Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

Table Abstraction Tool (TAT)

User Manual

Raghav Padmanabhan

2/3/2009

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

CONTENTS

I. Important Notes 4

II. Table Abstraction 5
1. Select Title 6
2. Select Caption 7
3. Augmentations 8
   a. Footnotes 8
   b. Aggregates 10
   c. Units 12
   d. Other 13
4. Table Analysis 15
   a. Verification 15
   b. Analysis 17
5. Highlight Cells & Indented Notation 19
   a. Highlight Cells 19
   b. Indented Notation 21
6. Generate XML 22

III. Transformations on Tables for Processing..... 23

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

LIST OF FIGURES

Figure 1: TAT.....5

Figure 2: Paste Error 6

Figure 3: Highlighted title & Caption 7

Figure 4: Footnote Cell Selection 8

Figure 5: Footnote Citation 9

Figure 6: Footnote Reference 10

Figure 7: Aggregate Cell Selection-I 11

Figure 8: Aggregate Cell Selection-II 12

Figure 9: Units 13

Figure 10: Other Augmentations-I 14

Figure 11: Other Augmentations-II 15

Figure 12: Canonicalized Table 17

Figure 13: Category Prompt 18

Figure 14: TAT highlighting category cells 19

Figure 15: TAT highlighting delta cells associated with a category 20

Figure 16: TAT highlighting one delta cell associated with two categories 21

Figure 17: TAT indented Notation Error 22

Figure 18: Table requiring transformations to make it 'TAT-friendly' 25

Figure 19: TAT bad table error 26

Figure 20: Transformed 'TAT-friendly' table.....27

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

TABLE ABSTRACTION TOOL - MANUAL

The Table Abstraction Tool (TAT) is an interactive tool used to generate an abstract notation of a table in the form of an XML file that records the relationships between the cells in the table. The tool extracts the logical structure of the table from its Excel representation with some user input. Each user action is time-stamped and logged in a file. The tool is built in Microsoft Excel utilizing VBA, an event driven programming language built into Microsoft Office Applications. It is required that the user of TAT be fairly proficient with Microsoft Excel. The process consists of four steps some of which require multiple user actions.

I. IMPORTANT NOTES:

To run TAT:

The application is a set of macros written in VBA. Therefore, enable macros before running the application. In Excel 2003, the Security Settings in the menu path **Tools > Macros > Security** need to be set to **Medium** or **Low** to allow the macros to run. In Excel 2007, the Macro security changes can be made from **Code** group > **Developer** tab > **Macro Security**. If the Developer tab is not displayed then click **Microsoft Office Button** & click **Excel options** and then in the **Popular** category, under **Top options for working with Excel**, check **Show Developer tab in the Ribbon.**

To modify TAT:

Before editing a macro the user should be familiar with the Visual Basic Editor - an environment in which one can write new and edit existing Visual Basic for Applications code and procedures. The Visual Basic Editor contains a complete debugging toolset for finding syntax, run-time, and logical errors in the code. The Visual Basic Editor can be used to write and edit a macro that is attached to a Microsoft Office Excel workbook. To view the code, click **Visual Basic**. Code is present in the following code modules: *Sheet1*, *Input_Form* , *Module 1* & *public Class*.

Other Notes:

The program works through a number of subroutines that store temporary variables. Some data whose scope is restricted to one subroutine and whose values cannot be retrieved are stored in a separate sheet called “MySheet” which is hidden. It contains data that the program uses in various subroutines. The log for the process is recorded by the system as and when the actions are performed by the user, in a hidden sheet called “Log”. Once the user clicks the [GENERATE XML] button, the program creates the file in the file system. On clicking the [START] button, values in both the sheets are cleared. This manual is also embedded in the Excel Workbook itself and appears as an Adobe Acrobat icon in the worksheet in the first column. Clicking the icon opens this Help manual.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

The process of *abstracting* a table is described below:

II. TABLE ABSTRACTION:

The first step is to open the excel workbook – "TAT.xls", which has all the code in the code modules (This can be verified using the Visual Basic Editor). Then, copy the original HTML table into Microsoft Excel. To preserve the table structure in Excel, the contents in the web page outside the table (some content which is not related to the table) should also be copied and pasted directly. The user can later remove the extraneous content although it is not necessary. Then click the [START] button in the Excel sheet (which should be integrated into Eclipse framework eventually). This draws the borders for the cells occupied by the table and pops up a User Form with the following options (displayed as buttons) for the user:

1. **SELECT TITLE**
2. **SELECT CAPTION**
3. **AUGMENTATIONS**
4. **TABLE ANALYSIS**
5. **HIGHLIGHT CELLS**
6. **GENERATE XML**

Figure 1 displays an example table:

	A	B	C	D	E	F
1	Trips by Canadians in Canada, by province and territory					
2	(Trips (destination))					
3		2000	2001	2002	2003	2004
4		Trips (destination)				
5		thousands				
6	**Canada**	**178,628**	**182,092**	**187,890**	**172,244**	**175,084**
7	Newfoundland	3,955	3,902	3,784	3,236	3,107
8	Prince Edward Island	977	966	1,125	897	911
9	Nova Scotia	7,034	7,019	8,287	7,164	7,066
10	New Brunswick	4,794	5,344	6,075	5,613	5,038
11	Quebec	40,842	40,608	45,928	47,216	48,484
12	Ontario	65,220	67,160	70,257	62,168	65,290
13	Manitoba	6,542	6,621	6,265	5,938	6,009
14	Saskatchewan	8,222	8,139	8,029	7,413	7,451
15	Alberta	20,022	21,256	19,186	15,775	15,890
16	Columbia	20,893	20,984	18,842	16,742	15,738
17	Yukon	F	92[E]	113[E]	83[E]	99[E]
18	[E] : use with caution.					
19	F : too unreliable to be published.					
20	**Notes:**					
21	- Estimates are based on the 1996 Census population counts.					
22	- 80 km or more.					
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.					
24	Last Modified: 2006-04-06.					
25						
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and rela					

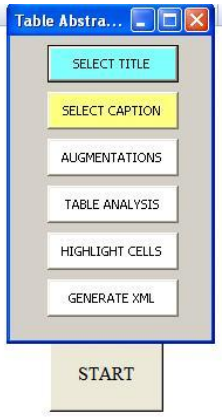


FIGURE 1. TAT

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

Note: In **Figure 1**, the information in row 26 **– "Find information related to this table ……."** is not part of the table. But for a faithful representation of the table in Excel, that line from the HTML is also copy-pasted in the Excel. This is the way Excel works.

Figure 2 shows the Excel representation of the table if text outside the <table>… </table> tags is not copied.

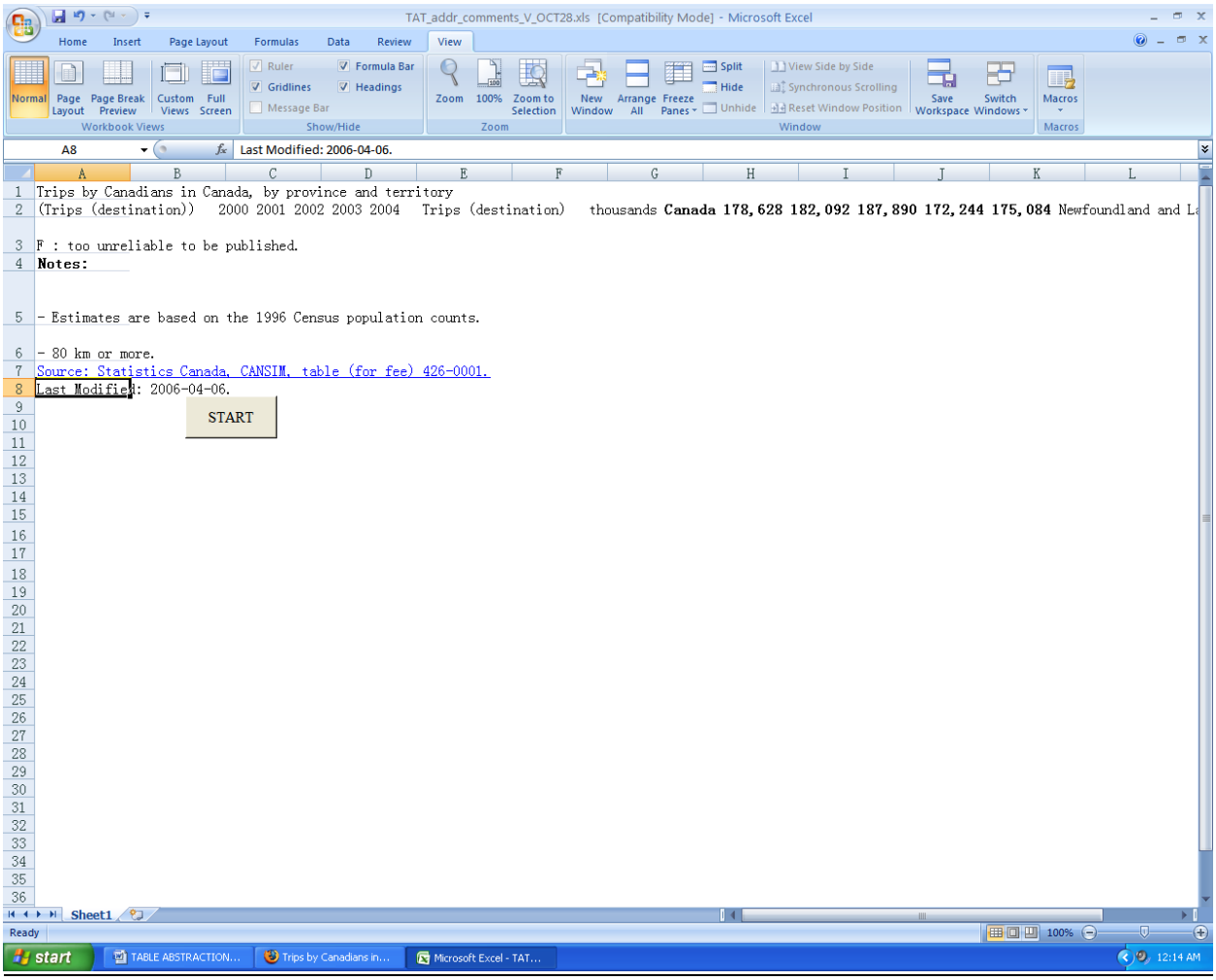


FIGURE 2. PASTE ERROR

1. SELECT TITLE:

The blue [SELECT TITLE] button displays a pop up that asks the user to select the cells containing the title. The cells selected by the user as the title are also colored. In the above example, the title would be "**Trips by Canadians in Canada, by province and territory**"

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

2. SELECT CAPTION:

The yellow [SELECT CAPTION] button, when clicked, displays a pop up that asks the user to select the cells representing the table caption. The cells selected by the user as caption are also colored. The caption in this case would be "**(Trips (destination))**".
For tables without title and caption, the user can directly press the [TABLE ANALYSIS] button. However, if title and caption are present and the user does not select them, the XML representation will not contain those details and these cells will also be canonicalized in the 'Table Analysis' action. The order of selecting the title and caption does not matter. The title can also be selected after selecting the caption.

