BROWN UNIVERSITY
Department of Computer Science
Master's Project

CS-94-M7

"Displaying Multivariate US Census and Migration Data Using Three-Dimensional Graphics and Animation"

by

Nisha D. Thatte-Potter

Displaying Multivariate US Census and Migration Data Using Three-Dimensional Graphics and Animation

by

Nisha D. Thatte-Potter

Department of Computer Science

Brown University

Submitted in partial fulfillment of the requirements for the

Degree of Master of Science in the Department of

Computer Science at Brown Univerisity

May 1994

Professor John F. Hughes

Advisor

Abstract

A system to exhibit multivariate demographic data is presented. This system was designed to overcome the limitations of static two-dimensional displays. It allows up to nine variables (including placement along the *x, y, z* axis) to be represented in a single icon. The system uses a human stick figure for its icon because the stick figure is both familiar to the user and allows the variables to remain distinct. It draws on the separate features of a stick figure such as arm position, head shape, and color to clearly show each variable's value. This presentation makes it easy for the user to quickly determine trends in the data. It includes animation to display migration between regions of the country using the same icon representation.

1.0 Introduction

Graphs can only present two pieces of information for every data point in the graph - the *x* and *y* coordinates on the graph. In a graph of U.S. census data, for example, the user can only see the number of people in each employment status at one time. U.S. census data, however, contains many variables such as sex, race, age, employment status, and educational attainment. In order to present all of this information simultaneously, several separate graphs must be overlaid. Unfortunately, such overlays can be difficult to read, particularly when there are many variables involved.

I propose using a basic icon of a human figure to represent all the information in one location. By using the icon to simultaneously represent different pieces of information about groups of similar persons (sex, race, employment status, education status, and age), I create an overall picture of these persons. I then place the icon on the screen, varying its location according to the region of the country that the represented persons live in. Finally by creating a "stack" of these icons, I give the user yet another piece of information; how many people in this location fit these specifications.

This system of representation is a great improvement over existing methods. Currently someone examining this type of data has two choices. Either they can read through piles of paper and compare results by flipping back and forth through the data or they can create a two-dimensional graph such as a pie chart, bar chart, or line graph. These graphs will typically have the quantity on the y axis and the values of a single variable on the x axis. This gives a very limited graph that does not allow for much analysis. Occasionally, the x axis will have multiple variables for each value. This creates many combinations, however, and will most likely span over several pages. By giving the user more information simultaneously, my system allows the user to take in all the details without having the distraction of moving between pages and remembering what s/he has previously viewed.

My system also displays animation of migration data. In standard two-dimensional displays, the data is static. The user sees the distribution of the people before migration, the number of the people who moved, and the distribution of people once the migration has finished. As with standard graphs for general data, readers only see one data classification at a time and therefore must spend time browsing through several pages to get the "larger picture". By animating icons similar to those used for the general population, but represent the migrants, I show how people move between regions. This animation of migration provides a whole new way to visualize migration since, as I said earlier, migration data is usually viewed in a static way.

The remaining sections of this thesis will detail the development and features of my system and how users reacted to it. Section 2 will look at related work. Section 3 explains the basic structure of the program. Section 4 presents the results of user studies. Section 5 gives possible future directions. Section 6 gives a conclusion.

2.0 Related Work

2.1 Perception

An important factor to consider when creating information graphics is how they will be perceived by users. It is important to understand how people use their perceptual and cognitive processes to understand the data being shown to them in order to present data to them. These ideas are explored in a paper by Jerry Lohse [2]. Lohse found that people can comprehend and absorb some visual primitives more easily than others. Texture and color are readily detected while shape detection can be slower. One way to help in discriminating among the shapes is to make them distinct and to not allow overlap. If a program has features that can be quickly discerned, the time the user spends deciphering these symbols is greatly reduced. As a result of these and other findings, Lohse created a program to model how people decode information from a graph by using eye fixations.

