CHAPTER 2

Methodology

This chapter describes the studying process of two articles. The first article used the primitive statistics methods and graphics such as bubble plots, confidence intervals and residual plots to show injuries and deaths of non-Muslims in the three southern most provinces of Thailand, including 4 districts of Songkhla, and fitted models using: 1) Poisson, 2) negative binomial, and 3) log-normal (log-transformed) regression. The second application started with analysis of the data with the similar process of the first article, then tried to use a new technology to make interactive graphics with use Google Earth by writing R program script to generate KML files, then using Google Earth to interpret, display and interact with the user. Two applications used the same set of data obtained from DSCC. For years 2004 - 2010, the situation in the three southern-most provinces of Thailand has caused a lot people death and injuries. DSCC has been established to collect those data and help families of victims. The data are collected in case by case of incident. The following methods show data in graphical format.

2.1 Modeling Incidence Rates of Terrorism Injuries in Southern Thailand

2.1.1 Data and Variables

The target area of data are 37 districts in southern of Thailand from Pattani, Yala, and Narathiwat provinces, and four districts of Songkhla province. We aggregated some

adjoining districts together because some district has less non-Muslim populations for accurate statistical analysis. We called them regions instead of districts. Thus, only 23 regions were left, with populations ranging from 54,039 to 154,634. We focused on incidence rates per 100,000 population from victims of terrorism, classified by gender of victims, the age-group, region of incidence, and year (seven years from 2004 to 2010).

For preliminary analysis of annual injury rates by region and year, Figure 2.2 shows different patterns of Muslim and non-Muslim residents, which makes it difficult to accommodate the entire data set using a single model. Furthermore, the incidence rates of non-Muslim and Muslims are more than double. This is the reason that we concentrated on non-Muslims.

But since some of the 23 regions contain quite small non-Muslim populations, which was not sufficient population to make valid statistical estimates of relative risks, we then further combined some adjoining region into 15 regions, at the end. Table 2.1 shows the final aggregated regions.

Province	Region ID: Districts	non-Muslim	Muslim	Population
Songkhla	1: Chana / Thepha	62,621	94,178	156,799
	2: SabaYoi / Na Thawi	62,236	48,271	110,507
Pattani	3: Mueang Pattani	41,122	67,149	108,271
	4: Kok Pho / Mae Lan	34,812	40,816	75,628
	5: Nong Chik / Thung Yang Daeng / Mayo / Kapho / Yaring / Yarang	19,878	284,020	303,898
	6: Panare / SaiBuri / Mai Kaen	19,717	88,471	108,188
Yala	7: Mueang Yala	75,291	79,343	154,634
	8: Betong / Than To	36,706	31,487	68,193
	9: BannangSata / Krong Pinang / Yaha / Kabang / Raman	17,845	174,865	192,710
Narathiwat	10: Mueang Narathiwat	31,950	72,665	104,615
	11: TakBai	15,376	45,781	61,157
	12: Bacho / Yi-ngo / Rueso / Ra-ngae / Si Sakhon / Chanae	23,560	255,362	278,922
	13: Sukhirin / Waeng	11,624	52,141	63,765
	14: Su-ngaiPadi / Cho-airong	13,563	75,688	89,251
	15: Su-ngaiKolok	23,323	41,317	64,640

Table 2.1 Regions used for analysis of injuries from terrorism in southern Thailand

2.1.2 Statistical Methods

We used Poisson, negative binomial, and linear (based on log-transformed data) regression models. Using four factors: gender, age-grouped with three levels (under 25 which are school ages, and 25-44 which are working ages, and 45 and over which are

seniors), year (7 levels from 2004 – 2010), and region (15 levels from Table2.1). From here, the linear model for log-transformed data is simply called the "log-normal" model. The formula of the Poisson and negative binomial models are based on two sets of predictors, which are gender and age, and year and region. Following is the formula.

$$\log(\lambda/P) = \text{factor(gender.AgeGrp)} + \text{factor(year.Region)} \qquad (2.1.2.1)$$

