



PERSPECTIVES DE L'IMT

En quête du graal de l'IMT : des données locales, granulaires, fréquentes et actuelles

Principales conclusions

- L'**étude d'opinion publique** du CIMT sur les besoins des Canadiens en matière d'information sur le marché du travail a mis en lumière l'importance d'une information pertinente et adaptée à la décision à prendre. Pour de nombreux Canadiens, cela correspond à une information spécifique à leur ville ou leur emploi, et pour d'autres, comme nos parties prenantes, cela signifie une information récente et à jour. En fait, nos partenaires et intervenants réclament depuis longtemps une information plus granulaire à l'échelle locale afin de soutenir la prise de décisions de meilleure qualité et plus informées, un point saillant de notre tout **premier numéro de Perspectives de l'IMT**.
- C'est dans cette optique que le CIMT et Statistique Canada, en collaboration avec leurs partenaires, ont évalué de nombreuses approches pour en arriver à une information sur le marché du travail plus locale (plus petites régions), granulaire (plus de détails), fréquente (plus souvent) et actuelle (plus à jour). Les approches examinées et débattues dans ce numéro comprennent : 1) des options fondées sur des enquêtes, 2) le couplage de données administratives et 3) des méthodes de modélisation.
- Les trois approches proposées varient considérablement dans leur façon d'améliorer l'information sur le marché du travail. Aucune approche n'est une panacée et chacune implique des inconvénients différents. On peut arriver à des données plus granulaires, mais au détriment de la localité, par exemple. Les trois approches ont aussi des coûts très différents.
- En se fondant sur notre évaluation préliminaire, dans un premier temps, le CIMT et Statistique Canada exploreront la faisabilité d'une méthode de modélisation particulière : celle d'estimation sur petits domaines (EPD). À cet égard, les sources existantes et variées d'information sur le marché du travail – comme l'Enquête sur la population active (EPA), les données de recensement et d'autres ensembles de données – sont des outils essentiels pour tirer parti de cette approche.
- Après avoir reçu les commentaires de nos intervenants et partenaires, nous réévaluerons dans quelle mesure cette nouvelle information répond à leurs besoins. Nous ajusterons le projet si nécessaire, soit pour intensifier la mise en œuvre de la production de données locales et granulaires, soit pour revoir les approches présentées ici.

Introduction

Le **premier numéro de Perspectives de l'IMT** faisait état du manque de données locales, granulaires, fréquentes et actuelles comme étant une lacune importante du système d'information sur le marché du travail (IMT) au Canada. Devant ce défi, un **plan stratégique** a été établi par le CIMT, priorisant la collaboration avec ses partenaires et intervenants afin d'explorer la faisabilité de différentes options pour fournir une IMT plus granulaire et localisée. Préparée par le CIMT et Statistique Canada, la présente édition de *Perspectives de l'IMT* définit et évalue trois grandes approches puis expose les prochaines étapes à suivre pour combler cette lacune. L'amélioration de l'information dans ce sens aidera les décideurs, les éducateurs, les professionnels en développement de carrière et les Canadiens en général à prendre des décisions plus éclairées en matière de carrière, de formation et d'éducation.

Dans le but de jeter les bases de cette publication, un récent **billet de blogue** du CIMT clarifiait les quatre critères distincts, souvent confondus, utilisés pour comparer l'information sur le marché du travail (voir le **tableau 1**). Lorsqu'on réfléchit à la façon dont les diverses approches se mesurent à ces critères, il est important de noter les inconvénients inhérents de chacune. Par exemple, une IMT plus granulaire pourrait être possible, mais seulement au détriment de son actualité. Dans chaque cas, le facteur limitatif principal est la fiabilité de l'information. Par exemple, il n'est jamais recommandé de maximiser le caractère local au détriment d'une estimation fiable du taux de chômage. Un degré raisonnable de fiabilité doit donc être atteint lorsqu'on améliore l'un ou l'autre des critères ci-dessous.