Edward Tufte [3] also described how people evaluate graphical pictures. He states that small multiple figures that can be viewed simultaneously are more effective than larger images that are scattered over a large region or presented as successive images on separate pages. As we will see below, issues of perception are even more important when presenting multivariate data.

2.2 "Chernoff Faces"

Herman Chernoff [1] was one of the first people to propose displaying multivariate data in one representation. His work has become the standard reference cited in regards to the display of multivariate information. Chernoff used faces to display multivariate data. His work was based on the rationale that humans are used to studying faces. Moreover, people easily detect small changes in faces, because they are conditioned to react to them. Finally, the faces do not have to be realistic in order to be useful, because people are used to cartoons and caricatures. The positive aspect of this application is that groups and trends are automatically recognizable.

In his program, Chernoff associated the variables with different attributes of the face, such as the width of the mouth, vertical size of the face, length of the nose, and slant of eyebrows. A human feature like the eyes could be associated with six (or possibly more) variables. Vertical position of eyes, separation of eyes, slant of eyes, eccentricity of eyes, size of eyes, and position of pupils were all used. The program then evaluated all of these variables and drew the corresponding drawing. I find this multiple assignment of variables to a single feature to be a flaw in the design. Overloading can make it difficult to discern what the different individual values are. Another disadvantage of using faces to represent data is that the user may incorrectly interpret them. As Chernoff points out, users may assume a happy face means that the data represented is positive, and assume a frown is something bad. This association may not be representative of the data and is difficult to control, because many different combinations of variables can create recognizable expressions.

2.3 Exvis System

Chernoff's original concept has been extended in the Exvis System [4,5] from the University of Massachusetts at Lowell. Exvis does not use an icon that represents a known physical object. Instead, Exvis uses a basic icon consisting of short line segments known as "sticks". These sticks can be joined at their ends. A variable is assigned to each stick and its value determines the angle at which the stick appears. Additionally, the placement of the icon on the x and y axis can represent one to two variables. When the icons are packed into a two-dimensional array, textures form that the user can perceive. This display appeals to the human capacity to understand and sense a texture.

The main disadvantage of this system is that there is no immediate recognization of the icon, because the sticks do not have any automatic interpretation to the user. While that makes the system flexible in the types of data it can present, it means the user must become familiar with the shape in order to quickly interpret data. Similarly, while the dense packing of the icons allows the user to readily see a texture, it makes it very difficult to pick out individual variable values. This problem is magnified by the abstract shape of the icons which makes it arduous to discern which stick belongs to which variable, since the connection of the sticks can appear very arbitrary to the user. Nonetheless, a very positive aspect of this work is the separation of variables. If the users do separate out the sticks, they can determine the value of the associated variable fairly quickly.

2.4 ArcView

ArcView [6, 7] is a product released by Environmental Systems Research Institute, Inc. It is a program to visualize data with a geographic database. It seems to be becoming the industry standard. ArcView consists of two-dimensional maps that the user can choose along with data sets. I chose to examine the 1988 population statistics. Using color, this system displays various characteristics of the population one at a time. If the user wishes, s/he can pick many geographic pieces of information such as rivers, cities, counties, or states to appear on the screen also. The system also gives the user a legend and allows the user to set the color and interval associated with the attribute. For instance, I could pick random colors and ask the program to set each color to represent a population density for Whites in America. The system then displays the color-coded information on the map.

The obvious drawback is the fact that it can only show one variable at a time and cannot display changes over time. It is, however, widely available, allows printing of the graph, can make queries of the data, has zoom features so that you can focus in on a region of the country, and has many types of data files to use.

3.0 Specifics of my system

My system was designed to improve upon current representations of population data by using the simultaneous presentation of multiple variables. In doing so, I combined elements of Chernoff's program with those of the Exvis system by using an icon shape that is both familiar to a user and relates to the associated information. At the same time I keep each variable distinct. Since I am representing population data, the icon I chose was a human figure. A too abstract shape such as the squiggles used in the Exvis system can discourage the user from making quick decisions, while an icon where the changes are intertwined (as in the Chernoff faces) makes it hard for the user to pick out individual data values. Keeping the variables separate, as Lohse detailed, allows the user to spend less time processing them. Finally, I use Tufte's conclusion that small multiple figures represent data better.

