C.8

Methodology Branch
Time Series Research and Analysis Division

Direction de la méthodologie
Division de la recherche et de l'analyse des chroniques

MINIMIZATION OF THE MEAN SQUARE ERROR OF THE PRELIMINARY ESTIMATES OF THE CANADIAN BENEFICIARIES SERIES

by

Guy Huot

HA-514
no. 93-04
c.2

Statistics Canada Statistique Canada

Canada

*53715

WORKING PAPER NO. TSRA-93-004E

TIME SERIES RESEARCH & ANALYSIS DIVISION
METHODOLOGY

MINIMIZATION OF THE MEAN SQUARE ERROR OF THE PRELIMINARY ESTIMATES OF THE CANADIAN BENEFICIARIES SERIES

by

Guy Huot

Statistics Canada

STATISTICS CANADA STATISTIQUE CANADA
JUIN JUNE 22 1991
LIBRARY BIBLIOTHÈQUE

MINIMIZATION OF THE MEAN SQUARE ERROR OF THE PRELIMINARY ESTIMATES OF THE CANADIAN BENEFICIARIES SERIES

GUY HUOT

Time Series Research and Analysis Division, Statistics Canada

January 1993

Abstract: There are sometimes systematic biases in preliminary observations. The statistical agencies producing the figures make adjustments to obtain unbiased preliminary observations. These figures are first published and then revised one or two months later. The revised figures are the true values. This paper uses the structure of the data revisions to obtain unbiased preliminary observations. The results encompass the in-sample analysis that support the monthly release by Statistics Canada of the unbiased preliminary observations of ten Canadian unemployment insurance beneficiaries series, and a summary of the performance of the out-of-sample released values from January 88 to December 91. The unbiased preliminary observations have reduced level and direction errors due to bias, in 95.4% of the 480 cases studied in the out-of-sample analysis.

Keywords: Data revisions, Modelling the structure of the data revisions, SARIMA models, Combining forecasts.

1. Introduction

Statistical agencies publish preliminary estimates (of economic time series based upon incomplete information) which are then subject to revisions. The use of preliminary data can result in substantial level errors and, in turn, estimate-of-change errors. The problem arises when estimating the month-to-month or quarter-to-quarter change between the preliminary estimate for the current period and the "final" (revised) estimate of the previous period.

The existence of measurement errors has been recognized in early studies by Zellner (1958) and Morgenstern (1963). More recently, Howrey (1978), Harvey *et al.* (1983), and Rao, Srinath and Quenneville (1989), among several others, have proposed methods to produce "optimal" preliminary estimates of variables for which a preliminary estimate is already available. Optimality is defined in the sense of minimizing the mean square error of the preliminary estimates. The approach used by those authors consists of modelling and forecasting the final estimates as well as the measurement errors. The optimal preliminary estimates result from a combination of the unrevised preliminary estimates, and the forecasts of final estimates and of the measurement errors. The improvement of the preliminary figures is viewed as both a forecasting and a combining forecast problem.

Huot and Plourde (1987) generalized the method proposed by Howrey (1978) and Harvey *et al.* (1983) and applied it to ten Canadian beneficiaries series. The Huot and Plourde (HP) method was implemented by Statistics Canada in 1988. A close monthly monitoring of the performance of both the optimal preliminary estimates and the unrevised preliminary estimates has systematically shown a greater accuracy in the former.

This paper presents the HP generalized method, the results of the in-sample analysis that supported the release in 1988 of the optimal preliminary figures, and a summary of their performance over a four year period.

Section 2 discusses the characteristics of the preliminary data, the final data and the measurement error (also referred to as missing information). Section 3 describes the method used to calculate the optimal preliminary estimates. Section 4 presents the numerical analysis done before and after the implementation of the optimal preliminary estimates. Section 5 gives the concluding remarks.

2. The Canadian beneficiaries series

The ten Canadian beneficiaries series discussed in this study represent a count of all the persons who qualified for unemployment insurance benefits from January 1975 to April 1992. The preliminary estimates are a systematic undercount of the number of beneficiaries. The final revised estimates are available 2 months after the release of the preliminary estimates. Exhibit 1 presents the mean percentage error (MPE) of the preliminary data from January 1988 to December 1991. That is

$$MPE = \frac{1}{48} \sum_{t=1}^{48} \left(\frac{Y^{P}_{t+2} - Y_{t+2}}{Y_{t+2}} \right) x\ 100 \qquad (1)$$

where Y^{P}_{t+2} is the unrevised preliminary estimate for the current period (t+2) and Y_{t+2} is the final estimate obtained after revision. The figures in exhibit 1, which show the undercount, are analysed later in section 4.2.