In this formula λ is the mean of the incidence rate and P is the person-years of exposure risk per 100,000 resident population for each data cell. The Poisson model has only the single parameter λ, and the variance is also λ. The negative binomial model generalizes the Poisson model through an additional dispersion parameter θ. It has the same mean as the Poisson distribution, but its variance is $\lambda\ (1+\lambda/\theta)$. The formulation for the log-normal model is

$$\log(y) = \text{factor(gender.AgeGrp)} + \text{factor(year.Region)} \qquad (2.1.2.2)$$

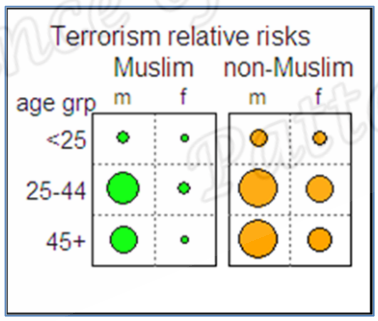
Here $y = \log(n^{+}/P)$, and n^{+} is defined as the number of injuries in a cell if this number is positive, and 0.5 otherwise. When computing adjusted incidence rates based on the negative binomial and log-normal models, a further correction is needed to ensure that each model gives adjusted incidence rates that preserve the total number of injuries observed.

2.1.3 Statistical Graphics

We used a bubble plot to show a risk for Muslim and non-Muslim persons per 100,000 population, and also confirmed between males and females shown in Figure 2.1. The bubble plot showed 25 years old and over has the most risk of injuries, particularly,

non-Muslims, who are mostly Buddhist. In term of data visualization, if we compare the two components in Figure 2.1 below, the bubble plot tells the information that both Muslim and non-Muslim males 25 years and older have higher risks, and that males have higher risks than females. This is because the human eye can distinguish an image more easily than the numeric value.

The bubble plot in Figure 2.2 shows a lot more information. For example, Pattani and Narithiwat rural area are dangerous almost every year but thus it is not so for Yala rural areas, and year 2007 seems to have been the most dangerous for both Muslim and non-Muslim residents. Furthermore, in TakBai, year 2004, you can clearly see the incident that sparked the terrorism incidence, where 286 Muslim protestors were killed or injured.



Terrorism relative risks

Age group	Muslim		non-Muslim	
	m	f	m	f
<25	27	8	49	31
25-44	157	24	229	110
45+	109	12	225	89

Figure 2.1 Comparison of risks of injury from terrorism in southern Thailand for Muslim and non-Muslim residents by gender and age group

The bubble plot in Figure 2.2 shows a lot more information. For example, Pattani and Narithiwat rural area are dangerous almost every year but thus it is not so for Yala rural areas, and year 2007 seems to have been the most dangerous for both Muslim and non-Muslim residents. Furthermore, in TakBai, year 2004, you can clearly see the incident

that sparked the terrorism incidence, where 286 Muslim protestors were killed or injured.

We used a confidence interval plot to compare between models shown in Figure 2.3, which shows adjusted incidence rates with 95% confidence intervals based on the negative binomial model, with corresponding adjusted incidence rates for comparing the region-year effects based on the log-normal model superimposed.

