ABSTRACT

RAMACHANDRAN, KARTHIK. Visualizing and Comparing Multivariate Scalar Data over a Geographic Map. (Under the direction of Dr. Christopher G. Healey).

Recent technological advances and innovations have given us ways to easily and quickly extract large sets of data, but the increasing amounts of raw information only highlight the lack of good visualization or pattern recognition techniques to interpret the data. The objective of the research is to build techniques to effectively visualize multivariate scalar entities over a topographical map. Our goals are; a. rapid interpretation of the magnitude of a scalar entity at a particular spatial location; b. rapid comparison of the magnitudes of different scalar entities; c. rapid comparison of the scalar entities across different regions of the map. Based on past research, I chose to investigate creating a texture of symmetrical units called texels. Each texel contains a fixed number of color-mapped hexagonal blocks representing each scalar entity. Users can dynamically choose the static variables to be visualized and the size of the texels. The research started as an experiment to visualize the United States Election results to represent the degree of variation in the results and the votes shared among the contestants. In addition to the election data my technique has also been applied to the United States census data, geographical and meteorological data to highlight interesting results.
Visualizing and Comparing Multivariate Scalar Data over a Geographic Map

by
Karthik Ramachandran

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Computer Science
Raleigh, North Carolina
2009

APPROVED BY:

__________________________  ________________________
Dr. Ben Watson             Dr. Robert St.Amant

__________________________
Dr. Christopher G.Healey
Chair of Advisory Committee
BIography

Karthik Ramachandran was born April 30, 1986 in Chennai, India. He received a Bachelor of Engineering degree in Computer Science and Engineering from Sathyabama university, Chennai, India in 2007. He is currently a Graduate student in the department of Computer Science at North Carolina State University.
ACKNOWLEDGMENTS

This thesis would not have been possible without the help and support of so many people. I would primarily like to thank Dr. Christopher Healey for his immense support and countless advice on my thesis. I would also like to thank the other two committee members, Dr. Ben Watson and Dr. Robert St. Amant for their time and support.

I would like to thank every single member of the Knowledge Discovery Lab for all their help and encouragement, and for making the lab very comfortable for me to work. I wish to thank all my friends here in the United States and in India for their support. Finally I would like to thank my parents and my family for their love and encouragement, and for supporting my education in the United States.
# TABLE OF CONTENTS

## LIST OF FIGURES
- vi

## 1 Introduction
- 1.1 Visualization ......................................................... 1
- 1.2 Problem Description .................................................. 4
- 1.3 Application Examples ................................................ 6
- 1.4 Proposed Approach ................................................... 7
- 1.5 Thesis Organization .................................................. 10

## 2 Visualization
- 2.1 Background of Scalar Visualization .................................... 11
- 2.2 Scalar Data visualization ............................................... 11
  - 2.2.1 Color Mapping ...................................................... 12
  - 2.2.2 Contour Mapping ................................................... 12
  - 2.2.3 Glyphs .............................................................. 13
  - 2.2.4 Additional methods ................................................. 13
- 2.3 Multi-Dimensional Issues ............................................... 14
- 2.4 Existing Techniques .................................................. 15
  - 2.4.1 Chernoff faces and Star Glyphs .................................... 15
  - 2.4.2 Parallel Coordinates ............................................... 16
  - 2.4.3 Bivariate and Trivariate Choropleth Maps .......................... 19
  - 2.4.4 Composition of Color with Texture .............................. 23
  - 2.4.5 Attribute Blocks .................................................. 24
- 2.5 The issues to be addressed ............................................ 27

## 3 Design
- 3.1 Motivation ............................................................... 29
- 3.2 Visualization Technique ................................................ 31
  - 3.2.1 Color Mapping ...................................................... 32
  - 3.2.2 Texels .............................................................. 32
  - 3.2.3 Hexagon Versus Squares .......................................... 32
  - 3.2.4 Advantages of the design ....................................... 41
  - 3.2.5 Disadvantages of the design .................................... 41

## 4 Implementation
- 4.1 Technical Details ..................................................... 43
- 4.2 Data Collection ........................................................ 44
- 4.3 Maps ................................................................. 45
- 4.4 Texture Mapping ........................................................ 46
- 4.5 User Interface ........................................................... 46
5 Results ................................................................. 47
  5.1 Election Results  ............................................ 47
  5.2 Census Results  ............................................. 51
  5.3 Meteorological Results ..................................... 60

6 Conclusion and Future Work  .................................... 66
  6.1 Future Work .................................................. 67

Bibliography ........................................................... 69
LIST OF FIGURES

Figure 1.1 Visualization of Arizona’s statistics ........................................... 9

Figure 2.1 Visualization using Chernoff faces ............................................. 17
Figure 2.2 Visualization of Quality of Life in South Carolina Using Polygonal glyphs 18
Figure 2.3 Parallel coordinates of random data .......................................... 20
Figure 2.4 Number of Olympic Athletes by County ..................................... 21
Figure 2.5 Color Model for Trivariate Soil Classification ................................ 22
Figure 2.6 Compositing Color with Texture to Visualize Multivariate Data ......... 24
Figure 2.7 Visualization of Water Balance using Attribute Blocks ................. 25
Figure 2.8 Legend for figure 2.7 ................................................................. 25
Figure 2.9 Uncertainty Visualization of January Temperature in India .............. 26

Figure 3.1 Democratic Presidential primary for the state of Wyoming ............. 30
Figure 3.2 Regular ways of tiling a 2D space .............................................. 34
Figure 3.3 Linear grouping of five attributes .............................................. 35
Figure 3.4 Symmetric grouping of five attributes in a brick-like square grid and hexagonal grid ................................................................. 36
Figure 3.5 Comparison of a brick-like square grid and hexagonal grid ............. 37
Figure 3.6 Grouping of three attributes in a hexagonal grid ............................ 38
Figure 3.7 Grouping of four attributes in a hexagonal grid ............................. 38
Figure 3.8 Grouping of five attributes in a hexagonal grid ............................. 39
Figure 3.9 Grouping of six attributes in a hexagonal grid ............................. 39
Figure 3.10 Grouping of seven attributes in a hexagonal grid ....................... 40
Figure 5.1 Republican Presidential Primaries for the state of Connecticut .......... 48
Figure 5.2 Democratic Presidential Primaries for the state of South Carolina ....... 49
Figure 5.3 Republican Presidential Primaries for the state of Utah ................. 50
Figure 5.4 Statistical results for the state of New Mexico ............................ 52
Figure 5.5 Statistical results for the state of New Hampshire averaged by the state ... 53
Figure 5.6 Statistical results for the state of New Hampshire averaged by the nation . 54
Figure 5.7 Statistical Results for the state of Oregon ................................. 56
Figure 5.8 Variations in average values based on population distribution .......... 57
Figure 5.9 Statistical results with population for the state of Oregon ............... 58
Figure 5.10 Statistical results for the country of India ............................... 59
Figure 5.11 Visualization of the Precipitation dataset of the United States ........... 61
Figure 5.12 Visualization of the Temperature dataset of the United States .......... 63
Figure 5.13 Visualization of the Temperature dataset of Australia .................. 65
Chapter 1

Introduction

The world is moving quickly with technological advancements and creative innovations but in doing so it leaves behind large amounts of unused and unanalyzed data. These data often provide profound insights into various fields of study if analyzed and evaluated in the right manner. But the task is not as simple as it sounds because of the sheer vastness of available data and the confusion of how to represent it. Collecting the data and representing it in a visual format is what constitutes the broad area of research called Visualization.

1.1 Visualization

In a nutshell visualization is the study of converting large scale numerical data and text into easily understandable visual representations. The procedure of visualization starts with identifying and acquiring the required data. Parsing the data to make it readable by the computer is the next step. Unwanted and unnecessary data should be removed from the original dataset. The data can be manipulated in various ways ranging from simple tasks like rounding off whole numbers to complex tasks like derivation of new results. Next come the most important tasks of representing the data in a visual form and refining the visualizations as and when needed. Additional work is needed, if the visualization designer wants the user to interact with the visualization [15]. Visualization is a complex process.
that requires significant effort and thinking to produce functionally efficient and aesthetically effective visual representations.

One important advantage of visualization is its use of our capabilities of our human visual system to interpret and understand astronomical amounts of data in a small amount of time. Interpretation of data does not stop with locating individual values in the data, but also identifying hidden patterns, which would not be visible if not in a visual format. Research says that humans can acquire more information through vision than through all the other senses combined[39]. Finally, visualization helps in formulating new hypothesis, and aids in the genesis of new ideas and concepts [16]. A visualization could mean something as simple as a two dimensional bar chart representing simple data or a three dimensional animation representing complex data. Selecting the right kind of representation depends on various issues like the type and quality of the data, availability of software/hardware to create the visualizations, amount of time needed to complete the visualization, and of course, efficiency and effectiveness of the visualization technique for the particular problem. The objective of producing visualizations is to allow the user to interpret large sets of data in a visually compelling and easily comprehensible form, and this demands visualizations to be perceptively simple and vivid. Human vision and perception influences many visualization designer’s choices, because it is believed that visualization is the analysis of the mental image of the visualization rather than the physical image itself [14]. The efficiency of an algorithm can be found by measuring its running time and storage requirements, but it is more difficult to extend the same standards for measuring the efficiency of visualization.

