Working Paper 21.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24–26 April 2017)

Automated Data Editing and Imputation for Surveys of Multinational Enterprises, a Banff Implementation

Prepared by Mark Xu, Andy K. Kim, and Larkin Terrie, Bureau of Economic Analysis, the United States

I. Introduction

In its annual and benchmark surveys of U.S. direct investment abroad and of foreign direct investment in the United States, the Bureau of Economic Analysis (BEA) collects operating and financial data from multinational enterprises.[1] These are the largest surveys conducted by BEA, and the resulting statistics are widely used in government, business, and academia. While BEA has traditionally edited these survey data by manually reviewing each individual completed survey, it has been searching for technologies that can produce more efficient and cost effective statistics without sacrificing data accuracy. Among the automated editing systems currently available from various agencies and organizations, the Banff system created by Statistics Canada was chosen for testing because (1) it has a long history and has been used by several national statistical agencies around the world (Barboza and Turner 2011, Johanson 2012, Kosler 2012, Winkler 2006), (2) the functionalities Banff offers are general and flexible enough to be adapted by the BEA, and (3) Banff's low maintenance cost.

The Banff system, which has its origins in the Generalized Edit and Imputation System developed by Statistics Canada, has existed for over three decades (Mohl 2007). The system is based on the work of Fellegi and Holt (1976), and in particular their finding, that better imputations are made when maximal use is made of the valid parts of questionnaires and as few fields as necessary are imputed. Banff is thus built on the assumption that the data in each record should be made to satisfy all edits by minimizing the number of fields edited. Banff offers nine independent SAS procedures that can be run in any order, including an error localization procedure that identifies which data items to correct for records that fail edits and several imputation procedures that can be used to fill missing data or replace the erroneous data identified in the error localization step.

[1] The Benchmark surveys are quinquennial surveys in which all companies in the sample frame are required to report, whereas for annual surveys, only companies notified by the BEA, and that meet certain thresholds, are required to report.

In 2015, BEA developed an auto-editing system with Banff as its core module. This auto-editing system uses six of Banff's nine procedures, including the error localization procedure as well as multiple imputation procedures. Two major tests of this auto-editing system were conducted using data from BEA surveys of multinational enterprises. The first of these was carried out using a subset of data from the 2011 Annual Survey of Foreign Direct Investment in the United States (BE-15), and the second was carried out with a subset of data from the 2009 Benchmark Survey of U.S. Direct Investment Abroad (BE-10). In both cases, the tests of Banff were judged to be successful, as the aggregate estimates produced from the Banff-processed survey data were highly similar to the aggregate estimates obtained when the same data were manually processed. With these successful test results, the system was used in production for a subset of data from the 2014 BE-10 survey in 2016, contributing substantially to the timely release of quality statistics from the survey.

The remainder of the paper is organized as follows: Section II explains the motivation behind BEA's adoption of an automated editing system by discussing the advantages of auto-editing vis-à-vis manual editing. Section III discusses constraints of the Banff system and the strategies that BEA developed to work around them. Section IV explains exactly how BEA adapted Banff to its needs, discussing which Banff procedures it has used for auto-editing and exactly how it combined them. Section V presents the results of two tests of the Banff auto-editing system, both of which were conducted with data from BEA surveys of multinational enterprises that had previously been manually edited. Section VI summarizes and concludes.

II. The Advantages of Automated Editing

BEA has traditionally relied on subject matter experts, survey editors, to manually correct instances where records fail edits. An edit is a validity condition that governs the relationships between, and the values that can be taken by, one or more survey items. This traditional approach to survey editing has both advantages and disadvantages. A major advantage is that, when a record fails an edit, editors are often able to directly contact the survey respondent and acquire the information necessary to correct the record in question. Editors also possess subject matter expertise and can perform research that allows them to correct items that an auto-editing system might miss or not be able to correctly change. These advantages are especially important in the case of high-priority[2] respondents whose survey responses have a large impact on final reported aggregate results.

However, a disadvantage of manual editing is that it requires substantial time and human resources and BEA receives far too many surveys for it to be feasible for editors to individually contact all respondents whose surveys fail edits. Indeed, in the case of low-priority respondents, editors often have to make imputations in order to force survey items to be consistent. These corrections, which can be subjective and thus variable from editor to editor, can also be time-consuming because inconsistent data items have to be checked against other items from the same survey form and against the same company's survey form from the previous year. A key part of the rationale for adopting an automated editing system is that it will allow BEA to efficiently and accurately edit a large number of surveys without requiring the close

[2] Priorities in business survey are usually based on company size, measured by, for example, the value of assets, net income, or sales. In auto editing, they can be set based on resource availability.

attention of editors to each individual survey. In particular, automated editing can reduce the amount of effort spent manually reviewing survey responses without damaging the quality of aggregate estimates, help make the correction of errors in survey responses more consistent while preserving the joint distribution of the data, and enable editors to focus manual effort on the accuracy of the survey responses that most strongly affect the overall estimates (Barboza and Turner 2011, Johanson 2012, Winkler 2006). When used for editing records that are not of the highest priority, auto-editing can thus help ensure the high quality of final data (Granquist and Kovar 1997, Lawrence and MacKenzi 2000).

