# Visual Discovery of Patterns in Census Data *

Zaixian Xie †, Charudatta Wad, Do Quyen Nguyen, Qingguang Cui, Di Yang,
Matthew O. Ward, and Elke A. Rundensteiner
Worcester Polytechnic Institute

## 1 Approach

One of the biggest challenges when analyzing census data is that many, if not most, of the data attributes are nominal, rather than ordinal. As most graphical mappings in visualization assume numbers are controlling the graphical attributes, and that users assume that similar graphical attributes imply similar data attributes, the mapping of nominal variables to numbers is critical. The above mentioned problem gets much worse when the number of possible values a nominal variable can take on is large.

Our solution involves structuring and interactively modifying the nominal variables of this dataset that are most critical for the contest tasks. We use interactive selection, filtering, and reordering to enable users to quickly and easily move through the dimensions associated with ethnicity and languages, using a geospatial layout in conjunction with multivariate glyphs, where instead of multiple dimensions controlling the glyph shape, the distribution of languages and ethnicities is conveyed. For the dimension conveying the industry of employment, we use the existing industry hierarchy in conjunction with a hierarchy visualization method, previously used to show hierarchies of dimensions and records, to instead show a hierarchy within a single dimension. Linking this view to other visualizations and supporting linked selection between all views enables users to see what jobs are performed where.

## 2 Data Preprocessing and Augmentation

To provide a geospatial reference point for each Super-PUMA geographic unit (which we call a region), we first acquired a list of cities and towns that fall within the region from the auxiliary files provided with the Contest data. We then matched each city with a longitude and latitude from a dataset used in last year's InfoVis Contest. Finally, we computed the average of these positions to generate the representative location for the region. While this is a coarse approximation, we felt that precision was not essential in conveying the areas in which certain ethnicities or industry types were located.

To enable the generation of more abstract views of the data, we computed aggregations for each region, for each ethnicity, for each spoken language, and for each industry category. We also converted the hierarchy of industry categories into a format that could be readily imported into an existing hierarchy visualization tool (InterRing) that had been developed in our lab.

## 3 Resulting Visualizations

We explored one state's worth of data using an existing multivariate visualization tool (XmdvTool), which supports interactive visual analysis of multivariate data using several techniques. Some patterns were quickly discovered, including the identification of regions where certain languages were not spoken and where certain industries were not present. We could also see some correlations between industries and ethnicities. However, our existing code did not allow us to see the relative sizes of populations with varying characteristics, and two of the dimensions, language spoken and industry, had too many distinct values to enable detailed analysis. To overcome these deficiencies, we implemented two new visualizations, as described below.

To study the distribution of spoken languages within and across regions, we developed a profile/histogram glyph to show the percentage of the population speaking each of the 400 languages listed in the census. As this number made detection of individual languages difficult to examine, we allow the user to interactively specify a subrange of the languages. To accomplish this, the user selects a region of interest from the geo-referenced glyph view. A detailed histogram is then shown in a separate window with languages automatically sorted from most to least common. A shaded region is shown depicting the subrange selection, which the user can slide up or down the histogram or widen the size of the selection. In all glyphs, only the bars corresponding to the selected languages are shown, and in all glyphs these bars are sorted in the same order as the selected glyph. This allows users to see similarities and differences in language profiles between different regions. See Figures 1 and 2.

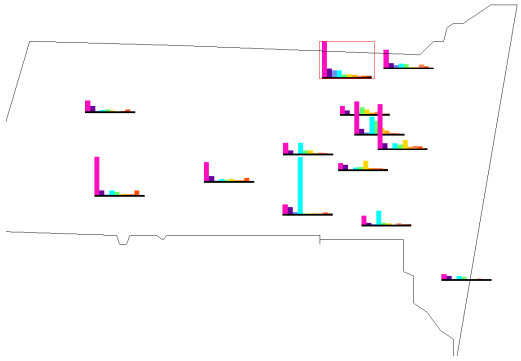

Figure 1: Profile/histogram glyphs showing a the frequency and distribution of a subset of languages spoken in Massachusetts, sorted by one of the northmost regions.

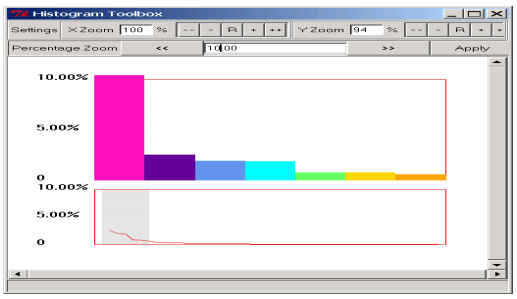


Figure 2: Histogram toolbox showing a detailed view of the full histogram and selected subrange for a selected region.

*This work was supported under NSF grants IIS-0119276 and IIS-00414380.

