CHAPTER 1

Introduction

1.1 Background and rationale

WYSIATI or *What You See Is All There Is,* is a concept originated by a Nobel Prize Laureate, Daniel Kahneman in his much acclaimed book, *Thinking, Fast and Slow* (2011). It describes how over-confident humans are in terms of believing in only what they see and assuming that all the facts are presented, while they ignore other unknown or unfamiliar information that might be more relevant or true to the story or event. Human behaviors and thoughts are the focus of the book, in which Kahneman proposes two systems of thinking modes – System 1 and System 2. System 1 is fast, automatic, uncontrolled, and subconscious, while System 2 is slower, effortful, deliberate, and rational.

The ideas from this book come to our attention and they are linked with what we have been interested in – data presentation. When people are presented with data, it can be presumed, based on System 1 from the book, that their first impression stimulates a rapid response and they tend to believe only what they can observe from the data presented. The problem can occur when the presentation cannot convey all the important information. Hence, people cannot extract the facts from the data as they fall into the *WYSIATI* situation, where *"all there is"* is not actually in the presentation.

How to overcome the lack of real *"all there is"* becomes an interesting topic. One of the approaches to present the data is to deliver it in a graphical style. It has been

recognized that graphical display or data visualization is a very useful tool to obtain information and analyse data (Tufte, 1990). This inspiring topic is the rationale for this thesis, investigating the approaches on creating a clear and accurate data visualization and fetching as much important information as possible.

We transform this idea into practical implementation using real world data. In this study, topics of interest are chosen based on their characteristics and some previous research. For this reason, we select four different, but related applications to be demonstrated in the thesis.

The first application employs a basic visualization – graphing. In this case, our interest is demographic data, normally in age-specific form. Even though such demographic data are usually discrete for each age range of 5- or 10-year intervals, a demographer requires a graph to be continuous for further analysis on cohorts (persons born in the same year). This task can be done by data interpolation and one of the preferred interpolations for demographic data is a cubic spline function because of its smoothness optimality. There are two requirements for this spline function. The first is that first and second derivatives often need to be zero for boundary conditions at the extremes, while the latter requires a fitting spline to be a non-decreasing monotonic function because populations, births, and date rates are normally non-negative. McNeil, Trussell, & Turner (1977) proposed a method for satisfying these two requirements. However, the proposed idea fails to satisfy the non-negativity assumption and it forces a fitting spline to be of degree 5 or higher. A further study by Smith, Hyndman, & Wood (2004) demonstrated the failure to satisfy the non-negativity assumption in the aforementioned proposal, and suggested a modification

by using the Hyman filter (Hyman, 1983). However, the computation for this part is complex and a special computer program is required to run the calculation. For this application, we propose a simple idea of adding artificial points of data to fit a simple cubic spline to demographic data and satisfy the boundary conditions.

The second application still relates to demographic data. In this case, we are dealing with registered mortality data in Thailand between 1999 and 2001. There is an indication that the death registration system in Thailand is of low quality, because almost half of registered deaths are considered ill-defined (Mathers et al, 2005). A death record outside a hospital is frequently coded by the head of the village is either lacking in proper medical training or having a little background of the deceased (Rukumnuaykit, 2006). In this study, we apply a statistical method to approximate the proportion of ill-defined deaths outside hospital in Thailand to evaluate the quality of the data. The difference between registered mortality data in this application, compared to the previous one, is that registered mortality data are not only age and gender-specific, but cause and district-specific as well. Therefore, a graphical display for this application involves spatial mapping with locations or districts.

The third application remains spatial mapping, but in this case we attempt to display information on more than one orientation and compare them over the location. To achieve this, a map needs to be in a three-dimensional system instead of a two-dimensional system as in the second application. The data of choice are incidence rate of terrorism events in three southern provinces of Thailand from 2004 to 2009. The incidence rates on a two-dimensional thematic map can illustrate information inadequately since the map cannot handle the magnitude of the rates between each

location clearly. Another problem with a two-dimensional map is that it lacks space to display enough information and it requires another display dimension to do so.

Our last application returns to a two-dimensional map. We select Thailand's national election results as data in this example. Instead of using tables and figures like a conventional presentation for election data, our aim is to display election results in a thematic map with additional information for elected candidates. According to Thai election rules, electoral districts in the parliament election are created by using the population density. The problem arises with the capital of Thailand, Bangkok, and its surrounding provinces, covering only 6800 km^2, but having a number of possible members of parliament of around 60, while the largest province in area, Nakorn Ratchasima, covers more than 20,000 km^2, but only has around 15 possible candidates (Department of Provincial Administration, 2011). The implication is that graphical symbol in the Bangkok area will be clustered and not detailed. We propose the idea of "local projection", which in this case stretches a map around a particular area to increase its visibility.

All applications are implemented using methodology introduced in the following chapter. The results and all four published manuscripts, for the four applications, are combined in Chapter 3. Chapter 4 is a conclusion of this thesis with discussion and further suggested work.

1.2 Literature review

A Literature review in the section merely focuses on the topic of data visualization. Related literature reviews of four applications studied in this thesis are included in published manuscripts in Chapter 3.

