CHAPTER 2

Methodology

2.1 Spline functions to fit demographic data

Demographic data are normally presented in a tabular form as discrete data as they are in age-specific of 5- or 10-year intervals. However, the graphical presentation of demographic data shows a rate of change of population, for example, the birth rate and the death rate. It means that the data need to be differentiated and thus require a continuous and differentiable function. To achieve this requirement, data at the non-defined ages need to be determined for the whole range by means of interpolation. A preferable interpolation for demographic data is a cubic spline function.

A spline interpolation is a piecewise polynomial function. Each function fits into a subinterval between adjacent knots, or known data points. It is designed to avoid the increase of oscillation form the use of a single high-degree polynomial interpolation over the entire interval. A cubic spline interpolation is a polynomial of degree three or less between adjacent knots (Kreyszig, 1998). Spline functions (Greville, 1969) are generally useful for smooth interpolation of data. A cubic spline with n knots $x_1 < x_2 < \cdots < x_n$ is any function $s(x)$ with continuous second derivatives comprising piecewise cubic polynomials between and beyond the knots. Denoting by $x+$ the function taking the value x for $x > 0$ and 0 elsewhere, $s(x)$ may be written as:

$$s(x) = d_0 + d_1 x + d_2 x^2 + d_3 x^3 + \sum_{i=1}^{n} c_i (x - x_i)_+^3 \qquad (2.1)$$

A natural cubic spline is a cubic spline satisfying the additional requirement that the function is linear for values of x outside the knots. This function has the property that among all functions with specified values at the knots, the natural cubic spline minimizes the integral of its squared second derivative over the interval (x_1, x_n). Since $s(x)$ is linear for $x < x_1$ if d_2 and d_3 are both 0, this requires that the cubic and quadratic terms in $s(x)$ must also disappear for $x < x_n$, so to be a natural spline the n+4 coefficients in the cubic spline must satisfy the two sets of equations.

$$d_2 = 0, \sum_{i=1}^{n} c_i = 0, \tag{2.2}$$

$$d_3 = 0, \sum_{i=1}^{n} x_i c_i = 0. \tag{2.3}$$

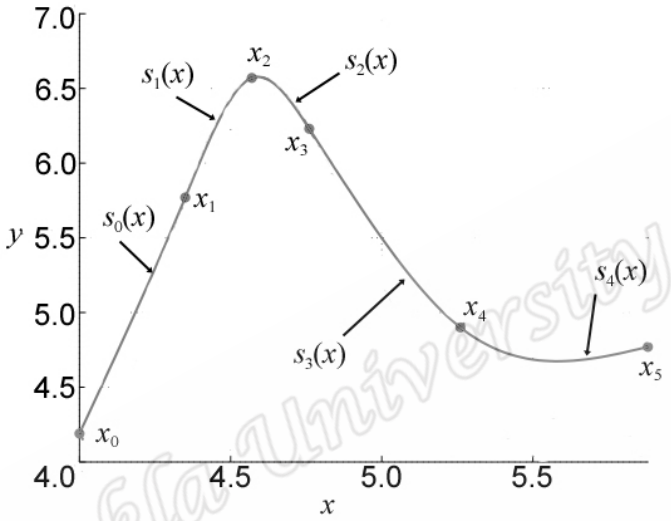
More generally, a natural spline of degree 2m+1 comprises piecewise polynomials of degree 2m+1 with continuous derivatives of order m+1 reducing to polynomials of degree m outside the knots, and has the property that among all functions with specified values at the knots, it minimizes the integral of the squared derivative of order m+1 over the interval (x_1, x_n). Equations (2.2) and (2.3) are then replaced by two sets of m+1 equations.

