From Data to Decision:
Non-map-based Visualization of
Geospatial Data Sets

Borzu Talaie

A thesis presented to OCAD University
in partial fulfillment of the requirements for the degree of
Master of Design in
Digital Futures

Toronto, Ontario, Canada, April 2014

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I authorize OCAD University to lend this thesis to other institutions or individuals for the purpose of scholarly research. I understand that my thesis may be made electronically available to the public. I further authorize OCAD University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature _____

From Data to Decision: Non-map-based Visualization of Geospatial Data Sets

Master of Design
2014
Borzu Talaie
Digital Futures
Ontario College of Art and Design University

Abstract

The most common means of visualizing data sets with geospatial data is by employing map-based visualization models such as graduated symbol and choropleth. I argue that these models are in fact not the most efficient and effective visualization models for processing geospatial data, especially when the data set holds a notable quantity of location data. To support my argument I designed and developed two alternate models, one that did not use a map, and one that used an abstract version of one: a scatterplot model with geographic references, and a hexagon model. User tests were then performed to evaluate these models. This thesis describes the process, outcomes and future directions of my research. It also provides a literature review that addresses the definition and attributes of data visualization, a taxonomy of data visualization models, and a description of the mechanics of the visual cognition and colour theory employed.

For Mirkka

Table of Contents

List of Figures

All Figures are by the author unless indicated otherwise.

Chapter 1

Introduction

Chapter One – Introduction

According to IBM every single day we create 2.5 quintillion bytes of data. IBM argues that the exponential growth of data means that 90 percent of the data that exists in the world today has been created in the last two years. "This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, e-commerce transaction records, and cell phone GPS coordinates, to name a few" ("Big Data at the Speed of Business" par. 1).

To provide a sense of scale for some of the most common resources for the data being gathered, consider these facts. Every minute of every day we create (James):

- more than 204 million email messages.
- over 2 million Google search queries.
- 48 hours of new YouTube videos.
- 684,000 bits of content shared on Facebook.
- more than 100,000 Tweets.
- $272,000 spent on e-commerce.
- 3,600 new photos shared on Instagram.
- nearly 350 new WordPress blog posts.

Data that is within this scale, and which is available in data warehouses and databases, requires new approaches to data handling and new analysis for people to make sense of this data. As most persons are not interested in the data specifically, data itself needs to be presented in a form that conveys useful information (Yau Visualize xvi).

Experience and Motivation

Over the past twenty-one years, I have earned diverse theoretical and practical experience in visual communication design. Through print production, user experience design, motion graphics and physical computing, I have tried, because of the need to achieve effective results, to question the fundamental theories of function and aesthetics. For example, I questioned the purpose of a promotional newsprint when I was commissioned in January 2012 by an architectural design office, atelier rzlbd, to design a newsprint in support of an exhibition of the atelier at the Harbourfront Centre in Toronto.

The theme of the exhibition was, as it was called, "Big Enough?" This exhibition challenged the existing paradigms of urban development and housing in cities by asking this question: How much space do we really need?

To emphasize this question, I decided against a straightforward promotional brochure, as such portfolio pieces usually showcase projects by means of images and supporting copy. Instead, I provided readers with pertinent architectural data for each project. Through this means readers had the opportunity to examine and explore the raw data to derive their own respective interpretations and therefore arrive at their own personal conclusions.

In terms of questioning the relationship between desired and required living spaces in an urban setting, this project enjoyed success. This success was mainly due to the exploratory nature of the visual materials presented, such as pie charts and bar graphs.

This project was a great motivator for me to decide upon data visualization as the field of research for my graduate studies.

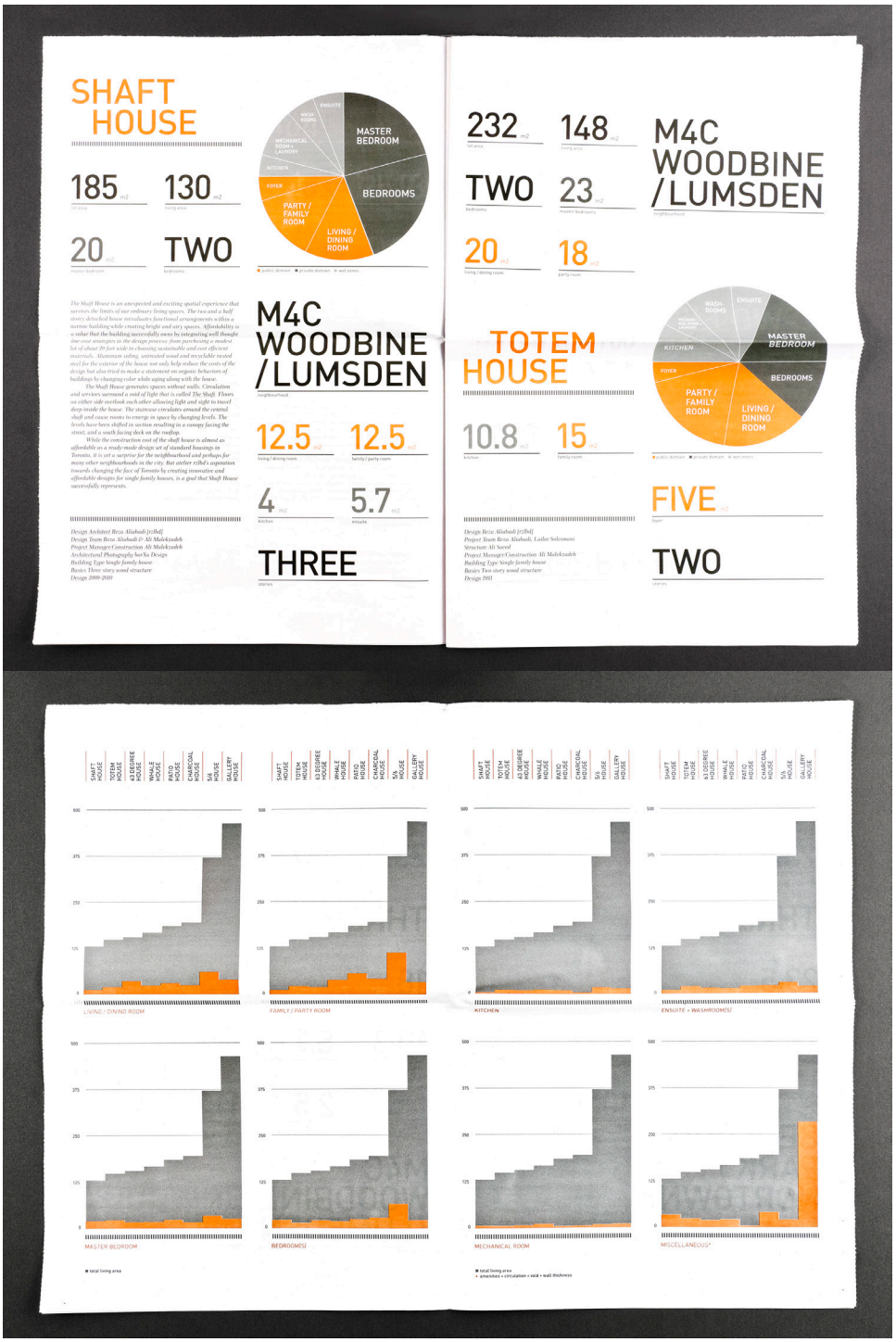

Fig. 1: Spreads from *rzlbd POST* 4.1 (Jan. 2012). Architectural data in tabular format was visualized by means of pie charts and bar graphs. This visualization was performed to facilitate readers in understanding the relationship between lot size and different living areas in projects with varying scales.

Scope of the Study

The study that follows examined a series of two-dimensional data visualization models through only basic interactivity—specifically, use of mouse hover to display information. The focus of this study was on the inherent qualities of each model and not the possible outcome of introducing interactive features. Such features might include zoom or filtering, both of which are designed to enhance the usability of interactive models—as Ben Shneiderman described this approach, "[o]verview first, zoom and filter, then details-on-demand" (337).

I evaluated at an overview level how people used these models to process the data: what type of data was available and where that data was located. The introduction of a feature such as data filtering would yield various biases, biases which might have enhanced our understanding of the data, but which were irrelevant to the concerns of this study.

None of the models examined here can provide the user with any ultimate solutions or final answers. Rather they function as aids in deriving information from large data sets.

In the following chapter I will talk briefly about the definition of data visualization and also how it is different from information visualization. Then I will visit the current classification of data visualization methods, placing an emphasis on mapping frameworks for data sets with geospatial data.

Chapter Three describes the research methods that I employed in this study. In Chapter Four I discuss my prototyping and user testing process in detail. Chapter Five illustrates the results of my user tests and analysis, followed in Chapter Six by conclusions and future directions for research.

Chapter 2

Literature Review

Chapter Two – Literature Review

Definition

The definition of data visualization (or data graphics) is best explained by Tufte (9): "Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and colour."

Data Visualization: How It Differs from Information Visualization?

Data visualization communicates a message by visualizing quantifiable data objectively, whereas information visualization is used to communicate any information—regardless of whether it is quantifiable or not—with a specific goal or message in mind.

It has been suggested that the term "information visualization" is useful for referring to any visual representation of data that is (Iliinsky and Steele 5-7):

- manually drawn (and therefore a custom treatment of the information);
- specific to the data at hand (and therefore nontrivial to recreate with different data);
- aesthetically rich (as its strong visual content draws the eye and holds interest); and
- relatively data-poor (because each piece of information must be manually encoded).

Because of their manual origins, information visualizations also tend to be limited in the amount data they can represent—the practical limitations of manipulating many data points create the problem. Similarly, it is difficult to change or update the data as any change must be implemented manually.

By contrast "data visualization" is useful for referring to any visual representation of data that is:

- algorithmically drawn (it may have custom touches but is largely rendered with the help of computerized methods);
- easy to regenerate with different data (the same form may be repurposed to represent different datasets with similar dimensions or characteristics);
- often aesthetically barren (the data is not decorated); and
- relatively data-rich (large volumes of data is welcome and viable, in contrast to information visualization).

Data visualizations also hold the advantage of being easily updated or regenerated with more or new data.

Data Visualization: A Storytelling Tool or a Discovery Tool?

A compelling observation on the value of data visualization is expressed by Tukey (vi): "The greatest value of a picture is when it forces us to notice what we never expected to see."

It is only through the visualization of data that we can easily recognize patterns and exceptions, and depict stories in data sets. Most data visualization models fit into one of two categories: *explanatory* or *exploratory*—in rare cases they can belong to both categories. Each category requires its own respective approach and respective process. It is therefore important to understand the difference between the two categories (Iliinsky and Steele 7).

Explanatory Data Visualization

Similar to information visualization, explanatory data visualization is used to tell a story that is already known to us, but not yet to the person to whom we are telling it. In this case certain editorial decisions are required. These decisions include the determination of which information should stay, and which information should be removed as it is irrelevant to the explanation and distracting to the viewer. This selection process brings focus to the data that support the story.

This type of data visualization is usually used as a presentation tool, and is commonly supported through at least some narrative that provides further explanation.

Exploratory Data Visualization

This model is used to identify attributes of a data set and what it might hold. Translating data into visuals that are easy to understand in turn enables us to identify quickly such features as patterns, trends and anomalous outliers.

This type of data visualization requires, in most cases, much cleaning and filtering, and is usually used in the analysis phase at a high level of granularity.

The statistician Francis Anscombe in the 1970s developed a demonstration that supports this idea. He conducted an experiment involving four sets of data, each exhibiting almost identical statistical properties, including mean, variance, and correlation. These data sets were known as Anscombe's Quartet (Kirk 10).

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Fig. 2: Sample data sets recreated from Francis J. Anscombe, "Graphs in Statistical Analysis," *American Statistician* 27.1 (Feb. 1973), 17-21.

Nothing much of interest is evident if we present these data sets in the format that is shown above—no indication of any pattern or trend is clear except perhaps the sequence of eights in the fourth data set.

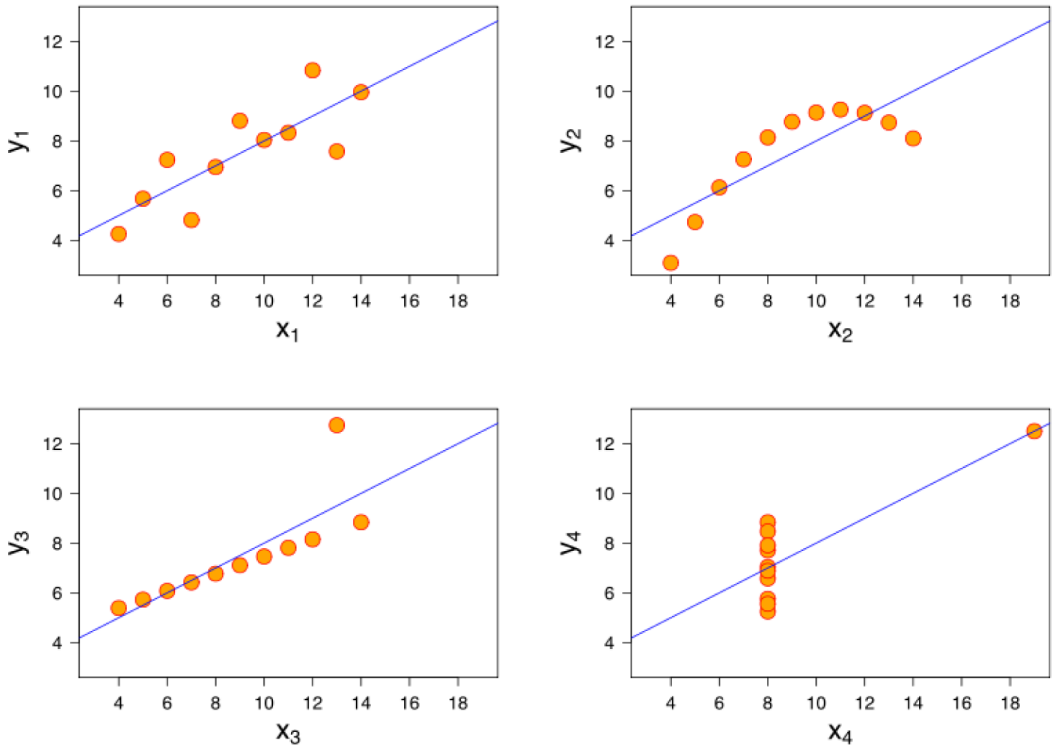

Fig. 3: Anscombe's Quartet plotted graphs. Image published under the terms of "Creative Commons Attribution-ShareAlike." Web. 22 Apr. 2014. Source: http://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3.svg

But visualizing these data sets graphically will make it much easier than would otherwise be the case to discover and confirm the presence or absence of patterns, possible relationships and anomalous outliers:

- likelihood of a trend line in X1, Y1.
- rise and fall pattern in X2, Y2.
- strong linear pattern with one distinct outlier in X3, Y3.
- strong similarity pattern with one distinct outlier in X4, Y4.

Taxonomy of Data Visualization Methods

One of the most critical parts of data visualization design is choosing the appropriate visualization method. In most cases the key factors in choosing a specific method are the physicality of the data, their organization, the relationships among the variables, and the intent of the visualization.

Kirk's classification proposal for certain data visualization methods, based on their primary communication purpose, has been illustrated in the next page (120). The reader should keep in mind that a creative field such as data visualization does not lend itself to the finite classifications presented here. For example, the most common way of visualizing data sets with spatial data is by positioning their values on a map using a geographic coordinate system. But, Yau argues, "[a] map isn't always the most informative way to visualize spatial data. Often, you can treat regions as categories, and a bar graph might be more useful than seeing a location" (Data 165). The scatterplot model that is explained in detail in Chapter Four tries to validate this argument.

