

THE OUTLIER DETECTION AND TREATMENT STRATEGY FOR THE MONTHLY WHOLESALE AND RETAIL TRADE SURVEY OF STATISTICS CANADA

Steve Matthews and H el ene B erard¹

ABSTRACT

The Monthly Wholesale and Retail Trade Survey (MWRTS), conducted by Statistics Canada, produces estimates at various geographic and industry levels based on monthly data collected for sales and inventories. The sales trend is used as an important economic indicator, and the monthly sales estimates form a substantial portion of the estimates for the Gross Domestic Product (GDP). The MWRTS is currently being redesigned, in part to produce estimates according to the new North American Industry Classification System (NAICS) and to take full advantage of the availability of administrative data from the Goods and Services Tax (GST) program. Although many improvements will be implemented, such as a reduction of frame mis-classifications by the use of administrative data and an innovative sample update procedure, influential units will continue to occur as a result of mis-classifications and specific procedures must be put in place to treat them. This paper presents the overall strategy developed to reduce the effect of these influential units. The results from an empirical study that compares the efficiency of four proposed methods to identify and treat outliers are presented and the implementation of these methods in the context of a monthly survey production is discussed.

KEY WORDS: Influential units, Monthly Survey, Simulation study

R ESUM E

L'Enqu ete mensuelle sur le commerce de gros et de d etail (EMCGD) produit des estimations mensuelles ventil ees par r egions g eographiques et groupes industriels en utilisant des donn ees recueillies pour les ventes et les stocks. Les variations dans le niveau des ventes d'un mois   l'autre constitue un important indicateur  conomique. De plus, les estimations mensuelles des ventes forment une portion substantielle des estimations du Produit int rieur brut (PIB). L'EMCGD est actuellement dans un processus de remaniement en partie pour produire des estimations suivant le nouveau Syst eme de classification des industries de l'Am erique du Nord (SCIAN), et pour b en eficier de la disponibilit e de donn ees administratives provenant du programme de la Taxe sur les produits et services (TPS). Dans la nouvelle enqu ete, plusieurs proc edures seront mises en place afin de r eduire les erreurs de classification dans la base de sondage dont une meilleure utilisation des donn ees administratives ainsi que l'introduction d'un processus innovateur pour la mise   jour de l' chantillon. Toutefois, malgr e ces nouvelles proc edures, des erreurs de classification subsisteront et des proc edures sp eciales doivent  tre  labor ees afin de traiter les unit es qui sont mal class ees. Cette communication pr esente la strat egie d evelopp ee pour r eduire l'effet de ces unit es influentes. Les r esultats d'une  tude empirique qui compare l'efficacit e de quatre m ethodes propos ees pour identifier et traiter les valeurs aberrantes sont pr esent es, de m eme que l'int egration de ces m ethodes dans le contexte de la production d'une enqu ete mensuelle.

MOTS CL ES : Enqu ete mensuelle,  tude de simulation; valeurs aberrantes; valeurs influentes

1. INTRODUCTION

The Monthly Wholesale and Retail Trade Survey is a large-scale business survey conducted by Statistics Canada. The survey produces estimates of total sales and inventories for businesses in Canada involved in wholesale or retail trade. These estimates, based on monthly data collected directly from businesses, are published at various province and industry levels. The estimates derived from the survey form a substantial portion of the Gross Domestic Product (GDP) and the month-to-month sales

trend is used as a performance indicator of the economy. Furthermore, the MWRTS serves as the frame for the Quarterly Retail Commodity Survey (QRCS).

The survey was last redesigned in 1988. Many steps have been taken since then to adapt the design to new demands and to maintain the quality of the survey estimates. For example, in 1997, a restratification exercise took place to deal with the growing number of businesses that had changed size and were no longer stratified to the proper size strata (Tr epanier *et al.*, 1998). However, a complete

¹ Steve Matthews and H el ene B erard, Business Survey Methods Division, Statistics Canada, 11th Floor, R.H. Coats, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6, steve.mathews@statcan.ca, helene.berard@statcan.ca.

redesign is now necessary to deal with the aging 1997 stratification variables and to meet new requirements. These new requirements include producing estimates according to the North American Industry Classification System (NAICS), making use of newly available administrative data, and reducing response burden (Bérard, 2001).

As part of the redesign, new methods were investigated to identify and treat influential units that result from frame misclassifications. Section 2 of this paper gives an overall description of the MWRTS methodology. Sections 3 to 5 describe the process developed for this purpose, and present results from a simulation study to compare various methods.

