CHAPTER 4

Conclusion

The statement “A picture is worth a thousand words” means just only one still image can explain a lot of information. An interactive image should provide much more information than words. Presenting data for viewers is a science and an art, which means the combination of statistics and computer graphics is necessary. Statistical software both commercial and free can only produce a still graphic, such as a bar chart, bubble plot, or confidence interval plot. At present, we need a system that can produce an interactive graphic, which will be able to zoom-in, zoom-out, and rotate. The purpose of this thesis is to make a system that can produce an interactive graphical mentioned above.

The first application is to use statistical methods and graphics to analyze and fit models. Firstly, we used a bubble plot to consider the pattern of data, which includes Muslim and non-Muslim populations. From Figures 2.1 and 2.2, we found that non-Muslim citizens are exposed to more dangerous at greater than Muslims, and both Muslim and non-Muslims 25 years of age or more are risk than those younger. Also males are at higher risk than females. Additionally, rural areas have greater risk than urban areas. A bubble plot can distinguish the pattern of data easier than numerical data, because human eye can digest such information in what they see better than looking numbers. The result from the bubble plot made us focus on non-Muslims rather than Muslims. However, a bubble only tells us about pattern of the data but we need more sophisticated statistical methods and graphics to continue the analysis. At the next step, we fitted models to the data. This is the way to find less obvious information in the data. We cannot tell instantly which model is better without seeing a residuals plot. Figure 4.1 shown that the log-normal model fits better than others by looking at plots that are closes to the red line. This is one more example of the power of pictures.

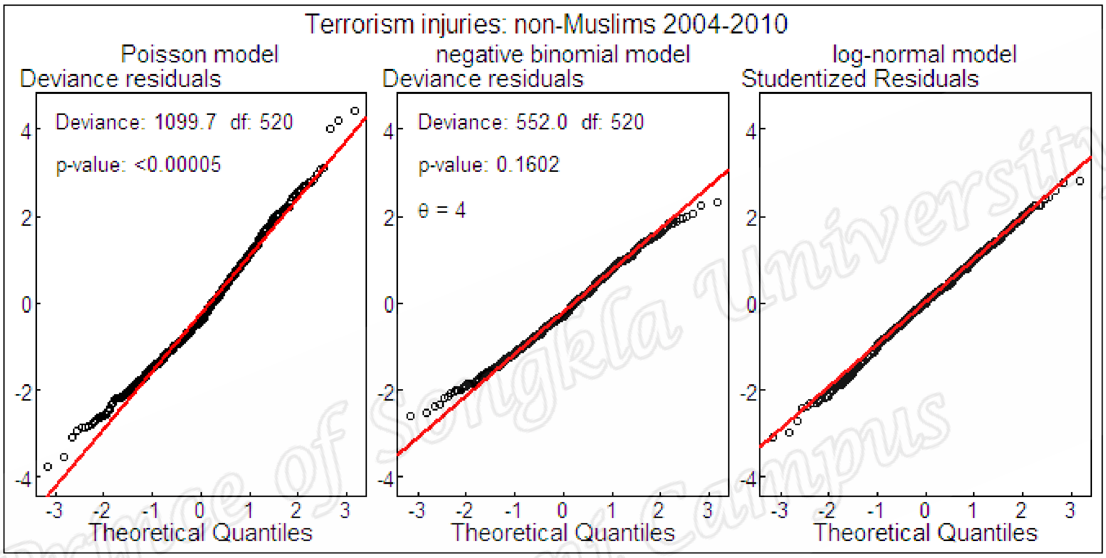

Figure 4.1: Residuals plots for Poisson, negative binomial and log-normal models.

There were three statistical models we tried to fit, Poisson, negative binomial and linear (after taking logarithm). The log-normal was the best fit for this kind of data. The Poisson model is not satisfactory because the residual deviance (1099.7) is more than double its number of degrees of freedom (520), so two models remained under consideration. Furthermore, we plotted confidence intervals comparing negative binomial and log-normal models, shown in Figure 2.3.

The result clearly shows that the log-normal model give higher estimates than the negative binomial model when the rates themselves are lower.