Note: Please refer to the Table Analysis section for Canonicalization process.

After selecting the title and caption, the table is highlighted with the title and caption (**Figure 3)**. TAT is now ready to process 'Augmentations'

	A	B	C	D	E	F
1	Trips by Canadians in Canada, by province and territory					
2	(Trips (destination))					
3		2000	2001	2002	2003	2004
4		Trips (destination)				
5		thousands				
6	**Canada**	**178,628**	**182,092**	**187,890**	**172,244**	**175,084**
7	Newfoundland	3,955	3,902	3,784	3,236	3,107
8	Prince Edward Island	977	966	1,125	897	911
9	Nova Scotia	7,034	7,019	8,287	7,164	7,066
10	New Brunswick	4,794	5,344	6,075	5,613	5,038
11	Quebec	40,842	40,608	45,928	47,216	48,484
12	Ontario	65,220	67,160	70,257	62,168	65,290
13	Manitoba	6,542	6,621	6,265	5,938	6,009
14	Saskatchewan	8,222	8,139	8,029	7,413	7,451
15	Alberta	20,022	21,256	19,186	15,775	15,890
16	Columbia	20,893	20,984	18,842	16,742	15,738
17	Yukon	F	92[E]	113[E]	83[E]	99[E]
18	[E] : use with caution.					
19	F : too unreliable to be published.					
20	**Notes:**					
21	- Estimates are based on the 1996 Census population counts.					
22	- 80 km or more.					
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.					
24	Last Modified: 2006-04-06.					
25						
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; an					

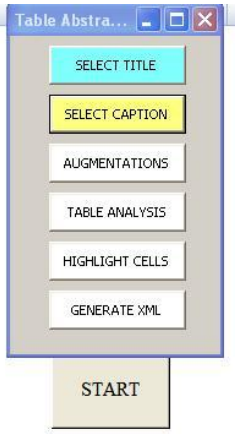


FIGURE 3. HIGHLIGHTED TITLE & CAPTION

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

3. AUGMENTATIONS:

The [AUGMENTATIONS] button, when clicked, displays a pop up which helps the user to select the augmentations for the cells in the table. There are four kinds of augmentations that a user can currently choose (using radio buttons):

1. Footnotes
2. Aggregates
3. Units
4. Other

3 (a). Footnotes:

To specify a footnote, the user must make three selections:
(1) Cells to which the footnote applies. (These cells also contain the footnote reference.)
(2) Footnote Citation- which specifies the footnote text.
(3) The symbol used as the footnote reference.
When the user clicks selects the 'Footnote' option (after clicking the [AUGMENTATIONS] button) and clicks the [OK] button, the system displays a pop up that asks the user to select the cells with the footnote reference (**Figure 4)**.

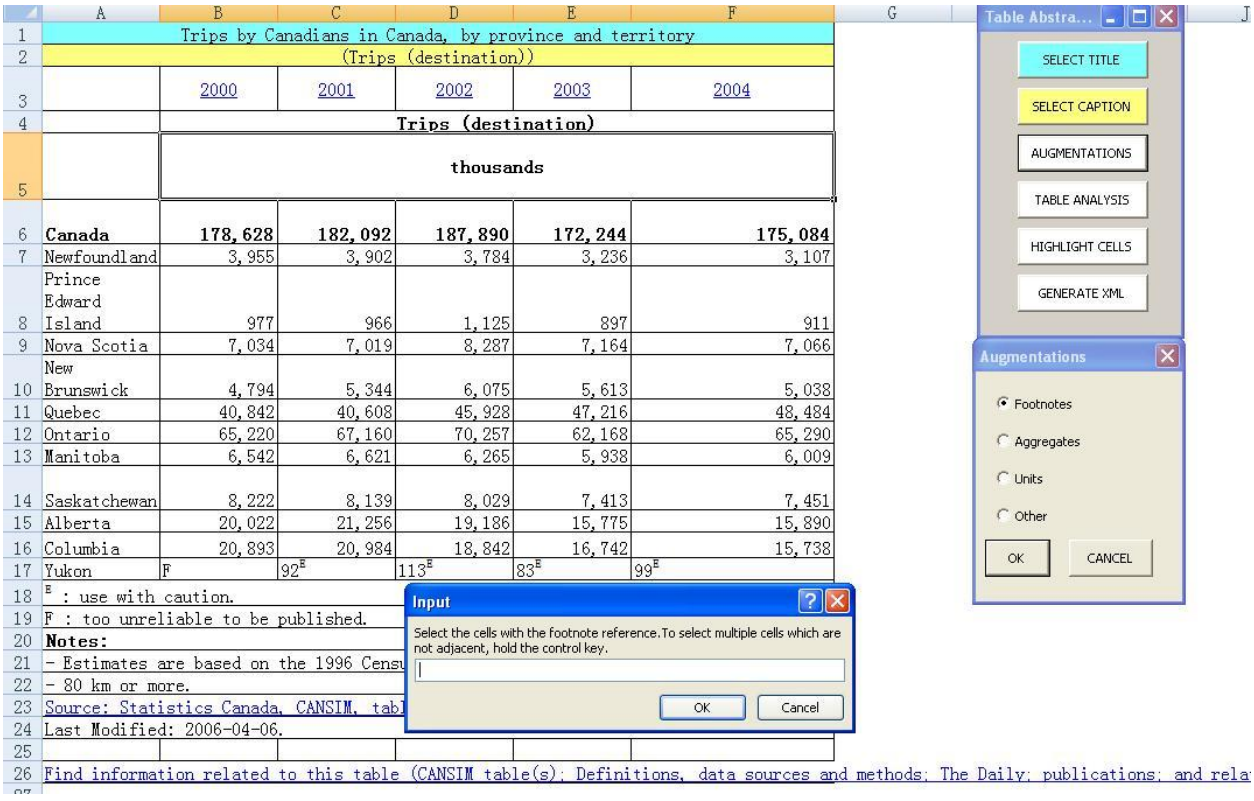


FIGURE 4. FOOTNOTE CELL SELECTION

In the above table there are two footnotes. The first one corresponds to the cell with address "**B17**" i.e. the value of number of trips to Yukon territory in the year 2000 is **"F"** which is '*too unreliable to be published'*. Thus, the user selects cell "**B17"** for the first pop up where the cells with the footnote reference are requested and clicks the [OK] button (**Figure 4**). This causes another pop up to be displayed which asks the user to select the cells with the corresponding footnote citation (**Figure 5)**.

.

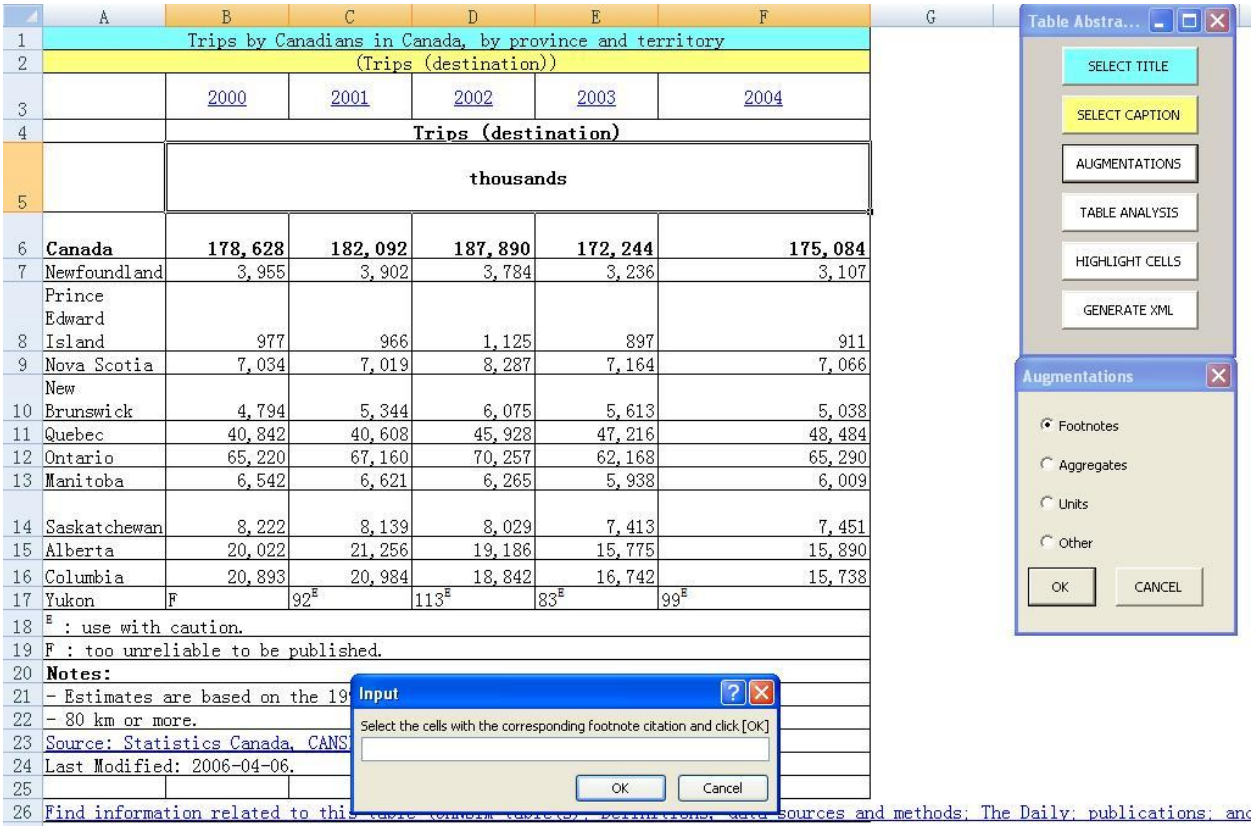


FIGURE 5. FOOTNOTE CITATION

The user then selects cell **"A19" –** with the text '*too unreliable to be published*', as the footnote citation, and clicks the [OK] button. This causes another pop up to be displayed which requests the user to specify the footnote reference (which is "**F**" in this case). After finishing this process of specifying one footnote, the following text appears as a comment for the cell "**B17"–**

(footnote)F : too unreliable to be published.(/footnote)(f.reference)F(/f.reference)

The string within the *(footnote)* and *(/footnote)* is the footnote text and the string between *(f.reference) and (/f.reference)* is the footnote reference. This allows the user to verify the actions performed.