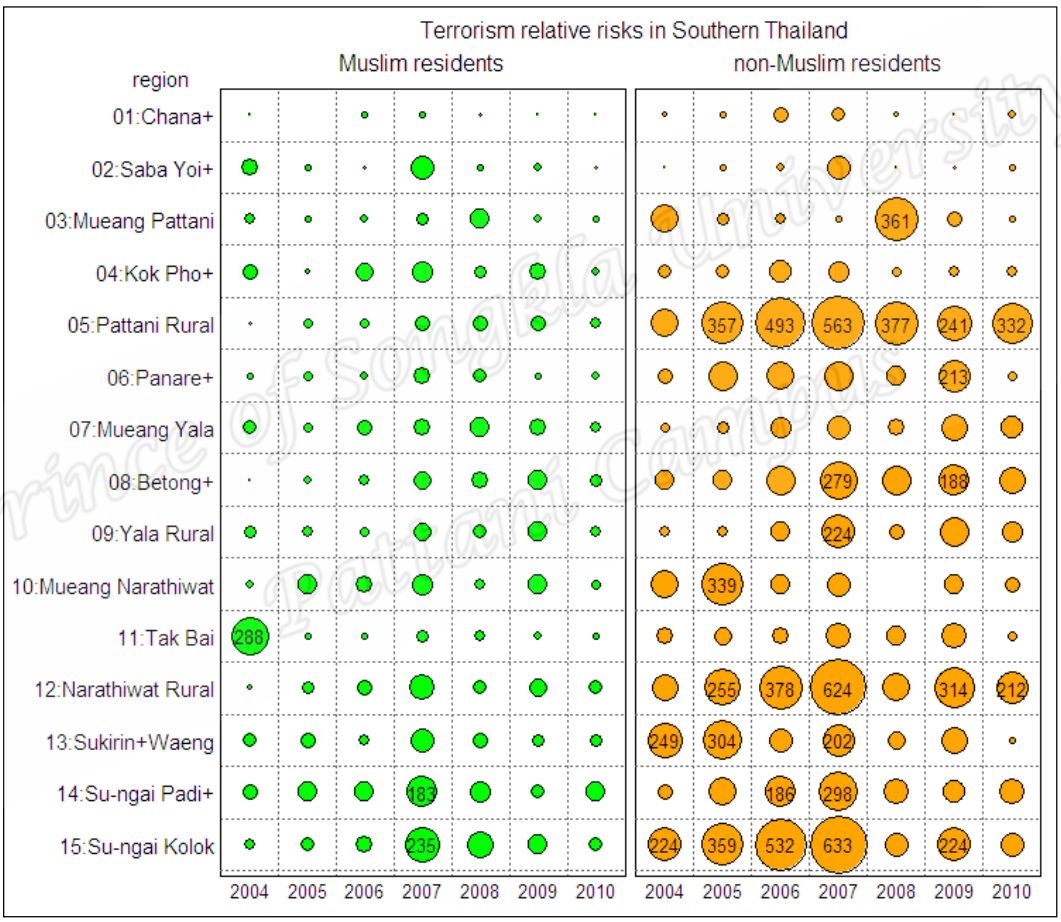


Figure 2.2 Bubble chart comparing risks of injury from terrorism in Southern Thailand for Muslim and non-Muslim residents by region and year

The adjusted incidence rates with 95% Confidence intervals from Figure 2.3 is the comparison of the negative binomial and zero-corrected log-normal models with

corresponding adjusted incidence rates for comparing the region-year effects based on the log-normal model superimposed. The log-normal model gives higher estimates than the negative binomial model when the rates themselves are lower.

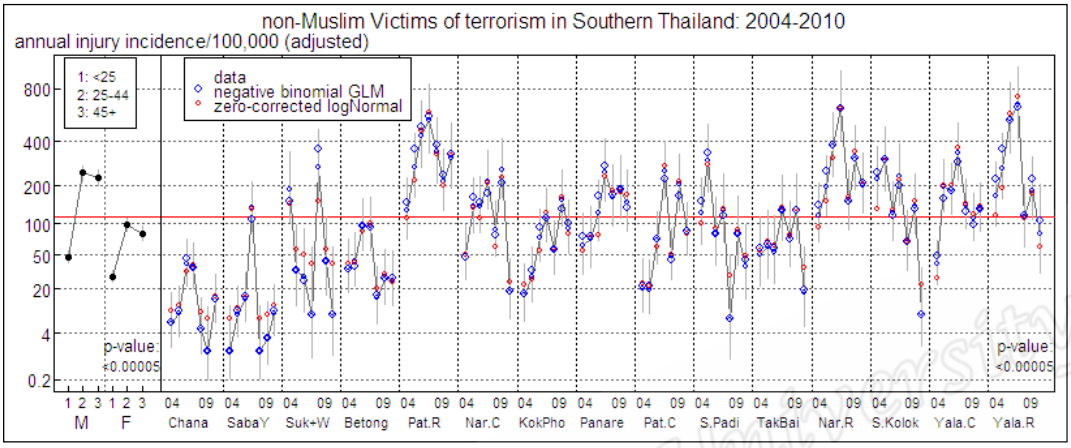


Figure 2.3: Adjusted incidence rates with confidence intervals from negative binomial model

Figure 2.4 compares these rates for the two models. There were some outliers due to the lower incidence rates given by the log-normal model (left panel).