Automated editing can also help to avoid certain pitfalls of manual editing that have been identified in the literature. Granquist and Kovar (1997) suggest that over-editing of survey data may occur, which is not just expensive but may also create errors that significantly affect the final estimate, especially when "soft edits"[3] are overly enforced. They point out that "data cleaning" does not necessarily identify and correct all erroneous data and leave true data unchanged and that re-contacting only respondents whose responses failed edits is an inefficient way of collecting high quality data. Furthermore, survey estimates from over-edited data can be skewed and are often not significantly different from estimates obtained from less heavily edited data. Lawrence and McKenzi (2000) recommend focusing manual editing efforts on the errors that have the greatest influence on the final data, known as significance editing, which is more efficient and can reduce respondent burden. By reducing the work burden of editors, the adoption of an auto-editing system can have precisely this effect, freeing editors to focus their efforts on correcting errors that have the largest influence on final reported estimates.

III. Banff's Constraints and BEA's Solutions

Despite its numerous advantages, there are also constraints inherent in Banff, some of which required BEA to develop creative work-arounds. In particular, three major constraints imposed by Banff are that edits must be linear, "hard" (meaning that at least one item is wrong if an edit is violated), and only applied to continuous numeric data. In contrast, many of BEA's traditional edits are nonlinear in form, "soft" (in that they are used to identify items that are suspicious but not necessarily wrong), and designed for categorical variables (such as industry classification codes). BEA adopted a variety of strategies, including the linearization of non-linear edits and pre-editing of data, to ensure that Banff could be used to process its multinational enterprise surveys.

From BEA's perspective, the most significant constraint imposed by Banff is that edits must be linear. Many of BEA's traditional edits are nonlinear, most commonly because they contain "if-then" statements but also frequently because they include logical operators (AND, OR), the not-equals sign (≠), ratios of variables, and/or absolute values. In many cases, it was possible to transform non-linear edits into linear edits. Table 1 summarizes when and how these transformations were made. Greater detail on these transformations, including examples, is provided in the Appendix.

[3] Soft edits are edits that are used to identify "suspicious" items rather than erroneous items.

Table 1: Transformation of Non-Linear Edits

Source of Non-Linearity	Solution
If-Then Statement	Apply edit (i.e., the "then" statement) only to records to which the "if" statement applies.
Ratio of Two Variables	Algebraic transformation (e.g., where *x* and *y* are variables and *a* is a constant, $x/y > a \rightarrow x > ay$).
Not equals (≠)	A transformation was only possible if the ≠ appeared in an if-then statement, in which case it was often possible to eliminate the ≠ by finding the contrapositive and treating it as the original edit.
Logical AND	Conditions separated by the AND operator were divided into separate edits.
Logical OR	A transformation was only possible if the OR appeared in an if-then statement, in which case it was often possible to eliminate the OR by finding the contrapositive and treating it as the original edit.
Absolute Values	An inequality containing an absolute value could be written as two separate inequalities joined by the logical AND if it took the form $\|x\| < a$. However, if it took the form $\|x\| > a$, no transformation was possible.

The treatment of edits containing "if-then" statements is a particularly important component of BEA's implementation of Banff. In general, Banff is set up so that all edits are applied to all records. However, BEA developed a procedure in which, after the transformation of non-linear edits, a subset of edits was selected and then applied to each individual record based on which if conditions were met by the record in question. As discussed below, this approach led to necessary modifications in how BEA ran two key Banff procedures, Proc Verifyedits (used to check whether edits are consistent with one another) and Proc Errorloc (used to identify which records fail edits and which fields within them should be imputed).

The other major strategy for addressing Banff's constraints was pre-editing of data. Pre-editing involves writing and executing SAS code to directly recode selected data items before the data are fed into Banff. We have found that effective pre-editing requires constant interaction between the statisticians who are responsible for the SAS code and survey editors, who are experts on the surveys themselves. Moreover, since each survey has its own set of edits, the SAS code developed for pre-editing is survey-specific and it often involves employing ad hoc methods to ensure that data are recoded in an appropriate manner. Pre-editing is also used to identify special cases where fields cannot be recoded based on predetermined criteria and have to be turned over to the survey editors for manual editing.

The largest portion of pre-editing involves correcting errors and inconsistencies in categorical variables. Banff is only able to auto-edit data that are numeric and continuous, and hence not categorical, since the error localization algorithm is based on a cardinality-constrained linear program (Sande 1979). During pre-editing, categorical items were often recoded based on other items in the same record. For instance, if a respondent reported its industry code as 8130 (meaning a religious, grant-making, civic, professional, or similar organization), it should report its form of organization as "individual, estate, or trust." When the industry code and form of organization were inconsistent, the form of organization was recoded to make them consistent. The recoding of data in the pre-editing stage was always based on the advice of survey

editors, who, in the case of this example, indicated that respondents tend to report their industry codes more accurately than their form of organization.

In addition, when an edit was not linearizable, it was applied during the pre-editing phase of data processing (i.e., before data were fed into Banff). For example, suppose there were a record for a holding company that reported zero assets. This record would violate the edit stating that, for holding companies, assets $\neq 0$, a statement that cannot be linearized. In this case, pre-editing would consist of assigning a null value to assets for this record, thereby ensuring that this value would later be imputed by Banff.[4] A similar method was used for other edits that could not be linearized.

