AN EXAMPLE OF USING TWO-PHASE SAMPLING TECHNIQUES TO INTEGRATE DISJOINT SURVEYS AND DATA SOURCES INTO ONE SURVEY

Jack Lothian and Tracy Tabuchi, Statistics Canada, RHC Bldg, 11th floor, Ottawa, Ontario, Canada, K1A 0T6

Key Words: two-phase sample design, administrative data, survey design

1. INTRODUCTION

Current North American business practices demand tight inventory controls and a quick and efficient transportation system. When retail businesses place an order for products they expect quick responses and delivery times of less than a week are becoming common. Trucks ship from door-to-door using an extensive high quality road system that extends to every corner of the continent. They can move large or small quantities short or long distances at costs that compete with rival systems. In this environment, truck freight has achieved supremacy over its traditional rivals – air, rail, and water freight. Fifty years ago rail and water were the primary modes of transportation but today trucking is king. These changes have led to a phenomenal growth in the trucking industry and many new entrants to the industry. The growth has spawned a richer and more complex industry than existed 50 years ago.

Over the last 50 years the trucking industry has experienced significant deregulation and globalization. The border between the United States and Canada is no longer a major impediment to businesses and deregulation has led to more out-sourcing and new business models. The changing business environment has changed the character of the industry and the views of the people who study the industry.

Statistics Canada's surveys of the trucking industry have to adapt to these changes in the industry. The estimates for the output of several sectors of the trucking industry has recently displayed volatility suggesting that a review of the surveys and their basic concepts was necessary.

Section 2 outlines the actual problems encountered, while sections 3 and 4 reviews the current situation and sections 5 and 6 summarizes the recommendations for addressing the problems

2. THE SCENARIO

The North American Industry Classification System (NAICS) has two sub-dimensions for the trucking industry 1) a product dimension: General Freight; Used Household and Office Goods Moving; or Specialized Freight and 2) a distance dimension: local versus long distance. (Table 1) Additionally, the industry recognizes another dimension. The industry is further sub-divided into "owner-operated" and "for-hire" firms. The naming convention for this division can be misleading because the standard definition of these two sub-industries defines owner-operated firms as subcontractors while for-hire firms are general contractors. Thus owner-operators receive subcontracts from for-hire firms. It should be noted that a firm that neither receives nor gives subcontracts to another trucking firm is a for-hire firm. A firm that both gives and receives subcontracts is an owner-operator.

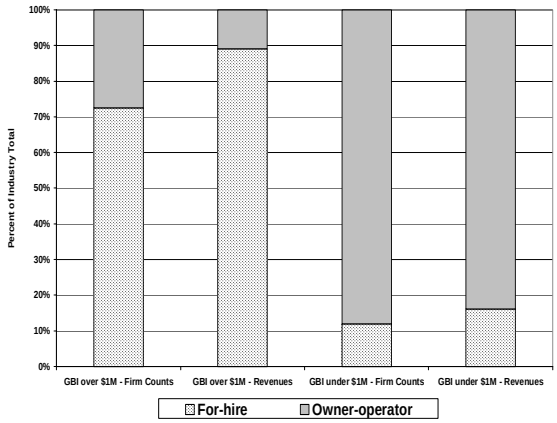
Table 1 - Trucking Industry - NAICS Classification

484	**Truck Transportation**
4841	**General Freight**
48411	Local
484110	Local
48412	Long Distance
484121	Full Truck-load
484122	Less than Truck-load
4842	**Specialized Freight**
48421	Moving
484210	Moving
48422	Local
484221	Bulk Liquids
484222	Dry Bulk Materials
484223	Forest Products
484229	Other
48423	Long Distance
484231	Bulk Liquids
484232	Dry Bulk Materials
484233	Forest Products
484239	Other

Various regulations, licenses and insurance issues within the trucking industry can impose significant costs on firms and a firm must absorb such cost when shipping inter-provincially or internationally or for shipping specialized goods. The costs are high enough that in general only larger more complex firms can afford to meet the requirements for this type of shipping. The smaller firms who wish to participate in this type of shipping must subcontract with these larger firms so they can be covered by the "for-hire" firm's licenses or insurance. Traditionally the majority of these smaller firms are one-man firms who own their own truck and this is why this category is referred to as "owner-operators". This classification system (owner-operated vs. for-hire) is widely accepted within the industry even though such a division of the industry has serious shortcomings when it is used within the context of a statistical survey. Additionally, this classification system is a requirement of the Canadian System of National Accounts (SNA) because it allows the SNA accountants to eliminate double counting of output within the industry. This double counting exists because both the owner-operator and for-hire firms count the same contract as a revenue source. Thus, the output from "owner-operators" should not be included in estimates of the Gross Domestic Product (GDP).

