CHAPTER 4

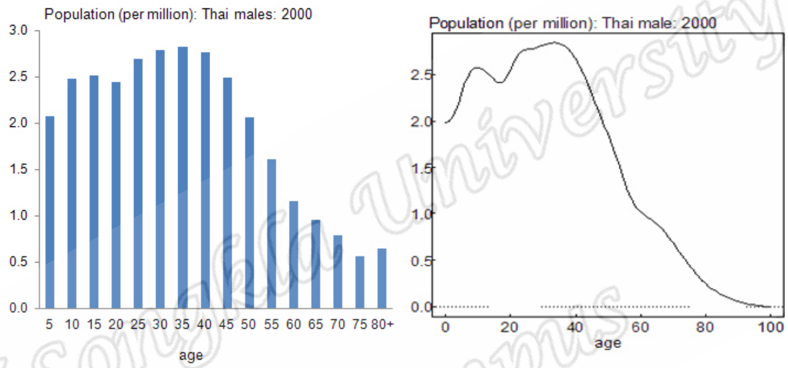
Conclusion

Based on the concept *WYSIATI* and System I of thinking, human comprehension can be induced by perception through the eyes. Humans tend to believe what they can see and understand, even though there is more convincing information than what they can observe. It seems to be difficult to change the way humans instinctively perceive, because such instinctive behaviour is what enables them to survive and they pass on this genes to successive generations. However, data presentation, especially data visualization, can play an important role in conveying more facts from available data and thus pre-empt evolution in changing human perception. This scheme is the rationale of this thesis, obtaining as much information as possible by means of presenting and visualizing data. Moreover, providing clear and accurate data visualization can expand the power of the human minds to judge information and make correct decisions and thus assist System I to respond not only rapidly but accurately as well. In this chapter, the benefits of our techniques are discussed by comparing our innovations with existing techniques and illustrating that using our techniques can produce more comprehensible information.

The first application is selected to express the data in the form of graph. In this case, demographic data are used as sources because the data are normally in an age-specific form and require data interpolation for unknown points for the whole range to be able to plot a graph. The comparison between a bar chart and our interpolating natural cubic spline curve from the same data, Thai male population distribution in 2000 is shown in Figure 4.1. It can be seen that the cubic spline interpolation approximates missing age points and thus gives more details on the trend of population in broader ranges, while a bar chart merely expresses the same information as shown in the table. This can help in decision making regarding population-related planning or actuarial planning management.

Age	*Male population in millions*	
	Cumulative	*Specific*
0	0.000	0.000
5	2.081	2.081
10	4.570	2.489
15	7.092	2.522
20	9.536	2.444
25	12.230	2.694
30	15.018	2.788
35	17.849	2.831
40	20.623	2.774
45	23.122	2.499
50	25.195	2.073
55	26.809	1.614
60	27.968	1.159
65	28.924	0.956
70	29.721	0.797
75	30.287	0.566
80+	30.936	0.649

(a)



(b) (c)

Figure 4.1: Thai male population in 2000 in (a) a tabular format, (b) a bar chart created from the data, and (c) a natural cubic spline fitting the data

Another example using the same cubic spline interpolation technique is shown in Figure 4.2. The age-specific population data in some provinces of Thailand in 2000 are used to interpolate the required age data with a natural cubic spline and added knots. The fitting population data can be seen for each province, indicating the discrepancies between population distributions. The populations go down steadily in some provinces, while there are holes and plateaus in early ages before going down later in other provinces. The improved graph immediately alerts a viewer to ask what could be the cause of the holes in the distributions around age 20. (A possibility is that the HIV-AIDS epidemic was responsible for reduced fertility among women as well as mortality.)

