Regional Analytics Laboratory **PRÉSENTE**

Janvier 2022

DES DONNÉES AUX DÉCISIONS :
Qualité des données

Présentation appelée en anglais : « Data to Decisions : Data Quality »

Rédigée par Meghan Eibner et Jamie Ward

Introduction

L’initiative des données aux décisions est présentée par le Regional Analytics Laboratory (RAnLab) et est parrainée en partie par le Future Skills Centre.

RAnLab est l'unité de données et d'analyse du Leslie Harris Center of Regional Policy and Development de la MemorialUniversity. RAnLab analyse les données et la géographie pour fournir un aperçu et une modélisation prospective pour des éléments tels que la démographie, l'offre de main-d'œuvre, les demandes de services, les prix des produits de base et d'autres indicateurs socio-économiques. Certains exemples de leur travail comprennent la production de projections démographiques communautaires et régionales à long terme, l'évaluation de la demande locale de logements et la réalisation d'analyses et de modélisations de données locales détaillées aux municipalités et aux régions, fournissant des informations essentielles pour la prise de décision fondée sur des faits.

Le but de l’initiative des données aux décisions est d'aider les personnes au Canada qui proviennent de divers horizons à apprendre comment appliquer les données à leurs propres projets. L’initiative utilise un langage simple, des exemples et des présentations vidéo pour améliorer l'expérience d'apprentissage.

Si vous avez des questions, veuillez contacter Meghan Eibner (meibner@mun.ca) ou Jamie Ward (jward@mun.ca).

Tout au long de ce chapitre, vous verrez des boutons cliquables qui renvoient à des pages Web contenant des informations supplémentaires.

Le bouton de droite vous mènera à la présentation « Des données aux décisions » sur la qualité des données.

La présentation n'est disponible qu'en anglais.

Table des matières

Qualité des données

Lors de la collecte et de l'utilisation des données, il est important d'être conscient des différents problèmes qui peuvent avoir un impact sur la qualité, et donc l'interprétation, des données.

Certains des problèmes de qualité des données les plus courants sont liés à la disponibilité, à l'uniformité et à l'exactitude.

Il est important d'effectuer des contrôles de la qualité de toutes les données recueillies et de documenter tous les problèmes de données que vous relevez (et toutes les mesures prises pour résoudre ces problèmes). Cela permet de garantir que l'analyse et l'interprétation sont aussi transparentes que possible.

Lors de l'examen d'un ensemble de données pour déterminer les problèmes potentiels, il convient d'être attentif aux éléments suivants :

- changements importants dans les données ;
- point de données aberrant ;
- interruptions ou lacunes dans les données.

Des exemples de ces problèmes de qualité des données, utilisant des données hypothétiques, sont présentés dans les pages suivantes.

Disponibilité des données

La mesure dans laquelle les données sont facilement accessibles aux utilisateurs quand et où ils en ont besoin.

Uniformité des données

La facilité d'utilisation des données applicables. Le processus de maintien de l'uniformité des données lorsqu'elles sont déplacées dans les diverses applications et entre ces dernières.

Exactitude des données

La mesure dans laquelle les données sont exemptes d'erreurs et peuvent être utilisées comme une source d'information fiable.

Changement dans les données:

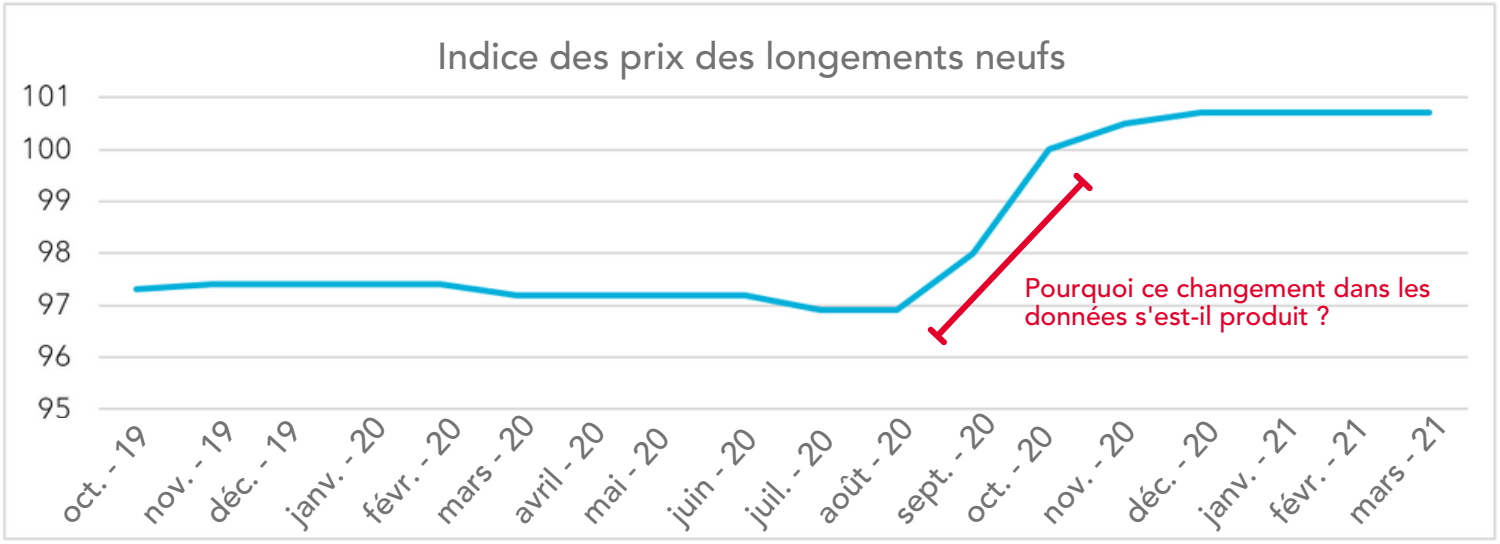
Le graphique ci-dessous montre une série de données chronologiques avec un changement observable dans les données en août 2020. Avant cette date, les données étaient relativement stables depuis un certain temps. Après une forte augmentation entre août 2020 et novembre 2020, les données se stabilisent à nouveau à un niveau beaucoup plus élevé.

Un changement aussi radical dans les données doit faire l'objet d'une enquête. Nous devons poser des questions pour tenter de déterminer la cause de ce changement, par exemple :

- S'est-il passé quelque chose sur le marché du logement ou dans l'économie pour justifier une augmentation légitime des données ?
- Y a-t-il eu un changement dans la définition de l'indice des prix des logements neufs autour d'août 20 pour entraîner un changement dans l'ensemble des données ?

Changement dans les données

Un changement dans la distribution des données (p. ex. un changement de classification ou de géographie).



Donnéesaberrantes :

Le graphique ci-dessous est un exemple d'un ensemble de points de données aberrant. Tous les points de données, à l'exception de de pour T.-N.-L., suivent un modèle similaire.

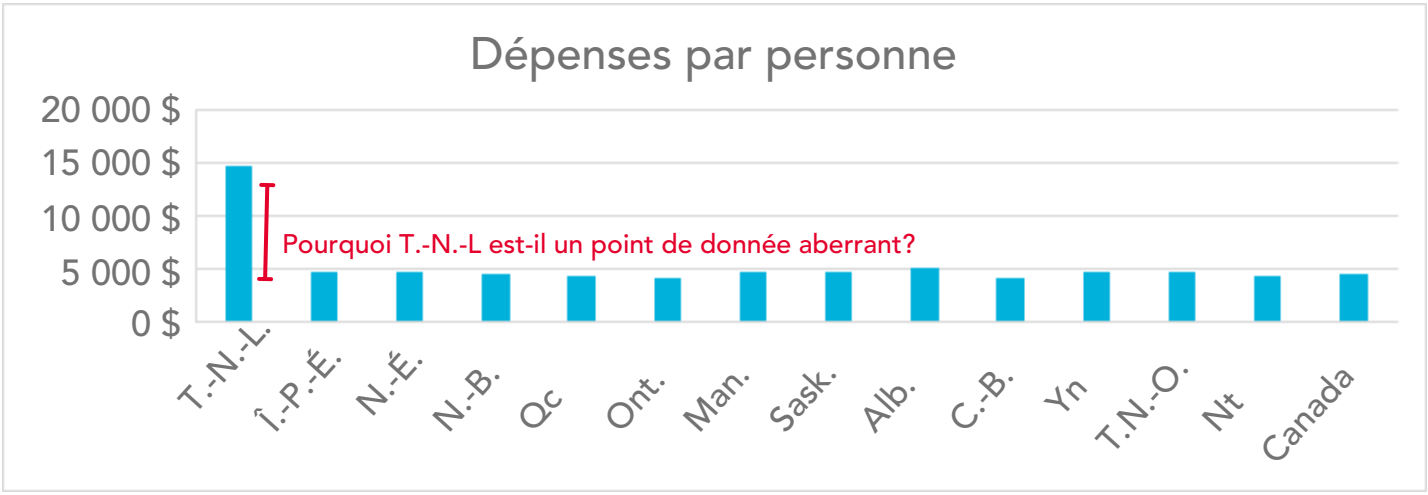
Le point de données aberrantes doit faire l'objet d'une recherche pour savoir s'il est le résultat d'une erreur ou s'il représente précisément ce qu'il mesure.

