Comparison of Business Revenues from two Administrative Files

Guylaine Dubreuil, François Brisebois, Patrice Martineau, Julie Girard, Caroline Rondeau
Business Survey Methods Division, Statistics Canada,
100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6
(guylaine.dubreuil@statcan.gc.ca, francois.brisebois@statcan.gc.ca, patrice.martineau@statcan.gc.ca, julie.girard@statcan.gc.ca, caroline.rondeau@statcan.gc.ca)

Abstract

A portion of the administrative data used at Statistics Canada (StatCan) comes from the T2 database, which is related to corporations and is made up of financial and fiscal data. These administrative data are provided by the Canada Revenue Agency (CRA). StatCan also receives from the CRA information related to the Goods and Services Tax (GST) that businesses have been remitting to the CRA since 1991. The GST information provided includes total revenue and taxes collected on products and services over a given reporting period.

Most corporations are present on both sources of data and it is therefore possible to compare their revenue variables and assess the strength of the relationships between them. This paper will focus on the different results obtained from the comparison and discuss improvements that can be brought into the methodologies used to process these data.

Key Words: Administrative data, comparison of variables, influential data, regression, survey replacement.

1. Introduction

In order to reduce response burden and data collection costs, StatCan has been constantly increasing its use of administrative data. These data are used as replacement in annual and sub-annual business surveys. One of the most widely used variables is the revenue generated by enterprises. In parallel, StatCan has put more effort into improving the coherence in estimates produced by various surveys. In particular, it has become a methodological issue between annual and sub-annual surveys covering the same type of industry.

A significant portion of the administrative data used at StatCan comes from the T2 database, which is related to Canadian corporations and is made up of financial and fiscal data. These administrative data are annual data provided by the Canada Revenue Agency (CRA).

At the same time, StatCan also receives information from the CRA that is related to the Canadian Goods and Services Tax (GST) collected by both incorporated and unincorporated businesses. This tax is levied on the principle of a value added tax. The information provided by the CRA to StatCan includes the total revenue as well as the GST collected over a given reporting period. Depending on the size of the business reporting periods are monthly, quarterly or annually. A calendarization process, however, converts all data to calendar month periods.

In general, annual business surveys use T2 data as replacement for sampled units, while sub-annual business surveys use GST data. Since different sources of administrative data are used in these surveys, the coherence partially depends on the coherence between the tax data sources. This paper focuses on comparing T2 revenue and GST revenue. One of the goals is to better identify the elements which may explain potential differences between the two sources. Also, we will try to evaluate the precision of the revenue concepts based on the two sources. At the same time, we will evaluate the possibility of using one source to improve the quality of the other in order to help increase data coherence.

Section 2 describes the two databases and Section 3 presents the target population. Sections 4 and 5 discuss the different types of analyses conducted to compare the data and show some of the results. For more details on the comparison analysis, see Dubreuil and al. (2008). A summary of the results concludes the paper.

2. Description of the Databases

2.1 The T2 Database

T2 data are provided at StatCan on an annual basis and refer to fiscal periods associated with a fiscal year. For a given fiscal year *Y*, the T2 fiscal period might end anytime between April *Y* and March *Y*+1. Each year, StatCan applies an edit and imputation process to the T2 data before saving the resulting data on an annual database. On the database, the businesses are identified by a business number (BN). For further details on the T2 processing, see Andrews and al. (2007).

The T2 database is made up of financial and fiscal data. There are in fact many revenue variables available for the comparison. After analysing the different revenue variables, it was decided to use Total Revenue which appeared to be the most appropriate variable when covering all industries globally. By definition, Total Revenue includes the sales of goods and services, investments and some extraordinary elements. In some cases, the total revenue can be negative. For the comparison study, we will refer to this revenue variable as T2_Revenue.

2.2 The GST Database

In parallel, StatCan receives GST information in the form of declarations including total revenue as well as the GST collected over a given reporting period. Then, StatCan applies different processing steps, such as edit and imputation, followed by a calendarization process which converts the GST declarations to calendar month declarations by taking into account the seasonal patterns. At the end, the GST data are stored on a full longitudinal calendarized database. For all months when a business was active, full monthly data are available. On the GST database, the businesses are identified by a BN. For further details on the GST processing, see Dubreuil et al. (2003), Brodeur and Pierre (2003), Quenneville and al. (2003), as well as Beaulieu and Quenneville (2008).