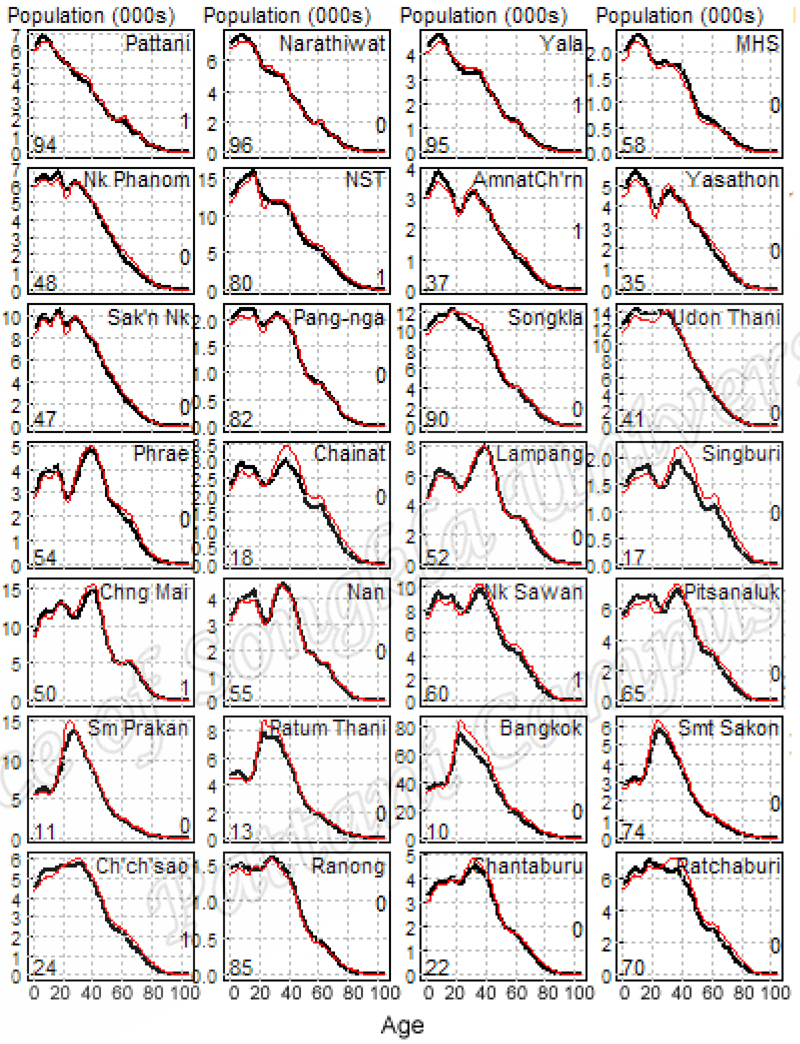


Figure 4.2: Thai populations in 2000 for some provinces, interpolated by a natural cubic spline functions with added knots

The advantages of our techniques in adding extra knots for a natural spline interpolation are shown in Figures 4.3 and 4.4. In Figure 4.3, we compare our result in interpolating age-specific fertility rate for Italian females in 1955 with the graph generated by McNeil, Trussell, & Turner (1977). The left panel of Figure 4.3 shows the quintic spline interpolation by McNeil, Trussell, & Turner (1977) (solid curve) and a single-piece polynomial interpolation (dashed curve), while the right panel displays our work. At a first sight, both spline interpolations look almost the same. However, our work follows the characteristics of a natural cubic spline where the curves outside the extremes tend to approach the horizontal axis smoothly as specified by the boundary request, while the quintic curve tends to hit the horizontal axis more abruptly.

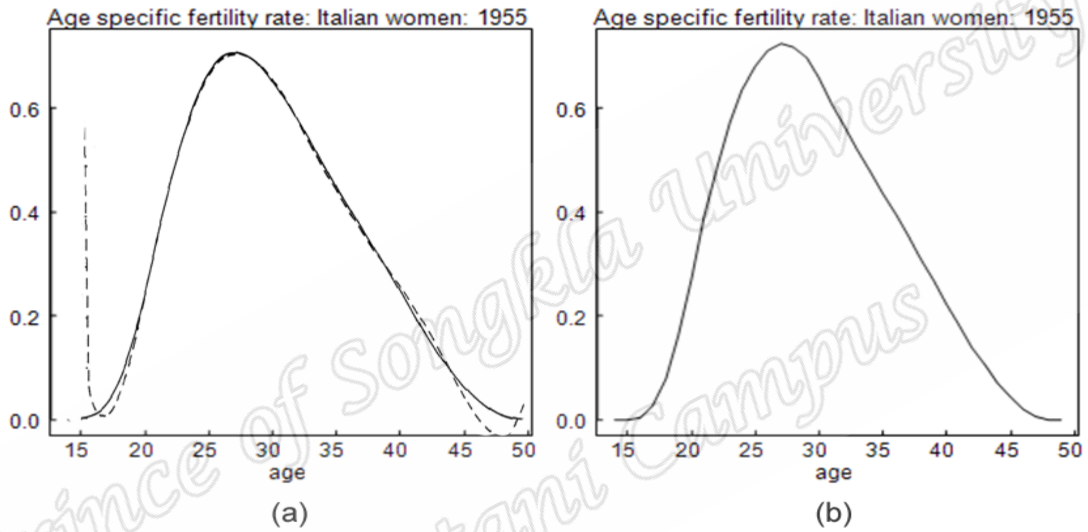


Figure 4.3: Italian fertility rate, 1955 with (a) re-production of McNeil, Trussell, & Turner (1977) in a solid curve and a single-piece polynomial in a dashed curve, and (b) a cubic spline function with extra knots

