WP. 16
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iii): Macro editing methods

The Common Editing Strategy and the Data Processing of Business Statistics Surveys

Invited Paper

Prepared by Etienne Saint-Pierre and Mario Bricault, Statistics Canada

I. Introduction

1. In 2010, Statistics Canada launched the Corporate Business Architecture initiative which entailed a review of Statistics Canada’s business methods and systems to achieve efficiencies, enhance quality assurance and improve responsiveness in the delivery of new statistical programs. The development and mandatory use of shared and generic corporate services and systems for collecting, processing, disseminating and storing statistical information for business and household programs is a key component of this initiative. To meet the objectives of the Corporate Business Architecture, Statistics Canada is currently undertaking a major integration project for its Business Statistics surveys, the Integrated Business Statistics Program (IBSP). The IBSP will provide a common survey framework for the various business surveys conducted at Statistics Canada. By 2016, nearly 120 surveys in ten different programs will be integrated into this new harmonized framework. The surveys under the umbrella of the IBSP will use Statistics Canada’s Business Register as a common frame. They will adopt common sampling, collection and processing methodologies driven by a common metadata framework and they will share common tools to analyse, edit and correct data.

2. Two new initiatives are currently developed under the IBSP and will be discussed in this paper. The first initiative is the implementation of an iterative processing model called Rolling Estimates, where estimates are produced and analysed on a regular basis in a cycle until an acceptable level of quality is reached. The second one is the development of a Common Editing Strategy (CES) for most business surveys. Both initiatives are central to the IBSP project. The objectives of the CES are to contribute to the reduction of operational costs via the harmonization of the editing methods and tools, the expansion of automation in the editing activities and also the reduction of respondent follow-up activities and manual interventions on non-influential units. It is also expected that this common editing framework and the new processing approach will be beneficial to data quality with improved timeliness of survey results. The accuracy and the coherence of statistical information are other elements of the data quality that should be positively affected by re-directing subject matter editing efforts toward analytical activities.

3. This initiative is also in line with the implementation in several statistical agencies of editing processes centred on the concept of selective data editing in an integrated framework. In fact, the development of the CES was guided in part by the Annual Integrated Collection (AIC) core editing strategy of the Australian Bureau of Statistics – see Brinkley, McDonald & Bismire (2007) and Mackay (2007).

4. In the second section of this paper, the new iterative processing model adopted for the IBSP - the Rolling Estimates - is presented. The iterations will allow the continuous realignment of the micro and macro-editing strategies based on the most recent available data. In the subsequent sections, the components of the CES are described. In section 3, the new editing rules framework under the CES is presented. An overview of the quality indicators and their use in the CES are discussed in section 4. The links between the CES and the active management of the collection activities as well as with the active management of the analysis are described in Section 5 and 6 respectively. The plan to simulate the Rolling Estimates approach and the CES with real data is presented in the last section of the article.

II. A new iterative data processing model: the Rolling Estimates

5. Recent studies on the volume and impact of manual interventions on the Unified Enterprise Survey (UES)[1] and interviews with analysts from all programs under the umbrella of the IBSP on their data editing practices, emphasize the need for rationalization and harmonization of these activities.

6. Chepita (2006) measures the prevalence of manual imputations on at least one of the two key variables (Total Operating Revenues or Total Operating Expenses) in eight key surveys from Services Industries. The conclusion was that the implementation of pre-programmed adjustment formulas and changes to the edit and imputation processor could have reduced by almost 70 per cent the number of questionnaires manually edited. Yeung (2007) demonstrated that about one third of the records in five UES surveys (Food Services, Primary Metals, Retail Chains, Retail Stores and Wholesale – for RY 2004) were manually adjusted either before or after the automated edit and imputation process. Nearly 13 per cent of the records adjusted manually were changed more than once. Naud (2009) looked at item response rates for 13 important financial variables for all the UES surveys. The high level of manual imputation was demonstrated. Sometimes a high percentage of units were manually imputed but these records only represented a small portion of the estimates. Saint-Pierre (2010) and Bricault (2010) consulted with subject matter areas to get a good understanding of the reasons behind manual interventions. Both concluded that a large portion of systematic manual interventions could be automated in the edit and imputation processor. They also came to the conclusion that a lot of manual interventions were done before edit and imputation and could be avoided with a better understanding of the systems, procedures and methodologies behind this process.