The third popup asks the user for the symbol used for footnote reference (**Figure 6).**

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

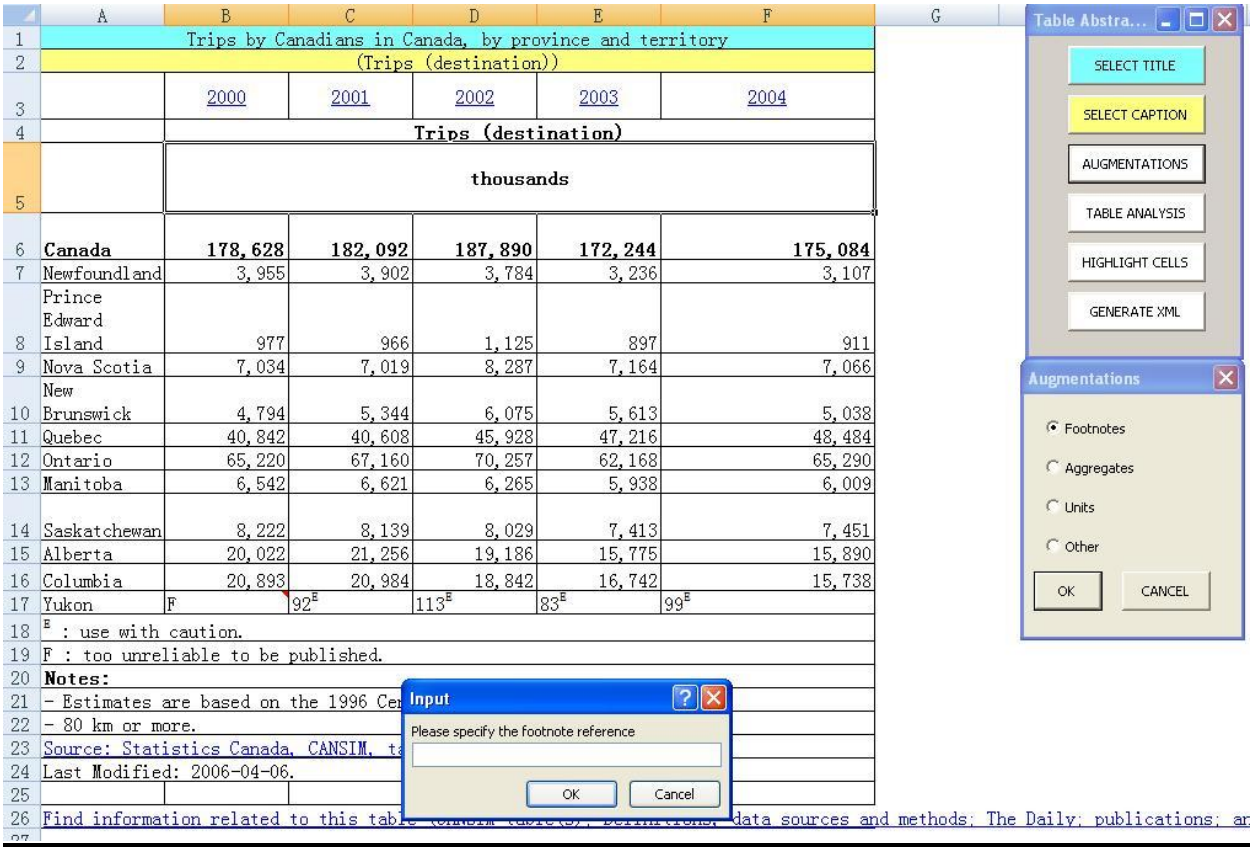


FIGURE 6. FOOTNOTE REFERENCE

To specify the values for the second footnote- cells '**C17: F17'**, the user has to make sure the 'Footnote' radio button in the menu is selected and click the [OK] button.

The following are the values for the second footnote:

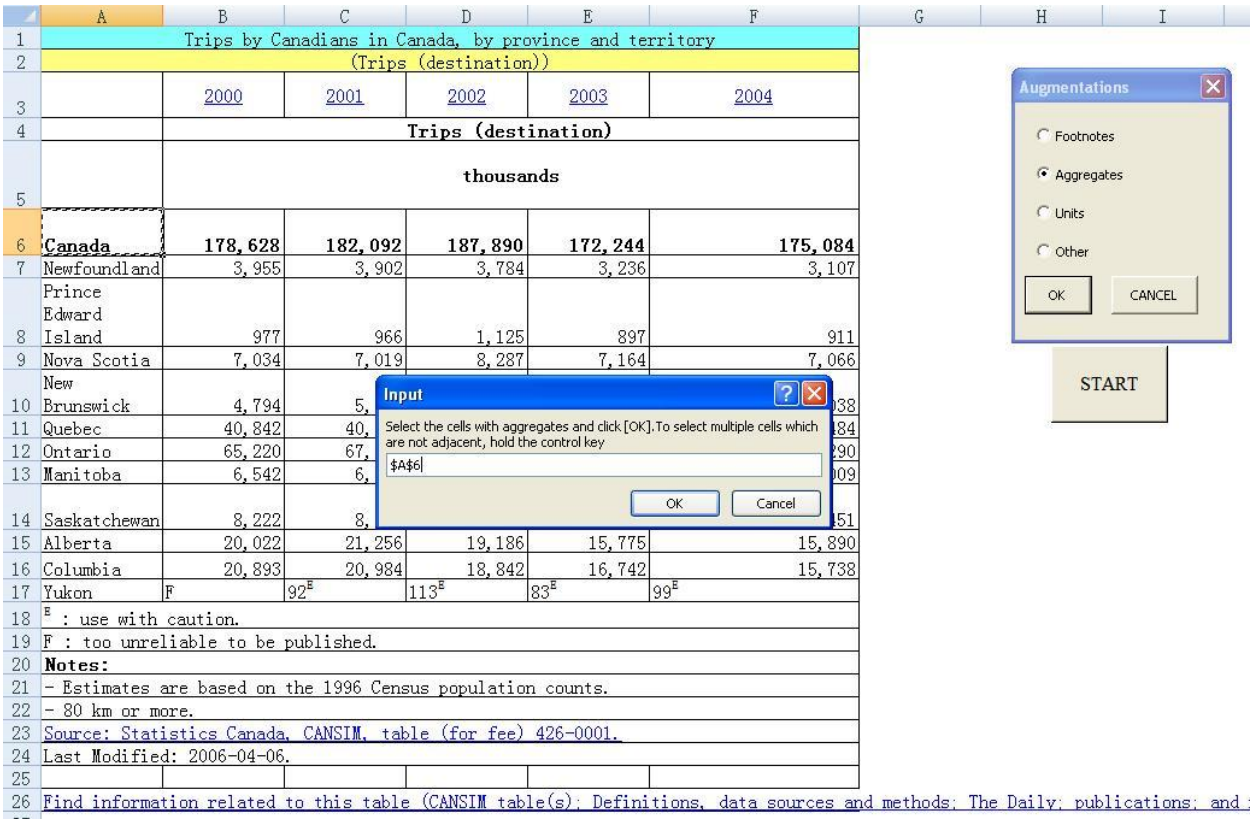
1. Cells with the footnote reference: **C17:F17**
2. Cells with the footnote citation: **A18 ("E: use with caution")**
3. Footnote reference: **E**

3(b). Aggregates:

TAT allows the users to specify the aggregate cells in the table. The aggregate cells are the category/sub-category cells whose corresponding delta cells are an aggregate (like sum or average) of delta cells of other sub-categories. In the example table, the delta cells corresponding to "**Canada**" i.e., the delta cells in the same row as Canada are actually a summation of all the delta cells below them. Thus, **Canada** is an aggregate of the sub-category cells – Newfoundland and Labrador, Prince Edward Island, Nova Scotia, etc.

When the user selects the 'Aggregate' option and clicks the [OK] button, the system asks the user to select the aggregate cells (**Figure 7)**.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

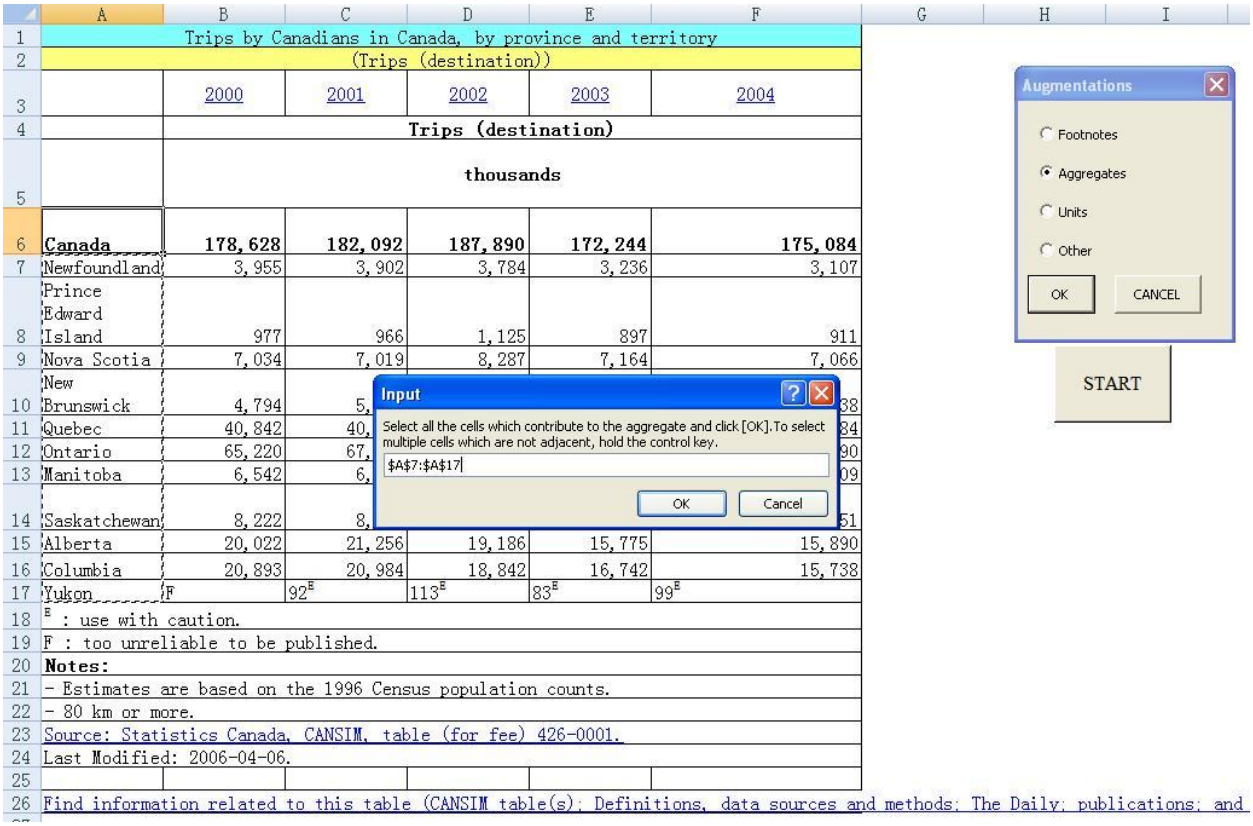


	A	B	C	D	E	F
1	Trips by Canadians in Canada, by province and territory					
2	(Trips (destination))					
3		2000	2001	2002	2003	2004
4		Trips (destination)				
5		thousands				
6	Canada	178,628	182,092	187,890	172,244	175,084
7	Newfoundland	3,955	3,902	3,784	3,236	3,107
8	Prince Edward Island	977	966	1,125	897	911
9	Nova Scotia	7,034	7,019	8,287	7,164	7,066
10	New Brunswick	4,794	5,			38
11	Quebec	40,842	40,			84
12	Ontario	65,220	67,			90
13	Manitoba	6,542	6,			09
14	Saskatchewan	8,222	8,			51
15	Alberta	20,022	21,256	19,186	15,775	15,890
16	Columbia	20,893	20,984	18,842	16,742	15,738
17	Yukon	F	92[E]	113[E]	83[E]	99[E]
18	[E] : use with caution.					
19	F : too unreliable to be published.					
20	Notes:					
21	- Estimates are based on the 1996 Census population counts.					
22	- 80 km or more.					
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.					
24	Last Modified: 2006-04-06.					
25						
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and					