In Figure 4.4, we compare our work in interpolating Australian female mortality data in 1901 to the work of Smith, Hyndman, & Wood (2004). In this case, both interpolations are achieved by using a cubic spline function. However, while our work has extra knots to complement the interpolation, the other applies the Hyman filter (Hyman, 1983). The probability density curves obtained by differentiating the cubic spline functions, shown in Figure 4.4, clearly indicate that our work (solid dark curve) is far superior to that based on the Hyman filter (light curve) and a conventional natural cubic spline (dashed curve). A further advantage of using a cubic spline with extra knots is that the computation is much simpler.

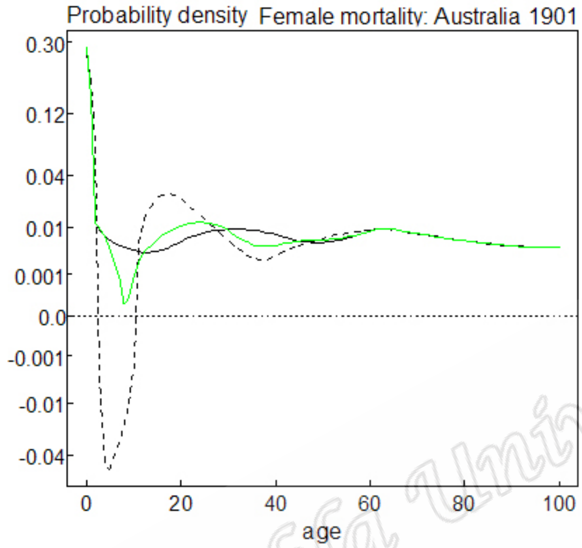


Figure 4.4: Probability density for Australian female mortality in 1901 from three techniques, a conventional natural cubic spline in a dashed curve, a natural cubic spline with Hyman filter (re-production of Smith, Hyndman, & Wood, 2004), and a natural cubic spline with an extra knot

The second and third applications illustrate that we can replace the confidence intervals with other graphical displays and provide readily digestible information for a viewer with minimal statistical background. The reason for using confidence intervals is to display the accuracy of the plotted results. However, this concept at times confuses a viewer with little statistical knowledge because the lengths of the intervals vary inversely with the accuracy, and thus the accuracy increases with the sizes of the samples. A viewer also has to compare these confidence intervals with an overall mean to see whether an interval is entirely above, crossing, or entirely below this mean and thus rank the estimated values. Besides, the data used in both applications are associated with geographical location and thus portraying them on a map is appropriate.

