjeux: précision de l'information sur le lieu de résidence

INRS Institut national de la recherche scientifique

Niveaux de la Classification géographique type de StatCan



Code de la ville de Brossard

Région	PR	DR	SDR	Nom
2				Québec
	24			Québec
	24	58		Longueuil (Territoire équivalent)
	24	58	007	Brossard (Ville)

Source: https://www.statcan.gc.ca/fr/sujets/norme/cgt/2021/introduction

Institut national de la recherche scientifique

Lieu de résidence: exemples de projets

- Analyses à une petite échelle géographique

- Migration permanente et temporaire, incluant chez certaines sous-populations
  - Travailleurs interprovinciaux
  - Établissement des immigrants en région
  - Migration interne des jeunes et des apprentis

- Avenues à explorer:
  - Proximité résidentielle
  - Trajectoires résidentielles de certaines sous-populations: étudiants et diplômés, personnes à faible revenu, etc.

INRS Institut national de la recherche scientifique

Intégration de données fiscales et d'autres bases de données

- Recensement et enquêtes: variables sur le revenu et autres informations auxiliaires

- Couplages enquête-T1FF ou données de programme-T1FF
  - Sélection et recodage de certaines variables → Différent contenu à travers les sources de données

INRS Institut national de la recherche scientifique

Données de programme: un aperçu

- **PLEMT**: au-delà des inégalités d'accès à l'éducation, les parcours éducatifs
  - Cohortes d'entrants au sein du système d'éducation postsecondaire depuis 2007
  - Couplage avec T1FF, Recensement 2016, END, BDIM, etc.
  - Information sur le programme d'étude, l'institution d'enseignement, et autre

- **BDIM**: programmes d'arrivée des immigrants et trajectoires de revenu post-immigration
  - Cohortes d'immigrants depuis 1978 (1952 pour information partielle)
  - Programme d'arrivée, information sur caractéristiques pré-arrivées, pays d'origine, etc.
  - Couplage avec T1FF depuis 1982
  - Nouvelle plateforme de couplage: LISE

INRS Institut national de la recherche scientifique

Couplages de données: Un aperçu

- Modèles de couplages:
  1. Correspondance (T1 et T4 avec le FDLMO)
  2. Rétrospectif (ESG-T1FF)
  3. Prospectif (BDIM et ELNEJ-T1FF)

- Exemples d'utilisation des couplages:
  - Complément d'information (par exemple, variables sur l'identité provenant de données d'enquête)
  - Validation de mesures indirectes ou de couverture d'une source de données administratives

- Exemples de couplages innovants:
  - BDIM et PLEMT
  - ELIA couplée au T1FF, T4, RPC et BDIM
  - Couplages entre données de programme provinciaux et données fiscales

INRS Institut national de la recherche scientifique

Limites des données administratives pour la recherche en sciences sociales

- **Manque certaines variables**:
  - Caractéristiques sociodémographiques: identité de genre, orientation sexuelle (attraction, attitudes et comportements), minorité visible, identité autochtone, religion, etc.
    - Certaines sont plus fluides que d'autres et plus difficile à inférer de manière indirecte
  - Structures familiales ou de ménage complexes
  - Emploi: professions; inactivité vs chômage; observations sous-annuelles (heures et semaines travaillées)
  - Épargne et consommation
  - Questions subjectives (aspirations, motivations, attitudes, opinions, bien-être, etc.)

- **Sous-populations moins bien couvertes**, dépendant de leur interaction avec les institutions (pas nécessairement unique aux données administratives)
  - Ex.: déclaration fiscale et participation à l'économie formelle; interaction avec système de santé; déclaration de l'émigration dans rapport d'impôt; résidence temporaire

INRS Institut national de la recherche scientifique

Limites des données administratives pour la recherche en sciences sociales

- **Peu de contrôle sur le contenu** puisque pas collecté par ou pour chercheurs
- **Obstacles à la collecte, le maintien et l'accès externe** (chercheurs) à ces données, surtout lorsque les organisations productrices ou dépositaires des données n'ont pas les ressources humaines, matérielles et financières nécessaires à l'interne
- **Utilisateurs et fournisseurs peu expérimentés/familiers** avec enjeux d'accès, de conservation, de partage et d'analyse de ce type de données
- Certaines informations probablement jamais disponibles – pourquoi?
- Enjeu de **reproductibilité/réplication**

INRS Institut national de la recherche scientifique

Pour devenir un expert des données administratives…
Participez au BADA!

- Le **Bootcamp en analyse de données administratives (BADA)** est une école d'été destinée aux chercheuses et chercheurs souhaitant réaliser des analyses avec des données administratives ou couplées de Statistique Canada.

- Objectifs
  1. Développer une connaissance avancée du contenu des principaux ensembles de données administratives canadiennes.
  2. Identifier le potentiel et les limites des données administratives pour la recherche en sciences sociales.
  3. Développer des techniques de manipulation de données et des méthodes d'analyse applicables aux données administratives à l'aide de Stata (avec des compléments possibles en SAS et R).
  4. Développer des compétences en analyse de données longitudinales.

- Détails
  - Date: 10-14 Juin 2024
  - Participants ciblés: Utilisateurs avec expérience préalable en méthodes quantitatives

INRS Institut national de la recherche scientifique