Par exemple, nous devons nous demander :

- Est-ce qu'il s'est passé quelque chose à T.-N.-L. à l'époque qui a entraîné des dépenses beaucoup plus élevées que dans le reste du pays ?
- Les données de T.-N.-L. ont-elles été mal saisies dans la base de données, ce qui a donné lieu à un point de données très différent de tous les autres ?

Point de données aberrant

Un point de données qui diffère significativement des autres observations



Lacunes dans les données:

Le tableau ci-dessous est un exemple d'un ensemble de données comportant une lacune; les données pour l'année 2006 sont manquantes. Il convient d'en rechercher la raison :

- S'agit-il simplement d'une erreur : les données pour cette année ont-elles été accidentellement omises ?
- Les données pour cette année sont-elles non disponibles? Si oui, pourquoi ?
- L'enquête n'a-t-elle pas été menée pour cette année-là ?
- Les pratiques de confidentialité exigeaient-elles que les données de cette année-là soient supprimées ?

Lacunes dans les données
Les données pour des éléments ou des ensembles de données particuliers sont sciemment ou inconsciemment manquantes.

Année	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Population	366	376	388	400	412		425	429	438	444	454

Pourquoi y a-t-il un manque de données ?

S'il est impossible de répondre à des questions concernant les données, il convient d'utiliser ces données avec une extrême prudence.

Les données publiées sans documentation ou avec des questions sans réponse peuvent en dire long sur la crédibilité/fiabilité de la source.

Si vous ne comprenez pas vos données d'entrée, il est peu probable que vous compreniez vos résultats.

Disponibilité des données

La disponibilité des données peut varier selon les périodes, la géographie et le sujet. Par exemple, un ensemble de données provenant d'une enquête introduite dans les années 1980 n'existera pas avant cette période car ces données n'avaient pas encore été collectées. Le format dans lequel les données sont disponibles peut également changer avec le temps. La plupart des données de Statistique Canada dans l'histoire récente sont disponibles en ligne pour téléchargement dans un format utilisable. Toutefois, certaines données historiques plus anciennes (données historiques de recensement, par exemple, antérieures aux années 1970) peuvent n'être disponibles qu'en format PDF numérisé, ce qui peut rendre la sélection et l'utilisation des données difficiles.

Si le niveau géographique ou organisationnel d'un ensemble de données est suffisamment petit pour qu'il soit possible de reconnaître des réponses individuelles à une enquête, la Loi sur la statistique exige que ces données soient supprimées. Cette mesure vise à protéger la vie privée des personnes et à garantir que les données recueillies sont privées, sécurisées et confidentielles.

Il en résulte des lacunes occasionnelles où les données ne sont pas disponibles en raison de leur suppression. Par exemple, un ensemble de données particulier peut être disponible pour toutes les municipalités d'une région, sauf une, en raison de critères de confidentialité.

Il se peut aussi que des années de données manquent dans un ensemble de données chronologiques. Dans ce cas, si le manque de données est relativement faible, il peut être possible de le combler en faisant une hypothèse sur la nature des données manquantes.