Method classification	**Communication purpose**
Comparing categories	To facilitate comparisons between the relative and absolute sizes of categorical values. The classic example would be the bar chart.
Assessing hierarchies and part-to-whole relationships	To provide a breakdown of categorical values in their relationship to a population of values or as constituent elements of hierarchical structures. The example here would be the pie chart.
Showing changes over time	To exploit temporal data and show the changing trends and patterns of values over a continuous timeframe. A typical example is the line chart.
Plotting connections and relationships	To assess the associations, distributions, and patterns that exists between multivariate datasets. This collection of solutions reflects some of the most complex visual solutions and usually focuses on facilitating exploratory analysis. A common example would be the scatter plot.
Mapping geo-spatial data	To plot and present datasets with geo-spatial properties via the many different mapping frameworks. A popular approach would be the choropleth map.

Fig. 4: Kirk's classification proposal for some of the data visualization methods based on their respective primary communication purposes (120).

Many advantages may be found in the use of existing methods—such as the familiarity on the part of many viewers to read these types of visualization, and the fact that these methods have been proven to work effectively for specific type of data sets. But many data sets have unique characteristics which in turn require new means of communication.

Mapping Geospatial Data

As mentioned in the description of Kirk's classification proposal, many different mapping frameworks are available for geospatial data sets. Yau separates these frameworks into three major subcategories: *locations*, *regions* and *cartograms* (Data 166). Locations are a direct translation of latitude and longitude to two-dimensional space. They are straightforward and intuitive, but can become a challenge when there are many locations, as such a situation could result in overlapping points. In such cases displaying the density of points across regions might be a more informative approach. A good example of this approach in the region subcategory is choropleth maps, in which defined physical regions are coloured based on a scale. In contrast to choropleth maps, cartograms display entire regions using proportionally-sized symbols such as circles, instead of using the actual physical regions.

For this study, I have examined the two most common methods in the locations and regions subcategories: the graduated symbol maps and the choropleth maps.

Graduated Symbol Maps

Graduated symbol maps, also known as location maps, use the visual variable of size to represent data. The size of a symbol is proportional to the range of values in the data set. As for the position of the symbols, they are usually placed at the centre of the area they represent. According to Meirelles:

> There are two main variables to consider when designing a graduated symbol map: the shape of the symbol and the scaling. The shape of the marks can vary, and the most common shape is the circle, although we see rectangular bars as well as triangles being used. There have been attempts at three-dimensional symbols, where the scaling is done to the cube rather than to the square root. But, if area perception is already hard in two dimensions, and often underestimated, then it gets more problematic judging relative sizes of quantities provided by volumes (138).

Meirelles continues:

> Selecting the scaling method is perhaps the biggest challenge in proportional symbol maps, as well as in choropleth maps. There are two ways to scale the size of symbols: *classed*, when size is range graduated, and *unclassed*, when sizes follow a proportional system. In unclassed systems, the number of categories is equal to the number of data values (138).

Fig. 5: A graduated symbol map of the Greater Toronto Area.

Choropleth Maps

Choropleth maps are based on statistical data aggregated over previously defined regions. Visual variables such as colour value, colour saturation, and texture (or all three, or a combination of two[1]), are used to encode regions in proportion to the measurement of the statistical data. Choropleth maps provide an easy way to visualize how measurement varies across a geographic area. They can also display the level of variability within a region ("Choropleth Map" pars. 1-2).

A challenge with choropleth maps is that larger regions receive more visual attention regardless of the data. They occupy more space in the physical world and on the computer screen (Yau Data 175). Meirelles also argues that:

> Because visual encoding is uniformly distributed within the regions of choropleth maps, the impression is that the phenomena represented are also uniformly distributed, which most often is not the case. The overall impression of the phenomena will be more meaningful if the statistical areas are of similar shape and small in size (142).

1 "By compositing color with texture, we can increase the number of co-located variables that can be effectively visualized, beyond what would be possible using either just texture or just color compositing alone" (Shenas and Interrante 446).

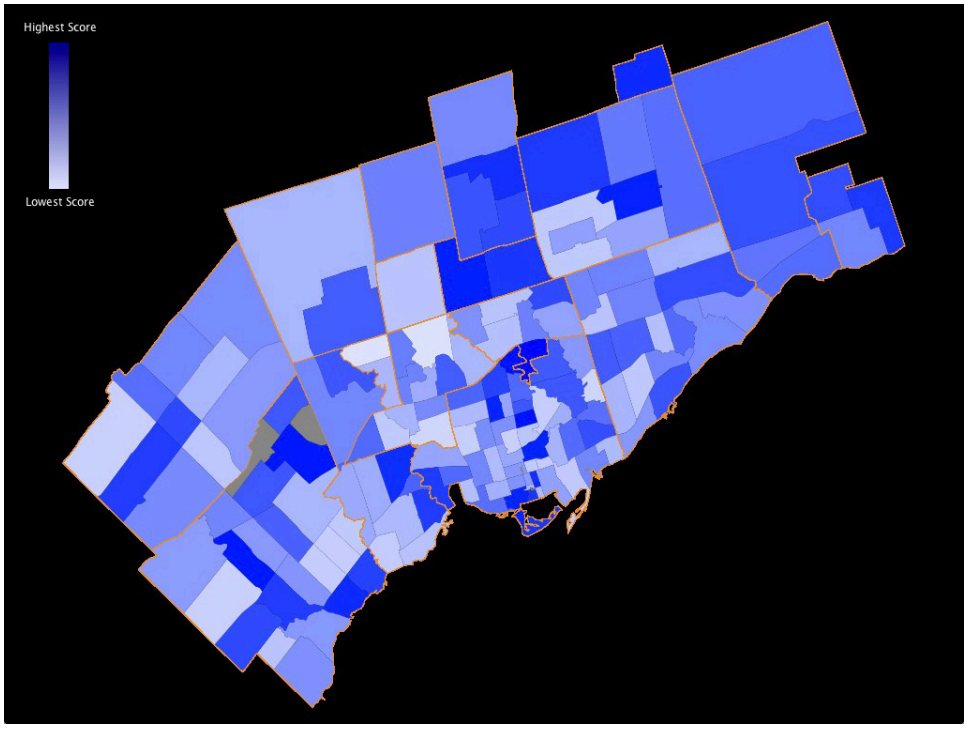

Fig. 6: A choropleth map of the Greater Toronto Area.

Distorting Maps

A classic example of distorting the geographic location of data points in maps is Harry Beck's re-design of London Underground Map in 1933.

> Beck displayed the system on an octagonal grid so that lines met at right angles or at 45 degrees, with the stations placed to show their relationships within the system rather than their actual distance from each other (Hollis 94).

Abstracted from the relative physical layout of the city, Beck's map showed only the connection between stations. He used this distortion to focus on the rider's goal: to go from one station to the other.

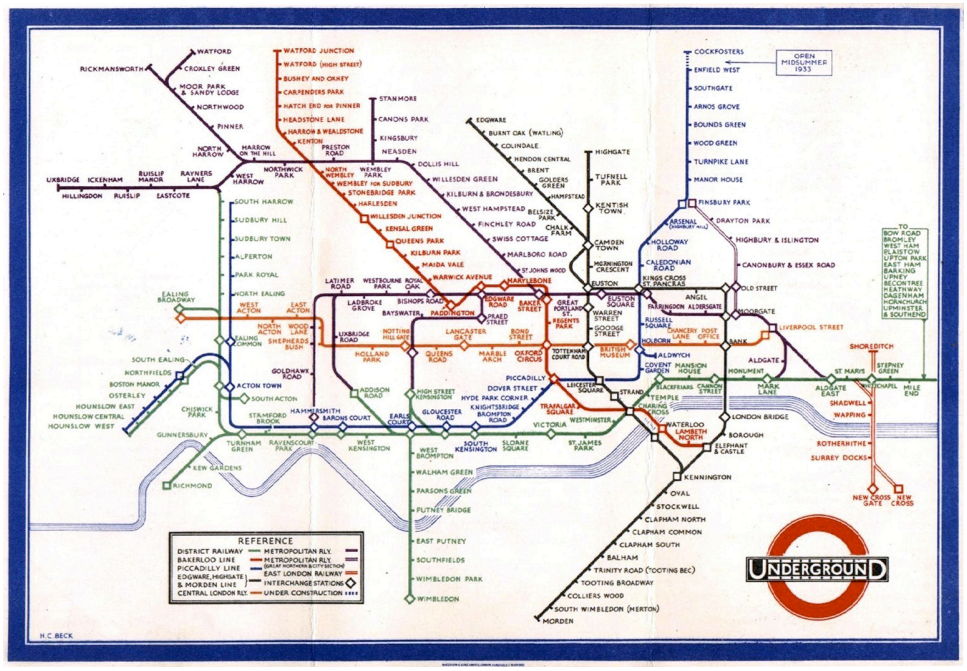

Fig. 7: London Underground Map designed by Harry Beck in 1933. Web. 24 Apr. 2014. Source: http://www.experiencecard.co.uk/blog/wp-content/uploads/2012/10/Experience-Card-Map.jpg

Visual Perception and Cognition

Although visual information is available to us all the time, our visual system extracts features separately and over different stages. Ware proposes a three-stage model for perception. In Stage One, billions of neurons work in parallel to extract millions of basic features—such as colour, texture and orientation—that are processed rapidly and simultaneously. In Stage Two, patterns and structures—such as regions of the same colour or texture—are processed and extracted serially and slowly. In Stage Three, a sequential goal-oriented process reduces the information to a few objects, then holds them in the working visual memory to form the basis for visual thinking (Ware Information 20-22).

> Working memory is where we wrestle with information to understand and process it so that it can eventually be assimilated into long-term memory; but the problem is that working memory is like a sieve. It is weak, can't wrestle for long, and can't wrestle with much (Evergreen 11-12).

Research shows that we can only hold between 3 to 5 chunks of information in working memory at any one time, and even that number varies by the environmental context (Baddeley 281-288). "When a viewer's working memory is overloaded, it drops some chunks of information, and then misunderstanding or frustration results" (Woodman et al. 80-87).

Pre-attentive Processing

When something catches our eyes, it taps into our earliest stages of attention. This process happens very fast—usually in fewer than 10 milliseconds—and is so subtle that some researchers call it pre-attentive processing (Ware Visual 27; Callaghan 300).

Designers can make intentional use of pre-attentive variables to enhance detection and recognition of relevant visual elements.

> Pre-attentive variables can increase the performance of the following tasks: target detection, boundary detection, region tracking, and counting and estimation. However there are factors that might impair the detection of pre-attentive-designed elements, such as number and variety—the degree of differentiation—of distractors in the representation (Meirelles 22).

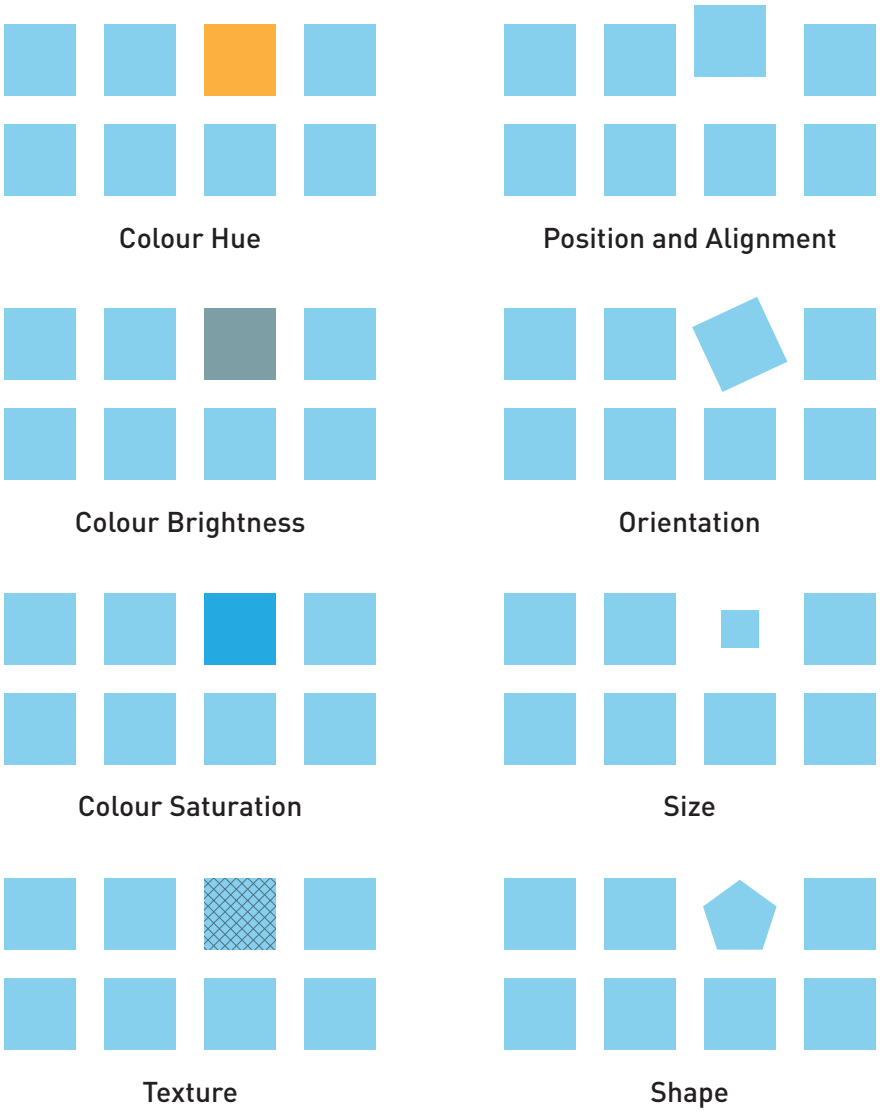

Fig. 8: Pre-attentive variables. Image redrawn by author. Inspired by Jenifer Tidwel's illustration of pre-attentive variables (286).

Retinal Legibility

Retinal legibility is considered one of the general *rules of legibility,* also known as the *rules of separation*. These rules ensure a needed separation between the visual variables and their respective steps. According to Bertin "[a] visualization must utilize the range of perceptible differentiation afforded by the visual variables, in such a way that the eye can separate the steps" (175).

To achieve maximum differentiation we must ensure that two basic features are in place. First, the size must be sufficient that the smallest signs are visible, that they stand out from the background, and that they overcome the visual noise. The size must also be limited, thereby ensuring that the largest signs do not overlap, and that they are each separate from the other. Second, we must also obtain the greatest amount of differentiation by utilizing the entire perceptible range of a given variable (Bertin 180).

The Titchener Effect

The Titchener effect is an optical illusion of relative size perception.

> The apparent size of a circle is dependent on its surroundings. A circle surrounded by a ring formed from other circles will appear smaller if the surrounding circles are enlarged. This effect is especially striking in cases with the inducing circles respectively smaller and larger than the referent circles (Surkys, Bertulis, and Bulatov 673).

Figure 9 demonstrates this illusion clearly. All the black circles are the same in size, but the one on the left, surrounded by a larger ring, appears to be smaller in size compared to the remaining two black circles. The same effect affects the black circle on the right. It is surrounded with a smaller ring, allowing it to appear bigger in size when compared to the remaining two black circles.

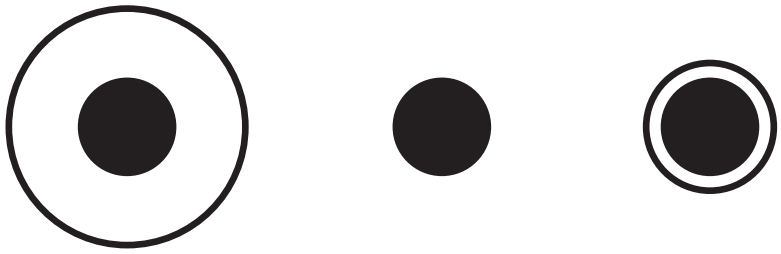

Fig. 9: Circles with the same size surrounded with different ring sizes.
The figure demonstrates the Titchener optical illusion effect.