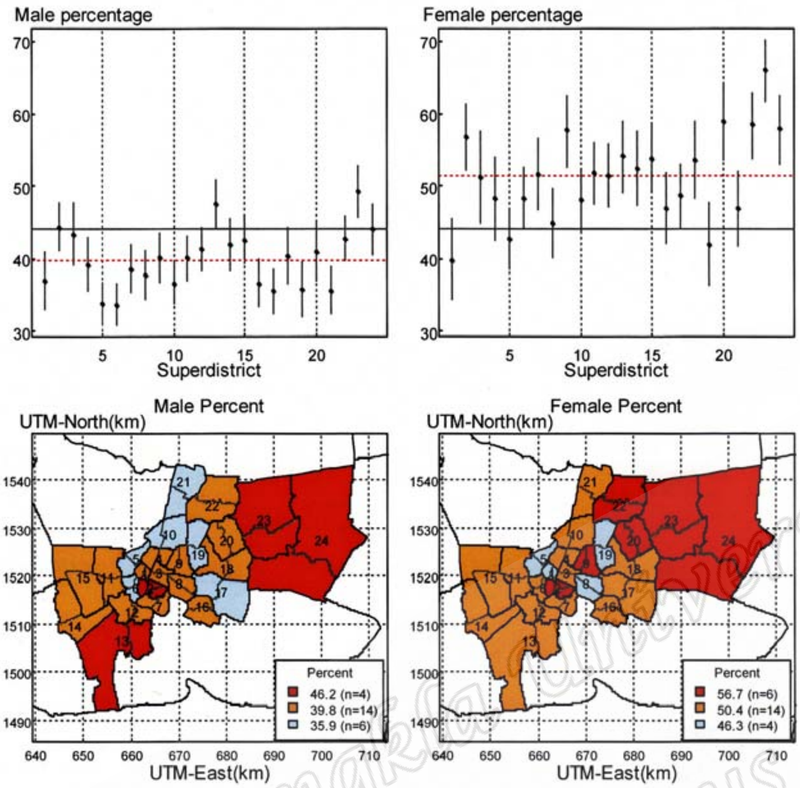


Figure 4.5: The comparison between the use of confidence intervals and thematic maps for ill-defined mortality rates in Bangkok between 1999 and 2001

In the second application, we transfer from confidence intervals as line segments to colors of regions in a thematic map as shown in Figure 4.5. Confidence intervals above, across, and below the overall mean are coded as three different colors, and these colors are filled in corresponding regions to create a thematic map. This enables a viewer to straightforwardly compare results in different regions. However, the magnitudes of the incidences themselves are not displayed because of the limitation of two-dimensional maps and there have to be two different maps for the same regions to compare the incidences between males and females. In the third application, we solve these problems by creating three-dimensional maps that show both magnitudes and comparisons of incidences in the same map. Since we employ freely available software (Google Earth), this can be achieved without complex modeling or rendering three-dimensional objects. With Google Earth, three-dimensional stacked bar charts displaying magnitudes of incidences can be placed in a corresponding region for both males and females and other graphics can also be added as a screen overlay. A viewer can change the preferred viewing angles by zooming-in, zooming-out, or rotating the viewing screen as well. As illustrated in Figure 4.6, there is much greater improvement in comprehension than using only confidence intervals and thematic maps.

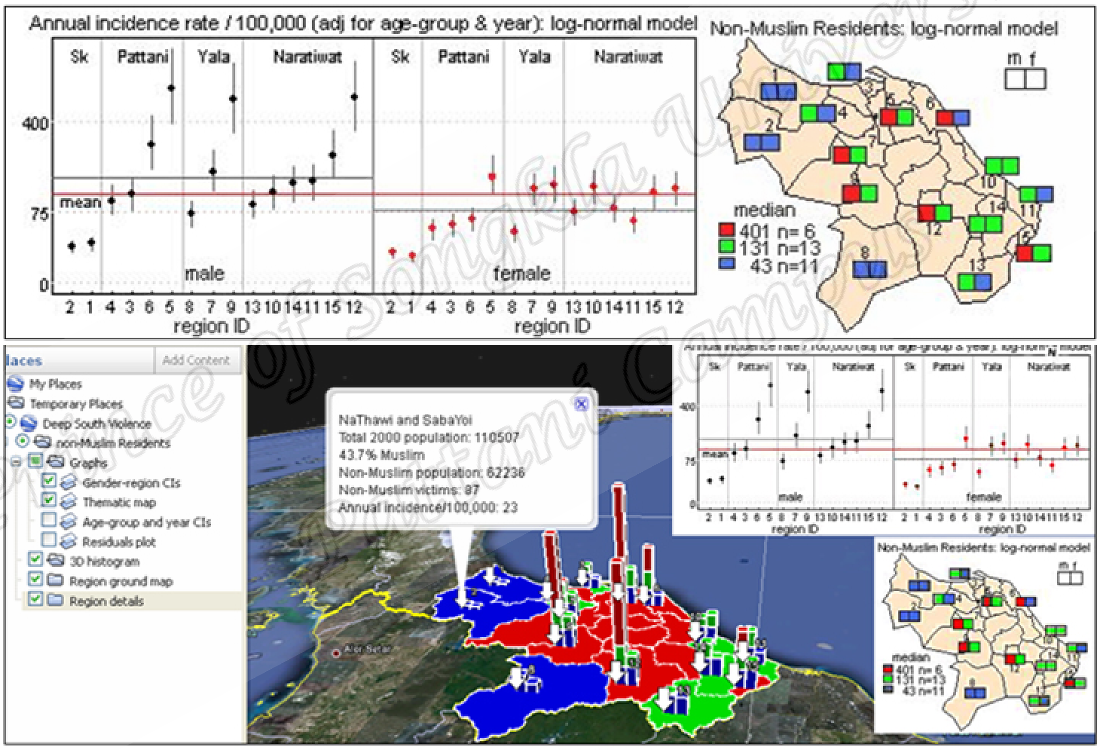


Figure 4.6: The comparison between the use of confidence intervals with thematic maps and a three-dimensional map on Google Earth with stacked bar charts and screen overlays the incidence rates of civilian injuries from terrorism events in the southern part of Thailand from 2004 to 2009