My system also improves on previous systems by using three-dimensions and animation. Three-dimensions add another variable to the display and allows the user to view the display from different vantage points. This is important because different patterns become visible from different views. The animation lets the migration flows be shown in a dynamic way. Most conventional displays of migration data are static. They either have arrows indicating the flow or present "before and after" pictures of how migration changed the population. By watching the migration happen in front of them through animation, users are given time to absorb the trends and see how the population changes each represented group of people.

The overall system shows four sections of the country (Northeast, South, Midwest, and West), each displayed as a platform. Groups of icons are placed on these platforms. For every individual group of people (for example, Female, High school educated, employed, over 25 years of age, and White), a "stack" of icons representing all of these attributes is placed. The stack height represents how many people make up this group.

Once the display is completed, users can move around the view by manipulating dials attached to the computer. By navigating through the display, they can view the data from many angles and zoom in and out of areas that they find interesting. Additionally, when they click the mouse on a stack of icons, the information that each stack of icons represents appears in a pop-up window.

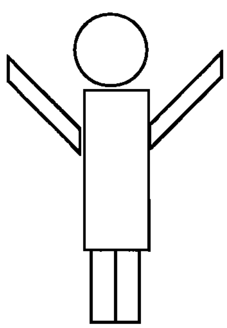
3.1 Basic shape of the icon

The icon is broken up into six features: the head, right arm, left arm, legs, color, and height. An example of the icon is pictured in Figure 1. The head can be a sphere or cube. The right and left arms can independently be in the up position, down position, or horizontal. The legs can be together, apart at an angle, or straight out (as if the icon was "doing a split"). Currently the icon can be colored either blue, white, or black, but additional colors can be added. The height of each icon stack can range from one to unlimited. Each icon is placed by the application at an <*x,y,z*> location. This position can be used to represent one to three variable attributes. Figure 2 shows the icon features with their possible values.

Assignment of only one variable to each feature of the icon lets the user quickly focus in on what is important to him/her and dismiss the other features. Alternatively, the user can take in the icon as part of the larger picture and search for patterns in the data. Each icon is created by specifying the characteristics or position of each "body" feature. The icon object does not know its relationship to the data that it represents. This makes it easy to change what each feature represents.

