Trades Division
Statistics Canada

Report on the methodology used for converting the Quarterly Retail Commodity Survey (QRCS) historical time series from SIC to NAICS

Creation Date : August, 23 2005
Created By : Norman Fyfe

Last Updated : October 7, 2005
Updated By : Norman Fyfe

Date Printed : October 7, 2005

1. Introduction

The signing of the North American Free Trade Agreement (NAFTA) in January 1994 created a need for a common economic classification system between Canada, the United States and Mexico. The creation in 1997 of the North American Industry Classification System (NAICS) by the statistical agencies of the three countries was designed to fill this need. The NAICS is built on a production oriented or supply conceptual framework in that establishments are grouped into industries according to similarity in production processes used to produce goods and services. The new system allows a better comparison of industrial statistics among the three countries[1].

The Quarterly Retail Commodity Survey (QRCS) has been conducted by Statistics Canada since the first quarter of 1997. It breaks down retail sales into more than 100 commodity groups. It is a sub-sample of – and is benchmarked to – the Monthly Retail Trade Survey (MRTS), a parallel survey that measures sales by store type. The redesign and subsequent adoption of the 2002 version of the NAICS in 2004 by the MRTS led to the redesign of the QRCS and its adoption of the NAICS, as its sample consists of a subset of retailers in the MRTS. Since 1948, Statistics Canada had primarily used the Canadian version of the Standard Industrial Classification system (SIC) and its various revisions (in 1960, 1970 and 1980). QRCS had been using the 1980 version of the SIC since the survey's inception in January 1997. The passage to NAICS in 2004 meant that the QRCS' new NAICS time series effectively began in the first quarter of 2004 and had no prior history.

This document presents an outline of the methodology that was used to convert the QRCS historical time series from SIC to NAICS. In particular, it describes: how new NAICS commodity-by-store type data series were generated from historical SIC-based survey data; how these series were adjusted to the new NAICS-based QRCS survey and how the adjusted backcasted series were benchmarked to ensure consistency with the MRTS.

2. Methodology used to convert QRCS historical series to NAICS

As part of its efforts to convert its historical SIC based time series to NAICS, the MRTS created dual-coded micro files of its in-sample establishments. This was possible as statistical establishments in Statistics Canada's Business Register (BR) have been classified since 1998 under both SIC and NAICS industrial classification systems. The BR is the frame for the MRTS and for most of Statistics Canada's business surveys. This double classification made it possible to design a methodology for the MRTS to obtain domain estimates according to the NAICS even though the survey frame was SIC based.

[1] Naics Canada 1997, Statistics Canada, Catalogue No.12-501-XPE

The dual-coded files were made up of records containing the establishment's unique identifier, its sample weight, total sales and NAICS and SIC codes. For establishments mapping to more than one NAICS code (referred to as "splits") the record contained a flag and the percentage of the establishment's total sales going to each NAICS code.

As the QRCS is a sub-sample of the MRTS, the same dual-coded files which were used for the backcasting of the MRTS were used to assign a NAICS code to each record in its old SIC based sample. Under this approach, every unit in the sample was individually re-classified to NAICS.

QRCS NAICS-based commodity estimates by trade group[2] were then obtained by summing the weighted recoded commodity data by NAICS domain and by aggregating the NAICS domain estimates by trade group.

2.1 QRCS historical conversion: quality issues and sources of error

Caution should be exercised when analysing and interpreting the converted historical data series. This section presents a number of issues to be considered with respect to data quality.

2.1.1- Coding errors

Various sources of error impact the effectiveness of the micro approach that was used for converting the SIC based historical estimates to the NAICS classification. An initial potential source of error is the survey frame itself. A classification error on the Business Register under the SIC would yield an incorrect corresponding NAICS code. Moreover, as dual classification was introduced in 1998, the early years of the backcasting are subject to a higher rate of coding errors than the later years, where most classification issues had been resolved. In order to reduce the impact of units that were incorrectly classified, large contributors to the estimates were verified manually and recoded where necessary.