Finally, Banff does not accept "soft" edits (i.e., edits used to identify items that are suspicious but not necessarily wrong). BEA's traditional edits are composed of both "hard" and "soft" edits. Edits for the Banff system need to be "hard" edits, which require corrections whenever they are violated. Pre-existing "soft" edits were carefully reviewed by experts, and they were either deleted or converted to "hard" edits depending on their importance. For the test of auto-editing conducted on data from the BE-10, a majority of "soft" edits were ultimately deleted. The conversion of "soft" edits to "hard" edits sometimes involves adding additional, more stringent, "if" conditions. For example, the BE-15A contains a "soft" edit stating that, for all companies, assets > 0. This edit was converted to a "hard" edit by adding an "if" condition specifying that it only applies to companies that are not holding companies.

IV. Approach to Implementing Banff

The Banff system provides nine independent SAS procedures that can be run in any order or even iteratively if desired. The independence of the procedures thus creates considerable flexibility in regard to how Banff is used. BEA developed an approach to implementing Banff that relies on six of the nine Banff procedures: Proc Verifyedits, Proc Errorloc, Proc Deterministic, Proc Donorimputation, Proc Estimator, and Proc Prorate. For a detailed description of the Banff system, see the *Functional Description of the Banff system for Edit and Imputation* V2.05. (Banff Support Team, 2012).

Once pre-editing of the data was complete, the first step was to run Proc Verifyedits to ensure that none of the edit rules contradicted one another. The Verifyedits procedure was run once for every record because the set of edits for each record was (potentially) different. These differences arose from the fact that many of the edits were initially in "if-then" form and, as discussed above, when edits were in this form, they were only applied to records to which the condition in the if clause applied.

The next step was to run Proc Errorloc, the error localization procedure, to identify fields to impute (FTIs). Similar to Proc Verifyedits, Proc Errorloc was run for each record with only a subset of all edits that are applicable to that record. Prior to running Proc Errorloc, weights were assigned to fields according to their relative importance and reliability in order to ensure that the least important or least reliable fields were chosen for imputation when Banff had to choose an FTI from among multiple potential fields. Weights were chosen in close consultation with editors who have extensive experience with BEA survey

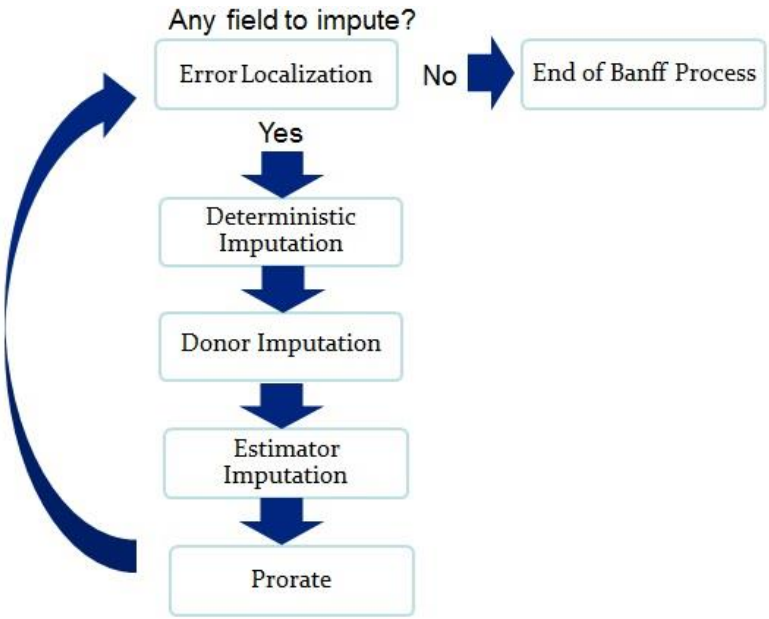
[4] For all missing items (or items with a null value), Banff imputes a new value.

forms and therefore knowledge of which fields are more important and, for fields of similar importance, which field is most likely to be reported correctly. The assignment of weights enabled Banff to mimic the behavior of subject matter experts when faced with a choice regarding which field(s) to impute.

Note that the error localization process requires that each variable processed by Banff has at least an upper bound or a lower bound given the set of edits. When there is no feasible region for a variable given the set of edits, the user can arbitrarily specify upper and lower bounds and add those to the set of edits. For our test runs, the upper and lower bounds were set based on the historical minimum and maximum for each variable. We discovered that setting the bounds for each variable may require some trial-and-error. Making the bounds too narrow can distort the original edits. However, if they are made too wide, they may not add any useful information in regard to setting the feasible region. For this reason, we customized all boundary conditions as a part of edits in Proc Errorloc and did not use the Banff Proc Outlier.

Once FTIs were identified with the Errorloc procedure, the process of imputing new values for the FTIs was begun. In general, the imputation procedures were run in the following order: 1) Proc Determinstic (deterministic imputation), 2) Proc Donorimputation (donor imputation), 3) Proc Estimator (imputation estimators), and 4) Proc Prorate (pro-rating). The rationale behind this ordering was that the procedure that produces the most reliable imputations, Proc Deterministic, should be run first and that Proc Donorimputation and Proc Estimator should be used to impute values only if they cannot be imputed by Proc Deterministic. Proc Donorimputation was run second for the robustness and flexibility of donor imputation (Chen and Shao 2000, Beaumont and Bocci 2009). Proc Prorate was run following Proc Estimator because Proc Estimator sometimes produces imputations that violate summation rules.

