Insights into Population Segmentation Using the InfAlign Map

David Fan[1], Kenneth Blake[2], Jason Reineke[2], Robert Wyatt[2]
[1]University of Minnesota, 1445 Gortner Avenue, Saint Paul, MN 55108
[2]Middle Tennessee State University, P.O. Box 64. Murfreesboro, TN 37132

Abstract

This paper shows an enhanced version of the InfAlign map that can be used for a visual exploration of relationships among tens of variables for a survey population of several hundred respondents. The map provides not only all the values for the variables examined but also the geographic location of each person. The map is interactive in allowing the user to see relationships among individuals by simply dragging a computer mouse to select and cluster individuals of interest. The capabilities of the InfAlign map are discussed along with comparisons to some other methods such as decision trees and heat maps.

Key Words: Clusters, graphical, visual, surveys, maps, political

1. Introduction and Data

A Middle Tennessee State University poll was conducted in February, 2010 for Tennessee residents with 634 respondents. The poll included a question asking whether the respondent was a Tea Party member (variable tpmember). "The Tea Party movement was a fiscally-conservative socio-political movement that emerged in the United States in 2009 through a series of locally and nationally coordinated protests." (Wikipedia, 2010).

A common approach to interpreting data from this type of question is to identify the types of people likely to be Tea Party members.

One method for such population segmentation is to use classification trees in statistical software packages based on CHAID (*Chi*-squared Automatic Interaction Detector, Kass, 1980) and related algorithms. Such trees are inherently hierarchical. That is, if approval of the president (variable apprez) is the strongest predictor of tpmember, then apprez is the first branch in the tree. After that, the algorithm searches for the predictors of apprez and no longer for variables like gender (variable gender) that can also predict tpmember well but not quite as well as apprez.

Instead, a tree algorithm would go on to look for predictors of apprez, the first branch in the tree. However, a user can explore a variable aside from apprez as a direct predictor of tpmember by manually forcing that variable as the first branch of the tree. Doing so makes the method not purely automatic but introduces a manual component. The user needs to navigate and interpret a number computer screens for each variable tested as the first branch.

The impact of different predictive variables can be explored more rapidly using the graphical InfAlign map (Fan and Fan, 2009). This map has been enhanced and combined with a geographical map to show where respondents with particular characteristics are likely to be located as well as the reverse, namely the types of people found in specific geographic regions.

2. InfAlign Map

In an InfAlign map (Figure 1, lower portion), all respondents can be considered to stand side by side with each respondent holding a flagpole with a series of flags from top to bottom. Each flag is shown on the map as a vertical blue bar with the height giving the value of the response.

To see the values for a particular respondent, the user hovers a red computer cursor over the vertical line representing that respondent. The red cursor line overlaps all the blue bars for that person. The response values of the indicted respondent are then provided both by the heights of the blue bars and the legend at the top of the map.

In Figure 1, the legend was "ID=161 A:1 B:2 C:4 D:2 E:2 F:3: G:12 H:2 I:9 J:-0.01 +". Reading this legend from left to right, the user sees that the respondent had unique identifier 161 as indicated by ID=161. After this information, the value for each variable follows the initial letter in the legend at the left of the map. Thus A:1 in the top legend was for variable "A: $tppopnow_$" with value 1; B:2 was for variable "B: $tpmember_$" with value 2 and so on through the map. The last variable was "J: $AGE_$" with value -0.01 to indicate that the respondent did not provide a valid age for whatever reason.

The top legend ended with a plus sign to indicate that the user could further display a short text string with additional information about the respondent so long as the string did not overflow the space for the legend. As an example, the user might provide the text of a short answer to an open ended question.

The heights of all the blue bars for respondent 161 corresponded to the values in the legend. The maximum height of the blue bars in any row was the highest value among all the responses for that row.

The width of each bar is narrow enough that all 634 respondents could be displayed on a single map. The map was somewhat wider than the computer screen shown in Figure 1 but the user could scroll the screen to the right to see the entire map. Obviously, the size of the map displayed depends on the screen. With ten variables and 634 respondents, the InfAlign map contained 6340 variable values.