7. As part of a preliminary investigation for the CES, Ratime (2009) showed that each survey area had developed its own set of tools and practices for performing data adjustments. The harmonization of systems, processes and practices was recommended. Cloutier (2009) presented the principles behind the CES in relation to the current UES process model at Statistics Canada. With the integration of new

[1] Currently 58 surveys covering the Manufacturing industries, the Services industries and the Distributive Trade industries are already part of an integrated survey framework inside the so-called Unified Enterprise Survey.
The 58 surveys already use common tools and systems for E&I and for data analysis and micro-editing, but their editing strategies are relatively heterogeneous.

business programs[2] into the IBSP framework, a revamped processing model driven by an efficient editing strategy will be put in place.

8. In the standard linear survey process currently used in the UES, collection activities (including follow-ups for failed collection edits) are performed first, followed by processing tasks, such as edit and imputation and finally the generation of estimates. After each of these steps, data validation and manual editing is performed by the subject matter analysts. In a non-selective editing environment, the same record can be looked at and modified several times without a clear knowledge of its influence on the results. The IBSP wants to optimize the editing work done by both collection services (follow-ups for failed collection edits) and subject matter divisions via the CES. Indicators and decision rules will be provided to better target units that need a follow-up and assess the relevance of manually editing a record (see Section IV). The efforts devoted to editing are expected to decline considerably.

9. The CES is part of a new iterative data processing approach called Rolling Estimates being developed for the IBSP. Under this iterative approach, collection, processing and micro/macro editing activities will be done in parallel and not in a standard sequential way. Iterative estimates will be produced in the cycle for each domain of estimation as soon as an acceptable level of survey and administrative data will be available. Systematic and regular integration of tax data for financial variables of non-complex enterprises for planned tax replacement or imputation purposes is also part of the new approach.

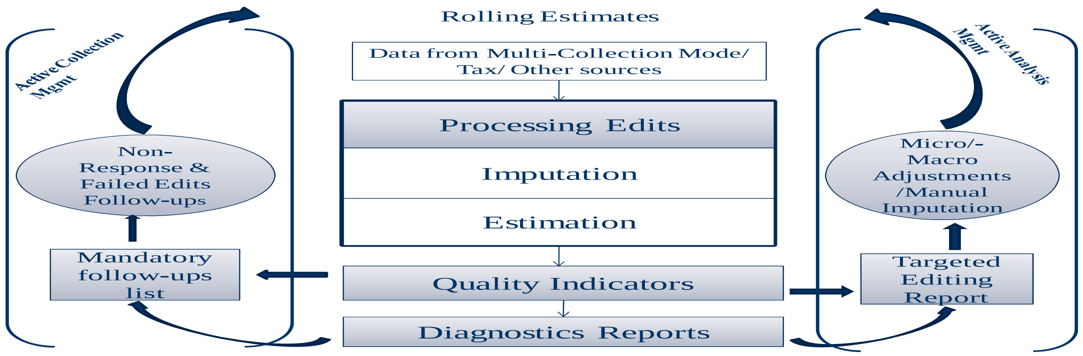
10. No manual intervention will be possible either before or after edit and imputation during an iteration as it is the case in the current UES processing framework. The edits, imputation (or re-weighting), allocation and estimation functions are performed on the data without any manual intervention. Data is reviewed only after the estimates are available in the analytical tool. The estimates will be assessed based on macro quality indicators after each iteration. In addition, a score will also be derived for each sampled units to measure their impact on 1) the estimates, and 2) its quality.

11. Based on both the macro quality indicators and the micro record impact score, data editing resources in the Collection Services and subject-matter areas could be assigned to units with a high impact in their domains of estimation. On one side, Collection Services will conduct follow-ups based on a priority list of units based on the current survey results (Active Collection Management). On the other side, subject matter areas will be able to review influential records in relation to the estimates and their quality (Active Analysis Management). Figure 1 illustrates the workflow of the Rolling Estimates, the two main components of the data editing (Active Collection and Active Analysis) and highlights (in gray) functions affected by the CES.

12. In subsequent iterations, corrections and manual adjustments on influential records are integrated along with the new survey and administrative data available to produce updated estimates with adjusted indicators. Another set of influential records is identified and re-prioritized for failed edit follow-ups and manual interventions. Estimates are run periodically until the quality indicators reach their established target for all estimation domains or until the end-of-survey date is met.

