Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Composer avec des données administratives, volumineuses et d'enquête : une évaluation de la qualité des bases de données des terres humides au Canada

Herbert Nkwimi Tchahou, Claude Girard et Martin Hamel [1]

Résumé

Bien que les milieux humides occupent seulement 6,4% de la superficie de notre planète, ils sont primordiaux à la survie des espèces terrestres. Ces écosystèmes requièrent une attention toute particulière au Canada puisque près de 25% de leur superficie mondiale se retrouve en sol canadien. Environnement Canada (EC) possède des méga-bases de données où sont rassemblées toutes sortes d'informations sur les milieux humides provenant de diverses sources. Avant que les informations contenues dans ces bases de données ne puissent être utilisées pour soutenir quelque initiative environnementale que ce soit, elles se devaient d'abord d'être répertoriées puis évaluées quant à leur qualité. Dans cet exposé, nous présentons un aperçu du projet pilote mené conjointement par EC et Statistique Canada afin d'évaluer la qualité des informations contenues dans ces bases de données, elles qui présentent à la fois certains des attributs propres aux données volumineuses (« *Big Data* »), aux données administratives et aux données d'enquête.

Mots Clés : Analyse de données ; Évaluation de qualité ; Environnement.

1. Introduction

Traditionnellement dans les sondages, les données exploitées afin de produire des inférences étaient issues d'un processus planifié de collecte auprès d'un échantillon (probabiliste) d'une population d'intérêt. Puis, progressivement, des données externes au processus de collecte sont devenues de plus en plus disponibles. Ainsi, ces données sont utilisées dans les inférences bien qu'elles aient initialement été recueillies à d'autres fins. La réduction du fardeau de réponse, la qualité des données externes disponibles et les contraintes budgétaires sont des raisons couramment évoquées pour justifier ce changement de cap. Par conséquent, l'utilisation des dossiers administratifs, nécessaires à l'administration de divers programmes non statistiques, est devenue une option de plus en plus exploitée par les agences statistiques nationales pour compléter, voire remplacer les données d'enquêtes.

Une autre option qui suscite l'intérêt des sondeurs concerne l'utilisation des données volumineuses. Pour cause, la révolution des technologies de l'information et l'utilisation accrue des réseaux sociaux qu'on a connue ces dernières années ont rendu disponible une masse importante de données variées qui ont la capacité de se renouveler constamment. On assiste à une transformation profonde dans la pratique des sondages : le statisticien d'enquête est de plus en plus appelé à produire des statistiques à partir à la fois de données d'enquête, de données administratives et de données volumineuses. Un exemple de cela est fourni par une activité de consultation menée par Statistique Canada pour le compte de Environnement Canada (EC) afin d'évaluer la qualité des informations contenues dans des méga-bases de données que possède EC en lien avec les terres humides. Pour se retrouver parmi la masse d'informations que contiennent ces bases, nous avons utilisé les attributs propres aux données volumineuses, aux données administratives et aux données d'enquête comme repères.

Nous présentons ici le travail de réflexion ayant précédé l'analyse statistique comme telle, qui était l'objet premier de cette consultation; il montre comment nous avons composé avec ces trois types de données dans un contexte de consultation statistique. Dans la prochaine section, nous discuterons brièvement de la consultation statistique, qui est une activité en marge des activités traditionnelles d'enquêtes que nous menons à Statistique Canada. Ensuite, nous présenterons les défis qu'on peut rencontrer en présence de données très riches mais non structurées, et comment on peut s'y retrouver en se fixant trois grands repères. Une fois ce cadre dressé, nous présenterons, en guise d'application, la consultation que nous avons menée pour le compte d'EC sur les terres humides. Et puisque les

[1]Division des Méthodes d'enquêtes auprès des entreprises, Statistiques Canada, 100, promenade Tunney's Pasture, Ottawa ON, Canada, K1A 0T6.

terres humides ne sont pas un sujet courant, nous dirons d'abord quelques mots sur ce sujet, avant de nous appesantir sur l'analyse et la synthèse de l'information disponible.

