SSC Annual Meeting, June 2009
Proceedings of the Survey Methods Section

CHANGING THE METHODOLOGY FOR GENERIC TO DETAIL ALLOCATION FOR INCORPORATED BUSINESS TAX DATA

Jessica Andrews[1]

ABSTRACT

Tax data is steadily becoming a more important tool for Statistics Canada in the fight to reduce response burden and increase response rates. Before tax data can be used, however, many fields must be imputed as only 8 of 700 on the financial statements are required by the Canadian Revenue Agency. The methodology which imputes missing fields has recently been changed resulting in many improvements to the data but also several implementation issues. This paper will discuss the impact of the new methodology and how new problems have arisen and been solved.

KEY WORDS: Cluster analysis, Discriminant analysis, Generic to detail allocation, T2 tax data.

RÉSUMÉ

Les données fiscales deviennent un outil de plus en plus important pour Statistique Canada dans sa lutte pour réduire le fardeau de réponse et améliorer les taux de réponse. Avant que ces données puissent être utilisées, plusieurs champs doivent être imputés parce que l'Agence du Revenu du Canada n'exige seulement que 8 des 700 champs des états financiers soient remplis. La méthodologie pour imputer ces champs manquants a changé récemment et a entraîné plusieurs améliorations aux données mais aussi quelques problèmes lors de sa mise en place. L'impact de la nouvelle méthodologie et les solutions aux problèmes rencontrés sera présenté.

MOTS CLÉS: Analyse discriminante; analyse par grappes; désagrégation des données génériques aux détails; données fiscales T2.

1. INTRODUCTION

Every year the Canadian Revenue Agency (CRA) collects financial statements from incorporated (T2) businesses in Canada. These statements, including a balance sheet and income statement, are received from CRA by Statistics Canada on a monthly basis in order to help surveys reduce cost and response burden and improve data quality. The T2 tax data base is a census representing roughly 1.5 million legal entities with 700 fields available in the balance sheet and income statement. Of these fields only eight are deemed mandatory by the CRA, including the section totals: non-farm revenue, non-farm expenses, farm revenue, farm expenses, assets, liabilities, shareholder equity and net income/loss. Most businesses report annual statements, though some may report for shorter time periods, with the fiscal period chosen by the business.

Upon entering the Statistics Canada database the data pass through many edits in order to improve data quality. Businesses, for which statements are missing or have serious inconsistencies, have their statements imputed in order to have a complete database. However as only eight fields are mandatory and only certain field can be filled through the editing, many fields in the files will be blank at the end of the process. This causes problems for surveys wanting to use the tax data as tax concepts do not always align with survey concepts, and thus detail level fields are necessary in order for surveys to use the data.

The generic to detail allocation (GDA) system used for T2 tax data aims to fulfill the needs of surveys by imputing values for as many details of the financial statements as possible. Originally the aim of this process was to provide good results at the macro level, however, with increasing replacement of survey data with tax data, good results at the micro level are needed. With this aim in mind the methodology for the generic to detail allocation process was changed for reference year 2006 for the Income statement (see Andrews et al. (2007a)). As the methodology was significantly changed, there has continued to be a large volume of work on the project during the previous two years. Creating the system has taken about

[1] Jessica Andrews, Statistics Canada, Jessica.Andrews@statcan.gc.ca

a year for each of the income statement and balance sheet (changed for reference year 2008), explaining why the change was not implemented for both at the same time.

This paper will address some of the issues that resulted from the changes and how these were resolved. The first set of difficulties encountered were the run times required by the new methodology. Also, the change in methodology inspired users to more closely examine the data they were using. This has resulted in a series of special edits which needed to be added to the system. The new system is also more susceptible to outliers, i.e. businesses that report huge amounts, as these businesses can change results for the whole universe if mistakes are made in their allocation to details. However, on the positive side a stability study has shown that the new system has improved stability of results at the macro level. Finally we will discuss how the lessons learnt from the income statement are being used to implement the new methodology for the balance sheet.