2. THE MONTHLY WHOLESALE AND RETAIL TRADE SURVEY

2.1 Survey Frame and Sample Design

The frame for the MWRTS is drawn on a monthly basis from Statistics Canada's Business Register (BR). The BR is a dynamic database that contains information on all active businesses in Canada. The BR maintains a four-level statistical structure, which represents each business through units called, from top to bottom, enterprises, companies, establishments and locations. The database contains many variables, including the NAICS code, geographic classification and various size measures such as the Gross Business Income (GBI), annual sales collected via the Goods and Services Tax (GST) program, and the revenues reported on the corporate tax declarations. For more information on the BR, please see Laniel *et al* (1996).

The sampling unit for the redesigned MWRTS is the cluster of establishments consisting of all establishments belonging to the same enterprise that are coded to the same industry group (derived from the NAICS) and geographic region. The populations for the retail and wholesale sectors consist of approximately 190,000 and 100,000 clusters respectively.

The retail and wholesale populations are each divided into two portions, the surveyed and non-surveyed portions. All clusters below a pre-determined size for each industry and geographic region are assigned to the non-surveyed portion. To reduce the response burden of small businesses, auxiliary data are used to produce estimates for this portion of each population.

All other clusters are assigned to the surveyed portion, and are stratified according to their industry group and geographic region. Within the intersection of each

industry group and geographic region, three strata are created based on the estimated size of the business. The stratum containing the largest businesses is sampled with certainty and the other two size strata are sampled by simple random sampling according to sampling fractions resulting from a square root allocation based on the estimated size measure. The same sample is retained from month to month, and each month, the births to the population are sampled according to the sampling fraction of their stratum. For more information on the frame and sample update procedures see Majkowski (2001).

2.2 Estimation

The estimator considered for the new design is the Horvitz-Thompson estimator. The estimate of the population total Y of a variable of interest for a given domain is then of the form:

$$\hat{Y}_d = \sum_{s_d} w_{hi} y_{hi}$$

where, s_d is the set of units selected in the sample that fall into domain d , w_{hi} is the sampling weight (the inverse of the selection probability) of the i^{th} unit in stratum h , and y_{hi} is the observed value of the variable of interest (sales or inventories) for the i^{th} unit in stratum h .

3. OUTLIERS

Each unit selected in the sample contributes the amount $w_{hi} y_{hi}$ to the estimate. Units contributing a large portion to the total estimate are called influential units as they have a large impact on the estimates of population totals and their variances. Note that at the estimation stage of the survey, we assume that we are dealing with true values, as erroneous data has been treated previously in the edit and imputation process.

Although the estimate of the population total is unbiased with respect to repeated samples, in practice our population estimates are based on one realised sample. Influential units in the take-all strata represent only themselves and hence are not targeted by the outlier detection and treatment processes. However, we could overestimate the population totals if the realised sample includes many influential take-some units that are not representative of other units in their stratum. Given that the sample will consist of the same units from month to month, the presence of these influential units in the sample can cause recurring problems in the estimates of monthly population totals.

Influential units in take-some strata that are not representative of other units in their stratum are a result of

mis-classification (i.e. differences between stratification information and the real classification of the units). In general, mis-classification is rare in terms of geography. Mis-classification by industry is more common. If a unit is stratified in a certain industry group with a high sampling weight but upon collection, the unit falls into the domain of a different industry group (associated with lower sampling weights), the unit can be influential as its contribution to the total estimate can be high relative to other units in the domain.

Historically, the most problematic cases have been mis-classifications by size. For example, a unit will be deemed influential if its reported sales are much larger than the other units in the same size stratum. This situation may arise if the reported sales value is unexpectedly large compared to the size measure. A new size measure was developed for the redesign that will make use of independent survey data and various administrative sources (Bérard, 2001). This new size measure will contribute to reduce the occurrence of mis-classification by size. Inevitably, some misclassification by size will occur, and a strategy, described in Section 4, is required to identify influential units and propose a means to reduce their impact.

3.1 Types of Influential Units

Before describing the strategy, it is important to note that for the purpose of our survey, we have classified the influential units into two categories; ‘temporary’ and ‘permanent’.

Temporary influential units report an irregular value for a particular month, or for a short series of months. Within the edit and imputation system, units with large differences from month-to-month are identified, reviewed, and not used to impute others. In case subject matter experts wish to adjust these units for estimation, diagnostic reports from edit and imputation will be provided to them for review and possible correction.