Exhibit 1
Mean percentage error associated with the preliminary data

Beneficiaries Series	MPE
Canada	- 1.24%
Nova Scotia	- 0.31%
Quebec	- 0.93%
Ontario	- 2.13%
Manitoba	- 0.71%
Saskatchewan	- 1.10%
Alberta	- 2.51%
British Columbia	- 1.48%
Yukon	- 5.18%
North West Territories	-12.43%

The systematic undercount is obtained by subtracting the final from the preliminary data and is here seen as missing information. We assume that the missing information is stochastic and will model it accordingly. Exhibits 2a and 2b show the missing information for the Canada and Yukon series, the other series display similar characteristics except for two series with no seasonal pattern. It is apparent from the graphs that the missing information is biased downward with an increasing variance during the period 1990 - 1992 which has been characterized by a structural economic recession.

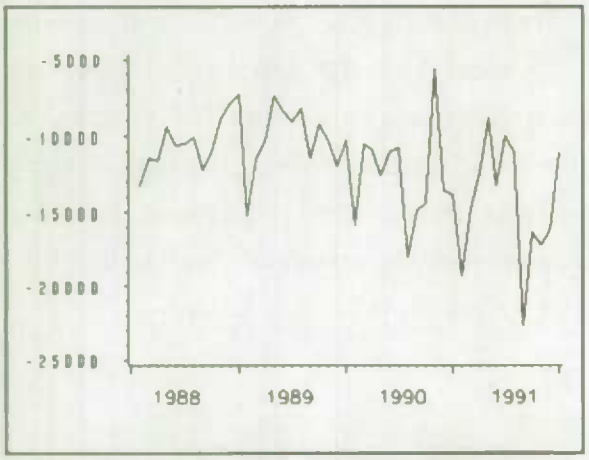

Exhibit 2a: Missing Information - Canada

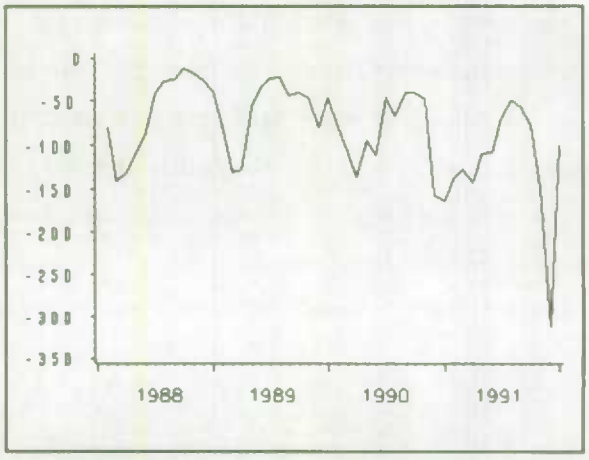

Exhibit 2b: Missing Information - Yukon

3. The Analytical Framework

Let Y^P_{t+1} be the preliminary estimate for the current period $(t+1)$, Y_t the most recent final figure available, $\hat{Y}_t(1)$ the forecast of Y_{t+1} given $Y_1, \ldots, Y_t$, $\hat{Y}^*_t(1)$ the one-month-ahead forecast of the missing information defined as $Y^*_t = Y^P_t - Y_t$, and $\hat{Y}^O_t(1)$ the "optimal" preliminary estimate. The approach proposed by Howrey (1978) and Harvey *et al.* (1983) and applied by Rao, Srinath and Quenneville (1989) is based on a linear combination of both the one-month-ahead forecast of the final series, and the missing information added to the current preliminary estimate

$$\hat{Y}^O_t(1) = \lambda \hat{Y}_t(1) + (1-\lambda)\,(Y^P_{t+1} + \hat{Y}^*_t(1)) \qquad (2)$$

Autoregressive models are fitted to Y_t for any $p \geq 1$ and Y^*_t for $p = 1$. The use of autoregressive (AR) terms was suggested by Harvey *et al.* (1983) given certain technical advantages, at the time, in setting up a state space model. $\lambda = \sigma^2_u/(\sigma^2_u + \sigma^2_*)$ where σ^2_* and σ^2_u are the variances of the residuals of the models fitted to the final series Y_t and to the missing information Y^*_t, respectively. λ is in the interval $[0,1]$.