The GST reporting frequency depends on the size of the businesses. Businesses having annual sales over $6 million are required to report monthly. Those with annual sales between $500,000 and $6 million have to report at least quarterly. Finally, businesses with annual sales between $30,000 and $500,000 can report annually. When their annual revenue is below $30,000, businesses are not subject to the GST and therefore not required to collect and remit the tax. There is only one revenue variable on the GST database, which by definition includes sales and other revenue and it cannot be negative. However, there are exempted goods and services for which businesses do not need to report related sales. These exempted goods and services are concentrated in specific industries (ex: health services). For the comparison study, a list of the industries that are most likely to have GST exempted goods and services has been established and more attention is given to them.

3. Target Population

For the comparison study, 2004 T2 data were used, meaning T2 fiscal periods ending between April 2004 and March 2005. Knowing that the GST data are available on a calendar month basis, only the T2 fiscal periods starting on the first day of a month and ending on the last day of a month were kept for the comparison study. Initially, 1,560,188 T2 records were kept.

Then, T2 records were matched to monthly GST data by BN. Only monthly GST data for the matched businesses having the exact corresponding period are used for the comparison. The GST revenue, called GST_Revenue is then derived using the sum of the GST monthly revenues covering the T2 fiscal period. That way, the comparison of the two revenues associated to the same business is done over exactly the same period. At the end, the total population used for the comparison includes 961,638 records, representing 61.6% of the initial number of T2 records.

There are many reasons why a T2 record did not have a GST match. There are records for which the business matched but covered different periods in the two data sources. There are also businesses which were identified as inactive in the GST database but not necessarily in the T2 database. Finally, there are businesses which are not subject to the GST because they are under the revenue size threshold or because most of their revenues come from exempted goods and services. Nevertheless, having almost one million records was considered sufficient to conduct a good quality comparison analysis.

4. Descriptive Analysis

4.1 First Overview on the Total Population

Before starting any comparison analysis, differences in revenue were anticipated since there are differences in the underlying concepts. In order to have insight on the kind of differences that might be observed between the two sources of business revenues, the first analysis simply compared the T2_Revenue and the GST_Revenue of the same business over the total population. Table 1 presents the distribution of the T2_Revenue over the GST_Revenue ratios after rounding the revenues to the nearest thousand.

Table 1: T2 /GST ratio distribution (in terms of record)
(Revenues rounded to the nearest thousand)

T2_Revenue/GST_Revenue	Total population	
	Number	**%**
Less than 0	4,165	0.43
Between 0 and 1	235,856	24.53
Equal to 1	125,327	13.03
More than 1	417,874	43.45
GST revenue = $ 0	84,264	8.76
T2 revenue = $ 0	39,321	4.09
GST revenue= 0$ and T2 revenue= $0	54,831	5.70
TOTAL	**961,638**	

We observe that T2_Revenue and GST_Revenue are equal for only 18.7% of the cases, with 30% of these having revenue equals to $0 from both sides. A small portion has negative ratios because of a negative T2_Revenue, and approximately one quarter of the ratios is smaller than one, meaning that GST_Revenue is larger than T2_Revenue. For the majority of records, however, T2_Revenue is larger than GST_Revenue (52% of records). This last result is not surprising given that the GST revenue concept might exclude some revenues, for example, revenues related to the sales of GST exempted goods and services.

This ratio analysis includes all records but does not indicate the size of the differences. Also, the relationship between the two revenue concepts probably varies from one industry to another.

4.2 Influential Data

Since the comparison analysis might be affected by some extreme differences, it was decided to identify influential data. More precisely, influential data were identified when they failed at least one of the tests used to detect them from a regression analysis applied on the total population. For more details on the detection method used, see Ladiray and Ramsey (2003). These data are influential in the sense that they affect strongly the correlation between the two revenues. Excluding influential data could give a better idea of the relationship between T2_Revenue and GST_Revenue for the remaining records. It is also important, however, to pay particular attention to these influential data to better understand why they are so different.