[2] In addition to the current UES programs, the Energy, Transportation, Finance, Research and Development, Capital Expenditures and Agriculture sectors will be part of the IBSP.

Figure 1 – Rolling Estimates and the Common Editing Strategy.


Rolling Estimates
Data from Multi-Collection Mode/ Tax/ Other sources
Processing Edits
Imputation
Estimation
Quality Indicators
Diagnostics Reports
Active Collection Mgmt
Non-Response & Failed Edits Follow-ups
Mandatory follow-ups list
Active Analysis Mgmt
Micro/-Macro Adjustments /Manual Imputation
Targeted Editing Report

13. This model provides at least four advantages:

(a) On the collection side, the failed edit follow-up and the non-response follow-up prioritization will not only be based on past information available prior to collection (e.g. revenues of the units on the Business Register) as it is now the case in the UES. The priorities will be re-adjusted based on the most current survey results and the evolution of the quality of the information.

(b) On the analytical side, estimates and quality indicators will be available periodically in the initial stages providing analysts a first-hand picture of the estimates thus allowing them to monitor the evolution of quality. Analysts can also prioritize and focus their manual editing.

(c) This model should also significantly improve the timeliness since, at the end of the collection process, estimates would have already been reviewed several times and influential records validated and edited. The estimates should then be almost final and ready for dissemination. This model is developed with the objective of disseminating results of the annual business surveys earlier than the current 15-month target. This may be achieved without reducing the collection, processing or analysis time since the three functions would be conducted simultaneously.

(d) Cost reduction is the main objective of this model. It is hoped to achieve this through the reduction of manual interventions and the optimization of follow-up activities from the CES.

14. The CES includes three major components:

(a) A new framework for editing rules to maximize the automated corrections of erroneous and inconsistent data and to reduce the delay between the reception of data and their use in the post-collection operations.

(b) The Active Collection Management to direct, prioritize and adapt non-response and failed edit follow-up efforts of the collection group based on the most recent available results.

(c) The Active Analysis Management to provide a standard framework, measures and tools to optimize, prioritize and adapt the analysis and micro-level editing performed by the subject matter analysts on influential records based on the most recent available results.

III. Edits rules framework in the CES

15. In the current UES model, resolution of failed collection edits is an element causing significant delays in the post collection process. Numerous attempts to establish contact with respondents, the complexity of the information required and the important volume of so-called "critical" edits to resolve are elements delaying the availability of the data for the post-collection phase in the traditional linear model. In 2009, the average delay between reception of the data and their availability to the post collection phase was 52 calendar days for the UES surveys. In addition to the time to capture the information, the positive resolution of edits contributes to this delay. The delay between the reception of the data and the availability of data for processing should be kept to a minimum to produce estimates and improve timeliness. In order to minimize these delays, the CES standardizes and redefines the model used for collection edits.

16. First, the emphasis should be put on replacing missing or inconsistent values with non-missing and consistent values using the data available within the records and the relationships between the data values as specified in the related metadata. The verification of the sum of parts, equivalencies, zero filling or systematic relationships between variables should be done via the deterministic edit rules and systematic correction methods for a given error should be automated as much as possible via deterministic imputation.

17. A large number of systematic manual interventions done by subject matter analysts to resolve equivalencies issues or logical relationships could be dealt with automatically should the metadata be set-up accordingly. In the current model, the subject matter specialists of the UES review the data transmitted from collection prior to imputation. A lot of the corrections that would have been automatically applied by the deterministic imputation in the edit and imputation stage are manually done by subject-matter areas due to the misunderstanding of the functionalities of the pre-processor. In the Rolling Estimates model, the analysts will look at the data only after the estimates are produced and after the deterministic imputation, thus eliminating unnecessary manual interventions.

18. Second, the main purpose of the edit rules should be to validate the data of key financial and commodity variables as defined by subject-matter areas and the Canadian System of National Accounts. The volume of edit rules should be managed carefully to minimize the validation of non-critical edits and to avoid imposing a non-essential response burden. Non-critical edits will be followed-up only if a unit is prioritized for the resolution of critical edits.

