Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

# Mesures de la qualité pour le rapport mensuel sur le pétrole brut et le gaz naturel (PBGN)

Par Evona Jamroz, Lihua An et Sanping Chen[1]

## Résumé

Le rapport mensuel sur le pétrole brut et le gaz naturel (PBGN) constitue un élément essentiel du PIB mensuel du Canada. Il regroupe trois catégories de données d'entrée, soit les données déclarées lors de multiples enquêtes « d'apport », les données administratives des organismes gouvernementaux et les profils de répartition historiques fondés sur les « opinions d'experts ». Dans cet article, nous résumons nos travaux en cours et les défis qui restent à relever pour élaborer des mesures de la qualité pour les estimations dans le nouveau rapport mensuel sur le PBGN. Pour les trois sources de données, les données administratives du gouvernement, pour lesquelles nous ne présumons pas d'erreur, sont fournies en format agrégé. Pour les données d'enquête, la variance attribuable à l'échantillonnage et à l'imputation peut être estimée au moyen de méthodes conventionnelles. Il est particulièrement difficile d'estimer l'erreur associée à un paramètre fondé sur l'opinion d'experts pour laquelle nous proposons une approche bayésienne. Pour intégrer les trois types de données, nous présentons un processus fondé sur la propagation d'erreurs afin de déterminer un seul coefficient de variation (c.v.) pour les estimations finales du PBGN. Les situations dans lesquelles le c.v. n'est pas une mesure de qualité adéquate seront également examinées.

Mots-clés : Inférence bayésienne; propagation d'erreurs; échantillonnage jackknife; mesures de la qualité; combinaison des données; opinion d'experts.

## 1. Introduction

La prolifération des données accessibles en ligne et les efforts déployés par les organismes de statistique pour réduire le fardeau de réponse ont donné lieu à des techniques créatives de remplacement des données. Les enquêtes qui combinent les données provenant de sources multiples permettent aux organismes de statistique d'élargir la portée des estimations en fournissant des estimations relatives à de nouveaux concepts sans le (fastidieux) traitement supplémentaire de questionnaires. Un tel exemple à Statistique Canada est le rapport mensuel sur le pétrole brut et le gaz naturel (PBGN). Ce rapport, une composante essentielle du PIB mensuel, fournit un profil exhaustif du secteur pétrolier et gazier au Canada à un moment précis en combinant les données administratives aux données existantes issues d'enquêtes mensuelles.

Or, les occasions de publier à propos de concepts nouveaux soulèvent des questions sur la manière d'attribuer un indicateur de la qualité ou une mesure de l'incertitude qui intègre les composantes des incertitudes des enquêtes et des données administratives. L'approche que nous proposons dans le présent article retire les composantes des estimations et leurs incertitudes du paradigme d'enquête et les place dans le cadre de l'erreur de mesure expérimentale. Le principe central utilisé pour calculer l'erreur attribuable aux mesures de différentes quantités est celui de la propagation d'erreurs, où les incertitudes sont transmises à la dernière valeur calculée au moyen de règles normalisées dérivées de la simple addition de variances ou de la linéarisation de Taylor (Harris, 2016).

Dans le cas du programme du PBGN, la propagation d'erreurs est proposée comme principale solution pour attribuer un indicateur de qualité, qu'il s'agisse d'un coefficient de variation (c.v.) ou d'une étiquette catégorique. Cependant, d'autres développements sont nécessaires pour deux types d'estimations du PBGN. Dans le premier cas, les