Rauch & Stockie (2008) illustrated an example on how to calculate the cubic spline functions and fit them to the given data as shown in Figure 2.1. The left panel of the Figure 2.1 shows the six given data points or knots, x_0, x_1, x_2, x_3, x_4, and x_5, while the right panel shows the fitted cubic spline function corresponding to the data provided. The resulting cubic spline consists of five cubic polynomials, denoted by $s_0(x)$, $s_1(x)$, $s_2(x)$, $s_3(x)$, and $s_4(x)$, interpolating data in the intervals between the given knots, such

as $s_2(x) = 6.57 + 0.79(x - 4.57) + 20.201(x - 4.57)^2 + 34.8675(x - 4.57)^3$ and $s_3(x) = 6.23 - 3.1102(x - 4.76) - 0.3266(x - 4.76)^2 + 2.4541(x - 4.67)^3$.

	x	y
x_0	4.00	4.19
x_1	4.35	5.77
x_2	4.57	6.57
x_3	4.76	6.23
x_4	5.26	4.90
x_5	5.88	4.77

(a)



(b)

Figure 2.1: The re-production of the work by Rauch & Stockie (2008) for a calculation of cubic spline function to fit given data. (a) is a given data and (b) is a result of fitted cubic spline function

In this application, we illustrated the method with three examples, (a) Italian fertility (Festy, 1970) described by McNeil, Trussell, & Turner (1977), (b) Thai male population from the 2000 population census (National Statistical Office, 2000) and (c) Australian female mortality in 1901 described by Smith, Hyndman, & Wood (2004).

In the first case, it is required that the spline function is 0 at x_1 and has first and second derivatives 0 at both x_1 and x_n to ensure that a cubic spline has continuous second derivatives. To simplify the calculation, the x-axis is shifted to accommodate the first knot at $x_1 = 0$, and two additional knots are placed at selected points a, b, with coefficients g, h, respectively. The functional form is as follows.

$$s(x) = \sum_{i=1}^{n} c_i (x - x_i)_+^3 + g(x-a)_+^3 + h(x-b)_+^3 . \quad (2.3)$$

A total of n+4 coefficients are then be determined uniquely using corresponding linear equations.

In the second case, the derivative condition is set to be 0 at only x_n, so the function is formed as follows.

$$s(x) = dx + \sum_{i=1}^{n} c_i (x - x_i)_+^3 + h(x-b)_+^3 \quad (2.4)$$

A total of n+3 coefficients are then determined in this case using corresponding linear equations.

With the mortality data, there are no end-point requirements, but one or more artificial knots may need to be included to ensure that the function is monotonic. The values of the function at the artificial knots can be estimated by trial and error.

2.2 Registered mortality data in Thailand using thematic maps

Thailand's mortality data from 1999 to 2001 are retrieved from the vital registration database for the 926 districts. We are particularly interested in ill-defined causes which are coded R00-R99 in the ICD-10 system issued by the World Health Organization (1992).

Since there are differences in population in each district of Thailand, the idea of, super-districts (aggregated adjacent districts with a combined population of

approximately 200,000) proposed by Lim & Choonpradub (2007) are applied. The mortality incidence rates are analysed for each super-district.

The adjusted proportions of ill-defined mortality by age group and super-districts are determined using a logistic regression model. A confidence interval for the proportion is constructed to show the magnitude of ill-defined mortality for each super-district.

An example of the conventional display of confidence intervals is shown in Figure 2.2. The horizontal dotted lines separate the means into three groups, according to the position of confidence interval based on weighted sum contrasts (Tongkumchum & McNeil, 2009), which is (a) entirely above the average, (b) contains the average, or (c) is entirely below the average.