Tableau 1 : Critères de comparaison de l'IMT

Localité :	Le plus petit niveau de détail géographique
Granularité :	Nombre de catégories selon lesquelles les données sont organisées (p. ex. : CNP, âge, niveau de formation, statut d'immigration, etc.)
Fréquence :	Intervalles auxquels les données sont disponibles, p. ex. mensuel ou annuel
Actualité :	Délai entre la période de référence et la publication

Trois approches

En fonction de ces quatre critères, la présente édition de *Perspectives de l'IMT* examine un certain nombre de façons possibles d'améliorer la qualité générale de l'information sur le marché du travail. Plus précisément, nous explorons ces trois grandes approches :

1. Enquêtes nouvelles ou élargies
2. Création de nouveaux couplages entre les ensembles de données administratives existants

3. Application de techniques de modélisation et de statistique pour produire des estimations plus précises

Dans le cadre de chaque approche, des options précises ont été définies et développées. La suite de cet article se concentre sur l'évaluation de chaque option selon la localité, la granularité, la fréquence et l'actualité, et prend aussi en compte les coûts et la rigueur statistique. Le **tableau 2** présente un résumé de cette évaluation.

Tableau 2 : Évaluation de trois grandes approches

	LOCALITÉ	GRANULARITÉ	FRÉQUENCE	ACTUALITÉ	COÛ
Aperçu actuel des données					
Enquête sur la population active (EPA)	★★	★★★	★★★★	★★★★	moyen
Recensement (questionnaire long)	★★★	★★★★	★	★	élevé
Profil vectoriel de l'assurance-emploi (PVAE)	★★	★★	★★★★	★★	bas
Fichier des familles T1 (FFT1)	★★★★	★	★★	★	bas
1. Approche par enquêtes					
1. a) Augmentation de la taille de l'échantillon de l'EPA	★★	★★★	★★★★	★★★★	élevé
1. b) Élargissement de l'EPA pour ne poser les questions de base que tous les 3 ou 6 mois	★★	★★★★	★★★	★★★★	élevé
1. c) Conception d'une nouvelle enquête de base axée sur les régions locales	★★	★★★	★★★	★★★	élevé
1. d) Ajout de questions de base de l'EPA à l'enquête existante	★	★★★	★★	★★★	élevé
2. Couplage de données administratives					
2. a) Couplage du recensement (questionnaire long) avec les fichiers fiscaux et du PVAE au fil du temps	★★★	★★★★	★★	★	bas
3. Méthodes de modélisation					
3. a) Modèles d'estimation sur petits domaines	★★★	★★	★★★	★★★	moyen
3. b) Techniques de modélisation avancées combinant recensement et EPA	★★★	★★★	★★★	★★★★	moyen
	Localité :	Granularité :	Fréquence :	Actualité :	
★★★★	Les subdivisions de recensement sont fiables	CNP4, SCIAN4 et variables sociodémographiques disponibles	Mensuelle	1 mois de délai	
★★★	Les divisions/agglomérations de recensement sont fiables	Un de ces éléments est moins détaillé ou manquant	Trimestrielle/semestrielle	6 mois de délai	
★★	Les régions métropolitaines/économiques de recensement sont fiables	Deux de ces éléments sont moins détaillés ou manquants	Annuelle	18 mois de délai	
★	Les provinces/territoires sont fiables	Les trois éléments sont moins détaillés ou manquants	< Annuelle	> 18 mois de délai	

Approche basée sur des enquêtes nouvelles ou élargies

Augmentation de la taille de l'échantillon de l'EPA

La première option à envisager est d'augmenter la taille de l'échantillon de l'Enquête sur la population active (EPA). L'élargissement de l'enquête (actuellement à 54 000 ménages) a toutefois ses limites. Étant donné sa structure actuelle, il est raisonnable de penser qu'en augmentant la taille de l'échantillon, on pourrait

estimer de façon fiable les indicateurs actuels pour les villes de 10 000 à 100 000 résidents (soit les agglomération de recensement ou AR). Cela améliorerait le caractère local par rapport aux estimations actuelles uniquement disponibles pour les régions économiques (RE) et les régions métropolitaines de recensement (RMR)¹ Ainsi, cette option améliore la localité de l'information tout en maintenant sa granularité et son actualité.