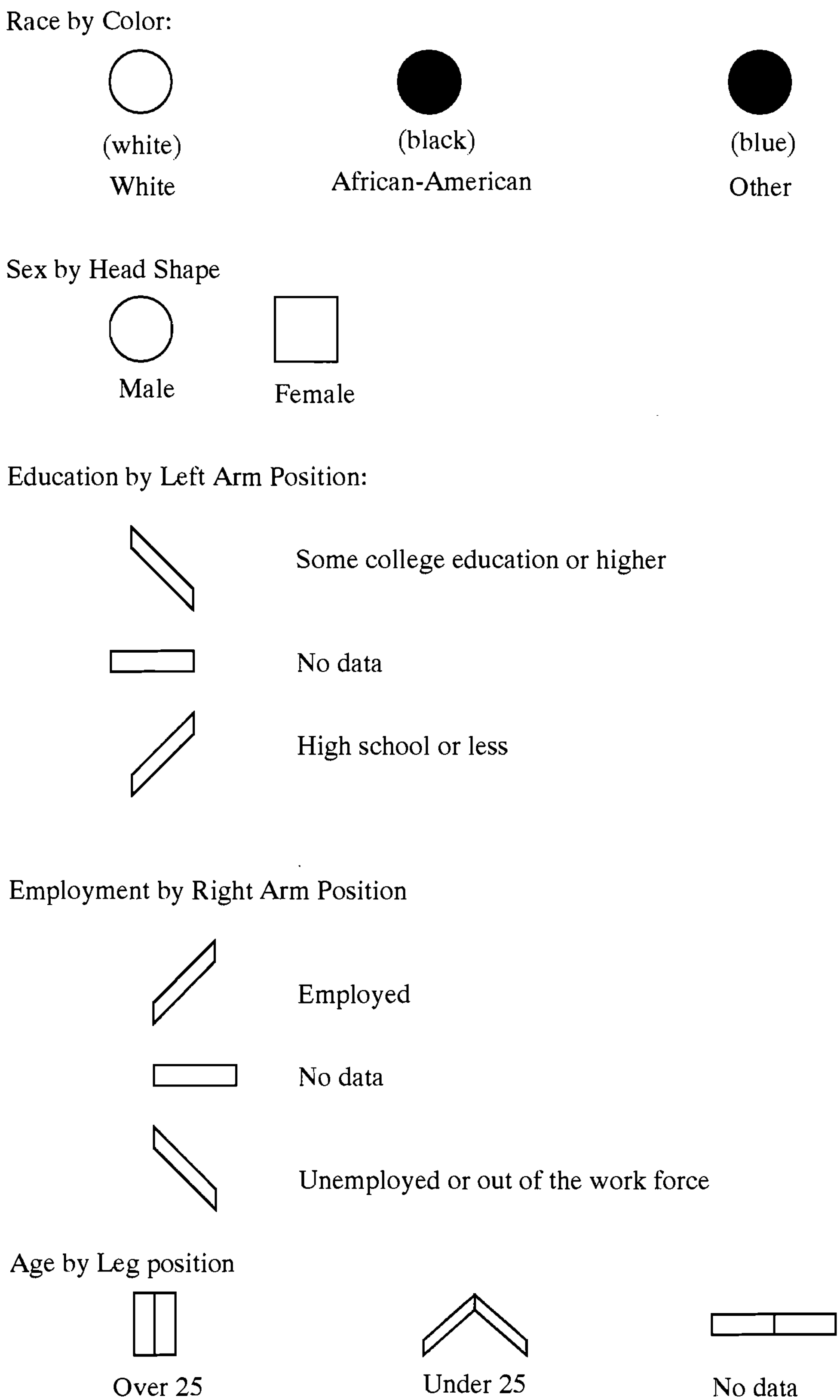
Originally the icon stack was a list of individual small structures that each represented 500,000 people. The icons were then grouped together and stacked on top of each other. This was more visually appealing but the system ran slowly since each icon had to be evaluated individually before being drawn. Since speed is very important in an interactive program, I chose to draw only one icon for each stack, and to set the height equal to what the height would have been had all the individual icons been stacked. To represent the number of total people more clearly, I create individual heads for every 500,000 people that the stack represents. This allows the user to note the individual heads to give some idea of volume while improving speed by not having a large list of individual icons. Running the program on a more powerful computer could allow the restoration of the original display.

FIGURE 1. Example of icon



This icon represents a White, male, with a college education or better, who is employed (either through in the civilian workplace or in the military), and is 25 years old or older.

FIGURE 2. Key to Representation of People



3.2 Data Acquisition

The data is based on 1990 U.S. census data [8, 9] and the 1990 - 1991 General Population Report on Geographic Mobility [10]. Since the migration is only tabulated on a regional level for the country, I chose to also aggregate the general census data on a regional level.

I represented the U.S. census data by rounding the numbers to the nearest multiple of 500,000. That is because, for demonstration purposes, this provided a manageable number of icons. The program reads in the data from files that specify the characteristic breakdown of people to generate the display. The census data I obtained was not in the form of groups of 500,000 people having several attributes so I had to massage it. First, I determine the number of people of each sex and age and create the correct number of data structures. Then I randomly set values for the other variables of the icon, based on the raw numbers that I obtained. For instance, let us say that the data was in the following form: 20,000,000 Whites over the age of 25 and 4,000,000 Whites over the age 25 who had a college or better education. First I create 40 data structures that each represent an icon that is White and over the age of 25. Then I us a random number generator to pick an icon data structure from among those 40 data structures. If the educational variable has not already been set on the one that I choose, I set the variable to indicate a college education. I continue in this manner until I choose and set 8 icons from among the 40. Once all the data has been set, the data structures are sorted to group identical icons together. Then the visual icon can be created with the correct height to indicate the number of people that the final icon represents.