The AR(1) model fitted to Y^*_t assumed that the mean is equal to zero thus implying that the preliminary estimates are unbiased. It was expected that statistical agencies would make the necessary adjustments to the preliminary estimates if systematic biases appeared. Consequently, optimality reduces to minimum variance, that is, a gain in efficiency.

Huot and Plourde (1987) extended the method discussed above by explicitly taking into account the bias present in the missing information. These authors assume, firstly, that additional information may be available in the past behavior of Y_t^* and, secondly, that Y_t^* will not necessarily move just like the final or the preliminary series in the sense that the missing information Y_t^* is characterized by its own trend-cycle, seasonal and irregular variations. In order to model Y_t^* a broader class of seasonal autoregressive integrated moving average (SARIMA) models (Box and Jenkins, 1970) was used. Furthermore, no constraints were imposed on the order of the parameters used to model Y_t^*. The two-month-ahead "optimal" preliminary estimate is then given by:

$$\hat{Y}_t^O(2) = \lambda \hat{Y}_t(2) + (1-\lambda)\,(Y_{t+2}^P + \hat{Y}_t^*(2)) \qquad (3)$$

where $\hat{Y}_t(2)$ is the forecast of Y_{t+2} given $Y_1, \ldots, Y_t$ obtained from the SARIMA model

$$\phi_p(B)\,\Phi_P(B^s)\,\nabla^d\nabla_s^D Y_t = \theta_q(B)\,\Theta_Q(B^s)\,a_t \qquad (4)$$

and $a_t \sim N(0,\sigma_a^2)$. Equation (4) represents a general multiplicative SARIMA model of seasonal periodicity s, and of order $(p,d,q)(P,D,Q)_s$. Here, B is the backshift operator $B^iY_t = Y_{t-i}$, $\nabla^d = 1-B^d$, $\phi_p(B) = 1-\phi_1B-\ldots-\phi_pB^p$ and $\Phi_P(B^s) = 1-\Phi_1B^s-\ldots-\Phi_P\,B^{sP}$ are the ordinary and seasonal autoregressive polynomials, $\theta_q(B) = 1-\theta_1B-\ldots-\theta_qB^q$ and $\Theta_Q(B^s) = 1-\Theta_1B^s-\ldots-\Theta_Q\,B^{sQ}$ are the ordinary and seasonal moving average polynomials. Similarly, $\hat{Y}_t^*(2)$ is the forecast of the missing information Y_{t+2}^* given $Y_1^*,\ldots,Y_t^*$ obtained from a general multiplicative SARIMA model

$$\alpha_p(B)\,A_P(B^s)\,\nabla^d\nabla_s^D Y_t^* = \omega_q(B)\,\Omega_Q(B^s)\,u_t \qquad (5)$$

and $u_t \sim N(0,\sigma_u^2)$.

Exhibit 3 shows the models fitted to the Canada and Yukon series (from January 1976 to December 1986), their parameter values estimated with the maximum-likelihood method, the results of the portemanteau test (Ljung and Box, 1978) and the residual variance. The Q statistics accept the null hypothesis of randomness of the residuals in each case. The non-zero parameters are identified using the Akaike's information criterion (AIC). The variance of the residuals of the corresponding optimal models is used to calculate λ in equation (3).

The idea of combining forecasts assumes that each forecast captures different aspects of the information available for prediction (Clemen, 1989). Thus the combination of forecasts achieves an increase in accuracy. In equation (3) both the forecast of the missing information added to the preliminary estimate and the forecast of the final series are aggregated to produce a single forecast.