The last application illustrates that graphical displays can show both categorical and scaled data. The voting results involve both types of data as the numbers of elected candidates represent categorical data and the percentages of votes represent scaled data. For categorical data, we use thematic maps with filled colors and plotted color points to represent the party with a majority of votes in a province and the elected candidates, respectively. We also employ a variable-scaled map to increase the clarity of a focused area and distinguish otherwise dense plotted points. The benefit of using this technique over the use of a conventional inset map is shown in Figure 4.7, where our technique clearly increases the visibility of a focused area without requiring more space and thus blocking extra information like an inset map does.

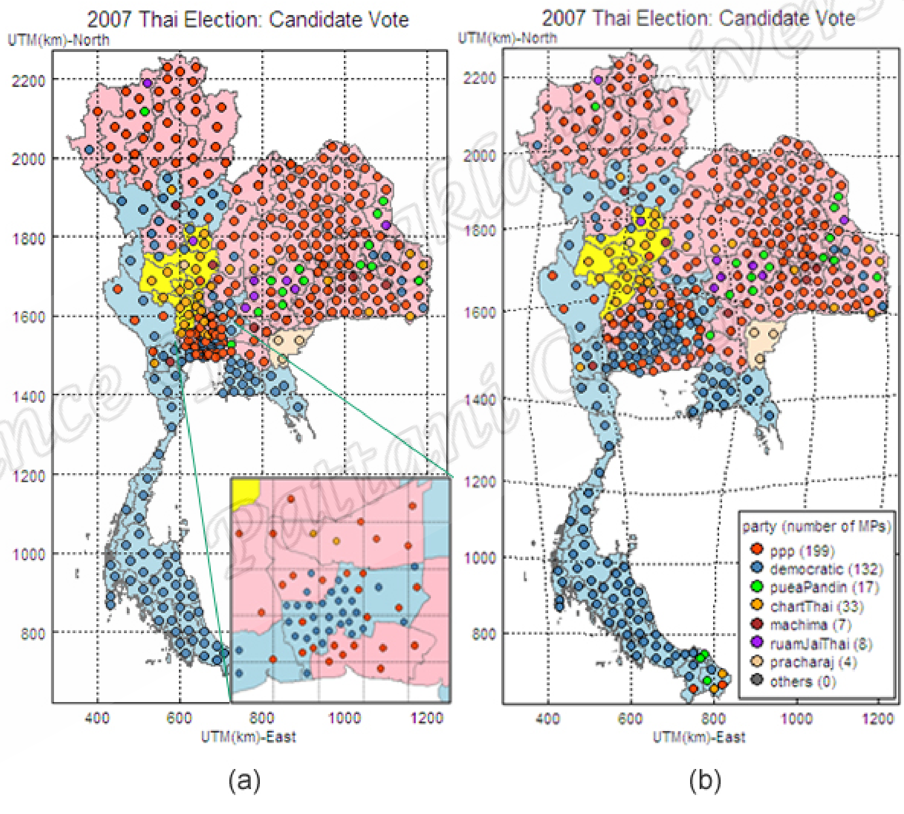


Figure 4.7: The comparison between (a) an inset map and (b) a local projection to focus on a clustered area around Bangkok for Thailand's Election in 2007