19. Finally, the majority of the edits will be applied during the processing of the data and not at the collection stage. The primary mode of collection is expected to be via electronic questionnaires. Only a few key edits will be embedded in the e-questionnaire to inform respondents of potential errors and inconsistencies. The number of edits will be kept to a minimum not to burden respondents. Furthermore, failed edits will not prevent the respondent from completing and sending the questionnaire. Data will then be available for processing without any delay and all the edits will be checked in the processing phase. For mail-out mail-back questionnaires, data will be available for processing immediately after being captured. Edits will be applied and assessed in the processing stage and only influential units in domains requiring improved quality will be selected for follow-ups. For telephone surveys, it would be more appropriate to apply and resolve edits interactively with respondents as the data is reported. Under the IBSP, the volume of collection edits will vary according to the collection mode.

IV. Quality indicators in the CES

20. In order to realign the strategies for both the Active Collection and Active Analysis, it is essential to have 1) quality indicators by domain of estimation (QI) to measure the level and progression of the quality of the estimates and 2) a score to measure the impact (MI) of each sampled unit in its domain of estimation in order to establish priorities.

21. QIs are traditionally used as direct measures of data quality and are often calculated and integrated at the very end of the process (analysis and dissemination) to help users in data interpretation. A score function for each micro-record is also currently in use in the UES survey to prioritize both non-response and failed edits follow-ups. As described in Pursey (2003), the score function is established based on the revenue available on the Business Register for the non-manufacturing units. When the targeted coverage rate is reached for a domain, follow-ups are stopped for the domain and resources reassigned where coverage rates are lower. This method helps prioritize work and reduces follow-up activities; however, it is based on a single static variable – the revenue from the Business Register. Enhanced score functions will be developed under the IBSP. It will take into account the significance of the units and their impact on the estimates and on the quality of the estimation for the domain. Results from the current cycle and their quality will be used to dynamically adjust the MI scores. Furthermore, the significance of the units could be established based on variables that are unique to a survey.

22. Under the new model, in combination with their traditional role, the QIs will serve two additional purposes. First they will help to trigger the signal to end active collection or deem the estimates final when the pre-defined targets are reached. Target quality levels will be established for each estimation domain prior to the collection and processing cycles. Second, the QI will help to allocate and prioritize the collection and analysis activities. Resources will be reassigned after each run of Rolling Estimates in domains where the quality and the representativeness of the survey population needs to be improved.

23. The potential QIs will be output-oriented and include various types of rates (response, coverage, reported or imputation rates, percentage of records with data for key variables). The estimated variance is also another indicator to be considered since it is directly related to the estimates. Few quality measures could be calculated based on one or two variables[3]. The calculation of a global QI by domain of estimation is essential to derive efficient prioritization rules.

24. To properly prioritize the units to follow-up inside each domain of estimation, a score function is needed. MI scores will be calculated based on a few relevant variables unique to each survey. A global MI score will be derived for every unit in each domain of the survey. This measure will give the impact of both reported and non-response units on the value and the quality of the estimates. Units above a pre-defined threshold will be considered as high impact units. The methodological details related to both the QIs and MI scores are discussed in Godbout, Beaucage and Turmelle (2011).

25. While the methodology to calculate the QIs and the MI scores will be similar across all surveys, the key variables and the quality targets will not be the same. The key variables used in the calculations of QI, MI score and in the analytical reports should be defined for each survey by the subject matter specialists in agreement with the primary user of the business survey, the Canadian System of National Accounts. The selection of key variables is a central element of the CES.

[3] An important element that Statistics Canada wants to improve in its Business Statistics Program is the representativeness of survey response. This element could be integrated in the QI via the R-indicator. See Schouten, Cobben and Bethlehem (2009).

V. Active Collection Management

26. Based on the QI and MI score and the underlying decision rules, each record will be assigned a path after an iteration of Rolling Estimates. If the QI for a domain reaches the target quality, the active collection will be closed for this domain and no more follow-ups will be performed whether the units are influential or not. If the QI is under the target level, only units identified as influential will be followed-up. Amongst the influential units flagged for follow-up, the priority will be assigned based on the MI score. The non-respondent unit with the highest MI score (above a pre-defined threshold) in a domain of estimation with a QI under the target level will be assigned the highest priority for the non-response follow-up. Units with failed edits on key variables, with a high MI score and in a domain with a low QI will be the primary focus of the failed edits follow-up. Outcomes from the follow-ups will be integrated into the next run of Rolling Estimates. Non-response of non-influential records (low MI score) will not require a follow-up (except if required to measure the non-response bias) but will automatically be resolved via imputation or re-weighting. Failed edits for non-influential records would most likely be overridden and the original data provided by the respondent kept intact (if coherent) since the impact on the estimates will be negligible. These records will be identified as “fit for use”. Figure 2 summarizes the various possible actions based on the combinations of QI and MI score.