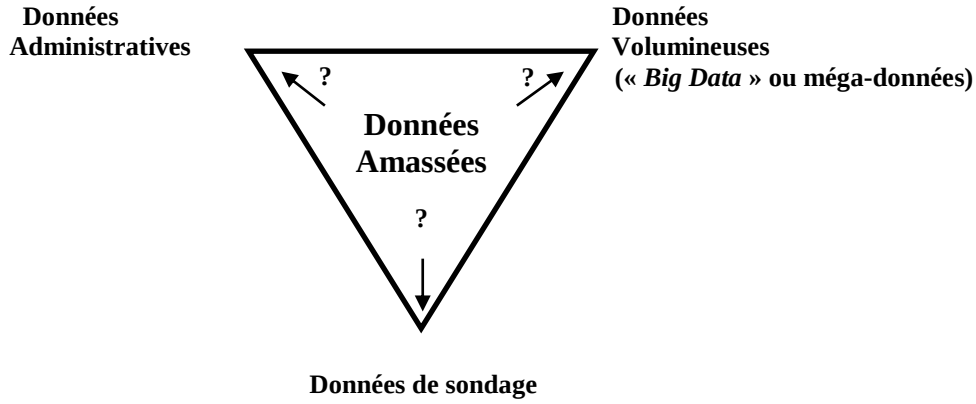
2. L'ABC d'une consultation

D'une manière générale, le déroulement d'une consultation statistique peut s'articuler autour de trois points principaux, chacun recelant un défi particulier. La première étape d'une consultation consiste à comprendre la question scientifique posée. En général, c'est une question qui parfois est plus ou moins claire même pour le client sollicitant la consultation. Une ou plusieurs rencontres peuvent être nécessaires pour cerner la problématique. Une fois le problème saisi, la deuxième étape consiste à trouver une modélisation statistique adéquate. Il faut s'assurer qu'on a des outils statistiques (explorer les différentes méthodes) et logistiques (disponibilité des logiciels) pour répondre à la question posée. Ceci dépend fortement des données disponibles et leur fiabilité (cohérence des données, valeurs manquantes, valeurs aberrantes) ainsi que les ressources allouées au projet. À cette étape, le consultant doit trouver un compromis entre efficacité et opérationnalité, et surtout établir un contrat avec le client dans lequel il est expliqué de manière détaillée ce qu'il envisage de faire et s'assurer avec celui-ci que cette solution répond à ses objectifs. Une fois les analyses statistiques terminées, la dernière étape, non moins importante, est la communication des résultats au client. Cette étape, qui peut paraître évidente de prime à bord, peut s'avérer la plus ardue de tout le processus. Les groupes de travail sont souvent constitués de personnes de sensibilité scientifique différente. Le rapport doit alors être allégé du contenu très technique afin d'être accessible à tous les intervenants

Comme nous l'annoncions à l'introduction, l'avènement du « *Big Data* », apporte un défi supplémentaire aux analystes statistiques en général. L'utilisation à des fins statistique des données conçues et collectées pour un usage administratif par exemple ne peut se faire sans surmonter un certain nombre d'enjeux préalables. Avant, les chercheurs collectaient des données s'articulant autour d'une question de recherche claire et précise. Les analyses statistiques qui s'ensuivaient étaient somme toute naturelles bien que parfois compromises par la manière dont la collecte a été planifiée. Dans le contexte actuel, le principal défi n'est donc plus l'absence de données au départ, mais l'avalanche de données maintenant disponibles : comment s'y retrouve-t-on? Pour passer de cet état brut d'information à de l'information synthétisée et exploitable, il faut d'abord être capable de s'y retrouver. Nous voulons insister sur le fait que cette abondance de données peut aussi bien conduire à de l'inférence de très grande qualité, comme à des conclusions totalement erronées si on fait fi de certaines précautions préalables. Dans la section suivante, nous abordons la question de comment s'y retrouver dans toute cette masse d'information en introduisant trois grands repères.