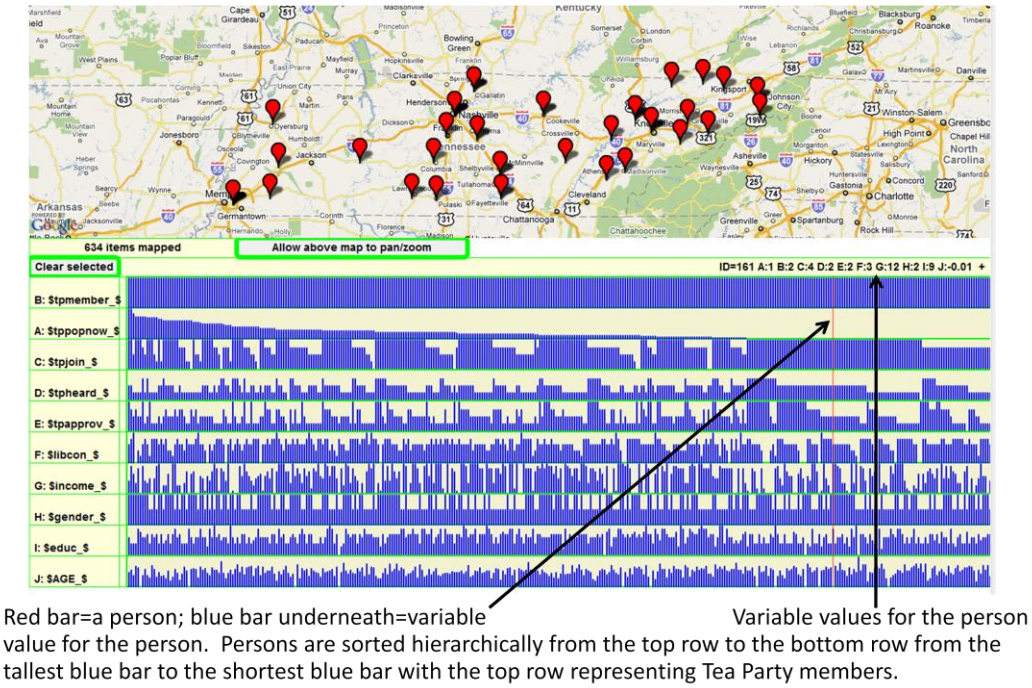
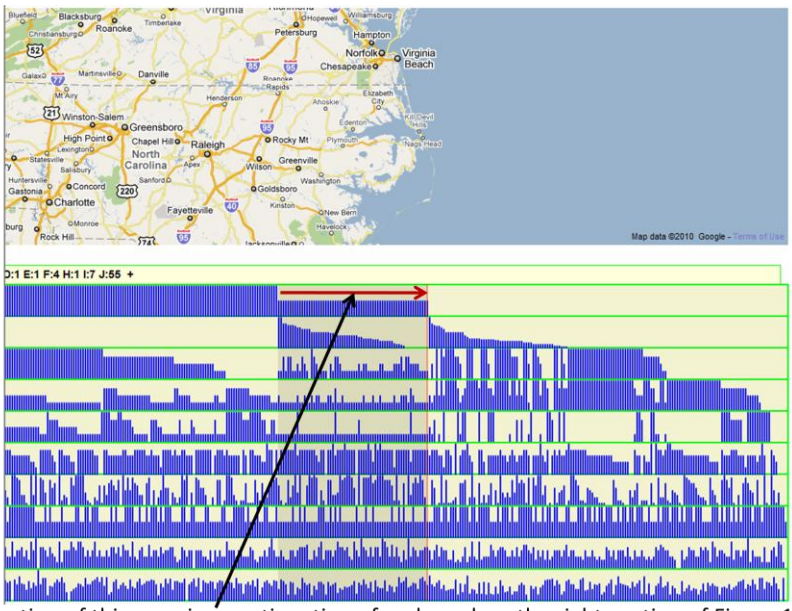


Figure 1: The lower portion of the figure shows an InfAlign map with the height of the blue bars indicating the values of all the mapped variables (legend on the left) for each respondent. The values for a particular respondent are indicated by a red bar that tracks the computer mouse that a user can move across a computer screen. The actual values of all variables for a red bar are written in the legend above the InfAlign map. The geographic map has red markers indicating the locations of a random sample of the mapped individuals.