Figure 2 –The ACM is driven by QI and MI.

		Active Collection - Actions	
		QI > Target Level	QI < Target Level
MI > Threshold (influential records)	Response - No Failed Edits	No Follow-up	No Follow-up
	Response - Failed Edits	No Follow-up	Follow up
	Non-response	No Follow-up	Follow up
MI< Threshold (Non-influential records)	Response - No Failed Edits	No Follow-up	No Follow-up
	Response - Failed Edits	No Follow-up	No Follow-up
	Non-response	No Follow-up	No Follow-up[4]

27. If non-response follow-ups on influential records are not successful enough to obtain the target QI, some non-influential records units will have be promoted as influential units in subsequent iterations and then followed-up. Once all domains in a survey reach their quality target, the active collection can be closed for this survey.

VI. Active Analysis Management

28. Under the IBSP, the analysis will remain under the responsibility of subject matter areas. The wide variety of subjects and industries requires a high level of expertise to interpret, analyse and proceed with macro/micro editing of the data. The integration of the analytical tasks under the IBSP will be done through the provision of an analytical framework consisting of a common workflow, measurements tools, systems and the access to various data sources in order to perform the analysis.

29. Standardised Macro-Analytical/Editing reports identifying which dominant domains to review and domains with potential anomalies in the estimates will be produced after each iteration. These common reports could be customized by adjusting the parameters (variable, value range, ratios). QIs will be included with the estimates to help analysts interpret results and prioritize their macro-editing efforts.

[4] Follow-ups could be performed to measure the non-response bias.

30. For domains of estimation that needs to be investigated, a list of high impact records based on the MI score will be available. Records with a MI score above the pre-defined threshold will be flagged for subject matter review since they impact the results. Other records could be flagged for review through pre-defined standard diagnostics reports (outlier ratios, historical comparison, top contributor, etc.) that subject matter analysts could customize by specifying parameters (variable, value ranges, ratios) via metadata tables. The combination of the information from the different reports will help subject matter to efficiently target their micro-editing. Manual interventions of subject matter analysts will be taken into account in the following iteration of the Rolling Estimates.

31. In order to facilitate understanding of the survey results and improve the detection of data anomalies during the editing and analysis phases, it is proposed that subject matter areas devote more time and effort to improve their knowledge of the data, their industries, of the major contributors and of economic and non-economic factors that could affect the results. The objective is to develop solid expectations[5] of survey results prior to reviewing the data. In the context of the Rolling Estimates approach, analysts could rapidly assess, after each iteration, if the results converge or not toward the initial expectations. This knowledge is also essential to obtain feedback from analysts on the performance of the QI, MI score and the underlying prioritization rules to properly direct follow-up activities, macro and micro-editing towards the maximisation of quality under work resources constraints.

VII. Future Work

32. The integration of 120 surveys into a single processing framework is a significant challenge. Tarassoff (2010) identified key issues in regards to the introduction of the Rolling Estimates and the CES. In order to address these issues and to develop and test an efficient strategy, 4 iterations of Rolling Estimates will be produced for the 58 UES surveys in 2011, in parallel to the current linear approach. Based on the results of these simulations, the principles, methodologies and parameters of both the Rolling Estimates and the CES will be validated, tested and adjusted if needed:

(a) Since no manual intervention will be allowed during an iteration, a complete set of data should be available at the end of the imputation process. New edit and imputation strategies combined with re-weighting will be put in place and tested.

(b) After each iteration, basic QIs (weighted response rates and coefficients of variation at the domain level will be used for the simulation) and MI scores will also be calculated for the 58 UES surveys based on one or two key variables. The selection of the QIs, the methodology to derive the MI scores and the relevance of the selected key variables in assessing the quality of the estimates will be carefully tested. The simulation will be useful to derive the thresholds for the global QIs and MI scores that will be used to establish the decisional scheme to prioritize the follow-ups and the editing work. The minimization of the differences in the aggregate results between the proposed approach and the current processing and editing approach should serve as the basis of evaluation in regards to the establishment of thresholds[6]. This exercise will help to assess the feasibility of the new proposed approach and identify the adjustments needed and how it can be efficiently implemented in the regular survey process.