3. Comment se retrouver dans une masse de données très riches mais non structurées?

Dans cette section, nous abordons une question fondamentale et centrale dans toute démarche statistique : la nature des données à analyser. Elle est déterminante dans le choix des méthodes à utiliser. À quel type de données avons-nous à faire? Cette question est d'autant plus d'actualité lorsque la planification de la collecte des données échappe complètement au contrôle du statisticien. Pour se retrouver dans une masse importante d'information, nous pensons qu'il est crucial pour le statisticien de bien comprendre les attributs des données qui lui sont soumis afin de juger de leur qualité et du rôle qu'ils pourront jouer dans une éventuelle analyse statistique. Ceci permettra aussi de mieux préciser le but poursuivi par le chercheur et les limites d'utilisation de telles données. Une façon d'y parvenir est de se fixer par rapport aux trois grands repères suivants :



Chacun des types de données précédentes va jouer un rôle différent dans une analyse statistique. Par exemple, les données administratives vont généralement être renseignées pour toutes les unités de la base (à quelques données manquantes près) tandis que les données de sondage sont limitées à l'échantillon. Si dans le premier cas il est assez direct de produire une estimation d'un paramètre de la population d'intérêt à partir du fichier administratif, le second cas est beaucoup moins immédiate. L'inférence sur les quantités d'intérêt est menée en utilisant uniquement les unités de l'échantillon et de l'ensemble des poids de sondage qui leurs sont associés. Afin de dissiper tout doute, il est important de préciser la représentation que nous nous faisons de chacun de ces types de données avant d'amorcer le cas pratique.

3.1 Données administratives

Les données administratives sont des données qui sont collectées le plus souvent pour des raisons de service et non à des fins statistiques. Comme exemple, l'Agence du revenu du Canada renseigne et met régulièrement à jour des informations au sujet des citoyens et résidents afin de faciliter le recouvrement des taxes. L'utilisation des données administratives pour des fins de statistique est une pratique de plus en plus attrayante parce qu'elle permet de réaliser un gain énorme en coût de collecte et permet aussi une réduction du fardeau de réponse. Cependant les concepts utilisés dans ces données administratives sont propres au service à rendre et donc diffèrent souvent d'une source à une autre. Ceci soulève un enjeu potentiel lié à leur qualité, par exemple la couverture de la population d'intérêt. Comme le mentionne Michael Brick (2011), l'utilisation des données administratives depuis de nombreuses années n'a pas produit l'effet escompté. Faisant référence à Jabine et Scheuren (1985) qui avaient décrit six objectifs fixés par le gouvernement américain pour améliorer l'utilisation des données administratives sur une période de dix ans, il constate que vingt-cinq ans plus tard, certains de ces objectifs ne sont toujours pas atteints.

3.2 Données volumineuses

Les données volumineuses sont de plus en plus présentes dans les enquêtes. On les décrit souvent en recourant aux trois V : Volume, Variété et Vélocité. Douglas Laney fut l'un des premiers auteurs à employer le concept trois V en 2001 pour décrire ce type de données. Le volume dans cette description caractérise le fait que ces données sont une avalanche d'informations qui nous tombe dessus (de l'ordre du zettaoctet : 10^{21} octets). Puisque les données analysées ne sont plus forcément structurées comme dans les analyses statistiques traditionnelles, mais peuvent être du texte, des images, du contenu multimédia, des traces numériques, des objets connectés, etc., on parle de données variées. Les données volumineuses à certains égards sont similaires aux données administratives dans le sens qu'ultimement les concepts mesurés par ces données ne correspondent pas nécessairement à ceux mesurés par les enquêtes. Malgré les promesses qu'elles revêtent quant à la réduction des coûts de collecte, elles peuvent occasionner des coûts quant au stockage de l'information et la difficulté du traitement et de l'extraction de l'information. Enfin, la vélocité décrit la capacité qu'ont les données volumineuses de se renouveler très rapidement. Dans un contexte d'inférence, les données volumineuses soulèvent également des enjeux liés à la qualité comme encore la couverture ou la cohérence.

3.3 Données de sondage

Ces données sont orientées vers l'analyse statistique. Elles sont recueillies autour des concepts clairs pour des besoins de sondage et, règle générale, leur collecte est planifiée et bien contrôlée assurant une qualité élevée. C'est le type de données le plus couramment rencontré à Statistique Canada. À l'issu de la collecte, les données sont formatées, structurées et vérifiées dans un processus automatisé. Le produit final est généralement de bonne qualité et prêt à l'emploi pour des fins d'inférences. Le principal bémol vient du fait que les unités sont renseignées pour une partie de la population seulement : l'échantillon. Il faut par la suite faire de l'inférence sur la population entière. Ce qui induit généralement une erreur dans les estimations due à l'échantillonnage.