The field of visualization can be broadly classified into two different categories called Information Visualization and Scientific Visualization. Information visualization is often described as the study of creating graphical techniques to help people understand abstract or non-spatial data [17]. Various concepts drawn from different fields like computer graphics, human–computer interaction, cognitive psychology, semiotics, graphic design, cartography, and art coupled with interactive computation has led to the increase in use of information visualization [18]. Information visualization often employs methods to spatialize textual data and implement user-interfaces, and is increasingly used in public arenas and forums to communicate information effectively.
Scientific visualization involves in building visualization techniques to visualize scientific or “real world” data, which usually include an explicit spatial dimension. It is extensively used in research laboratories and R&D facilities to observe computational representations of real phenomena. The significant characteristics of scientific visualization are the representation of large sets of data, the aim to achieve the verisimilitude of natural phenomena, the use of animated sequences to show the progression of time, and the demands for powerful computers to perform high speed computation [19]. Scientific visualization is an active field of research, and some of the current research issues include improved volume graphics, multiresolution modeling, visualization of tensor fields and automation of visualization design [20].

There is often a debate in the visualization community of whether Scientific visualization and Information visualization are completely exclusive and distinct subfields of visualization. The main difference between the two being that scientific visualization is used to represent physically based or spatial data, and Information visualization is used to visualize non-physical data [21]. Some researchers claim that Scientific visualization is aimed at an eclectic technical and mostly small audience while Information visualization is aimed at a wider audience group that includes the public [22]. These researchers make a distinct classification between the two because of these differences, while others claim that the choice of spatialization of a particular dataset is the explicit choice of the visualization designer [18]. The necessity for a new classification of visualization based on the types of spatialization, and not on whether spatialization is used or not has also been suggested [23].

Geovisualization or Geographic Visualization is a popular concept of visualization, which supports the visualization and analysis of geospatial data. Geovisualization does not specifically fall under scientific visualization or information visualization but borrows concepts from both to convey geographic information [24]. A very rudimentary example of a geovisualization would be a world atlas. A world atlas contains all the countries and oceans, with the neighboring countries shaded with different colors and the ocean and water bodies colored blue, enabling us to easily identify individual countries. A World atlas is a visualization of a data file, which would contain the names of countries and oceans with their
geographic coordinates. Current challenges in geovisualization include using it as a tool to address critical issues in public health, environmental science and crisis management, by collaborating concepts from cartography and GIS with those from computer graphics, information and scientific visualization, computer-supported cooperative work, diagrammatic reasoning, cognitive science, HCI, cognitive systems engineering, and other domains, to produce effective geovisualization tools [25]. Geovisualization is a rapidly growing science, thanks to the increase in the quality and quantity of geospatial data, locate-aware computing, advanced databases and data-mining, human-computer interaction with geospatial information [28]. Geospatial data can be described as any data that combines the geographic location with its characteristics [29]. If the geospatial data contains multiple information about each geographic location, then it is called multivariate geospatial data.

Current technologies allow us to collect data in a plethora of ways, from satellites to small thermometers fixed to the ground, but it is almost impossible to guarantee the quality of the data or the verity of the data to be 100% accurate. In some cases these errors, if small and negligible can be tolerated, but in high risk situations, which affect a large number of people, a comprehensive study of the uncertainty in the data is mandatory [26]. There are three possible sources of uncertainty: incomplete information, inadequate understanding and undifferentiated alternatives [27]. A simple example of uncertainty in data would be weather information from two or more sources. One weather source might report a particular temperature reading at a place, but another weather source might report a different temperature reading for the same place at the same time. We would have no idea which is reliable. One possible solution would be to average the data, but as the sampling size increases, so does the probability of coming up with the wrong average. Visualization of uncertainty can often result in better decision making and better understanding of a problem [2].

1.2 Problem Description

The process of visualizing multivariate data is called multivariate data visualiza-
tion, and is primarily used in spatial analysis. It deals with the visualization of high- dimensional data in a low dimensional environment such as a computer screen or a sheet of paper which is often achieved by attaching features such as color, texture properties like spatial locations and size, to each data element [30]. Two questions have to be answered before pursuing multivariate data visualization; first, is it reasonable or logical to display all data elements with all their associated values? Second, how can all the data elements be displayed in a low-dimensional environment? [31]. There are four major issues with multivariate data visualization. First, visualizing the high dimensionality of the data set. Second, deduction of the different hidden relationships existing between the variables. Third, a requirement of experts to provide insight on the derived relationships. Fourth, a requirement for exceptional computational means to support fast execution time [32].

This thesis mainly concentrates on showing multivariate scalar data over an underlying geographical map. This technique is used to visualize multivariate data in which one would expect the attributes to correlate with one another. There has to be an expected correspondence to how the attributes vary across the geographic region with respect to one another. The primary goals of the thesis are; a. rapid interpretation of the magnitude of individual attributes across a geographical region; b. Rapid comparison of the magnitudes of two or more attributes; c. Comparison of the attributes across different regions of the map. The resulting visualization should capture the essence of the dataset in terms of the various relationships existing between the scalar attributes. The user should be able to clearly identify and single out a scalar variable and deduce its magnitude at all points of the geographic map and at the same time explore relationships and extract patterns in the multivariate dataset. Comparisons should be possible across the different regions of the map, to get a holistic sense of how the data values are distributed across a particular geographic region.

The interesting point about this kind of multivariate scalar visualization is the fact that the variables have to be understood as an individual entity and also as a group of related values. This demands a design which uses the same type of mapping method for all the variables. Only then would one be able to compare the magnitude of two separate variables. Care should be taken to ensure that interpreting the magnitude of the individual
variables does not interfere with interpreting the underlying relations between the variables. In order to support spatial data, it is absolutely mandatory that the design built by the visualization designer does not obscure the underlying relationships and patterns existing between the variables but subtly highlight them.

1.3 Application Examples

Finding the right kind of data to test the proposed visualization technique was just as interesting as developing the visualization technique itself. Any evaluation of a visualization technique requires appropriate data sources. The data we needed had to be multivariate and be related in some way. This part might need more clarification. Let there be a dataset that contains two scalar meteorological data attributes, average temperature and average windspeed for points sampled at $\frac{1}{2}^\circ$ steps in latitude and longitude. This is a legitimate multivariate dataset with two scalar data variables but this would not be effective for this research because temperature and windspeed do not correlate with each other. The resulting visualization will have nothing interesting to offer. On the other hand if we had the altitude for every sampling point instead of windspeed, then this technique can be used to visualize the dataset because there is an expected correlation between altitude and temperature. Temperature decreases as altitude increases and this technique can be used to find interesting patterns and anomalies with the data.

The summer of 2008 I was working on a project to visualize the United States election data. The objective was to build visualization techniques to represent the degree of variation in the results, and the votes shared among the contestants in each electoral constituency, rather than just merely representing the winning party. The visualization of the election data clearly showed us the uncertainty that existed amongst the voting demographic of each state. Glaring differences between the victorious and the defeated candidates, tight races in counties, similar trends of party inclinations between the neighboring counties, and rural/urban differences, were all conspicuously visible and decipherable without much effort from the viewers part.
Although useful, the election results data do not offer much in terms of uncertainty or multi-variation. For example, in every county there were clearly only two front-runners with the third candidate well behind or non-existant. This prevented the visualization of three scalar variables which are close in magnitude. Census data provided this missing quality of the election data in terms of uncertainty and multi-variation. Theories of social balances have existed for a long time. For instance, the idea that the level of affluence and employment opportunities are high in an area which has a high literacy and education rate, or the idea that an area with a highly educated population would indicate towards a higher median household income, seem convincing and for the most part true. But these relationships are not always the same and sometimes the differences can be so astronomical that it could defy all rules of social sciences. The visualization of these multivariate census data provided for some very interesting observations of the geographical areas exhibiting predictable results, as though validating social theories, and those exhibiting surprising anomalies. Finally, our technique has also been used to visualize various meteorological grid datasets at $\frac{1}{2}^{\circ}$ resolution.