The two models give different confidence intervals. Since the design is balanced, the log-normal model gives confidence intervals of equal size when plotted on a log scale. However, for the negative binomial model the confidence intervals vary according to the number of observed events. In the right panel of Figure 2.4 confidence intervals for the log-normal model cover a wider range than those for the negative binomial model, but most of them are shorter.

Figure 5 compares these rates for the two models. There were some outliers due to the lower incidence rates given by the log-normal model (left panel). The red dots show that region 14 (Su-ngai Padi and Cho airong) in 2008 and region 13 (Sukirin and

Waeng) in year 2007 and 2010 give a low incidence rates. Also region 1 (Chana and Thepha) in year 2008, and region 2 (Saba Yoi and Na Thawi) in year 2004 and 2008 has a low incidence rates as well.

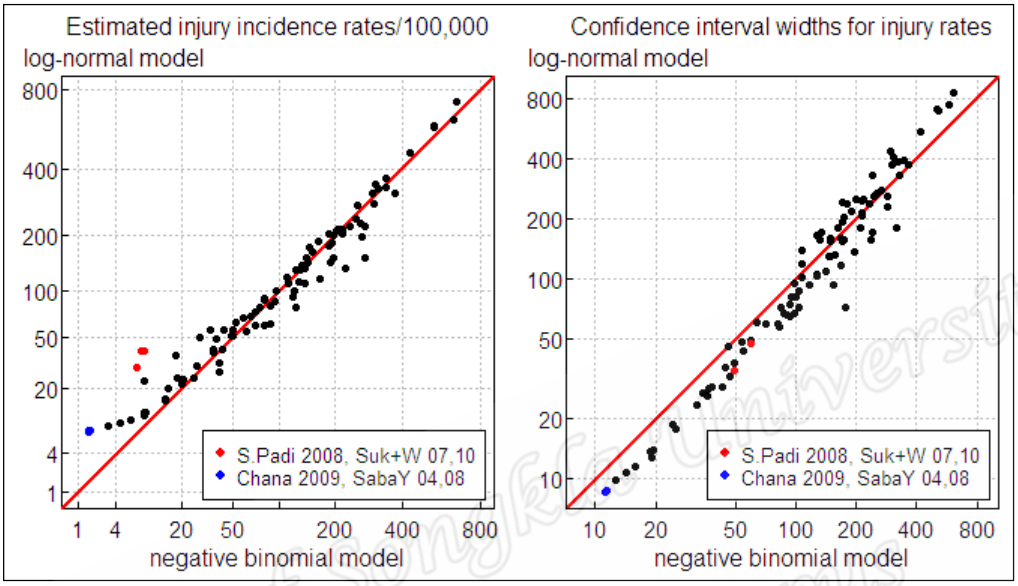


Figure 2.4: Plots of incidence rates comparing negative binomial and log-normal models (left panel) and corresponding confidence interval widths (right panel)

2.2 Graphing Incidence Rates over Regions using R and Google Earth

This article tried to develop new methods to display data in term of thematic maps and three-dimension maps, and a GIS system. As for the first article, the data were obtained from the Deep South Coordination Center as well as the first article, except that this article does not include data for year 2010 (2004 – 2009). The method is the combination of a statistical model and GIS thematic map, which is constructed by using the Keyhole Markup Language (KML) and Google Earth. KML and Google Earth are both freely available for downloadable from Google. Also an R program is used to

construct a KML file dynamically. KML is just a language that can only display data over Google Earth; it cannot compute data. This is the reason that we have to combine KML and an R program. The R program takes the responsibility to analyze data and KML displays the result over Google Earth or a Google Earth plugin over the web browser on the Internet.

