WP 13

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (i): Development and use of data editing quality indicators

Impact of Data Processing on Unified Enterprise Survey Micro Data
- Route to the Final Data Point

Contributed Paper
Submitted by Statistics Canada[1]

1. Background

1.1 Purpose of Study

Assessing data quality is an area of survey methodology that is as important as any other although it does not lie along the critical path of production for most surveys. Data quality is something generally discussed before a survey is designed, and then assessed after the survey is completed. In this regard, data quality is much different than other methodological survey steps such as sampling, questionnaire design and variance estimation. One cannot complete a survey without these steps yet assessing data quality can be delayed or reduced depending on the resources available to a survey manager.

The Unified Enterprise Survey (UES) is the large business survey that was created from the Project to Improve Provincial Economic Statistics (PIPES) so assessing data quality was imperative. Our goal was to obtain a complete and thorough assessment of data quality. The aspect of data quality that will be discussed in this paper is the impact that data processing has on the micro data of the UES. Specifically, to answer questions such as: “What happens to the data between the stages of survey processing?” Are we undoing previous changes often?”; “What is the impact of these changes?”; “Is there a trend in the changes made?”.

1.2 Unified Enterprise Survey (UES)

This large annual business survey combines a wide range of industries including services, wholesale, retail and manufacturing to name a few. It seeks to measure various provincial and national economic variables including revenue, expenses and employment. It began in 1997 as a pilot project with only 7 industries and now, in its sixth cycle includes over 30. Being such a large survey, it’s design and production routine is inherently complex and the stages of data processing are no exception. There are 7 basic stages of data processing in the UES.

[1] Prepared by Fred Hazelton [Frederick.Hazelton@statcan.ca]

1. Data capture
2. Post data capture correction (pre-processing)
3. Edit and Imputation
4. Post edit and imputation correction (Manual Correction)
5. *Allocation*
6. *Calendarization*
7. *Estimation*

The main goal of this project was to follow records at the micro data level as they travelled through the stages of data-processing. For reasons of simplicity our assessment was limited to the first four stages above.

2. The Concept

2.1 The Challenge

Our first attempts at assessing the impact of data processing included traditional simple techniques. Tabulating the average percent changes between stages and counting the number of positive and negative changes at each stage are some examples. These techniques provided some useful results but failed to present a complete picture of what was happening to the data from its initial collection to its final resting place after stage 4. We then began to explore new ways of examining the data. Was there a way that we could get the "Big Picture"? Was it possible to come up with a new way to represent the path that a particular data point travels and would that representation give us a better sense of what was happening to the data and how it impacted both the survey process and the survey results?