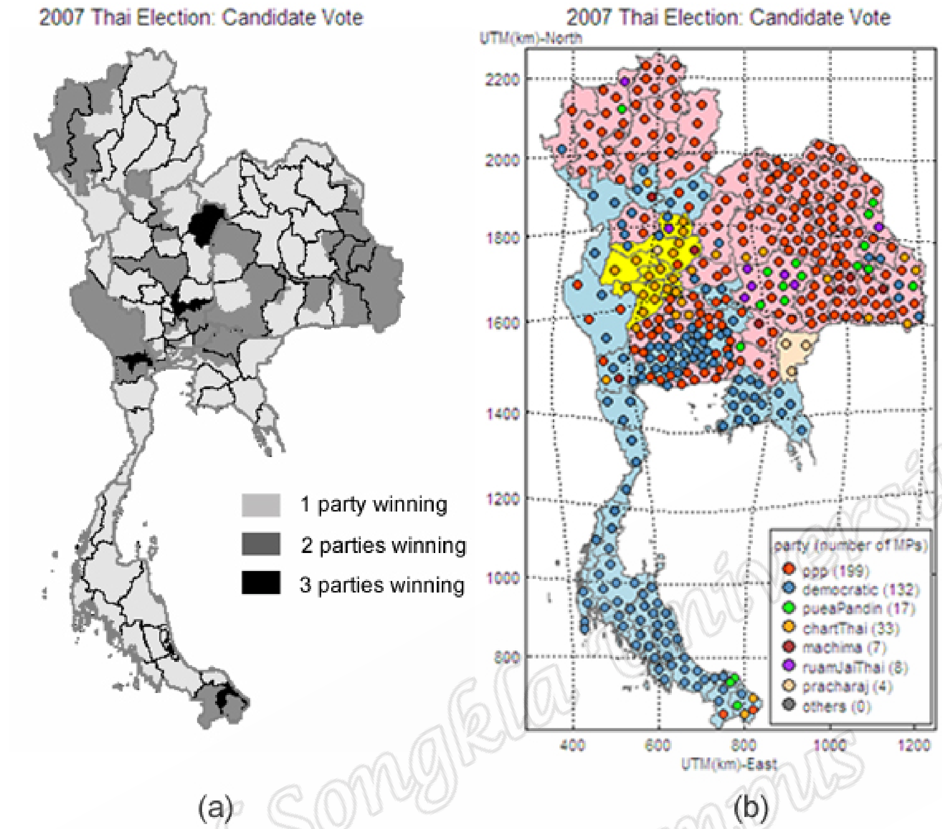


Figure 4.8: The comparison between (a) re-production of a thematic map by Thailand Political Data (2012) and (b) a thematic map using plotted color points for Thailand's Election in 2007

The use of plotted color points on a thematic map expresses more information than the use of a thematic map only. Figure 4.8 shows the comparison between the re-production of a thematic map from Thailand Political Database (2012) and the thematic map using our technique for the election result in 2007 on the left and right panels of the figure, respectively. The left panel of Figure 4.8 shows the provinces with different shades resulting from all elected candidates from one party, elected candidates from two parties, and elected candidates from three parties. This thematic map does not portray all information available from the election result as it shows merely a level of competitiveness, while our thematic map evidently shows more informative displays including the majority of the votes in each province and the parties of elected candidates (by color points). The use of different colors instead of grey shadings also guide the mind more directly to the information provided.

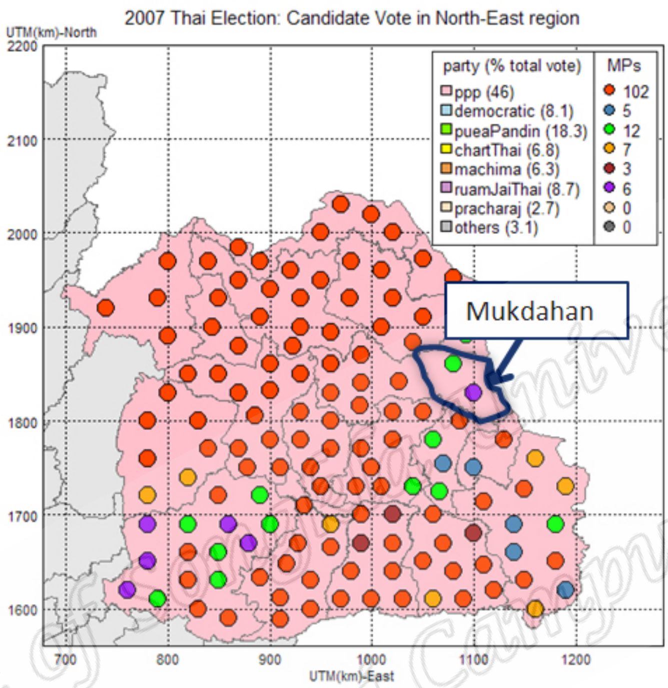


Figure 4.9: The thematic map with plotted points of the election result in the north-east region of Thailand in 2007, highlighting at Mukdahan

Moreover, the implementation of plotted points reveals interesting information, for example, Mukdahan, the province where the party with majority votes did not have any single elected candidates as shown in Figure 4.9. The analysis using our techniques clearly indicates the tendency of two major-party system in Thailand as shown in Figure 4.10.