Migration data was similarly massaged. It is rounded to the nearest multiple of 10,000 since the number of people moving is much smaller. In addition to using the random method of setting the data as described above, I also have to make sure that I am not creating a migration icon that does not have a matching general population icon. For instance, if I create a migration icon that is other, female, college educated, employed, and over 25 migrating from the West, I have to check that there is a stack of icons in the West with these same exact attributes. The reason behind this is that the migration icons first appear on the screen on top of a stack of existing icons. If there isn't a general population stack with these identical characteristics, then there will be no place to put the migration icon. If it turns out that there isn't a matching stack of icons in the general population, the migration icon is not used. This, however, means that there is a possibility of data loss since some of the created migration icons might not appear on the screen. In practise this turns out to be a small number.

I used the random method because of the way the data is set up and the way that the data structures are created. The first set of data structures contains 1 to 15 year olds, the second contains 16 to 20 year olds, etc. If I start at the beginning of the structures instead of picking one at random, there is the possibility of setting all the 15 to 25 year olds to have one attribute and all the 50+ year olds to have the opposite of this attribute. This could make all young aged icons be employed, but all older aged icons to be unemployed. My method of setting the attributes randomly does have the drawback of possibly creating an icon when there are not people who hold those characteristics in the real population. For the purposes of this system, this level of accuracy is sufficient.

3.3 Migration

I present migration here as the path from one region to another that represents the physical movement of people. In this display of migration, the user sees the following: at the beginning of the migration from region A to region B, all the migration icons appear on the stacks of the corresponding icons from region A. They rise and file out one row at a time. This ensures that the process is orderly (like filing out after a wedding ceremony) and that the icons do not walk into one another. They next follow a pre-defined path from region A to region B. The paths have been designed so that even if every possible migration occurs there will be minimal interference between groups of migrating icons. When they reach their destination they either lie on top of a target stack in region B, if one matching the migration icons' characteristics is present, or if not, they form a new stack.

The visual display of the migration icons is the same as that of the general population icons but the migration icons are scaled smaller to indicate the fact that the number of people that each migration icon represents is much smaller. The user can select from among four options for the animation of migration. The display can show all migrations, all the migrations to one region from the three other regions, from one region to all the other regions, or from one specific region to one other specific region. The user can suspend the migration animation to examine the figures more closely and can reset the migration information and rerun a migration.

3.4 Code

The three-dimensionality of this program meant that I had to choose both a programming language and a library package that would display three-dimensional objects. I chose to code in the C++ programming language using the Brown University Computer Graphics Group UGA packages [11]. This combination allowed me to easily develop a three-dimensional program with user interaction. UGA allows for the creation of a three-dimensional window and simple polygonal shapes and reports on events from external devices such as the dials and the mouse.

The code is an application layered atop UGA packages. It is broken up into several components: camera actions, dialog actions, light actions, mouse actions, the icon object, the assignment of data characteristics, the platform/map object, the overall coordinating, and the drawing routines. Each of these components is a separate C++ object that is encapsulated from the others. An improvement could be made by breaking down the drawing object into smaller objects. This could create separate animation objects, migration objects, and general population objects making the code more modular.

4.0 User Studies

4.1 Overview of users

Thirteen people participated in the study. They were shown the program and shown how to use it. They were allowed to use it as long as they wanted and were timed to see how much time they spent on it. This was not a "hands-off" study since I gave suggestions while each user was trying to do something (moving around the data and examining the data) and when the users asked questions. I recorded their comments and actions as they tried out the program. Occasionally I asked them questions during their use based on what I saw them trying to do or if they exhibited any confusion. At the end I asked the following basic questions:

1. How would you describe the characteristics of the data you just saw and what sort of relationships did you see in the data? Did you notice any trends in the data?