Colour

Colour has three perceptual dimensions: hue, value and intensity.

> Hue is the colour itself, for example a red or a green. Value or lightness is the intrinsic lightness or darkness of a colour. The value scale ranges from black to white. Intensity or saturation is the term applied to the relative purity and amount of the colour in a given area. A full scale of intensities would range from hue on one end to a neutral grey on the other (Robinson 81-82).

The Human Eye

There are two kinds of receptors in our eyes: rods and cones. Rods are responsible for registering values. Cones perceive hues. There are three kinds of cones in the eye: those that respond primarily to red, those that respond primarily to green, and those that respond primarily to blue. Many more rods function in the eye than cones, and they respond slowly to changes in levels of light. The information gathered from these rods and cones is transferred to the optic nerve and sent to the brain, which is where we actually perceive colour (Howlett 124). Because of the fact that there are more rods than cones in the eye, value assumes greater importance in vision than do either hue or intensity.

> The sensitivity of the eye to value differences is great over a wide range, but the sensitivity decreases as both extremely high and extremely low values are approached. Of somewhat more significance in application, however, is the fact that with decreasing value the sensitivity diminishes more rapidly for colours of longer than of shorter wave length (Robinson 90).

Simultaneous Contrast

Simultaneous contrast is an illusion that occurs when different hues, values or intensities lie beside one another. Therefore colours that are not particularly strong when viewed by themselves might look brilliant when placed beside other stronger colours (Ware Information 75). For example, in figure 10 all the squares—placed on top of the rectangle with a blue to white gradient—have an intensity of 70%, but the square on farthest left looks brighter than the square on farthest right.

Fig. 10: A colour composition demonstrating the simultaneous contrast effect.

Also according to Howlett:

> Some colours placed next to each other in certain configurations can't be focused on at the same time (such as red and green) and create a vibrating effect—the colours seems to move, which can be very disturbing. Colours that are very similar in both hue and value will also cause visual discomfort when seen next to each other, because the eye has to strain to distinguish the slight differences (133).

Colour Blindness

"About 10% of the male population and about 1% of the female population have some form of colour vision deficiency. The most common deficiencies are explained by lack of either the long-wave-sensitive cones or the medium-wave-sensitive cones resulting in missing the red-green channel" (Ware Perception 98). Therefore to ensure that colours can be distinguished by most individuals, selecting colours in the red-green channel should be avoided.

Research Question

The two common mapping strategies—that of the graduated symbol map and of the choropleth map—have revealed shortcomings that are linked to the realities of geospatial representation: scaling methods and distribution. Both of these shortcomings demonstrate the limitations of these map-based representations.

These observations led me to consider: Does the absence of a map in visual representation of data sets—that is, one with geospatial data—affect negatively users' ability to *process* the data?

I used the word *process* (e.g., the recognition of patterns or of outliers) rather than understand (e.g. why patterns emerge or why there are outliers) because understanding data depends on so many factors other than the visual representation of the data. These are factors that are beyond this study.

Chapter 3

Methodology

Chapter Three – Methodology

Given the complexity of data, using data to provide a meaningful solution to a visualization problem requires insights from three diverse fields: statistics, data mining and graphic design. However, each field has evolved in isolation from the others. Fry argues that we should reconcile each of these fields as part of a single process. Graphic designers can learn the computer science necessary for visualization, and statisticians can communicate their data more effectively by understanding the visual design principles behind data representation (5).

To gain knowledge about the processes involved in data visualization design, especially the required computer skills, I chose practice-based research as my initial research method.

> Practice-based research is an original investigation undertaken in order to gain new knowledge partly by means of practice and the outcomes of that practice. Claims of originality and contribution to knowledge may be demonstrated through creative outcomes which may include artefacts such as images, music, designs, models, digital media or other outcomes such as performances and exhibitions. Whilst the significance and context of the claims are described in words, a full understanding can only be obtained with direct reference to those outcomes (Candy 3).

The creative outcome of this study was a series of standalone data visualization prototypes—explained in detail in Chapter Four. In later stages, to document and demonstrate my progress I graduated to reflective practice as a research method. Finally, I employed user testing to evaluate the prototypes and validate the outcomes of my study. A group of twelve persons from the Digital Futures graduate program and the Visual Analytics Lab (both units are at OCAD University) were recruited to evaluate the prototypes. The recruitment was approved (approval number: 2013-36) and overseen by the university's Research Ethics Board. The committee determined that the recruitment met all the requirements for ethical treatments of research participants. Approval documents are included in Appendix A.

The Process

Fry suggests a seven-step framework for visualizing data sets (5). I followed these steps in the order presented here except for the last two steps—basic interactivity was introduced in my prototypes before I refined them.

Acquire
Obtain the data, whether from a file on a disk or a source over a network.

Parse
Provide some structure for the data's meaning, and order it into categories.

Filter
Remove all but the data of interest.

Mine
Apply methods from statistics or data mining as a way to discern patterns or place the data in mathematical context.

Represent
Choose a basic visual model, such as a bar graph, list, or tree.

Refine
Improve the basic representation to make it clearer and more visually engaging.

Interact
Add methods for manipulating the data or controlling what features are visible.

Fig. 11: Ben Fry's seven-step data visualization framework (5).

Acquire

In the summer of 2013, I had the opportunity of holding an internship position at *The Globe and Mail*. In two months at the newspaper, I worked with the Data Analytics Team. This team provided me with two different data sets: a Demographics data set and a Business data set.

The data the newspaper provided me assisted in initiating this study. No parts of this data have been documented or otherwise identified as part of this study in any form—visual, verbal or tabular.

This aspect of my study is in compliance with the terms of my signed non-disclosure agreement with *The Globe and Mail*. The terms and conditions of my internship at *The Globe and Mail* can be reviewed in Appendix B.

Parse

To understand the data, I carefully examined the relationships between different values then organized them into categories. Most of the data was categorized in individual rows based on Forward Sortation Areas (FSAs) in Canada. An FSA is a geographic area in which all postal codes start with the same first three characters.

Filter

Both data sets included several columns that were used to calculate different values irrelevant to this study. I removed all the columns except the column that showed the final desirability score for FSAs in the Demographics data set and the one showing profit and loss in the Business data set. I have also filtered down the number of rows in either data set from 1023 to 152 (total number of rows with data for FSAs in the Greater Toronto Area).

Mine

The values for the Demographics data set had been normalized—transformation of the variable to a new range, in this case from 0.0 to 1.0—to treat all variables equally. Through this process one column did not yield more influence over another when the final score was calculated. As for the Business data set, variables were already in Continuous format. "A continuous variable is a variable where an infinite number of numeric values are possible within a specific range. An example of a continuous value is temperature where between minimum and maximum temperature, the variable could take any value" (Myatt and Johnson 21).

Represent

To represent these data sets I employed two common models from the geospatial mapping category in the taxonomy of data visualization methods: graduated symbol model and choropleth model. I also devised two custom models that are not in the taxonomy of data visualization methods: hexagon model and a modified scatterplot model with geographic references on the x-axis. These models are described in detail in Chapter Four.

Interact

As mentioned in the Introduction, this study does not concern itself with the possibility or the outcomes of the introduction of advanced interactive features—such features might include zoom or filtering. The focus was instead on the inherent qualities of each model at an overview level. The only interactive feature used in these models was a simple mouse hover, one that revealed more information about individual data points.

Refine

A series of prototypes were initially developed to visualize the data sets. Upon further consultation with my advisors, I narrowed down these prototypes to four: two models based on commonly used data visualization models, and two custom models. All four of these models were refined (1) to enhance usability, and (2) to reduce cognitive load based on my findings in the literature review and also based on the feedback that I received from my advisors—e.g., integration of legends or use of single visual variables to display values for data points.

Chapter 4

Prototyping

Chapter Four – Prototyping

For the prototyping stage I was required either to use one of the available software packages commonly employed in data visualization, or learn a programming language through which I would develop working models for evaluation. I decided on learning a programming language, using the Processing environment. Despite its steep learning curve, Processing proved to be a flexible tool with none of the limitations of the ready-made applications. I provide in Appendix C a list of solutions, with their respective strengths and limitations enumerated for both categories.

In this chapter I will explain in detail the evolutionary process of designing and developing a series of prototypes that I used in this study. As mentioned in the previous chapter, four of these models were selected and refined to enhance usability and clarity before they were evaluated in the user testing stage. The initial models with no refinement are referred to as version one (V 1.0) and refined models are referred to as version two (V 2.0).

The first two models introduced here—the graduated symbol and the choropleth—use actual geographic maps for visualizing data sets. The other two models—the hexagon and the scatterplot—use either a distorted and abstract version of the actual geographic map or only geographic references for representing the location of data points.

Graduated Symbol Model

This model uses the centre point of each FSA and draws a coloured circle to visualize the desirability score values extracted from the Demographics data set.

In version one, two different visual variables—colour and size—were used to show different desirability score values. For example, a big dark blue circle indicated an FSA with a high desirability score, and a small red circle indicated an FSA with a low desirability score.

To reduce cognitive load in version two, I used only size as the differentiating visual variable, with a semi-transparent blue colour. Also in this version the boundaries between different regions on the map were outlined using an orange line. Doing so made the boundaries more recognizable.

In both versions, users can obtain more information—such as the postal code, region, and desirability score in numeric format—by hovering the mouse over the centre of these circles.

The biggest challenge that I faced with this model was the fact that I was not able to obtain the geographic coordinates for the centre of FSAs in the Greater Toronto Area (GTA). To work with this model I had to manually calculate the position of each centre point on the map and register its x and y coordinates in a separate data set.

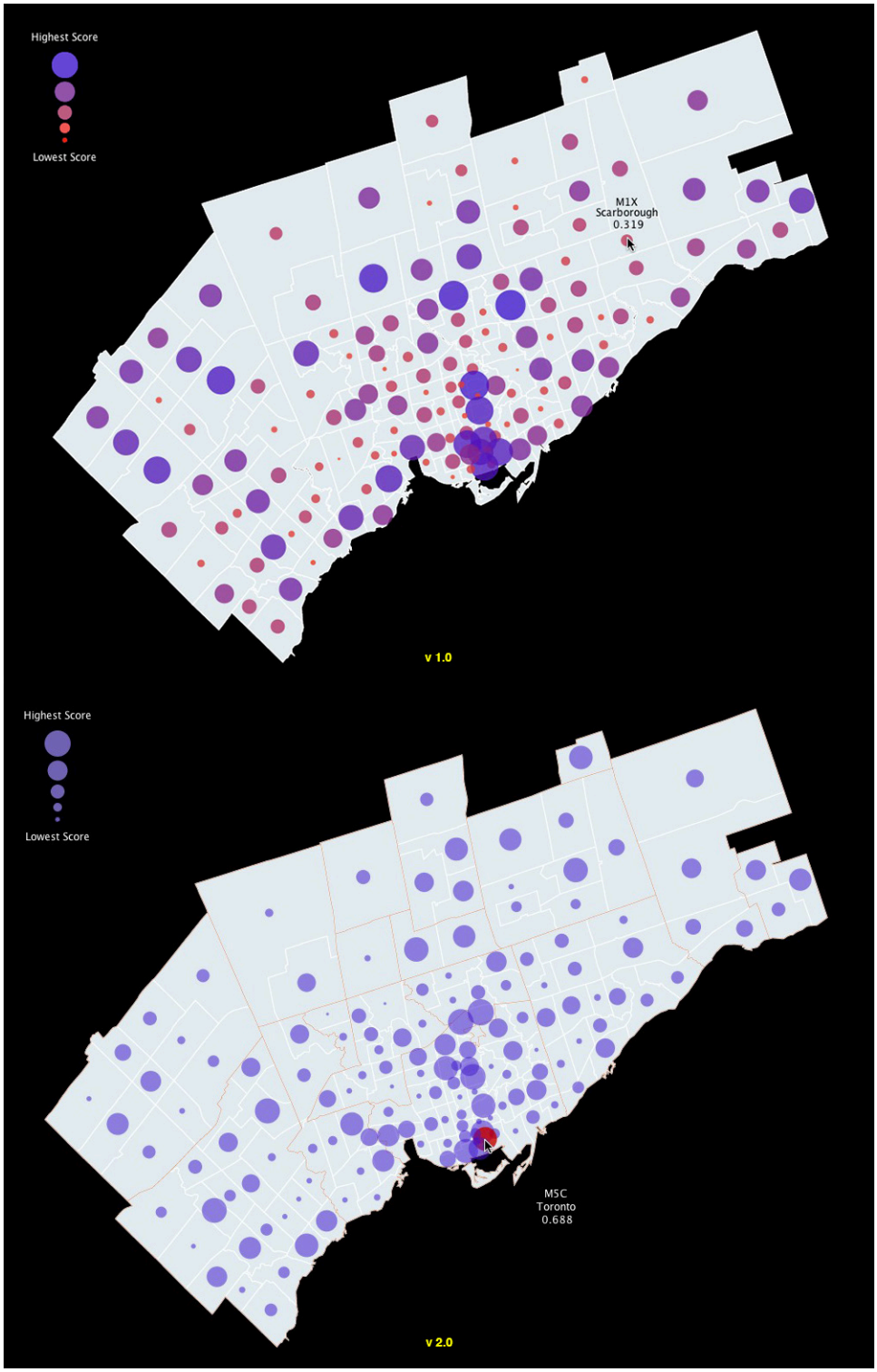

Fig. 12: Version one (top) and version two (bottom) of the graduated symbol model.

Choropleth Model

In this model each polygon represented an FSA in the GTA. In version one, each polygon was assigned a colour from the spectrum of available colours between blue and red. The higher the desirability score, the closer the colour would be to blue, and the lower the desirability score, the closer the colour would be to red.

In version two, different intensities of the colour blue were used to represent desirability scores. I employed this method to reduce cognitive load. The amount of these intensities were calculated by mapping the values derived from the data sets to a range between 0 to 255. The value 0 defined the colour as entirely transparent (minimum intensity) and the value 255 as entirely opaque (maximum intensity). The values between these extremes resulted in differing intensities respectively (Reas and Fry 29). For example, polygons with higher intensities—darker blue—represented FSAs with higher desirability scores, and polygons with lower intensities—lighter blue—represented FSAs with lower desirability scores. Not all of the FSAs in GTA had values in the data set. Those FSAs were shown in grey.

Users can obtain more information—such as the postal code, region, and the desirability score in numeric format—by hovering the mouse over each polygon. Also in version two the boundaries between different regions on the map were outlined using an orange line.

This model required a Scalable Vector Graphic[1] map of the GTA for visualizing FSAs and their values which I created manually. Each polygon was placed on an individual layer in the SVG file before they were ordered to match the order of FSA rows in the data sets.

1 Scalable Vector Graphic (SVG) is an XML-based vector image format for two-dimensional graphics that has support for interactivity and animation ("Scalable Vector Graphics" par. 1).

