Increasing Efficiency in Generating and Comparing National Travel Behavior Estimates over Time

Anthony Fucci and Alex Cates, Westat

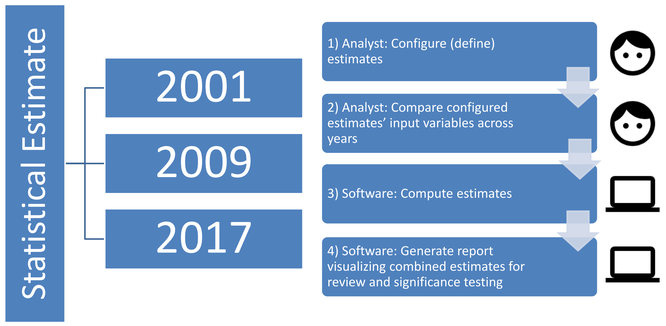
Background

Generating estimates to make statistically representative claims about the population from a National Household Travel Survey (NHTS) dataset is typically done with statistics-focused software, usually within SAS, SPSS, Stata or R. Like other complex sample surveys, a significant task the software assists one with is correctly performing variance estimation for a desired statistic, needed to make statistically significant claims about the population. This presentation will demonstrate a framework using R software that simplifies generating national travel estimates across the three most recent NHTS programs, covering 2001, 2009 and 2017. The principal aim is to increase efficiency and quality of travel trend analysis by reducing programming burden to increase focus on exploring statistics and interpreting data. The presentation will explore automating reports with side-by-side faceting of multi-year estimates in tables, charts and maps, and it will automate some summarization of the differences between the NHTS years of interest and the provided estimate. The resulting report will demonstrate the method's potential for increasing our ability to study national travel behavior over time using a case example analysis topic: national regional geographies and travel behavior to access food.

Method Implemented

The following graphic is a representation of the method broken into four high level processes. Notice that the first two processes are human derived and the subsequent two are computer/software derived.

Figure 1. High Level Representation of Method



1) Configure Estimates

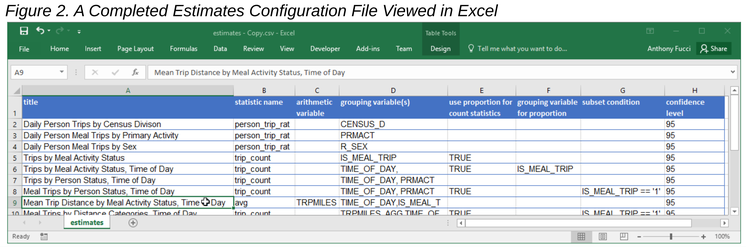
The first step is to develop a list of configuration parameters that can be organized and understood by software designed for generating statistical estimates. A text (.csv) file with descriptively-named fields was developed to serve as the analyst's estimate configuration file. Table 1 lists the parameters of an estimate to be completed by the analyst.

Table 1. Components that Define a Statistic for Analysis

Item	Description
Title	The title of the configured statistic being defined
Base Statistic	Desired base statistic: count, proportion, mean, sum, median
Variable	Variable to use for numeric statistic, if selected
Grouping Variable(s)	Variable(s) by which to group the statistic
Proportion	Display count statistic as percent, if selected
Proportion Variable	Grouping variable to compute proportion over
Subset	Include or exclude certain observations
Confidence Level	Probability to compute an interval estimate from the standard error

Below, Figure 2 demonstrates the text (.csv) configuration file being completed in Microsoft Excel. Each row in the file is an estimate that will be generated for 2001, 2009 and 2017. The resulting estimates will be visualized and summarized in the final automated report.

Figure 2. A Completed Estimates Configuration File Viewed in Excel



2) Compare Estimate Inputs Across Years

Analysts have to assume caution when duplicating a query over multiple datasets. While the analyst configures their desired estimate, they should review the data and metadata of the estimate's input variables. Table 2 demonstrates an analysis variable that defines whether a trip is a meal trip or not. Take note of 2017, where differences in the survey require one to normalize a new variable that respects those differences.