2. T2 DATA PROCESSING

Before the generic to detail allocation process can begin the tax data undergo a number of edits and control processes which are essential for preserving data quality. This starts with the load of the data once Statistics Canada receives it from CRA. Edits are run to ensure that all sections of the financial statements are complete and balance, negative values are corrected, corrections are made for cases of overlapping fiscal statements and an outlier detector is used to find any extremely large values within a given financial statement. These edits lead to automatic corrections, apart from any that have outlying values detected which are corrected manually. Any records which cannot be corrected, either automatically or manually are set aside in an error database.

Records not in error undergo deterministic imputation after these edits on key fields such as inventories, depreciation, and salaries and wages. These imputations are done using information in the fiscal statements or from other administrative sources. The number of such edits is expanding as more fiscal data becomes usable at Statistics Canada.

Imputation is performed once a year to complete the database by imputing any records which are in error or missing. There are two different methods used, historical imputation with a trend and donor imputation by nearest neighbour. These imputations are done in the September following the end of the T2 tax data reference year which lasts from April 1st of one year to March 31st of the following year. Thus a record from reference year 2008 will only be imputed as of September 2009.

Historical imputation is used whenever there is good quality data from the previous reference year, and imputes the whole of the record. The one exception is when either the balance sheet or income statement is available when only the missing part of the record will be imputed. The balance sheet and income statement each break naturally into two parts (liabilities and assets; revenues and expenses) thus four separate trends are calculated, with over 90% of records being imputed from a trend calculated at the industry(NAICS 6) and province level. The vast majority of records needing imputation are imputed using this method. Studies have found that this imputation performs very well.

If a record is missing or in error and there is no previous good data, then donor imputation is used. This is done by matching businesses to nearest neighbours using estimates for revenue, assets and expenses and comparing variables such as NAICS, region, and fiscal period. With donor imputation the whole record is imputed and scaled to match the values for revenue and assets.

One final imputation is then used which sets some businesses to inactive by nullifying their income statements. This is only done for businesses which have been historically imputed for at least two years, which are expected to report GST amounts but which have not reported GST or other financial information in the current year (i.e. have not reported GST or payroll (PD7) amounts).

Once the data base is complete generic to detail allocation takes place, the method will be explained in the next section. After this step is complete the cells in the financial statements are combined into chart of account variables which match survey concepts. The records are then rolled up to the enterprise level and data is allocated to the establishment level. For more information on all of these processes please see Andrews et al (2007b) and Hamel and Martineau (2007).

3. GDA PROCESSING

The aim of generic to detail allocation is to fill in as many of the details in the financial statements as possible by assigning values based on total, subtotal or generic amounts. The structure of the data is such that most generic (or "other" fields) are in blocks including detailed fields, which then add up to a subtotal with the subtotals adding up to give large section totals. The new methodology is concerned solely with moving amounts reported in the generic field to the detail fields (see Table 1). Businesses may report amounts in a generic field when they choose or are unable to report amounts in specific detail fields for a block, and in blocks not treated by GDA when none of the specific details are appropriate for the amounts being reported.

For reference year 2006 a new methodology was introduced for the income statement at the generic level. The old methodology, in use for the balance sheet until reference year 2008, grouped all businesses by their industry code (the first two digits of the NAICS code) and their size based on three different revenue classes. Although this method produced good results at the macro level, investigation showed that distributions within imputation classes were not homogenous and that the imputation was not providing good micro data (Huang and Ladiray (2005)).

As a result a new methodology based on cluster analysis was introduced (for details on the selection of the methodology see Andrews et al 2007a) which instead formed imputation groups on the basis of detail distribution. The groupings were chosen so that detail distributions within imputation classes were very similar, with the possible exception of one take all group which was used for detail distributions which are not common enough to be separated into different imputation groups. Once the groups are decided, businesses with unknown imputation group (i.e. reporting a generic amount) are assigned to a group through discriminatory models. The models were created using pairs of years as test data and finding the models that were best able to predict the known distributions of data. The new methodology was shown to produce much better micro results and similar macro results.