1.4 Proposed Approach

Visualization has a long history of using colors as one of the primary visualization tools. The visualization designers have taken advantage of our ability to recognize and distinguish colors. The number of different ways colors have been exploited in the rich history of visualization stands testament to the fact that color-mapping is the most used tool of visualization. Three primary goals have to be met for a successful visualization of data using colors. The use of colors should allow for rapid and accurate identification of individual data elements. The color of the target element must be easily found with respect to the color of other non-target elements. Lastly, the right number of colors that can be displayed while still allowing for rapid and accurate target identification must be found [33]. The reason color is so commonly used is because as human beings we are familiar with colors. We can say how far or how close two colors appear to one another. In some cases we also
tend to think of certain entities in terms of colors. For example we tend to associate Heat with the color Red and Cold with the color Blue. This is one reason why we often visualize temperature by assigning shades of red for higher temperatures and shades of blue for lower temperatures.

The technique that I have designed also uses color-mapping as the primary technique. I started this research without any fixed ideas or preferences for a method. I explored different kinds of techniques, ranging from glyph based techniques to contour mapping, analyzing an entire gamut of visualization styles. Assigning color values to represent data is called color-mapping or pseudo-coloring, and it was finally chosen as the most effective and appropriate technique to visualize the multivariate scalar data.

The proposed technique uses texture mapping to map a texture over a geographic map. The texture itself consists of tightly knit symmetrical blocks called texels. Each texel in the texture contains a fixed number of hexagons arranged in a particular order to maintain the symmetry of the design. The individual hexagons themselves are shaded with particular color values representing the magnitude of the scalar data at that particular spatial location. Unique color ranges are chosen for the multivariate data based on the data’s type and values. The number of color-mapped hexagons in a texel is equal to the number of multivariate scalar data variables. The position of an individual hexagon inside a texel uniquely identifies the scalar variable that it represents. All the hexagons in the visualization are shaded with colors from the same color range to ensure easy comparison between the scalar data.

Figure 1.1 is the result of applying this technique to the federal statistics of the state of Arizona. Four statistical attributes are used: Number of persons above the age of 25 with a bachelor’s degree, Median Household income, level of affluence and employment. The block patterns in the map are the texels and they contain the four hexagonal blocks representing the four attributes as shown in the legend. The attributes of each county are compared to the state averages. If the values of the statistical attributes are higher than the state averages, then they are shaded with the colors from the positive side of the color scheme (Red), or otherwise from the negative side (Green). Counties with all the four
Figure 1.1: Visualization of Arizona’s statistics
attributes in shades of red indicate a progressive and affluent trend because the level of education, median household income, level of affluence and employment are higher than the state averages. Counties with all the four attributes in shades of green represent a statistically negative trend because the values of the attributes are lower than the state averages. Counties with a mix of red and green shades are particularly interesting because their statistics deviate from the expected social and economical relationships. In the figure 1.1, the counties in the center of the state indicate a very positive statistical presence because shades of red are prevalent across them. All the other counties are predominantly colored with shades of green which indicate a negative statistical trend. Expectedly the counties with large cities like Phoenix are the most statistically progressive of the state with higher number of persons with a Bachelor’s degree, higher median household income, higher employment and higher level of affluence than the state averages.

1.5 Thesis Organization

The remainder of the thesis is structured as follows. Chapter 2 describes significant past research in the field of scalar visualization and multivariate visualization. Chapter 3 details the concepts of this technique’s design and the advantages and disadvantages of this technique. Chapter 4 describes the technical implementation of the technique. Chapter 5 provides the results of applying this technique to different datasets. Chapter 6 discusses the conclusion and the future work.
Chapter 2

Visualization

2.1 Background of Scalar Visualization

Visualization can be used to represent different kinds of data. There are many ways in which data can be classified. There is a distinct dichotomy between dependent variables and independent variables. There are also other classifications based on the dimensionality and type of data. In the domain of scientific visualization, visualization can be broadly classified into three types in terms of the data used: scalar data visualization, vector data visualization, and tensor data visualization. Exclusive techniques exist to represent scalar data, vector data, and tensor data, and some researchers have also built designs to represent both scalar and vector data in the same display [41, 44, 43, 42]

2.2 Scalar Data visualization

A scalar variable is an entity that has only a magnitude, or only one single value as opposed to a vector variable which has both a magnitude and direction. Examples of scalar variables are attributes like temperature, humidity, poverty rate etc. The temperature at a given location has a numeric value in degree Celsius or degree Fahrenheit, but no direction. On the other hand vector variables like wind have both a magnitude and
direction. The wind at a particular location is quantified by the force of the wind measured in kilometers/hour or miles/hour, and the direction in which the wind flows. This research deals only with visualizing scalar data. Numerous ways exist to visualize simple scalar data and a brief description of some of them follows.

### 2.2.1 Color Mapping

Color is the visible wavelengths of electromagnetic energy ranging from 4000 Angstroms to 7000 Angstroms. The wavelengths near 4000 Angstroms are violet in color and the wavelengths near 7000 Angstroms are red in color. The 4000 to 7000 Angstroms range forms the Violet, Indigo, Blue, Green, Yellow, Orange and Red (VIBGYOR) color scale [1]. The most commonly used color model to produce colors in computer graphics is the RGB model. RGB or Red-Green-Blue is an additive color model which manipulates the amount of Red, Green and Blue added to produce different colors. 100% of Red, Green and Blue produces white and 0% of Red, Green and Blue produces black. The use of color in scientific visualization and cartography can be classified into two broad categories, choropleth maps and filled isoline maps. In Choropleth maps, the presence or absence of different colors at different places indicate the presence or absence of particular scalar attributes. In filled isoline maps, bands of color represent the intensity of a scalar attribute [52].

### 2.2.2 Contour Mapping

Contour lines or contour mapping is a popular technique in scientific visualization and cartography. A contour line or simply a contour is a line connecting points of equal magnitude, or a line or curve along which a function has an equal value [5]. It is most often used in generating topographical maps to represent the elevation at various locations of the map. The contour line connects points of equal elevation. The empty space between two successive contour lines is defined as the contour interval, and it represents the difference in the magnitude between the two contour lines [3]. Traditionally contour lines were used to
represent elevation, so much so that contour mapping actually meant topographical mapping. But modern cartographers and scientists use contour lines to represent a variety of geographical features and meteorological data, mainly because of the flexibility of contour lines to represent spatial locations of equal magnitude. Generating contour lines is fairly straightforward. It starts with superimposing a rectangular grid over a geographical map, noting the magnitude estimates at each node of the grid, calculating a weighted average obtained from the data points, and finally interpolating the values at the grid nodes to connect points with equal magnitude with straight or curved lines [4].

2.2.3 Glyphs

A glyph in the context of visualization is any small shape or symbol employed as a part of the design. Glyphs are often used in the visualization of vector fields for their ability to accommodate the dimension of direction. Glyphs in the form of arrows are traditionally used to represent vector fields with the size of the arrow mapped to the magnitude and the direction at which the arrow is pointing, mapped to the direction of the vector field. Glyphs are also used in representing scalar attributes. The extensive usage of glyphs stems from the fact that it can represent multivariate data. Various attributes of a glyph like its size, shape, color, orientation, density etc can be connected to different variables [6]. Even simple changes to glyphs can be noticeable. Glyphs can be any geometrical shape, and so they can be custom designed to cater to particular kinds of data attributes to increase the efficiency of the visualization.

2.2.4 Additional methods

The methods discussed above represent the simplest scalar visualization approaches. Technology has paved the way for new and exciting features for visualization like animation and sound. Sound is something that appeals to our senses just as color does, but through different organs. The presence of a sound or the change in pitch or timbre of a sound can
often be mapped to a scalar attribute to produce innovative visualizations. The advantage with sound is that it augments our optical capabilities, because we can see and hear things at the same time. We can also process them at the same time [7]. Sound can even be used to represent multivariate data by mapping different data elements to different attributes of sound like pitch, damping, loudness, duration, etc [8]. Animation has been widely used in the field of visualization. Animation can be taken to mean any kind of movement or the visible appearance and disappearance of certain entities of the visualization. Time-lapse variations display the change in values of attributes over time. Animation can be used to play the results at each time interval in frames like a video to observe the changes over time.

2.3 Multi-Dimensional Issues

The variety of scalar visualization techniques discussed in the previous sections are all efficient techniques, but the success of extending these techniques to represent multiple scalar attributes remains dubious. The obvious way to handle the multiplicity issue would be to use a combination of two or more of the previously discussed techniques. As mentioned before glyphs can accommodate multiple scalar values by mapping each of the variables to the different features of the glyphs. But it is not always easy to compare the magnitude of one scalar attribute with another because of their varied representations. For instance, it may be hard to compare the color of a glyph with its height or regularity, in terms of determining the exact attribute values used to produce the glyph’s color, height and regularity.