Creating a sampling frame including this characteristic (owner-operated vs. for-hire firms) is not possible because there is no reliable administrative source for this information and because the concept is somewhat grey in practice. Analogous to the construction industry it is possible for an entity to be either a subcontractor or general contract depending on the size or type of job. For this reason, the definition can be fluid over time and a firm can be an owner-operator in year 1 then for-hire in year 2 and return to owner-operator in year 3.
Statistics Canada's answer to this problem was to create two sub-industries using an empirical division of the universe and sampling frame. It was assumed all firms with revenues over $1 million were for-hire firms and the over-whelming majority of the firms under this threshold are owner-operators. Two separate surveys with two separate frames were created. The division is created using the gross business income (GBI) estimates produced for every business unit on Statistics Canada's business register. All units with a GBI greater or equal to $1M (except known owner-operators) were included in the Quarterly Motor Carriers of Freight (QMCF) survey frame. All units with GBI less than $1M (and known owner-operators with revenue over $1M) were included in the Annual Survey of Owner-Operators and Small For-Hire Motor Carriers of Freight (AMCF) survey frame. As can be seen in Chart 1, this empirical division appears to be a reasonable heuristic division of the universe.

Chart 1: Estimated Distribution of Firms by Firm Type - Reference Year 2002



Over the last five years the estimates from both of these two surveys have become volatile and occasionally at odds with each other. After study it was found that the volatility was attributable to short-coming in the basic assumptions underlying these two surveys. The first problem was survey jumpers who were causing serious swings in the estimates for the two surveys. Jumpers were caused by two mechanisms. The most significant mechanism was attributable to small instabilities in the estimates of GBI. From one year to the next, movements of about 20 to 40 units with GBI very close to $1M can cause significant changes in the estimates. The second mechanism is a result of how survey responses were handled. If an owner-operator is discovered from a QMCF response it is demoted to the AMCF survey and if a for-hire unit with real revenues over $1M is discovered in the AMCF survey it is promoted to the QMCF survey. These promotions and demotions have been increasing over time. Both situations can create significant changes in either of the two published series of estimates.

The second problem relates to significant shifts in the estimates of the population of for-hire and owner-operators in the AMCF survey. There are two mechanisms causing this. First, responses to questions relating to this concept can be relatively few in some strata and this implies that in some strata three or four responses could be defining the division of these two populations. Second the non-specificity of the concept is contributing to the volatility. We are seeing units, including large take-all units in the AMCF survey, switching between the two populations from one year to the next.

3. THE CONSTRAINTS

Statistics Canada has limited funds available for re-engineering the two surveys and both the QMCF and AMCF surveys were re-engineered within the last ten years and there is a reluctance to abandon all the recent work. Thus the solution to the problems requires that we maintain as much of the current survey systems as possible.

4. THE CURRENT SURVEYS

Both the QMCF and AMCF surveys could be seen as two phase surveys. In the QMCF survey the Phase II sample is the Q5 (quarter 5) supplement to QMCF and is almost identical to the Phase I sample. In the AMCF survey the Phase II sample is approximately $1/6^{th}$ the Phase I sample.

In both surveys, firm output and expense information is collected in the first phase while additional commodity, truck type, fuel consumption, and owner-operator/for-hire information is collected in the second phase. In both phases, the QMCF survey requests more detailed breakouts of the financial information plus more details by geographical regions. Phase I of the AMCF survey draws a sample from administrative income tax data but Phase I and II of the QMCF survey and Phase II of the AMCF survey are done through a mail-out with follow-up.