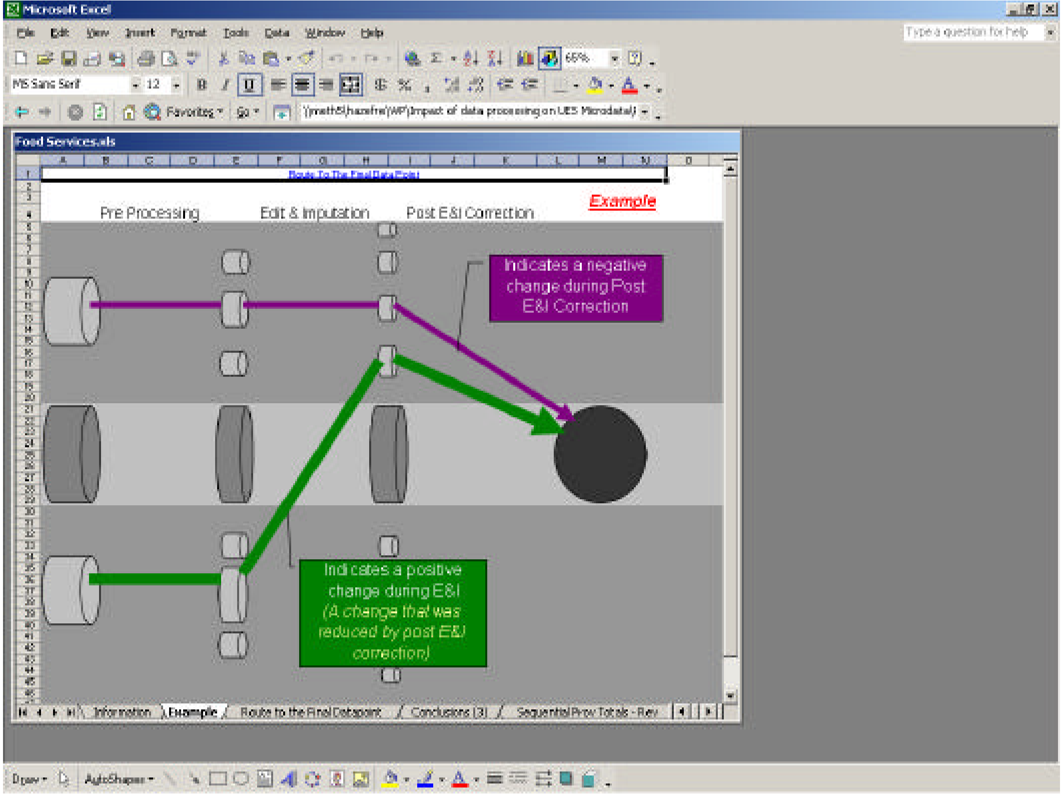
2.2 The Experiment

One idea that seems possible was that of a data point map. A picture of the four processing stages and a visual representation of how the data points travelled though them. This map could provide easy access to supplemental statistics through hyperlinks and perhaps use 2 or three dimensions to explain the impact of data processing. With more than 30 surveys including thousands of respondents and hundreds of fields each, it was impossible to represent each data point individually. We would need a solid concept to simplify the picture and make it useful for analysis.

2.3 The Concept

We came to the realization that for our purposes of analysis we could define each and every data point at any given stage of processing according to one of three categories. Suppose we treat the fourth stage (our last stage of analysis) as the fixed resting place of each data point. Now at any stage previous to that fourth stage the data point can be higher than, equal to or lower than its final fixed value. Similarly, between any 2 stages a data point can be affected by a positive change, no change or a negative change. Figure A shows an example of how we apply this concept and present the results in an excel diagram.

Figure A


Microsoft Excel
File Edit View Insert Format Tools Data Window Help
Type a question for help
65%
MS Sans Serif 12
Favorites Go
\\math5\hazefre\WP\Impact of data processing on UES Microdata\
Food Services.xls
Route To The Final Data Point
Pre Processing
Edit & Imputation
Post E&I Correction
Example
Information
Example
Route to the Final Datapoint
Conclusions (3)
Sequential Prov Totals - Rev
Draw
AutoShapes

Indicates a negative change during Post E&I Correction

Indicates a positive change during E&I *(A change that was reduced by post E&I correction)*

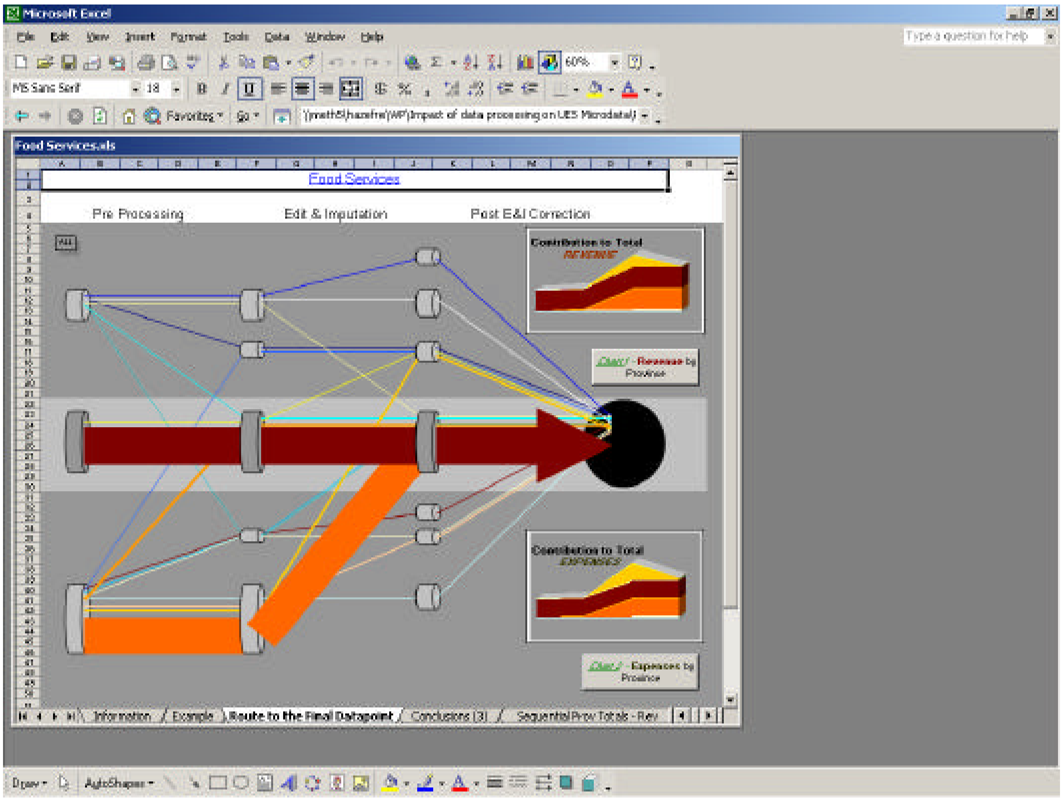
We see in Figure A the four stages of interest moving from left to right beginning with collected data, travelling through pre-processing, edit and imputation and then post edit and imputation correction before arriving at the big black dot which represents the final value for every micro data record. The direction of the line segments indicates whether the change was positive, negative or zero and their thickness indicate the number of data points that travelled that path. Recall that in each segment, the location of each line shows its relationship to the final value. For example, the thicker bottom line in the pre-processing stage indicates that these data points began at a value lower than the final value. Similarly, in the edit and imputation stage we see that the value was increased from a value lower than its final value to a value that was higher than its final value.