Figure 1: Iterative Banff Process



As shown in Figure 1, the preceding steps make up the first iteration of the Banff process. After the first iteration, Proc Errorloc was run on the output from the first iteration to find out whether there were any FTIs left. If there were no FTIs left, the Banff process was considered complete. If there were FTIs left, then another round of the four imputation procedures was run. This process was repeated until there were no FTIs left. We found that, in general, two iterations sufficed to complete the auto-editing process. If a record could not be cleaned after two iterations, it was usually because it contained multiple erroneous items and was thus not cleanable using the auto-editing process. Such records were handed over to the editors for manual review.

V. Testing BEA's Implementation of Banff

BEA has conducted two major tests of its implementation of Banff for its surveys of multinational enterprises.[5] In both cases, the data used for the test had been previously manually edited, thus making possible a comparison of the results obtained from Banff auto-editing with the results obtained from manual editing. BEA conducts surveys of both U.S. direct investment abroad and foreign direct investment in the United States, and one of each type of survey was tested. The first test was conducted on a survey of inward foreign direct investment, the 2011 BE-15 survey. The test was limited to the U.S. affiliates of foreign multinational enterprises that reported using the BE-15 A form, which requires affiliates to be majority foreign-owned and have total assets, sales, or net income of over $275 million (positive or negative), resulting in 1,510 records for the analysis. The analysis was conducted using only the Balance Sheet and Change in Retained Earnings sections of the BE-15 A form. These sections were chosen for the test because they exist in both outward and inward direct investment surveys and the edits involved are sufficiently complex to test the feasibility of the auto-editing system. The complexity stems from the fact that each field is governed by multiple edits and thus relationships with other fields, meaning that each value for a given record must satisfy multiple equalities or inequalities.

It took approximately one year to set up the Banff auto-editing system for the initial test using the 2011 BE-15 A form. Linearizing edits, assigning variable weights, and programming pre-edits required multiple meetings with editors as well as repeated trial and error, as results from the auto-editing system were compared against the estimates produced with manual editing to fine-tune the system. Statistical analyses were also conducted to find appropriate models for use with Banff's Estimator procedure. However, once the auto-editing system was in place, modifying it for the second test was relatively straightforward, in part because the SAS programs written for the first test were modulized with separation of codes and data inputs that could be reused with, in general, only minor changes.

The second test was conducted using data from a survey of U.S. direct investment abroad, the 2009 BE-10 survey. This test was confined to U.S. parent companies of U.S. multinational enterprises that reported assets, sales, and net income of less than $300 million (i.e., enterprises that filed using the BE-10

[5] BEA's multinational enterprise surveys include quarterly, annual, and benchmark surveys of U.S. direct investment abroad and surveys of foreign direct investment in the United States. Surveys of U.S. Direct Investment Abroad include the BE-10, BE-11, and BE-577 and Surveys of Foreign Direct Investment in the United States include the BE-12, BE-13, BE-15, and BE-605. For detailed information, refer to http://www.bea.gov/surveys/pdf/a-guide-to-bea-direct-investment-surveys.pdf.

Short A form), resulting in 1,176 records for the analysis.[6] We chose to conduct a test on benchmark data because, if found feasible, we could use the system to edit part of the 2014 BE-10 benchmark survey that was in production in 2016 and would benefit from additional resources for editing. The year 2009 was chosen because it was the most recent benchmark year for the U.S. direct investment abroad surveys. In addition, the BE-10 Short A form was selected for testing because it is a relatively simple form (in terms of the complexity of it edits), and we wanted to be able to complete the test in time to use the auto-editing system for the processing of 2014 BE-10 data. The BE-10 Short A form is comprised of simplified versions of the Balance Sheet and Income Statement, Employment, and Exports and Imports of goods sections.

Tables 2 and 3 compare the number of fields[7] that were manually edited versus the number of fields identified for imputation and then imputed by Banff in each of the two tests. These tables provide two key takeaways. First, in both tests, editors edited over twice the number of fields imputed by Banff – 12.8% edited fields versus 6.1% imputed fields in the case of the 2011 BE-15 A and 3.5% edited fields versus 1.0% imputed fields in the case of the 2009 BE-10 Short A. Second, there is significant overlap in the fields edited by editors and the fields identified for imputation by Banff. In the case of the 2011 BE-15A, of the 6.1% fields imputed by Banff, more than two-thirds of these were also edited by editors; and in the case of the 2009 BE-10 Short A, of the 1.0% fields imputed by Banff, more than half were also selected for editing by the editors. These two findings indicate that Banff may be more efficient than traditional editing, in that it alters fewer total fields, but that it still identifies many of the same fields for correction as the editors (which is likely a result of the weights assigned in the error localization process).

Table 2: Percent of Fields Edited Manually vs. Auto-Edited, 2011 BE-15 A

		Manual Editing		
		Edited	**Not Edited**	**Total**
Auto-Editing	**Edited**	4.1%	2.0%	6.1%
	Not Edited	8.7%	85.3%	93.9%
	Total	12.8%	87.2%	100%

Table 3: Percent of Fields Edited Manually vs. Auto-Edited, 2009 BE-10 Short A

		Manual Editing		
		Edited	**Not Edited**	**Total**
Auto-Editing	**Edited**	0.6%	0.5%	1.0%
	Not Edited	2.9%	96.0%	99.0%
	Total	3.5%	96.5%	100%

[6]Although the results are not presented in this paper, the auto-editing system was also tested on 2009 BE-10 D forms, required of certain, relatively small foreign affiliates of U.S. enterprises. The results of this test were similarly satisfactory to those presented in this paper.