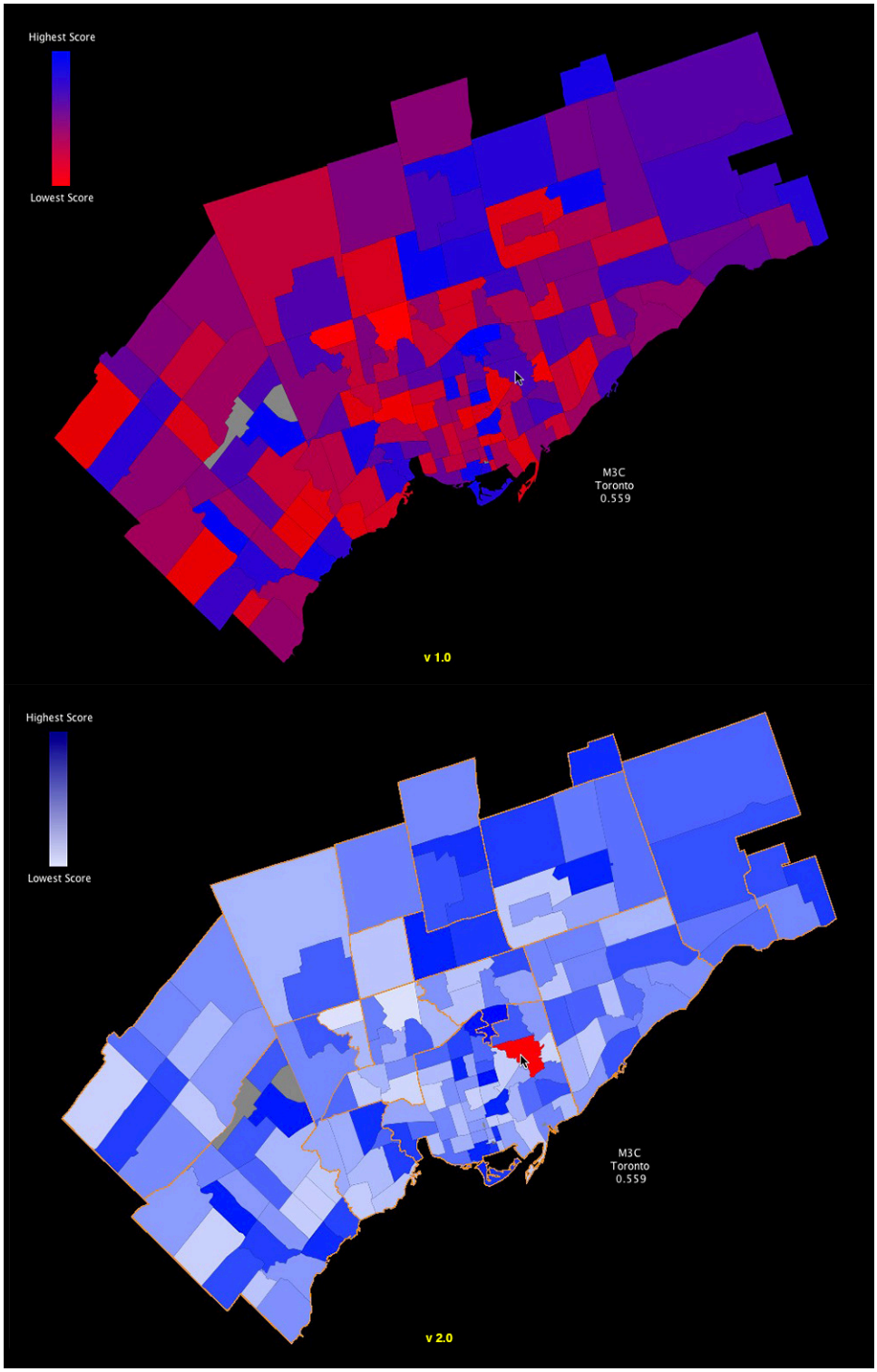

Fig. 13: Version one (top) and version two (bottom) of the choropleth model.

Hexagon Model

This model is similar to the choropleth model. The only difference is that a series of hexagons are used to replace the polygons in the choropleth model.

The advantage of this model over the choropleth model is that each FSA receives the same visual attention because the size for all the hexagons are same. For example, some of the FSAs located in downtown core that were barely visible in the choropleth model are clearly illustrated in this model.

The downside of this model is that the exact geographic location and the proximity between different FSAs are distorted because of the transformation from polygons with varying shapes and sizes to uniform hexagons with the same size.

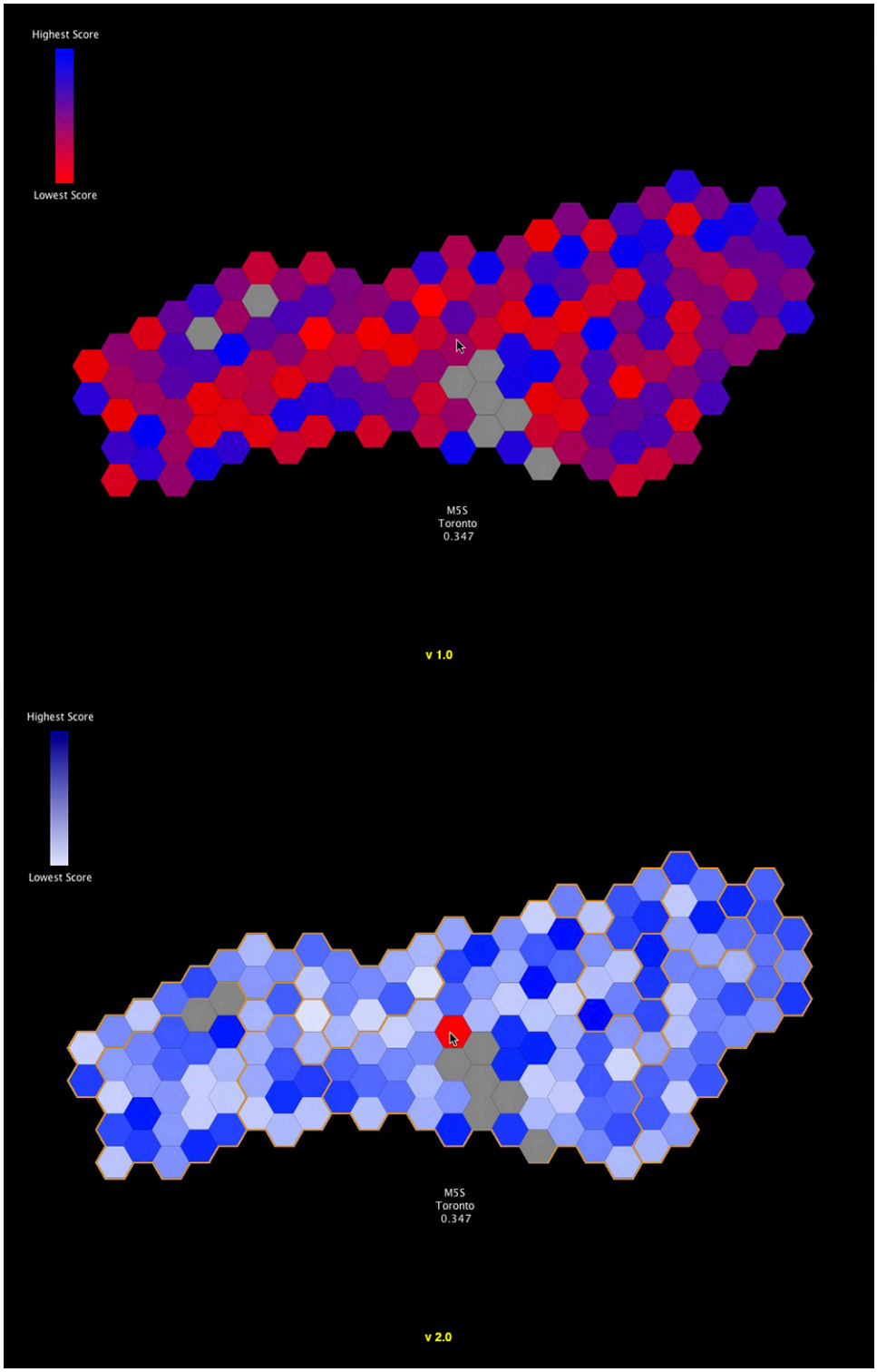

Fig. 14: Version one (top) and version two (bottom) of the hexagon model.

Scatterplot Model

This model is a visual representation of the desirability scores for all the FSAs in the GTA. Each data point in this model represented an FSA. The desirability score for individual FSAs were represented by their respective positions on the vertical axis. The higher the score, the closer the points were positioned to the top of the screen and the closer their colour was to blue. The lower the score, the lower the points were positioned to the bottom of the screen and the closer their colour was to red. To display these relationships clearly a scale was introduced in version two, and the changes in colour variable were removed from the model to reduce cognitive load. A horizontal rule guide—one that was activated by pressing and holding the left mouse button—helped the user to understand better the relative position of FSAs to each other.

In version one, data points were positioned from left to right on the horizontal axis in the same order as their row numbers in the data set. Since FSAs in the data sets were ordered in an alphabetical order, this visualization method created some confusion regarding the ordering and positioning of FSAs. To fix this issue, in version two, the approximate geographic location of FSAs and their regions were first mapped from left to right on the x-axis (an illustration of the method used is provided in Appendix D). I then reordered rows of data in the data sets so that the position of each FSA on the x-axis would match the approximate geographic locations calculated in the previous step.

Also in version two, depending on the position of the mouse, a region with all of its FSAs was highlighted. The name of each region was shown at the bottom of the screen. Users could obtain more information—such as the postal code, region and desirability score in numeric format—by hovering the mouse over each data point.

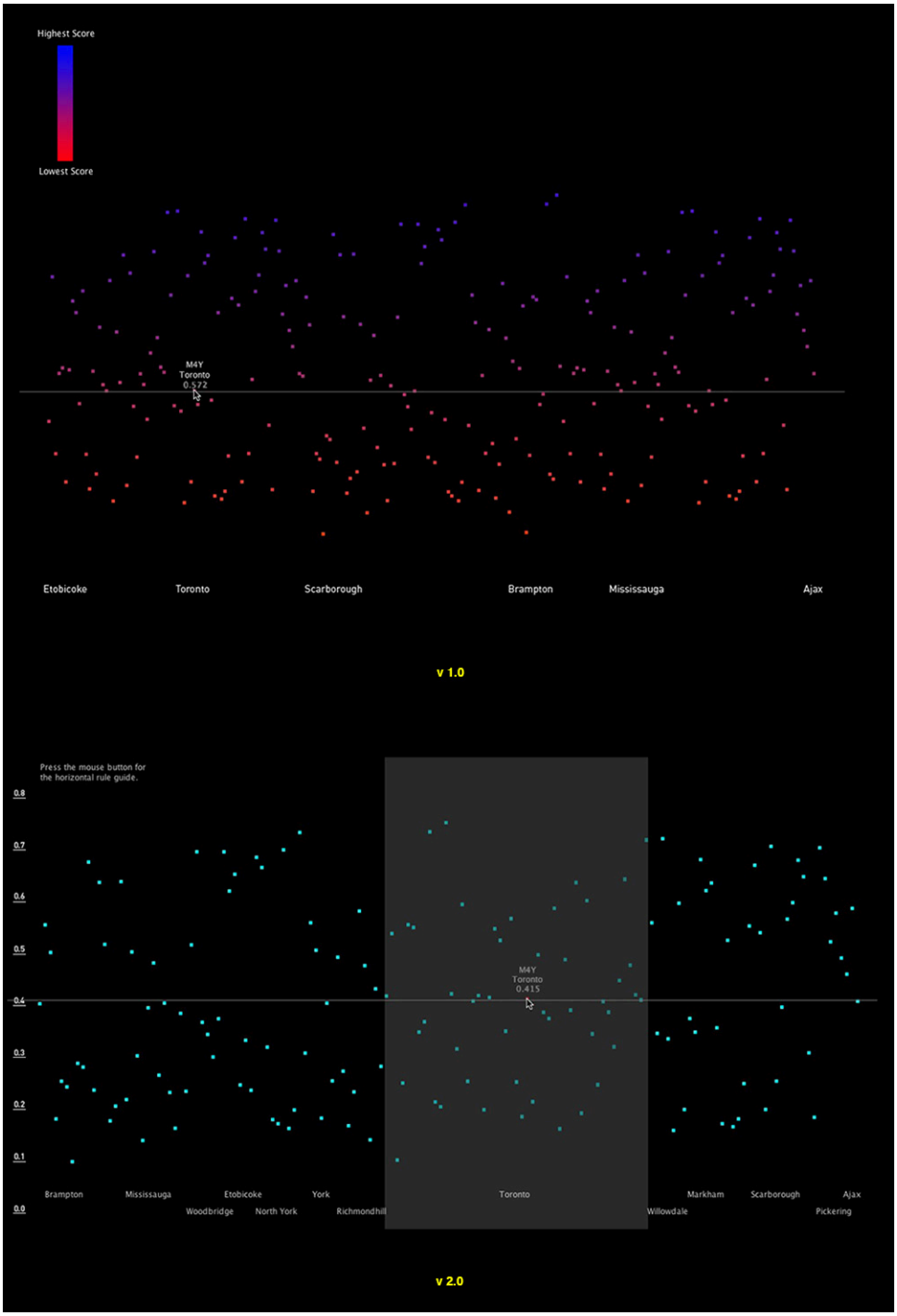

Fig. 15: Version one (top) and version two (bottom) of the scatterplot model.

User Testing

To evaluate the speed and accuracy of these prototypes, a series of script-based user tests were conducted with twelve participants. This sample size was determined using the following method ("Permutations" par. 3):

1. There are four ways to select the first prototype.
2. After we select the first prototype, there are three ways to select the second prototype.
3. After we select the second prototype, there are two ways to select the third prototype.
4. And finally there is only one prototype left to select.

Therefore the prototypes selected for this study could be permuted 4 x 3 x 2 x 1 ways, or 24 ways.[1] But since choropleth model and the hexagon model are not significantly different—except in the use of an actual map—I therefore divided the above result by the number of similar models: 24 ways divided by 2 equals 12 ways—one for each participant. These permutations were then further reordered to avoid order-effect bias.[2]

Eight of the participants were male and the remaining four female, with an age range between twenty and fifty-one years. Seven of the participants were living in the GTA and the remainder had each lived in Toronto for at least eighteenth months. Most of the participants were familiar with visual analytic tools, and all of them had, on a regular basis, been using some form of a screen-based map for navigation and finding specific locations in the city. None of the participants had been diagnosed with colour blindness.

1 The number of ways you can change the order of a set of things is called the number of permutations of that set of things ("Permutations." par. 1).

2 The relative position of an item in an inventory of questions or stimuli may uniquely influence the way in which a respondent reacts to the item. This phenomenon is referred to as order-effect bias (Perreault 544).

Two different data sets were used with these prototypes: a univariate set holding desirability score values, and a bivariate set holding net profit and net loss values. The only difference in the prototypes that used the bivariate data set was that two different colour hues were used to differentiate between the positive and negative values—blue for net profit and red for net loss.

For each prototype, participants were asked to perform a series of similar tasks, such as locating a data point with a specific attribute or recognizing a pattern in data points. User testing scripts and user tasks are documented in Appendix E.

Chapter 5

Analysis

Chapter Five – Analysis

I have conducted both quantitative and qualitative analyses of the user testing results. These results are presented in this chapter.

Quantitative Analysis

Figure 16 displays the user testing results for all the prototypes that used a univariate data set. The task was to find and register the top three data points—that is, those with high values—starting with the highest. The scatterplot model produced the lowest average in response time (12 seconds). The choropleth model produced the highest average in response time (1 minute and 2 seconds). The scatterplot model was the only model with no participant errors. The choropleth model produced the highest participant error rate (3/12) of all the models.

Figure 17 displays the user testing results for all prototypes that used a univariate data set. The task was to find and register the bottom three data points with low values, starting with the lowest. Scatterplot model produced the lowest average in response time (14 seconds). The response time in the other three models was approximately 3 times higher. Scatterplot and graduated symbol models produced no participant errors. Hexagon models provided the highest participant error rate of all the models (2/12).

Figure 18 displays the user testing results for all prototypes that used a univariate data set. The task was to recognize a region with clustering of data points with the highest values. The average response time was lowest for the graduated symbol and choropleth models (1.5 seconds) and highest for the hexagon model (6 seconds). In this task there were no participant errors for any of the models.