Statistique Canada indique quand les données ont été supprimées. Vous pouvez voir un numéro en exposant à côté d'un point de données qui correspond à une note sous l'ensemble de données. Voir la page 14.

Interpolation

Le processus consistant à combler un manque de données sur la base d'une hypothèse est appelé « interpolation ». Les exemples suivants montrent trois façons différentes de combler une année de données manquante dans un ensemble de données en utilisant trois hypothèses différentes. Notez qu'il s'agit d'un exemple simple avec une seule année de données manquantes. Plus il y a de lacunes dans un ensemble de données, plus le processus pour tenter de les combler est compliqué.

Interpolation
Le processus consistant à combler un manque de données en utilisant des hypothèses raisonnables, justifiables et bien documentées

Année	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Population	366	376	388	400	412		425	429	438	444	454

Comment pouvons-nous combler ce manque ?

Hypothèse 1 : Moyenne de deux points

Les données pour l'année 2007 sont manquantes. Une façon de combler les données manquantes est de prendre simplement la moyenne entre les deux points de données situés de part et d'autre de l'année manquante. La population de 2006 était de 412 habitants et celle de 2008 de 425 habitants. On peut donc supposer que la population de 2007 était de 419 habitants.

$$\frac{412 + 425}{2} = \frac{837}{2} = 419$$

Hypothèse 2 : Variation annuelle moyenne en pourcentage

La deuxième façon de combler les données manquantes est de déterminer la variation annuelle moyenne en pourcentage de la population et de l'utiliser pour calculer les données pour 2007.

Année	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Population	366	376	388	400	412		425	429	438	444	454
Variation annuelleen %		2,7%	3,2%	3,1%	3,0%			1,7%	2,1%	1,4%	2,3%

Calculez la variation annuelle en pourcentage pour chaque année pour laquelle des données sont disponibles (voir le tableau ci-dessus), puis prenez la moyenne de cet ensemble de données calculées. Dans ce cas, la variation annuelle moyenne en pourcentage est de 2,3 %.

$$moyenne \quad (2{,}7\%,\ 3{,}2\%,\ 3{,}1\%,\ 3{,}0\%,1{,}7\%,\ 2{,}1\%,1{,}4\%,\ 2{,}3\%) = \quad 2{,}3\%$$

On peut alors supposer que la population de 2006 a augmenté de 2,3 %, ce qui donne une population de 422 habitants en 2007.

$$412 * (1 + 2{,}3\%) \quad = \quad 412 * (1 + 0{,}023) \quad = \quad 412 * 1{,}023 \quad = \quad 422$$

Hypothèse 3 : Variation linéaire en pourcentage

Cette approche est similaire à celle de l'exemple précédent, sauf qu'au lieu de calculer la variation en pourcentage pour chaque année, on suppose une seule variation en pourcentage sur l'ensemble de la période. Pour commencer, calculez la variation en pourcentage entre la première année de données et la dernière année de données dans l'ensemble de données. La variation en pourcentage entre 366 (en 2002) et 454 (en 2012) est de 24,0 %.

$$\frac{454 - 366}{366} = \frac{88}{366} = 0{,}24 \quad ou \quad 24\,\%$$

Divisez ensuite ce pourcentage par le nombre d'années de l'ensemble de données (dans cet exemple, 11 ans) pour obtenir le pourcentage de variation linéaire supposé. Le résultat est de 2,2%.

$$\frac{0{,}24}{11} = 0{,}022 \quad ou \quad 2{,}2\,\%$$

Ce changement de 2,2 % peut maintenant être appliqué à la population de 2006 pour calculer le point de données manquant de 2007. On obtient ainsi une population présumée de 421 personnes en 2007.

$$412 * (1 + 2{,}3\,\%) = 412 * (1 + 0{,}022) = 412 * 1{,}022 = 421$$

La méthode utilisée pour combler une lacune dans les données dépend des caractéristiques des données de l'ensemble de données :