Élargissement de l'EPA pour ne poser les questions de base que tous les trois ou six mois

Une autre option consiste à poser un sous-ensemble des questions actuelles de l'EPA à un plus grand échantillon, mais seulement sur une base trimestrielle ou semestrielle plutôt que mensuelle. Cela permettrait également d'ajouter de nouvelles questions sur des sujets comme la qualité d'emploi, sans compromettre la collecte et le traitement mensuels des données de l'EPA actuelle. Cette option fournit des données plus granulaires, mais moins actuelles que celle de l'EPA originale. Elle pourrait être conçue de façon transversale (des échantillons différents à chaque cycle) ou longitudinale (des échantillons suivis pour quelques cycles).

Conception d'une nouvelle enquête de base axée sur les régions locales

La troisième option fondée sur une enquête est d'en concevoir une nouvelle, trimestrielle ou semestrielle, axée sur de plus petites régions géographiques. Les questions de base existantes de l'EPA seraient réutilisées afin de maintenir la comparabilité des résultats de l'enquête. Cette option améliorerait principalement le caractère local de l'information (ainsi que la granularité pour les questions de base). Les données de la nouvelle enquête seraient toutefois collectées moins fréquemment (tous les trois ou six mois plutôt que chaque mois).

Ajout de questions de base de l'EPA à une enquête existante

Enfin, nous avons également examiné la possibilité de tirer parti d'une structure existante d'échantillonnage en ajoutant un sous-ensemble de questions de l'EPA. Il serait important de se servir d'une enquête qui pourrait améliorer le caractère local de l'EPA sans compromettre la fréquence ni l'actualité (ni les résultats de l'enquête originale). Toutefois, il n'a pas été

possible de trouver une enquête appropriée qui répondrait à ces exigences, en particulier en ce qui a trait à l'amélioration du caractère local qui prévaut déjà dans les estimations de l'EPA.

On résume : l'approche par enquêtes

Parmi les options fondées sur des enquêtes, celle de l'élargissement de l'EPA est la plus prometteuse, du moins pour ce qui est d'améliorer les critères établis. Cette option utilise l'infrastructure d'enquête existante, et les nouvelles données, disponibles chaque mois, seraient comparables aux données historiques de l'EPA. Toutefois, la modification de la taille de l'EPA pourrait compromettre la fiabilité et l'actualité de l'enquête, en plus d'alourdir le fardeau de réponse. En outre, tout changement structurel apporté à l'EPA ne pourra être effectué qu'avant le prochain cycle d'échantillonnage, dont le début est prévu en 2025. Enfin, comme pour toutes les options basées sur des enquêtes, l'élargissement de l'EPA a un coût élevé et offre moins d'améliorations à l'ensemble actuel de données que les autres approches examinées ici.

Approche de couplage de données administratives

Couplage du recensement (questionnaire long) avec les fichiers fiscaux et du PVAE au fil du temps²