FIGURE 7. AGGREGATE CELL SELECTION-I

The user selects the cell – "**Canada**" and clicks the [OK] button. The system displays another pop up asking the user which subcategory cells are aggregated by the previously selected cell (**Figure 8**). The user selects all the cells from "**Newfoundland**" to "**Yukon**", as the values for the delta cells corresponding to "**Canada**" are the sum of the delta cells corresponding to these subcategories.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009



	A	B	C	D	E	F
1	Trips by Canadians in Canada, by province and territory					
2	(Trips (destination))					
3		2000	2001	2002	2003	2004
4		Trips (destination)				
5		thousands				
6	Canada	178,628	182,092	187,890	172,244	175,084
7	Newfoundland	3,955	3,902	3,784	3,236	3,107
8	Prince Edward Island	977	966	1,125	897	911
9	Nova Scotia	7,034	7,019	8,287	7,164	7,066
10	New Brunswick	4,794	5,			38
11	Quebec	40,842	40,			84
12	Ontario	65,220	67,			90
13	Manitoba	6,542	6,			09
14	Saskatchewan	8,222	8,			51
15	Alberta	20,022	21,256	19,186	15,775	15,890
16	Columbia	20,893	20,984	18,842	16,742	15,738
17	Yukon	F	92[E]	113[E]	83[E]	99[E]
18	[E] : use with caution.					
19	F : too unreliable to be published.					
20	Notes:					
21	- Estimates are based on the 1996 Census population counts.					
22	- 80 km or more.					
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.					
24	Last Modified: 2006-04-06.					
25						
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and					

FIGURE 8. AGGREGATE CELL SELECTION-II

On selecting the subcategory cells and clicking the [OK] button, the system displays the following text as a comment for the cell "**A6"** containing information about the cells it aggregates.

(aggregate)$A$7:$A$17(/aggregate)

The string between the *(aggregate) and /(aggregate)* identifiers specifies the cell addresses of the cells it aggregates.

3.(c). Units:

The units of the delta cells are important in interpreting the table contents. TAT allows the users to specify the cells in the table which are units. In the above table, the spanning cell "**thousands"** specifies the units of the Number of trips by Canadians. In the Augmentations menu, if the user selects the 'Units' radio button and clicks the [OK] button, the system displays a pop up asking the user to select the cells which denote units in the table (**Figure 9**)

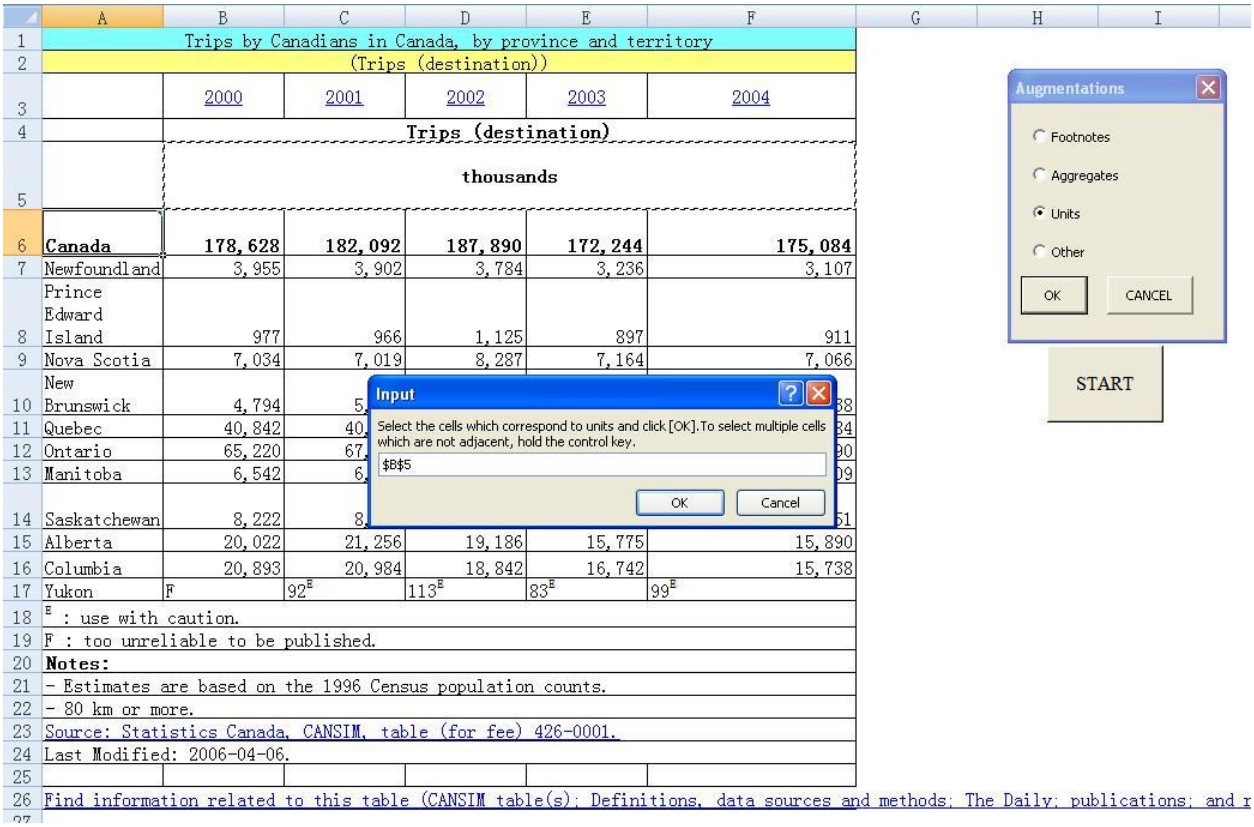
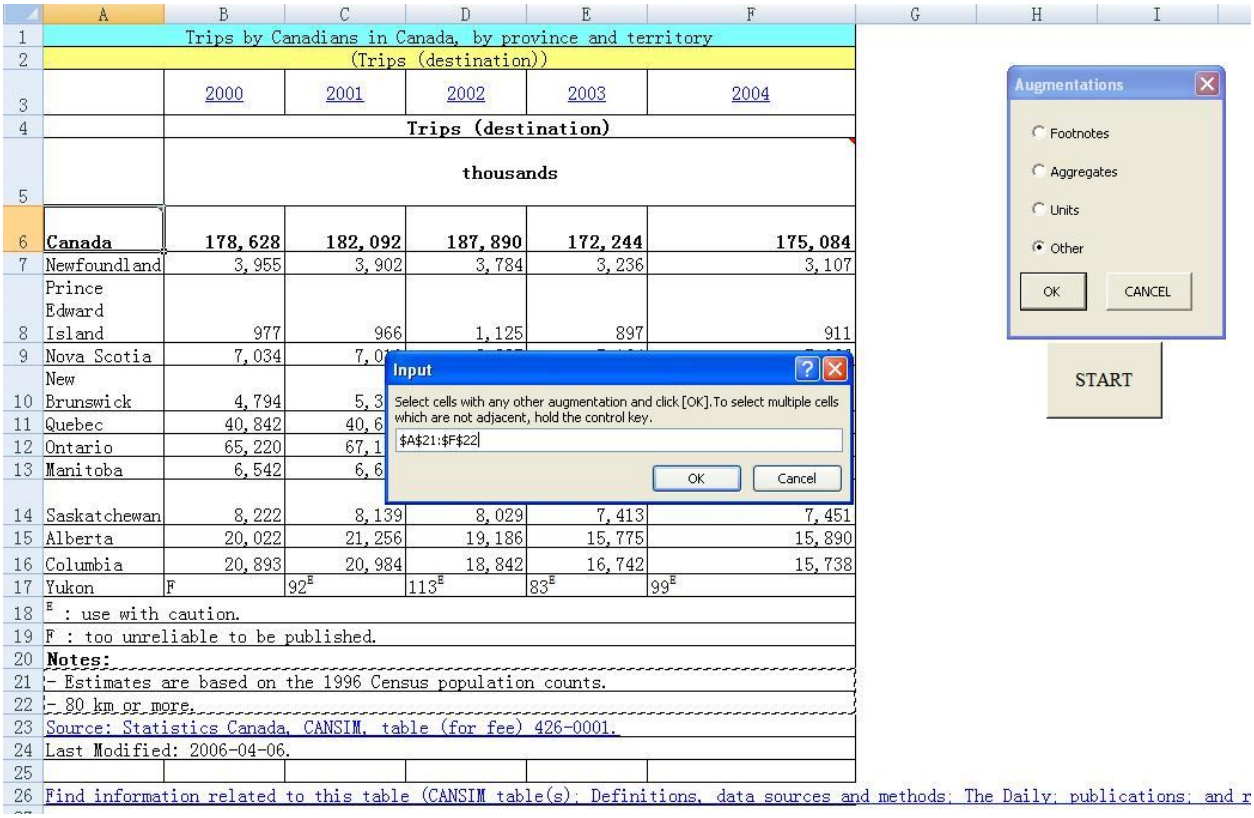


FIGURE 9. UNITS

The user selects the cell "**thousands**" and clicks the [OK] button. This causes the text (*Units)* to appear as a comment for the cell.

3.(d). Other:

The user can also specify any other kind of augmentations for any cells in the table. For example, the cells in the **rows 21 & 22** with the text– **"Estimates are based on the 1996 Census population counts" and "80 km or more"** are special notes for the table. TAT allows users to specify these kind of additional notes as well. In the Augmentations menu, the user selects the 'Other' option and clicks the [OK] button. The system then displays a pop up to the user and asks the user to select the cells with any other augmentation (**Figure 10**).