4. Des principes à la pratique: Étude de cas

Dans cette section, nous présentons en guise d'étude de cas, la consultation que nous avons menée pour le compte d'EC sur les terres humides. En effet, EC possède une masse importante de données diverses décrivant le territoire canadien sous toutes ses facettes et regroupée dans des méga-bases de données. Dans le cadre de l'un de ses projets en lien avec la protection de l'environnement, EC aimerait utiliser ces informations pour assurer le monitoring des terres humides. Mais avant d'utiliser celles-ci, elle aimerait au préalable évaluer leur qualité. Comment exploiter au mieux ces données? Quels crédits pouvons-nous accorder à cette masse d'information? Telles sont entre autres les questions préoccupantes pour EC. Avant d'aller plus loin, présentons ci-dessous des raisons justifiant cet engouement pour les terres humides.

4.1 Terres humides « 101 »

Bien que les terres humides représentent seulement environ 6,4 % de la superficie de la planète, elles jouent un rôle clé dans la survie des espèces qui nous entourent. Ces milieux humides ont un intérêt particulier pour le Canada car 25 % de celles-ci se trouvent en territoire canadien. Nous présentons ci-dessous une liste non exhaustive des rôles clés que jouent les terres humides pour l'environnement.

- Éponge géante naturelle : les terres humides stockent et restituent de l'eau permettant ainsi la régulation et l'alimentation des cours d'eau ;
- Filtre naturel : leur composition favorise la transformation biochimique des éléments organiques et minéraux;
- Thermorégulateur naturel : le stockage du carbone permet la régulation du climat;
- Réservoir naturel de vie : on y note un grand nombre d'espèces vivant car elles constituent un milieu de vie idéal pour de nombreuses espèces.

Dans la littérature géophysique, les experts s'accordent généralement pour dire qu'il existe cinq grands types de terres humides : les eaux peu profondes, les marais, les marécages, les tourbières ombrotrophes (BOG) et les tourbières minérotrophes (FEN).

Après cette brève introduction sur les terres humides, examinons à présent comment la masse importante d'informations qu'EC dispose se situe par rapport à ce triumvirat d'axes. Avons-nous en tout ou en partie à faire à des données administratives? À des données volumineuses ? Ou simplement à des données de sondage? Nous insistons sur le fait que bien se situer dans ce repère nous aidera à mieux comprendre les forces et les faiblesses de ces données, et juger de leur utilité ultérieure dans une analyse statistique.

4.2 Synthèse de l'information disponible

4.2.1 Données administratives

Nous avons déjà noté qu'une partie importante de l'information dont dispose EC a les attributs de données administratives. Ces données ont été collectées au travers de différents ministères et organismes provinciaux partenaires d'EC. Elles ont été utilisées dans ce projet pour effectuer le découpage du territoire canadien en parcelles géographiques appelées polygones de terres humides. Ces polygones servent d'unités ou d'entités d'enregistrement dans notre base de données. Il est important de souligner que ces informations n'ont initialement pas été collectées pour assurer le monitoring des terres humides. Pour la plupart, les raisons de la collecte étaient la facilitation de la gestion des différents services gouvernementaux. Le défi ici est donc énorme. On doit définir les polygones de façon cohérente et de façon à maximiser l'information disponible. Prenons l'exemple de la pente ou la déclinaison du terrain. Sans la pente, on peut penser que la géographie du Canada peut être représentée en deux dimensions. Mais la pente apporte une troisième dimension à la représentation. Si cette information n'est pas renseignée pour toutes les unités, ou qu'elle est incohérente d'une unité à l'autre, elle ne pourra pas être prise en compte dans ce découpage.

Un autre objectif visé par EC était, une fois le découpage effectué, de classer chacun des polygones résultant dans une des cinq catégories prédéfinies présentées dans la section 4.1. Une façon définitive d'y parvenir aurait été par exemple d'embaucher des experts qui auraient visité chacun des polygones individuellement et décidé dans quelle catégorie le classer. Cette solution bien qu'attrayante d'apparence est évidement irréaliste, étant très couteuse en temps et en ressources. Par exemple le découpage a donné lieu à des centaines de polygones pour la plus petite des bases, celle employée en guise de projet pilote. Pour palier à cette insuffisance, une classification approximative issue de la modélisation des données volumineuses a été adoptée. La section suivante est consacrée à cet effet.