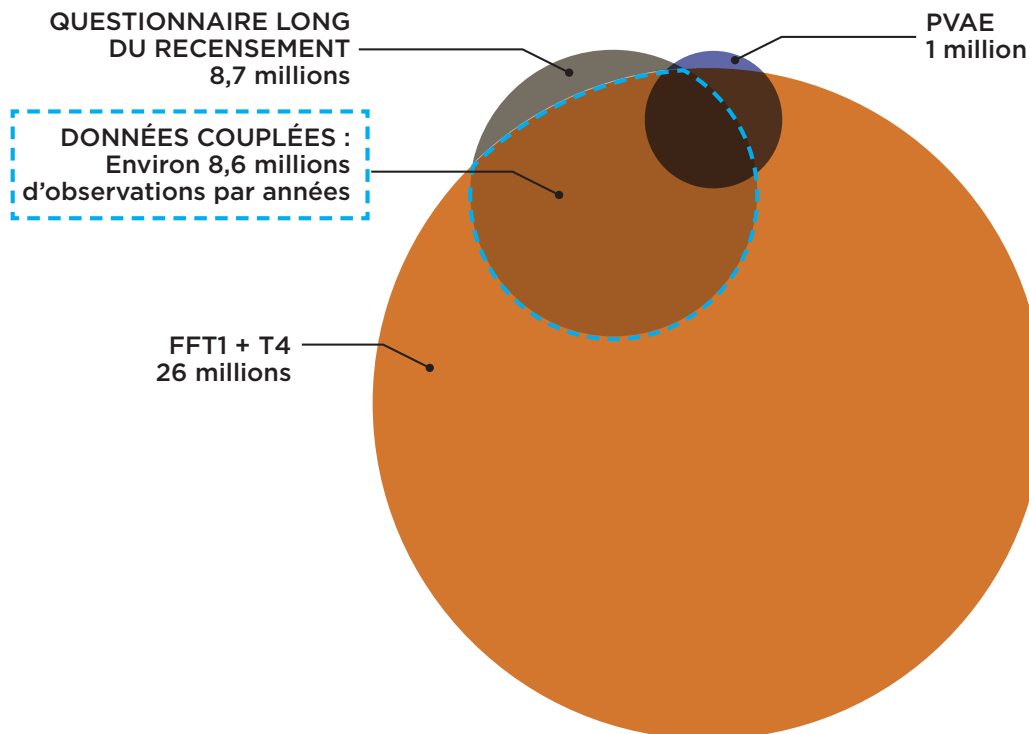
En reliant divers ensembles de données administratives, on tire profit des données de la plus grande qualité provenant de chaque source. Du plus grand au plus petit échantillon, les trois sources de données les plus prometteuses pour améliorer l'information sur le marché du travail sont les suivantes : 1) le Fichier des familles T1 (FFT1) et les fichiers T4, 2) le questionnaire long du recensement et 3) le fichier du Profil vectoriel de l'assurance-emploi (PVAE).

Le Fichier des familles T1 et les fichiers T4 contiennent des données anonymes sur tous les contribuables canadiens (environ 26 millions par année) que Statistique Canada obtient de l'Agence du revenu du Canada (ARC). Ces données contiennent des renseignements sur le revenu lors de chaque année civile, y compris le revenu d'emploi, les intérêts, les dividendes et les revenus de location. Le questionnaire long du recensement est un formulaire obligatoire distribué à 25 % des ménages canadiens en 2016, soit environ 8,7 millions de personnes. Il recueille des données granulaires sur de nombreux indicateurs démographiques, sociaux et économiques comme l'origine ethnique, le niveau de scolarité et l'activité actuelle. Enfin, le fichier du Profil vectoriel de l'assurance-emploi comprend les dossiers hebdomadaires des participants au programme de l'assurance-emploi (estimés à 1 million de personnes par année). En plus du soutien au revenu pour les chômeurs, le PVAE contient des

observations sur les personnes qui ont recours à l'assurance-emploi dans le cadre d'événements précis de leur vie (p. ex. : maladie, grossesse, proche aidant, etc.).

Bien que le questionnaire long du recensement soit l'enquête la plus vaste et détaillée au Canada (couvrant 8,7 millions de personnes en 2016), les dossiers fiscaux couvrent un nombre de personnes beaucoup plus important (environ 26 millions). Par conséquent, la taille de l'échantillon du recensement est un facteur limitant dans le couplage avec les fichiers FFT1 et T4 (voir la **figure 1**). Puis, lorsque les données du PVAE sont ajoutées, des détails supplémentaires sur les bénéficiaires de l'assurance emploi sont obtenus. Afin de ne pas limiter les données liées à celles concernant les bénéficiaires, il est recommandé que les individus n'ayant pas utilisé l'assurance emploi soient conservés dans l'ensemble de données, bien qu'avec moins de détails, tels que présentés dans la **figure 1**.