- L'hypothèse de la ligne droite fonctionne bien avec des données uniformes qui suivent un modèle prévisible, ce qui permet de supposer que la variation annuelle en pourcentage entre la première et la dernière année de données est relativement stable.
- L'approche du pourcentage annuel moyen peut mieux refléter les changements au sein d'un ensemble de données qui est moins uniforme, mais encore relativement prévisible (c.-à-d. sans valeurs aberrantes ou grands changements).
- L'approche de la moyenne en deux points peut fonctionner mieux avec des données plus volatiles afin de saisir les changements d'une année sur l'autre

Exactitude des données

L'exactitude des données peut devenir un problème lors de la saisie, de la copie, de la conversion ou du transfert de données. Lorsque des erreurs se produisent dans les données, celles-ci peuvent ne plus être une source d'information fiable. Il est donc essentiel d'effectuer des contrôles de qualité sur toutes les données qui ont été manipulées. Un exemple d'erreur humaine entraînant l'inexactitude des données peut être une simple faute de frappe lors du transfert de données PDF numérisées dans un format numérisé (p. ex. Excel).

Exactitude des données
La mesure dans laquelle les données décrivent correctement ce qu'elles ont été conçues pour mesurer

TRUC

Les points de données aberrants peuvent être un indicateur de problèmes d'exactitude des données.

Uniformité des données

Des problèmes d'uniformité peuvent survenir lors de la collecte de données provenant de sources, de provinces ou territoires et de périodes qui diffèrent. Il est important d'examiner la définition et la description des ensembles de données recueillies afin de déterminer s'ils sont comparables et peuvent être utilisés comme un ensemble de données valide et uniforme. Même si un ensemble de données porte le même nom (p. ex. "Dette nette"), il peut être défini différemment selon les zones géographiques, les administrations et les sources.

Uniformité des données
La facilité d'utilisation des données connexes. Le processus de maintien de l'uniformité des données lorsqu'elles sont déplacées dans les applications et entre ces dernières.

TRUC De grands changements dans les données peuvent être un indicateur de problèmes d'uniformité des données.

Un ensemble de données recueillies auprès d'une source unique peut également présenter des changements de définition au fil du temps. Par exemple, les données sur la population de la ville << x >> peuvent être disponibles pour les dernières décennies, mais, au cours de cette période, la ville a fusionné avec des communautés voisines, ce qui entraînerait à la fois un changement dans les données, puisque les chiffres de population sont maintenant combinés en un seul, et un changement dans la définition de l'ensemble de données lui-même.

Pertinence des données

Une fois que vous avez effectué le contrôle de la qualité de vos données et que vous êtes prêt à les utiliser, vous devez également vous assurer que vous les utilisez de manière appropriée. L'un des problèmes les plus courants résultant d'une utilisation inappropriée des données est appelé le sophisme écologique.

Il existe plusieurs erreurs spatiales pouvant résulter d'une utilisation inappropriée des données, mais l'erreur écologique est la plus courante.

Lorsque des données à un niveau géographique spécifique ne sont pas disponibles, il peut être tentant d'appliquer des données provenant d'autres zones/niveaux géographiques. Par exemple, appliquer des données provinciales ou nationales à un niveau local. Pour la majorité des localités, le tableau dressé par les données nationales est très différent de la réalité locale. Utiliser des données de cette manière conduira presque certainement à une interprétation trompeuse. C'est ce qu'on appelle le sophismeécologique.

Vous souhaitez en savoir plus sur les sophismes? Essayez de faire des recherches sur les sujets suivants :

- **Longueur connue**
- **Erreur de localisation**
- **Sophisme atomique**
- **Problème de l'unité aréolaire modifiable**

Nous traiterons du sophisme écologique dans la section suivante.

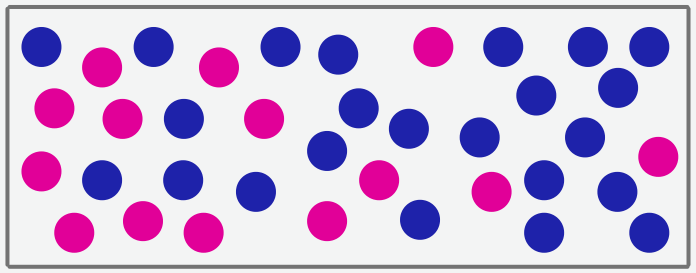
Sophisme écologique

Le sophisme écologique est une erreur qui se produit lorsque des conclusions sont tirées à propos d'individus ou de petites zones sur la base de données relatives à un groupe ou une zone géographique plus large. Supposer que ce qui est vrai pour un groupe l'est également pour les membres individuels du groupe peut entraîner une analyse et une interprétation trompeuses ou inexactes des données. C'est pourquoi il est essentiel d'utiliser des données au niveau géographique approprié.