Exhibit 3

SARIMA models

Series	SARIMA Models	Q(24)	$\hat{\sigma}^2$
Canada Final	$(1 - 0.68B - 0.37B^2)\nabla\nabla^{12} Y_t =$ $(1 + 0.22B^4 - 0.58B^{12})a_t$	25.35	324018749
Canada Missing Information	$\nabla Y_t^* =$ $(1 - 0.35B)$ $(1 - 0.45B^2 - 0.29B^5 + 0.32B^{12} + 0.33B^{24})u_t$	21.83	3575997
Yukon Final	$(1 - 0.47B + 0.27B^{12})\nabla\nabla^{12} Y_t =$ $(1 + 0.21B^3)a_t$	24.94	4635
Yukon Missing Information	$(1 - 0.70B)\nabla^{12} Y_t^* =$ $(1 + 0.25B - 0.16B^6)(1 - 0.77B^{12})u_t$	18.60	515

$\lambda = \sigma_u^2/(\sigma_u^2 + \sigma_a^2)$ where σ_a^2 and σ_u^2 are the variances of the residuals of the SARIMA models above. λ could have been defined as $\sigma_u^2(2)/(\sigma_u^2(2) + \sigma_a^2(2))$ where $\sigma^2(2)$ is the variance of the two-step-ahead forecast error. The value of λ (between 0 and 1) determines the proportion in which the two sources of information on the right hand side of equation (3) enter the aggregation. The extreme case $\lambda = 1$ implies that the preliminary estimate Y_{t+2}^p is not useful at all. The smaller λ is, the more reliable Y_{t+2}^p is.

4. Empirical results

This section briefly presents the results of the in-sample analysis that supported the publication of the $\hat{Y}_t^o(2)$ estimates obtained from equation (3), and a summary of the performance of the out-of-sample $\hat{Y}_t^o(2)$ values over the last four years.

4.1 In-sample analysis

The purpose of the analysis was to find a predictor of the final figure Y_{t+2} that would perform better than the actual preliminary estimate Y_{t+2}^p which is biased. The two other predictors available

on the right hand side of equation (3) are the forecast $\hat{Y}_t(2)$ obtained from the final series Y_t and the actual preliminary estimate Y^p_{t+2} adjusted for the missing information, that is, $Y^p_{t+2} + \hat{Y}^*_t(2)$. $\hat{Y}^o_t(2)$ on the left hand side of the equation results from the linear combination of these predictors. A comparison of the predictors on both sides of the equation has been conducted over an 18-month period from July 85 to December 86.

Exhibit 4 displays the performance of the four predictors in terms of the Root Mean Square Error (RMSE), for each of the ten beneficiaries series. The RMSE is used to show the relative contribution of the error of the component series to the total error of the Canada series in columns 2 and 3. The RMSE for $\hat{Y}^o_t(2)$ is,

$$RMSE = \left[\frac{1}{18}\sum_{t=1}^{18}\left(\hat{Y}^O_t(2) - Y_{t+2}\right)^2\right]^{1/2} \qquad (6)$$

and similarly for $\hat{Y}_t(2)$, Y^p_{t+2} and $Y^p_{t+2} + \hat{Y}^*_t(2)$.

Exhibit 4

Root Mean Square Error for the four predictors

Beneficiaries Series	RMSE for $\hat{Y}_t(2)$	RMSE for Y^p_{t+2}	RMSE for $Y^p_{t+2}+\hat{Y}^*_t(2)$	RMSE for $\hat{Y}^o_t(2)$
Canada	15184	12104	2475	2482
Nova Scotia	1085	148	71	70
Quebec	5876	3017	796	847
Ontario	7573	5275	946	963
Manitoba	1108	351	84	83
Saskatchewan	1090	373	115	115
Alberta	6594	1750	420	489
British Columbia	4849	1181	416	398
Yukon	129	79	29	27
North West Territories	78	183	24	29

The figures in the third and fourth columns show that the most accurate of the four estimators for Y_{t+2} are $\hat{Y}^o_t(2)$ and $Y^p_{t+2} + \hat{Y}^*_t(2)$, the third best being Y^p_{t+2}. Finally, $\hat{Y}_t(2)$ ranks last after Y^p_{t+2} which is biased, that is, the two-month-ahead forecast errors in column 1 are bigger than the bias effect in column 2 except for the North West Territories series. Equation (3) took into account the poor performance of $\hat{Y}_t(2)$ by setting λ to almost zero except for the Alberta, Yukon and North West Territories series as shown in exhibit 5. That is, equation (3) practically reduced to

$$\hat{Y}_t^O = Y_{t+2}^P + \hat{Y}_t^*(2) \qquad (7)$$

However, the British Columbia series indicates that $Y_{t+2}^P + \hat{Y}_t^*(2)$ may not systematically be the best estimator. Further research is required before using equation (7).