The vertical lines for the respondents were sorted hierarchically from left to right and from top row to bottom row.

In this way, values for the variable in the top row decrease in order from tall to short. Then, in the row below, the order of bars is again from tall to short within any zone where the bars are of the same height in the row above.

The top portion of Figure 1 is a geographical map with red markers corresponding to the locations of a random set of the persons on the InfAlign map with resolution at the level of the county. A random set is selected because the geographical map can become too cluttered when too many markers is displayed.



The left portion of this page is a continuation of and overlaps the right portion of Figure 1 that was too wide to fit on the screen. The user highlights Tea Party members by mouse down at the red arrow tail, mouse drag right to highlight (indicated by dark vertical band), and mouse up at the red arrow head.

FIGURE 2: Selection of individuals on the InfAlign map. This figure is the map in Figure 1 scrolled to the right because the map was too wide to fit on the user's computer screen. The user dragged the mouse to select the individuals in the darkened region indicated by the red arrow.

Figure 2 displays the InfAlign map of Figure 1 scrolled to the right to see the right portion of the map that was off the computer screen of Figure 1. As indicated by the horizontal red arrow, the user used a sequence of steps of mouse down at the tail of the red arrow, drag to the right in the direction of the red arrow, and mouse up at the arrow head. The dragging selected and highlighted the dark vertical zone in Figure 2.

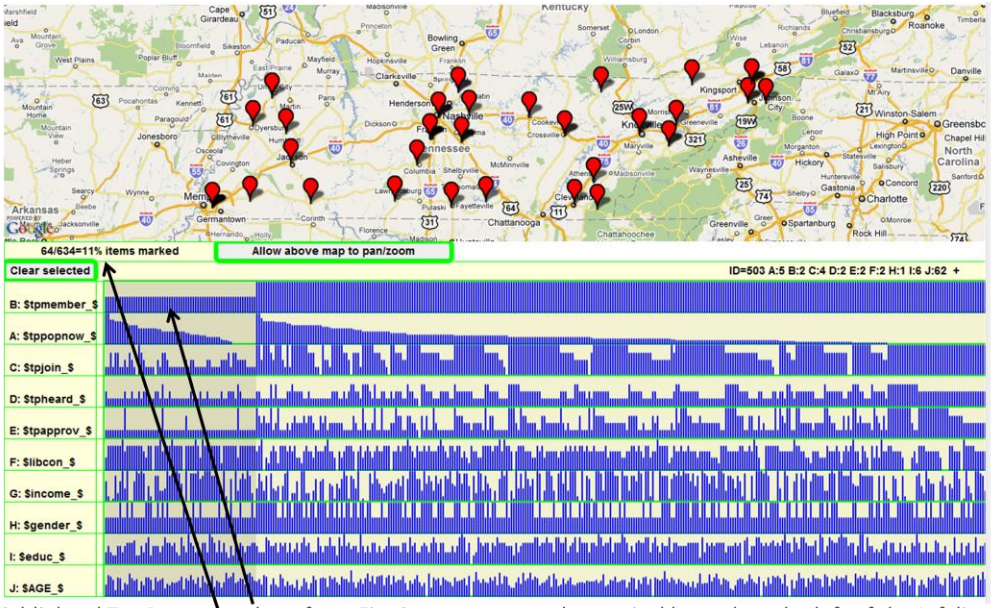


Figure 3: Repositioning of selected individuals to the left in the InfAlign map. The highlighted individuals, the Tea Party members in Figure 2, automatically moved to the left of the map after the selection process of Figure 2.