3. A Visual Tool

3.1 Applying the concept to Food Services data

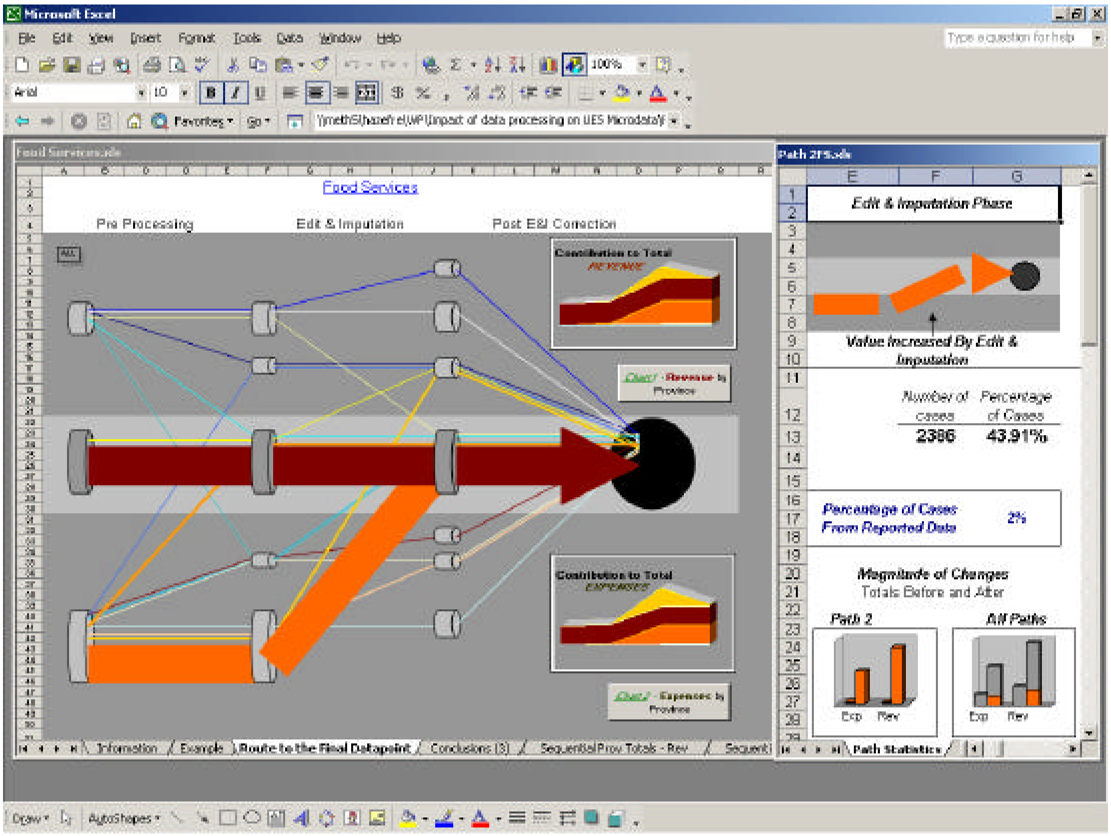
We apply this concept to the data stored from the various stages of processing on the UES database. As an example, we choose the Food Services industries within the UES. We focus on 2 main variables of interest, total revenue and total expenses from reference year 1999. We obtain the data from each of the four stages for both variables and apply the same concept to get a complete diagram of paths for this industry.