As a general rule, if an establishment's SIC code on the QRCS micro file was not the same as the one on the dual-coded MRTS file, then the QRCS' SIC code was taken as the default value, as it reflected the latest sampling information. One of the reasons behind the classification discrepancy comes from the fact that QRCS micro files are revised annually and contain the most up-to-date information, whereas the dual-coded files reflect the most recent information available on the survey frame at the time of the dual-coded files creation.

[2] Trade groups are special aggregations of the North American Industry Classification System (NAICS) industries. There are 19 trade groups which are further aggregated to 8 trade group sectors.

2.1.2 – Dealing with establishments with multiple SIC-to-NAICS relationships

As in any concordance, a certain number of "splits" were produced. Splits are SIC- based establishments that do not have a one-to-one relationship with a NAICS code. Most of the larger splits were, after research, reclassified to a simple dominant NAICS, reflecting current reporting arrangements to the QRCS. Larger splits are defined as units accounting for at least 1% of a TG total. In general, there were about a dozen of these large units in any given month. They accounted for about half of the estimates attributable to splits. The number of these large units was fairly constant throughout the backcasting period.

The remaining splits were too small to have an impact on the estimates (or were all part of the same TG) and were ignored. These smaller "splits" accounted for only approximately 2% of the overall sales estimates.

The trade group containing the largest number of splits was clothing stores and a large share of the estimates (45% in December 2003) in clothing stores were attributable to splits. However as the majority of the splits happened within the clothing stores TG, the effects of the splits on the estimates were cancelled out at the TG level. For all other trade groups, splits accounted for less than a quarter of a percent of the trade group total estimates in 2003.

2.1.3- How to account for establishments moving from the wholesale sector to the retail sector?

With the adoption by the QRCS of the 2002 version of NAICS, some establishments which were previously classified as wholesalers under the SIC 80 are now classified as retailers or "**new entrants**". Examples of such establishments are computer stores, home centres, building material dealers and office supplies and stationery stores.

Overall, in retail trade, according to a study conducted by the MRTS for the 1998-2001 period, 96.3% of SIC80 retail sales remained within the retail sector under the NAICS classification, with 3.7% transferred to the manufacturing and services sectors. Similarly, 94.3% of wholesale sales within the SIC80 classification remained in the wholesale sector under NAICS, with 4.2% transferred to retail and 1.5% moving to other sectors.

To compensate for the fact that the QRCS cannot provide information in the NAICS domain estimates for the new entrants, 2004 QRCS data was used to re-create commodity breakdowns and seasonal patterns for each backcast period for those establishments that overlapped[3] the old wholesale SIC and the current retail NAICS samples.

[3] Overlapping units are wholesale records that were BOTH in the Monthly Wholesale Trade Survey (MWTS) SIC survey and in the new NAICS MRTS survey

Data from selected establishments only were retained. They included those establishments with a NAICS code associated with a high proportion of new entrants: automotive parts and accessories stores, tire dealers, computer and software stores, home centres, paint and wallpaper stores, building material dealers, outdoor power equipment stores, nursery and garden centres and office supplies and stationery stores. Together, these 9 NAICS code groupings accounted for the majority of the new entrants. Computer and Software stores and Building and Outdoor Home Supplies stores were most affected, as almost 100% and 79% of the estimates in these respective trade groups came from new entrants in 2003.

The percentage of QRCS estimates attributable to new entrants varied during the backcasting period, but generally remained high for the 9 NAICS codes mentioned above. Domain estimates of other NAICS codes affected by new entrants and which were not retained for the backcasting exercise were adjusted during the benchmarking process to MRTS. This was the best option available as there did not exist reliable commodity breakdowns of sales for the new entrants over the backcasting period. The disadvantage of this method was the application of 2004 monthly commodity distributions to previous years.

“Non-overlapping units” had their commodity distribution imputed using a 2004 twelve month average of collection records taken from establishments in the same NAICS. The impact of these units on the backcasted series was negligible as they contributed very little to the estimates, given their relatively small size. Both overlapping and non-overlapping units that had been selected from the Monthly Wholesale Trade Survey (MWTS) sample were then added to the QRCS micro files.