Since Color-mapping has been shown to be an efficient technique to visualize scalar data, we chose to investigate using colors alone to represent our multivariate attributes. To use color-mapping for our technique, the color range used to visualize one attribute should be the same as the ones used to visualize the other attributes. Only then would one be able to rapidly compare the magnitude of two or more of the scalar attributes. This creates the problem of how to order the multivariate attributes on the map and how to differentiate them all. The attributes are continuous which means that they have a finite value across all
parts of the map. The easiest solution would be to create multiple maps, each map shaded according to the magnitude variation of one of the multivariate attributes. Comparisons between the attributes could be made by physically comparing maps with one another, but side-by-side comparisons have been shown to be ineffective.

To tackle the problem of multiple maps some researchers suggest the concept of integral and separable displays. The multivariate data should be checked to see if its visually integral or visually separable before choosing designs to build a geographic multivariate visualization technique [48]. Integral displays are used to represent integral variables which are seen as a collective entity rather than as individual variables. Separable displays are used to represent separable values which do not relate to one another but exist only as separate attributes [50]. Integral displays usually involve integrating all the variables into one single object while separable displays separate each variable into separate objects [10]. Several techniques exist to represent visually integral and visually separable data but there are many fewer designs to represent data that must be integrally and separably understood at the same time.

2.4 Existing Techniques

Although this kind of multivariate scalar data visualization is not ubiquitous, researchers have come up with visualization techniques to analyze multivariate scalar data in a spatial context. A brief description of some of the popular ones follow.

2.4.1 Chernoff faces and Star Glyphs

Chernoff faces are small face like glyphs that use facial features to convey different ranges of data [11]. Chernoff developed Chernoff faces in 1973 with 18 different kinds of features representing 18 dimensional data. Change in parameters like the length of the nose, radius of the face, size of the eyes, etc were mapped to different scalar variables to perform
the visualization [12]. It was widely appreciated for its creative use of human faces, and also easy to understand, because of our capacity to quickly identify different kinds of facial features. Unfortunately it does not provide sufficient scope for use in a detailed analysis, since the features are constrained to a small fixed range of positions. To express a variable with a large range is not practical and can result in inaccurate analysis. Chernoff faces also require significant space. To effectively view all the qualities of the facial features each glyph has to be clearly visible. Another limitation of Chernoff faces is that comparison of two values can be difficult since the variables are mapped to different types of parameters. A final drawback of Chernoff faces is that there are interdependencies in the parameters, that is, when certain parameters exhibit high magnitudes of data, parameters exhibiting lower magnitudes can be completely obscured [13]. Figure 2.1 is the result of using Chernoff faces to visualize four attributes. The size of the eyes is mapped to the percentage of service employees, size of the nose mapped to the percentage of electronic voting, the shape of the mouth mapped to the percentage of adult employment and the shape of the face mapped to the median housing prices.

A “whisker plot” is a glyph in which each variable is represented by the length of a line segment originating from the center of the glyph [11]. A variation of the whisker plot is the star glyph where the ends of the adjacent line segments are joined together. Star glyphs can accommodate any reasonable number of variables [14] like Chernoff faces though star glyphs have drawbacks as well. The range of values represented by the variables is directly proportional to the size of the star glyphs. If the variables have a large range of values to represent, the length of the rays of the star glyphs have to be long to accommodate the range, which in turn increases the size of the star glyphs. Extensive scrutiny of the star glyphs is required to correctly interpret and compare the magnitudes of the attributes. Patterns and hidden relationships do not stand out to the human eye due to the often uneven and geometrically varying shapes taken by the star glyphs. Figure 2.2 is the result of using star glyphs to visualize the quality of life in the state of South Carolina.

2.4.2 Parallel Coordinates

Parallel coordinates is a technique that was proposed in the 1980s for represent-
Figure 2.1: Visualization using Chernoff faces
Figure 2.2: Visualization of Quality of Life in South Carolina Using Polygonal glyphs
ing high dimensional data in a low-dimensional environment. It is a fairly straightforward method in which each data dimension or data variable is represented as a vertical axis. Parallel coordinates for $n$ dimensional data has $n$ axes as uniformly spaced lines. A data element is represented as a polygonal line crossing each axis at a point proportional to its value for that dimension [34]. Parallel coordinates can be used to analyze data distribution and functional dependencies [35]. One of the main limitations of parallel coordinates is the limited space available for each parallel axis [36]. In the case of visualizing a large dataset the visualization can become very cluttered which can render identification and interpretation impossible [38].

There have been extensions to parallel coordinates to support more specific forms of multivariate data. A simple variation of parallel coordinates is to transform it into a circular version in which the axes originate from the center and extend to the perimeter of the circle. It can be thought of as a form of overlapping star glyphs [37]. Hierarchical parallel coordinates visualize large data sets by hierarchical clustering the data, and displaying derived aggregation information [35, 34, 38]. The polygonal lines representing data dimensions are usually colored for the easy identification of the data. Figure 2.3 is the result of using the technique of Parallel Coordinates to visualize a 6 Dimensional random data.

2.4.3 Bivariate and Trivariate Choropleth Maps

Bivariate and trivariate choropleth maps are used to visualize bivariate and trivariate scalar data. Colors are used to represent the magnitude of the variables and to distinguish them from one another. A color scheme is built to accommodate all possible magnitude combinations of the variables. Each spatial location of the map is shaded with the color corresponding to the magnitude pairs (bivariate data) or the magnitude triplets (trivariate data) of the color model. The colors chosen should be distinguishable and the transition of the colors must be smooth and visually coherent [45]. Choropleth maps facilitate recognizing patterns, as well as easy interpretation of the variables as individual entities, and as a set [46]. Bivariate maps have been found to be effective tools to recall patterns, and are preferred to two distinct univariate maps [45]. The color schemes for bivariate choropleth maps usually take the form of a square whereas the color scheme for
trivariate maps take the form of a tri-axial graph [47]. Researches have performed various extensions to trivariate choropleth maps by using textures instead of colors for one or two variables [49]. Figure 2.4 is a bivariate map representing the number of Olympic athletes in each county. The $x$ axis of the color scheme is mapped to the ratio of winter Olympic athletes to summer Olympic athletes and the $y$ axis is mapped to the per capita of athletes. Figure 2.5 shows a tri-axial color scheme developed based on [47] to represent a trivariate dataset to classify soil.

Although bivariate and trivariate maps are frequently used to communicate information, they have often been criticized for their inability to effectively present individual distributions and the correlations between them [51]. The legend has to be extensively referenced to properly interpret the color scheme of the maps [45]. A trivariate map can be used only for three variables that add up to 100% [51, 52]. The interpretation of a tri-axial graph to study the color scheme of a trivariate map can be difficult compared to other trivariate visualization techniques. Normally multivariate choropleth maps cannot be used to represent more than three variables.
Figure 2.4: Number of Olympic Athletes by County
Figure 2.5: Color Model for Trivariate Soil Classification
2.4.4 Composition of Color with Texture

Researchers continue to study the problem of multivariate visualization and new techniques are frequently introduced. [40] proposed a novel technique of combining colors and textures to represent multivariate data. The research suggests two methods to segment the texture to add color, Intensity-based Texture Segmentation and Texture Segmentation into Meaningful Blobs. Intensity-based texture segmentation uses the histogram of the texture pattern intensities to divide the texture image into \( n \) segments, \( n \) is the number of scalar data in the dataset. The second method decomposes the texture into multiple components based on the local orientation of directional elements of the pattern. The research proposes coloring the multiple segments of the texture with different colors to represent different scalar variables. The concept of ‘color weaving’ [41] is suggested to efficiently combine the colors with the texture segments. The result would be a texture with various segments shaded with different colors representing different scalar variables. Figure 2.6 is the result of combining four colors with two different textures to visualize five scalar variables. The first variable is mapped to the choice of the texture pattern and the remaining four variables are mapped to the presence or absence of colors in the texture. In figure 2.6, one of the textures (over the state of Colorado) is segmented based on the histogram of the texture intensity, and the traits 3 and 4 are represented by the colors in the two directions. The other texture is segmented based on the local orientation model, and the traits 3,4,5 and 6 are represented by the colors in the four directions.

This technique can be used to build aesthetic visualizations by combining color with natural textures. The implementation of the technique is quite simple. The viewer also can easily discern the information represented in the data. One drawback of this technique is that it can be used to represent only Boolean data, that is, the presence or absence of a particular color or texture only reveals a presence or absence of an attribute and not a particular magnitude. A simple extension to deal with this problem would be to use color ranges instead of discrete colors. In figure 2.6, color ranges from white to blue or white to green can be used to represent the magnitude of a data variable. To enable the user to easily determine magnitude, the texture segments must be above a certain size. It is also
possible that color bands overlap at some parts of the texture as shown in trait 2. These locations would be particularly difficult to interpret. Identifying patterns and recognizing spatial relationships among variables is an important feature of multivariate visualization, but this technique will preclude the user from locating patterns due to the multitude of colors and textures used in the visualization.