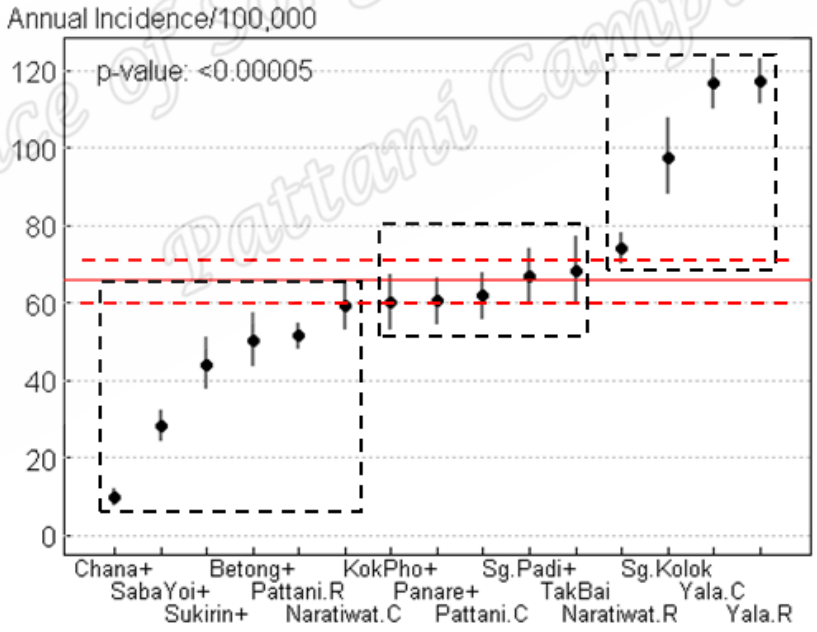


Figure 2.2: Grouping of means according to the locations of their comparison confidence intervals

We propose an alternative to comparing the incidence rates instead using a conventional bar chart. In Figure 2.3, three groups of the means in Figure 2.2 are

transferred to colored components. The components of this bar chart below the lower horizontal line are shaded blue, those between the two horizontal lines are shaded green, and those components above the upper horizontal line are shaded red. The heights of the bars denote the magnitude of the incidence rates.

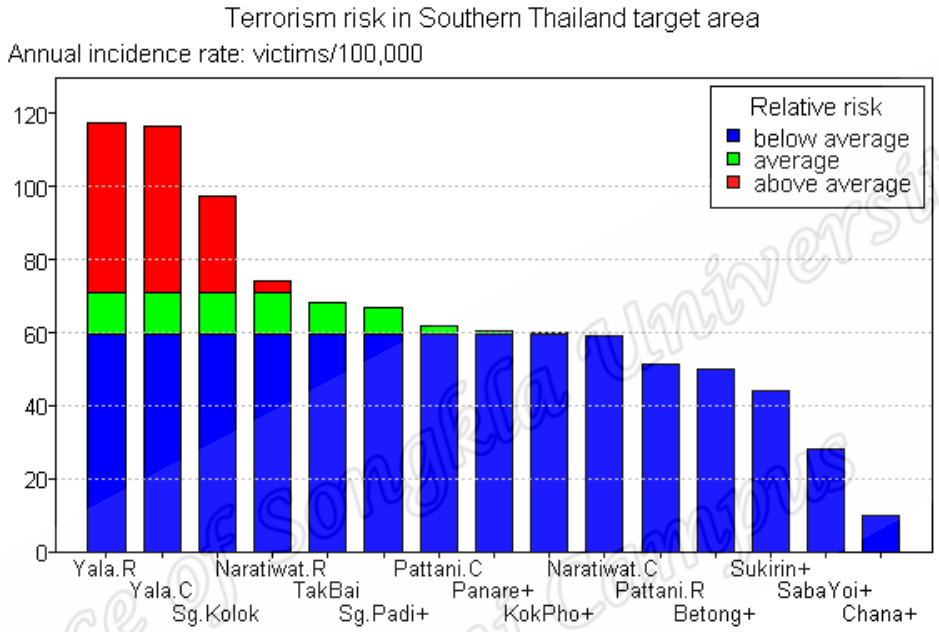


Figure 2.3: An example of a stacked bar chart for comparing confidence intervals

We apply the aforementioned idea to this application. To distinguish the differences of mortality levels for different super-districts, the color of the components associated with each super-district is used to fill the background of a corresponding super-district and create thematic map of ill-defined mortality in Bangkok and Thailand.