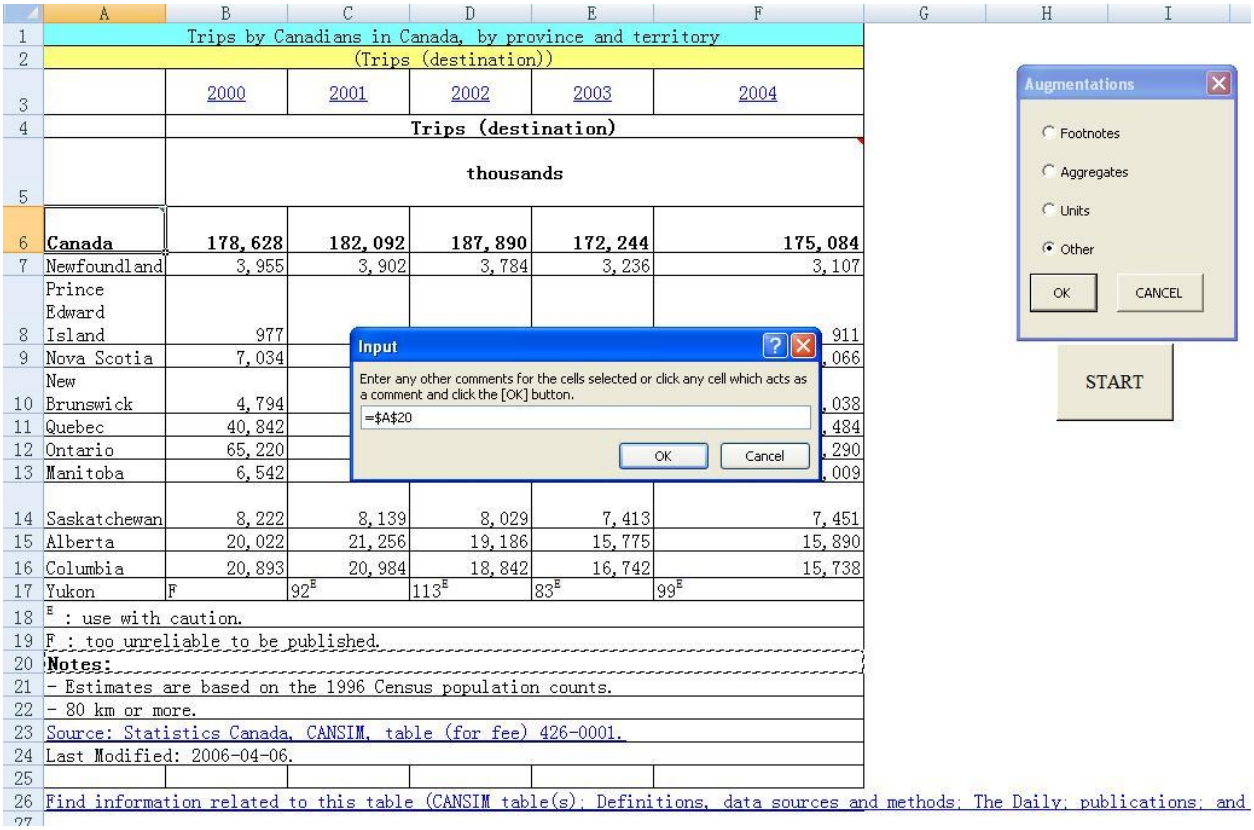


	A	B	C	D	E	F
1	Trips by Canadians in Canada, by province and territory					
2	(Trips (destination))					
3		2000	2001	2002	2003	2004
4		Trips (destination)				
5		thousands				
6	**Canada**	**178, 628**	**182, 092**	**187, 890**	**172, 244**	**175, 084**
7	Newfoundland	3, 955	3, 902	3, 784	3, 236	3, 107
8	Prince Edward Island	977	966	1, 125	897	911
9	Nova Scotia	7, 034	7, 0			
10	New Brunswick	4, 794	5, 3			
11	Quebec	40, 842	40, 6			
12	Ontario	65, 220	67, 1			
13	Manitoba	6, 542	6, 6			
14	Saskatchewan	8, 222	8, 139	8, 029	7, 413	7, 451
15	Alberta	20, 022	21, 256	19, 186	15, 775	15, 890
16	Columbia	20, 893	20, 984	18, 842	16, 742	15, 738
17	Yukon	F	92[E]	113[E]	83[E]	99[E]
18	[E] : use with caution.					
19	F : too unreliable to be published.					
20	**Notes:**					
21	- Estimates are based on the 1996 Census population counts.					
22	- 80 km or more.					
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.					
24	Last Modified: 2006-04-06.					
25						
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and r					

FIGURE 10. OTHER AUGMENTATIONS-I

The user selects the cells in the rows 21 and 22 and clicks the [OK] button. The system displays another pop up to the user asking to enter any comments for these cells or select any cells which serve as comments. In the above case, these augmentations appear below the heading "**Notes**" in row 20, so the user selects the merged cell **"A20"**. The "Other" option allows the user to enter any arbitrary comments that pertain to the whole table by selecting the entire table for the domain of "Other".

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009



	2000	2001	2002	2003	2004
Canada	**178,628**	**182,092**	**187,890**	**172,244**	**175,084**
Newfoundland	3,955	3,902	3,784	3,236	3,107
Prince Edward Island	977				911
Nova Scotia	7,034				066
New Brunswick	4,794				038
Quebec	40,842				484
Ontario	65,220				290
Manitoba	6,542				009
Saskatchewan	8,222	8,139	8,029	7,413	7,451
Alberta	20,022	21,256	19,186	15,775	15,890
Columbia	20,893	20,984	18,842	16,742	15,738
Yukon	F	92[E]	113[E]	83[E]	99[E]

[E] : use with caution.
F : too unreliable to be published.
Notes:
- Estimates are based on the 1996 Census population counts.
- 80 km or more.
Source: Statistics Canada, CANSIM, table (for fee) 426-0001.
Last Modified: 2006-04-06.

Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and

FIGURE 11. OTHER AUGMENTATIONS-II

The text: *(Other) Notes :(/Other)* appears as a comment over the cells.
It should be noted that in all the above cases, if the augmentations need to be overwritten, they have to be deleted first before adding the new augmentation (footnote, aggregate, units or other) as new augmentations are just appended to the existing ones. This can be done by right-clicking on the cell and selecting the "**Delete Comment**" option from the popup menu.

4. TABLE ANALYSIS:

This is the most important part of the program that actually *abstracts* the table. It consists of two stages: Verification (Steps 1-5) and Analysis (Steps 6-11).

4 (a) Verification:

In the first part, TAT checks if the table is well-formed or not. For this the user is first prompted to click the top-leftmost and the bottom-rightmost delta cells (which will be done automatically in future versions). The system colors the cells selected by the user in orange and does the following actions:

1. Deactivates any hyperlinks present in the cells (because of a direct copy-paste from the web page) that might cause unnecessary actions when the user clicks those cells.
2. Checks for any empty delta cell in that particular range of delta cells selected by the user. If there is an empty delta cell, the system colors the empty cell, enters **"D?"** in the cell and prompts the user to enter a value for that cell. If the user does not enter any value, the delta cell is given the value "**D?**"
3. The system performs the '*Canonicalization'* process, which splits all the merged cells in the table and repeats the text across the split cells. The canonicalization process is restricted to the category and delta cells. For the above table, the process splits the merged cells with the text "**Trips (destination)**" & "**thousands**" and repeats the value/text over the entire span. The title and caption cells are not split as they were correctly assigned before performing the table analysis. (**Figure 12**)
4. The next action is to form the *list-rows* for the column and row headers by looping through the cells ***above*** and to the ***left*** of the delta cells – categories + sub-categories. The *list-row* notation is a simple one- or two-dimensional array notation of the cells.
   The *list-rows* for the current table would be

   List-row notation for the categories above the delta cells -

   Thousands – Trips (Destination) – 2000
   Thousands – Trips (Destination) – 2001
   Thousands – Trips (Destination) – 2002
   Thousands – Trips (Destination) – 2003
   Thousands – Trips (Destination) – 2004

   List-row notation for the categories to the left of delta cells -

   Canada
   Newfoundland and Labrador
   Prince Edward Island
   Nova Scotia
   New Brunswick
   Quebec
   Ontario
   Manitoba
   Saskatchewan
   Alberta
   British Columbia
   Yukon Territory

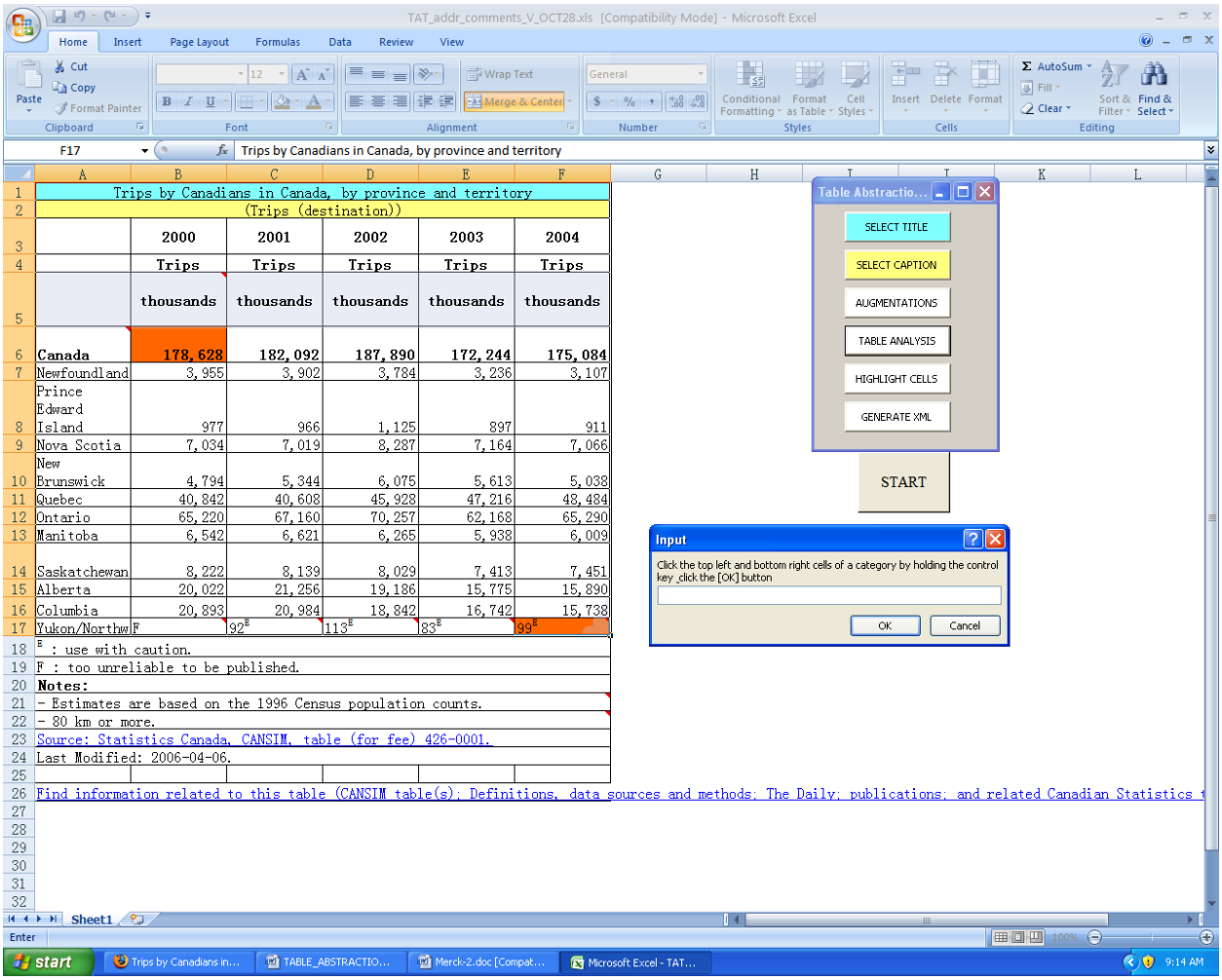


FIGURE 12. CANONICALIZED TABLE

5. The *list-rows* are then checked to determine if the table is well-formed or not. If the system determines that the table is not well-formed because of repetitions in its headers, it highlights them in red and gives an appropriate error message. In such a case, the user must correct the table by manipulating it with Excel operations and then start over from the beginning by clicking the [START] button. In the current example, the table is well-formed as there are no repetitions of rows in the *list-row* arrays. Future versions of TAT will incorporate a more robust check of the conditions for a well formed table.