Figure 3 shows that the highlighted bars in Fig. 2 moved to the left to be next to the legend in the InfAlign display. The text on the left hand side just below the geographic map shows that 64 of the 634 members were selected in Figure 2 and had moved to the left in the map in Figure 3.

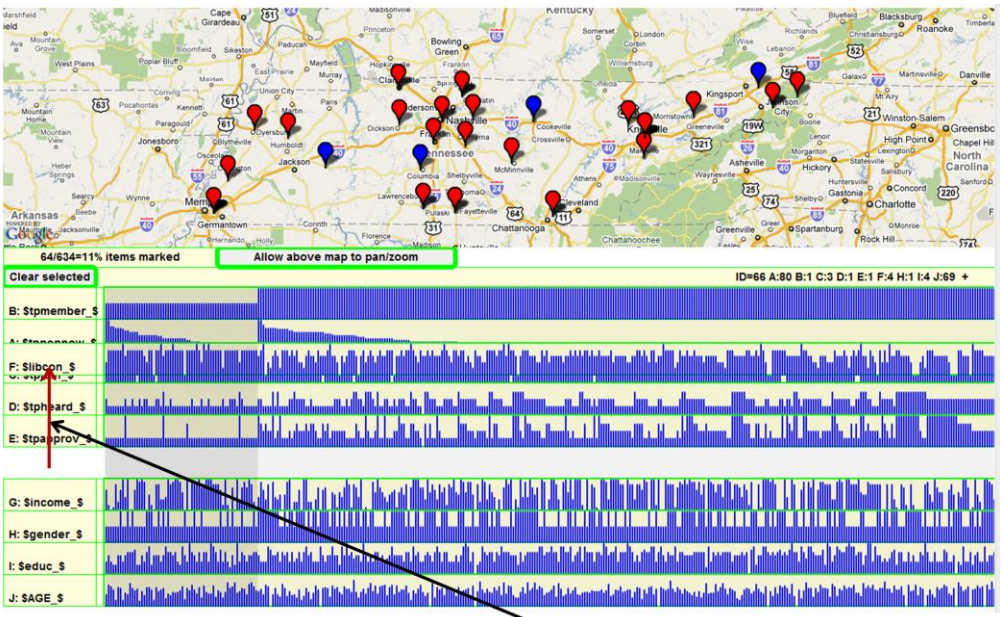


FIGURE 4: Reordering of the rows in the InfAlign map. The user placed the computer cursor over the legend on the left to drag row "F: $libcon_$" upward to the position of the row below "B: $tpmember_$"

Figure 4 shows Figure 3 with row "F: $libcon_$ dragged to the row below row "B: $tpmember_$" using the sequence of steps of mouse down, mouse drag in the direction of the top of the map, and mouse up. The geographical locations of a random sample of persons represented in the InfAlign map are indicated by red and blue markers on the geographical map in the top portion of Figure 4. The red markers correspond to non-highlighted persons. The blue markers correspond to highlighted persons.