Figure 2.6: Compositing Color with Texture to Visualize Multivariate Data

2.4.5 Attribute Blocks

Miller designed a new technique to visualize continuously defined attributes across a geographic region using what he called “attribute blocks” [53]. $n$ attributes are arranged into an array of dimension $k_r \times k_c$, the “attribute block” array. Each attribute is assigned a color range. The attribute block array acts as a “screen door” of lenses tiled across the entire map. The attribute value at a specific location can be interpreted from the color of the attribute block in the array. One goal of this effort was to see how many attributes...
could be visualized using a single planar projection. The \( k_r X k_c \) blocks of the attribute arrays need not necessarily be unique. A single attribute can take up multiple (mostly consecutive and adjacent) blocks in the array. Users are allowed to dynamically alter fields of view, attribute block sizes and positions, and configurations. These dynamic interactions can reveal hidden attribute values and hidden patterns. The size of the attribute blocks can be as small as 1 pixel to display a visualization with the feel of pointillism. Miller also talks about using attribute blocks to tackle uncertainty visualization as shown in figure 2.9.

![Figure 2.7: Visualization of Water Balance using Attribute Blocks](image1)

![Figure 2.8: Legend for figure 2.7](image2)
Figure 2.7 shows the result of visualizing water balance using four attributes; Potential Evapotranspiration (Cyan), Soil-Moisture Holding Capacity (Green), Temperature (red) and Precipitation (Blue) for the month of January. Figure 2.8 is the legend for figure 2.7. This visualization successfully highlights that precipitation has a strong effect in the west and northwest because of the dark blue blocks in these regions; soil-moisture holding capacity has a strong influence in the upper Midwest because of the dark green blocks in the upper midwest; Potential Evapotranspiration has an influence in the water balance only in Florida, Mexico and the Baja because of the highly saturated cyan colored blocks. Figure 2.9 shows the visualization of January temperature uncertainty in India. Temperature readings from various sources are arranged in the attribute block array and are assigned the same color scale. The disagreement in the temperature values can be identified by the appearance of irregular block patterns, which mean that the temperature readings vary from one another. As is evident from the figure, considerable disagreement exists in the Northeast area of the map and some small disagreements occur in the northwestern parts.

To understand a multivariate dataset in its entirety the user has to change the size of the attribute blocks and continuously move the attribute block arrays to find the underlying patterns and hidden data. A systematic study of color models is required and care should be taken to ensure that a color mapped to one variable does not belong to the
color range of another variable. This implementation will be most effective for continuous variables which have prominent values across a limited area of the entire map. For example in figure 2.7, we can identify the high precipitation influence in the northwest because other attributes are not dominant in those areas. Having an area with multiple bright colors could make it more difficult to correctly interpret the values of the attributes and analyze patterns in the area. This implementation will provide symmetric patterns only when \( k_r \) equals \( k_c \). Attribute rows and attribute columns can be used instead of attribute block arrays but their effectiveness is untested. Assigning the same attribute to multiple blocks in the array can result in uneven and asymmetric patterns with attributes that dominate obscuring the other attributes. The uncertainty visualization using attribute blocks is effective only for two variables. For three or more variables its hard to say which particular variable differs from normal values.

2.5 The issues to be addressed

Many multivariate visualization techniques have been discussed in the previous section. Most of these techniques cater to very general kinds of multivariate data. To date there are still few multivariate visualization techniques that represent a dataset with expected correlation between the attributes, while still providing the ‘aesthetics’ that visualization designers desire. In the pursuit of such quests it is easy to lose sight of the most important quality that a visualization technique must exude, easy and effective communication of the intended information.

The time taken to interpret a choropleth map is directly dependent on the background color of the map [56] and on how it interacts with the rest of the colors in the map. Traditionally lighter colors have always been used to represent lower magnitudes and darker colors have been used to represent higher magnitudes regardless of their natural implications [54]. For example, if the number of sunny days was visualized across a region, it will be more effective to represent higher numbers with shades of white and lower numbers with shades of black which would help us interpret the information more easily by connecting
the color range with our mental picture of a sunny day and a cloudy day. The proposed visualization design should have a flexible color scale to suit the different kinds of datasets used and as much as possible should try to assign color values that will support a mental connection between the attribute and the color.

Past research has shown that multivariate visualization techniques using glyphs do not always allow an accurate interpretation of the data. The accuracy is higher for the analysis of an entire glyph with all its attributes at a specific location than across a region. On the other hand the accuracy is lower for the analysis of a part of the glyph (for just one attribute) at a specific location [55]. This could lead to serious problems in comparing trends between an individual location and a region. These issues have to be addressed in our proposed design.

Victoria Interrante’s ‘Composition of color with texture’ [40] technique can be used to visualize multivariate traits, that is, to show the presence or absence of particular characteristics but it is not effective for visualizing the magnitude of the attributes. Even if it were extended it would still be difficult to compare the magnitudes of two attributes, one of which is represented by color and another represented by texture. The proposed design should be able to represent the presence and absence of certain attributes and also be able to represent the magnitude of the attributes. James Miller’s “Attribute Blocks” [53] technique produces asymmetric attribute block arrays for certain number of variables, which affects recognizing patterns in the dataset. The “Attribute Blocks” technique does not address the issue of representing multivariate uncertainty variables. Figure 2.9 shows the disagreement in the temperature readings from various sources but it does not tell us which sources differ from each other and by how much they differ. Since attribute blocks use the same color ranges, it is also not possible to identify an individual source. These issues must be addressed by the proposed design. Finally our proposed design must accomplish the goals presented in the problem description.
Chapter 3

Design

This chapter will briefly explain the various methodologies involved in the design phases of the thesis. Just like any other research, there were numerous issues that arose in conceptualizing the design. This chapter will address these issues to provide a holistic understanding of the problem. An explanation of why this design was chosen over other choices is also provided. The chapter ends with a recounting of the advantages and disadvantages of our new design.

3.1 Motivation

As previously stated, this research started as an offshoot while working on the visualization of election results. An array of colored blocks was used for the visualization. Each block assigned to one of the top two democratic or republic presidential candidates. The colors of the blocks ranged from the background color for zero votes to fully saturated blue or red for 100% of the votes for democratic and republic candidates respectively. Some blocks in the texture contained small symbols to identify which candidate they represented. When the map of a state is viewed, each county contains blocks colored according to its particular voting results. The counties where the results were close are predominantly light blue or light red in color, indicating a tight race between the two contestants. The counties
where one of the contestants won by a huge margin appear tiled with blue (or red) and white squares, indicating a sweeping victory. In any county the winning candidate can be found by seeing which block is more saturated. This map served the purpose of indicating the winning candidate in each county and also the degree of variation in the results. Figure 3.1 shows the visualization of the top two Democratic primary candidates in the state of Wyoming.

![Visualization of the top two Democratic primary candidates in the state of Wyoming](image)

**Figure 3.1**: Democratic Presidential primary for the state of Wyoming

The same concept of using textures of continuously arranged square blocks could not be used for the visualization of the top three candidates in each county. The arrangements of three square blocks results in an asymmetric design. The arrangement of three squares produces straight lines or diagonals of the same color which dominates the texture, distracting focus away from the patterns and the relationships that exist between the three attributes. This was the main motivation to start building designs that could comfortably
support any number of multivariate attributes and yet be conducive to extract the information in a simple way.

3.2 Visualization Technique

The basic principle of this visualization is to generate geographic maps and apply a symmetric arrangement of texels over each geographic region to display its multivariate attributes. The design is a combination of color-mapping techniques and textures to create intuitive hexagonal patterns that highlight individual data attributes and the patterns they combine to form. The texture is automatically created for each geographic region by taking the multivariate variables at that geographic region as input.

There can be two types of visualizations that we support. One contains spatial regions with constant attribute values within each region. For example, election data and census data with state-wise or county-wise totals fall into this category. A mapping region is the geographic area to which a texture is mapped. In this case, the entire state or county is the mapping region and a common texture is applied repeatedly over that particular region. These visualizations are usually easy to interpret because the mapping regions are large enough to comfortably accommodate all the data variables.

The second type of visualization involves a dataset with continuously varying attributes. For example, meteorological datasets are continuously varying datasets with samples taken at small latitude and longitude intervals. This means that every geographic area with a given interval will contain unique attribute values. It is usually impossible to represent all the attribute values in that small amount of space, so only a subsample of the attribute value is represented. Visualizations of continuously varying datasets try to take advantage of the human eye to mentally interpolate missing values.
3.2.1 Color Mapping

Color-mapping is the main technique used in our design. The hexagonal structures of the texels are shaded with specific colors representing the attribute values at a particular location. The shaded colors are selected from a unique predefined color scheme. The colors are selected to reflect the characteristics of the data used, in order to visually and mentally associate the color with the data attributes. The colors are formed by assigning scalar functions of the data attribute to the red, green and blue components of the texture image. The RGBA texture mode is used to create the texels, with the alpha value varied to produce color ranges from the background color of the map to the full color taken from the color scale.