Order	**Graduate Symbol** (A)	**Choropleth** (B)	**Hexagon** (C)	**Scatterplot** (D)
CBAD	00:42.750	00:45.900	00:22.630	00:07.530
DCBA	01:31.190	01:24.650	00:25.530	00:15.380
BCDA	00:42.030	02:44.630	00:45.510	00:24.260
DACB	00:28.700	00:26.250	00:19.350	00:14.280
ACBD	01:45.310	01:38.730	00:33.050	00:11.360
CADB	01:03.350	00:50.520	00:49.900	00:11.110
CDAB	00:30.260	01:16.570	00:31.270	00:10.550
ADCB	00:36.130	00:37.170	00:13.380	00:12.650
BDAC	00:34.510	01:02.430	00:14.430	00:11.480
BADC	00:32.050	00:44.250	00:11.010	00:07.650
CABD	00:37.030	00:18.330	00:33.560	00:12.850
CDBA	00:24.800	00:33.780	00:20.480	00:12.910
Average	**00:47.343**	**01:01.934**	**00:26.675**	**00:12.668**

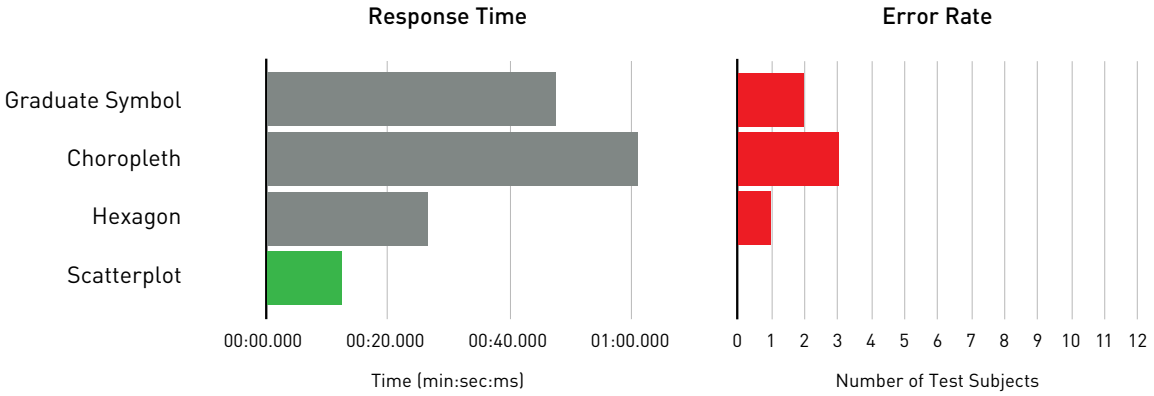

Fig. 16: User test results for finding and registering top three data points with highest values.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:16.330	00:42.870	01:08.600	00:09.140
DCBA	01:11.220	01:04.580	01:03.030	00:20.660
BCDA	01:16.130	01:35.720	00:45.380	00:21.560
DACB	00:37.350	00:20.450	00:27.080	00:13.360
ACBD	01:04.280	01:01.300	00:56.170	00:10.650
CADB	00:39.470	00:25.180	01:18.140	00:12.150
CDAB	00:17.780	00:36.880	00:45.350	00:13.760
ADCB	00:32.100	00:32.060	00:19.730	00:15.000
BDAC	00:17.610	00:53.350	00:16.150	00:13.200
BADC	00:17.030	00:51.020	00:13.530	00:11.000
CABD	00:25.130	00:10.300	00:48.420	00:12.200
CDBA	00:21.830	00:19.130	00:23.180	00:13.510
Average	**00:36.355**	**00:42.737**	**00:42.063**	**00:13.849**

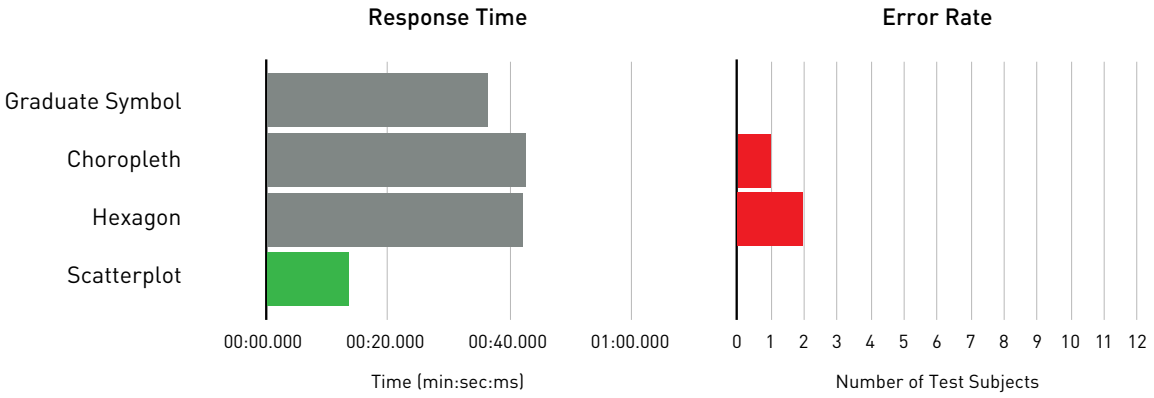

Fig. 17: User test results for finding and registering bottom three data points with lowest values.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:00.500	00:01.990	00:04.030	00:02.400
DCBA	00:01.880	00:01.900	00:02.940	00:03.560
BCDA	00:00.160	00:01.190	00:04.460	00:02.960
DACB	00:02.900	00:00.780	00:04.160	00:02.810
ACBD	00:03.160	00:02.860	00:21.480	00:01.240
CADB	00:01.010	00:01.000	00:11.680	00:01.450
CDAB	00:00.410	00:00.330	00:03.510	00:00.180
ADCB	00:02.060	00:00.350	00:01.480	00:01.610
BDAC	00:01.560	00:04.010	00:03.800	00:08.110
BADC	00:01.110	00:01.360	00:00.960	00:01.600
CABD	00:01.410	00:01.350	00:05.410	00:11.980
CDBA	00:00.910	00:00.860	00:03.440	00:04.110
Average	**00:01.423**	**00:01.498**	**00:05.612**	**00:03.501**

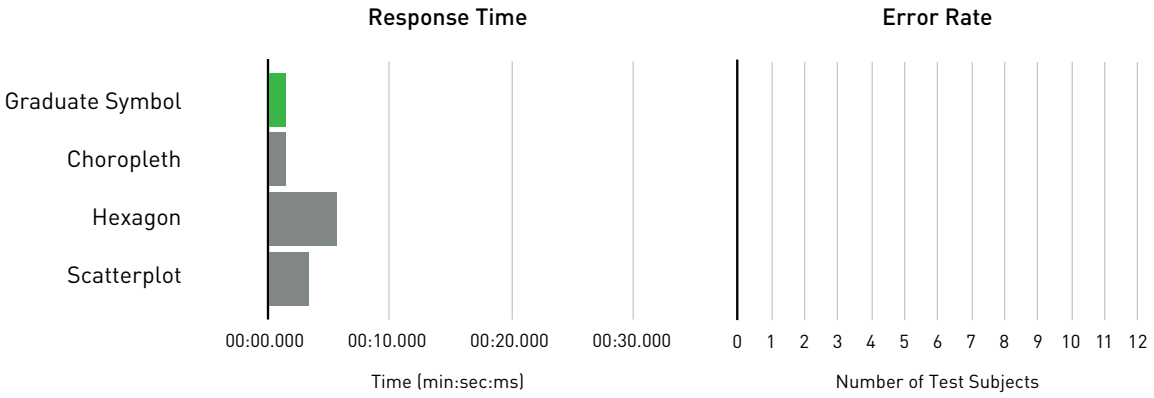

Fig. 18: User test results for recognizing a region with clustering of data points with the highest values.

Figure 19 displays the user testing results for all prototypes that used a univariate data set. The task was to recognize a region with a clustering of data points that held the lowest values. There were no significant differences in the average response time between the models (1.6 to 3.4 seconds). In this task there were no participant errors for any of the models.

Figure 20 displays the user testing results for all prototypes that used a univariate data set. The task was to compare the size of a data point of interest in the graduated symbol model, the intensity of colour in the choropleth and hexagon models, and the vertical position of the same data point in the scatterplot model, in all cases with the rest of the data points. The scatterplot model had the lowest average response time (4 seconds). The choropleth model had the highest average response time (11 seconds). The participant errors were high in the graduated symbol (11/12), choropleth (9/12) and hexagon models (10/12), whereas there were no participant errors in the scatterplot model.

Figure 21 displays the user testing results for all prototypes that used a bivariate data set with a series of positive values and a series of negative values. The task was to find and register top three data points with high values in the positive series, starting with the highest. The scatterplot model had the lowest average response time (18 seconds). The average response time for the rest of models was approximately 2 times higher. The scatterplot and choropleth models had no participant errors in finding and reporting the top three data points in the correct order. Two out of twelve participants made an error in the graduated symbol model and one participant made an error in the hexagon model.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:00.400	00:00.680	00:02.240	00:00.300
DCBA	00:22.760	00:02.830	00:00.480	00:09.280
BCDA	00:00.360	00:17.000	00:01.010	00:00.350
DACB	00:02.900	00:01.510	00:00.210	00:09.910
ACBD	00:03.730	00:08.730	00:02.250	00:00.430
CADB	00:01.150	00:00.780	00:01.880	00:00.930
CDAB	00:00.400	00:01.000	00:07.350	00:00.200
ADCB	00:02.060	00:00.210	00:00.350	00:00.890
BDAC	00:00.510	00:03.500	00:00.460	00:00.810
BADC	00:00.650	00:01.330	00:00.530	00:00.280
CABD	00:03.730	00:00.580	00:01.050	00:03.550
CDBA	00:02.430	00:00.550	00:01.690	00:02.660
Average	**00:03.423**	**00:03.225**	**00:01.625**	**00:02.466**

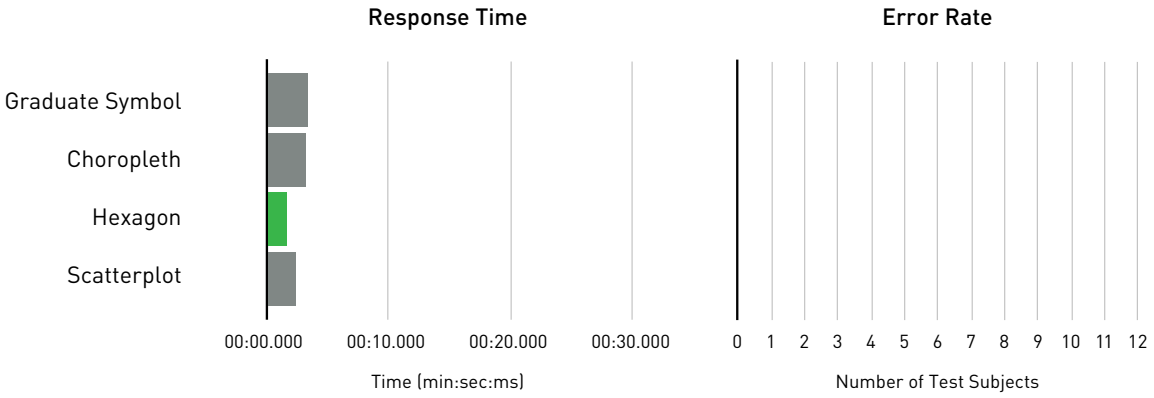

Fig. 19: User test results for recognizing a region with clustering of data points with the lowest values.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:04.280	00:30.530	00:07.350	00:10.110
DCBA	00:09.400	00:02.330	00:05.380	00:04.360
BCDA	00:09.800	00:09.580	00:01.380	00:00.510
DACB	00:00.480	00:03.160	00:05.160	00:00.760
ACBD	00:08.080	00:15.660	00:05.710	00:03.330
CADB	00:04.260	00:01.660	00:04.660	00:05.580
CDAB	00:01.330	00:00.380	00:06.550	00:04.040
ADCB	00:05.480	00:06.360	00:04.000	00:00.330
BDAC	00:17.130	00:38.410	00:16.960	00:04.550
BADC	00:09.110	00:05.230	00:05.750	00:10.460
CABD	00:09.510	00:19.000	00:02.830	00:00.410
CDBA	00:07.760	00:03.280	00:08.680	00:00.760
Average	**00:07.218**	**00:11.298**	**00:06.201**	**00:03.767**

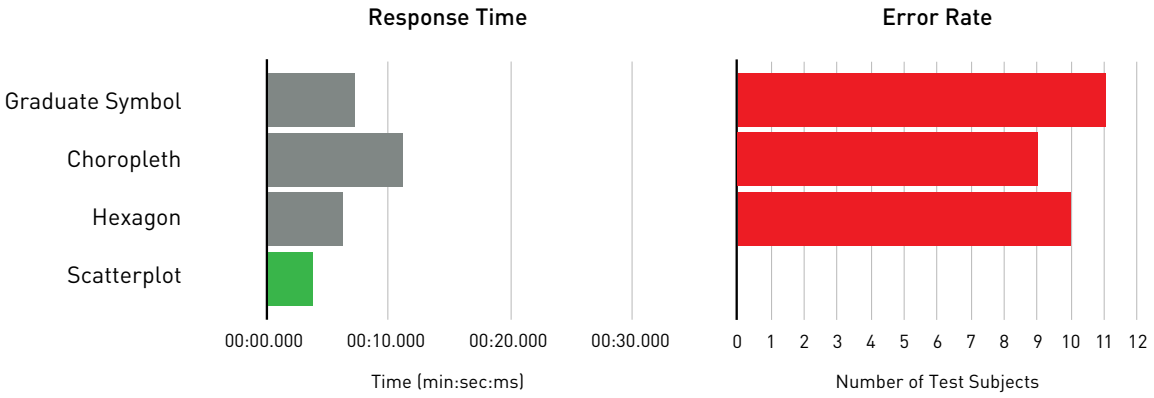

Fig. 20: User test results for comparing a data point of interest with remainder of the data points.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:12.060	00:15.300	00:32.030	00:07.010
DCBA	00:53.880	01:02.730	01:06.000	00:23.560
BCDA	00:47.500	01:02.280	00:52.400	00:29.170
DACB	00:23.400	00:16.030	00:25.060	00:23.860
ACBD	00:54.530	00:32.350	00:31.610	00:12.700
CADB	00:38.030	00:34.200	01:00.000	00:19.450
CDAB	00:24.200	00:18.880	00:36.120	00:22.230
ADCB	00:19.430	00:25.360	00:26.000	00:16.110
BDAC	00:18.230	00:32.400	00:23.300	00:11.310
BADC	00:24.850	00:38.430	00:20.330	00:13.510
CABD	00:17.310	00:18.980	00:36.170	00:15.900
CDBA	00:14.550	00:20.560	00:30.010	00:18.030
Average	**00:28.998**	**00:31.458**	**00:36.586**	**00:17.737**

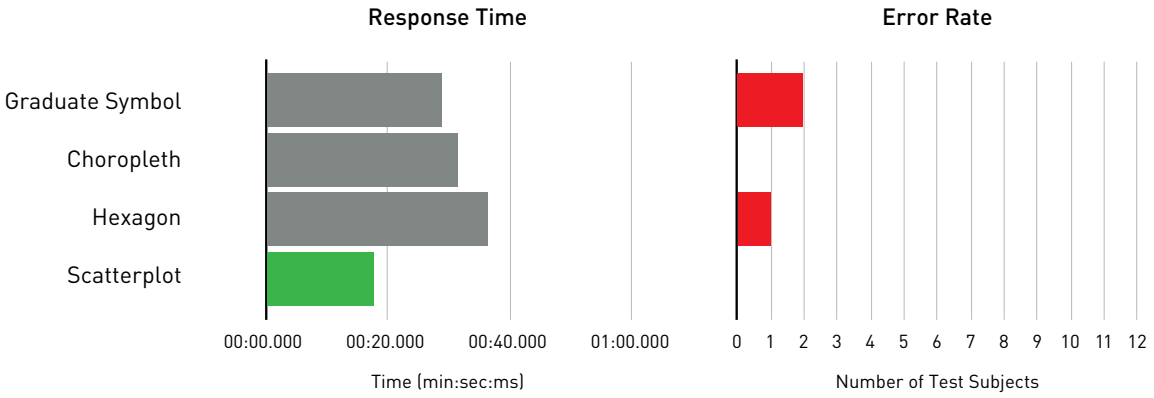

Fig. 21: User test results for finding the top three data points with high values in the positive series.