3. How were the historical backcasted series adjusted to the current survey levels?

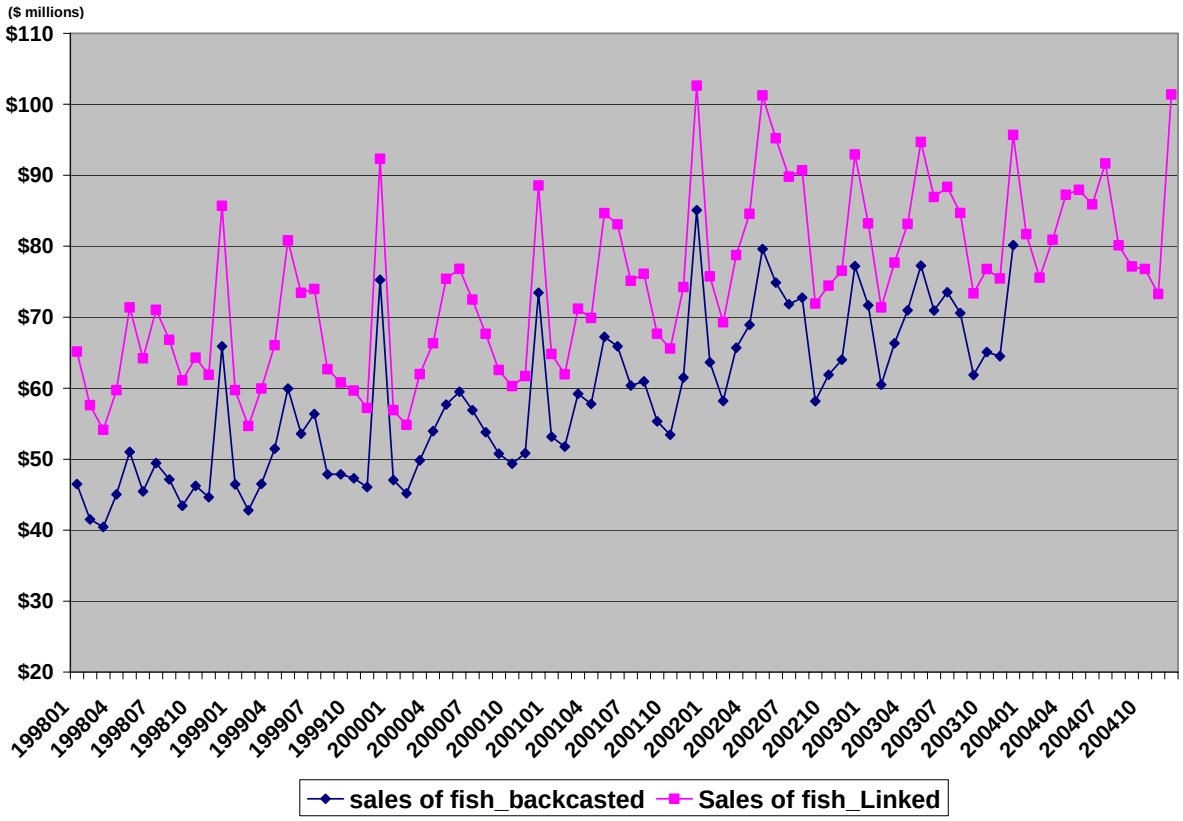
A graphical analysis of the backcasted series revealed important level shifts for a number of commodities between the last backcasted period and the first NAICS sample period. The level shifts can be explained by a number of reasons, including sample changes, classification changes and other methodological changes. The methodology used to correct these level shifts is explained in detail in this section.

The last month of collection on a SIC basis was November 2003. In order to have a complete, publishable set of SIC based estimates for the last quarter of 2003 and to partially assess the impact of the new sample on the QRCS estimates, a set of SIC-based estimates was derived from the NAICS sample for December 2003 (the first month of collection under NAICS) and the first quarter of 2004. This set of estimates served as a “parallel run” for the QRCS that was used to estimate the impact on estimation caused by the introduction of the new sample.

These estimates were generated by first copying data from records identified as "overlappers" in the NAICS sample into corresponding records of the SIC sample. For "non-overlappers" the process was more complex. An establishment's total sales figure was obtained from the MRTS parallel collection (that is to say that data collected by the MRTS under the new NAICS sample were for a period of 5 months coded according to both the SIC and the NAICS) and its commodity distribution was derived through imputation using the most recent/appropriate historical data for the establishment. The resulting SIC based commodity estimates were converted to NAICS by recoding the micro data, using a similar process as for the backcasting of the QRCS historical time series.