Exemples :

- Si une province a une population en déclin, il est incorrect de supposer que toutes les communautés de cette province connaissent un déclin démographique.
- Si une communauté est composée d'une importante population hautement qualifiée, il est incorrect de supposer que chaque personne de cette communauté est hautement qualifiée.
- Si une région a un taux de chômage élevé, il est incorrect de supposer que toutes les communautés de cette région ont un taux de chômage élevé.

Région1



= 10 personnes avec une formation post-secondaire

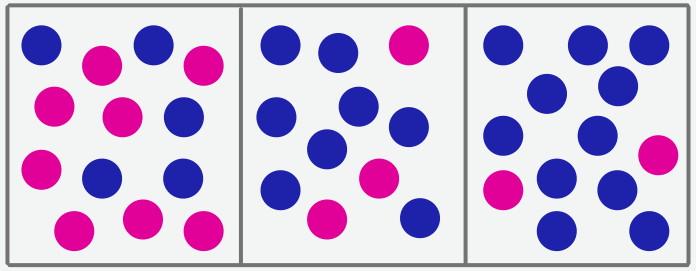
= 10 personnes sans formation post-secondaire

Ce que nous savons :

- Les statistiques montrent que la majorité des personnes vivant dans la région 1 ont suivi des études postsecondaires
- La région 1 est composée de trois villes

Peut-on supposer que la majorité des personnes vivant dans chaque ville ont suivi des études postsecondaires? Non, ce raisonnement est un sophisme écologique.

Région 1



En examinant les données au bon niveau géographique, nous pouvons constater que la majorité des personnes vivant dans la ville A n'ont pas suivi d'études postsecondaires

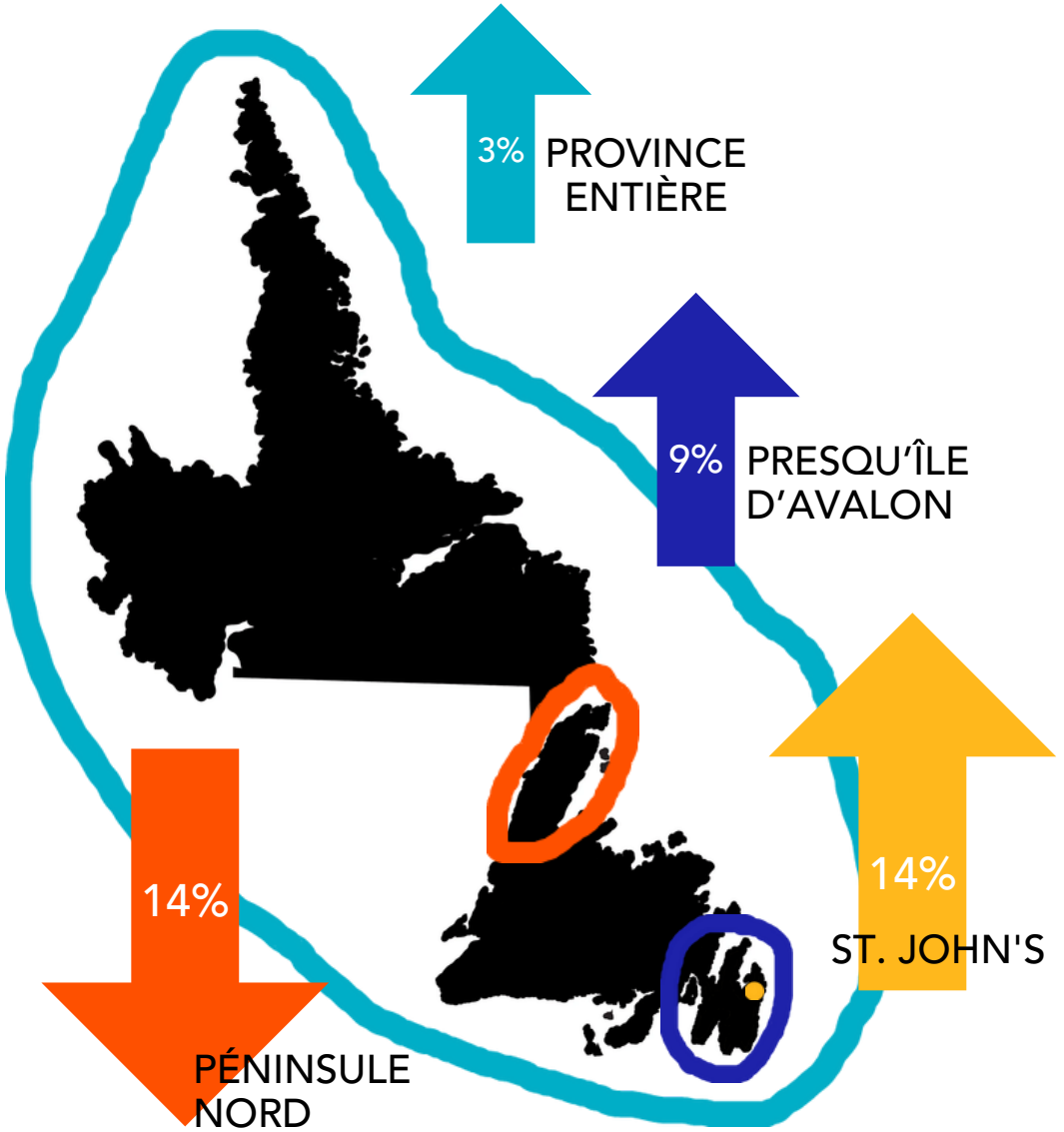
Exemple de sophismeécologique:

De 2006 à 2016, la province de Terre-Neuve-et-Labrador a connu une croissance lente, la population totale n'ayant augmenté que de 3% au cours de cette période de dix ans.

La péninsule d'Avalon a toutefois connu une croissance démographique de 9% au cours de cette même période et la population de St. John's a augmenté de14 %.

Pendant ce temps, la population de la péninsule nord a diminué de 14%.

Cet exemple montre comment les données provinciales peuvent ne pas refléter avec précision les expériences plus nuancées des petites régions géographiques de la province.



Ressources supplémentaires

Statistique Canada a élaboré ses six dimensions de la qualité des données, qui donnent plus de détails sur les facteurs pris en compte pour déterminer la qualité des données.

Cliquez ici pour plus de détails sur les six dimensions de la qualité des données

En fonction de ces critères, Statistique Canada signalera les ensembles ou points de données en fonction de leur pertinence. Par exemple, un point de données peut être marqué comme « à utiliser avec prudence » ou « non fiable ». Ces notes doivent être prises en considération, et documentées, lors de la collecte et de l'utilisation de données marquées.

Cliquez ici pour consulter la boîte à outils sur la qualité des données de Statistique Canada

Statistique Canada a également mis au point une boîte à outils sur la qualité des données qui vise à sensibiliser aux pratiques d'assurance de la qualité des données.

Résumé du chapitre

Des données aux décisions

Qualité des données

- L'analyse de données de qualité repose sur l'utilisation de données de qualité. Il est donc important de s'assurer que les données sont disponibles, uniformes et exactes.
- Les problèmes de qualité des données à surveiller comprennent les changements importants dans les données, les points de données aberrants et les lacunes dans les données.
- Dans certains cas, ces problèmes peuvent être résolus par interpolation, en utilisant des hypothèses documentées et justifiables pour combler les données manquantes.
- Parmi les méthodes d'interpolation figurent la moyenne à deux points, la variation annuelle moyenne en pourcentage et la variation linéaire en pourcentage.
- Le sophisme écologique, une erreur qui se produit lorsque l'on suppose que ce qui est vrai pour un groupe est vrai pour une personne au sein de ce groupe, ce qui peut entraîner une analyse et une interprétation des données trompeuses.