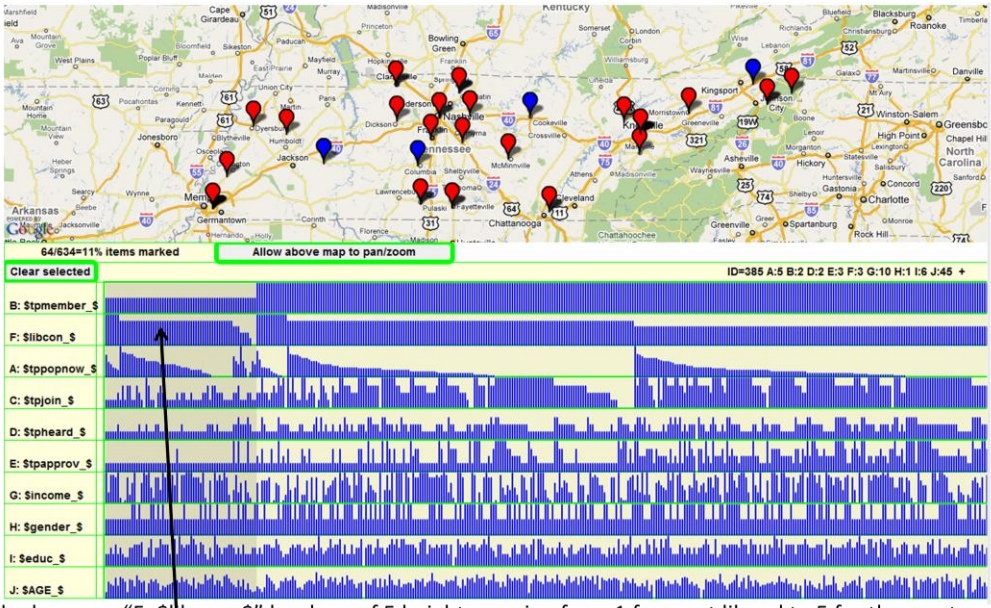
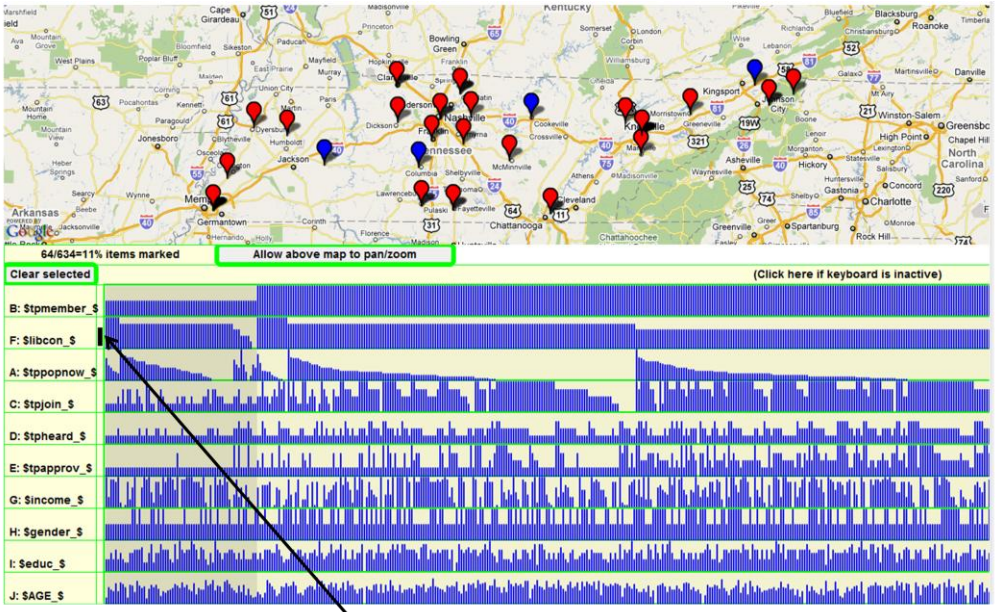


Figure 5: Result of the row drag step of Figure 4. The new order of rows has "F: $libcon_$" as the second row from the top. All intervening rows in the drag step of Figure 4 are pushed one row down.

Figure 5 shows the result of the dragging and dropping in Figure 4. Ideology row "F: $libcon_$" was moved to a position just below row "B: $tpmember_$" with intervening rows all shifted down by one row. After the movement of the rows, the InfAlign map is redrawn after a hierarchical sort.



Use mouse down, mouse drag vertically, then mouse up to draw a vertical black bar to delimit the zone of values to be combined into a common pool (the black bar pools heights ranging from 1 through 3; the two values 4 and 5 are pooled into a separate zone).

Figure 6: Selection of a vertical zone within which all variables are assigned to the same value. The selection is performed in a mouse down, mouse drag, and mouse up movement of the computer mouse in the narrow region between the blue bars of the InfAlign map on the right and the legend on the left.

