Updated January 5, 2004

Use of Tax Data for Replacing Business Survey Data – How to Identify Problematic Tax Records?

Sanping Chen and Paul Cascagnette
Statistics Canada
Ottawa, Canada K1A 0T6 chensan@statcan.ca, cascpau@statcan.ca

Abstract

Using various administrative and tax data sources to replace key variables collected in business surveys can significantly reduce both survey cost and response burden. Data collected from administrative sources, however, are generally not subjected to the same data quality and consistency checks as those for survey data, and often do not correspond exactly with the variables targeted in the survey questionnaire. Successful use of administrative sources to replace survey data depends to a certain extent on how effectively we can address the issues of ensuring inherent data quality and adequate correspondence with the survey variables.

In this work, we present some preliminary results on using multivariate methods, Mahalanobis distance in particular, for this purpose in business surveys. Our emphasis is on the examination and utilization of the internal relationships and structures of the tax data themselves for identifying records that have large differences between the two sources and that should be treated in producing business statistics. Our results demonstrate the utility of multivariate methods in at least three key areas: direct replacement of survey response by tax data, modelling of survey response by tax data, and identification of misclassified tax records.

Key words: tax data, survey replacement, data quality, multivariate method, Mahalanobis distance.

1. Introduction

One of the major trends in statistical agencies is the increasing use of administrative records for producing various statistics (Brackstone 1987). Business surveys are not an exception in this movement. This trend is stimulated and accelerated by several factors, especially the demand for more business and economic statistics and the ever-increasing cost of conducting business surveys. These increasing demands further lead to heavier response burden and deteriorating response rates. In the mean time, the pervasive application of information technology has made more and more administrative data readily available to statistical agencies, often at little additional cost, making them an attractive alternate source of information for producing business statistics.

However, despite their increasing abundance and relative low cost, at least in their raw form, administrative data is collected for different purposes than survey data, often having different variables and concepts from that of the latter. These, and other reasons, may lead to differences between what is measured by the two processes. Thus, additional quality control processes may be needed to ensure comparability in the estimates, particularly while the transition to more administrative data is made.

In general, survey and administrative data tend to correlate well after a small percentage of outliers are removed. Once the switch to administrative data is made, however, survey data is no longer available as a reference to identify potential outliers, in terms of correspondence of survey and administrative data. Therefore, to make administrative data a practical substitute of survey data, one will have to find methods based only on the former for identifying records that may not correspond well to survey responses. This is one of the primary objectives of the current work. Note that the two conspicuous strengths of most administrative data, namely a large number of data fields and a near-complete coverage of the populations of concern, provide us with new possibilities for quality control measures.

2. Restaurant Revenues and Sales Taxes – An Illustrative Example

Much of the current work is motivated by actual projects in which the authors have been involved. One of the projects was an initiative to reduce the cost and response burden of the Monthly Restaurants, Caterers and Taverns Survey (MRCTS) conducted by Statistics Canada and uses the Goods and Services Tax data.

The MRCTS is a Canada-wide survey of the restaurant industry. The primary objective is to estimate the monthly business volume of the industry. Thus, the principal variable of the survey is the total sales of the most recent calendar month.

The Goods and Services Tax (GST) is a federal value-added tax levied on most goods and services in Canada, normally at 7% of the value of the goods and services. In some Canadian provinces, it has been combined with the provincial sales tax to form a so-called harmonized sales tax (HST), with the combined rate usually around 15%. The taxes are remitted by individual businesses periodically to various federal tax processing centres across Canada. Both the GST amount and the total sales amount (including both taxable and non taxable sales) are reported in the GST submission. It seems very desirable to use the reported sales to replace the sales figure from the MRCTS, thus significantly reduce the survey sample.

We identified these sample units which remitted GST monthly and matched their historical GST sales with survey data of the same period. Note that the GST data used for the study was raw data before any edit rule be applied to detect errors. The study shows that in some cases, the correlation with survey data and the resulting substitution estimate could be relatively low. As shown below, the standard Pearson correlation coefficient of the supposed same quantity, namely monthly total sales between two data sources, could be below 10%. Also the substitute weighted sum could be more than 400% of the survey result.