Exhibit 5 displays the average λ values for the ten beneficiaries series. Since $Y_{t+2}^P + \hat{Y}_t^*(2)$ is the most accurate of the predictors on the right hand side of equation (3), it has received almost all of the weight. Furthermore, the larger values of λ tend to be associated with the larger bias errors measured in terms of the Mean Percentage Error. The standard deviations in column 2 show that the λ values have smoothly evolved over the 18-month period. This characteristic is important. A lack of weight stability does not garantee that the combined forecasts will have the smallest variance or the best accuracy (Kang, 1986).

Exhibit 5

Average λ values and standard deviations

Beneficiaries series	λ	(Std. Dev.)	$(1 - \lambda)$
Canada	.015	(.0004)	.985
Nova Scotia	.003	(.0001)	.997
Quebec	.016	(.0002)	.984
Ontario	.010	(.0003)	.990
Manitoba	.007	(.0001)	.993
Saskatchewan	.010	(.0003)	.990
Alberta	.046	(.0075)	.954
British Columbia	.014	(.0002)	.986
Yukon	.099	(.0027)	.901
North West Territories	.190	(.0042)	.810

Statistical agencies generally publish both the original and the seasonally adjusted data. A test has also been conducted to assess the benefits obtained from the seasonal adjustement of $\hat{Y}_t^O(2)$ obtained from equation (3) instead of Y_{t+2}^P using the X-11-ARIMA program (Dagum, 1988). The RMSEs for the seasonally adjusted Y_{t+2}^P and $\hat{Y}_t^O(2)$ figures calculated with respect to the corresponding seasonally adjusted final estimates Y_{t+2} are shown in exhibit 6. The superiority of the seasonally adjusted figures $\hat{Y}_t^O(2)$ is confirmed.

Exhibit 6

Root Mean Square Error for the seasonally adjusted series

Beneficiaries series	RMSE for Y^{P}_{t+2}	RMSE for $\hat{Y}^{O}_{t}(2)$
Canada	13067	3095
Nova Scotia	167	65
Quebec	3313	893
Ontario	5699	1088
Manitoba	338	85
Saskatchewan	413	93
Alberta	1707	465
British Columbia	1381	267
Yukon	84	40
North West Territories	195	38

4.2 Out-of-sample analysis

The out-of-sample analysis covers a period of 48 consecutive months ending in December 1991. A record of the Y^{P}_{t+2}, $\hat{Y}^{O}_{t}(2)$ and Y_{t+2} figures has been kept for that period in order to get a global assessment of the performance of $\hat{Y}^{O}_{t}(2)$. Moreover $\hat{Y}^{O}_{t}(2)$ has been closely monitored each month before its release. The performance is measured using three error measures: the mean percentage error (MPE), the standard deviation of the percentage errors, and the root mean square percentage error (RMSPE). The RMSPE for $\hat{Y}^{O}_{t}(2)$ is,

$$RMSPE = \left[\frac{1}{48} \sum_{t=1}^{48} \left(\frac{\hat{Y}^{O}_{t}(2) - Y_{t+2}}{Y_{t+2}} \right)^{2} \right]^{1/2} x\ 100 \qquad (8)$$

and similarly for $\hat{Y}^{P}_{t+2}$.

The comparison of the MPEs in columns 1 and 3 shows that $\hat{Y}^{O}_{t}(2)$ is practically an unbiased estimator. The RMSPEs in column 4, which reduces to the standard deviations displayed in column 3, support this conclusion. On the other hand, the standard deviations in columns 1 and 3 are similar. The reason is that λ is almost zero except for the Alberta, Yukon and North West Territories series, and the standard deviation of $\hat{Y}^{*}_{t}(2)$ is negligeable in comparison to that of Y^{P}_{t+2}.