For some purposes, is it is useful to assign people with different values for a variable to a common category. Figure 6 shows how a user dragged the computer mouse to select and assign a set of individuals with different values to a single category. The steps are a mouse down, mouse drag, and mouse up sequence of steps to specify a vertical zone in the InfAlign map for assigning blue bars to a common height. The zone is indicated by the black vertical bar.

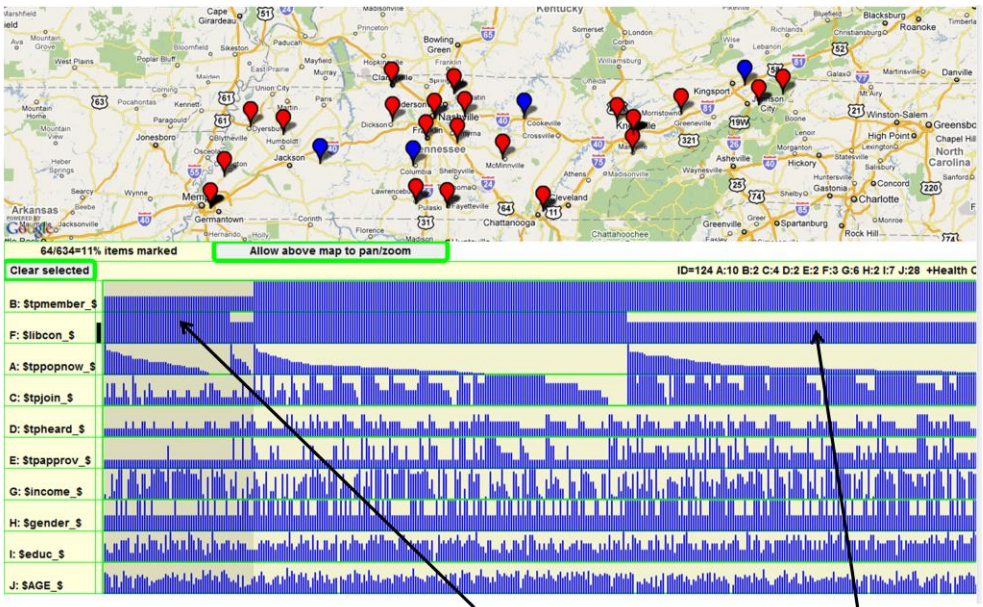


Figure 7: Result of the category selection step of Figure 6. All blue bars with values 4-5 are assigned to the category with the tall bars. All blues bars with values 1-3 are assigned to the category with the short bars.

Figure 7 shows the result of the placement of the black vertical bar in Figure 6. All blue bars within the specified vertical limits indicated by the black vertical bar are assigned the same height. All remaining blue bars are given a different height. Again, the display is redrawn after a hierarchical sort. After the dragging step, there are only two zones in row "F: $libcon_$" with tall bars corresponding to the variable having the values of 4-5 and the short bars representing the values 1-3 in Figure 6.