[1]Evona Jamroz, Statistique Canada, 170, promenade Tunney's Pasture, Ottawa, Canada K1A 0T6 (evona.jamroz@canada.ca); Lihua An, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Canada K1A 0T6 (lihua.an@canada.ca); Sanping Chen, Statistique Canada, 100, promenade Tunney's Pasture, Ottawa, Canada K1A 0T6 (sanping.chen@canada.ca)

estimations dérivées de la différence entre des valeurs corrélées entraînent des c.v. exagérément élevés. Les c.v. s'avèrent également problématiques lorsqu'ils sont appliqués aux proportions. Diverses solutions de rechange, y compris les intervalles de confiance, ont été proposées (Neusy et coll., 2016). Dans le second cas, l'estimation d'une proportion est fondée en grande partie sur l'expertise sujet matière et aucune dérivation mathématique formelle de l'incertitude n'est possible. Puisque le paramètre en question est une proportion, il peut être modélisé efficacement au moyen d'une distribution bêta (Wang, 2014) dont les paramètres peuvent à leur tour être estimés au moyen d'une distribution empirique discrète *a priori*, ce qui justifie le recours à l'échantillonnage jackknife.

Le reste du présent article est organisé comme suit : dans la section 2, on décrit le rapport mensuel sur le PBGN plus en détail, et l'on fournit un exemple d'estimation du PBGN; dans la section 3, on décrit la méthode de propagation d'erreurs; dans la section 4, on discute des enjeux soulevés par l'attribution d'un indicateur de qualité pour la différence entre deux valeurs; dans la section 5, on décrit l'utilisation de l'échantillonnage jackknife et le choix d'une distribution bêta comme distribution *a priori*, accompagné de quelques résultats; enfin, dans la section 6, on présente quelques conclusions.

## 2. Rapport mensuel sur le pétrole brut et le gaz naturel

Le rapport mensuel sur le PBGN rassemble des données provenant d'une source administrative et de plusieurs enquêtes mensuelles pour créer de nouvelles estimations permettant de brosser un tableau complet du secteur pétrolier et gazier au Canada. Le rapport fournit ainsi une grande valeur ajoutée aux utilisateurs de données sans imposer un fardeau d'enquête supplémentaire aux répondants et aux opérations de traitement d'enquête à Statistique Canada. La source administrative utilisée est un organisme de réglementation provincial chargé de percevoir les redevances des sociétés pétrolières et gazières et qui, en tant que tel, on peut raisonnablement le supposer, fournit des données exhaustives et exactes. Plusieurs enquêtes sont également utilisées pour produire ce rapport : une enquête mensuelle sur les oléoducs, trois enquêtes mensuelles sur le gaz naturel et une enquête mensuelle sur les produits pétroliers raffinés. Ces enquêtes traditionnelles sont sujettes aux erreurs et aux incertitudes attribuables à l'échantillonnage et à l'imputation de la non-réponse.

Les stocks mensuels de pétrole en début de période d'une province comme l'Alberta sont un exemple d'estimation du rapport mensuel sur le PBGN. L'estimation est réalisée en faisant la somme des trois concepts suivants provenant de trois différentes sources de données :

1. le stock mensuel en début de période dans les gisements pétrolifères et les usines pétrolières, qui provient de la source administrative (organisme de réglementation);
2. le stock mensuel en début de période contenu dans les oléoducs, qui provient de l'enquête mensuelle sur les oléoducs;
3. le stock mensuel en début de période dans les raffineries de pétrole et les usines de valorisation, provenant de l'enquête sur les produits pétroliers raffinés.

## 3. Propagation d'erreurs

Lorsque des mesures sont réalisées en laboratoire pour calculer la valeur d'un nouveau paramètre qui n'est pas directement mesurable, des valeurs d'incertitude ou d'erreur sont normalement rattachées à ces mesures. Ces incertitudes doivent être propagées dans le dernier résultat calculé à l'aide de règles normalisées qui sont bien documentées. Par exemple, lorsque deux quantités sont additionnées, l'erreur résultante de la somme (ou de la différence) calculée découle de la somme des variances, en faisant l'hypothèse que les deux quantités additionnées sont indépendantes. Lorsque deux quantités sont multipliées ou divisées, la linéarisation de Taylor est utilisée afin d'obtenir une expression de l'erreur relative du produit ou du quotient. Une fois la variance (ou l'erreur type) de la somme (ou de la différence, du produit, du quotient) obtenue, un indicateur de qualité peut être attribué à l'estimation, tel qu'un c.v. ou une étiquette catégorique (A, B, C, etc.).