Table 1: This table shows one of the blocks which is part of the GDA process. Case 1 is used to establish ratio distributions, whereas case 2 and case 3 need to undergo GDA so as to move the amount in 8810 to the detail fields

Block 8810		Case 1	Case 2	Case 3
Generic Field	8810 Office Expenses		100	50
Detail Fields	8811 Office Stationary and Supplies	50		60
	8812 Office Utilities	15		
	8813 Data Processing	15		

The ratios for GDA are calculated once a year in April using all data in the survey universe that have arrived for the current reference year, received data for records from the previous year which have not been received for this reference year, and all received data from two reference years ago. Businesses which have responded only detail amounts within a block and 0 for the generic field are used to calculate the ratios for the imputation groups for the block. Then each month, after data have been loaded and passed through all the edits, records are assigned their proper cluster in each block and ratio distributions for their generics are imputed.

An interesting distinction between the old and new methodologies is that most of the error for the old methodology came from the calculated ratio distribution, due to averaging over companies whose detail allocations were quite distinct, while very little came from assignment to imputation group due to the easy definition of imputation groups. The new methodology is almost the opposite, as most imputation groups have very homogenous detail allocation patterns there is

little error due to the distribution calculations. Errors occur more frequently, however, due to a business being assigned to an inappropriate imputation group by the discriminatory model. Thus where as the old methodology provided many records with an unrealistic looking distribution of details (an average distribution of regular distribution patterns), the new methodology assigns much more realistic distribution patterns to businesses, and generally gives better estimates.

4. IMPLEMENTATION ISSUES

The first issue which arose on implementation of the new methodology was run time. The old methodology had one ratio creation run in the spring, and thereafter businesses were easily assigned to the correct imputation group using their NAICS and size. The new methodology while taking a similar run time for ratio creation in spring requires a more complicated process be run each month in order to assign businesses to their correct imputation group. Initially part of the problem for the system was that the creation of the data set for calculating distributions was done from a database where businesses had already undergone the GDA process, thus resulting in many more businesses that could be used for the ratio calculation. As nonparametric discrimination in SAS takes a long time to create the tree of known data, significantly increasing this amount of data had a large impact on run time. Once this original mistake was discovered this reduced run time considerably. This lesson also helped to solve the problem for some of the slowest running blocks. In all cases where a parametric model performed equally or almost as well as a nonparametric model the method was changed. This produced significant time savings, cutting hours off the run time for each block which was remodelled. Finally the program was divided into separate sub programs so that different blocks could be run concurrently. The end result was a decrease in the spring run from four days to less than two, and a decrease in the monthly run time of a day to approximately four hours.

The original method of performing GDA had been criticized by users for not working well. Focus groups on why the method was working badly, however, were fairly unsuccessful with little in the way of detail being given by users. One unexpected positive impact of implementing the new system was that users began to critically examine the data and give much more feed back on which aspects they were unhappy with. As a result specific rules could be implemented.

Several initial complaints dealt with details which were allocated incorrectly. In some cases these had not been problematic before either because the detail was fairly rare, and thus was not included by the old methodology; or because allocation to the detail was either industry or size related. An example of the first kind of problem for a rare detail was field L8238 on the income statement which is an Alberta royalty tax credit. Obviously this detail should only be allocated if the business is doing a significant proportion of its business in the province. Thus new rules were added to ensure that this detail was only allocated for Alberta companies.

An industry related problem that occurred was with details relating to freehold mineral tax expenses and mining tax expenses. Again it is clear that these details are only relevant for businesses which are involved with mining. It is rare that these details would be responded to by a business of any NAICS other than a mining NAICS, and thus under the old GDA system, there had been few problems with allocation to these details. However under the new system, there were more non mining companies being allocated this detail, so a simple rule for the cluster determination was created which barred non mining companies from allocating amounts to these details.

A size related problem also occurred for a block covering unspecified interest bank charge expense amounts. Initially the new methodology was found to be over allocating to bank charge expense amounts, rather than to details such as long term debt interest amounts. Upon examination, this was mainly due to a few large companies which resembled smaller companies in their behaviour in other blocks. Obviously, if several hundred thousand was located in this block, it is more likely it would be due to mortgage or long term debt expenses than bank charge expenses. Thus the cluster assignment for this block was changed to reflect this user knowledge to improve the system.