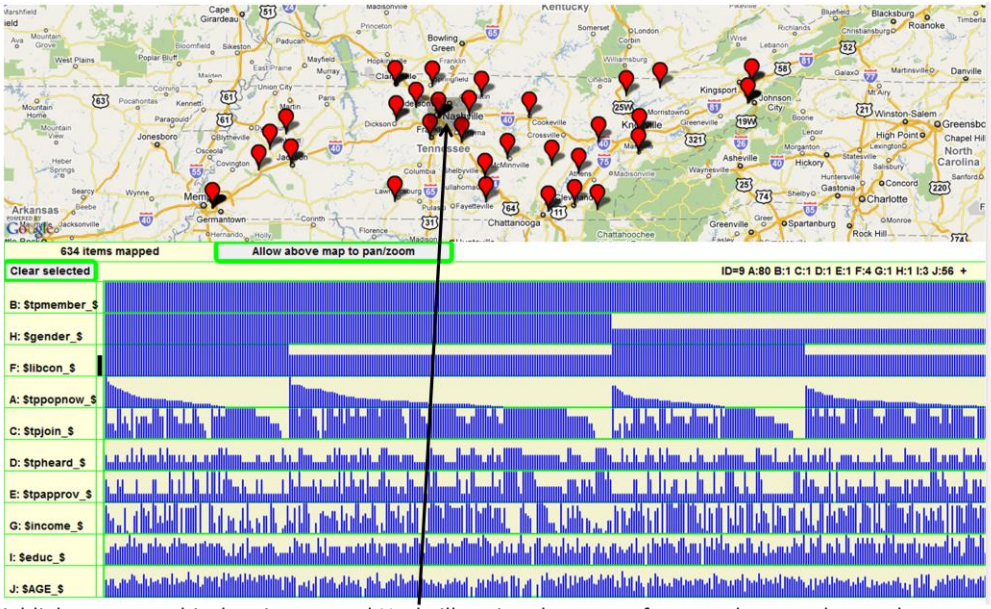


Figure 8: Selection of individuals in a geographical area. The user selects an area in the geographical map using a mouse down, mouse drag, and mouse up movement. The selected area is indicated by the darkened zone on the geographical map.

Figure 8 shows how the user performed the steps of mouse down to the northwest of Nashville, mouse drag to the southeast of Nashville, and then mouse up to highlight a dark colored rectangular region around Nashville.

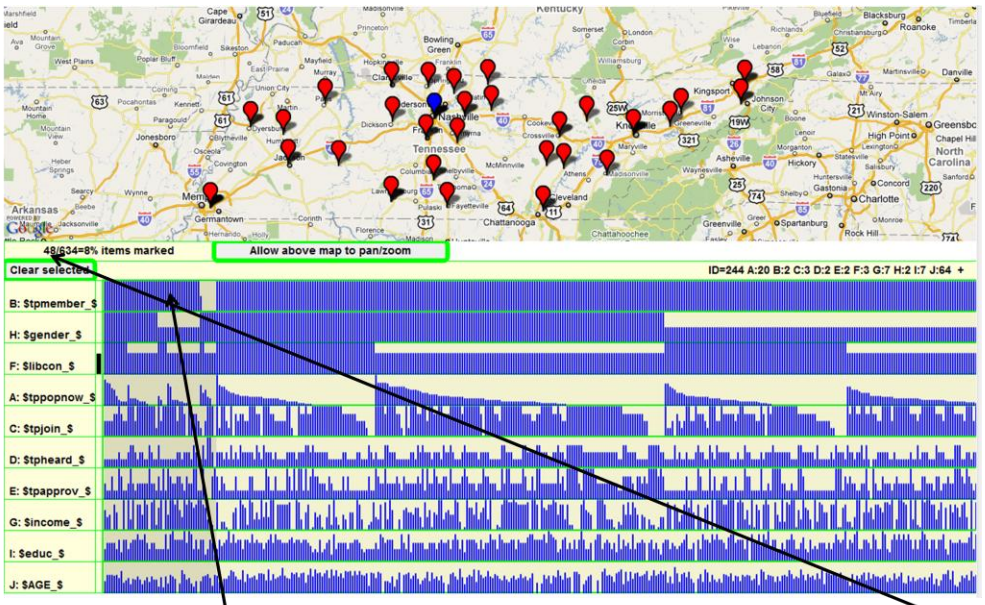


Figure 9: Individuals selected by the region selection step of Figure 8 are indicated by the blue marker on the geographical map and by the selected individuals in the darkened region on the left hand side of the InfAlign map.

Figure 9 shows that the bars in the InfAlign map corresponding to the persons from the selected region in the geographical map of Figure 8 all moved to the left in the highlighted region of Figure 9 and that only one of the 48 persons in the selected region around Nashville was a Tea Party member corresponding to the single short blue bar in row "B: $tpmember_$" in the highlighted region. The marker in the geographical map in the Nashville region is blue to indicate that the highlighted bars all corresponded to persons in the Nashville area.