Les valeurs en entrée de chacune des sources de données, ainsi que la nouvelle estimation du PBGN, sont affichées dans le tableau 3-1 pour l'exemple des stocks mensuels en début de période du PBGN décrit à la section 2, pour les données de l'Alberta de la période de novembre 2016.

**Tableau 3-1**
**Estimation du Rapport mensuel sur le pétrole brut et le gaz naturel pour les stocks mensuels du pétrole brut en début de période, Alberta, novembre 2016**

| Paramètre | Estimation ($m^3$) | Variance |
|---|---|---|
| Stock en début de période, gisements pétrolifères et usines pétrolières | 2 518 613 | 0 |
| Stock en début de période, oléoducs | 6 608 803 | 205 511 287 004 |
| Stock en début de période, raffineries et usines de valorisation | 40 537 | 78 925 494 111 |
| PBGN Stock de pétrole brut en début de période en Alberta | 9 167 953 | 284 436 781 115 |
| | | c.v. = 5,82 % |

## 4. Variance attribuable à une différence

La variation nette des stocks de pétrole brut dans une province donnée est un autre paramètre calculé dans le rapport mensuel du PBGN. La variation nette des stocks est la différence entre les stocks au début et à la fin de la période de référence du même mois. Il y a clairement une corrélation entre les valeurs des stocks en début et en fin de période, ce qui contredit l'hypothèse de l'indépendance des deux variables en entrée. Il faut alors soustraire un terme de covariance lors du calcul de la variance de la variation nette des stocks. Dans certains cas, si les variables sont de même ordre de grandeur, comme c'est souvent le cas pour les stocks de début et de fin de période, la différence est infinitésimale par rapport à l'erreur type et la valeur du c.v. devient alors très grande. Le c.v. peut ainsi s'avérer une mesure qui est loin d'être idéale pour attribuer un indicateur de qualité. Le tableau 4-1 fournit un exemple, pour les données de la Saskatchewan de novembre 2016, où l'utilisation du c.v. comme fondement de l'indicateur de qualité entraînerait la suppression de la valeur lors de la publication en raison de la taille considérable de l'erreur.

**Tableau 4-1**
**Calcul du c.v. pour la variation nette des stocks, Saskatchewan, novembre 2016**

| Concept du PBGN | Valeur | Variance | Erreur type | c.v. |
|---|---|---|---|---|
| Variation nette des stocks, Saskatchewan | 4705 m3 | 54149180 | 7359 | 156 % |

Il y a d'autres situations où les c.v. se sont avérés problématiques, comme dans le cas des valeurs qui représentent des proportions. Dans ce dernier cas, d'autres auteurs ont proposé des intervalles de confiance comme mesure alternative sur laquelle peut être fondé un indicateur de qualité catégorique.

## 5. Variance attribuable à l'opinion d'experts

Dans certains cas, un paramètre peut être estimé entièrement, ou en grande partie, par un expert en la matière qui se fie sur ses connaissances du domaine et des événements récents qui influencent le paramètre en question. En ce qui concerne le rapport mensuel sur le PBGN, la proportion *p* du pétrole qui est dévié vers un oléoduc particulier doit être estimée par l'expert en la matière. Ce paramètre n'est pas obtenu d'une enquête ni disponible dans les données administratives. L'analyste formule plutôt une « opinion éclairée » en fonction des événements qui influencent le transport du pétrole, comme les embargos et les accidents d'oléoducs. Ces événements tendent à être non stationnaires,

c'est-à-dire qu'ils n'affichent aucun comportement cyclique et, par conséquent, qu'ils n'ont pas tendance à être bien représentés par les séries chronologiques.