2.2.1 Illustrative Data

We considered incidence rates per 100,000 population of terrorism events classified by gender, age-group (<25, 25-44, 45 or more) region of incidence, shown in Table 2.1, and year (six years). Because the overall victim incidence rates for non-Muslims were very much higher than for Muslims, we restricted the study to non-Muslim. For statistical analysis purpose, we aggregated those districts into 23 regions. However, some regions only contained less than 5,000 non-Muslims. We then aggregated again into 15 regions as shown in Table 2.1. The population of regions ranged from 54,039 to 154,634.

2.2.2 Statistical Model

A regression model was used to fit the data by using the four factors noted above. A log-transformation was used for incidence rates because the skewness and to satisfy statistical assumptions. There were 104 zero counts in 540 cells. We then replaced zeroes with 0.5 to enable log-transformed incidence rates to be calculated for all cells.

To assess the risks for subgroups of residents, we fitted an additive model with three factors: age-group, year, and gender-region, using “sum” contrasts to obtain confidence

intervals for comparing incidence rates for each level of each factor with the overall mean after adjusting for other factors (see Venables and Ripley 2002, chapter 6 for the definition of “sum” contrasts).

Figure 2.5 shows confidence intervals of annual incidence rates of male and females. Clearly see, that females are more risk than females. Also rural areas are the worst comparing to the city.

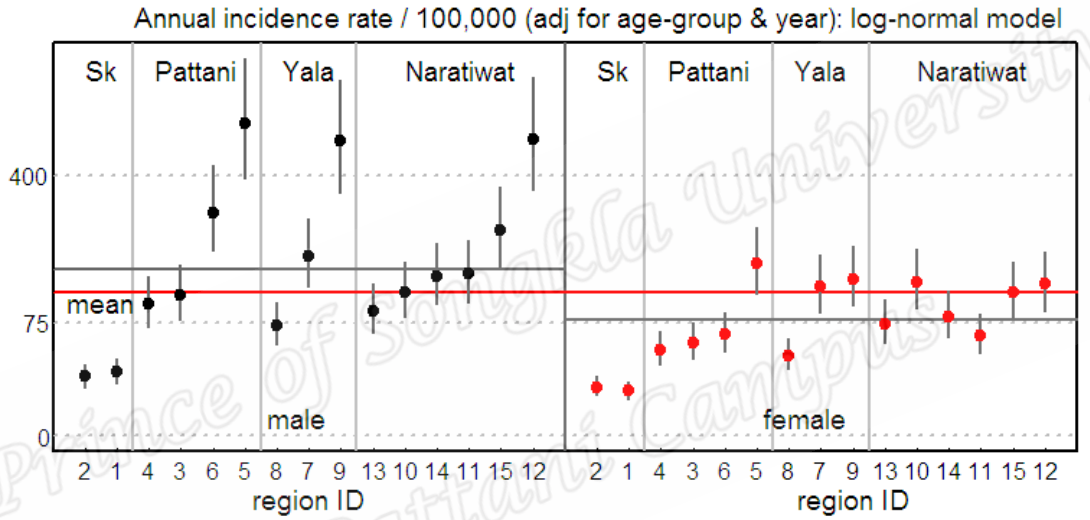


Figure 2.5: Estimates of annual incidence rates per 100,000 population of injury to civilian non-Muslim residents of the terrorism target area in Southern Thailand, classified by gender and region after adjusting for year (2004-2009) and age group (<15, 15-44, 45+). The horizontal red line denotes the overall mean incidence rate and the grey horizontal lines denote the rates for males and females, respectively. The vertical lines denote 95% confidence intervals for differences between the incidence rates and the overall mean.

2.2.3 GIS Thematic Map

Figure 2.6 is the result from a GIS thematic map constructed by the R program. The idea is to draw an overlay with coloured boxes over the surface of the real location on Google Earth. There are two boxes for each region; the left box represents the terrorism risk for males and the right box for females. There are three colours for each box. The red colour belongs to confidence intervals entirely above the mean, the blue color means the risk is entirely below the mean, and the green color means the confidence interval contains the mean.

We used polygon to create borders of the map, and filled with appropriate colours.