We have a ready tool, namely the GST rate—the percentage at which the tax is levied. Our argument is that it is not very likely that both the GST and the total sales fields would simultaneously have the exact same type of error during tax remission and processing. Therefore, if the sales amount field is incorrect, it would very likely lead to an abnormal ratio of the GST amount to sales. Note that the reverse is not necessary true in general. That is, due to various reasons such as rebates, deductions, sales that may not be subjected to taxation, whether provincial sales tax is included, etc., one does not have a universal tax rate for all records in the GST database and therefore, a ratio of the GST amount to sales different from 7% (or 15%) does not mean that the sales and/or GST are necessary incorrect. Fortunately, for restaurants, caterers and taverns, there are basically no non taxable sales and thus, the ratio of the GST amount to sales is likely to be very close to the GST rate. When a calculated tax-to-sales ratio is significantly either below the general 7% or significantly above the combined federal-provincial rate of 15%, then there is a sign that there might be something wrong with the tax record, hence it should be excluded from replacing survey results for the MRCTS industries.

Analyses show that this is indeed the case, namely that if we exclude tax records that have an abnormal GST-to-sales ratio, the correlation between monthly restaurant sales from tax and from the survey improves, as shown by Table 1.

Table 1. Correlation between Monthly Restaurant Sales from Tax (Raw Data) and from the MRCTS Response (Year 2000, Non-Complex Restaurants which Remitted GST Monthly)

GST-to-sales ratio	n	Pearson correlation	Spearman correlation
No restriction	5217	0.096	0.794
0.069 – 699	4921	0.583	0.868
0.69 - 69	4890	0.645	0.875
1 - 30	4872	0.647	0.878
5 - 20	4730	0.649	0.879

It should be noted that the improvement of correlation is achieved at the expense of decreasing the number of tax records usable for replacing survey data.

The GST example clearly shows how internal relationships within the administrative data can be used to help identify records that have large differences between the two sources that should be excluded from or otherwise treated before being used to replace survey response. However, this example benefits from the existence of a strong internal relationship, namely the more or less fixed ratio of sales tax to sales, especially for restaurants, caterers and taverns. While such a strong relationship may not exist in general, the principle of examining the internal data structure and identifying records that significantly deviate

from this structure can, in some situations, be extended to much administrative data, as we endeavour to show in the following.

3. Application of Multivariate Mahalanobis Distance

We now move on to another project in which we try to use business tax data to replace financial information collected in Statistics Canada's annual surveys of Service Industries. Of primary interest are several major financial variables: total annual revenue, total expenses, total salaries, wages and benefits (SWB), etc., which are also reported in annual business tax returns. We soon found that once a sufficient (and usually relatively small) number of tax records that do not correlate well with survey responses are excluded, tax data did provide fairly accurate replacement estimates.

Due to the wide range of industries and business practices, we do not have strong and simple internal relationships like the GST rate in the general business tax database. But we believe that, given that all companies operate under the same set of business rules, laws, environment and market mechanism, at least within each industry or sub-industry, business tax data should demonstrate some general structure specific to that industry. An advantage of the general business tax database is its near-completeness, as by law all live companies are required to file a tax return every year. It provides us with a much greater amount of information than a typical survey sample for establishing this internal structure.

Given the current practice in Statistics Canada that emphasizes the production of tax-data-based estimates that are comparable with that from survey data, it seems reasonable to use the difference between corresponding values from the two data sources as a major measure for indicating potentially troublesome records.

We concentrate on the multivariate structure of major business and financial fields contained in the tax database. There are several technical issues. First, because business data are known for their skewness, when used directly, most such fields are first transformed using a logarithm function to make them better suited for multivariate analysis. Secondly, many fields in the tax records are sparsely populated, since they only apply to a small subset of the businesses, usually excluding these fields from multivariate analysis. For this reason, one may not try to use too many variables in order not to significantly reduce the total number of tax records available for constructing multivariate data structure, or one may aggregate certain variables. Thirdly, there are fields derived from other variables, which cause co-linearity or near-co-linearity in the multivariate data structure. These variables would typically have to be removed from the analysis set. Finally, there is the key question of the best or better selections of variables and/or their transformations that would best help identify records that have large differences between the two sources. This issue calls both relevant business accounting expertise and extensive exploratory data analysis, which we have not had sufficient time to explore.

For our multivariate analysis, we have identified the following ten fields reported by most businesses . The first four were derived from multiple tax fields and defined differently for different industries, and the Number of Employees variable came from a separate administrative source.

Total revenue
Total expenses
Total salaries, wages and benefits (SWB)
Total depreciation
Number of employees
Total shareholder equity
Retained earned deficit
Total liability
Net income loss
Total asset

It should also be stressed, as in most studies conducted so far, due to the technical difficulties in allocating numbers belonging to a large complex enterprise to its various branch operations and establishments, our analysis is restricted to the subpopulations consisting of all simple/single enterprises, where plans for tax replacement/modelling are currently focused.