Out of the 961,638 compared records, 1,136 records were identified as influential, representing 0.1% of the records. We observe that 76.4% of these influential data belong to the fourth quartile, hence the largest businesses group according to both revenue values. T2_Revenue and GST_Revenue are in this quartile despite the significant difference between them. Moreover, these few data represent 28.8% of the total T2_Revenue and 20.8% of the total GST_Revenue. These large businesses frequently have a complex structure with revenue coming from several sources (ex: sales, investments, grants, etc.). Furthermore, 48.4% of these data are imputed for at least one of the administrative data source as opposed to 29.3% in the total population. Note, however, that the imputation could be done for the whole fiscal period, or just partially for the GST data (ex: one month) since many of the data are processed by sub-annual reporting periods. Another characteristic that is particularly important relates to the GST reporting periods. In the total population, only 11.4 % of records were reported on a monthly basis while 82.8% of influential data are reported monthly. This last observation is related to the fact that the influential data are usually large businesses and thus have to provide their GST information on a monthly basis.

From this influential data analysis, we were able to identify some potential problems in the T2 and GST processing systems, some of which are now fixed. Also, it gives us an idea of the type of businesses for which we should use administrative data carefully.

For the rest of the comparison analysis, we will often compare the results before and after the exclusion of the influential data, so that we will be able to evaluate their impact on the quality of the relationship between T2 and GST revenues.

4.3 Relative Difference by Industry

Now that we have identified the influential data, we would also like to have an idea of the size of the differences between T2_Revenue and GST_Revenue and analyse the impact of these differences on total estimates. Since the relationship between the two revenue concepts probably depends on the industry, relative differences of the total GST_Revenue with respect to the total T2_Revenue were calculated by industry before and after excluding the influential data. The industries are identified by the first three digits of the North American Industry Classification System (NAICS3). Note that for the multi-activity businesses, the industry was coded to the dominant industry. The NAICS was coded the same way and from the same source on both administrative databases. Table 2 presents a summary of the results in terms of number of NAICS3 within different size categories of relative differences.

Table 2: Number of industries (NAICS3) by relative difference categories
(With and without influential data)

Relative difference category (with respect to T2)	Number of industries	
	Including influential data	Excluding influential data
More than 50%	10	4
Between 20% and 50%	15	15
Between 10% and 20%	20	15
Between 5% and 10%	14	17
Between 0% and 5%	27	45
Between -5% and 0%	7	4
Less than -5%	11	4
TOTAL	**104**	**104**

We observe that for 34 industries, the relative differences are between -5% and 5% when we include influential data. This number increases at 49 industries after excluding them. This is almost the half of the industries. Thus for many industries, the differences between the total revenue from T2 data and from GST data are small as long as we are able to exclude some influential data.

For a majority of the industries, the total T2_Revenue is larger than the total GST_Revenue, and in some cases the relative differences are higher than 20%. Even after removing influential data, there are still 19 industries for which the relative difference stays high. Most of these industries are more likely to have GST exempted goods and services. Furthermore, some of these industries with large relative differences are sectors where the proportion of the T2 revenue coming from the sales of goods and services is less than 50%. Considering all industries together, the global relative difference is 14.5% before excluding influential data and 5.6% after excluding them.

Even if T2_Revenue is generally higher than GST_Revenue, there are a few industries where it is the opposite. In particular the relative difference might be higher for some agriculture sectors since the T2 revenue variable chosen for the analysis does not include agriculture revenues.

5. Regression analysis

From the descriptive analysis, we already know there are differences between T2_Revenue and GST_Revenue but these differences seem to depend on the industry. However, we know very little about the kind of relationship existing between the two revenues at the business level. A linear regression analysis is now used to estimate the correlation between the revenues and also to estimate the intercept and the slope. The linear model is used only for analytic purposes, not to predict the revenue of one source from the other one. We have decided to use this approach

because it could detect a relationship between the two revenue concepts even if the hypothesis that they are equal is rejected.

Let the linear model be the following:

$$GST\_Revenue_i = \beta_0 + \beta_1 * T2\_Revenue_i + \varepsilon_i .$$

By using the least squares method, we can estimate the parameters β_0 and β_1, the intercept and the slope. If the concepts underlying the two definitions of revenue are the same, we would then expect that the slope is close to 1 and the intercept close to 0. Because of the large size of our population the usual statistical tests were always significant, thus, we must use our own judgment to determine if the intercept and the slope are close to 0 and 1 respectively.