Figure 1 : Aperçu de l'option du couplage de données administratives



Il est important de noter que le couplage entre le questionnaire long du recensement et les fichiers administratifs fiscaux et du PVAE devrait être établi au fil du temps afin d'obtenir des observations entre les années de recensement.

On résumé : l'approche de couplage des données administratives

En raison de la taille considérable des ensembles de données administratives (particulièrement le recensement et les fichiers fiscaux), l'option du couplage de données administratives offre de l'information très localisée. De plus, puisque cette option tire profit de données existantes, elle est relativement peu coûteuse à mettre en œuvre. Le défi que posent les ensembles de données couplés est que les quatre critères d'IMT sont soumis à l'ensemble le plus limité. Par exemple, le nombre de fichiers fiscaux est assez important pour que des données fiables et anonymes puissent être produites au niveau des **subdivisions de recensement** ou des régions de tri d'acheminement (RTA)³. Toutefois, la granularité de l'information est limitée au sexe et à l'âge.

Par ailleurs, le recensement offre des données moins localisées, au niveau des **division de recensement** et des agglomérations de recensement, mais extrêmement granulaires : l'origine ethnique, la scolarité, l'activité et d'autres caractéristiques sont observables. La liaison de ces deux ensembles de données limite le caractère local au plus petit dénominateur commun (c'est-à-dire les divisions et les agglomérations de recensement).

Lors des années de recensement, les données liées du recensement et des fichiers fiscaux fournissent une information extrêmement détaillée. Pour toutes les autres années cependant, seules les caractéristiques fixes (p. ex. pays de naissance, ethnicité) provenant du recensement sont applicables aux observations provenant des fichiers fiscaux. Une solution à ce manque de granularité lors des années où il

n'y a pas de recensement est de mettre à profit d'autres sources d'information. Par exemple, des caractéristiques variables telles que le niveau d'éducation ou l'occupation provenant du PVAE ou de l'EPA pourraient être liées aux données des fichiers fiscaux lors des années où il n'y a pas de recensement. Ces sources de données sont plus fréquentes, mais couvrent une plus petite partie de la population et ne peuvent donc être liées qu'à un nombre limité d'observations au sein des fichiers fiscaux. Un exemple de ceci est présenté dans la figure 1 pour les cas où les données du PVAE sont liées aux données de recensement et de FFT1.

La principale limite de cette approche est son actualité. En raison de leur taille et de leur complexité, les données du recensement et des fichiers fiscaux sont présentement traitées de 12 à 18 mois après la période de référence. Le couplage de ces ensembles de données ajouterait à ce délai et chaque ensemble de données supplémentaire augmenterait encore davantage le temps de traitement.

Approche des méthodes de modélisation

Modèles d'estimation sur petits domaines

Cette option appliquerait les **estimations sur petits domaines** (EPD) aux données de l'EPA dans des régions où le nombre de répondants est trop faible pour permettre des estimations fiables de l'information sur le marché du travail. La technique consiste à recourir à une source de données complémentaire ou auxiliaire (avec un échantillon plus grand et des informations connexes) pour améliorer le caractère local des données de l'EPA. L'amélioration de la localité serait fonction des données auxiliaires. Par exemple, l'EPA contient de l'information sur les salaires regroupée par profession. Les estimations sur petits domaines utilisant l'EPA comme ensemble de données principal et les fichiers fiscaux comme ensemble auxiliaire offrent une information beaucoup plus

localisée par occupation. Il est important de noter que la fréquence des résultats des estimations sur petits domaines peut aussi être limitée par la fréquence des données de l'ensemble auxiliaire.