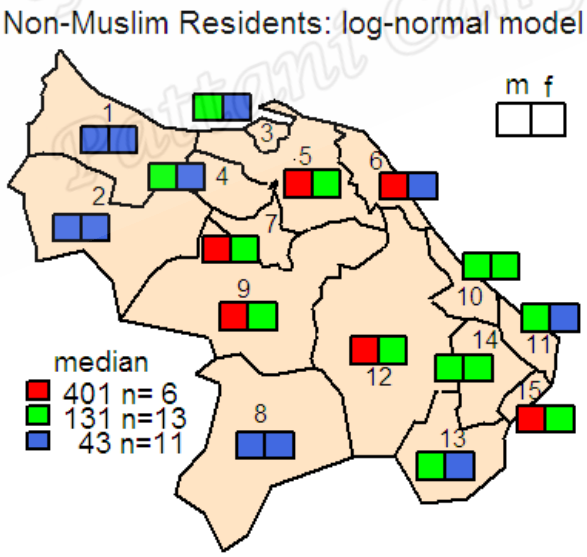


Figure 2.6: Simple thematic map of the 15 regions in the Southern Thailand terrorism target area, using three colours to classify the terrorism risk for males (left boxes) and females (right boxes) as above average (red), below average (blue) or not evidently

different from average (green). The numbers in the legend denote the median incidence rates per 100,000 populations in each of the three groups.

2.2.4 Interactive GIS Graphic System

In this section we create an interactive graphics system for displaying and controlling the position of graphs and maps. This system also creates three-dimensional histograms as shown in Figure 2.7.

The system is developed by KML and the basic Google Earth program. The KML source code is created dynamically using R (R Development Core Team 2007). We also used R to fit the statistical model and to create Figures 2.3, 2.4 and 2.5.

Keyhole Markup Language (KML) is a tag-based language based on Extensible Markup Language (XML) standard, used to display geographic data in Google Earth, Google Maps, and Google Maps for mobile phone (KML Tutorial 2013).

For our application, the R source code reads the following data tables.

(a) A contingency table of events classified by gender, age-group, region and year;
(b) a table of population denominators classified by gender, age-group and region;
(c) shape files containing the longitudes and latitudes of the boundary arcs for each region; and
(d) longitude and latitude coordinates specifying the locations within each region for placing the three-dimensional histograms.