3. Conclusion

This paper's main goal was to present a novel tool for graphically visualizing relationships among individuals in a population. The InfAlign map shows all values for ten or more traits for each individual. In the maps in this paper, there were 634 individuals each with quantitative values for 10 traits on the Infalign map plus the two geographical coordinates of longitude and latitude for the geographical map. Therefore, the total display could display 7608 quantitative data values in way that was visually accessible.

Relationships are rapidly seen by clustering individuals in different patterns with a simple mouse move. The InfAlign map can also be linked to a geographical map to display not only quantitative values of traits on the InfAlign map but also the geographic locations of the individuals.

The technology clusters individuals with the same or similar characteristics in close proximity on an InfAlign map. That allows for a quick visual inspection to see groups of individuals with the same traits.

In the hierarchical ordering of the InfAlign map, individuals are clustered by common traits with the greatest clustering occurring for the traits at the top of the map.

A key feature of the map is the ability of the user to reorder the rows in a drag and drop operation with a computer mouse. After each such process, the individuals on the map are resorted hierarchically from left to right and from top to bottom. The sorting allows the user to see individuals with the same trait grouped in one row and also the extent to which characteristics in nearby rows associate with the row of interest.

The resorting further allows individuals to be re-clustered by different traits simply using a computer mouse. This rapid and efficient re-clustering is more flexible than initially clustering individuals based on prior conditions and then performing analyses on the clusters.

The re-clustering permitted the exploration of the political ideologies of Tea Party members by dragging the ideology row to be just below the Tea Party membership row. Similarly, the user can drag another row like gender to replace ideology. Such a procedure showed that most members were male. Then, scrolling the screen to the right would show a comparison of these values for the non-Tea Party members.

Furthermore, the numbers of people in any visually identified cluster can be counted by using the computer mouse to select the cluster. The number of people in the cluster is then written just below the geographical map.

The values of all traits for a person represented by a moveable red line are given in a legend just above the map so that the user can read the actual values. And, not shown on the maps in this paper, the user can further click the mouse and a web page will open giving yet further information about the individual indicated by the red bar.

The user can additionally reassign and remap the values for a trait using a drag and drop motion of the mouse to select a vertical zone within which all values will be assigned to a common value. That permits a resorting of individuals so that the traits in the rows below will be more influential in the clustering.

Throughout the manipulation of the InfAlign map by a computer mouse, the geographic locations of the respondents are displayed on the geographical map above the InfAlign map. A random set of individuals displayed on the InfAlign map is indicated by colored markers in the geographical map with blue indicating selected and red indicating non-selected. There was substantial overlap in the geographical markers in this paper because geographic locations were only specified at the county level.

In a logically comparable process, the user can select a region on the geographical map using a drag and drop motion of the computer mouse. Then, the individuals in the geographically selected region are clustered to the left in the InfAlign map and shown in the dark color used to indicate selected individuals.

Perhaps the closest relative to the InfAlign map is the heat map where values are indicated by colors and their intensities. The InfAlign map uses the heights of blue bars to indicate these values, instead. The advantage of height over color is that it is easier for the human eye to see the presence of absence of a pixel than a gradual change in intensity or color.

The use of height to indicate value further increases the types of information that can be presented on a map by allowing color to be used for other purposes. Thus a red bar could track an individual and a group of selected individuals could be shown by a dark colored zone.

Although, the population in this paper consisted of human survey respondents with the traits being responses to items on a questionnaire, the same methodology can be applied to any type of population and traits. For example, the methodology has also been used with individuals being text documents rather than persons and with the traits being words in the text rather than questionnaire responses.

The InfAlign map is sufficiently different from other mapping techniques that it can reveal relationships that are less easily seen using other methods. The map's full capabilities will become more evident as it is used for other studies. Such studies can readily be performed because the InfAlign map can be accessed through a web browser from remote sites.

References

Wikipedia. 2010. Tea Party movement. http://en.wikipedia.org/wiki/Tea_Party_movement. Accessed September 13, 2010.

Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics.* 119-127.

Fan, D. P., and R. S. Fan. 2009. Population Analysis using Linear Displays. *U.S. Patent 7,519,521.*