7 A field in Banff refers to a data cell regardless of whether it contains data or is null.

Tables 4 and 5 present percentage differences in the estimates of key aggregate statistics that resulted from auto-editing versus traditional manual editing and the percentage shares of fields edited by auto-editing and manual editing for each item. The most important conclusion to be drawn from these two tables is that, in general, the final aggregate statistics produced by auto-editing are highly similar to those produced by traditional manual editing. The estimates resulting from auto and manual editing are, in most cases, within one or two percentage points of one another. In addition, Banff was able to produce these final statistics in a more efficient manner than manual editing. For almost all items, the percentage share of data points that Banff altered to produce the final estimate was significantly lower than that altered in traditional editing to produce the final estimate.

Table 4: Comparison of Auto-Edit and Manual Edit by Data Item, 2011 BE-15A

Variables	% Difference Auto vs. Manually Edited*	% Share of Total Fields Auto-Edited**	% Share of Total Fields Manually Edited***
Assets	0.67	2.52	12.91
Inventories	0.74	1.19	4.04
Equity Investment	-0.47	7.95	14.24
Net PPE	-3.36	1.66	7.02
Other Assets	1.06	8.01	24.37
Liabilities	1.07	12.19	19.80
Owners' Equity	-1.73	1.79	14.44
Capital Stock	0.47	2.85	12.45
Retained Earnings	-58.02	2.32	18.34
Treasury Stock	25.73	0.33	0.60
Acc. Comp. Income or Loss	-35.54	0.86	13.58
Trans. Adj.	107.26	0.20	4.90
Other Acc. Comp. Income or Loss	-11.43	2.05	13.05
Other Owners' Equity	-2.51	0.66	11.66

*(Auto-edited/Manually edited – 1)*100

** (# of fields auto-edited/total # of records)*100

*** (# of fields manually edited/total # of records)*100

Table 5: Comparison of Auto-Edit and Manual Edit by Data Item, 2009 BE-10 Short A

Variables	% Difference Auto vs. Manually Edited	% Share of Total Data Points Auto-Edited	% Share of Total Data Points Manually Edited
Sales	0.62	3.57	8.93
Employment	8.87	5.44	17.52
Net Income	149.34	0.34	6.97
Assets	1.48	0.26	14.03
Liabilities	0.66	0.60	6.55
Exports	-3.79	6.29	5.44
Imports	-0.44	3.23	3.57

There is a small number of items in both tables for which the difference between the estimate from auto-editing and the estimate from manual editing appear to be substantial (for example, retained earnings and translation adjustment in table 4 and net income and employment in table 5). However, these differences are due in large part to special circumstances of the test cases that will not necessarily pertain when Banff is used in the actual production of statistics. Recall that the analysis of data from the 2011 BE-15A was limited to items in the Balance Sheet section of the survey. Retained earnings has relationships, which are governed by edits, with items that are in other sections of the survey, such as the Change in Retained Earnings and Income Statement sections, that were not included in the test.[8] The initial test of the BE-15 A did not take these edits into account, which helps to explain why the final estimates from auto-editing differ from the final estimates obtained from manual editing.

In regard to the BE-10 Short A, the relatively large discrepancies in the estimates of net income and employment can be attributed to the fact that the editors used the Securities and Exchange Commission's 10-K data to verify the accuracy of net income and employment reported on the BE-10.[9] In a subsequent test of Banff on the BE-10 Short A, 10-K data were taken into account and the percent difference for the estimate of net income was reduced from 149 to 63 percent.[10] Another source of discrepancies in the statistics estimated from the BE-10 Short A data is that editors examine survey forms for the U.S. enterprises' foreign affiliates when editing the BE-10 Short A data, which are for the U.S. parent companies. Data from foreign affiliate reports are related to the parent reports and editors make revisions

[8] For example, net income is a component that is in the retained earnings section as well as the income statement section.

[9] 10-K data refer to data in the annual report submitted by public companies to SEC that provides a comprehensive overview of the business and financial condition of the companies, including audited financial statements. Although most of definitions of variables differ, there are a few variables from the BEA surveys that are either defined the same way by the SEC or have a close relationship with variables from 10-K data. For instance, Net Income from the BE-10 Short A form must be equal to that from 10K data. Also, Assets and Liabilities from the BE-10 Short A cannot be greater than those from 10-K data.

[10] We assumed that data from 10K are correct, so when items from BE-10 Short A were inconsistent with 10K, we pre-edited the BE-10 Short A data so that they are consistent with 10-K data.

to data reported on the parent forms for consistency with affiliate forms. Foreign affiliates' data have not yet been incorporated into Banff processing of the BE-10 Short A.[11]

Tables 6 and 7 present information on how frequently each imputation method was used in each test of the Banff auto-editing system. The most important takeaway from these two tables is that in both tests deterministic imputation[12] was used far more frequently than any of the other imputation methods. This finding provides additional evidence that Banff is producing valid imputations because deterministic imputation is the most reliable of the methods available through Banff. Auto-editing of the 2011 BE-15A, in particular, involved a high proportion of deterministic imputations, over 94 percent of all imputations (when including all rounds/iterations of the auto-editing process). The higher proportion of deterministic imputations for the BE-15A than for the BE-10 Short A is likely due to the greater complexity of the BE-15A. When a form is highly complex, each data point is governed by multiple edits and thus relationships with other data points. As a result, there is considerable scope for a computer to find solutions for each data point that a human might not be able to discover. On the other hand, form complexity increases data heterogeneity and reduces the number of potential quality donors that leads to the very small percentage of Donor Imputation for the BE-15A compared to the BE-10 Short A.