Figure B



There are 17 unique paths that appear but we see from the thickness that a large majority of them fall on 2 main paths. The straight path that travels through the middle and the large path at the bottom that increases in the edit and imputation stage. The path through the middle represents about 45% of all data points in this industry for the two variables. These data points are completely unchanged throughout the 4 data processing stages. The large path at the bottom represents 44% of all data points and its path signifies data points that begin at a lower value than their final value, remain unchanged through pre-processing, are increased during E&I and remain unchanged through post E&I correction. In order to view statistics on the paths we assign hyperlinks to each path segment that link to other excel spreadsheets. For example, suppose we clink on the middle segment of the large path at the bottom.

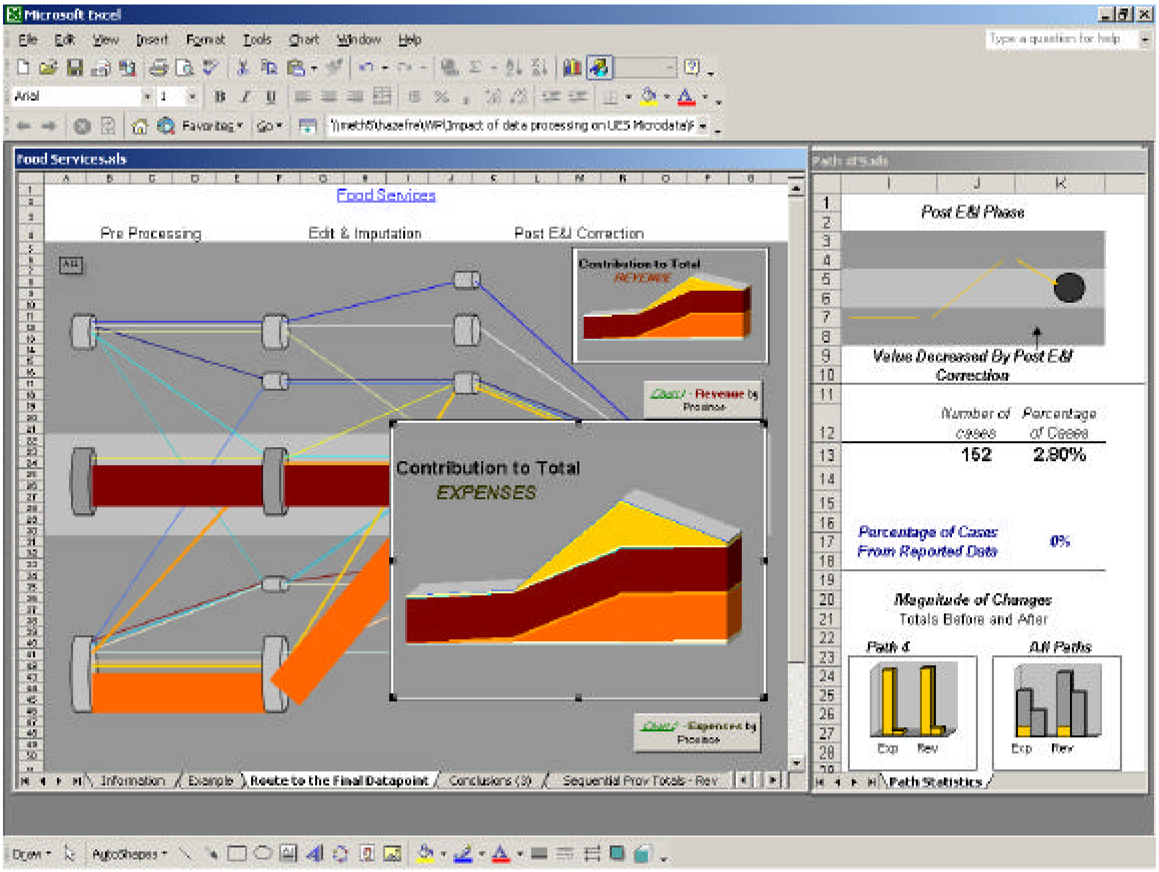
Figure C


Microsoft Excel
File Edit View Insert Format Tools Data Window Help
Type a question for help
Arial 10
100%
Favorites Go
\\meth5\hazel\rel\WP\Impact of data processing on UES Microdata\
Food Services.xls
Food Services
Pre Processing
Edit & Imputation
Post E&I Correction
Contribution to Total REVENUE
Chart - Revenue by Province
Contribution to Total EXPENSES
Chart - Expenses by Province
Information / Example / Route to the Final Datapoint / Conclusions (3) / SequentialProv Totals - Rev / Sequent
Path 2FS.xls
Edit & Imputation Phase
Value Increased By Edit & Imputation
Number of cases
Percentage of Cases
2386
43.91%
Percentage of Cases From Reported Data
2%
Magnitude of Changes
Totals Before and After
Path 2
All Paths
Exp
Rev
Path Statistics
Draw AutoShapes