†E-mail:xiezx@cs.wpi.edu

For industry categories, we needed an approach that would scale better to the 300 job categories. Fortunately, as the categories are already hierarchically structures, we could extend our radial space-filling hierarchy viewer (InterRing), previously used to study clusters of dimensions and records, to now support viewing and interactive exploration of hierarchies within a single dimension. When the user selects a specific region, all industry categories found in that region are highlighted in the InterRing display. To convey the actual counts of people in the highlighted professions, we implemented two strategies. The first distorts the radial angle assigned to each wedge of the InterRing display so that industries with large numbers of employees occupy more screen space than those with few or no employees (see Figure 3). The second color codes each block of the original hierarchy display so that those with similar numbers of employees will have similar colors. We feel that both strategies have merit.

Within InterRing, the user is able to select subsets of industry groups of interest and use this to affect the other visualizations. Thus it is possible to identify linkages between industries, languages, location, and any other variable within the census dataset. We feel that seamless linkage, specified via whichever visualization is most appropriate to the user and task, is a powerful mechanism to support visual exploration. Data points related to agriculture jobs in Figure 4 are selected showing in red. We can find that some races in some specified areas do not do these jobs.

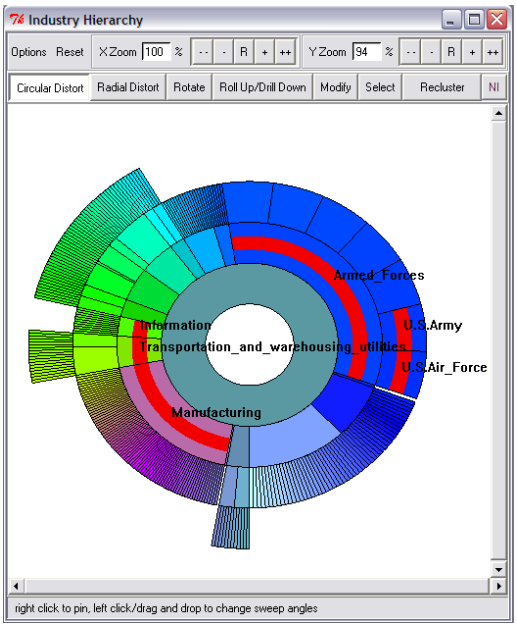


Figure 3: Distorted InterRing view conveying employee distribution across industries for one Super-PUMA area in Massachusetts. The radial angle reflects the number of people employed in that industry.)

## 4 Tiny Displays

Each of the displays within XmdvTool can be arbitrarily scaled to fit the application and device. Figure 5 shows examples of some of our new visualizations scaled to approximately 220x176 pixels. As most visualizations support both scaling and distortion, detail is available on demand. The biggest problem we encountered is that our GUI, with menus, toolbars, and buttons, did not scale as well as the visualizations. In fact, only the menus could be used as is. The logical solution, though not yet implemented, would be to ensure that every toolbar icon and button have a corresponding

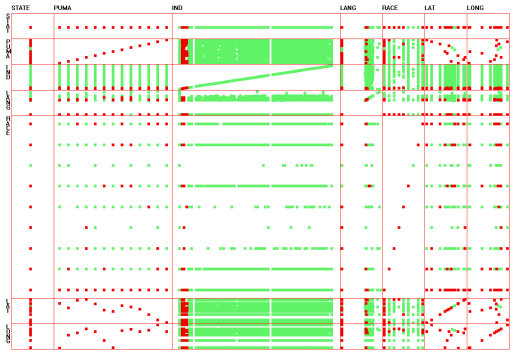


Figure 4: Profile/histogram glyphs showing a the frequency and distribution of a subset of languages spoken in Massachusetts, sorted by one of the northmost regions.

menu entry, and then just eliminate the GUI components that cannot be included in a tiny display system.

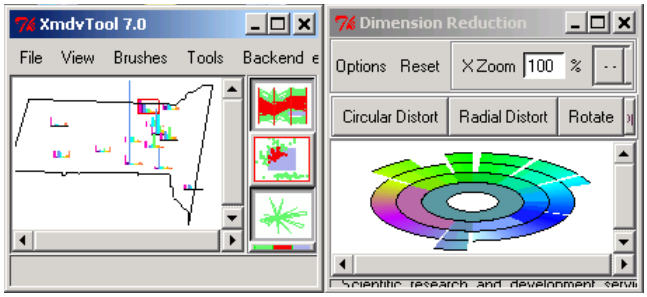


Figure 5: Visualizations on small displays (220x176 pixels).

## 5 Weaknesses and Future Work

We feel one significant weakness of our solution is that text labels on several visualizations quickly clutter the display. While the text attributes of a selected record are clearly presented in a separate message box, we think a better solution would be to use font scaling and distortion to enable users to see labels only in their areas of interest.

Other areas we would like to explore involve the emphasis of differences between regions rather than just similarities. Thus, for example, in the glyph views the user could be given the option of replacing all glyphs with ones computed as the difference between the original glyphs and a user-selected glyph. Color coding could be used to differentiate positive and negative differences. A similar strategy could be used with the InterRing display to show how industry profiles differ between regions.