4(b) Analysis:

The second stage consists of forming the *indented notation* for the categories in the table based on user input.

6. If the table is well-formed, the system prompts the user to select the top-leftmost and bottom-rightmost cells of a category. Depending on the location of the clicked category

cells, the system determines if the category has a column-based header (header above the delta cells) or row-based header (header to the left of delta cells). The system colors the category cells selected in green. The system then forms the *list-row* notation for that category alone and checks for the root of the category tree. The root is repeated through all the rows of the *list-row* array as it is canonicalized.

7. If no root is found for the category, then a virtual header is added for the category. For the above table, since in the *list row* notation for categories above the delta cells, '**Trips (Destination)**' and "**thousand**s" are repeated, either of the two can be the root for that category. For the category to the left of the delta cells, there is no repetition in the columns. Hence, it requires a virtual header. The virtual header is a unique string for every table processed and is "**VHxxx**" where xxx represents a unique number.
8. The *indented notation* for that category is formed in a separate sheet in the workbook. (**Figure 13**). Each cell in *the indented notation has* a comment which refers to the address of that cell in the original table.
9. The system then prompts asking the user if there are more categories present in the table with a Yes – No message box. If the user responds Yes then steps 6-8 are repeated.

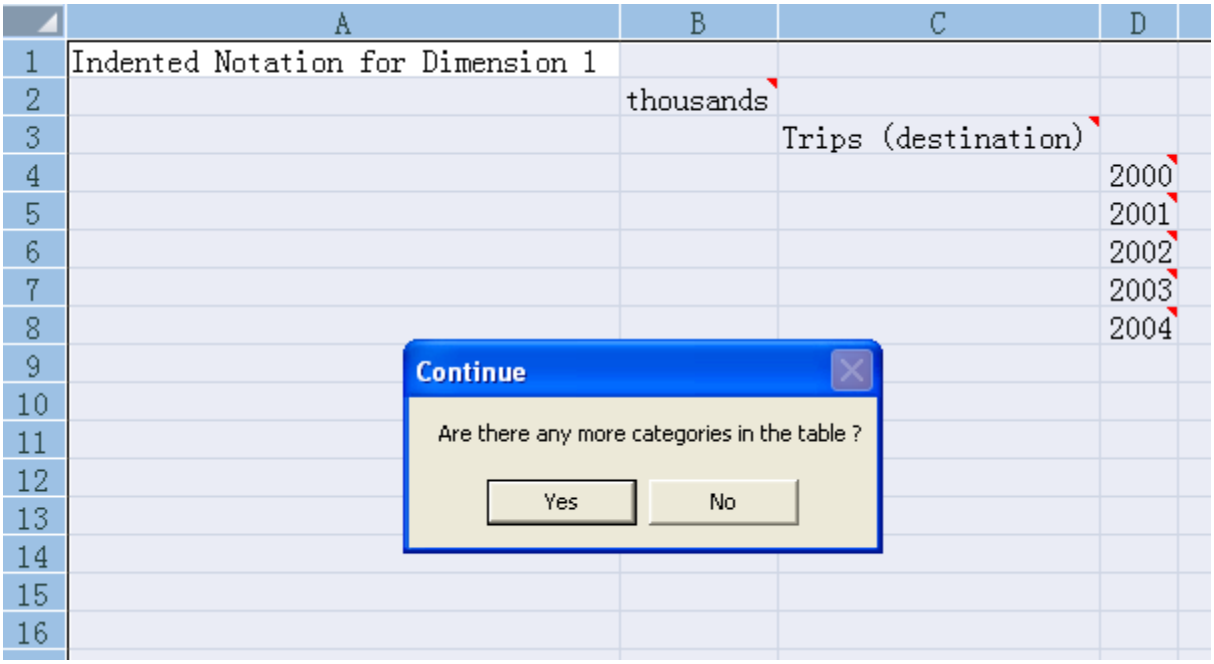


FIGURE 13. CATEGORY PROMPT

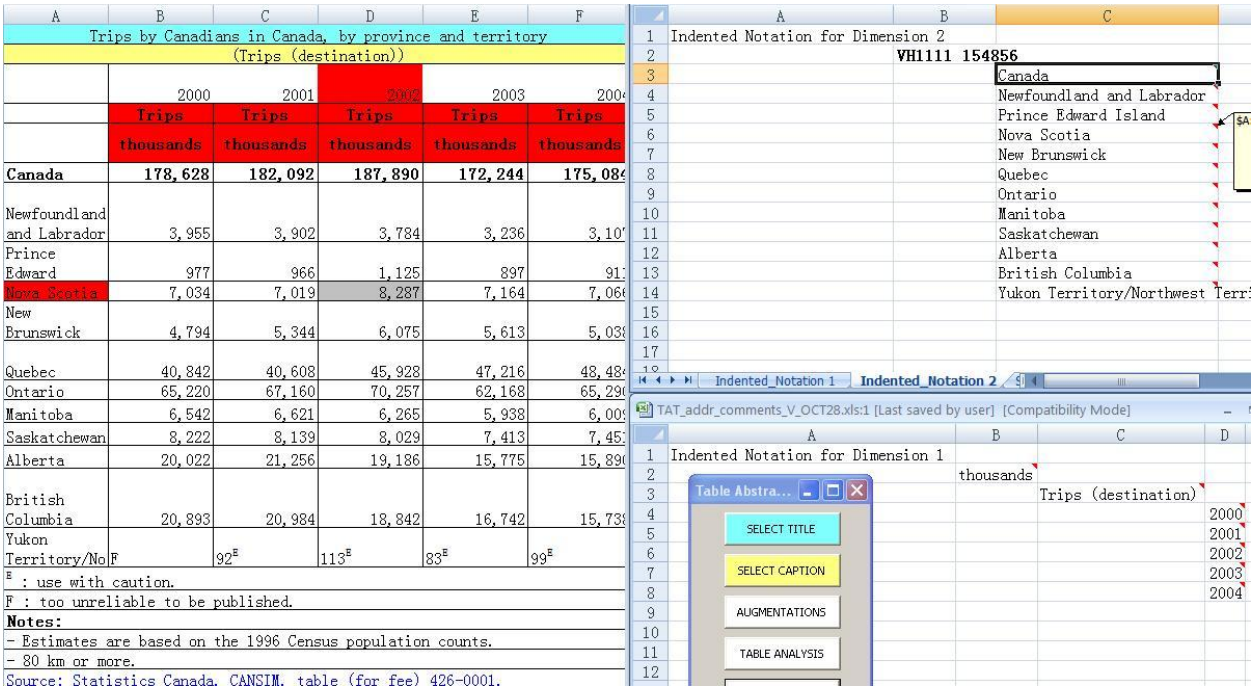
10. If the user clicks the No button, the system displays the table along with all the Indented Notation sheets arranged in tiles. The user has the chance to correct the indented notation if necessary.
11. The user can now check if the table has been interpreted correctly by looking for the category cells associated with a single delta cell and the delta cells associated with a single category or sub-category by clicking the [HIGHLIGHT CELLS] button.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

5. HIGHLIGHT CELLS & INDENTED NOTATION

5(a). Highlight Cells

After the table has been *abstracted*, the system displays the indented notation of the trees along with the table (**Figure 14**). If the user clicks the [HIGHLIGHT CELLS] button and clicks a delta cell, the system highlights all the category cells associated with it in red. The button text changes to "STOP HIGHLIGHTING".



Trips by Canadians in Canada, by province and territory
(Trips (destination))

	2000	2001	2002	2003	200
	Trips thousands	Trips thousands	Trips thousands	Trips thousands	Trips thousands
Canada	**178,628**	**182,092**	**187,890**	**172,244**	**175,08**
Newfoundland and Labrador	3,955	3,902	3,784	3,236	3,10
Prince Edward	977	966	1,125	897	91
Nova Scotia	7,034	7,019	8,287	7,164	7,06
New Brunswick	4,794	5,344	6,075	5,613	5,03
Quebec	40,842	40,608	45,928	47,216	48,48
Ontario	65,220	67,160	70,257	62,168	65,29
Manitoba	6,542	6,621	6,265	5,938	6,00
Saskatchewan	8,222	8,139	8,029	7,413	7,45
Alberta	20,022	21,256	19,186	15,775	15,89
British Columbia	20,893	20,984	18,842	16,742	15,73
Yukon Territory/No	F	92[E]	113[E]	83[E]	99[E]

[E] : use with caution.
F : too unreliable to be published.
Notes:
- Estimates are based on the 1996 Census population counts.
- 80 km or more.
Source: Statistics Canada. CANSIM. table (for fee) 426-0001.

1 Indented Notation for Dimension 2
2 VH1111 154856
3 Canada
4 Newfoundland and Labrador
5 Prince Edward Island
6 Nova Scotia
7 New Brunswick
8 Quebec
9 Ontario
10 Manitoba
11 Saskatchewan
12 Alberta
13 British Columbia
14 Yukon Territory/Northwest Terr

Indented_Notation 1 | Indented_Notation 2

TAT_addr_comments_V_OCT28.xls:1 [Last saved by user] [Compatibility Mode]

1 Indented Notation for Dimension 1
2 thousands
3 Trips (destination)
4 2000
5 2001
6 2002
7 2003
8 2004

Table Abstra...
SELECT TITLE
SELECT CAPTION
AUGMENTATIONS
TABLE ANALYSIS

FIGURE 14. TAT HIGHLIGHTING CATEGORY CELLS

In **Figure 14**, the system highlights all the category cells associated with a delta cell in red and highlights the delta cell selected itself in gray. The Indented Notations for the above table are also arranged in the form of tiles as shown in the above figure.