2.3 Terrorism incidence rates in southern Thailand using dynamic maps

We consider incidence rates per 100,000 populations of terrorism events classified by gender, age, district of residence and year. Each adverse outcome corresponds to a civilian victim suffering injury or death as a result of a defined violent terrorism event in the three southernmost provinces of Thailand (Pattani, Yala and Naratiwat) as well as four districts on the eastern side of Songkla Province. These data are retrieved from Deep South Coordination Centre, Thailand (2009). Since the overall victim incidence rates for Muslims are very much lower than those for other residents in the area, we restricted the study to non-Muslim victims. For purposes of analysis we first aggregated the 37 districts in the target area into 23 regions with populations ranging from 50,000 to 150,000.

We employ a regression model for incidence rates using gender, age group, and region. The incidence rates were log-transformed to avoid skewness and thus satisfy statistical assumptions. We then fitted an additive model with age group, year, and gender-region, again using weighted sum contrasts to obtain confidence intervals for comparing incidence rates for each level of each factor with the overall mean.

All statistical analysis and graphics are produced using the R package. For the mapping part, we are not producing three-dimensional maps directly, but using three-dimensional and dynamic map in Google Earth instead. The benefit of using Google Earth, apart from its free availability on the internet, is its function of importing graphics and adding layers of polygons and the ability to zoom in, zoom out, and rotate the view to suit a viewer's choice. The R package is then utilized to produce a KML file, a file format used in Google Earth to add all graphics from analysis and

three-dimensional bar charts of the incidence rates on the involved regions. We use the idea proposed in the previous application to show a stacked bar chart presenting the incidence rates of terrorism events. Unlike the previous application, three-dimensional displays allow three-dimensional stacked bar charts to be placed on a corresponding region instead of using coloring fill. This enables a viewer to see the magnitudes of events and can compare them among the regions dynamically. An example of different viewings of Google Earth adding polygon bars and a layer of images is shown in Figure 2.4.

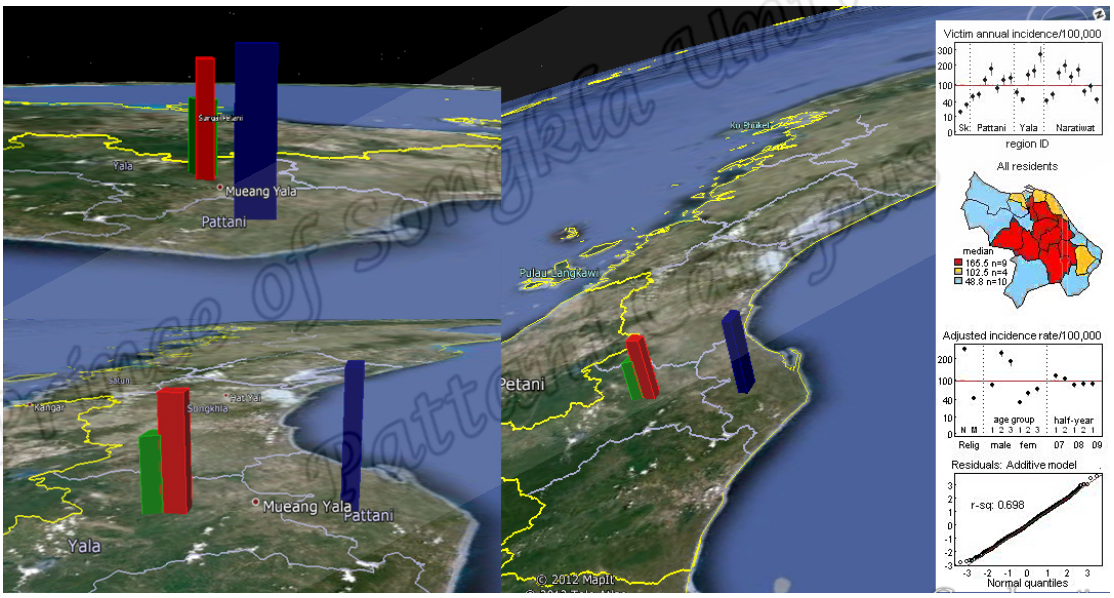


Figure 2.4: The screenshots of Google Earth with different viewing angles and added graphics such as polygon bars and images

2.4 Thailand's election data on thematic maps using local projection

The election data are retrieved from official national election results in 2005, 2007, and 2011 maintained by The Election Commission of Thailand (2011). Due to differences between electoral districts, we aggregate votes at provincial level instead.