Permanent influential units report an irregular value for a given month, and are expected to continue to report irregular values in the future. If the weighted value is high, this can have a large impact on estimated totals for certain domains. The decision was made to develop a strategy to treat influential units, particularly those that are influential due to size mis-classification. If the estimated size were accurate for each unit on our frame, we would not encounter these problems as relatively homogeneous size strata would be formed. Thus, units would contribute similar amounts to estimates. However, influential units may result from mis-classifications in variables used for

stratification, or from the growth of particular units since stratification.

4. OUTLIER DETECTION AND TREATMENT

The outlier detection and treatment strategy consists of four steps that focus on the primary output of the survey, the monthly sales.

4.1 Proposed Strategy

Step 1: Identification of Suspicious Domains

Since limited time is available to identify, investigate and treat outliers during the tight monthly production schedule, we seek a method that is efficient in finding and treating units that are influential at the published domain level. The first step identifies suspicious domains, and only these domains are subjected to further consideration. The set of suspicious domains is identified as follows.

The forecasted sales for the current month are calculated (using time series methods) based on data from previous months. The forecasted data are compared to the sample-based estimates. Any domain for which the sample-based estimate is greater than the forecasted estimate (beyond a specified tolerance) is considered suspicious. Domains for which the estimate and the projection agree are not considered further. At the time of this article, the methodology for Step 1 has not been fully developed. Until this is completed, all domains are considered suspicious by default.

Step 2: Identification and Treatment of Influential Units within Suspicious Domains

Within each suspicious domain, an outlier detection and treatment program is executed to identify influential units within the domain. A proposed treatment is derived for each influential unit. Four outlier detection and treatment methods have been evaluated for this step. Section 5 of this paper describes these methods and an evaluation study carried out to assess them.

Step 3: Subject Matter Review of Proposed Treatments

Once a treatment is proposed for influential units, a report is generated that includes the treatment, and its impact on both the level (total monthly sales) and the trend (month-to-month change) estimates. This report is provided to subject matter experts to assist in their analysis. The final treatment is determined by subject matter experts who select whether to implement the proposed treatment, or some other treatment based on other sources of information.

Step 4: Implementation of Treatments

Once a treatment is determined, it is applied to the unit at the estimation stage via a correction factor applied to the weighted sales used to produce the estimates. This correction factor is applied by default for the current month and all subsequent months until such time as a subject matter expert determines that it is no longer necessary.

5. EVALUATION STUDY

A simulation study was carried out to compare four methods described below for the identification and treatment of influential units within suspicious domains (Step 2).

5.1 Methods under Consideration

Modified Fuller: The Fuller “Test and Treat” method, described in Fuller (1991), was designed to minimise the Mean Square Error (MSE) under simple random sampling, thus no consideration is given to a stratified design with different sampling weights by stratum. This method rests on the assumption that values in the tail of the distribution follow a Weibull distribution. The method performs a hypothesis test based on this distribution to determine if outliers are present, and if so, how many. If the hypothesis test indicates that outliers are present, all outliers have their values reduced to the largest non-outlier value with an added adjustment. In the simulation study, the added adjustment of zero was used (i.e., outlier values were changed to largest non-outlier value). If the hypothesis test is not significant, no treatment is applied. An adaptation of the Fuller “Test and Treat” method was used in this study since the contributing units to the domain estimates may originate from different strata with various weights. The weighted values of sales were used as inputs to the method. The method was restricted to identify at most two outliers, as Fuller (1991) shows that, in practice, treating more than two outliers leads to very small efficiency gains.

Kokic and Bell : This method, detailed in Kokic and Bell (1994), was designed to minimise the MSE under a stratified simple random sampling design. An optimal cut-off point is estimated for each stratum through an algorithm and all units with values larger than this cut-off point have their values reduced to the cut-off point. This method always identifies at least one outlier, and can identify many.

Deflation Factor: This method is currently used in the MWRTS to treat outliers. A cut-off point for each domain is determined using the quartiles of the sample observations. For details on the calculation of the cut-off point, see the description of the quartile method in Lee *et*

al (1992). Units with a value larger than the cut-off point have their values reduced to the cut-off point through a deflation factor applied to the weighted sales. This method can identify any number (including zero) of outliers.

Take-All Promotion: This method is also currently used in the MWRTS to treat outliers (only for the most extreme cases). The same cut-off point is used as in the deflation factor method, but any unit with a value larger than the cut-off point is re-stratified to a take-all stratum (i.e., its weight is changed to one, and the weights of other units in its original stratum are re-calculated based on updated sample and population sizes).