Table 6: Frequency of Different Imputation Methods, 2011 BE-15A

Imputation Method	Status Code	Description	Imputed Fields %*
Deterministic	IDE	Deterministic (1st Round)	92.2
	IDE2	Deterministic (Subsequent Rounds)	1.9
Donor	IDN	Donor Imputation	0.4
Estimator	ILR1	Historical Regression Model	1.9
	IDIF	Residual Difference Between Two Variables	0.2
	IHRT	Historical Subcomponent Share	0.3
	ICA	Carry Forward	0.4
Prorate	IPR	Prorated (Previously Imputed Fields)	2.7
All			100.0

[11] The differences for Assets, Net Income, and Employment are mainly due to the use of survey forms from foreign affiliates during the traditional editing process, which has not yet been taken into account in the auto-editing process. For example, Net Income of U.S. parents should be greater than the income of their directly held foreign affiliates, which is reported in the BE-10 B, C, and D forms. When this rule is violated, editors conduct further investigation and edit Net Income based on the research. Similar processes are taken for Assets, Liabilities, Exports, and Imports. For Employment and Sales, 10-K data, when available, are used to edit the A, B, C, and D forms.

[12] Deterministic Imputation identifies the cases in which there is only one possible value that will allow the record to pass the original edits (Banff Support Team, 2012).

Table 7: Frequency of Different Imputation Methods, 2009 BE-10 Short A

Imputation Method	Status Code	Description	Imputed Fields %
Deterministic	IDE	Deterministic Imputation	46.8
Donor	IDN	Donor Imputation	10.7
Estimator	ICOM	Imputation by Regression	15.7
	ISUM	Imputation by Sum	7.7
	IPRO	Imputation by Proportion	4.7
Prorate	IPR	Prorated	10.2
	EDT	Editors' recommendation	4.2
All			100.0

Finally, tables 8 and 9 present evidence regarding the reliability of the auto-editing system. Generating reliable estimates is as important as generating accurate estimates. If the system generates accurate estimates for one dataset and inaccurate estimates for another, the system is not reliable. One way of measuring reliability is to calculate the variance of the difference between estimates from the auto-editing system and estimates from traditional manual editing. If the system is reliable, the variance should be low relative to the magnitude of the difference in the estimates. To measure the variance of the difference, the jackknife resampling method was used on the data from the tests of the 2011 BE-15A and the 2009 BE-10 Short A.[13] Jackknife resampling involves estimating a parameter by systematically leaving out each observation in the dataset and calculating the parameter based on all of the other observations in the dataset. When there are N observations in a dataset, the jackknife estimate of a parameter is obtained by aggregating the estimates from these N subsamples, each of which is of size $N - 1$.

It is not desirable to use the jackknife resampling method on the data after they have been processed by Banff because some values in the data may be the result of auto-editing imputations that were based on the observations left out during jackknife resampling. To get around this problem, jackknife estimates were derived by leaving out one record at a time, running the auto-editing process,[14] and then calculating the final estimates of interest. Tables 8 and 9 present jackknife estimates of the differences between auto-editing and manual editing for each variable in tables 4 and 5 as a percentage of the aggregates obtained from manual editing when the surveys were originally edited.

[13] The jackknife method is especially useful in estimating bias-corrected estimation and variance estimation. Jackknife estimators are found by leaving out one record at a time and calculating the estimate without the record and then averaging the estimates at the end.

[14] For each individual jackknife estimate using a subsample of size n-1, a single iteration of the Banff process was run (Proc Errorloc, Proc Deterministic, Proc Donorimputation, Proc Estimator, Proc Prorate) and then the aggregate estimate was calculated based on this single iteration. Only one iteration of the Banff procedures was run in order to save processing time and because the vast majority of FTIs are imputed on the first iteration.

Table 8: Jackknife Estimates of the Range of Percent Difference between Auto-Edit and Manual Edit, 2011 BE-15A

Variables	Range of Percentage Difference	
Assets	-0.63%	1.80%
Inventories	-0.55%	2.03%
Equity Investment	-9.95%	8.52%
Net PPE	-8.47%	1.75%
Other Assets	-0.37%	2.38%
Liabilities	-0.16%	2.21%
Owners' Equity	-6.39%	2.95%
Capital Stock	-3.67%	1.53%
Retained Earnings	-143.58%	29.68%
Treasury Stock	-21.93%	59.81%
Accum. Comp. Inc. Loss	-89.08%	0.75%
Translation Adj.	-75.54%	289.92%
Other Acc. Comp. Inc. Loss	-43.07%	5.80%
Other Owners' Equity	-45.04%	35.50%

Table 9: Jackknife Estimates of the Range of Percent Difference between Auto-Edit and Manual Edit, 2009 BE-10 Short A

Variables	Range of Percentage Difference	
Sales	-0.20%	1.15%
Employment	2.94%	15.18%
Net Income	-33.01%	157.33%
Assets	0.87%	3.46%
Liabilities	-2.30%	0.92%
Exports	-7.79%	-1.30%
Imports	-8.22%	3.46%

The results presented in tables 8 and 9 have two sets of implications. First, the 95% confidence intervals for the difference between estimates from the auto-editing system and estimates from traditional editing include zero for all fourteen variables in table 8 and for four of the seven variables in table 9. For these variables, the differences between the results obtained from manual editing and those obtained from auto-editing are not statistically significant at the 95% confidence level. To be sure, in table 9, assets, employment, and exports do have confidence intervals that do not include zero, indicating a systematic difference between the manual editing and auto-editing results. However, these differences were

accounted for above and should thus not be seen as an indication of a systematic problem with the auto-editing system.