Friendly (2009) explains that data visualization is an important branch of science. It can also be seen as information abstracted in attributes or variables for information units. This topic of data visualization can be focused on two main schemes - statistical graphics and thematic cartography. These schemes are related to the visual representation of quantitative and categorical data, but with different objectives. Statistical graphics are suitable to applications in which graphical methods are utilized to support statistical analysis, while thematic cartography mainly represents spatial data. However, these two schemes have become overlapping in visual representation for exploratory and discovery studies, including simple location mapping, spatial distributions of geographic characteristics, and graphical methods used to portray patterns, trends, and indications.

Data visualization shows practical benefits. Tufte (1997) gave examples of how data presentations can be used to investigate some high-profiled problems. One of the examples is the work of Dr. John Snow, who was investigating a cholera epidemic in London in 1854. Dr. Snow found the source of cholera by creating a map of the area where deaths had occurred. It could be shown that almost all the victims were using a well pump on a particular street (Broad Street) and Dr. Snow concluded that the water was causing the epidemic. Another example is from the tragedy of the space shuttle Challenger due to the problems in temperature of O-rings. The initial presentation was

from Morton-Thiokol engineers who opposed the launch of Challenger which eventually exploded and killed all the crew. However, they only presented a few previous flights with O-ring problems and ordered the data chronologically. The presentation was not convincing and Challenger was launched as the tragedy happened. Tufte re-exhibited the problems by showing all flights and ordered them by temperature. The results showed that the temperature was much colder than expectated in almost all flights and O-ring problems occured in all of them.

The idea of connecting graphical displays to data analysis can be dated back at least to 1962, when John Tukey published his paper, The Future of Data Analysis (Tukey, 1962). He pointed out that a simple graph can bring more information than other tools and it is capable of showing information quantitatively and revealing unexpected aspects of the data. He also suggested that graphing is not necessarily a tool to avoid substantial computation, but rather the result of computational effort.

Additionally, Tukey endorsed the idea with the same name as his book, Exploratory Data Analysis (Tukey, 1977). The idea is based on visualization instead of mathematical data analysis. The methodology of data exploration can be divided into two steps. The first step is to use appropriate tools to investigate the data and the next step is to confirm the validity and merits of the data. He also suggested that graphing can help store quantitative data, guide to conclusions, and discover further information. However, different types of graphs are suitable for different applications as he introduced the then-new categories of graphs for particular data.

Computer technology has advanced since Tukey's publications. Data visualization now not only includes data analysis methodology, but also visualization in scientific

computing and information visualization. Visualization in scientific computing focuses on scientific and engineering applications with huge amounts of numerical data. It requires a powerful computer-based system to manage more complex numerical models. Information visualization has attracted a lot of interest recently due to the demand for implementing growing data sets in other areas, such as finance, administration, or digital media (Post, Nielson, & Bonneau, 2003).

The development of data visualization occurs not only in a two-dimensional view, but in a three-dimensional view as well. The problem is that the display has to be on paper or a computer screen that can exhibit data in a two-dimensional plane. There have been attempts to overcome this problem by using many techniques, for example, perspective displays and stereo illustrations. Another method to display data two-dimensionally is to order all the graphics in an appropriate scale and place them in one whole display (Tufte, 1990).

As far as the technology evolves, data visualization can be combined with other displays to create broader viewing experiences. An example of such combination was the development of mapping described in the review from Kraak (2002). He illustrated how maps developed from forms of cartography created by hand to modern map created by computers. Additional information was then added in the layers covering a map such as scientific visualization and geo-visualization. Since the wide availability of the internet, a map can be presented within a website, catering to greater numbers of viewers. These new techniques, such as cartograms and interactive three-dimensional views, allow more refreshing experiences for a viewer.

To produce excellent statistical graphics, the displays have to convey complex ideas with clarity, precision, and efficiency and serve the clear purposes of description, exploration, tabulation, and decoration. Furthermore, graphical displays should show the available data, and make a viewer think about the contents, not the methods, design or technology. Good presentations should avoid distorting the data, such as, presenting too many numbers in a small space, and should encourage the viewer to compare different sets of data if possible. For large data sets, the graphics should display data coherently and reveal data structure for different levels. (Tufte, 1983)

Cleveland & McGill (1984) showed the results from their experiments which suggest that a viewer gains accurate perception of information presented regarding the shape of basic patterns. The accuracy ranges downward from position and scale, to length, to slope, to angle, to area, to volume, and to the least accurate, color. This can be used in design processes of data display to enable better viewer perception. They also recommended that preserving relevant ratio when comparing data is more important than setting the magnitude of the data.

Wainer (1983) presented the basic tendency of deteriorating quality of data displays. He stated that the points to be considered are categorized into the data, the clarity of the data, and the accuracy of the data. For the clarity of the data, he stated that the graph can be cluttered if data density is high and the data-ink ratio can be determined by the amount of data

.

1.3 Objectives

The objective of this thesis is to present techniques to facilitate better perception of data from four applications as follows.

(1) Using spline functions to fit demographic data;

(2) Showing registered mortality data in Thailand using thematic maps;

(3) Showing terrorism incidence rates events in southern Thailand using dynamic maps;

(4) Showing Thailand's election data on thematic maps using local projection.