3.2.2 Texels

Texels form the heart of the layout strategy of our design. Texels are a tessellated group of color-mapped hexagon blocks that vary depending on the number of multivariate data attributes and their arrangement for the visualization. The hexagons are positioned in ways that would maintain the symmetry of the design. The texels themselves are arranged in the map to produce a seamless symmetric image. A texel is similar to a Chernoff face or an attribute Block array, in terms of how they all present the values of the data attributes at a particular geographic region.

3.2.3 Hexagon Versus Squares

Traditionally square or rectangular grids have been used to divide a 2D space. Past research shows extensive usage of square grids in the field of computer graphics and visualization. Research has also shown that squares are not the most ideal method of dividing a 2D space. Squares have been used because of their simplicity, and because of the existing models and concepts on square tiling.
There are three ways of tiling a 2D space: rectangular tilings, triangular tilings and hexagonal tilings [57]. For a tiling to be uniform, a point \( p \) in one tile should form a lattice with the duplicates of \( p \) in all the other tiles [58]. Triangles do not exhibit this property. Alternate triangles must be rotated \( 180^\circ \) to produce a continuous triangular tiling, therefore, a point \( p \) does not form a lattice with its duplicates. It is highly desirable that the grid be uniform to guarantee symmetry. This leaves us to choose rectangular tilings or hexagonal tilings. There are two ways of tiling a rectangular grid: a simple square grid or a brick-like grid as shown in figure 3.2.

Hexagonal tilings are increasingly preferred to represent flow direction, because the number of neighbors is important for representing flow direction. A square in a square grid has 8 neighbors with which it shares an edge or a vertex. A hexagon in a hexagonal grid has only 6 neighbors, and it shares an edge with all these neighbors. The distance between the squares and its 8 neighbors varies but the distance between a hexagon and all its neighbors is the same [59]. The flow direction is defined by the neighbor with the lowest elevation. This is well pronounced with hexagonal grids because the distance between two neighboring hexagons is constant. The results of comparing a hexagonal grid and a square grid for representing flow directions revealed that there were fewer errors in the hexagonal grid. Hexagonal tilings are shown to function more accurately in exploration and search strategies, because approximation algorithms are more consistent on hexagonal tilings [57].

One of the main reasons to opt for a hexagonal grid is because of its symmetry properties. Symmetry is an object’s ability to remain invariant in position and orientation when subject to a spatial translation. The process of building a hexagonal grid is easier than building a rectangular grid because of the required symmetry constraints. The orthogonal co-ordinate system of a rectangle simplifies calculations and transformations on a rectangular grid but introduces certain ambiguities because of its inconsistent nearest neighborhood problem [61]. The lack of proper methods to decompose a hexagonal grid was one factor that led to the rise in the use of rectangular grids. But researchers have recently developed simple and implementable methods to decompose hexagonal grids.
Figure 3.2: Regular ways of tiling a 2D space
Framing a square grid and a hexagonal grid reveals that they are both symmetric and can be used to tessellate a 2D space. The problem of multivariate visualization requires grouping a set number of tiles in a grid. For example, if the multivariate dataset has five attributes, then five tiles have to be grouped to represent the data. These groups form the texels. Texels must tessellate the 2D space while providing symmetry to the design. This is again possible with both square and hexagonal grids. A linear grouping of any number of tiles in either grid can be formed to tessellate the texels as shown in figure 3.3. Unfortunately, a linear grouping is not intuitive and can produce poor designs making interpretation of the data difficult.

![Square Tiling and Brick-like Square Tiling](image)

![Hexagonal Tiling](image)

Figure 3.3: Linear grouping of five attributes

The same five attributes can be represented in a square grid by using a brick-like structure as shown in figure 3.4. The texels can tessellate the 2D space without requiring rotations. This is an acceptable texture because it preserves the symmetry of the design. The distance between a tile representing an attribute in a texel and another tile representing the same attribute in a neighboring texel is the same for all attributes. This makes it easier to associate the tiles representing the same attributes in different texels. The same property can be achieved using a hexagonal grid. The texel again contains five hexagon that evenly
tessellate the 2D space.

![Symmetric grouping of five attributes in a brick-like square grid and hexagonal grid](image)

Figure 3.4: Symmetric grouping of five attributes in a brick-like square grid and hexagonal grid

The brick-like square grid looks almost the same as a hexagonal grid. They have similar properties. However, there are useful advantages to hexagonal grids.

i ) The symmetry of a hexagonal grid is not ambiguous. There is only one way in which hexagons can be tessellated in a 2D space unlike square grids which can be tesselated in multiple ways.

ii ) Each of the edges in a hexagonal grid is a common side shared by two hexagons. This is not the case in a brick-like square grid in which some of the edges form only half of the side of squares.

iii ) The distances between the center of a hexagon and the center of its 6 neighbors are the same. The distances between the center of a square and the center of its six neighboring squares are not the same. Imagine a unit square grid with sides measuring 1
unit. The distance between the center of the square grid and the grids to its left and right will be 1 unit, but the distance between the center of the square grid and the other four grids to its top and bottom will be approximately 1.1180 units as shown (figure 3.5). The distance between the center of a hexagon and the centers of its six neighboring hexagons is the same in a unit hexagonal grid and is 1.732 units.

Figure 3.5: Comparison of a brick-like square grid and hexagonal grid

iv ) The angle between the lines connecting the center of an individual hexagon and the center of its six neighboring hexagons is the same, 60°. The angle between the lines connecting the center of an individual square and the center of its six neighboring squares is not the same (figure 3.5).

These points suggest that the brick-like square grid results in a mathematically and geometrically inefficient pattern. On the other hand, hexagonal grids result in a geometrically integral design with more identifiable patterns.
Figure 3.6: Grouping of three attributes in a hexagonal grid

Figure 3.7: Grouping of four attributes in a hexagonal grid
Figure 3.8: Grouping of five attributes in a hexagonal grid

Figure 3.9: Grouping of six attributes in a hexagonal grid
Figure 3.10: Grouping of seven attributes in a hexagonal grid
3.2.4 Advantages of the design

The advantages of the design can be summarized as points in the following ways.

i ) The use of the same color scale for all multivariate attributes makes it easy to both interpret the magnitudes of individual attributes and to compare magnitudes, both between attributes in a texel, and over the same attribute in different texels.

ii ) The texels can contain any number of hexagonal tiles, but will still tessellate the 2D space forming a symmetric collection of texels.

iii ) The hexagonal blocks inside a texel are arranged in an intuitive manner to reinforce the symmetry of the design.

3.2.5 Disadvantages of the design

The disadvantages of this technique are as follows.

i ) The design cannot be used to visualize bivariate data. There is no way in which the hexagonal blocks can be arranged to contain two scalar attributes. In the case of bivariate analysis a square grid or a triangular grid must be used.

ii ) The design can visualize only scalar attributes. The design can be extended to visualize multiple vector fields but the extension itself would be difficult.

iii ) The size of the texel can be problematic. The larger the number of attributes, the larger the texel. Texels have to be at least a minimum size to enable successful interpretation of the data. The percentage of a texel allocated to each of the \( n \) attributes inside
a texel would be \(1/n\). A larger texel covers more spatial area, hence more samples in the continuously varying dataset. This results in subsampling which reduces the level of detail that can be captured.
Chapter 4

Implementation

This chapter discusses the various steps involved in the implementation of our design. It starts with a brief survey of the technical details. It addresses the collection of data, which includes the shapefiles and the multivariate datasets, followed by a description about the creation of the underlying maps. The process of texture mapping is discussed and finally the chapter ends with a description of the user Interface developed to interact with the visualization.

4.1 Technical Details

The project was created and developed in a machine running the Solaris 8 on a SunBlade 150. The monitor used is a 17 inch color CRT monitor. The project is OS specific, and it can be run on other platforms only after making changes to the programs.

C++ was used as the main programming language to create and implement the technique. OpenGL graphics API was used to create the texture image containing the texels, and to map these textures to the geographic maps. The user interface was created using GLui, an OpenGL user interface library.
4.2 Data Collection

This section briefly discusses the different processes involved in collecting the required data for the thesis. Two types of data were needed: shapefiles and multivariate data files. Shapefiles are used to create the geographic maps over which uniform textures are mapped. The shapefile is a collection of three or more files: a .shp file containing primary geographic reference data stored as feature geometry; a .shx file holding the index of the feature geometry and a .dbf file in DBase format containing the attributes of the shapes [62]. Besides these regular files, some shapefiles also contain .prj files (Projection Format files), .xml files (XML Metadata files) and .txt files (Text Readme File). The election visualization used two different United States shapefiles, one containing the feature geometry of all 50 states and another containing the feature geometry of the counties within each state. The research also uses a shapefile containing all the countries in the world to create a worldmap and a shapefile for the Republic of India.

