Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Quality Measures for the Monthly Crude Oil and Natural Gas (MCONG) Report

Evona Jamroz, Lihua An, and Sanping Chen[1]

Abstract

The Monthly Crude Oil and Natural Gas (MCONG) report is a critical component of Canada's monthly GDP. It brings together three categories of input data: data reported by multiple "feeder" surveys, administration data from government agencies, and historical allocation profiles based on "expert opinions." In this paper, we summarize our ongoing work and remaining challenges for developing quality measures for the estimates in the new MCONG report. For the three data sources, the government administrative data are provided in macro format, for which we assume no error. For the survey data, the variance due to sampling and/or imputation can be estimated using conventional methods. A particular challenge is to estimate the error associated with a parameter that is based on expert opinion for which we propose a Bayesian approach. To integrate the three types of data, a process based on error propagation is presented to determine a single coefficient of variation (CV) for the final MCONG estimates. The situations in which CV is not an adequate quality measure will also be discussed.

Key Words: Bayesian inference; Error propagation; Jackknife sampling; Quality measures; Combining data; Expert opinion.

1. Introduction

Both the proliferation of online accessible data and efforts by statistical agencies to reduce response burden have spurred creative data replacement techniques. Surveys that combine data from multiple sources allow statistical agencies to widen the scope of the estimation landscape by providing estimates for new concepts without additional (and burdensome) questionnaire processing. One such example at Statistics Canada is the Monthly Crude Oil and Natural Gas (MCONG) report. This report, which is a critical component of monthly GDP, provides a comprehensive snapshot of the oil and gas sector in Canada by combining administrative data with already existing data from "feeder" surveys.

Opportunities for the publication of new concepts however raise questions about how to assign a quality indicator or measure of uncertainty that incorporates the constituent uncertainties of the feeder surveys and administrative data. The proposed approach presented in this paper removes the constituent estimates and their uncertainties from the survey paradigm and places them in the framework of experimental measurement error. The central principle for calculating the error due to measurements of different quantities is that of error propagation, where the uncertainties are transmitted through to the final calculated value using standard rules that are derived from straightforward addition of variances or Taylor linearization (Harris, 2016).

In the case of the MCONG programme, error propagation is proposed as the principle solution for assigning a quality indicator, whether it is a coefficient of variation or categoric label. For two types of estimates in MCONG however, further development is required. In the first case, estimates that are derived from the difference of correlated values result in unacceptably high CVs; CVs have also proved problematic when applied to proportions and various alternatives, including confidence intervals, have been proposed (Neusy et al, 2016). In the second case, an estimate of proportion was based largely on subject matter expertise and no formulaic derivation of the uncertainty is possible. Because the parameter in question is a proportion, it can be effectively modeled using a beta distribution (Wang, 2014) whose parameters in turn can be approximated by an empirical discrete prior distribution justified by the use of jackknife sampling.

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

The remainder of this paper is organized as follows: Section 2 describes the MCONG report in more detail, along with an example of an MCONG estimate; section 3 describes the error propagation method; section 4 discusses issues that arise with assigning a quality indicator for a difference of two values; section 5 describes the use of jackknife sampling assuming a beta prior distribution along with results; section 6 presents concluding remarks.

2. The Monthly Crude Oil and Natural Gas Report

The MCONG report gathers data from an administrative source and other "feeder" surveys to create new estimates which provide a more comprehensive picture of the oil and gas sector in Canada. The report thus provides much added value to data users without applying further survey burden to either respondents or to survey processing operations at Statistics Canada. The administrative source that is used is a provincial regulatory body which is involved in collecting royalties from oil and gas companies, and as such, can safely be assumed to provide comprehensive and accurate data. There are also multiple feeder surveys to the report: a monthly pipeline survey, 3 monthly natural gas surveys, and a monthly refined petroleum products survey. These traditional surveys are susceptible to errors/uncertainties due to sampling and non-response (imputation).