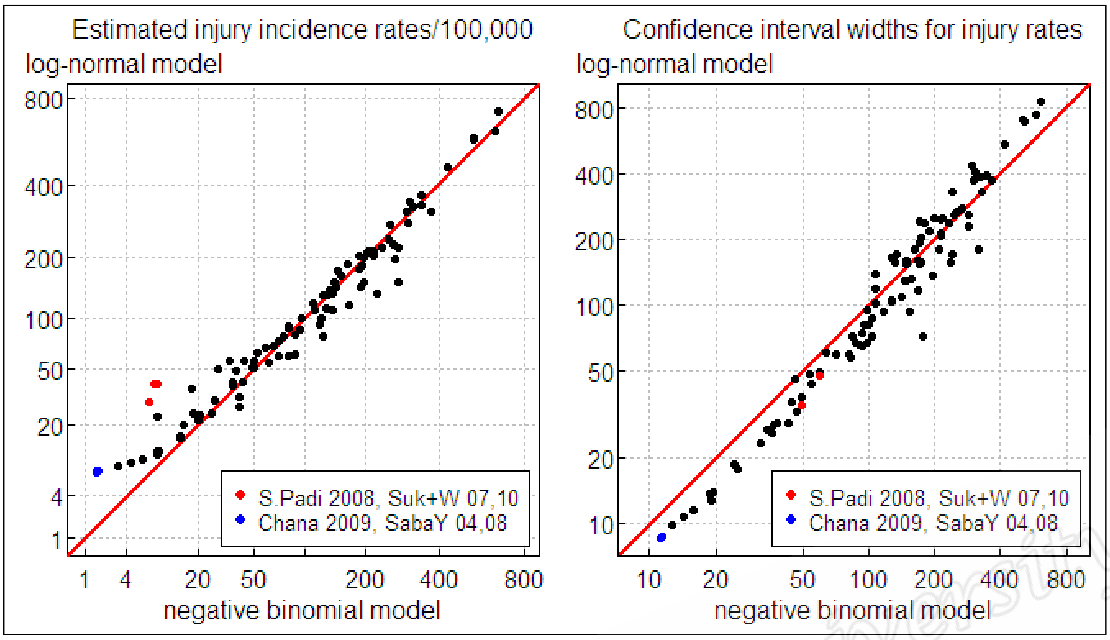

Figure 4.2: Plots of incidence rates comparing negative binomial and log-normal models (left panel) and corresponding confidence interval widths (right panel).

Figure 4.2 shows confidence interval plots comparing negative binomial against log-normal. The left panel shows estimates of incidence rates per 100,000 residents, and the right panel is the difference of two model (widths) for injury rates.

As we mentioned above, graphs are very important for displaying information, without it our life would be complex. Bubble plots give basic information such as the pattern of the data, which is very important to determine before further analysis. Residual plots tell us whether the model is acceptable fit. Confidence interval plots show the variation in injury risks. Statistical methods need those plots for showing complete information of the analysis results.

The first application focused on using both statistical and graphical methods. In the second application, we used the statistical methods as a basis for data analysis and then we developed the system with Google Earth employing KML and R programming to create an interactive GIS system. The system can show a bar chart over three-dimension space, and overlay the statistical graphic over the map.

A regression model for incidence rates was used to fit the data, but there were problems with the data set because some cells are zeroes. Zeros cannot be log-transformed. But we replace zero cells with 0.5, the log-transformed can be undertaken.

Look at Figure 2.4, showing incidence rates for all regions. Female had lower risks than males, Naratiwath (10) City and Sukirin/Waeng (13) are exceptions. The confidence intervals tell the story about rates of risk in each area but they do not tell the real location of the area, even on the map. Therefore, the GIS thematic map has been created to solve the problem. Figure 2.5 is the production of the GIS thematic that transforms the confidence intervals from Figure 2.4.

If we look at Figure 2.4, the statistician will understand but not for people who have no statistics knowledge. However, for Figure 2.5, all people can understand because it is a general image rather than a statistical graphic. The map represents the real location, and coloured boxes denotes the situation with respect to region and gender. Therefore, this is the system that helps viewers to get more information. Suppose, for example, if we distribute this thematic map to a tourist who wants to travel to the three southern-most provinces of Thailand. The tourist will see exactly where not to go, such as regions with coloured red, or they may go there extremely carefully. However, it would be better to colour the area with different colours corresponding to the confidence intervals in Figure 2.4, but the problem is that we cannot colour areas for the two sexes on the same map, otherwise, we have to create two maps with different sexes. This is the limitation of two-dimension maps. To solve this problem, the three-dimensional thematic map should be used.

We were searching for a way to create the three-dimension maps, and finally, we found that Google Earth is the best way to develop them because Google Earth is freely available, at least for the basic version. It has a pro version but we have to pay money. So we restricted to the free version. Google Earth also has their own language that can be displayed by Google Earth, also Google Maps as well. That language is Keyhole Markup Language (KML). Google is developing KML based on the XML standard so made us easy to learn. KML has the ability to markup the place over Google Earth, and write down information for the place, make shapes over Google Earth and more. Importantly, it is very easy to understand and code. The only problem with Google Earth for our application is that it has no statistical computation ability. Luckily, the R program which is also freely available, can create a result as a text file. We can create this KML file by using R instead of writing up KML by hand. Then we just run an R script to analyze the data and KML will be automatically created, and it can be opened by Google Earth.