The data for the election results were obtained from the CNN website. The website has detailed information about the 2008 election results separated by states, and inside each state by counties. The data that we required were results of the United States Presidential primaries by county. The results were copied into text files and parsed to generate an output file containing records of county name, its corresponding state name, and the names of the Democratic or Republic presidential primary contestants along with the percentage of votes they received.

A second requirement was to investigate our ability to visualize multivariate statistical data. The federal statistics website www.fedstats.gov provided all the information that we used. Fedstats provides a full range of statistical information collected from various federal agencies belonging to the federal government. The statistical information are grouped by states. Each state contains the average state statistics along with the statistics of its counties. The statistical attributes that we visualized were the percentage of population holding a Bachelor’s degree or higher (more than 25 years old), mean household income, percentage of the population above the poverty rate and the employment rate. A
data file was compiled using these statistics with each record containing the county name, its corresponding state name, and the four statistical attributes for the county. Similar statistical information was required for the Republic of India. The statistical handbook from the website of the state of Tamil Nadu, a southern state of India, provided the statistics for all the states in India.

Finally we compiled a collection of multivariate meteorological data. Global air temperature and precipitation data was obtained from the Center for Climatic Research, Department of Geography, University of Delaware. These global data sources were created by Legates and Willmott [63] from monthly and annual mean temperature and mean precipitation records from 24,941 stations for air temperature, and 26,858 stations for precipitation. The data contains the global monthly and annual mean temperature and precipitation values interpolated to 0.5° degree by 0.5° of the latitude/longitude grid. Additional meteorological data was obtained from the Climatic Research Unit, University of East Anglia [64, 65]. This is also a global dataset which contains eleven meteorological attributes, namely, diurnal temperature, frost frequency, precipitation, radiation, temperature maximum, temperature mean, temperature minimum, vapor pressure, wet day frequency, wind and cloud cover. These datasets were used to extract the meteorological data for the continental United States. This output datafile contains latitude-longitude pairs interpolated at 0.5° by 0.5° together with their corresponding meteorological attributes.

4.3 Maps

Maps were generated to draw the underlying geographic regions over which textures could be mapped. The shapefiles were used to generate the maps. I had access to already-existing files to read and parse the shapefiles, triangulate the shapes, and draw the maps. A separate program had to be written for reading and extracting the shapefiles and drawing them as line-loops to be used with grid datasets.
4.4 Texture Mapping

Once the maps are created, textures are applied over the maps. This completes the visualization process. The texture images for this visualization are automatically created after the program reads the multivariate input data. The texture image is a symmetric grouping of texels. A texel forms a group of neighboring color-mapped hexagons. A texel has a thicker line to identify its boundary with other texels.

The dimensions of the texture image is $512 \times 512$ pixels, enough space to correctly encode the texels into the texture image. Numeric values are assigned to the RGBA components of the texture to produce the required hexagon color. The A component is used to vary the color of the texel to the background color. The texture image is stretched or compressed to maintain symmetry based on the size of the underlying map.

4.5 User Interface

The user interface acts as a medium between the user and the visualization. It allows the user to choose and change various characteristics of the visualization. The user first chooses the map to be drawn. The multivariate data attributes are grouped into sets and represented using check boxes. Depending on the maps chosen different sets are enabled or disabled to reflect the multivariate datasets. The attributes within a dataset can also be selected individually, giving the user freedom to choose any combination of the attributes. The size of the texels can be increased or decreased. Regular transformation controls like translation, rotation and zoom are also provided. The user can also save the visualization as a JPG picture using a screenshot feature.
Chapter 5

Results

This chapter details the visualization results of applying our system to different types of multivariate datasets. The technique was applied to three types of data: election data, census data and meteorological data. A selected group of interesting results is displayed in the following three sections with brief explanations.

5.1 Election Results

This section highlights interesting visualizations from the 2008 Democratic and Republican presidential primaries. Figure 5.1 shows the results of the Republic primary in the state of Connecticut. The state has eight counties. The results of each of the county can be interpreted from its corresponding texels. The legend explains which hexagon block inside the texel represents which candidate. A close look at the figure reveals that Mitt Romney won only in the south-western county, but with a considerable margin. The northern county containing the city of Hartford has texels where two hexagons are shaded with the same color. This occurs because Mitt Romney and John McCain have the same percentage of votes. Mike Huckabee has a negligible presence across the entire state, with his strongest presence in the south-western county. The visualization shows John McCain winning six of the eight counties and tieing in another. The distribution of the results is similar across the state, with the exception of the south-western county. This is shown by the color
change of the hexagon representing John McCain along the border of the south-western and its neighboring counties.

Figure 5.1: Republican Presidential Primaries for the state of Connecticut

Figure 5.2 shows the result of the Democratic primary in the state of South Carolina. The three contestants Hillary Clinton, Barack Obama and John Edwards are shown in the legend. The map reveals that Barack Obama won in almost all of the counties. Hillary Clinton garnered more votes than John Edwards clinching second place in the state. Although John Edwards comes from South Carolina, he does not have a strong presence in the east-central and southern counties. He won only in the westernmost county. The dominant blue hexagons in the east-central and the southern counties indicate victory by a huge margin for Barack Obama. The counties in the northwest show a balanced vote share, but the victory for Barack Obama can be discerned by the slightly darker shade of blue on his Hexagons.
Figure 5.2: Democratic Presidential Primaries for the state of South Carolina
Figure 5.3 displays the result of the Republican primary in the state of Utah. The four contestants represented are Mike Huckabee, John McCain, Mitt Romney and Ron Paul as shown in the legend. Mitt Romney’s sweeping victory across the entire state is clearly evident. He won in every county by a huge margin. Any presence of the other contestants is not seen in the central or western side of the state, although John McCain is represented in some of the southeastern and northeastern counties with a light red color.

Figure 5.3: Republican Presidential Primaries for the state of Utah
5.2 Census Results

This section displays interesting results obtained from applying this technique to the statistical dataset. Figure 5.4 shows result of visualizing employment, affluence, number of people with a Bachelor’s degree and median household income for the State of New Mexico. The statistical attributes of each county are compared to the state averages. If the attribute value is higher than the state average, then it is colored red. If the attribute value is lower than the state average, then it is colored green. The saturation of the color represents how far it falls from the state average. Most of the counties seem to agree with the theory that affluence and employment are high in a county with a high number of Bachelor’s degrees and median household income. This can be verified by observing counties that are colored entirely with shades of red or green. There are counties whose texels contain both colors. Here the anomalous attribute is employment. Counties in the eastern and south-western parts of the state are colored in shades of green except for the hexagons representing employment. Despite the low education level, low median household income and low affluence rate, employment is unexpectedly high. Finally Santa Fe and its neighboring counties exhibit a progressive quality with most or all values above the state mean.

Figure 5.5 shows the results for the state of New Hampshire. Four statistical attributes as shown in the legend are represented: employment, affluence, number of Bachelor’s degree and median household income. Many counties in the state exhibit results close to the state averages, shown as hexagons colored with shades of white. The northernmost county exhibits statistically negative results. The south-eastern county exhibits average results except for the dark red hexagons that reveal a much higher level of affluence. Figure 5.6 shows the result for the same state, but compared to the national averages. This map makes it evident that most counties are above national averages, except for the northernmost county which still exhibits negative results. This suggests that New Hampshire is better off than the average US state.

Figure 5.7 shows statistical results for the state of Oregon. Oregon is a large state with 36 counties. Most of the counties are colored with shades of green exhibiting statis-
Figure 5.4: Statistical results for the state of New Mexico
Figure 5.5: Statistical results for the state of New Hampshire averaged by the state
Figure 5.6: Statistical results for the state of New Hampshire averaged by the nation
tically negative results. Only a few counties in the north-western parts of the state and one county in the center show positive results. This can be explained by the population distribution of the state. If the population is concentrated in one part of the state, then the colors present on the map can be heavily affected by this. The state statistical averages are calculated by averaging the statistical values over the population of the state and not over the number of counties. If a densely populated area exhibits a positive statistical trend then we expect to see a lot of green across the rest of the state to compensate. Figure 5.8 shows the variations in the average values of two county’s A and B based on their population distribution. The green line represents equal population in both the counties. It shows an equal decrease in the average value of county B as the average value of county A increases. On the other hand, the blue line represents a population distribution of 80% in county A and 20% in county B. The blue line shows a rapid decrease in the average value of county B for even small increases in the county A’s average value. This has a direct impact on the colors chosen to represent the counties. Figure 5.8 demonstrates how a heavily populated region can impact the statistical values at sparsely populated areas. More often than not, a densely populated county contains a large city. Cities often exhibit a positive statistical trend due to the numerous job opportunities and access to good education. This can be seen in figure 5.7. Counties in and around the cities of Portland and Salem are colored with shades of red indicating a positive statistical result.