An example of an MCONG estimate is the monthly opening inventory of oil in a particular province such as Alberta. The MCONG estimate is created by summing the following three concepts from 3 different data sources:

1. Opening monthly inventory at oil fields and plants, which originates from the admin source (regulatory body)
2. Opening monthly inventory contained in pipelines, originating from the monthly oil pipeline survey
3. Opening monthly inventory at refineries and upgraders, originating from the refined petroleum products survey

3. Error Propagation

When measurements are made in a laboratory setting in order to compute the value of a new (and not directly measurable) parameter, uncertainties or error values are typically attached to the measurement. These uncertainties must be propagated to the final computed result, and are done so using standard well documented rules. For instance, if two quantities are to be added together, the resultant error on the computed sum (or difference) derives from the sum of the variances, assuming the two quantities being added are independent. For two quantities multiplied or divided together, Taylor linearization is used to obtain an expression for relative error of the product. Once the variance (or standard error) of the sum/difference/product/quotient is obtained, a quality indicator can be assigned such as a CV or categoric label (A, B, C, etc).

Using the MCONG monthly opening inventory example described in section 2 with data from Alberta for reference period November 2016, the input values from each of the data sources, along with the new MCONG estimate are displayed in Table 3-1.

Table 3-1
Monthly Crude Oil and Natural Gas estimate for Monthly Crude Oil Opening Inventory, Alberta, November 2016

Parameter	Estimate (m^3)	Variance
Opening Inventory, Fields and Plants	2518613	0
Opening Inventory, Pipelines	6608803	205,511,287,004
Opening Inventory, Refineries and Upgraders	40537	78,925,494,111
MCONG Opening Inventory Crude Oil in Alberta	9167953	284,436,781,115
		CV=5.82%

4. Variance Due to A Difference

Another parameter that is calculated in the MCONG report is the net inventory change of crude oil in a given province. The net inventory change is the difference between the opening and closing inventory in the same month. Since the opening and closing inventory values are clearly correlated, the assumption of independence for the two feeder variables is violated and a covariance term must be subtracted when calculating the variance of the net inventory change. In some cases if the variables are similarly valued, as may often be the case for opening and closing inventories, the difference is vanishingly small compared to the standard error, and the CV value becomes very large. The coefficient of variation can thus be a less than ideal metric to use for assigning a quality indicator. Table 4-1 provides an example, using Saskatchewan data for November 2016, where using the CV as the basis for a quality indicator would result in suppression for publication because of the large error.

Table 4-1
Calculation of CV for Net Inventory Change, Saskatchewan, November 2016

MCONG Concept	Value	Variance	Std Error	CV
Net Inventory Change, Saskatchewan	4705 m3	54149180	7359	156%

Other instances where CVs have proved problematic include values that represent proportions. In the latter case, confidence intervals have been proposed by other authors as an alternative metric on which a categoric quality indicator can be based.

5. Variance Due to Expert Opinion

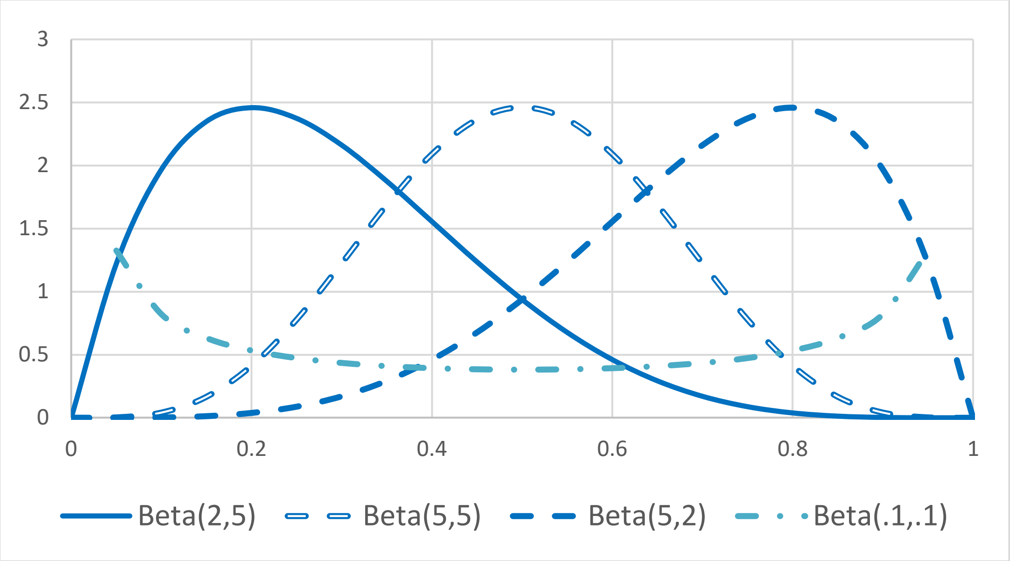
In some cases a parameter may be largely (or completely) estimated by a subject matter expert who relies on his or her knowledge of the field and recent influential events that affect the parameter. For the MCONG report, a proportion, p, of oil that is diverted to a particular pipeline requires estimation by the subject matter analyst. The parameter is not reported by feeder surveys, nor is it available through administrative data. Rather, the analyst makes an "educated guess" based on events that influence transport of oil such as embargos or pipeline accidents. These events tend to be non-stationary, meaning they do not exhibit cyclic behaviour, and thus tend not to be well represented as time series. To assign an error to this type of subjective estimate in a rigorous manner poses a considerable challenge. The subject matter expert may assign uncertainty to his subjective estimate, but this can be influenced by personal bias. We use jackknife resampling assuming a beta distribution for the proportion, to arrive at an empirically derived variance.