Table 2. Example of Normalizing an Input Analysis Variable across Years

YEAR	NAME	TABLE	TYPE	DOMAIN	VALUE	LABEL
2001	IS_MEAL_TRIP	trip	character	WHYTO == '82'	1	Yes
2001	IS_MEAL_TRIP	trip	character	WHYTO != '82'	2	No
2009	IS_MEAL_TRIP	trip	character	WHYTO == '82'	1	Yes
2009	IS_MEAL_TRIP	trip	character	WHYTO != '82'	2	No
2017	IS_MEAL_TRIP	trip	character	**WHYTO == '13'**	1	Yes
2017	IS_MEAL_TRIP	trip	character	**WHYTO != '13'**	2	No

3) Compute Estimates

We can now develop a software script that is able to interpret the configuration files we created in steps 1 and 2, process the estimates, and organize the results for rendering in an automated report. We will rely on the "summarizeNHTS" R software package because it consolidates high level data processing items required for generating common estimates with NHTS data. Having the ability to parameterize the study year for estimates is greatly simplified because there is available open source software that has functions and methods for understanding the items in Table 3.

Table 3. High Level Components Required for Generating Estimates

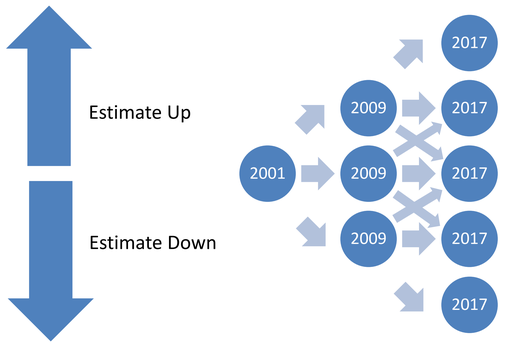
Item	Description
Core data	Data files with survey variables
Replicate weights	Data files with replicated sample weights needed to compute error
Standard error formula	The formula needed to compute error according to the weighting method
Analysis variables	Normalized variable definitions for comparing across programs
Statistical query	The definition of the statistic
Variable metadata	Meanings of coded values
Visualization object	Visualization preference for statistic

Once the results for each year's estimate is stashed, we compute 95% confidence interval estimates to measure and visualize significance from one year to the next in our report.

4) Generate Report

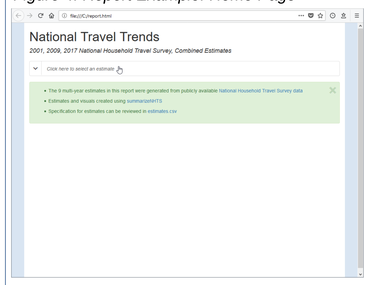
We will use the open source R software package "rmarkdown" to render an HTML document that visualizes the results of the configured estimates. The markdown syntax framework allows one to write text, tables and figures to an input text file which can then be translated (its input content rendered) to an output file (in this case, a .html file). We decide how we want to present the results for each multi-year estimate, write a function that creates that desired presentation input text using understood markdown syntax, and then loop through every estimate to create the complete input text that renders the desired content presentation. JavaScript and CSS are used in addition to HTML to give the report a dynamic page-navigation effect. A dropdown menu at the top of the report is how the analyst selects which multi-year estimate they would like to review. There are two main components being presented in the resulting content for each estimate. First, the results for each year are presented in independent bar charts, stacked side by side for comparison. Second, the report includes a trend significance table for each estimate. The trend significance table looks like a typical crosstab table but in each cell there is a nested line chart instead of a value. The line chart represents the trend of the requested estimate, from 2001 to 2017, using 95% confidence interval estimates. One can hover over the line chart with their cursor to view the estimate value for each year. This is done so that one can quickly visualize statistical significance across 3 survey years. Figure 3 demonstrates the nine possible trends an estimate can take that will be visualized in the trend significance table.