For scaled data, we illustrate the use of bubble plots to display vote changes among parties at successive elections. Figure 4.11 indicates that bubble plots are more convenient than a table for a viewer to judge the sizes of vote swings when comparing parties.

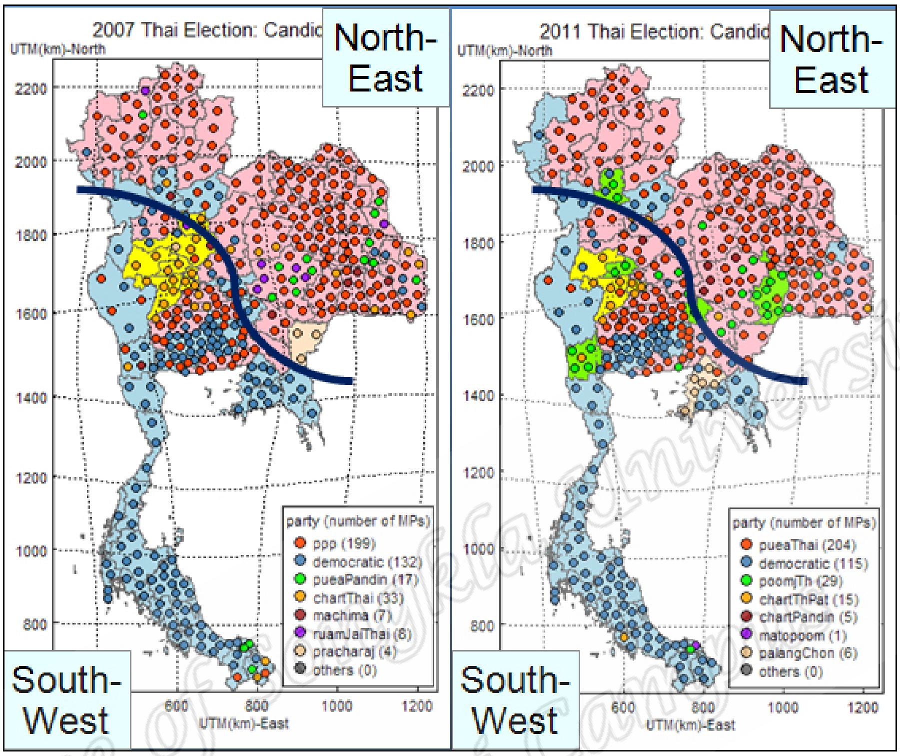
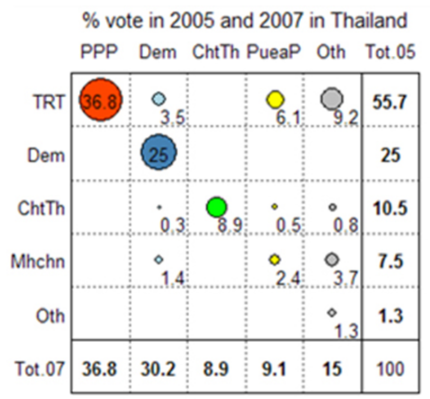


Figure 4.10: The thematic map with plotted points of the elections in 2007 and 2011 indicating the separation of political preferences between the north-east and the south-west of Thailand, and the tendency of two-party system

Parties in 2005	% of votes (2005)	Parties in 2007	% of votes (2007)
TRT	55.7	PPP	36.8
Democrats	25.0	Democrats	30.2
Chart Thai	10.5	Chart Thai	8.9
Mahachon	7.5	Puea Pandin	9.1
Others	1.3	Others	15

(a)



(b)

Figure 4.11: The comparison between (a) a table and (b) bubble plots to display vote swings from the elections from 2005 to 2007

Our innovations are proved to justify our objectives in order to create graphical displays to enhance the comprehension and conception of a viewer. It can be seen that using appropriate techniques as in our applications for displaying data can complete the task of conveying available information as much “as all there is”. Our work can complement System I of thinking stages with *WYSIATI* by means of offering available information from data and allow human minds to use this information to make a decision responsively and accurately.