2. What were you expecting to see in general? (To determine what preconceptions they had.)

3. How did this meet your expectations?

4. What did you like?

5. What did you not like?

6. What stands out the most?

7. What do you see no use for?

8. What was easy to understand? (This question and the next covered both the usability and the concept of the program).

9. What was hard to understand?

10. What suggestions to you have to improve this program.

4.2 General impressions

Most people used the program in similar ways and responded with similar evaluations. A few people had unusual reactions which have been noted below. The majority of the people concentrated on the original display of the general population and had to be almost prompted to use the animation. Most users had no preconception of what to expect other than what I had told them. A few were expecting similar program to what they were accustomed to: two-dimensional maps with two variables with arrows indicating migration flows. Some of the testers had watched over my shoulder or heard me discus the program during development so they based their expectations on that. In every case, the program met or exceeded their expectations. Everyone liked the use of color and many wished that there could have been more use of it. They stated that color was the easiest feature to notice right away. In fact, when asked to comment on trends most people mentioned ones that contained race which is represented by color.

As for viewing the program and camera control, all but one person liked the ability to rotate, tilt, and move around the data. It took everyone a little while to get used to using the dials for translate, rotate, and zoom but most everyone could operate the dials easily by the end of the session. The majority of the users appreciated the three-dimensional views and the feeling of volume.

Another well liked aspect was the animation to represent migration. Many enjoyed seeing the icons in a walking position and felt that the animation gave them time to notice details and a feeling that the information was dynamic. Virtually all of the people noticed that the migration icons were smaller but did not always realize that the smaller size represented the smaller number of people involved in migration. They noticed that the migration icons moved to the existing stacks or created their own at the end of the animation though they had to be asked about where they thought the migration icons were ending up as they did not remark about it out loud. One person mentioned that had she not been looking at the finishing region from the angle that she was currently viewing it, she might have missed seeing where the icons went.

People voiced other negative reactions about the system. There was a learning curve associated with deciphering the figures. People had to either use the mouse to select a stack of icons to determine the icons' attributes or check with the printed key to determine what the positions of the various limbs meant. Some felt that while having all that data available was useful, at times it felt like too much data. They would have liked to been able to occasionally filter out some data and only look at what they thought was important. One subject remarked that she would like to view one region at a time and filter out the other regions. The head was a difficult part of the body to decipher and a few people did not even notice that there was a difference in the head shapes between stacks. Many people remarked that they had to study the figures carefully to find the information that they were interested in (e.g., education or employment). Other problems were with the way the stacks were sorted for display. The fact that the stacks were sorted by race was obvious but not the fact that they were sorted by sex as a second sort within race. A few subjects felt that the lack of numeric data along with visual data was a hinderance and that while they could notice the general trends, they would like to have had the actual numbers visible.

Some people had comments about how the display was viewed. One person felt that rotation was not very useful and did not see much use for it. Several people remarked that they would like to have the ability to save some camera views and have some predefined ones since getting back to a certain camera angles could be difficult. Finally, many felt that the stacks could be hard to distinguish from one another. This complaint went away after I added better lighting.

A few problems with animation were also brought up. One person felt that while the animation was perfect for migration, it would be hard to extend animation to other problems. Another drawback pointed out was that the walking position during animation and the smaller size of the figures made it hard to see the position of the legs and some other features.

Many suggestions were given in conjuction with the criticism. There were suggestions for more variables. Two possibilities would be to use width (fat and skinny icons) and to apply patterns on top of the colors. More raw data and even a two-dimensional chart of what is on the screen, including current information about a migration, would be helpful. While most people got accustomed to the dials, many people requested that there be some default views to show off certain details such as volume and the regions of the country. One person thought that using the same migration paths regardless of the overall type of migration was confusing. She felt that the migration paths should be tailored to the type of migration requested (one to one, all to one region, etc.). Some people wanted more control over the assignment of variables. They wanted to be able to pick which variable went with which features of the icon. Many testers would have liked to be able to sort the stacks according to what they thought was important. One person wanted a help menu.