Figure 22 displays the user testing results for all prototypes that used a bivariate data set with a series of positive values and a series of negative values. The task was to find and register bottom three data points with low values in the negative series, starting with the lowest. The average response time in all the models were approximately the same (1 minute). One participant made an error using the scatterplot model. For the rest of the models the participant error rate was three out of twelve.

Figure 23 displays the user testing results for all prototypes that used a bivariate data set with a series of positive values and a series of negative values. The task was to recognize a region with a clustering of data points with high values in the positive series. The average response time was lowest in the choropleth model (17 seconds) and highest in the scatterplot model (26 seconds). The participant error rates were highest in scatterplot model (10/12) and hexagon model (9/12), and lowest in the choropleth model (4/12).

Figure 24 displays the user testing results for all prototypes that used a bivariate data set with a series of positive values and a series of negative values. The task was to recognize a region with a clustering of data points that held low values in the negative series. The average response times were lowest and approximately the same for the choropleth and scatterplot models (3-4 seconds), and highest and approximately the same for the graduated symbol and hexagon models (8 seconds). In this task there were no participant errors.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:28.450	00:42.000	01:46.160	00:18.130
DCBA	01:03.680	01:37.500	01:10.930	01:44.320
BCDA	00:44.330	03:05.680	01:02.030	01:17.670
DACB	00:39.210	00:22.700	00:38.450	01:15.850
ACBD	01:37.870	00:24.160	00:33.200	00:54.070
CADB	00:42.900	00:20.510	01:42.110	00:48.680
CDAB	01:32.490	00:56.970	01:45.090	01:00.230
ADCB	01:05.290	00:47.450	00:42.530	00:47.280
BDAC	00:45.970	01:33.700	00:57.870	01:31.970
BADC	00:47.980	00:56.620	00:22.880	00:19.930
CABD	00:36.760	00:27.000	01:15.120	00:34.870
CDBA	00:17.080	00:27.160	00:55.430	00:51.740
Average	**00:51.834**	**00:58.454**	**01:04.317**	**00:57.062**

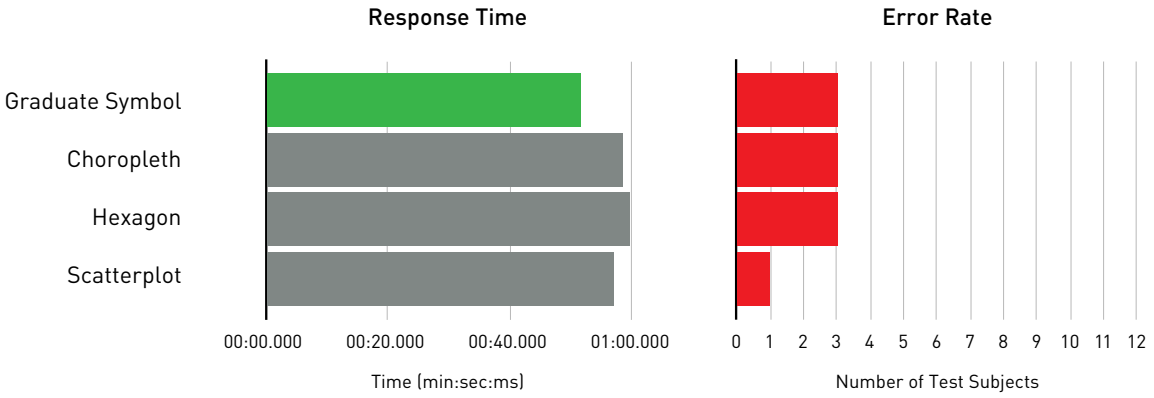

Fig. 22: User test results for finding the bottom three data points with low values in the negative series.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:15.460	00:14.730	00:34.320	00:11.530
DCBA	00:13.500	00:45.220	00:21.320	00:25.800
BCDA	00:05.730	00:07.910	00:09.500	00:15.310
DACB	00:21.550	00:08.510	00:03.430	00:04.960
ACBD	00:25.100	00:23.450	00:32.210	00:28.970
CADB	00:17.850	00:05.880	00:21.300	00:11.560
CDAB	00:15.410	00:16.080	00:19.750	00:07.560
ADCB	00:29.310	00:05.750	00:17.510	00:16.460
BDAC	00:16.060	00:21.210	01:16.820	02:13.290
BADC	00:25.860	00:07.730	00:02.230	00:21.310
CABD	00:33.960	00:27.830	00:09.380	00:17.050
CDBA	00:14.000	00:17.430	00:06.110	00:15.740
Average	**00:19.482**	**00:16.811**	**00:21.157**	**00:25.795**

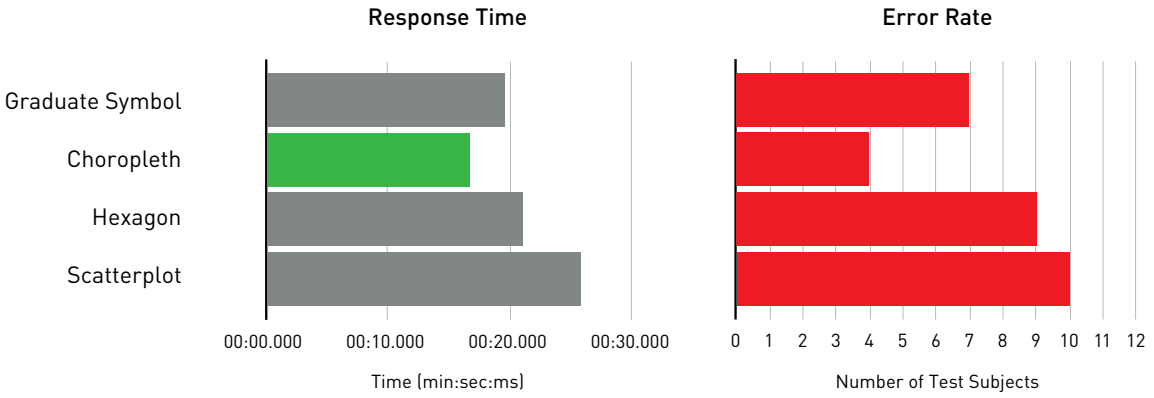

Fig. 23: User test results for recognizing a region with clustering of data points with high values in the positive series.

Order	Graduate Symbol (A)	Choropleth (B)	Hexagon (C)	Scatterplot (D)
CBAD	00:01.980	00:06.980	00:21.410	00:03.460
DCBA	00:10.380	00:00.910	00:04.930	00:02.360
BCDA	00:00.550	00:04.130	00:04.710	00:04.800
DACB	00:20.850	00:00.800	00:01.680	00:04.480
ACBD	00:07.730	00:02.950	00:09.160	00:04.580
CADB	00:06.230	00:04.150	00:07.560	00:02.800
CDAB	00:10.860	00:04.800	00:12.300	00:02.830
ADCB	00:06.180	00:01.910	00:05.860	00:02.400
BDAC	00:00.240	00:07.010	00:09.250	00:02.460
BADC	00:04.480	00:08.410	00:04.260	00:04.130
CABD	00:22.480	00:02.110	00:09.550	00:03.650
CDBA	00:02.110	00:03.530	00:10.830	00:03.240
Average	**00:07.839**	**00:03.974**	**00:08.458**	**00:03.433**

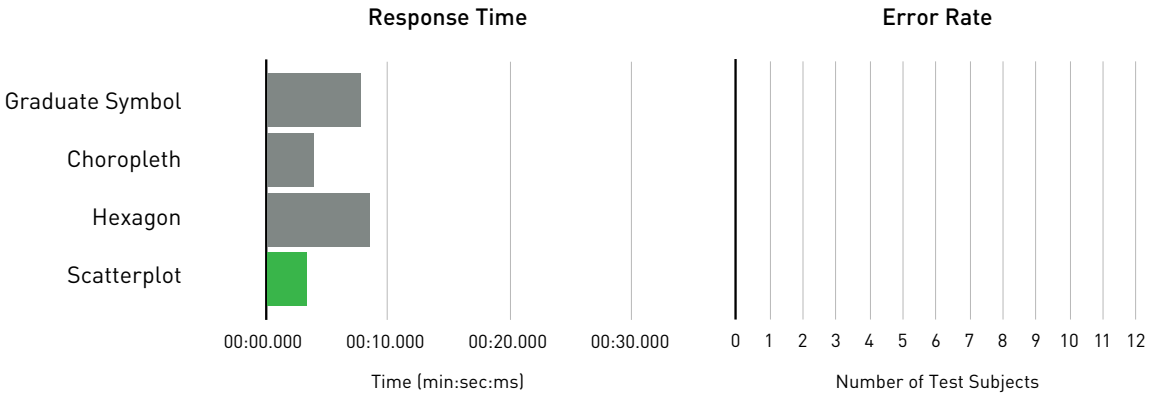

Fig. 24: User test results for recognizing a region with clustering of data points with low values in the negative series.

Qualitative Analysis

In this section I will report the results of my conversations with participants in regards to their experience using the different models.

Most of the participants reported that in both choropleth and hexagon models comparison of areas with low gradation in colour saturations—especially the colour red—was particularly hard. This process also depended on context: relations do in fact change if the same colour value is surrounded by other colours. Comparison was even harder in those cases where these areas did not share a boundary. On the other hand, recognizing patterns and clustering of data points was more evident in the choropleth and hexagon models.

Eight out of the twelve participants could read more information by looking at different saturations of colour compared to looking at the respective sizes of the circles in the graduated symbol model. In addition, the majority of participants reacted immediately to this fact: The processing of data points in the graduated symbol model was harder relative to the other models because of the overlapping of circles.

All the participants reported that the scatterplot model helped them to create a much more immediate and truer grasp of the data than did the other three models. For example, it was easier to find and register both the highest and lowest values. It was also easier to assess the overall picture because all the data points were the same size. However, in the scatterplot model it was hard to differentiate data points where the difference between their respective vertical positions was small. Another limitation with the scatterplot model was the lack of spatial logic, as there was no awareness of space and therefore of how places were located relative to each other; most of the participants did not, for example, notice or mention the ordering of the GTA regions in the scatterplot model.

Chapter 6

Conclusions and Future Directions

Chapter Six – Conclusions and Future Directions

Conclusions

The research question in this study was whether the absence of a map in visual representations of data sets with geospatial data affects negatively users' ability to process the pertinent data. To address this question four different models were developed and evaluated by means of a series of user testing sessions.

When the focus is not the geographic location but on the actual respective values of data points, the scatterplot model proved the most efficient with no participant errors. For example, when participants were asked to recognize a range of high and low values, the average response time was three times higher with some participant errors in the other three models (figs. 16 and 17). Also almost all the participants made errors in the same three models in another task where they compared a data point of interest with the rest of the data points (fig. 20).

When it comes to recognizing patterns, all the models performed more or less equally (figs. 18, 19 and 24) except when the number of data points with close values was relatively high. In this case, the choropleth model was marginally more efficient with only half the number of participant errors compared to the other models (fig. 23).

The argument for this study was that, in general, people can use map-based visualization models, such as choropleth and graduated symbol models, for all types of evaluation. What this study demonstrated was that the choice of models should support the task at hand. Although map-based visualizations might be well suited for geographic analysis and spatial navigation, they are not also optimal for representing the actual values of data points.

Future Directions

All the models described in this study are static. Further development of these models either to represent temporal data or to enable the users to manipulate the value of data points directly could be the subject of future studies. Figure 23 provides an early prototype for direct manipulation of values using the scatterplot model.

Another direction could be the inclusion of a map in a bar graph model. Doing so could show the geographic location of selected data points and therefore enhance the users' spatial logic and awareness. Figure 24 provides an early prototype for the integration of a map into a bar graph model.

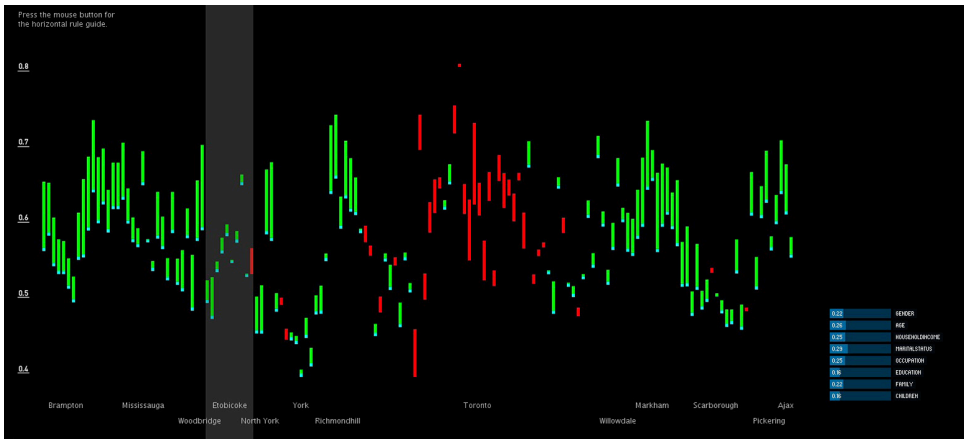

Fig. 25: An early prototype of direct manipulation of values using the scatterplot model.

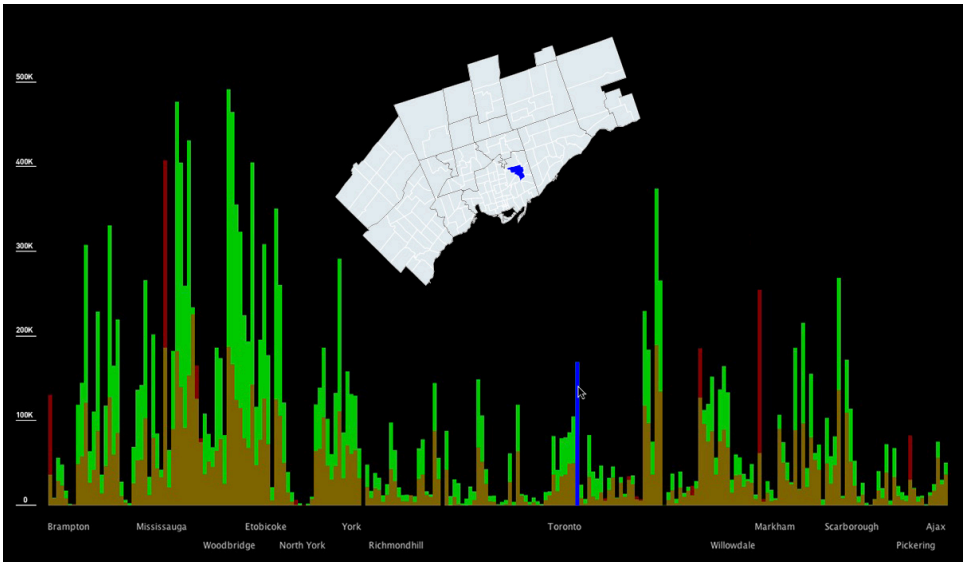

Fig. 26: An early prototype for integration of a map into a bar graph model.

Sources Cited

Baddeley, Alan. "Working Memory: The Interface Between Memory and Cognition." Journal of Cognitive Neuroscience 4.3 (1992): 281-88. Print.

Bertin, Jacques. *Semiology of Graphics: Diagrams, Networks, Maps.* Redlands: ESRI, 2011. Print.

"Big Data at the Speed of Business." IBM. n.d. Web. 15 Dec. 2013. <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>

Callaghan, Tara C. "Interference and Dominance in Texture Segregation: Hue, Geometric Form, and Line Orientation." *Perception and Psychophysics* 46.4 (1989): 299-311. Print.