Figure 4.3 is a simple sample of a KML file with represent a Placemark of the Faculty of Science and Technology, Prince of Songkla University, and the Figure 4.4 shows the result of running this KML file with Google Earth.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2" xmlns:gx="http://www.google.com/kml/ext/2.2"
xmlns:kml="http://www.opengis.net/kml/2.2" xmlns:atom="http://www.w3.org/2005/Atom">
<Document>
	<name>Faculty of Science and Technology.kml</name>
	<Placemark>
		<name>Faculty of Science and Technology</name>
		<LookAt>
			<longitude>101.237167535678</longitude>
			<latitude>6.880063188398867</latitude>
			<altitude>0</altitude>
			<heading>0.2272894646173007</heading>
			<tilt>15.17857511898 8</tilt>
			<range>2341.892900631792</range>
			<gx:altitudeMode>relativeToSeaFloor</gx:altitudeMode>
		</LookAt>
		<styleUrl>#m_ylw-pushpin</styleUrl>
		<Point>
			<coordinates>101.2368610119461,6.883196160005772,0</coordinates>
		</Point>
	</Placemark>
</Document>
</kml>
```

Figure 4.3 Sample of simple KML content that places mark of the Faculty of Science and Technology, Prince of Songkla University.

This is the way to create a three-dimension graphic, which is a bar chart that stacks over the surface of Earth (over Google Earth). Figure 2.5 is the final product of this thesis. Look closely at Figure 2.5, on the left panel is detailed list of whatever we want to display on the right panel. So we use Google Earth as a base of application and then create a colour based bar chart over the area, by stacking coloured-boxes together (red for above the mean, green for crossing the mean and blue for below the mean). Also, we can fill the area with colors that refer to the thematic map in Figure 2.5. And confidence interval plots can be overlaid over the map, including thematic maps and residual plots, this makes us simpler to make comparisons. At the end of the application, here is what we have created and stack over the Google Earth map: coloured bar charts filling the area with the different colours belonging to the risks of those areas, report of the population status of each region, confidence interval plots, residual plots, and thematic maps.

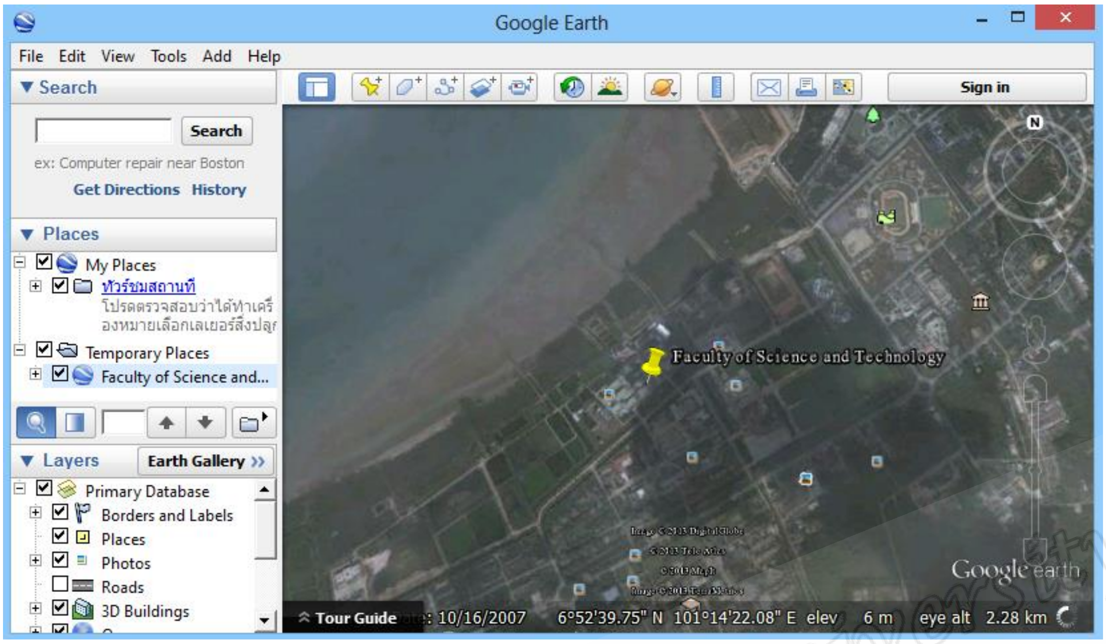