On clicking a category/sub-category cell, the system highlights the entire set of delta cells associated with it (**Figure 15)**. In Figure 15, the user selected sub-category **2001.**

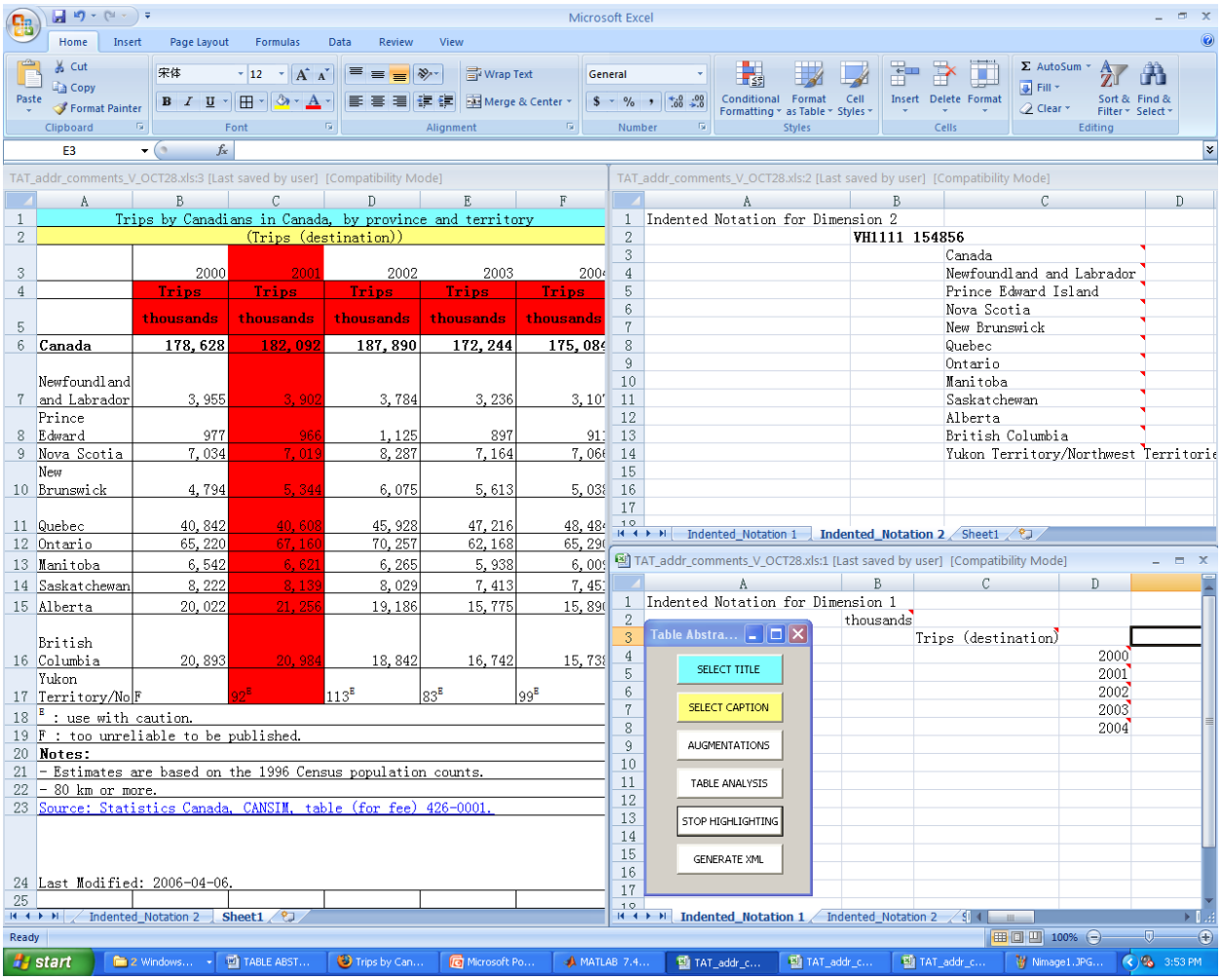


FIGURE 15. TAT HIGHLIGHTING DELTA CELLS ASSOCIATED WITH A CATEGORY

The system can also highlight a specific delta cell covered by two categories. If the user clicks a column-based category/sub-category cell followed by a row-based category/sub-category cell, the system highlights a specific delta cell covered by the categories (**Figure 16)**. The same cell is highlighted even if the user selects the row-based category/sub-category cell first followed by the column-based category/sub-category cell.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

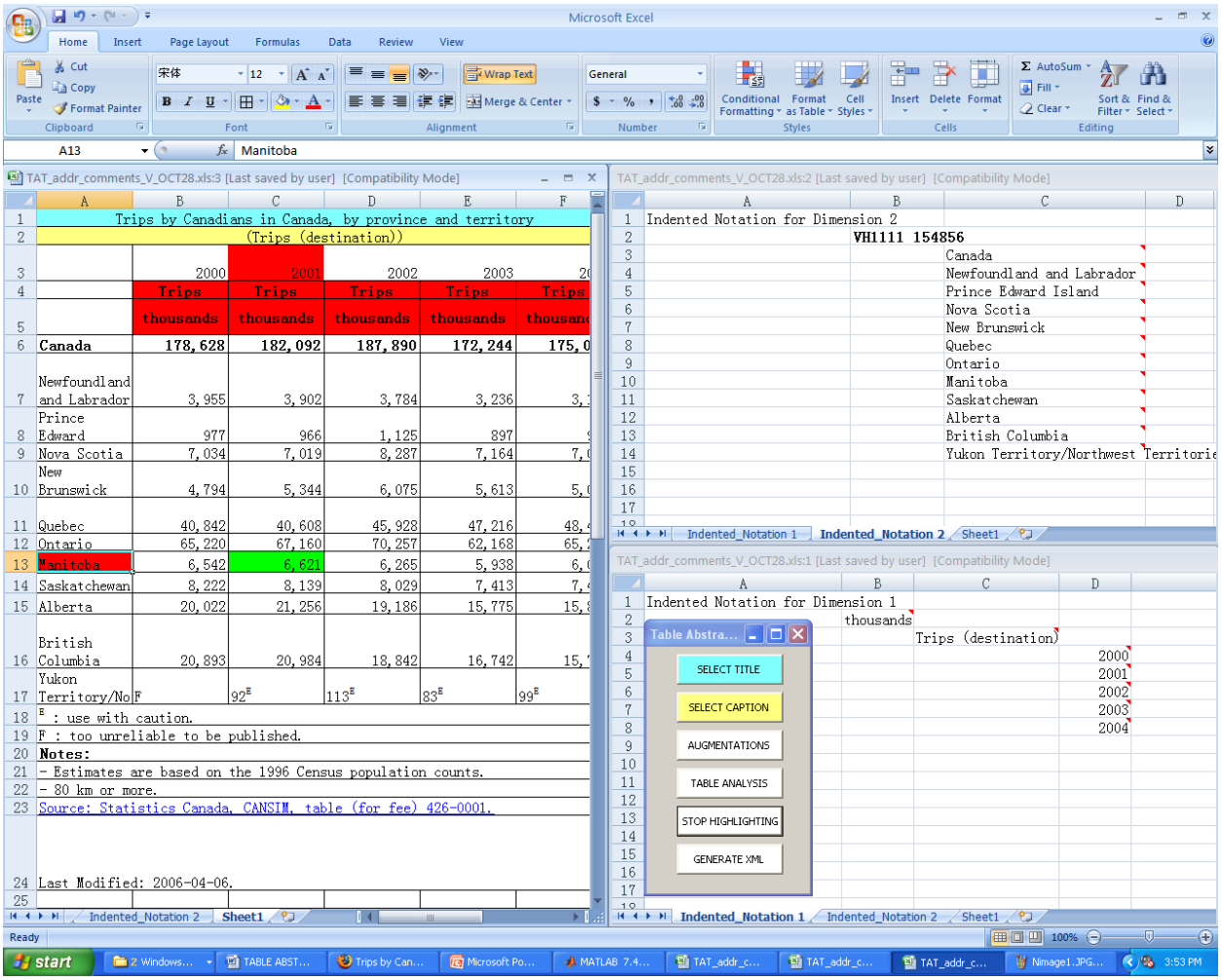


FIGURE 16. TAT HIGHLIGHTING ONE DELTACELL ASSOCIATED WITH 2 CATEGORIES

In the above case, the user chose the column-based sub-category '**2001**' and the horizontal sub-category '**Manitoba**'. The delta cell associated with these sub-categories is highlighted in green. To stop the highlighting process, the user should click the [STOP HIGHLIGHTING] button.

5 (b) Indented Notation

The advantage of TAT over WNT is that the corrections to the indented notation can be made by directly using Excel operations instead of MATLAB. The user can make all the changes which were allowed in WNT. The user can re-arrange the cells or delete any empty rows or columns between the cells. An important point is that the highlighting is done looking at the indented notation. So, if the user makes any changes to the contents of the indented notation with text that is not present in the table, then the system displays an error message (**Figure 17)**.

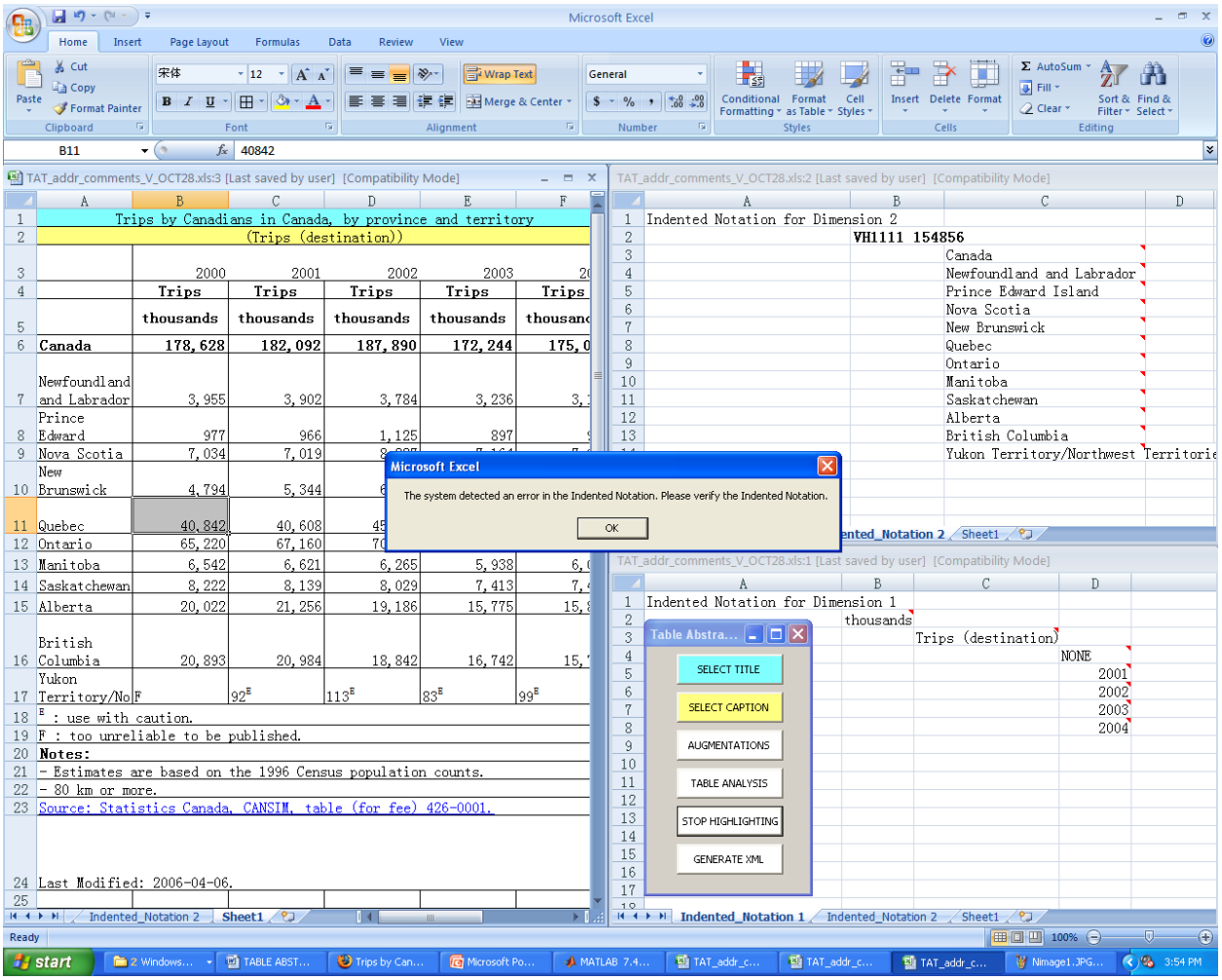


FIGURE 17. TAT INDENTED NOTATION ERROR

In the above screenshot it can be seen in the bottom right corner i.e. Indented Notation for Dimension 1 that the year "**2000"** has been changed to "**NONE".** When a delta cell –"B11" corresponding to the subcategory "**2000**" is selected to view the highlighting, the system displays an appropriate error message. The error can be corrected by changing the corresponding text in the table i.e., "**2000**" in cell "**B3**" to "**NONE**" or by changing the text in cell "**D4**" in the sheet "**Indented_Notation_1**" to "**2000**" again.