Table 2
Summary of Survey Populations & Sample Sizes

Survey	Frequency	Population Size	Sample Size
QMCF - Phase I	Quarterly	2,900	780
QMCF(Q5) Phase II	Annual	2,900	760
AMCF - Phase I	Annual	48,000	26,500
AMCF - Phase II	Annual	48,000	4,500

Each survey has a semi-independent frame and all survey functions such as the edit, imputation, outlier detection, sample selection, and estimation functions were independently designed in the two surveys. While there are strong common design elements between the two surveys they differ in detail in every survey sub-system and in the questions asked on the questionnaire.

5. THE RECOMMENDATIONS

The major issues are the volatility of the financial estimates from each of the two surveys plus the volatility of the estimates of the number of owner-operators in the universe. We are recommending the following changes:

1. We recommend an integration of the two surveys into one common survey with a phased in integration of all survey functions. Initially the frame should be integrated but over time the questionnaires, edit, imputation, outlier detection and estimation functions should all be integrated.

2. The two survey frames and universe should be integrated into a common frame. The integrated frame should contain all the characteristics required to run the two current surveys.

3. Instead of viewing the $1 million boundary as a division of the population into two distinct sub-populations, we propose viewing them as two strata in a common population. The current empirical boundary seems a reasonable division of the population.

4. The owner-operator/for-hire categorization should be viewed as a continuum rather than a binary variable. We should discontinue asking any questions about this variable instead we will seek information concerning the value of subcontracts received and/or given by each unit. This information should be collected in Phase II. The current binary classification could be derived from this data by specifying that an owner-operator was any unit that received subcontracts that were in excess of a specific percentage of their revenues.

5. Survey responses should no longer be used to demote and promote firms above and below the $1 million threshold. The stratification should be based solely on information derived from the BR and administrative data.

6. If data requirements show that estimates for the below $1M and above $1m are required then after collection survey data could be used to compute domain estimates for the above and below $1.

7. The preliminary size measure that is used for stratification should be defined as the maximum of: 1) GBI; 2) corporate tax revenues declared on their annual tax return if the firm is incorporated; and 3) revenue derived from the annualized Goods and Services Tax (GST) that was returned to the government by the firm, if a general sales tax (GST) return was filed. Tests suggest that this size measure will eliminate a significant number of the survey jumpers.

8. The old QMCF portion of the new integrated survey would undergo a paradigm shift and be viewed as a quarterly survey of all large and complex firms. The AMCF portion of the survey would contain small and simple firms. The old QMCF survey should be expanded to include all firms with size measure over $1M (i.e. by including owner-operator firms).

9. There should be a common questionnaire for Phase II in both the old QMCF and AMCF portions of the survey. (This questionnaire should contain questions on subcontracting.) There should be a common sampling strategy across the two portions of the survey that will minimize the overall variance for the full universe.

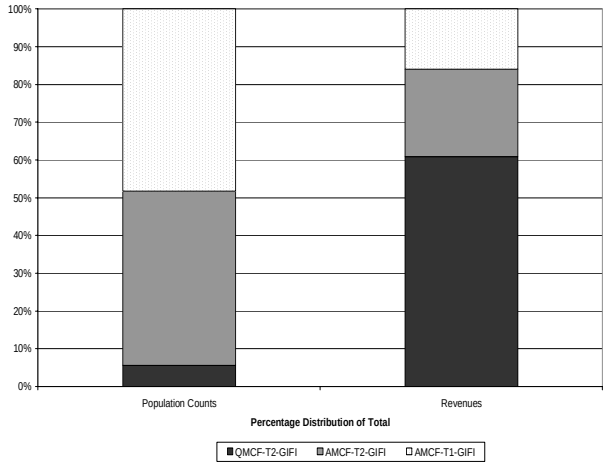
10. The AMCF estimates should be integrated with the QMCF estimates when they are published. The QMCF estimates should continue to be published quarterly.

11. Otherwise we should retain all current functions in the QMCF and AMCF surveys.

12. We should increase the use of administrative information.

Electronic reporting of annual administrative income tax data has become the norm in the Canadian business community. Statistic Canada is currently receiving from the Canada Revenue Agency (CRA) an electronic standardized balance and income statements for an annual census of incorporated for-profit universe. (T2-General Index of Financial Information or GIFI) In addition, Statistics Canada receives a significant portion of the unincorporated universe in an electronic form (T1-E-FILERS or T1-GIFI). Statistics Canada is increasingly trying to use this data to reduce the burden on respondents and decrease survey costs. As can be seen in Chart 2, the Census portion representing incorporate trucking firms represents approximately 50% of the trucking firms and 85% of the total revenue generated by Canadian trucking firms. In addition, each reference year, a representative sample of approximately 5,000 T1-GIFI trucking firms is draw by Statistics Canada. With this type of coverage, hopefully very high quality estimates of output of the industry could be obtained without any direct Phase I survey.