Figure 4.4: The result from KML code in Figure 4.3

Furthermore, the system provides the ability to zoom-in and zoom-out in each point that we need to focus on.

Furthermore, the KML file can be distributed over the Internet, and the web browser with Google Earth plugin can be opened as well. Therefore, our system can be used over computer networks.

Further study and development

Since we started the study, we have been thinking about presenting graphs over the web browser. HTML and JavaScript were what we thought of, because we felt in control of what we want to do, and we did some experiments. Problems were no plug-in to link R program to JavaScript and browser compatibility (some browsers display correctly, some do not). That is the reason we stuck to Google Earth and KML. At present, the JavaScript and web browsers are better, also there is a package for R, called "googleVis" (Gesmann et al 2013). The package provides the ability to connect between R program and the Google Visualization API. A favourite function is the Motion Chart (Gesmann et al 2011) with capacity of displaying a motion chart. There is software called "Gapminder" (http://www.gapminder.org/), with a screenshot show as Figure 4.5, which also has the ability to do a motion chart.

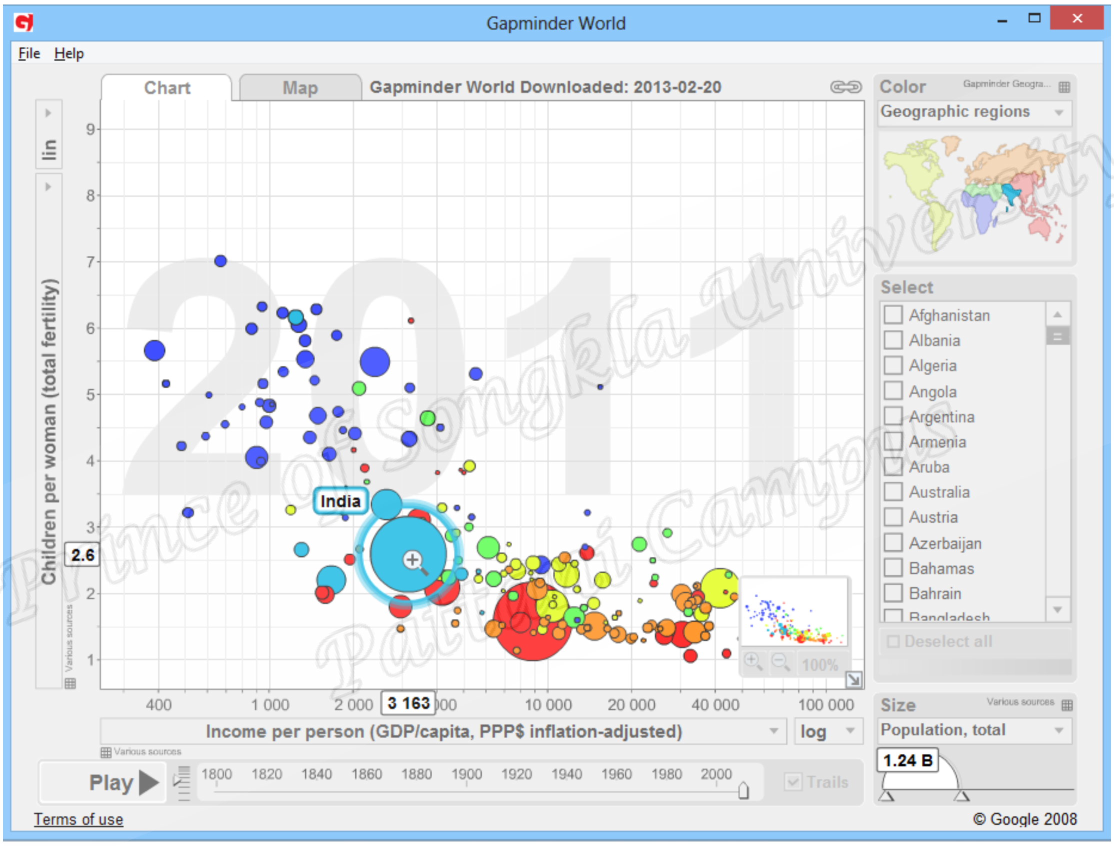

Figure 4.5 Gapminder presents the world's data.

Therefore, presenting incidence rates of injuries need newer methods for displaying information, such as the motion chart that will show a better pattern of data. There is a study (Bladh et all 2005), that addresses an animated transition navigator in 3D-maps to make users feel more comfortable. On the other hand, presenting graphs with animation like the Gapminder would be better than just still graphs. Also, if we present graphs in a tree-dimensional space, it should provide more hidden information in graphs.

Finally, there are now such packages and tools to develop new systems that can produce graphs like we mentioned above. For example, an animated bubble plot able to display a size of number of terrorism injuries in the southern Thailand from year 2004 to 2011. It will show the bubble size animation changing. That means we will notice instantly about the incidence pattern.