5.2 Simulation Study

The population for the simulation study is the survey frame for June of 2001. A sales value was modelled for every unit in the population using auxiliary data from the survey frame and historical data from the survey. For more details on the modelling of the sales, see Majkowski (2001). This allowed us to calculate the population total, and to simulate realisations of different samples from the population. From this modelled population, 500 samples were drawn according to the survey’s sample design, and each outlier detection and treatment method was applied to the sample data. Estimates of domain totals were produced from the treated data.

Since influential units in the population are a relatively rare event, we created alternate populations with more influential units in order to assess the methods in the presence of more outliers. We started with our modelled population described above (Population 0) and increased the sales values of randomly selected units to levels historically observed for units in the survey. We refer to the resulting dataset as Population 1. We repeated this step starting with Population 1 to generate additional outliers, and refer to this dataset as Population 2.

5.3 Evaluation

The methods under consideration are being compared using a number of criteria. We calculated the bias introduced by each method, for each population, and each domain, in terms of Relative Bias (RB) estimated by:

$$RB_{pm}(d) = \frac{1}{R} \sum_{r=1}^R \frac{\hat{Y}_{pmr}(d) - Y_p(d)}{Y_p(d)}$$

where $Y_p(d)$ is the sales total for the domain d in population p , $\hat{Y}_{pmr}(d)$ is the sales total estimate from simulation r and outlier detection and treatment method m in population p , and R is the total number of simulations.

We calculated the variance of the resulting estimates from each method for each population and each domain in terms of the Relative Standard Error (RSE) estimated by:

$$RSE_{pm}(d) = \sqrt{\frac{1}{R} \sum_{r=1}^R \left[\frac{\hat{Y}_{pmr}(d) - \hat{Y}_{pm}(d)}{Y_p(d)} \right]^2}$$

where $\hat{Y}_{pm}(d) = \frac{1}{R} \sum_{r=1}^R \hat{Y}_{pmr}(d)$

As well, we constructed histograms of the resulting estimates from each method for visual inspection of their distribution.

5.4 Results

The results presented here are for selected domains in the wholesale trade sector, but results for the retail trade sector are similar. Tables 1 and 2 give summaries of the estimated Relative Biases, and Relative Standard Errors associated with each method from the 500 simulations.

Results show that, no one method performs best in terms of RB and RSE for all populations and domains, however it is apparent that some have more desirable features than others. In particular, not applying any treatment leads to a very small estimated bias, but can result in high relative

Table 1 - Summary of Estimated Relative Bias (%) in Wholesale Trade

RB _{pm} (d)	Food Products	Motor Vehicles	Machinery and Equipment	Electronics
Population 0				
No Treatment	0.02	-0.08	0.12	-0.17
Modified Fuller	-0.59	-0.66	-0.53	-0.59
Kokic and Bell	-0.82	-0.91	-0.79	-0.89
Take-all Promotior	-1.65	-0.71	-2.37	-0.90
Deflation Factor	-0.83	-0.53	-0.94	-0.49
Population 2				
No Treatment	-0.03	-0.25	0.14	-0.26
Modified Fuller	-2.22	-3.93	-3.19	-3.26
Kokic and Bell	-1.53	-2.19	-1.90	-2.01
Take-all Promotior	-3.94	-4.39	-5.22	-3.75
Deflation Factor	-2.86	-3.51	-3.87	-2.86

Table 2 - Summary of Estimated Relative Standard Error (%) in Wholesale Trade

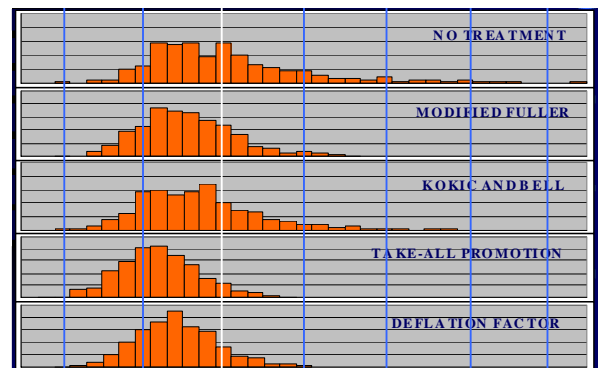
RSE _{pm} (d)	Food Products	Motor Vehicles	Machinery and Equipment	Electronics
Population 0				
No Treatment	3.55	2.64	3.57	2.58
Modified Fuller	2.46	2.73	3.08	2.61
Kokic and Bell	2.72	2.50	3.19	2.47
Take-all Promotior	2.54	2.74	2.97	2.68
Deflation Factor	2.44	2.65	2.95	2.60
Population 2				
No Treatment	4.72	5.78	6.78	5.29
Modified Fuller	2.93	3.71	3.77	2.93
Kokic and Bell	3.59	4.75	5.02	4.03
Take-all Promotior	2.48	3.38	2.87	2.63
Deflation Factor	2.55	4.02	2.92	2.90