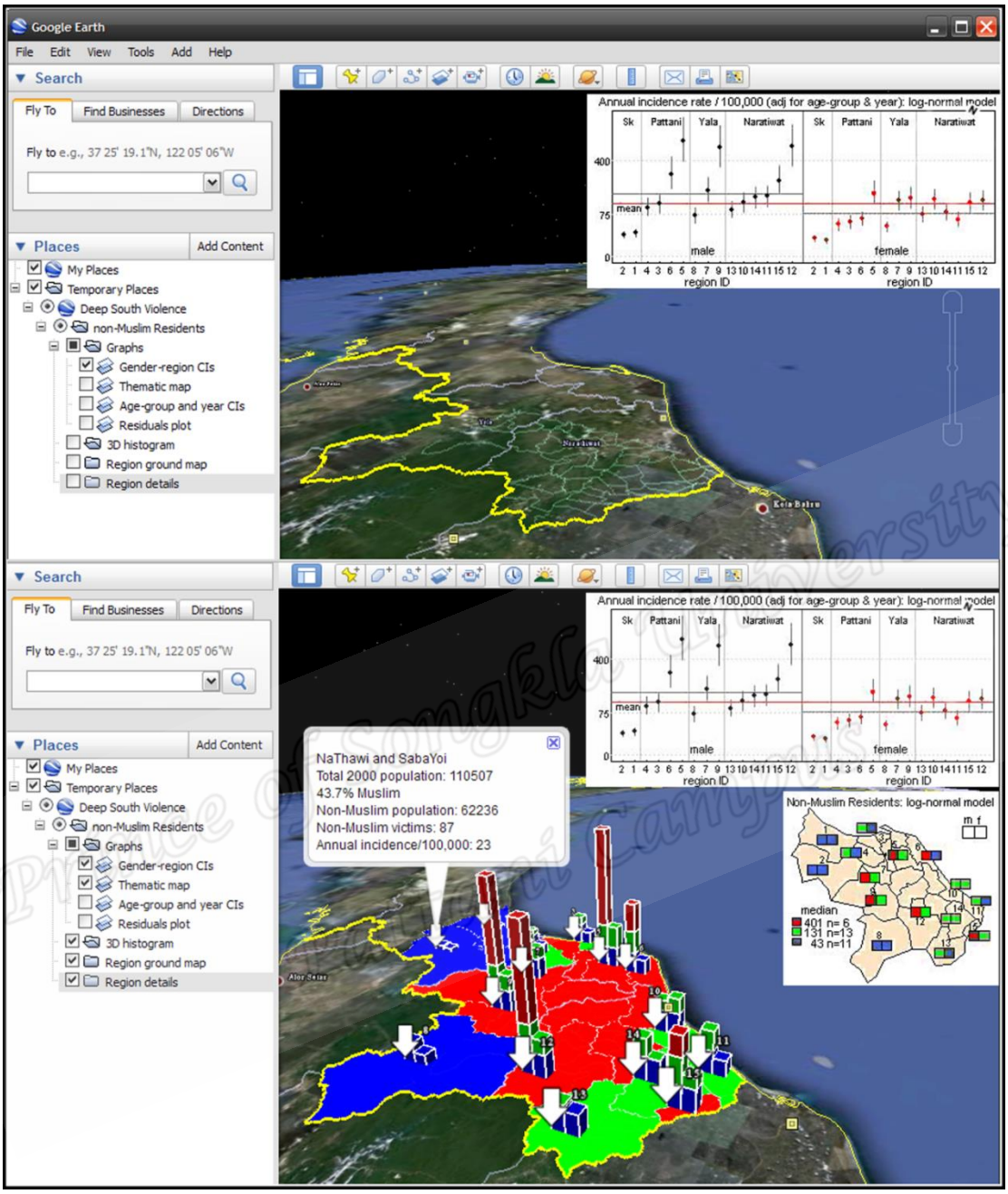


Figure 2.7: Google Earth displays of GIS map of terrorism incidence rates to non-Muslim residents in regions of Southern Thailand. The upper panel shows the location of the target area in the Malay Peninsula viewed from a specified location above the Earth's surface, with the graphical user interface in the sidebar to the left of the map that appears when a user double-clicks on the Google Earth icon, and with the graph in

Figure 2.4 appearing as a screen overlay. The lower panel shows additional features that appear when the user selects further menu options, including the simple thematic map in Figure 2.5, and corresponding colour-coded extruded polygons with ground overlays and pop-up boxes giving further details about region characteristics.

The first part of the R program fits the models, computes the confidence intervals, and creates the graphs corresponding to the confidence interval plot and thematic map shown in Figures 2.3, 2.4 and 2.5, which are stored as JPG image files.

The second part of the R program dynamically creates the KML text file, comprising header and body components.

The KML header contains style definitions, including the solid and transparent colors used in the ground-overlay maps, as well as the method for executing the rollovers from solid to transparent colors when a region or extruded polygon is cursor-selected. These extruded polygons have specified altitudes (in meters) and are necessarily anchored at ground level (or sea level or ocean floor level if preferred), but it is not possible in KML to anchor them at a specified altitude. However, you can circumvent this problem and thus create "floating" extruded polygons by putting them inside a sleeve defined as a similar polygon anchored at ground level and then coloring the sleeve with a transparent color having 100% opacity.

The body of KML consists of the code to create the 3-D histograms and the ground cover regions on Google Earth, based on the fitted model. It also requires reading the longitude and latitude coordinates of the boundary arcs of the regions, stored in a pre-processed text file.

When the resulting KML file is compiled, Google Earth will execute it immediately. The user can operate the system by selecting menu options from a sidebar and using the mouse to zoom, move, and rotate objects displayed in the Google Earth viewing window.

Details of R commands needed to fit the statistical models are given in Chapters 6 and 7 of Venables and Ripley (2002), and comprehensive details of R commands needed to create statistical graphs are described in Chapter 3 of Murrell (2006). Comprehensive information on KML can be found on Google's KML Code website (http://code.google.com/intl/th/apis/kml/).