Recall that in the Bayesian framework (Jackman, 2009), a prior distribution describes how a parameter of a random variable's distribution is itself distributed. In the case of the MCONG's proportion of oil diverted to a particular pipeline, the beta distribution with parameters (α, β) is a natural choice for describing the distribution of p (Wang et al, 2014). The beta distribution, like p, is defined on the interval [0,1] and through the choice of (α,β), can flexibly represent a range of distributions, including unimodal symmetric, unimodal right or left skewed, or U-shaped (see figure 5-1). The beta distribution is defined by:

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

where B(α,β) is the beta function.

Figure 5-1
Beta distributions for a variety of shape parameters



Next we assume that, largely due to the subjectivity of the expert opinion, the parameters (α,β) of the above beta distribution follow an unknown prior distribution $\pi(\alpha,\beta)$. We then use the historical data $\{p_1, p_2, \cdots, p_k\}$ to approximate this unknown prior distribution empirically through resampling. In this paper, we only present the results from the simple jackknife resampling method.

For i =1,2, …, k, after dropping the value p_i from the historical data, a beta distribution is fit, providing a pair of parameter estimates (α_i,β_i). Because each resample is considered equally likely, we obtain a discrete approximation of the unknown distribution $\pi(\alpha,\beta)$:

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

Then, by applying the Bayesian theorem, the posterior distribution of (α,β) given the latest expert opinion p_0 can be approximated by the following discrete distribution:

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

We can then calculate the posterior variance of $\frac{\alpha}{\alpha+\beta}$, the expected value of the expert opinion as follows:

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

The jackknife resampling and posterior variance calculation was performed for 4 different pipeline routes and is shown in table 5-1.

Table 5-1
Posterior variance calculation for different pipeline routes

Source_destination	p_0	Variance	Std. Error
AB_ON	0.2643	3.038E-06	0.001743
AB_SK	0.1107	8.111E-07	0.000901
AB_AB	0.5541	5.164E-06	0.002273
AB_BC	0.0708	2.875E-07	0.000536

6. Conclusion

When survey or administrative data is combined to produce new estimates, the uncertainties on the input data must be carried to the final estimate using error propagation techniques. If the new estimate is a difference or is a proportion, a quality indicator based on a confidence interval may be preferable to a CV. For estimates based on subject matter opinion, Bayesian methods together with jackknife resampling may be employed to calculate a posterior variance of the distribution's parameters, which reflects the uncertainty in the expert opinion-derived data.

References

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., and H. Mantel (2016), "Confidence Interval for Proportions Estimated from Complex Survey Data", *Proceedings of the Survey Methods Section, SSC Annual Meeting*.

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

Acknowledgements

The authors would like to acknowledge the assistance of Jack Gambino for his suggestions regarding the beta distribution and Bayesian framework, James Lakatosh for his assistance creating the jackknife samples, and Ahcene Ait Yahia and Michael Pupaza for providing background information on the MCONG program.