Chart 2 - Distribution of Population and Revenue by Administrative Data Type - Reference Year 2002



Unfortunately, the administrative tax data arrives about a year and half after the reference year ends and it lacks important information concerning breakouts on geographic location, distances traveled, and type of shipments. Additionally, this industry is a component in both the quarterly estimates and annual provincial estimates of the national accounts produced by the SNA. Thus administrative information cannot directly replace the quarterly QMCF-Phase I survey. The QMCF-Phase I survey, as redesigned, could cover nearly all the complex firms with multi-establishments or inter-provincial shipments or international shipments. The AMCF-Phase 1 survey population could consist of simple and small units and all financial information for this survey population could be derived from administrative data.

In the current QMCF-Phase II, a mini balance and income statement is collected and this could be discontinued and replaced with administrative tax data. This could eliminate about half of the current questions on QMCF-Phase II questionnaire and bring Phase II questionnaire in QMCF much closer to Phase II questionnaire in AMCF.

13. We should improve the stratification bounds.

Both the current AMCF and QMCF surveys are stratified by NAICS code and GBI size. This follows standard methodological practices at Statistics Canada. The standard reason given for this stratification is that it creates homogeneous sub-groups or sub-populations. While this is probably true for the industry or NAICS strata it is probably not correct for the size strata. These strata are created because all business surveys have a highly skewed size distribution with significant upper tails. In most business surveys two or three percent of the units contribute about 60% to the value being measured. Thus chopping the heavy tail from the distribution and declaring them to be take-all units in the sample significantly improves the quality of the estimation process. Further sub-dividing the size spectrum may further improve the situation and allow us to confidently use procedures that assume an underlying normal distribution.

Both the QMCF and AMCF survey populations exhibit size distributions that are highly skewed and thus a size stratification should continued to be performed.

6. CONCLUSION

Conceptually we should stop viewing the owner-operator/for-hire as two distinct categories. Instead, we propose viewing it as a variable to be estimated within the current domains. This paradigm shift allows us to integrate what was formally viewed as two disparate surveys into one survey with two major portions.

The current Phase I systems for both QMCF and AMCF would be retained except the QMCF survey population will be expanded to include all firms over a pre-determined threshold ($1 million if necessary) and the estimation system would produce integrated estimates of industry output that includes both QMCF and AMCF portions of the survey.

Phase II would undergo significant revisions so that one common questionnaire would be used for both the QMCF and AMCF portions of the survey. The sampling procedure would be integrated across both portions of the survey and optimized to ensure minimum variance for full universe estimates.

Current testing suggests the proposed changes will eliminate the volatility in the revenue estimates for the QMCF and AMCF populations. The proposed changes in way owner-operator/for-hire information is gathered will allow us to produce more stable and meaningful estimates of subcontracting in the industry.

The integration of the survey frames and Phase II questionnaires would lead to some efficiency gains and an overall decrease in the Phase II variance. The replacement of direct survey information with administrative data in Phase II of QMCF would lead to further efficiency gains.

7. ACKNOWLEDGEMENTS

The authors would like to thank Julie Trépanier for her valuable contribution to this project. The author would also like to thank Marietta Morry and François Gagnon for their comments that helped to improve the quality of the paper.

8. REFERENCES

Lavallée, P., and Hidiroglou, M. A. (1988), On the Stratification of Skewed Populations. Survey Methodology, 14, 33-43.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. Survey Methodology, 18, 209-217.

Statistics Canada (2002), Trucking in Canada - 2002, Catalogue no. 53-222-XIB.

Trucking Traffic Survey Development Project Team (2001), Data Needs and Requirements Document: Findings of the Consultation with Major Stakeholders. Unpublished report, Statistics Canada.

Matthews, Steve (2001) "Comparing the Tax Universe and the BR Universe for a Population of Small Businesses: Annual Motor Carriers of Freight Survey", Unpublished report, Statistics Canada.