The statistics appear in a smaller spreadsheet to the right. Included are the name of the segment, a brief description, the number of cases, the percentage of cases and the percentage of cases that come from reported data. This latter is included to determine what proportion of the data points actually contained data after the data collection stage. Any non-blank value that is present after data collection constitutes reported data. The charts at the bottom are a representation of something that is otherwise missing from the main diagram. In the main diagram, there is no indication of the magnitude of the changes that occur. A data point that is changed from 1,000 to 2,000,000 would lie on the same path as a data point that is changed from 1,000 to 1,500. The "Magnitude of Changes" charts indicate the total value of all data points before and after the stage being examined. The total expressed is the weighted total of all data points using the sampling weights. The left chart indicates the weighted total of all data points on this particular path before and after the stage for each of the two variables. The right one indicates the proportion that those data points contribute to the entire data set.

The results in these "Magnitude of Changes" charts are mirrored in the main diagram by two area charts, one each for total revenue and total expenses. These charts show how the weighted totals change from stage to stage as well as how each path contributes to the weighted totals at each stage.

Figure D



	Number of cases	Percentage of Cases
	152	2.80%

Percentage of Cases From Reported Data	0%

Figure D highlights the “Contribution to Total Expenses” area chart. In it we see how the total weighted expenses increases slightly after pre-processing and then to a larger degree after the edit and imputation phase, both not surprising results. However what is a surprising result is the sharp decrease in the final post edit and imputation correction stage. We noted earlier that the main diagram was dominated by two main paths but now there seems to another path that has an impact. This is the path that has a triangular shape on top of the area chart. We know that it was not a particularly thick path in the main diagram so it does not represent a large number of data points. This is confirmed in the statistics spreadsheet at the right (represents 2.80% of all cases). However, the “Magnitude of Changes” charts and the area charts show us that the impact of these data points is significant. After edit and imputation these accounted for a quarter of the total yet they were reduced in the post edit and imputation stage to account for barely 2% of the total. These are results that can stimulate discussion among those involved in the design and maintenance of the survey, certainly a goal of any data quality assessment tool.

4. Conclusions

4.1 The Answers (for Food Services)

Now that we have the visual tool we can attempt to answer those questions that were posed when outlining the purpose of the study (section 1.1).
"What happens to the data between the stages of survey processing?"

- Most data is unchanged or changed automatically by the edit and imputation stage.

"Are we undoing previous changes often?"

- Occasionally although with little impact.

"What is the impact of these changes?"

- 3 paths dominate the diagram, the main middle path, the large bottom path and the triangular path in the area chart.

"Is there a trend in the changes made?"

- No but their impact has a definite trend throughout the stages: increase, increase, decrease

4.2 The Outcomes

Our experiment led to taking a different approach to analysing the path of micro data through data processing and presenting the results. We have a dynamic visual tool that allows us to get a real sense of the data movement throughout the survey process. Some outcomes we discovered were expected and others were not. What we do know for sure is that this tool will present for us a relatively concise way of viewing the big picture. The results that we discover from the tool will surely lead to further questions and hopefully to further investigation with the ultimate goal of improving survey processing steps and overall data quality.

At present it is hoped that this visual tool will be used by survey managers to analyse the data processing of their respective surveys. We have created the excel diagrams for other industries and for other reference years. It would be advantageous to expand the tool to include other variables of interest and perhaps to expand the functionality. It would be particularly interesting to allow the thickness of each path segment to represent a measure of the cost of processing rather than the number of cases thereby creating an indication of the strain on recourses relative to their potential value. As well allowing a user to eliminate or combine path segments and then reconfigure the diagram would be a good addition to allow "what-if" analysis. Certainly, the tool appears to be an excellent way to represent the impact of micro data processing and its concepts will hopefully be continually improved so that the goals of the data quality analysts are met.