5.1 Linear Regression on the Total Population

As a starting point, we estimated the linear regression model on the total population. If we do not exclude any influential data, the estimated intercept is half a million dollars and the slope is 0.60. The coefficient of determination (R-Square) providing the variability explained by the model is equal to 0.68. However, as soon as we exclude the influential data, we estimate the following model:

$$GST\_Revenue_i = -74{,}467 + 0.99 * T2\_Revenue_i ,$$

with a R-Square of 0.99. Again, it is clear that we should be aware of influential data since we observe a strong linear relationship between the two revenues after excluding them. The slope is very close to 1 and the intercept is not necessarily significant compared to the mean GST_Revenue which is over $1.5 million.

5.2 Linear Regression by Industry

The previous model works well even if it includes all industries. Nevertheless, it is advisable to repeat the regression analysis by industry in order to identify sectors which are problematic and those where the relationship between the T2 and the GST revenue is strong. Furthermore, we expect that the estimated parameters depend on the industry.

For this regression analysis, the influential data were identified among each industry defined by the NAICS3. Overall, 5,623 influential data were detected by NAICS3. This set of influential data includes 81% of the influential data previously detected with the total population.

Of the 104 NAICS3, 73 have a R-Square greater than 0.98 after the exclusion of influential data, which indicates a very strong relationship in most of the industries. In fact, only four NAICS3 have a R-Square smaller than 80%. Again, the NAICS3 industries with weaker results were those that are most likely to have GST exempted goods and services. The estimated parameters vary with the industry but the slope is between 0.95 and 1.05 for 62 NAICS3 industries.

A graphic analysis would have helped to visualise the relationship between the two revenues but with the large set of data we use, it is unlikely that we would obtain interesting results from a graph. We do have two industries, however, with a smaller set of data. Figures 1 and 2 display examples of graphs of regression models between GST_Revenue and T2_Revenue for these two industries. The black line represents GST_Revenue=T2_Revenue (Y=X) and the blue line represents the estimated model (fitted line). The green points are the influential data.

In Figure 1, the regression model on Postal Services (NAICS3=491) is very close to the line Y=X. The intercept is around -$4,000 which is very close to 0, and the slope is at 0.96. The R-Square is equal to 0.86. There are many points directly on the line Y=X but the slope of the model is different from 1 due to some points with larger differences. From this graph, the residuals seem larger when the revenue increases but there are some very large points directly on the line (not showed on this graph). We might have some concerns about the heteroscedasticity but remember that here, the main purpose of the linear regression analysis is not to fit the best model. Even without estimating a model, we can observe from the graph that both revenues are often equal.