4.2.2 Données volumineuses

Les bases de données d'EC présentaient certains attributs propres aux données volumineuses. Afin d'effectuer la classification mentionnée ci-dessus, EC a utilisé la masse volumineuse d'information provenant des satellites. En effet, les satellites balaient le territoire de manière constante et périodique, et produisent toute sorte d'informations aussi riche que variée. Cette information a été exploitée dans ce projet pour effectuer une classification automatisée de chacune des parcelles dans l'une des cinq catégories de terres humides. L'avantage de procéder ainsi est qu'il est possible d'associer chacun des polygones à une catégorie particulière et ceci en un temps record malgré leur nombre important puisqu'il s'agit essentiellement là d'un traitement informatique. Cependant, ce traitement donne lieu à une classification qui est imparfaite parce qu'elle est le résultat d'une modélisation et non d'une étude experte des lieux.

4.2.3 Données de sondage

Enfin, nous avons noté que dans la masse d'information d'EC il y avait un troisième type de données, celle-ci ayant des caractéristiques différentes des deux premières. Une partie infime de la base de données avait été classifiée de façon experte (consultant indépendant) dans chacune des cinq catégories, classification concordant avec la modélisation dans certains cas et pas dans d'autres. Pour la base utilisée dans le projet pilote, on disposait d'un verdict expert pour un échantillon représentant environ 2% de tous les polygones. EC pressentait le potentiel de cette nouvelle source d'information, puisque cette classification du consultant résultait d'une étude plus approfondie voire même la visite des lieux. À première vue, le caractère «quelconque» de l'échantillon de polygones examinés par le consultant rendait difficile son exploitation. Est-il possible de quantifier de quelque façon la discordance observée entre cette classification et celle issue du modèle pour cet échantillon ? Si oui, pouvons-nous extrapoler ces résultats à l'échelle de la base tout entière? Une étude approfondie des rapports qu'a produit le consultant a permis d'établir que nous étions en présence d'un échantillon probabiliste. Plus précisément, le consultant avait pris le soin de découper le territoire canadien en différentes strates dans lesquelles il avait sélectionné aléatoirement un échantillon

de polygones. Dans ce contexte, l'inférence statistique au niveau de la base entière devenait possible. L'une des exploitations faite de ces données se résume dans le tableau suivant.

Tableau 4.2.3-1
Données synthétisées

id	X	Y
1	BOG	∅
2	FEN	
.	.	
.	.	Bog
.	.	Marais
N	Marécages	Fen

Les données administratives sont utilisées pour construire les unités d'échantillonnage ***id***, la classification automatique provenant des données volumineuses satellitaires est utilisée comme variable auxiliaire ***X*** (renseignée pour toutes les unités de la base) et enfin la classification du consultant est utilisée comme variable d'intérêt ***Y*** (renseignée pour le sous-ensemble de polygones choisis). Nous avons entre autres produit une estimation de la proportion de chaque type de polygone que le consultant aurait trouvé si celui-ci avait eu la chance de parcourir la base de données toute entière.

5. Conclusion

En somme, dans ce travail, nous avons montré comment on peut composer avec différent type de données dans le contexte d'une consultation statistique. Nous sommes partis d'une masse très riche d'information mais non structurée rendant son exploitation initialement difficile. Avant de mener les analyses statistiques qui étaient le but premier de la consultation, nous avons structuré, analysé et synthétisé l'information disponible. Comment nous y sommes parvenus ? En se servant des trois grands axes qui sont : les données administratives, les données volumineuses et les données de sondages.

6. Remerciements

Nous remercions Arthur Goussanou, Christian Olivier Nambeu, Nathalie Hamel et Wesley Yung pour leur effort dans la révision de cet article.

Bibliographie

Brick, J. Michael. (2011), "The future of survey sampling", Public *Opinion Quarterly 75(5)*, p. 872-888.

Jabine, Thomas, and Fritz Scheuren. (1985), "Goals for Statistical Uses of Administrative Records: The Next 10 Years", Journal *of Business and Economic Statistics* 3(4), p. 380-91.

Laney, D. (2001), "3D Data Management: Controlling Data Volume, Velocity, and Variety", Technical report, META Group.