Second, for eight of the fourteen variables in table 8 and five of the seven variables in table 9, the confidence intervals are relatively narrow, as the absolute value of both the upper and the lower bound is less than ten percent. The fact that the percentage differences of six variables in table 8 have relatively wide ranges might be seen as potentially troubling. However, the somewhat wide ranges for these variables do not indicate any systematic problems with the auto-editing system as they are likely due to the fact, discussed above, that the test using the BE-15A was limited to the Balance Sheet section of the survey and thus did not take account of the relationships between these variables and items in other sections of the survey.

VI. Discussion and Conclusion

BEA's auto-editing system provides an efficient and consistent means of processing survey data. Setting up this system did, however, require substantial effort, including the sustained efforts of experts in business accounting, statistical analysis, and SAS programming. BEA had to develop a systematic approach to overcoming certain constraints of the Banff system that could have, at first glance, made Banff seem incompatible with the needs of BEA. In particular, strategies had to be developed for 1) dealing with the many non-linear edits that BEA has traditionally used in editing its surveys of multinational enterprises, 2) editing categorical variables, and 3) incorporating "soft" edits into the BANFF survey editing process.

Once these obstacles were overcome, BEA developed a Banff-based auto-editing system that incorporates six of Banff's nine SAS procedures. This auto-editing system was tested using two subsets of data from previously manually edited BEA surveys of multinational enterprises, the 2011 BE-15 A and the 2009 BE-10 Short A forms. These tests showed that the final aggregate estimates produced with auto-edited data tend to be very close to those produced with manually edited data. In addition, auto-editing appears to be more efficient than manual editing in that, even though the final aggregate estimates of the two are similar, auto-editing ultimately arrives at these estimates by altering fewer total fields than does manual editing. The fact that the auto-editing system tends to make heavy use of deterministic imputation, a highly reliable imputation method, also provides reason to support the accuracy of the editing done through the auto-editing system.

Following the successful tests of the 2011 BE-15 and the 2009 BE-10, the auto-editing system was used for the actual processing of the 2014 BE-10 Short A and D survey forms, for which a combined 20,000 forms were submitted to BEA.[15] Without the auto-editing system, it would have taken an additional 20 accountants approximately two months to process these forms. Since the processing of these survey

[15] Not all 2014 BE-10 Short A and D forms were edited due to errors that cannot be fixed systematically. For example, when a respondent did not provide a valid four-digit NAICS (North American Industry Classification System) code for its industry with the largest sales, the record was turned over to the editors for manual review. Since industry codes are categorical variables, they cannot be imputed by Banff. Identifying the correct industry for a respondent when one is not provided often requires research on the part of survey editors.

forms represented the first time the auto-editing system was used for official data production purposes, the results were carefully reviewed by field experts. The results were found to be highly satisfactory and were accepted with only minimal additional manual editing. Given this outcome, BEA is planning to expand the use of the auto-editing system to other surveys in upcoming years.

Based on the test results discussed in this paper, we were also able to identify ways of improving BEA's auto-editing system. For example, differences between manually edited and auto-edited survey items can be reduced if the auto-editing system is set up for the entire survey, thus allowing the system to take account of all relationships among survey items instead of only those among items in a certain segment of the survey. Differences between manually and auto-edited data can also be reduced by incorporating 10-K data into the auto-editing system, which was in fact done when the system was used to process 2014 BE-10 forms. There is also scope, in the future, for improving the processing of the BE-10 Short A by taking account of data reported for directly held foreign affiliates on the BE-10 B, C, and D forms. However, one obstacle to doing this is the difficulty of simultaneously editing the BE-10A, B, C, and D forms. During the traditional manual editing process, an editor edits the BE-10A, B, C, and D forms concurrently for each U.S. reporter (i.e., each U.S. company that files an A form). This approach enables the editor to balance between items from A forms and items from B, C, and D forms, making A, B, C, and D forms consistent for each U.S. reporter. A new module would need to be developed either within Banff or outside of Banff based on the ownership structure of each multinational enterprise to ensure data consistency among all forms for the parent company and its foreign affiliates in each multinational enterprise.

References

Banff Support Team. 2012. Functional Description of the Banff System for Edit and Imputation. Version 2.05. Statistics Canada, Ontario.

Barboza, W. and Turner, K. 2011. "Utilizing Automated Statistical Edit Changes in Significance Editing." National Agricultural Statistics Service, Joint Statistical Meetings.

Beaumont J.F. and Bocci, C. 2009. "Variance Estimation When Donor Imputation is Used to Fill in Missing Values." *The Canadian Journal of Statistics.* Vol. 37, Issue 3. September 2009. 400-416.

Chan J. and Shao J. 2000. "Nearest Neighbor Imputation for Survey Data." *Journal of Official Statistics.* Vol. 16, No.2, 2000, pp. 113-131.