Figure 3. Representation of the Possible Trends of an Estimate from 2001 to 2017



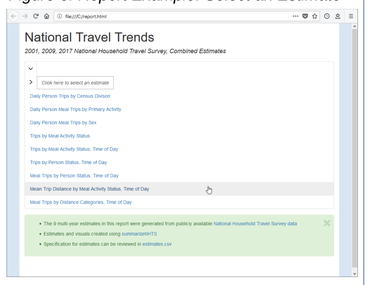
Results

Figure 4. Report Example: Home Page



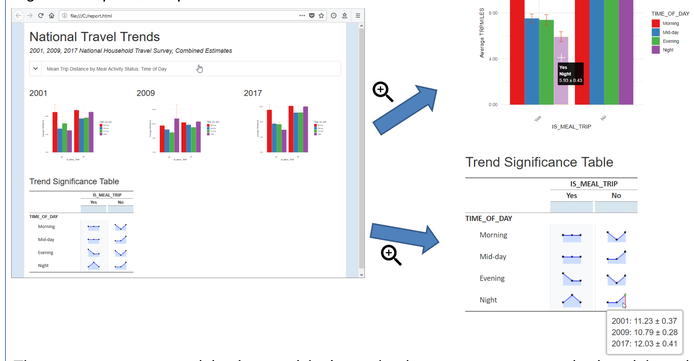
Resulting report is a single file for mobile or desktop web browsers. A message box on the home page provides the number of estimates and a downloadable link to the input configuration file.

Figure 5. Report Example: Select an Estimate



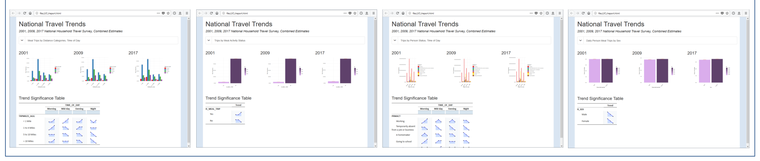
Multi-year estimate results are available in just two clicks. Notice how the estimate titles match those of the input configuration file.

Figure 6. Report Example: Multi-Year Results



The report may appear minimal at a quick glance, but in one page we are reviewing eight estimates across 3 years. Below the year-level charts, the 24 estimates in all are reduced into a single table where significant changes in 95% confidence interval estimates from 2001 to 2017 are visualized in miniature line charts. Let's summarize how we reached the point of consolidating this information into this one page using the figure 6 example estimate titled "Mean Trip Distance by Meal Activity Status, Time of Day." For each of the three NHTS programs, the software translated the estimate configuration parameters we specified in figure 2. Figure 2 shows we only specified five parameters: a title, statistic name, arithmetic variable, grouping variables and confidence level. This request resulted in eight individual estimates, as seen by counting the number of bars in a chart for a given year. 95% confidence intervals were then computed for each year's estimate so that a metric for significance using 2001 as a baseline could be computed for visualization. The resulting estimates and confidence metric were then processed by visualization functions for rendering into the using 2001 as a baseline report. Figure 7 shows the report results for other estimates to further demonstrate how much multi-year analysis potential there is with little to no programming.

Figure 7. Other Multi-Year Estimates Included in the Example Report



Conclusions and Future Research

Timeliness and efficiency of analysis is a regular challenge that is further complicated when looking across multiple survey program years. The method implemented in this presentation removes the data analysis bureaucracy of loading files, merging tables, labelling values, and programming statistics. Analysts focus on only two configuration files: 1) where they define the statistic they would like estimated; 2) where they normalize variable codes across years for comparability. By simply altering a single-field parameter to break apart an analysis group or change an aggregation level or statistic, transportation analysts are encouraged by automation to look at an increasing number of statistics to expand their knowledge in their topic of interest. Further research should build upon this framework to focus even more attention on new methods for comparing travel estimates over time.

Contact Anthony Fucci
Email: AnthonyFucci@westat.com
Phone: (404) 596-7955

Alex Cates
Email: AlexanderCates@Westat.com
Phone: (404) 596-8479