Exhibit 7

Accuracy of Y^{p}_{t+2} and $\hat{Y}^{o}_{t}(2)$ in terms of MPE and RMSPE

Beneficiaries series	Y^{p}_{t+2}	Y^{p}_{t+2}	$\hat{Y}^{o}_{t}(2)$	$\hat{Y}^{o}_{t}(2)$
	MPE (Std. Dev.)	RMSPE	MPE (Std. Dev.)	RMSPE
Canada	-1.24 (0.33)	1.28	-0.02 (0.35)	0.34
Nova Scotia	-0.31 (0.16)	0.35	0.02 (0.15)	0.16
Quebec	-0.93 (0.32)	0.98	0.06 (0.30)	0.30
Ontario	-2.13 (0.54)	2.20	-0.09 (0.65)	0.65
Manitoba	-0.71 (0.31)	0.77	0.07 (0.35)	0.36
Saskatchewan	-1.10 (0.41)	1.17	0.06 (0.41)	0.41
Alberta	-2.51 (0.44)	2.54	-0.06 (0.55)	0.55
British Columbia	-1.48 (0.54)	1.57	-0.15 (0.49)	0.50
Yukon	-5.18 (3.15)	6.04	-0.30 (3.22)	3.21
North West Territories	-12.43 (2.82)	12.74	0.26 (2.41)	2.40

The effect of combining forecasts is that the combined error is smaller on the average. $\hat{Y}^{o}_{t}(2)$ is usually much better than Y^{p}_{t+2}. $\hat{Y}^{o}_{t}(2)$ was worse than Y^{p}_{t+2} in only 22 out of 480 cases, specifically, 8 cases in Nova Scotia, 1 in Ontario, 3 in Manitoba, 4 in Saskatchewan, 1 in British Columbia and 5 in the Yukon series.

In 95.4% of the cases, $\hat{Y}_t^o(2)$ reduces the level error and consequently increases the accuracy of the change between the current period estimate and the final previous period figure. This is fundamental for the most recent figures which are often used for policy making.

5. Conclusion

It is important when preparing estimates based upon incomplete information, to take into account the behavior of the missing information. The missing information for the beneficiaries series is a systematic undercount of the number of beneficiaries. The missing information can be modelled and forecast. The method used by Huot and Plourde (1987) extends that of Harvey (1983) by modelling the bias and by lifting the restriction on autoregressive models and the constraints imposed on the order of the parameters of the models.

The results show a substantial reduction in the bias. The λ weight tends to be correlated with the size of the bias measured in terms of mean percentage error. The smaller the bias is, the more reliable the preliminary estimate is, which accordingly gets more weight.

The optimal preliminary estimates have reduced level and direction errors due to bias, in 95.4% of the 480 cases studied.

References

Box, G.E.P., and Jenkins, G.M. (1970): *Time Series Analysis: Forecasting and Control*, San Francisco; Holden Day.

Clemen, R.T. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, 5, 559-583.

Dagum, E.B. (1988), "The X-11-ARIMA/88 Seasonal Adjustment Method - Foundations and User's Manual," Time Series Research and Analysis Division, Statistics Canada.

Harvey, A.C., McKenzie, C.R., Blake, D. and Desai, M.J. (1983), "Irregular Data Revisions," in Zellner, A. (ed.), *Proceedings of the ASA-CENSUS-NBER Conference*, Washington D.C.: Bureau of the Census.

Harvey, A.C. (1984), "A Unified View of Statistical Forecasting Procedures," *Journal of Forecasting*, 3. 245-275.

Howrey, E.P. (1978), "The Use of Preliminary Data in Econometric Forecasting," *Review of Economics and Statistics*, 60, 193-200.

Huot, G., and Plourde, J. (1987), "The Impact of Actual and Estimated Preliminary Values on Seasonally Adjusted Series," paper presented at the 1988 Annual Meeting of the Statistical Society of Canada, (unpublished).

Kang, H. (1986), "Unstable Weights in the Combination of Forecasts," *Management Science*, 32, 683-695.

Ljung, G.M., and Box, G.E.P. (1978), "On a measure of lack of fit in time series models," *Biometrika*, 65, 297-307.

Morgenstern, O. (1963), *On the Accuracy of Economic Observations*, Princeton University Press, Princeton.

Rao, J.N.K., Srinath, K.P., and Quenneville, B. (1989), "Estimation of Level and Change Using Current Preliminary Data," in *Panel Surveys* (D. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh, Editors), Wiley, New York, 457-479.

Zellner, A. (1958), "A Statistical Analysis of Provisional Estimates of Gross National Product and Its Components, of Selected National Income Components, and of Personal Savings," *Journal of the American Statistical Association* 52, 54-65.

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010142149