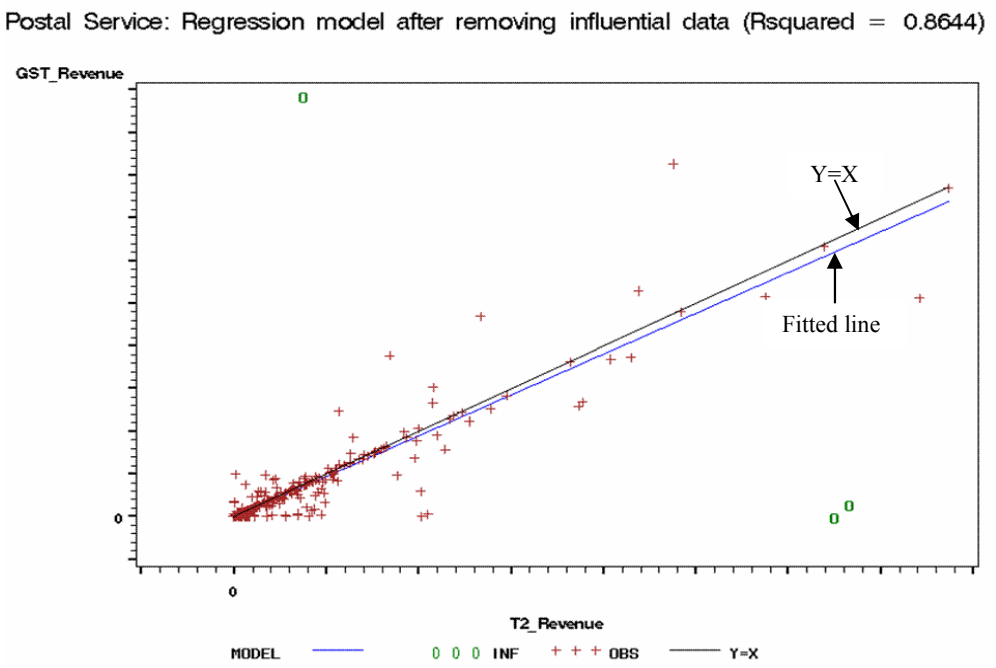

Figure 1: Postal Services: Regression model after excluding influential data (R-Square=0.86)

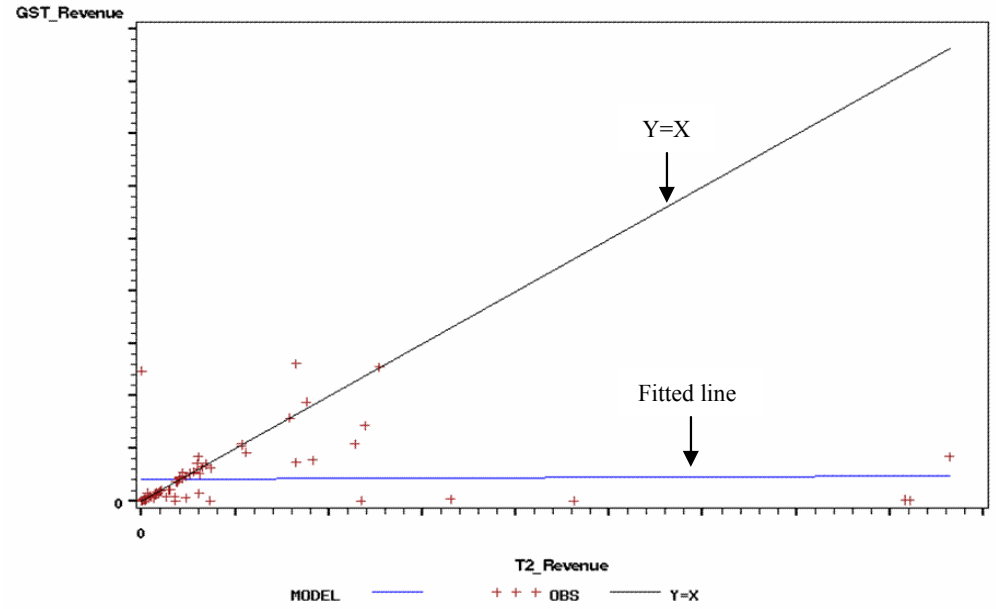

Figure 2: Hospitals: Regression model after excluding influential data (R-Square=0.05)

In Figure 2, the regression model on Hospitals (NAICS3=622) is one of the worst cases. We know however that most of the goods and services provided by this industry are GST exempted. The intercept is larger than $400,000, the slope is at 0.01 and the R-Square is equal to 0.05. Despite everything, there are many points around the line Y=X. The model is not good because of very extreme influential data (the influential data are not shown on the graph because they were too large). From the graph, we suspect some other points to be influential where most of the time, T2_Revenue is a lot higher than GST_Revenue. This is expected as for this industry, a large portion of the revenue is

coming from the sales of GST exempted goods and services, which is excluded from the GST revenue. These data were not identified as influential because of the other extreme influential data. A second run of influential data detection would probably have helped to have a better model. However, this was not our main objective.

6. Summary

The main goal of this study is to establish the relationship between business revenue from the T2 database (T2_Revenue) and from the GST database (GST_Revenue) when available from both sources and over the same periods of time. The chosen definition for T2_Revenue is total revenue. This variable seems to be the most appropriate when evaluating all industries combined, but would not always be the best if we were evaluating only specific industries.