4.3 Interest in the data

The interest category ran from professionals and graduate students in demography, to students in related fields who have had to study similar data, to people who were interested in demographics generally. Eleven people were directly interested in the data. Of the two other users, one had no interest at all in the data and one was interested in other graphs but not demographic data.

The group interested in the underlying data spent more time using the system. The eight that I timed spent approximately 26 minutes using the program. This set of users had the tendency to immediately start analyzing the data and to look for trends. While a few of those in the demography field felt that they could not use it in their line of work, they felt that it would be a good teaching tool. When one tester was pressed as to why she could not use the program for her studies, it was mainly based on the choice of underlying data.

Being accustomed to looking at two-dimensional graphs may have had an effect on how they reacted to the program. The users accustomed to looking at two-dimensional graphs were more likely to remark on how, to them, the data was better communicated in three-dimensions. I also attributed the situation where a few of the demographers were unhappy with the height of the icon representing the number of people to the newness of three-dimensionality. This small group of users liked the height factor but felt that an icon for each 500,000 should be placed directly on the platform. They stated this despite remarking that doing so might clutter the screen and were not sure if a single height icon was an overall better solution. Many two-dimensional displays represent data in this way and that might have contributed to their uneasiness with a new format.

As for the migration animation, a few of the demographers felt that net migration would be more useful to them than the total migration flow. In this case, regions would have to be grouped together and if 2,000,000 people moved from region A to region B but 500,000 had moved from region B to A, then only enough figures to represent that 1,500,000 people moving should be displayed as moving from A to B.

The two not interested in the data spent approximately 10 minutes with the system before moving on to the question part of the user study. They focused in on the visual aspects of the icons, and looked at the overall presentation of the screen, such as lighting and camera direction. They did not do any data analysis and did not look for trends in the data.

4.4 Level of computer science

All the users were computer literate. Everyone had used mainframe computers and several were knowledgable with personal computers. The amount of experience with computers ranged from non-programmers to those who use programs as part of their profession to experienced programmers. A subset of the experienced programmers were also very familiar with computer graphics and animation.

The users that did not normally use three-dimensional programs were more appreciative of the three-dimensional and animation aspects of the program. Many termed it as "fun", "entertaining", or "novel". They were more likely not to have imagined a three-dimensional application for the problem and were thinking of either a map with arrows or overlays of several maps. There did not seem to be a difference in the learning curve of moving around the figures with the dials and the different camera views by rotation and translation between the groups, even though many of the graphics people had had experience with camera manipulation and the dials. No one held back while trying the program; everyone seemed eager to try as much as they could.

5.0 Future directions

There are many directions that the development of this system could go in. The first would be to make the program more robust and add some of the suggestions made by users. These would include data file selection by the user, such as "elderly migration data." I could also add a user choice window which would allow the user to select the assignments of attributes to parts of the body. Letting the user choose the sorting of the icons on some attribute other than race and sex would be helpful. Allowing filtering and querying of the data so that only icons with certain characteristics would appear would give the user a chance to fine tune what was being displayed on the screen and improve comprehension. Along those same lines, applying that to migration would also be beneficial, since that way the user could see a subset of the migration with one type of variable controlling what animated (i.e., only men, only college-educated people, etc.). A zoom feature could be built in allowing users to take in the overall U.S. picture and then zoom in onto the level of detail that they found interesting such as state, counties or region. Some people wanted to have the numerical data available, while others wanted average statistics on the region, while others wanted the percentage statistics (e.g., whites are certain percentage of the population). All of this information could be made into a table on screen allowing the user to verify the trend that they just saw with the raw data.

Additionally, I do not think that the icons are limited to census data. I think that the icon is more general purpose and can be used in any application that studied characteristics of people. Other such topics include political party affiliation, voting, religious affiliation, and consumer buying.