After analyzing figure 5.7 we might expect that counties shaded with red are densely populated and all the other counties are sparsely populated. To check this hypothesis, figure 5.9 was created. Figure 5.9 shows the same results but with an additional attribute, population. Shades of red represent population higher than the average county population and shades of green represent population lower than the average county population. Figure 5.9 shows that the hypothesis is almost but not entirely correct. As expected counties that are predominantly red have a population higher than the average, while counties with shades of green have a population below the average. The one exception is the county in the south-western part of the state. This county has its values in shades of green but the population is red, indicating a population much higher than the average county population.
Figure 5.7: Statistical Results for the state of Oregon
Figure 5.10 shows the visualization results for the country of India. The statistical attributes employment rate, level of affluence, per capita income and literacy rate are represented as shown in the legend. The visualization reveals a clear separation between the states exhibiting progressive results and the states exhibiting statistically negative results. The states in the southern part of the country reveal very positive results with per capita income and literacy rate equal to or higher than the national average and the employment rate and affluence higher than the national average. The states in the central part of the country exhibit statistically negative results, identified by the predominant shades of green. The states in the northern and western parts of the country also reveal negative results, but the affluence rate in these parts is much higher than the national average even though employment, literacy rate and per capita income are lower than the national average. The states in the eastern parts of the country also exhibit statistically negative results. The most dominant attribute in these areas is employment shaded in dark green to show a value much lower than the national average.
Figure 5.9: Statistical results with population for the state of Oregon
Figure 5.10: Statistical results for the country of India
5.3 Meteorological Results

As a final example, our technique was applied to meteorological data for the United States and Australia. Figure 5.11 shows the results for the precipitation dataset: number of wet days per year, precipitation in millimeters, percentage of cloud cover, and vapor pressure. The data attributes are mapped to shades of white for values below the national average and shades of blue for values above the national average. A black spot indicates missing data. A white spot reveals the minimum value in the country and a completely blue spot shows the maximum value.

From figure 5.11, we see that all four attributes have shades of white in the south-western regions of the country, indicating low values. Cloud cover is higher than the national average in the eastern and western ends of the US and near average in the northwest and southeast regions. Vapor pressure is particularly low in the northeast and high only in southern Texas and Florida. It is interesting to note that precipitation is higher than the national average but number of wet days is lower than the national average in the southeast. This indicates a seasonal rainfall period with intense rainfall occurring over only a few days. On the other hand in the northwest regions, precipitation is below the national average but the number of wet days is near average. This indicates rainfall with low intensity spread out across many days. The four attributes have a similar range of values in the southwest regions and southeast regions, but differ significantly in the midwest. Figure 5.11 shows us that the four attributes do not correlate across the country but only in a few limited regions.
Figure 5.11: Visualization of the Precipitation dataset of the United States
Figure 5.12 shows the result of applying this technique to the temperature and radiation datasets. The four meteorological attributes in this dataset are maximum temperature, mean temperature, minimum temperature and solar radiation. The attributes are mapped to a color scale ranging from shades of white for values below the national averages to shades of red for values above the national averages. Missing values are drawn as black tiles. Temperature takes a white color only in the peaks of the Rockies in the state of Colorado, where solar radiation is average. There is a correlation between temperature and solar radiation in the southwestern regions of the country, particularly in the state of Arizona, in the midwestern regions and in parts of the great plains. The northeast and the northwest regions have solar radiation lower than the national average but temperature values at or above the national averages.
Figure 5.12: Visualization of the Temperature dataset of the United States
Figure 5.13 shows the result obtained from applying this technique to the temperature and radiation datasets for Australia. Australia has a relatively flat terrain so we expect to see more correlation between the temperature readings and the solar radiation levels. Figure 5.13 shows some occurrence of correlation between all four data attributes. There is high correlation between the four attributes in the northern parts of the country. The solar radiation is high in the central parts of Australia because of their distance from the sea, but the temperatures are much lower. The island of Tasmania exhibits the lowest values in the country, identified by the dominant shades of white.
Figure 5.13: Visualization of the Temperature dataset of Australia
Chapter 6

Conclusion and Future Work

We see the world changing in many different ways. The changes themselves, multifarious in nature, are the ramifications of previous changes, and the initiators of new changes. It is common to expect to see correlations and associations between various attributes. If history is anything to go by, the hidden relationships and correspondences in the world help us understand it better. There are currently a limited number of ways to explore these attributes and the relationships that exist between them. This visualization technique was designed to represent attributes and also to highlight their relationships and correlations in ways that would make them easy to identify.

The technique uses color and texture to represent the aforementioned attributes. Color-mapped hexagons are symmetrically arranged in blocks called texels, and these texels are used to continuously tessellate a map. The color of the hexagons are mapped to an attribute’s magnitude. The position of the hexagon inside a texel reveals the identity of the attribute. Users can interpret the distribution of the individual attributes. Users can also compare the distribution of two or more of the attributes because the same color scale is used for all the attributes. The design uses hexagonal blocks because any number of hexagons can be symmetrically grouped inside a texel, allowing the design to accommodate any number of attributes.

The technique can be used to represent multivariate scalar attributes in which we
expect to see some correlation between the attributes. The attributes must be related, that is, the magnitude of an attribute at some spatial location must have certain impact on the magnitude of the other attributes at the same spatial location. This technique was used to represent Democratic and Republic presidential primary results. The visualizations show the winning contestants in each county and also the percentage of votes shared among the contestants. The technique was also used to represent United States and India’s statistics. Four statistical attributes were visualized to study the correlation between these attributes. Two meteorological datasets containing four attributes interpolated at \( \frac{1}{2} \)° resolution of the latitude/longitude grid were also visualized across the United States and Australia to analyze the relationships in meteorological data.

6.1 Future Work

This design is adequate to represent the data used in this thesis but the design needs certain extensions to support more types of data. This design can support only scalar attributes. A complex extension to the design would be required to represent vector attributes. Since only colors and textures are used in this technique, various directional mapping techniques could be added to the design to support the directionality of vector attributes. Visualization of multiple vector data could be very useful to study the relationships and correlations between multiple vector fields.

Color is an integral part of the design and it aids in interpreting the data attributes and representing an attribute’s numeric values. This emphasis on color warrants a detailed study of the color models, which in turn will help in conceptualizing sophisticated techniques to build the color scheme for the visualizations. Since the same color scheme is attached to all the attributes, there has to be a considerable amount of harmony between the colors in the color scheme and background color. Analyzing multiple correlations using different color schemes in the same map is another feasible and practical extension to investigate.

It would be a real challenge to extend this design to represent three dimensional scalar data attributes. Applying texels to a three dimensional surface is fairly straightfor-
ward and could be useful in studying scalar attributes over a geographic elevation map. But studying attributes that overlap at different height or depth values will be particularly difficult. It would be an interesting experiment to test the effectiveness of this design in representing non-geographic but physically spatial data.

The technical aspects of the thesis can be extended to execute more quickly and efficiently. The texture takes time to create, especially for grid datasets. Faster methods can be introduced but care should be taken to ensure that the new methods do not compromise on the clarity of the texture. The texture will appear jagged and uneven at extreme levels of zoom. This can be handled by creating multiple mipmap images but this in turn reduces the speed of the program. Newer and faster ways of effectively creating multiple textures of different sizes must be explored.

The idea of the technique is to analyze and study interesting relationships and correlations between attributes that cannot be easily compared using a table of numeric values. Any extension or future work should be conducted keeping this idea as the first and foremost objective and should not in any way compromise the strengths of the design.
Bibliography


[9] Brent M. Dennis and Christopher G. Healey, Assisted Navigation for Large Information Spaces Proceedings of the conference on Visualization '02

[10] Patricia M. Jones; Christopher D. Wickens; Stuart J. Deutsch, The Display of Multivariate Information: An Experimental Study of an Information Integration Task Online Publication Date: 01 March 1990P


[12] Christopher J. Morris, David S. Ebert, Penny Rheingans - An Experimental Analysis of the Effectiveness of Features in Chernoff Faces - 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making, Proceedings of SPIE


[26] A Typology for visualizing uncertainty, Judi Thomson; Elizabeth Hetzler; Alan MacEachren; Mark Gahegan; Misha Pavel, Proceedings Vol. 5669, Visualization and Data Analysis 2005


[48] Shortridge, Barbara G, Stimulus Processing Models from Psychology: Can We Use Them in Cartography, Cartography and Geographic Information Science , Volume 9, Number 2, October 1982 , pp. 155-167(13), Cartography and Geographic Information Society


[63] Cort J. Willmott, Kenji Matsuura and David R. Legates, (with support from NASA’s Seasonal to Interannual ESIP), Global Air Temperature and Precipitation: Regridded Monthly and Annual Climatologies, November 18, 1998.