Fellegi, I.P., and Holt D. 1976. "A systematic approach to automatic edit and imputation." *Journal of the American Statistical Association*, 71, 17-35.

Granquist, L. and J. G. Kovar. 1997. "Editing of Survey Data: How Much Is Enough?" In *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley, 415-435.

Johanson, J.M. 2012. "Banff Automated Edit and Imputation on a Hog Survey." National Agricultural Statistics Service, American Statistical Association.

Kosler, J.S. 2012. "Survey Process Control with Significance Editing: Foundations, Perspectives, and Plans for Development." National Agricultural Statistics Service, Joint Statistical Meetings.

Lawrence, D. and R. McKenzi. 2000. "The General Application of Significance Editing." *Journal of Official Statistics*, Vol. 16, No. 3, pp. 243-253.

Mohl, C. 2007. "The Continuing Evolution of Generalized Systems at Statistics Canada for Business Survey Processing." In *Proceedings of the Third International Conference on Establishment Surveys* (ICESIII) (June 18-21, 2007), American Statistical Association, 758-768. Available at: http://www.amstat.org/meetings/ices/2007/proceedings/ICES2007-000135.PDF

Sande, G. 1979. "Numerical Edit and Imputation." Presented at the 42nd International Statistical Institute Meeting, Manila, Philippines.

Winkler, W.E. 2006. "Data Quality: Automated Edit/Imputation and Record Linkage." U.S. Census Bureau, Research Report Series (Statistics #2006-7).

Appendix: Linearizing Non-Linear Edits

Many of BEA's traditional edits are nonlinear, most commonly because they contain "if-then" statements but also frequently because they include logical operators (AND, OR), the not-equals sign (≠), ratios of variables, and/or absolute values. Some non-linear edits are relatively easy to transform into a linear form but some are more challenging or even impossible. Some examples are provided to aid in understanding the steps taken to transform non-linear edits.

A. If-Then

- Edits with an if-then structure may be the most common form of traditional edits. The Banff system does not accept if-then in edits because they are not in a linear form. These edits, however, are conveniently transformed into a linear form by breaking up the "if" statement and the "then" statement, and then applying the edit from the "then" statement only to records to which the "if" statement applies.
- Example: If industry ≠ holding company, then sales > 0. This traditional edit is transformed into sales > 0, and is applied to non-holding companies. Because this edit is applied only to a selection of records, separate Banff processes are executed for records for holding companies and non-holding companies.

B. Ratio of one variable to another

- Transforming edits that include the ratio of one variable to another involves simple algebra if it has the form of Var1/Var2 < c, where c is a constant. However, if there is a third variable in the edit, it cannot be transformed into a linear form.[16]
- Example: Exports/Sales < 0.15. This edit is transformed into Exports < 0.15*Sales.

C. Not equal to

- For numerical variables, there are an infinite number of ways one variable can be not equal to a linear combination of other variables or a constant. It is often impossible to transform such edits into a linear form. If an inequality edit existed in an if-then condition, then, depending on what was in the "if" statement, the edit was transformed using logical equivalency by switching statements in "if" and "then."
- Example: If Liabilities > 0 then Assets ≠ 0. This edit can be equivalently stated "if Assets = 0 then Liabilities ≤ 0." Since Liabilities cannot be a negative number, the edit becomes if Assets = 0 then Liabilities = 0.

D. AND

- If two conditions are joined by an "AND," it means the data must satisfy both conditions. So if an edit is composed of two or more conditions joined by "AND," they are separated into individual edits.

[16] It is possible to transform the edit into a linear form by taking logarithm of both sides, but if variables considered in the edit happen to be in another edit, it becomes extremely difficult to simultaneously edit the same variables in different forms.

E. OR

- Edits are sometimes satisfied if at least one of two or more conditions is met. They are usually expressed with two or more satisfactory conditions joined by "OR." Direct transformation of such edits is mostly impossible because satisfying one condition from multiple conditions is not allowed in linear programming. If the edit was in the form of "if-then," then it was transformed using logical equivalency as described earlier.

F. Absolute values

- Edits regarding bounds of linear combination of variables are often expressed with absolute values. Depending on how the bounds are set in an edit, it was transformed into a corresponding linear form.
- Example: |Net Income| < 300,000. This edit is equivalent to Net Income > -300,000 ***and*** Net Income < 300,000, which is equivalent to two separate individual edits.
- Example: |Net Income| > 300,000. This edit is equivalent to Net Income > 300,000 ***or*** Net Income < -300,000, which cannot be directly transformed into a linear form.

These examples do not cover all types of edits we encountered, but do cover a majority of them. Sometimes an edit is constructed with a combination of the listed examples, and it was treated in a logical way. In all cases, whenever an edit needed to be transformed into another form, subject matter experts were consulted to make sure the transformed edit was logically equivalent to the original edit.

The last comment regarding these transformations is on the treatment of greater than, or ">." Since the Banff system assumes numerical data are continuous, it takes ">" as "≥" given that there is practically no difference between the two for a continuous variable. In reality, there is a difference between ">" and "≥." For instance, in an edit, Liability > 0 means liability must be greater than zero and it cannot be zero. However, when the edit is fed into the Banff system, it is interpreted as Liabilities ≥ 0, implying as long as Liabilities is not zero, it passes the edit, which is not desired. To remedy the issue, the edit is changed to Liabilities ≥ 1.