At this stage, our analysis has mainly concentrated on the calculation of Mahalanobis distance from the L1 estimator of location (the multivariate equivalent of the median) of a selected set of business variables, based on the entire tax database for

the particular industry or sub-industry. The procedure and programs used are those developed by Thomas, Patak, and Franklin (1998).

3.1. Direct Replacement Schemes

For this scenario, our principal research question is: can we use measures derived internally within the tax database to help identify tax records that are substantially different from their corresponding survey responses?

The most direct way to address the research question is to examine whether the multivariate measure (Mahalanobis distance from the location vector in our case) based only on tax data is related to the difference between tax data and the survey response.

Our finding is that properly constructed Mahalanobis distances based on multiple tax variables are almost always significantly correlated with tax-versus-survey deviations, thus can help identify tax records that do not correspond to survey response. However, the strength of the correlation varies between variables and industries, and more importantly, depends on the choice of the multivariate set on which the Mahalanobis distance is based, along with their multivariate transformations. The industries are defined using the North American Industry Classification System (NAICS).

Table 2. Spearman Rank Correlation between Mahalanobis Distances and Absolute Tax-Versus-Survey Differences. (Annual Survey Of Traveller Accommodations, Reference Year 1999, NAICS=7211, N=1558, N=248)

Difference in variable	M-distance based on 10 tax variables	M-distance based on 5 tax variables	M-distance based on 4 tax variables
Total revenue	0.251	0.132	0.132
Total SWB	0.488	0.491	0.444
Total depreciation	0.264	0.202	0.155
Total expenses	0.334	0.304	0.256
Total profit	0.336	0.352	0.283

All correlations are highly significant ($p<0.001$).

The finding shows that the higher the Mahalanobis distance, the record is likely to have a bigger difference from its corresponding survey response. In other words, multivariate measures derived internally within tax data can contribute significantly to the identification of the records that deviate substantially from their corresponding survey response. While much additional work is needed to further enhance the correlations listed here, as well as to set up proper rules for using the relationship to identify tax records that have large differences between the two sources, sometimes the identification turns out self-evident as shown by the following graph (Figure 1).

Further work notwithstanding, here we may tentatively propose the exclusion of a small percentage of tax records that have the highest Mahalanobis distance (or another multivariate deviation score) from being used in replacing survey response, to avoid producing estimates that would differ substantially from that of the survey. We find in many cases when the initial tax-versus-survey correlation is weak, such exclusion tends to improve it. No less importantly, we also find in most cases, the tax-versus-survey correlation stabilizes after the exclusion of the top 10-15% tax records in terms of Mahalanobis distance (Table 3)

Figure 1. An Example of a Self-evident Outlier in Mahalanobis Distance and Survey-versus-Tax Absolute Deviation

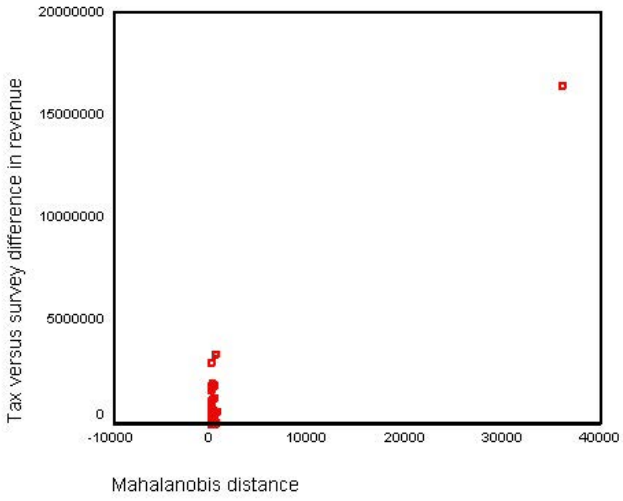


Table 3. Annual Survey of Arts and Entertainment, Reference Year 1999, All NAICS Combined. Pearson Correlation between Survey Response and Annual Tax Data after Excluding Some Top-Ranked Records in Terms of Mahalanobis Distance Based on 8 Tax Variable Ratios (Denominator Total Revenue)

Percentage excluded	*n*	Total revenue	Total SWB	Total depreciation	Total expenses
0	1503	0.762	0.552	0.705	0.770
0.5	1496	0.819	0.553	0.706	0.815
1	1486	0.822	0.554	0.706	0.818
2	1469	0.829	0.556	0.713	0.824
3	1456	0.840	0.556	0.711	0.824
5	1421	0.841	0.514	0.712	0.824
10	1342	0.847	0.579	0.696	0.823
15	1262	0.847	0.587	0.733	0.823

3.2. Tax-Data-Based Modelling

It is being recognized that, no matter how careful the mapping is done, direct replacement of survey response by administrative records is always subject to possible bias due to, among other things, conceptual, definitional and operational differences in nominally identical variables between the two data sources. Therefore, more and more efforts in this area are directed towards modelling the survey response by tax data.