standard errors when outliers are present. The Modified Fuller and Deflation Factor methods give reduced relative standard errors, while leading to small relative biases. The Kokic and Bell method appears to be conservative in treating units as a small bias is introduced, but the relative standard error is not reduced as dramatically as for other methods. The Take-all promotion seems to reduce the relative standard error, but introduces a larger bias. These diagnostics suggest that each method leads to a trade-off between increased bias and reduced variance.

To compare the methods in more detail, the distributions of the resulting estimates were examined. From the simulations, histograms of the relative error, $[\hat{Y}_{pmr}(d) - Y_p(d)]/Y_p(d)$, of the estimates were constructed for the domains under study. Figure 1 illustrates the 5 distributions of the relative errors after receiving no treatment and after treatment by each of the four methods under consideration (based on R=500 samples). The histograms represent the estimates from the simulations for the domain of Food Products in wholesale trade for Population 2, which was selected as an example of typical results. In the histogram, the white vertical line is positioned at a relative error of 0%, and grey vertical lines indicate each 5% interval on the horizontal axis. When no treatment was performed, some samples resulted in relative errors as large as 20%. Desirable features in the distributions include three aspects; absence of large over-estimates (no long tail to the right), distribution roughly centered near zero (indicates small bias), and a tendency to not further reduce estimates for which the relative error is negative without treatment (this would lead to increased errors).

We observe from these histograms that if no treatment is applied when outliers are present in the population, the shape of the distribution is quite flat, and the probability of having an estimate far from the true value is fairly high. The Modified Fuller and Deflation Factor methods yield good results in the three aspects mentioned earlier. The Kokic and Bell method does not affect the large

Figure 1 – Histograms of 500 Relative Errors for Wholesale Food Products in Population 2



overestimates enough as there are still a fair number of estimates with high positive errors after the treatment. The Take-all Promotion method does eliminate the large overestimates, but the overall shift of the distribution to the left indicates a larger bias.

6. CONCLUSIONS

Although the sample design is relatively efficient and robust to outliers (indicated by a low RSE with “No Treatment” in Table 2), we note gains from implementing outlier detection and treatment procedures, particularly when a relatively large number of influential units exist in the population. The diagnostics produced do not indicate a clear choice between methods, but examination of the histograms indicated that the Modified Fuller and Deflation Factor methods yield desirable properties. The Take-all Promotion leads to a larger relative bias than the other methods, and although the Kokic and Bell method does lead to a reduction in the Mean Square Error, it doesn’t perform as well according to our criteria on the resulting distribution.

REFERENCES

Bérard, H. (2001), The Redesign Of The Monthly Wholesale And Retail Trade Survey Of Statistics Canada. Proceedings of the Survey Methods Section SSC Annual Meeting, June 2001, 812-86.

Laniel, N., Mach, L., Finlay, H. and Dionne S. (1996), Measuring Errors on the Business Register. Proceedings of Statistics Canada Symposium, Non-Sampling Errors, 35-45.

Lee, H., Ghangurde, P. D., Mach, L., and Yung, W., Outliers in Sample Surveys. Methodology Branch Working Paper BSMD-92-008E, Ottawa: Statistics Canada.

Fuller, W. A. (1991), Simple Estimators for the Mean of Skewed Populations. *Statistica Sinica* 1, 137-158

Kokic P.N. and Bell P.A. (1994), Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. *Journal of Official Statistics*, Vol. 10, No. 4, 419-435.

Majkowski (2001) Maintaining Estimate Quality and Easing Response Burden in a Sub-annual Business Survey, Proceedings of the Section on Survey Research Methods, (American Statistical Association), to appear.

Trépanier, J., Babyak, C., Marchand, I., Bissonnette, J. and St-Pierre, M. (1998). Enhancements to the Canadian Monthly Wholesale and Retail Trade Survey. Proceedings of the Section on Survey Research Methods, (American Statistical Association), 487-492.