Techniques de modélisation avancées combinant recensement et EPA

La dernière option présentée ici est une technique de modélisation en deux étapes qui tire profit de la localité et de la granularité des données du questionnaire long du recensement 2016 ainsi que de la fréquence et de l'actualité des observations de l'EPA. À l'heure actuelle, des estimations fiables sur petits domaines (p. ex. pour les revenus) ne peuvent être faites directement à partir des données de l'EPA, puisqu'il y a trop peu de résultats. Inversement, le questionnaire long du recensement offre des résultats très locaux et granulaires, mais avec un écart de cinq ans entre les observations.

Pour surmonter ces limites, on pourrait explorer une combinaison d'estimation sur petit domaine et de techniques de prévision. On pourrait d'abord projeter les données du recensement jusqu'à présent (ce qu'on appelle aussi des prévisions immédiates) en incorporant des observations réelles tirées de données à jour comme l'EPA, les comptes macroéconomiques (p. ex. : le PIB provincial) et d'autres sources. Dans un deuxième temps, les variables projetées du recensement pourraient être utilisées comme données auxiliaires dans un modèle d'estimation sur petits domaines, ce qui améliorerait la localité et la granularité sans réduire la fréquence et l'actualité.

On résume : l'approche des méthodes de modélisation

Les options envisagées dans le cadre de cette approche utilisent des techniques de modélisation existantes et bien établies afin d'estimer de l'IMT granulaire pour des régions locales. Le modèle d'EPD est déjà mis en œuvre par Statistique Canada pour **estimer les taux de chômage dans les petites villes**. L'utilisation

des méthodes d'EPD à elles seules peut toutefois être quelque peu limitante, puisque les données auxiliaires sont le plus souvent des données administratives et impliquent donc les désavantages abordés précédemment, à savoir qu'elles sont généralement désuètes depuis plusieurs années et offrent peu de granularité sur le plan démographique ou du statut sur le marché du travail.

La deuxième option de modélisation tire profit de la granularité du recensement en établissant des prévisions jusqu'à présent. L'utilisation des données du recensement permettrait d'obtenir des ventilations beaucoup plus fines, y compris des estimations sur petits domaines pour les revenus, le sexe, l'activité et le niveau de scolarité, par exemple. De telles estimations ne seraient pas possibles avec des ensembles de données administratives standard. L'inconvénient ici est que l'approche n'a pas encore été testée, de sorte qu'il faudrait procéder à une évaluation et à une validation pour déterminer les meilleures techniques de prévisions à appliquer. De plus, toute approche de modélisation nécessitera une validation supplémentaire afin de réduire les incertitudes quant à la nature estimée de l'information, une question particulièrement préoccupante dans le milieu de la recherche. En plus des séries de données économétriques standard, de nouveaux algorithmes d'apprentissage machine devraient aussi être envisagés et testés.

Équilibrer tous les critères: test de viabilité des estimations sur petits domaines

Comme nous l'avons mentionné plus tôt, toute option dans le cadre de l'approche par enquêtes serait coûteuse, et les changements à l'EPA ne seront possibles que lorsque le prochain cycle d'échantillonnage débutera en 2025. L'approche fondée sur les données administratives, quant à elle, manque d'information granulaire entre les années de recensement. De plus, le couplage des données administratives avec celles du

recensement allongera le délai de traitement, ce qui réduira l'actualité des données. Compte tenu des inconvénients de ces options, particulièrement en comparaison avec les désavantages plus restreints des approches de modélisation présentées précédemment, nous croyons que les modèles d'estimation sur petits domaines, avec une certaine combinaison de méthodes de prévision et d'EPD, constituent des solutions nouvelles et efficaces pour produire des données locales et granulaires qui soient à la fois fréquentes et actuelles. De telles méthodes de modélisation ne sont pas trop coûteuses à explorer et à mettre en œuvre si elles s'avèrent fiables et efficaces.