We have found that Mahalanobis distance can also contribute to the identification of problematic records for modelling. In fact, if we replace the tax-versus-survey difference by the model residual (the difference between the survey response and the model-predicted response), then similar correlation exists between the residual and the tax-data-based multivariate Mahalanobis distance.

Additionally, in a modelling approach, the data quality problem is not only shown in large differences between the predicted and observed survey response, but also often manifested via the phenomenon that a few tax records would have undue large influence over the model itself, leading to unstable model predictions. In this direction, we find that the Mahalanobis distance can also be effectively used to help identify these tax records that would have led to unstable models.

More exactly, we find that properly constructed Mahalanobis distance based only on tax data is always significantly correlated to the influence of an individual record on a regression model for predicting survey response, as measured by the Cook's D influence statistic (see Table 4).

Table 4. Spearman Rank Correlation between the 10-Variable Mahalanobis Distance and Cook's D Statistic of Multiple Regression Model of a Survey Variable by Tax Variables, Annual Survey of Arts and Entertainment, Reference Year 1999, All NAICS Combined, N=721.

Survey variable modeled	4-variable (tax) model	10-variable (tax) model
Total revenue	0.220	0.273
Total expenses	0.122	0.208
Total depreciation	0.370	0.422
Total SWB	0.200	0.357
Total profit	0.142	0.277
Sum of ranked Cook's D (4 survey vars)	0.367	0.477
Minimum ranked Cook's D (4 survey vars)	0.347	0.418

Again, all correlation coefficients here are highly significant ($p < 0.001$).

The upshot of our finding is that Mahalanobis distance can be used to help identify tax records that would lead to either inaccurate model predictions or unstable models.

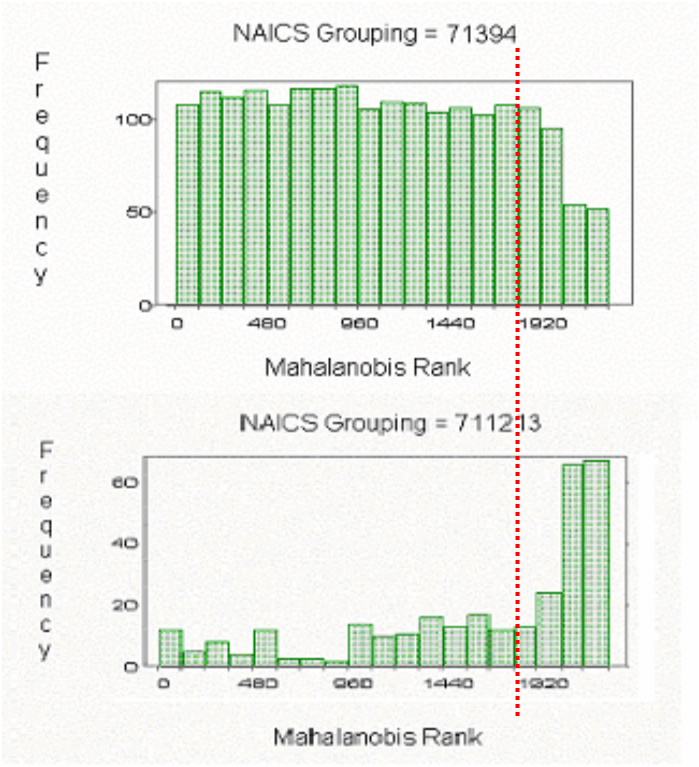
3.3. The Problem of Misclassification

An ongoing concern in business surveys is frame maintenance. In addition to deaths not identified in a timely fashion, the frame maintenance problem is often manifested in the inclusion of misclassified (out-of-scope) businesses in a sampling frame. As is well-known, regular surveys provide the additional value in survey responses and feedback that help correct and reduce the misclassification problem. Therefore, the increasing use of administrative data to replace survey responses raises the concern of deterioration in frame quality due to the decrease or disappearance of survey feedback.

We have found that the multivariate structure established through analysing the administrative data may provide help in alleviating this problem. In particular, by establishing a multivariate pattern particular to an industry, out-of-scope records can be identified by their violation of this pattern.