Further research can involve the other thinking stage, System 2, that is slow, effortful, deliberate, and rational. The connection between graphical display and System 2 then can lead to dynamic visualizations with complex sets of data to enable logical thinking and reduce the time needed to comprehend the information. An excellent example of this type of presentation is Gapminder (Gapminder Foundation, 2006), software initiated by Hans Rosling. Gapminder exhibits a new approach of presenting statistical data by turning complex global trends from many decades into animations. The trends in life expectancy, population, education, fertility, mortality, wealth and poverty rates, among others, are available for free access. An example of using Gapminder is shown in Figure 4.12, where the achievement in Mathematics for children in the eighth grade is compared between many countries. A viewer can dynamically change the periods of displaying data, change the sizes and types of indicators including population, education, wealth and poverty, and health. This can set cross-references between two aspects of data and the information is available on both the scale of the magnitudes and the world map.

Another similar work, following Gapminder, is presented by Gesmann and Castillo (2011). Their work combines the R package as a data analysis tool to Google Visualisation API as an internet presentation tool. They create the googleVis package to connect R data frames to the displaying service from Google such as Google Map, Google Geo Map, Google Motion Chart, and Google Table. The result of googleVis is a html code with a link to JavaScript created by Google. A user can analyze data in R and then transform the result into a display on a website.

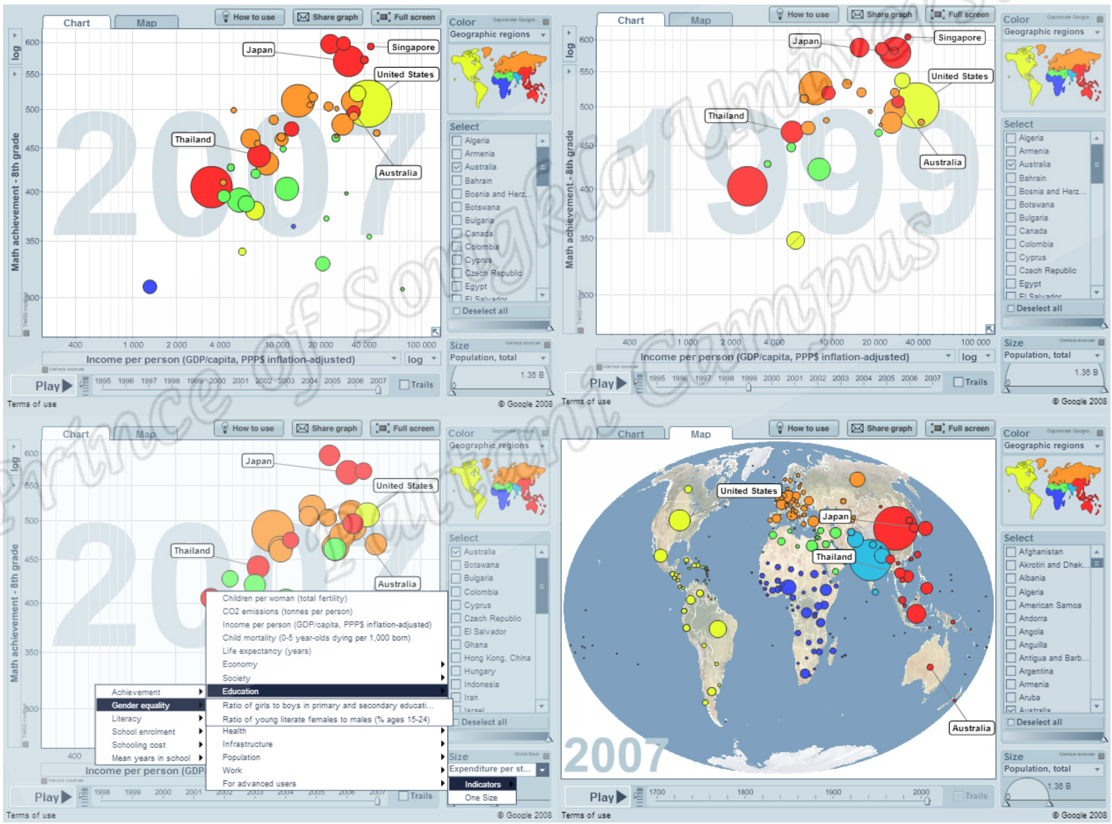


Figure 4.12: The screenshots of Gapminder show the achievement of eighth grade students in different countries. The periods of the data, the sizes of the bubbles, the variations of indicators, and the types of viewing can be changed dynamically. (Gapminder Foundation, 2006)