Étant donné les avantages et les possibilités qu'offrent les options de modélisation, le CIMT et Statistique Canada poursuivent conjointement un projet de recherche dans le but d'explorer la faisabilité de nouvelles estimations sur petits domaines pour de nombreux indicateurs importants du marché du travail. Le défi principal quant aux modèles d'EPD est de trouver les données auxiliaires qui peuvent améliorer la fiabilité des estimations locales. Notons que la source de données auxiliaires varie selon la variable estimée. Par exemple, les données d'assurance-emploi ont été utilisées pour estimer les taux de chômage dans les agglomérations de recensement dans le cadre de l'EPA. Une autre possibilité est d'utiliser des données administratives provinciales et territoriales comme source auxiliaire de données pour les modèles d'EPD.

La voie à suivre

Le Division de la statistique du travail au Statistique Canada et le CIMT ont évalué une variété d'options pour offrir de l'IMT plus locale, granulaire, fréquente et actuelle. Trois grandes approches ont été ciblées : les méthodes par enquêtes, le couplage de données administratives et les techniques de modélisation.

Après avoir examiné les désavantages des trois approches, y compris les coûts et conséquences possibles sur l'infrastructure statistique existante, nous avons commencé à travailler sur un projet conjoint afin d'évaluer toutes les possibilités que les techniques de modélisation d'EPD peuvent offrir. La première étape consiste à tirer profit des couplages de données administratives et des nouvelles techniques de prévision pour améliorer l'actualité de l'information produite. De cette façon, nous pouvons tirer parti des avantages des principales sources d'information sur le marché du travail, comme la fréquence de l'Enquête sur la population active et la granularité du recensement.

À la suite de ce processus exploratoire, nous collaborerons avec nos intervenants pour veiller à ce que les estimations produites soient pertinentes, selon les critères présentés plus haut. Pour ce faire, les résultats doivent combler adéquatement l'écart dans la localité, la granularité, la fréquence et l'actualité de l'IMT. Lorsque tous ces travaux préliminaires seront terminés, nous évaluerons si cette méthode est généralisable (si elle s'avère utile) ou réévaluerons les options présentées ici et établirons une marche à suivre en conséquence.

Acknowledgements

Le présent numéro de *Perspectives de l'IMT* a été préparé conjointement par le CIMT et Statistique Canada. Nous aimerions remercier nos partenaires fédéraux, provinciaux et territoriaux pour leurs commentaires et suggestions sur une version antérieure de ce rapport. L'équipe tient à souligner la précieuse contribution de Ted McDonald, de l'Université du Nouveau-Brunswick, et de Fraser Summerfield, de l'Université St Francis Xavier. Pour en savoir plus sur cette édition de *Perspectives de l'IMT* ou les autres activités du CIMT, veuillez consulter notre page de publications ou encore communiquer avec Behnoush Amery, économiste principale, à behnoush.amery@lmic-cimt.ca, Tony Bonen, directeur de la recherche, des données et de l'analytique, à tony.bonen@lmic-cimt.ca, Josée Bégin, directrice de la Division de la statistique du travail, à josee.begin@canada.ca ou Vince Dale, directeur adjoint de la Division de la statistique du travail, à vincent.dale@canada.ca.

Notes :

- 1 Les RMR et les AR représentent les centres urbains. Les RMR sont de grands centres urbains ayant une population de 100 000 ou plus. Le Canada compte actuellement 35 RMR. Les AR sont de plus petits centres urbains comptant toutes entre 10 000 et 100 000 habitants. Il en existe présentement 114 au Canada. Les 76 régions économiques (RE) du Canada varient considérablement en superficie et en population. Le Yukon est la RE la moins peuplée, avec environ 36 000 personnes, et Toronto compte plus de 6,2 millions d'habitants (2016).
- 2 À proprement parler, le recensement n'est pas une base de données administratives. C'est un sondage obligatoire. Toutefois, dû à son ampleur et importance, il partage un grand nombre de caractéristiques avec les données administratives. Donc, nous le présenterons en discutant du couplage de données administratives.
- 3 Les régions de tri d'acheminement (RTA) correspondent aux trois premiers caractères d'un code postal.