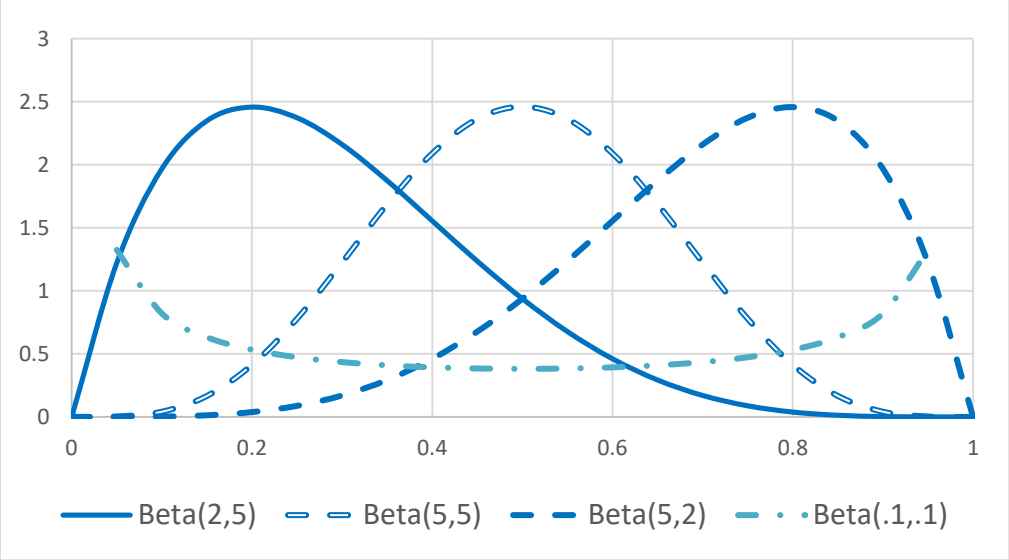
Attribuer une erreur de manière rigoureuse à ce type d'estimation subjective représente un défi de taille. Il est possible que l'expert en la matière attribue une incertitude à cette estimation subjective, mais elle peut être influencée par un biais personnel. Nous utilisons le rééchantillonnage jackknife en faisant l'hypothèse que la proportion *p* obéit à une distribution bêta afin d'obtenir à une variance dérivée empiriquement.

On se souviendra que dans le cadre bayésien (Jackman, 2009), une distribution *a priori* décrit la manière dont est distribué un paramètre de la distribution d'une variable aléatoire. Dans le cas de la proportion de pétrole dévié vers un pipeline particulier du PBGN, la distribution bêta de paramètres (α, β) représente un choix naturel pour décrire la distribution de *p* (Wang et coll., 2014). La distribution bêta, tout comme *p*, est définie dans l'intervalle [0,1] et, par le choix des paramètres (α, β), elle peut représenter avec souplesse toute une gamme de distributions, y compris des distributions unimodales symétriques ou asymétriques et des distributions en forme de U (voir la figure 5-1). La distribution bêta est définie de la manière suivante :

$$f(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{1}$$

où B(α, β) est la fonction bêta.

**Figure 5-1**
**Distributions bêta pour divers paramètres de forme**




Vu la subjectivité de l'opinion des experts, nous adoptons ensuite l'hypothèse que les paramètres ($\alpha$, $\beta$) de la distribution bêta ci-dessus suivent une distribution *a priori* inconnue $\pi(\alpha, \beta)$. Les données historiques $\{p_1, p_2, \ldots, p_k\}$ sont utilisées pour estimer la distribution *a priori* inconnue de façon empirique, par rééchantillonnage. Dans cet article, seuls les résultats de la méthode jackknife sont présentés.