In April of 2009 further changes were made to the system as the editing system was updated to incorporate new data which had become available. As a result three fields in different blocks were no longer processed by the GDA system as their values could be imputed deterministically from schedules that businesses supply along with the financial statements. At the same time one of the most significant blocks for surveys was dropped from the GDA process. After intense examination by subject matter specialists and a series of discussions, it was decided that this block did not contain an

exhaustive list of detail fields and thus the generic amount could not safely be set to zero and the amount moved to its detail fields.

5. IMPACT OF THE NEW SYSTEM

The implementation issues which arose for individual data fields were to be expected with the switch to a new and totally different imputation system. However, the new system seems to have performed well when considered over all fields. Imputed data appears more realistic, and although it is impossible to judge whether it is more accurately imputed, from the tests run on known data and the happiness of the users with the system it is felt that there has been a definite improvement in data quality. As hoped, macro results have remained fairly consistent with the old system. A question arose, however, on how stable the data base was under the new methodology versus the old.

A stability analysis of the income statement was done over the entire data base after the new methodology had been used for two years. This looked at any large changes in within block percentages of number of businesses reporting a particular detail after GDA and percentage amounts reported in details after GDA. There were a large number of such changes between the old and new methodology, a number of which were due to the rule used for the old methodology that no detail reported by less than 10% of businesses in a class being allocated an amount. When comparing between years for the new and old methodology we have seen a rise in stability of the data base with the new methodology having significant changes in details for half the number that the old methodology did. In particular, only four of the details undergoing the GDA process changed their block percentage amount by more than 5% between reference years 2006 and 2007 versus 10 details for reference years 2004 and 2005. Some of the changes for the new methodology were also easily explained by the economy, such as an increase in vehicle fuel expenses given the increasing price of gas over the years in question.

The results for the income statement have had a considerable impact on how we have proceeded with modelling the other half of the financial statements, the balance sheet. The balance sheet has provided some new problems, one being that there are some blocks which are very closely related such as the blocks for accounts receivable and allowance for doubtful accounts. The best models usually use fields from these related blocks, however, as there is also a link in probability of response to the blocks models had to be carefully assessed. We also frequently ran into the problem that the allocation amounts for similar details could not be well fitted, thus in certain cases we have aimed instead at good allocation to groups of similar details (which are then used in the same survey variables) rather than on concentrating on individual details. The experience of working on the income statement has resulted in a preferred use of parametric models, with non parametric models being used only in cases where no parametric model did reasonably well. Further, a heightened awareness of the types of issues that can arise with detail allocations resulted in the creation of clustering rules for businesses reporting large amounts in certain blocks, prior to implementation.

6. CONCLUSIONS

On initial impressions the old methodology for generic to detail allocation was a very good system. It was easy to implement with clearly defined imputation groups, and intuitively these imputation groups made sense. Furthermore they matched imputation groups used in surveys and provided good macro results. As the use of tax data at Statistics Canada has evolved, however, the weaknesses with this system particularly at the micro level have become more apparent. Thus a new system which was shown to have better micro results has been implemented, with clearly defined ratio distributions in imputation groups but more complicated rules for assigning businesses to imputation groups.

Implementing this new methodology has been a long process involving working with many different users of tax data. The implementation of the new methodology has led to increased participation of users in the system, and finally allowed more of the users' knowledge to inform the imputation system. The result has been a much improved system, which is still being changed as users continue to examine their data more closely. In building the system care was taken to allow the system to be changed as requests come in, resulting in a more flexible system capable of dealing with a number of special cases simultaneously.

REFERENCES

Andrews, J., F. Brisebois, and N. Hamel (2007a). "Methodology of Allocating a Generic Field to its Details." *ICESIII, The Third International Conference on Establishment Surveys, June 18 to 21, 2007*, Montreal, Canada. American Statistical Association, Alexandria (Virginia).

Andrews, J. et al. (2007b)**.** Methodology for the Processing and Imputation of Corporations Data (T2). Statistics Canada Working Paper BSMD-2007-009E .

Hamel, N. and Martineau, P. (2007). "Évaluation de la qualité des données sur les sociétés incorporées (T2) produites par la Division des données fiscales." Working paper, Statistics Canada, BSMD-2007-005F.

Huang, R. and Ladiray, D. (2005). "Imputing Distribution in Administrative Tax Data." *Proceedings of Statistics Canada's Symposium 2005* - Catalogue no. 11-522-XIE.