According to the analysis, we observe T2_Revenue is often greater than GST_Revenue. This can be explained by the fact, among others things, that sales in GST exempted goods and services are included in the T2 data and excluded from the GST data. We also suspect that certain types of revenues, such as grants and investment income, included in the T2 data, are not always included in the GST revenue. It also seems that there are differences in concepts depending on the industry. For these industries, another T2 revenue variable might have been closer to the GST revenue concept. However, since these administrative data are used as survey replacement, the revenue variable should be chosen in accordance with the survey definition.

The results of the analysis are greatly impacted by a few influential data. Excluding them has helped better understanding the relationship between both revenues. Unfortunately, it is not obvious which data are influential when we just use one of the sources, which is generally the case for real production work. It might be useful to evaluate if the BN associated with the influential data stays influential over time. An inventory of influential BN could be built in order to inform users of the situation. We have also paid attention to the influential data from this comparison analysis and we have identified some of their characteristics. Most of them are very large businesses. We do not yet have evidence that they are complex structures but this is likely. Currently, the use of administrative data as survey replacement is supposed to be done for simple structure businesses only and this suggests that survey replacement for complex businesses should then be done with caution.

When GST and T2 revenues are compared by industrial sector, the difference between the two sources is relatively small except for a few industries which are most likely to sell GST exempted goods and services. A list of those industries has been established and will be updated according to the results of the comparison analysis. It is currently not recommended to use GST data for those industries.

After the exclusion of influential data, a regression analysis shows a strong correlation between T2 and GST revenues for almost all industrial sectors (NAICS3). Because of the strong relationship between the two revenues, the possibility of using the GST revenue to impute the T2 revenue is under study. Improvements were also made on the T2 side to reduce the overestimation of selected BN for which GST data and other sources showed strong signals of inactivity. This would increase the coherence between the two sources of data. The use of T2 data to impute GST data is less practical since for a given period, GST data are processed many months before T2 data. However, the use of T2 data as benchmark for selected chronic outliers from the GST side could be investigated.

Finally, it would have been desirable to measure the impact of the differences between T2 revenue and GST revenue when they are used in business surveys but this kind of analysis could be quite complex. First of all, the annual surveys often use the fiscal period as their reference year which is equivalent to the T2 fiscal period. When we try to produce annual estimates from sub-annual surveys, we add-up the months of January to December in order to get a calendar year. Furthermore, the sample is not the same for the two types of surveys and the way the administrative data are used from both sides is not necessarily the same. For example, the T2 data are often used as direct replacement when the GST data are used with a ratio model. Different aspects of the complexity of this kind of analysis are presented in Brisebois and Yung (2007). Nevertheless, this comparison analysis has helped to increase trust in the quality of administrative data used by several business surveys.

Acknowledgments

The authors would like to thank Richard Laroche and Christian Wolfe for their contribution to the project, as well as Jessica Andrews and Louis Pierre for their precious comments helping improve the quality of this paper.

References

Dubreuil, G., Girard, J., Laroche, R., Martineau, P., Rondeau, C., Wolfe, C. (2008), Comparaison du revenue T2 au revenue de la TPS. Statistics Canada Working Paper.

Andrews, J., Hamel, N., Martineau, P., Rondeau, C. (2007), Methodology for the Processing and Imputation of Corporations Data T2, Statistics Canada Working Paper.

Dubreuil, G., Hidiroglou, M.A., Pierre, L. (2003), Use of administrative data in the modelling of monthly survey data, Proceedings of the Survey Methods Section of the SSC Annual Meeting.

Brodeur, M. ,Pierre, L. (2003), Use of Tax Data: An Application of Goods and Services Tax (GST) Data, Statistics Canada Symposium.

Quenneville, B., Cholette, P., and Hidiroglou, M.A. (2003), Estimating Calendar Month Values from Data with Various Reporting Frequencies, Statistics Canada Advisory Committee on Statistical Methods.

Beaulieu, M., Quenneville, B. (2008), Calendarization of the Goods and Services Tax (GST) Data: Issues and Solutions, Proceedings of Joint Statistical Meeting.

Ladiray, D., Ramsey, L. (2003), Statistical Evaluation of the CoA-based Comparison between Tax and Survey Data, Statistics Canada Internal Document.

Brisebois, F., Yung, W. (2007), Benchmarking – Why and When should it be Performed?, Statistics Canada Advisory Committee on Statistical Methods.