Pour chaque $i = 1, 2, \ldots, k$, la valeur $p_i$ est retirée de l'ensemble des données historiques et une distribution bêta est ajustée, laquelle permet d'obtenir une estimation des paires de paramètres ($\alpha_i$, $\beta_i$). Comme ces échantillons sont équiprobables, nous obtenons une approximation discrète de la distribution inconnue $\pi(\alpha, \beta)$ :

$$\pi(\alpha = \alpha_i, \beta = \beta_i) \approx \frac{1}{k} \tag{2}$$

Ensuite, en appliquant le théorème de Bayes, la distribution *a posteriori* de (α, β) tenant compte de la plus récente opinion des experts, $p_0$, peut être estimée par la distribution discrète suivante :

$$P(\alpha = \alpha_i, \beta = \beta_i | \, p = p_0) = \frac{\frac{p_0^{\alpha_i - 1}(1-p_0)^{\beta_i - 1}}{B(\alpha_i, \beta_i)}}{\sum_{j=1}^{k} \frac{p_0^{\alpha_j - 1}(1-p_0)^{\beta_j - 1}}{B(\alpha_j, \beta_j)}} \tag{3}$$

Nous pouvons alors calculer la variance *a posteriori* de l'espérance de la valeur de l'opinion d'experts, $\frac{\alpha}{\alpha+\beta}$, comme suit :

$$Var\left(\frac{\alpha}{\alpha+\beta}\right) = E\left(\left(\frac{\alpha}{\alpha+\beta}\right)^2\right) - \left[E\left(\frac{\alpha}{\alpha+\beta}\right)\right]^2 \quad (4)$$

Le calcul du rééchantillonnage jackknife et de la variance *a posteriori*, effectué pour quatre différents tracés d'oléoduc, est présenté dans le tableau 5-1.

**Tableau 5-1**
**Calcul de la variance *a posteriori* pour différents tracés d'oléoduc**

| Source_destination | $p_0$ | Variance | Erreur type |
|---|---|---|---|
| Alb._Ont. | 0,2643 | 3,038E-06 | 0,001743 |
| Alb._Sask. | 0,1107 | 8,111E-07 | 0,000901 |
| Alb._Alb. | 0,5541 | 5,164E-06 | 0,002273 |
| Alb._C.-B. | 0,0708 | 2,875E-07 | 0,000536 |

## 6. Conclusion

Lorsque les données d'enquête ou les données administratives sont combinées pour produire de nouvelles estimations, les incertitudes relatives aux données d'entrée doivent être transmises à l'estimation finale au moyen de techniques de propagation d'erreurs. Si la nouvelle estimation est une différence ou une proportion, il se peut qu'un indicateur de qualité basé sur un intervalle de confiance soit préférable à un c.v. Quant aux estimations fondées sur une opinion d'experts, les méthodes bayésiennes, de concert avec le rééchantillonnage jackknife, peuvent être utilisées pour calculer une variance *a posteriori* des paramètres de la distribution, ce qui permet de tenir compte de l'incertitude dans les données dérivées de l'opinion d'experts.

## Bibliographie


Harris, D. C. (2016), *Quantitative Chemical Analysis*, New York: W.H. Freeman and Company.

Jackman, S. (2009), *Bayesian Analysis for the Social Sciences*, Chichester: John Wiley and Sons.

Neusy, E., et H. Mantel (2016), « Confidence Interval for Proportions Estimated from Complex Survey Data », *Proceedings of the Survey Methods Section, SSC Annual Meeting*.

Wang, X.-F., et Y. Li (2014), « Bayesian Inferences for Beta Semiparametric-Mixed Models to Analyze Longitudinal Neuroimaging Data », *Biometrical Journal*, 56, p. 662-677.


## Remerciements


Les auteurs tiennent à remercier les personnes suivantes : Jack Gambino, pour ses suggestions au sujet de la distribution bêta et du cadre bayésien, James Lakatosh, pour son aide dans la création des échantillons jackknife, ainsi qu'Ahcene Ait Yahia et Michael Pupaza, qui ont fourni de l'information contextuelle sur le programme de PBGN.