6.0 Conclusion

This paper has presented a system for creating and displaying icons for multivariate data. It uses an icon that resembles a person and assigns various data attributes such as age, sex, and race to the head, legs, arms, and color of the icon. Additionally it uses animation to represent migration flows between regions of the country.

This system is deemed successful through the user studies. Trends in the data were recognized by most users from the mildly interested to the those who study this data every day. Everyone appreciated the animation and felt that it helped the display.

While this has been judged as being a first step in the right direction, it is only that: a first step. Many improvements can be made as we learn what people need to visualize and what appeals to them. Hopefully, one day users will be able to use the computer to make their data come to life in a way that it cannot on paper.

5. R. Daniel Bergeron and Georges G. Grinstein, "A Reference Model for the Visualization of Multi-dimensional Data," *Eurographics '89*, pp. 393-399, (1989)

6. Environmental Systems Research Institute, Inc., "ArcView User Guide," (1992)

7. Environmental Systems Research Institute, Inc., "ArcUSA 1:2 M User Guide and Data Reference," (1992)

8. Census of Population and Housing, 1990: Summary tape File 3 (United States Summary Regional Reports Except urbanized areas and their components) [machine readable data files] / prepared by the Bureau of the Census - Washington: The Bureau [producer and distributor], (1991, March 1993)

9. Census of Population and Housing, 1990 Summary tape File 1 on CD-ROM (United States Summary) [machine readable data files] / prepared by the Bureau of the Census - Washington: The Bureau [producer and distributor], (1991, February 1992)

10. U.S. Bureau of the Census, Current Population Reports, P20-463, *Geographic Mobility: March 1990 to March 1991*, U.S. Government Printing Office, Washington, D.C., (1992)

11. Brook Conner, "OD: Objects to Draw in WAV Windows," Brown University Computer Graphics Group, (1993)

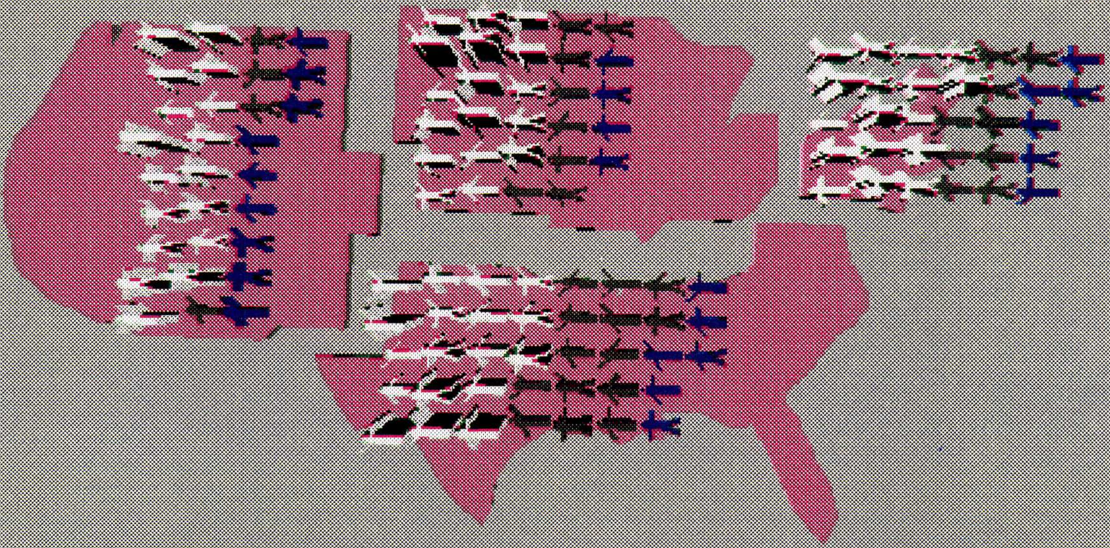


FIGURE 3. The general population of the United States.