We demonstrate this possibility by the following example in which we deliberately mixed up two different sub-industries (NAICS 711213 Horse Race Tracks, N=312, and NAICS 71394 Fitness and Recreational Sports Centres, N=1985) covered in the same annual industry survey conducted by Statistics Canada and calculated Mahalanobis distance of the joint business tax dataset. As shown in Figure 2, the two sub-industries manifested strikingly different distribution profiles in Mahalanobis distance. Therefore, a significant percentage of misclassified administrative records may be excluded by excluding the cases with the extreme Mahalanobis distances, at the expense of a small percentage of correctly classified records. This also suggests the use of other multivariate methods (e.g. cluster and discriminant analyses) to help identify misclassified administrative records.

Figure 2. Annual Survey of Arts and Entertainment, Reference Year 2001. Distribution According to Mahalanobis Rank (10 Tax Variables) For 2 Combined Groups



3.4. Limitations of the Study

There are, admittedly, several limitations of our study.

First of all, the study lacks input and guidance from business tax and accounting intelligence, as neither author has expertise in this area. Our extensive exploratory analysis was thus not very efficient in identifying sets and transformations of tax variables that may be sensitive to erroneous records. The next phase of the study will involve such expertise.

Secondly, most administrative data used in this study were not raw records, but came from the tax database made available by the Tax Data Division, Statistics Canada, for aiding regular annual business surveys. These datasets have already been subjected to regular edit and imputation processes, including some univariate outlier detection. Consequently, the previous edit and imputation process may have altered or eliminated many outliers, thus confounding some of our analysis results.

Thirdly, due to time and resource constraints, we were not able to expand our analysis beyond Mahalanobis distance, though some initial efforts indicate similar results from the use of a new probabilistic multivariate outlier algorithm (Jibrin, Pressman and Bell, 2003).

3.5. Percentage of Exclusion and Treatment of Problematic Cases

Our study suggests that multivariate analysis can help in identifying administrative records that have large differences between the two sources that would result in estimates substantially different from that of survey responses. Practical application of this approach, however, requires answers to two additional questions.

First, even with anticipated improvement in precision with better selections and transformations of administrative data variables for multivariate identification of outliers, such an algorithm will not be error-proof. In other words, the exclusion of multivariate outliers is likely to exclude both problematic and useful records. Therefore a key methodological question is to find the optimal percentage of exclusion of multivariate outliers that best balance the need to prevent administrative records

that have large differences between the two sources from contributing to the final estimates, and the inclusion of the largest number of useful records.

This question is perhaps best exemplified by the increasing talk of population-based estimation approach in which the entire administrative database is used to produce business statistics. In such a scheme, there may always be a certain proportion of the administrative records that will have to be excluded due to their quality problems. But, what is the optimal proportion of exclusion that would result in, say, the minimal mean squared error of the estimate?

The second question is how one handles the excluded administrative records. Treat them as non-responses? Or, if the timeline allows in a survey replacement scheme, send a questionnaire to these businesses? Or one may use multivariate imputation to force the records to conform to the average multivariate structure? This seems an issue intimately related to the operational aspects of the survey or estimation process.

4. Conclusions

This work has shown that multivariate analysis can contribute significantly to the identification of administrative records that have large differences between the two sources and that should not be included in the production of business statistics, or at least, subject to verification or treatment. The results apply to three important areas for using administrative data to replace survey responses, namely—direct replacement, multivariate modelling, and identification of misclassified records.

Much additional work is needed for improving the efficiency and accuracy of the multivariate methods proposed here, as well as for addressing the important question of the optimal exclusion proportion.

5. Acknowledgments

The authors thank Thomas, Patak and Franklin for use of their computer programs for calculating Mahalanobis distance, and to Mary March for her suggestion of this subject and her continuing encouragement during the study. The authors also wish to thank Pierre Lavallée and Linda Ramsey for reviewing this paper and providing their helpful suggestions.

6. References

Brackstone, G.J. “Issue in the Use of Administrative Records for Statistical Purposes,” *Survey Methodology*, 13(1987), pp.29-43.

Thomas, S., Patak, Z. and Franklin, S. “A Multivariate Outlier Detection Routine: Using the Stahel-Donoho Estimator to Calculate Mahalanobis' Distance,” Business Survey Methods Division, Statistics Canada, 1998.

Franklin, S, Thomas, S and Brodeur, M. “Robust Multivariate Outlier Detection using Mahalanobis Distance and Modified Stahel-Donoho Estimators,” *Invited Papers, The Second International Conference on Establishment Surveys,* 2000, pp. 697-706.

Jibrin, S., Pressman, I. and Bell, G. “A Probabilistic Method for Detecting Multivariate Outliers,” Technical Report, Dept. of Mathematics and Statistics, Carleton University, 2003.