[5] Shaping expectations for the data is an important element in the editing strategy of the Annual Integrated Collection project of the Australian Bureau of Statistics – see Mackay (2007).

[6] This is in line with the approach used for the Retail Sales Inquiry Survey of the UK Office for National Statistics. The choice of score thresholds in the new selective editing approach, above which businesses are selected for follow-ups, was determined in Silva (2009). The thresholds for each domain were selected to minimize the amount of editing under the constraint of having an absolute difference between the estimates using the selective editing and the estimates using the current edit rules under 1%.

(c) Based on QIs, MI scores and the thresholds used for prioritization, the reduction of the volume of failed edits and non-response follow-ups will be estimated. This is a key issue to estimate the impact on interviewer's workload and for resource planning of the Corporate Collection Services.

(d) Results will be available to subject matter for their analysis. Experienced analysts will be asked to contribute to the assessment of the results and the establishment of the prioritization rules for their respective survey. Simulations are also going to be conducted in 2012 and 2013 by incorporating the findings of the 2011 simulation. The Rolling Estimates and CES will be in production in 2014.

VIII. References

Bricault, Mario (2010). *Common Editing Strategy – Discussion Paper*. Statistics Canada. Internal document.

Brinkley, Eden, Andrew McDonald and Lynne Bismire (2007). *Integration of Annual Economic Collections in the Australian Bureau of Statistics*. Presented to the 3rd International Conference on Establishment Surveys, Montreal, Quebec, Canada, June 2007

Chepita, Ryan (2006). *An Analysis of Pre-Grooming and Manual Imputation in the Service Industries of the Unified Enterprise Survey*. Statistics Canada. Internal document.

Cloutier, Lucie (2009). *Selective Editing for Business Surveys at Statistics Canada*. Statistics Canada. Paper presented at the UNECE Work Session on Statistical Data Editing in Neuchâtel, Switzerland, October 2009.

Godbout, Serge, Yanick Beaucage, Claude Turmelle (2011). Quality and Efficiency Using a Top-Down Approach in the Canadian's Integrated Business Statistics. Paper to be presented at the UNECE Work Session on Statistical Data Editing in Ljubljana, Slovenia, May 2011.

Lewis, Daniel, Alaa Al-Hamad (2009). *Assessing the Impact* of *Selective Editing on Data Quality*. Office for National Statistics, UK. Paper presented at the UNECE Work Session on Statistical Data Editing in Neuchâtel, Switzerland, October 2009

Mackay, Bob (2007). *AIC core editing strategy*. Australian Bureau of Statistics. Internal document.

Naud, Jean-François (2009). *Item Response Rates for the Most Important Financial Variables UES, Reference Year 2007*. Statistics Canada. Internal document.

Norberg, Anders, Chandra Adolfsson, Gunnar Arvidson, Peter Gidlund and Lennart Nordberg (2010). *A General Methodology for Selective Data Editing*. Statistics Sweden.

Pursey, Stuart (2003). *Use of the Score Functions to Optimize Data Collection Resources in the Unified Enterprise.* Statistics Canada. SSC Annual Meeting, June 2003.

Ratime, Rufin (2009). *Common Editing Strategy: Harmonization of Tools and Processes*. Statistics Canada. Internal document.

Saint-Pierre, Étienne (2010). *Consultations Report: Manual Adjustment of Survey Data*. Statistics Canada. Internal document.

Schouten, B., Cobben, F., Bethlehem, J. (2009), *Indicators for the representativeness of survey response*, Survey Methodology, 35 (1), 101 – 113

Silva P.L.N. (2009). *Investigating selective editing ideas towards improving editing in the UK Retail Sales Inquiry*. Internal report for ONS.

Tarassof, Peter S. (2010). *Rolling Estimates: A Baseline Review*. Statistics Canada. Internal document.

Yeung, Chi Wai (2007). *Manual Interventions in the UES edit and imputation Process*. Statistics Canada. Internal document.