6. GENERATE XML:

Clicking this button generates the XML notation for the table based on its indented notation. The XML filename contains the time stamp of the process. TAT creates the XML file in the same folder as the Excel Workbook. TAT also creates a log file with the name containing the time stamp of the process in the same folder. It records the time taken by the user for each action and the idle time between successive actions. To change the path, please look at the *createxml* & *Writelog* procedures in the source code, which can be viewed in the Visual Basic Editor.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

III. TRANSFORMATIONS ON TABLES FOR PROCESSING:

There are three most commonly occurring templates of table layouts which TAT cannot process. They are shown below:

Template 1:

	A1		A2	
Stub	B1	B2	B1	B2
X				
C1	XX	XX	XX	XX
C2	XX	XX	XX	XX
Y				
C1	XX	XX	XX	XX
C2	XX	XX	XX	XX

Table 1

Table 1 is a 4 category table:

Category 1- Virtual header1 (root) – A1, A2
Category 2- Virtual header2 (root) – B1, B2
Category 3- Virtual header3 (root) – X, Y
Category 3- Virtual header4 (root) – C1, C2

This template cannot be processed by TAT as X and Y which are sub-categories of the same category are not contiguous and have sub-categories of another category between them.

Template 2:

	A1		A2	
Stub	B1	B2	B1	B2
X	XX	XX	XX	XX
C1	XX	XX	XX	XX
C2	XX	XX	XX	XX
Y	XX	XX	XX	XX
C3	XX	XX	XX	XX
C4	XX	XX	XX	XX

Table 2 is a 3 category table.
Category 1- Virtual header1 (root) : A1, A2

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009
Category 2- Virtual header2 (root) : B1, B2
Category 3 -Virtual header3 (root) : X: X, C1, C2; Y:Y,C3,C4
Note: In category 3, X and Y are sub-categories which have their own sub-categories. Also, generally, X and Y are aggregates.

In Table 2, we need a virtual header for X and Y. Also, X is the parent node of C1 and C2 while Y is the parent node of C3 and C4. This relationship between the sub-categories is expressed in the form of format changes like bold and italics which are not preserved by Microsoft Excel cannot be interpreted by TAT correctly.

Template 3:

	A1				
Stub	B1	B2	B3	B4	
X					
C1	XX	XX	XX	XX	
C2	XX	XX	XX	XX	Table 1
Y					
C3	XX	XX	XX	XX	
C4	XX	XX	XX	XX	Table 2

Table 3

Table 3 is actually a concatenation of two tables. This is also a very commonly found template for tables where related tables are concatenated for presenting a complete picture. However, logically these tables should be defined as two (or more) separate tables.

In order to make these tables '*TAT-friendly*', a few transformations on the tables need to be performed. These transformations preserve the logical structure of the table. The following section illustrates an actual table of **Template 1** being transformed using Excel operations.

Financial statistics for enterprises (quarterly)

	2nd quarter 2007[r]	1st quarter 2008[r]	2nd quarter 2008[p]	1st quarter 2008 to 2nd quarter 2008
				seasonally adjusted
		$ millions		%
All industries				
Operating revenue	751,104	777,204	**788,410**	1.4
Operating profit	66,618	67,721	**69,388**	2.5
After-tax profit	44,396	45,098	**47,991**	6.4
Non-financial industries				
Operating revenue	676,716	698,830	**712,278**	1.9
Operating profit	46,272	48,642	**50,631**	4.1
After-tax profit	31,114	33,285	**36,280**	9.0
Finance and insurance industries				
Operating revenue	74,388	78,375	**76,131**	-2.9
Operating profit	20,346	19,079	**18,758**	-1.7
After-tax profit	13,281	11,812	**11,711**	-0.9

[p] : preliminary.
[r] : revised.
Note: These quarterly statistics cover the activities of all corporations in Canada, excluding government controlled and not-for-profit corp
Source: Statistics Canada, CANSIM, table (for fee) 187-0002 and Catalogue no. 61-008-X.
Last Modified: 2008-08-21.

FIGURE 18. TABLE REQUIRING TRANSFORMATIONS TO MAKE IT 'TAT-FRIENDLY'

This is a three category table with Types of Industries (All Industries, Non-financial industries and Finance and insurance industries), Financial Statistics (Operating Revenue, Operating Profit and After-Tax profit) and the period (2nd quarter 2007, 1st quarter 2008 etc.) forming the three categories.

Raghav Padmanabhan – raghav.krishna@gmail.com
DocLab, RPI
2/3/2009

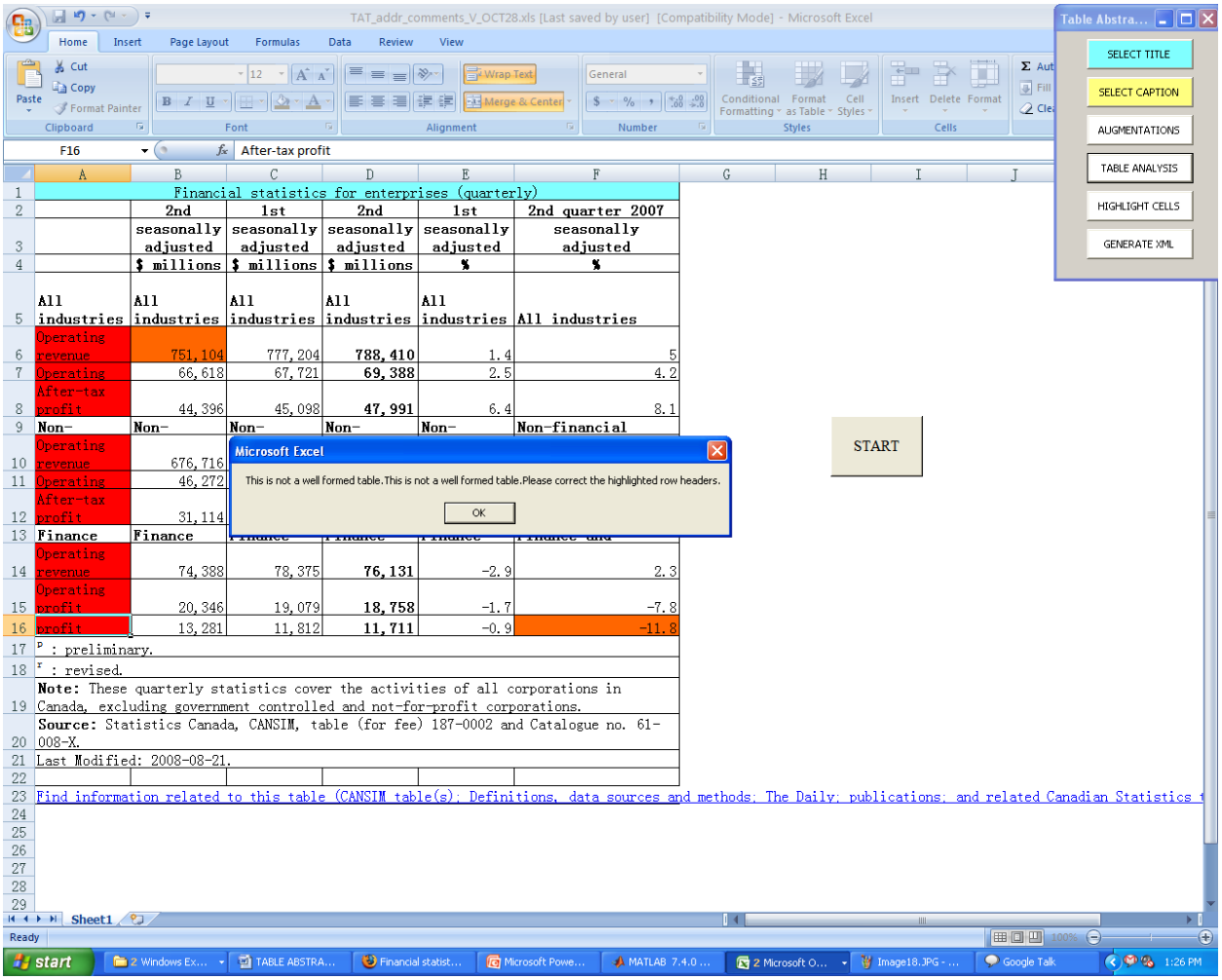


FIGURE 19. TAT BAD TABLE ERROR

In the above table, the cells Operating Revenue, Operating Profit and After-Tax profit are repeated under the type of industries and hence TAT assumes that it is a bad table, as shown by the error message (**Figure 19**). This table has to be modified to the form (**Figure 20)** before it can be processes by TAT.

	A	B	C	D	E	F	G
1		Financial statistics for enterprises (quarterly)					
2			**2nd**	**1st**	**2nd**	**1st**	**2nd**
3			**seasonally adjusted**				
4			**$ millions**			**%**	
5	**All industries**	Operating revenue	751,104	777,204	**788,410**	1.4	5
6		Operating	66,618	67,721	**69,388**	2.5	4.2
7		After-tax profit	44,396	45,098	**47,991**	6.4	8.1
8	**Non-financial industries**	Operating revenue	676,716	698,830	**712,278**	1.9	5.3
9		Operating	46,272	48,642	**50,631**	4.1	9.4
10		After-tax profit	31,114	33,285	**36,280**	9	16.6
11	**Finance and insurance industries**	Operating revenue	74,388	78,375	**76,131**	-2.9	2.3
12		Operating profit	20,346	19,079	**18,758**	-1.7	-7.8
13		profit	13,281	11,812	**11,711**	-0.9	-11.8
14		p : preliminary.					
15		r : revised.					
16		**Note:** These quarterly statistics cover the activities of all corporations in Canada, excluding government controlled and not-for-profit corporations.					
17		**Source:** Statistics Canada, CANSIM, table (for fee) 187-0002 and Catalogue no. 61-008-X.					
18		Last Modified: 2008-08-21.					
19							
20		Find information related to this table (CANSIM table(s); Definitions, data sourc					
21							

FIGURE 20. TRANSFORMED 'TAT-FRIENDLY' TABLE

Even though the statistics repeat in a column in this table, they are preceded by different industries, so there are no duplicate paths. TAT regards this type of a composition for a table as well formed. TAT will add virtual headers to both row categories.

Please note that the same operations are required for tables in **Template 2** form.
For tables in the form shown in **Template 3,** it is best to separately process the individual tables.