Candy, Linda. *Practice Based Research: A Guide.* Sydney: U of Technology, 2006. Print.

"Choropleth Map." *Wikipedia: The Free Encyclopedia.* Web. 1 Mar. 2014.

Evergreen, Stephanie D. H. *Presenting Data Effectively: Communicating Your Findings for Maximum Impact.* Los Angeles: Sage, 2014. Print.

Fry, Ben. *Visualizing Data.* Beijing: O'Reilly Media, 2008. Print.

Hollis, Richard. *Graphic Design: A Concise History.* New York: Thames and Hudson, 1994. Print.

Howlett, Virginia. *Visual Interface Design for Windows: Effective User Interfaces for Windows 95, Windows NT, and Windows 3.1.* New York: Wiley, 1996. Print.

Iliinsky, Noah P. N., and Julie Steele. *Designing Data Visualizations.* Sebastopol, CA: O'Reilly, 2011. Print.

James, Josh. "How Much Data is Created Every Minute?" *Domosphere.* 8 June 2012. Web.15 Dec. 2013. <http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/>.

Kirk, Andy. *Data Visualization: A Successful Design Process: A Structured Design Approach to Equip You with the Knowledge of How to Successfully Accomplish Any Data Visualization Challenge Efficiently and Effectively.* Birmingham: Packt, 2012. Print.

Meirelles, Isabel. *Design for Information: An Introduction to the Histories, Theories and Best Practices Behind Effective Information Visualizations.* Beverly, MA: Rockport, 2013. Print.

Myatt, Glenn J., and Wayne P. Johnson. *Making Sense of Data III: A Practical Guide to Designing Interactive Data Visualizations.* Hoboken, NJ: Wiley, 2011. Print.

"Permutations." Highline Advanced Math Program, n.d. Web. 22 Apr. 2014. <http://home.avvanta.com/~math/permutations2.html>

Perreault, Jr., William D. "Controlling Order-Effect Bias." *Public Opinion Quarterly* 39.4 (1975): 544. Print.

Reas, Casey, and Ben Fry. *Getting Started with Processing.* Beijing: O'Reilly, 2010. Print.

"Scalable Vector Graphics." *Wikipedia: The Free Encyclopedia.* Web. 15 Mar. 2014.

Shenas, Haleh H., and Victoria Interrante. "Compositing Color with Texture for Multi-Variate Visualization." *Proceedings: GRAPHITE 2005: 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia.* Ed. Stephen N. Spencer. New York: ACM, 2005. 443-446. Print.

Shneiderman, Ben. 1996. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations." *Proceedings: 1996 IEEE Symposium on Visual Languages.* Los Alamitos, CA: 1996. 336-343. Print.

Surkys, Tadas, Algis Bertulis, and Aleksandr Bulatov. "Delboeuf Illusion Study." Medicina 42.8 (2006): 673-81. Print.

Tidwell, Jenifer. *Designing Interfaces.* Beijing: O'Reilly, 2006. Print.

Tufte, Edward R. *The Visual Display of Quantitative Information.* Cheshire, CT: Graphics, 2001. Print.

Tukey, John W. *Exploratory Data Analysis.* Reading, MA: Addison-Wesley, 1977. Print.

Yau, Nathan. *Data Points: Visualization That Means Something.* Indianapolis, IN: Wiley, 2013. Print.

———. *Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics.* Indianapolis, IN: Wiley, 2011. Print.

Ware, Colin. *Information Visualization: Perception for Design.* Waltham, MA: Morgan Kaufmann, 2013. Print.

———. *Visual Thinking for Design.* Burlington, MA: Morgan Kaufmann, 2008. Print.

Woodman, Geoffrey F., Shaun P. Vecera, and Steven J. Luck. "Perceptual Organization Influences Visual Working Memory." *Psychonomic Bulletin and Review* 10.1 (2003): 80-87. Print.

Bibliography

Abrams, Janet, and Peter Hall, eds. *Else/where: Mapping: New Cartographies of Networks and Territories.* Minneapolis: U of Minnesota Design Institute, 2006. Print.

Andrienko, Gennady, et al. "Challenging Problems of Geospatial Visual Analytics." *Journal of Visual Languages and Computing* 22.4 (2011): 251-56. Print.

Cairo, Alberto. *The Functional Art: An Introduction to Information Graphics and Visualization.* Berkely, CA: New Riders, 2013. Print.

Hallisey, Elaine J. "Cartographic Visualization: An Assessment and Epistemological Review." *The Professional Geographer* 57.3 (2005): 350-364. Print.

Lima, Manuel. *Visual Complexity: Mapping Patterns of Information.* New York: Princeton Architectural, 2011. Print.

McCandless, David. *The Visual Miscellaneum: A Colorful Guide to the World's Most Consequential Trivia.* New York: Collins, 2009. Print.

Monmonier, Mark. "Lying with Maps." *Statistical Science* 20.3 (2005): 215-22. Print.

Pearson, Matt. *Generative Art: A Practical Guide Using Processing.* Shelter Island, New York: Manning, 2011. Print.

Rhyne, Theresa Marie, and Alan McEachern. "Visualizing Geospatial Data." *Proceedings of the Conference on SIGGRAPH 2004 Course Notes.* Print.

Rubin, Jeffrey. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests.* New York: Wiley, 1994. Print.

Shiffman, Daniel. *Learning Processing: A Beginner's Guide to Programming Images, Animation, and Interaction.* Amsterdam: Morgan Kaufmann, 2008. Print.

Steele, Julie, and Noah P. N. Iliinsky. *Beautiful Visualization.* Beijing: O'Reilly, 2010. Print.

Tufte, Edward R. *Beautiful Evidence.* Cheshire, CT: Graphics, 2006. Print.

———. *Envisioning Information.* Cheshire, CT: Graphics, 2011. Print.

———. *Visual Explanations: Images and Quantities, Evidence and Narrative.* Cheshire, CT: Graphics, 2010. Print.

Wilson, David C., et al. "Charting New Ground: Modeling User Behavior in Interactive Geovisualization." *GIS '08 16th International Symposium on Advances in Geographic Information Systems.* New York: ACM, 2008. 1-4. Print.

Wisniewski, Pamela Karr, et al. "Grounding Geovisualization Interface Design: A Study of Interactive Map Use." *Proceedings of CHI '09 Extended Abstracts on Human Factors in Computing Systems.* New York: ACM, 2006. 3757-3762. Print.

Zikopoulos, Paul C., et al. *Harness the Power of Big Data: The IBM Big Data Platform.* New York: McGraw, 2013. Print.

Appendix A

REB Approval Documents

Research Ethics Board

October 15, 2013

Dear Borzu Talaie,

RE: OCADU 124 “Data Driven Journalism.”

The OCAD University Research Ethics Board has reviewed the above-named submission. The protocol and the consent form dated October 15, 2013 are approved for use for the next 12 months. If the study is expected to continue beyond the expiry date (October 14, 2014) you are responsible for ensuring the study receives re-approval. Your final approval number is **2013-36.**

Before proceeding with your project, compliance with other required University approvals/certifications, institutional requirements, or governmental authorizations may be required. It is your responsibility to ensure that the ethical guidelines and approvals of those facilities or institutions are obtained and filed with the OCAD U REB prior to the initiation of any research.

If, during the course of the research, there are any serious adverse events, changes in the approved protocol or consent form or any new information that must be considered with respect to the study, these should be brought to the immediate attention of the Board.

The REB must also be notified of the completion or termination of this study and a final report provided before you graduate. The template is attached.

Best wishes for the successful completion of your project.

Yours sincerely,

Tony Kerr, Chair, OCAD U Research Ethics Board

OCAD U Research Ethics Board: rm 7520c, 205 Richmond Street W, Toronto, ON M5V 1V3
416.977.6000 x474

Research Ethics Board

February 7, 2014

Dear Borzu Talaie,

RE: OCADU 124 “Data Driven Journalism.”

AMENDMENT

The OCAD University Research Ethics Board has reviewed and approved your amendment request. Your final approval number remains **2013-36.**

Please remember that if, during the course of the research, there are any serious adverse events, changes in the approved protocol or consent form or any new information that must be considered with respect to the study, these should be brought to the immediate attention of the Board.

Best wishes for the successful completion of your project.

Yours sincerely,

Tony Kerr, Chair, OCAD U Research Ethics Board

Appendix B

The Globe and Mail Internship Terms and Conditions

THE GLOBE AND MAIL
CANADA'S NATIONAL NEWSPAPER

May 24, 2013

Borzu Talaie
c/o Adrian Norris
The Globe and Mail

Dear Borzu:

The purpose of this letter is to confirm our recent discussions regarding a summer internship opportunity at The Globe and Mail. I am pleased to present you with this offer, and I look forward to you returning (1) signed original, confirming your agreement with the following terms and conditions:

POSITION DETAILS:

1. Your position title is Research Intern, reporting directly to me.

2. Your summer internship with The Globe and Mail will commence effective July 2, 2013. This position is temporary and you will be employed up to and including August 30, 2013.

3. Your work schedule will be ~~five (5)~~ four (4) BT days a week (~~Monday to Friday~~) (Mon, Tues, Wed, Fri) BT. Any changes to your schedule must be discussed with and approved by your manager.

4. Upon completion of your summer internship, you will be paid an honorarium of $1,125.00 CAD (equivalent to $125.00 CAD per week from July to August).

NON-DISCLOSURE AGREEMENT:

In connection with the disclosure of personal information and other information (the "Confidential Information") relating to the business of The Globe and Mail Inc. ("The Globe") provided by you to The Globe, you hereby agree that the Confidential Information is being provided to you in connection with the Internship and may not be used by you for any other purpose without the prior consent of The Globe and, in the case of personal information, the individual in question. The Confidential Information must be kept confidential by you at all times and must be destroyed or returned to The Globe upon completion of your Internship.

This Agreement shall be governed and construed in accordance with the laws of the Province of Ontario and the federal laws of Canada applicable therein.

NON-COMPETE AGREEMENT:

In consideration of employment with The Globe and Mail, you are required to not compete against The Globe and Mail or any other division of Woodbridge Company Ltd. in business interests during your summer internship. Specifically, this means you agree not to solicit The Globe and Mail's customers either directly or indirectly for business purposes during your summer internship with The Globe and Mail;

To signify your acceptance of this offer including the Non-Disclosure Agreement and Non-Compete Agreement, please sign both copies of this letter, return one signed original to Elizabeth MacDonald in Human Resources and retain one for your records.

Borzu, I am delighted at the prospect of welcoming you to our team. We have many exciting challenges ahead of us. If you have any questions, please do not hesitate to contact me at (416) 585-5685.

THE GLOBE AND MAIL
CANADA'S NATIONAL NEWSPAPER

Sincerely,

Adrian Norris
Creative Director

Borzu Talaie

24/5/13
Date

/em

Appendix C

Data Visualization Tools

Appendix C – Data Visualization Tools

Unlike other design disciplines, in data visualization there is no one tool that can perform all the desired visualization functions. Therefore a portfolio of different technical solutions should be developed to satisfy the visualization phase. These solutions are categorized into two major groups: Out-of-the-box visualization tools and programming as a tool.

Out-of-the-Box Visualization Tools

These are by far the easiest solutions for beginners—the process could require only loading a data set and then choosing a predefined visualization method. At the same time they are not particularly flexible in producing exactly what the user might want to achieve. Also, some software applications provide many buttons and options that need to be learned.

Microsoft Excel or Apple Numbers
All the familiar spreadsheets that provide the user with all the standard chart types.

Google Spreadsheets
This is essentially the online version of the spreadsheet software mentioned above. The advantages of using Google Spreadsheets are the accessibility of the software—it may be retrieved through any web browser from anywhere—and the ease in sharing both the data and results with other users.

Many Eyes
Many Eyes is an ongoing research project by the IBM Visual Communication Labs. This online experimental application enables users to explore large data sets as a group. It is interactive, and provides several advanced customization tools.

Tableau
This relatively new software is designed mainly to explore and analyze data visually. It offers several interactive visualization tools and is excellent for data management. It also enables the user to mix and match different displays, link-in a dynamic data set for a custom view or a dashboard as a snapshot of what is taking place in the data set.

Programming as a Tool

Programming provides the user with greater flexibility and the ability to adapt to different data types. However, the noteworthy amount of time required to learn a new programming language compromises these advantages.

Python
Python provides a clean and easy-to-read syntax that can handle large volumes of data. It is also somewhat weak aesthetically, and its outputs, to be presentable, need modification through a graphic application.

PHP
PHP is one of the principal programming languages for the web. It provides easy setup for most web servers, and maintains a highly flexible graphic library called GD. The library enables the programmer to create visualizations from scratch or to draw with ease basic charts and graphs.

Processing
This is a lightweight open-source programming language geared towards designers and data artists. It permits the programmer to create an animated and/or interactive graphic with only a few lines of code. This language is regrettably also slow on the computers of some users, but this is its only limitation.

R
A free and open-source statistical computing software with a good statistical graphics function. There are many R packages through which the programmer can generate graphics with only a few lines of code.

D3.js
Launched in 2011 by the Stanford Visualization Group, D3.js is a JavaScript library. D3.js is particularly popular because of its countless libraries and plug-ins, and its ability to work in web browser environment.

Appendix D

Mapping the GTA FSAs on the X-axis

Appendix D – Mapping the GTA FSAs on the X-axis

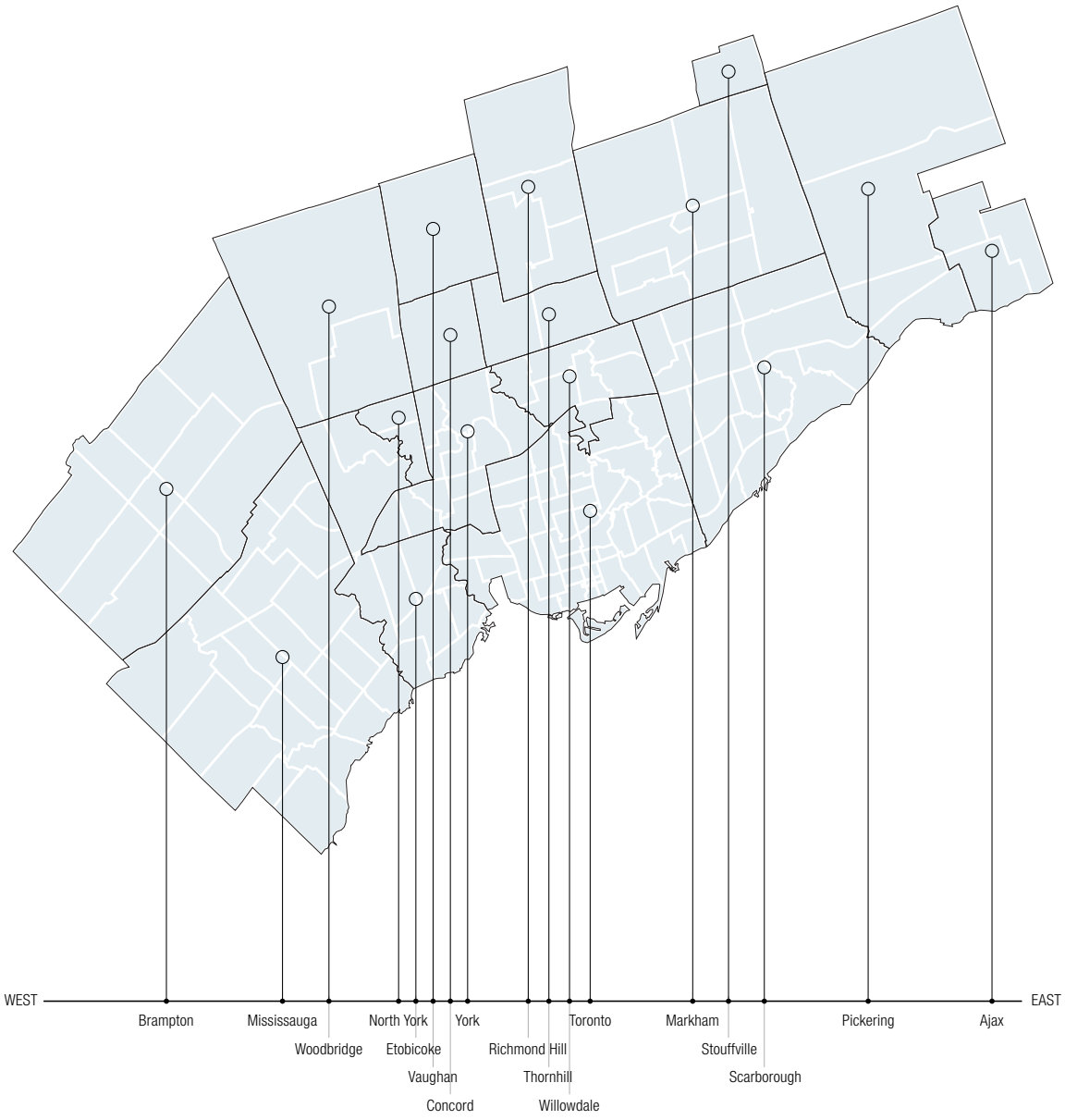

Fig. 27: Mapping the Greater Toronto Area FSAs on the x-axis.