The "parallel run" and the current NAICS data were then used to adjust the historical backcasted series to the published levels of the new survey. The historical backcasted time series were adjusted to the levels of the new NAICS based survey with the help of time constant, multiplicative ratios. The ratios were calculated by dividing the *current NAICS estimates* by *the corresponding backcasted* estimates for each commodity/trade group combination. Each backcasted time series was then multiplied by its corresponding, time constant, adjustment ratio. Figure 1 below provides a graphical illustration of how the levels of a backcasted commodity series were adjusted to the new 2004 NAICS levels using the adjustment ratios described above.

Figure 1: Sales of fresh fish and other seafood in Canada: comparison of linked and un-linked data.



It was decided to adjust the backcasted series to the March 2004 current survey levels as very seasonal commodities such as lawn and garden products were better adjusted using March ratios.

The level adjustment process stripped the system of its “additivity” which resulted in discrepancies between trade group totals and the sum of commodities within the trade groups. To restore additivity, the series were subjected to a reconciliation process which prorated the difference between a commodity aggregate and its components. To illustrate this process, let us look at sales of books, newspapers and periodicals. Books, newspapers and periodicals represent one commodity grouping for which sales estimates are computed. However these sales estimates can be further broken into 2 commodities: (1) sales of books and (2) sales of newspapers and periodicals. The adjusted estimate for the individual commodities should be equal to the adjusted estimate of the commodity grouping. However, discrepancies arise, as the adjustment process described earlier produces separate adjustment factors for the grouping and the component commodities.

Given the higher level of confidence in the more aggregate estimate (better reporting, lower imputation rate, etc.), a decision was taken to preserve the levels at the commodity grouping level. Therefore, a ratio of the commodity grouping total to the sum of the 2 commodities was applied to each commodity. The overall total was obtained indirectly by summing commodities across trade groups.

After the reconciliation process was completed the series were subjected to a benchmarking process. QRCS backcasted total sales were benchmarked at the MRTS trade group level. An exception to this is the Department Store trade group (QRCS includes the sales of department store concessions while MRTS does not).

4. Outlier and level shift corrections in the backcasted series

It would not have been feasible due to time and resource constraints to analyse all 2,736 QRCS backcasted time series (144 commodities within 19 trade groups) for anomalies. Therefore, a list of the most important commodity time series by trade group was established. The list was made up of time series which accounted for at least 1% of the total sales estimates in a given trade group. In most cases, the largest 10 commodities accounted for the vast majority of the estimates in a given TG. The final tally of backcasted series requiring an outlier and level shift detection methodology was approximately 225.

Those main series were subjected to a thorough outlier and level shift detection testing, which was conducted with the help of X-12-ARIMA computer program. This sophisticated statistical program allowed for the detection of anomalies in the backcasted data while accounting for trading days, moving holidays and seasonal pattern variations. There are many reasons for outliers and level shifts in time series and they can make interpretation of the data over time difficult. A level shift involves all observations before a certain point. The observations jump from one level to another level and stay at the new level. An additive outlier, on the other hand, involves only one observation. There is an unexpectedly large or small value.

In the case of retail economic series, outliers and level shifts may be caused by restructuring at a sector (e.g. chain opening or closing) or because a large company may move from one industry to another. Unusual events, such as the SARS outbreak, the mad-cow disease or a major blackout may also cause unusual values. Sample changes in the form of “births” or “deaths” may also introduce level shifts. In other cases, there may be survey non-sampling errors, such as coverage or measure issues. These latter errors are the ones that needed to be corrected in order to prevent the distortion of the upcoming seasonal adjustment of the QRCS historical time series. All other outliers or level shifts caused by unusual events or economic restructuring are preserved to reflect the economic reality but are taken into consideration in the calculation of the time series seasonal factors.

Whenever possible, unjustified discrepancies found in the main series were removed by correcting errors or improving statistical imputations contained in the backcasted micro data using prior knowledge of the data. When a micro adjustment was not feasible, macro type adjustments were used to correct

errors in the main backcasted data. The remaining series were not adjusted for such atypical values.