Appendix E

User Testing Scripts

Graduated Symbol Model Using Univariate Values

This is the map of the Greater Toronto Area. Each polygon represents a postal code. Points on the map, painted in blue, represent the centre of each postal code and their size has been scaled to display the score of that postal code. How the score is calculated is irrelevant to this study. What is important to know is that larger circles represent higher scores and smaller circles represent lower scores—you can always refer to the legend to understand these relationships.

In order to get more information about each of these points, such as 1- the postal code they represent, 2- their region and 3- their score in numeric format, you can hover the mouse over each point. Once a point is selected, its colour changes to red and extra information appears at the bottom right of the map. The orange lines define the boundaries of each region. Please note that scores in numeric format range between 0.0 and 1.0.

User tasks:

1. Name in order the top three postal codes with the highest score, starting with the highest.
2. Name in order the bottom three postal code with the lowest score, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with high scores?
4. Where in the map do you see a clustering of postal codes with low scores?
5. By considering this specific data point that I am showing you on the screen with the tip of my pen, overall would you say the rest of the postal codes shown here have a higher score than this data point, a lower score than this data point, or their average would be the same as this data point?

Fig. 28: The graduated symbol model with univariate values.

Choropleth Model Using Univariate Values

This is the map of the Greater Toronto Area. Each polygon represents a postal code. Each polygon has been painted with a different saturation of the colour blue. The amount of saturation reflects the score for that postal code. How the score is calculated is irrelevant to this study. What is important to know is that polygons painted with high saturation (darker blue) have higher scores and polygons painted with low saturation (lighter blue) have lower scores—you can always refer to the legend to understand these relationships. The orange lines define the boundaries of each region. Polygons painted in grey represent areas that their value is not known to us—data is missing from the data set.

In order to get more information about each of these polygons, such as 1- the postal code they represent, 2- their region and 3- their score in numeric format, you can hover the mouse over each polygon. Once a polygon is selected, its colour changes to red and extra information appears at the bottom-right side of the map. Please note that scores in numeric format range between 0.0 and 1.0.

User tasks:

1. Name in order the top three postal codes with the highest score, starting with the highest.
2. Name in order the bottom three postal code with the lowest score, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with high scores?
4. Where in the map do you see a clustering of postal codes with low scores?
5. By considering this specific data point that I am showing you on the screen with the tip of my pen, overall would you say the rest of the postal codes shown here have a higher score than this data point, a lower score than this data point, or their average would be the same as this data point?

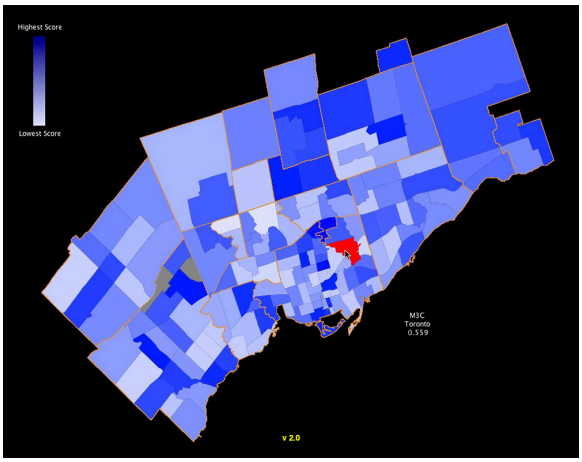

Fig.29: The choropleth model with univariate values.

Hexagon Model Using Univariate Values

This is a distorted and abstract map of the Greater Toronto Area. Each hexagon that has been painted with a different saturation of the colour blue, represents a postal code. The amount of saturation reflects the score for that hexagon. How the score is calculated is irrelevant to this study. What is important to know is that hexagons painted with high saturation (darker blue) have higher scores and hexagon painted with low saturation (lighter blue) have lower scores–you can always refer to the legend to understand these relationships. The orange lines define the boundaries of each region. Hexagons painted in grey represent areas that the desirability value is not known to us–data is missing from the data set.

In order to get more information about each of these hexagons, such as 1- the postal code they represent, 2- their region and 3- their score in numeric format, you can hover the mouse over each hexagon. Once a hexagon is selected, its colour changes to red and extra information appears at the bottom of the map. Please note that scores in numeric format range between 0.0 and 1.0.

User tasks:

1. Name in order the top three postal codes with the highest score, starting with the highest.
2. Name in order the bottom three postal code with the lowest score, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with high scores?
4. Where in the map do you see a clustering of postal codes with low scores?
5. By considering this specific data point that I am showing you on the screen with the tip of my pen, overall would you say the rest of the postal codes shown here have a higher score than this data point, a lower score than this data point, or their average would be the same as this data point?

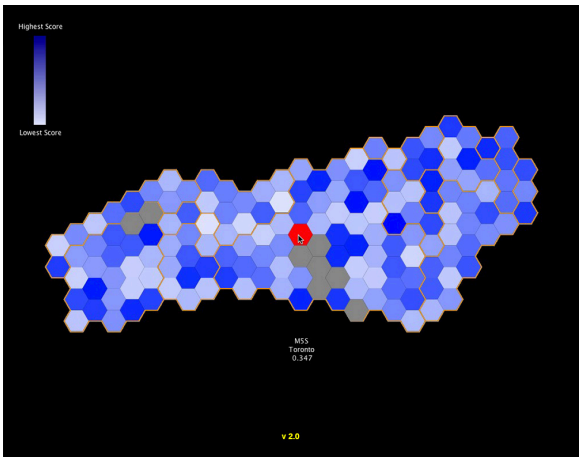

Fig. 30: The hexagon model with univariate values.

Scatterplot Model with Geographic References Using Univariate Values

This is a visual representation of postal code scores in the Greater Toronto Area. Each blue point on this model represents a postal code. The score for each postal code is shown by its position on the vertical axis. The higher the score, points are positioned closer to the top of the screen and the lower the score, points are positioned closer to the bottom of the screen—you can always refer to the scale on the left hand side of the screen to understand these relationships. You can also take advantage of the horizontal rule guide by pressing and holding the mouse button to better understand the position of each postal code on the scale.

When you move your mouse sideways on this model, depending on the position of your mouse a region with all of its postal codes gets highlighted. You can find the name of each region at the bottom of the screen. These regions have been ordered from left to right (west to east) based on their approximate geographic location.

In order to get more information about each of these points such as 1- the postal code they represent, 2- their region and 3- their score in numeric format, you can hover the mouse over each point. Once a point is selected, its colour changes to red and extra information appears above the mouse cursor.

User tasks:

1. Name in order the top three postal codes with the highest score, starting with the highest.
2. Name in order the bottom three postal code with the lowest score, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with high scores?
4. Where in the map do you see a clustering of postal codes with low scores?
5. By considering this specific data point that I am showing you on the screen with the tip of my pen, overall would you say the rest of the postal codes shown here have a higher score than this data point, a lower score than this data point, or their average would be the same as this data point?

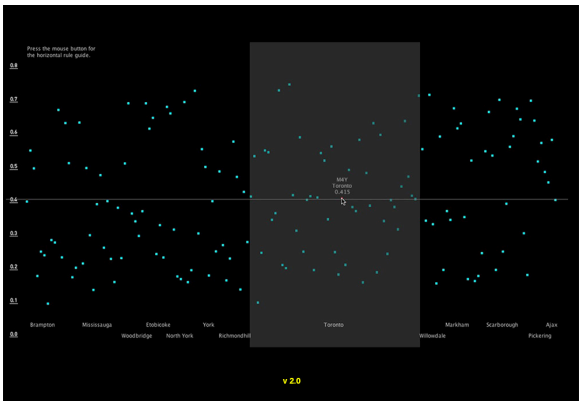

Fig. 31: The scatterplot model with univariate values.

Graduated Symbol Model Using Bivariate Values

You are already familiar with this model. The only difference between this model and the previous one is that two sets of values are being visualized here.

Points in blue represent postal codes with a net profit (positive values), and points in red represent postal codes with a net loss (negative values). The size of the points indicate the amount of profit or loss for that postal code. For example a blue circle with a big diameter represents a postal code with a very high net profit and a red circle with a small diameter represents a postal code with a very low net loss. You can always refer to the legend to understand these relationships.

User tasks:
1. Name in order three postal codes with the highest net profit, starting with the highest.
2. Name in order three postal codes with the lowest net loss, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with low net profit?
4. Where in the map do you see a clustering of postal codes with low net loss?

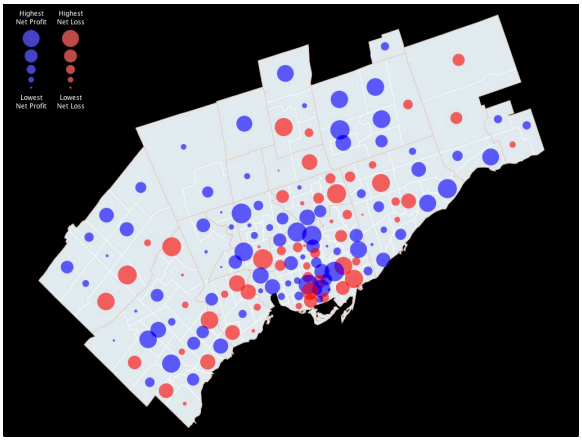

Fig. 32: The graduated symbol model with bivariate values.

Choropleth Model Using Bivariate Values

You are already familiar with this model. The only difference between this model and the previous one is that two sets of values are being visualized here.

Polygons in blue represent postal codes with a net profit (positive values), and polygons in red represent postal codes with a net loss (negative values). The amount of saturation for each colour indicates the amount of profit or loss for each postal code. For example dark blue represents a postal code with a very high net profit and light red represents a postal code with a very low net loss. You can always refer to the legend to understand these relationships. Polygons painted in grey represent areas that their value is not known to us—data is missing from the data set.

User tasks:

1. Name in order three postal codes with the highest net profit, starting with the highest.
2. Name in order three postal codes with the lowest net loss, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with low net profit?
4. Where in the map do you see a clustering of postal codes with low net loss?

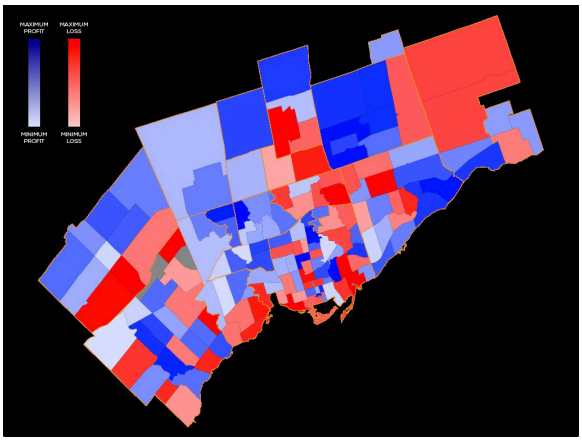

Fig. 33: The choropleth model with bivariate values.

Hexagon Model Using Bivariate Values

You are already familiar with this model. The only difference between this model and the previous one is that two sets of values are being visualized here.

Hexagons in blue represent postal codes with a net profit (positive values), and hexagons in red represent postal codes with a net loss (negative values). The amount of saturation for each colour indicates the amount of profit or loss for that postal code. For example dark blue represents a postal code with a very high net profit and light red represents a postal code with a very low net loss. You can always refer to the legend to understand these relationships. Hexagons painted in grey represent areas that their value is not known to us—data is missing from the data set.

User tasks:

1. Name in order three postal codes with the highest net profit, starting with the highest.
2. Name in order three postal codes with the lowest net loss, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with low net profit?
4. Where in the map do you see a clustering of postal codes with low net loss?

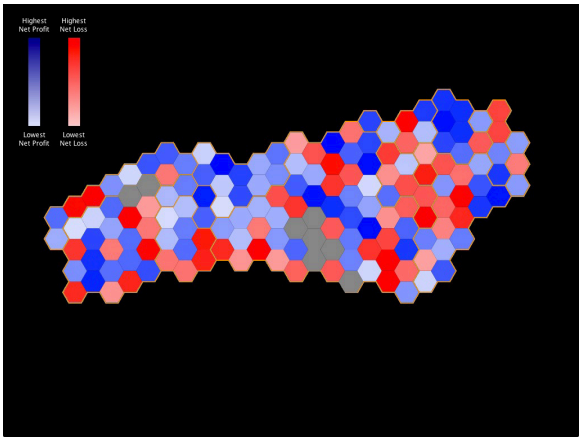

Fig. 34: The hexagon model with bivariate values.

Scatterplot Model with Geographic References Using Bivariate Values

You are already familiar with this model. The only difference between this model and the previous one is that two sets of values are being visualized here.

Points in blue represent postal codes with a net profit (positive values), and hexagons in red represent postal codes with a net loss (negative values). The vertical position of the points indicates the amount of profit or loss for that postal code. For example, postal codes with high net profit are located closer to the top of the screen and postal codes with low net loss are located close to the centre of the screen—zero marker on the scale. You can always refer to the scale on the left hand side to understand these relationships.

User tasks:
1. Name in order three postal codes with the highest net profit, starting with the highest.
2. Name in order three postal codes with the lowest net loss, starting with the lowest.
3. Where in the map do you see a clustering of postal codes with low net profit?
4. Where in the map do you see a clustering of postal codes with low net loss?

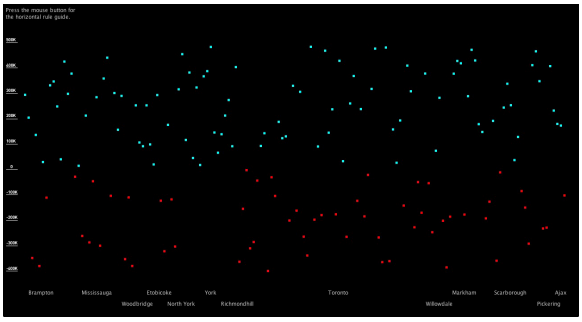

Fig. 35: The scatterplot model with bivariate values.

Appendix F

Sample Prototypes

Appendix F – Sample Prototypes

This Appendix—enclosed as a DVD—contains executable version of all the models for both Mac and Windows platforms. These models have been categorized into two subfolders: *Univariate folder* which contains prototypes using the univariate data set and *Bivariate folder* which contains prototypes using the bivariate data set.

List of Files

Univariate Folder
Choropleth Model
Graduated Symbol Model
Hexagon Model
Scatterplot Model

Bivariate Folder
Choropleth Model
Graduated Symbol Model
Hexagon Model
Scatterplot Model