All data are processed and all graphics are produced using the R statistical and graphical package. Graphical presentations in this presentation are divided into two parts – thematic maps and swing votes.

2.4.1 Thematic maps

To create each election result, total votes in each province are combined for every participating party. These totals are subsequently sorted and the party with the most numbers of votes in a province are selected. An individual color is assigned to all parties with elected candidates. The color of a winning party for each province is thus filled in a matching area on the map of Thailand and then points representing elected candidates are placed them to a corresponding province.

To overcome clustered points of data on Bangkok and surrounding areas in the map, we implement the idea of locally projected mapping or variable-scale mapping. It can be realized by decreasing a scale from the center to the edge of working area linearly (Fairbairn & Taylor, 1995). We use a simple method of converting Cartesian coordinates into Polar coordinates and then calculate the new scaled distances. The computed distances are then converted back to Cartesian coordinates and used to plot the map. To realize this, we generate a scaling formula by denoting x as the longitude and y as the latitude, in UTM kilometers, of a point in the region on the Earth's surface. The x_0 and y_0 are coordinates of center of Bangkok. We then convert from Cartesian coordinates to Polar coordinates to retrieve radius distance (r) and angular distance (θ) as follows:

$$r = \sqrt{(x - x_0)^2 + (y - y_0)^2} \text{ and } \theta = atan2(y, x) \tag{2.5}$$

The distance r for each map point is therefore increased to the stretched distance r_1. We define three original distance levels, a_1, a_2, and a_3, and three stretched distance levels b_1, b_2, and b_3, to convey the scaling gradually from the center of Bangkok with the distance above a_3 remains unchanged. The calculations used for scaling are as shown in (2.6).

$$r_1 = \begin{cases} (b_1 / a_1) \times r & \text{if } r \le a_1 \\ b_1 + (b_2 - b_1)/(a_2 - a_1) \times (r - a_1) & \text{if } a_1 \le r \le a_2 \\ b_2 + (b_3 - b_2)/(a_3 - a_2) \times (r - a_2) & \text{if } a_2 \le r \le a_3 \end{cases} \quad (2.6)$$

Consequently the new Cartesian coordinates are:

$$x_1 = x_0 + r_1 \cos(\theta) \text{ and } y_1 = y_0 + r_1 \sin(\theta) \quad (2.7)$$

The radial scaling for three distance levels is shown in Figure 2.5.

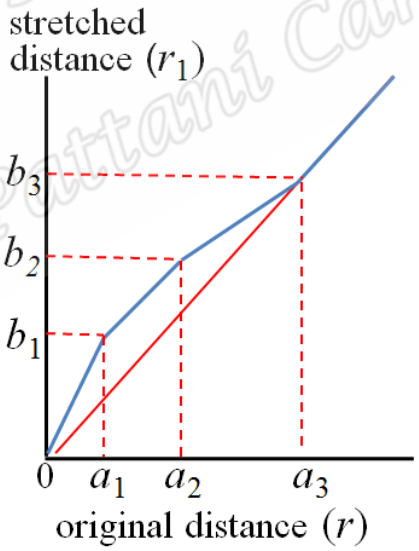


Figure 2.5: Radial scaling used in creating variable-scale map

The comparison of thematic maps with plotted points for a normal-scaled map and a variable-scaled map is presented in Figure 2.6. It shows that plotted points in the clustered area around Bangkok can be more distinguished when enlarging scales on the focused area.

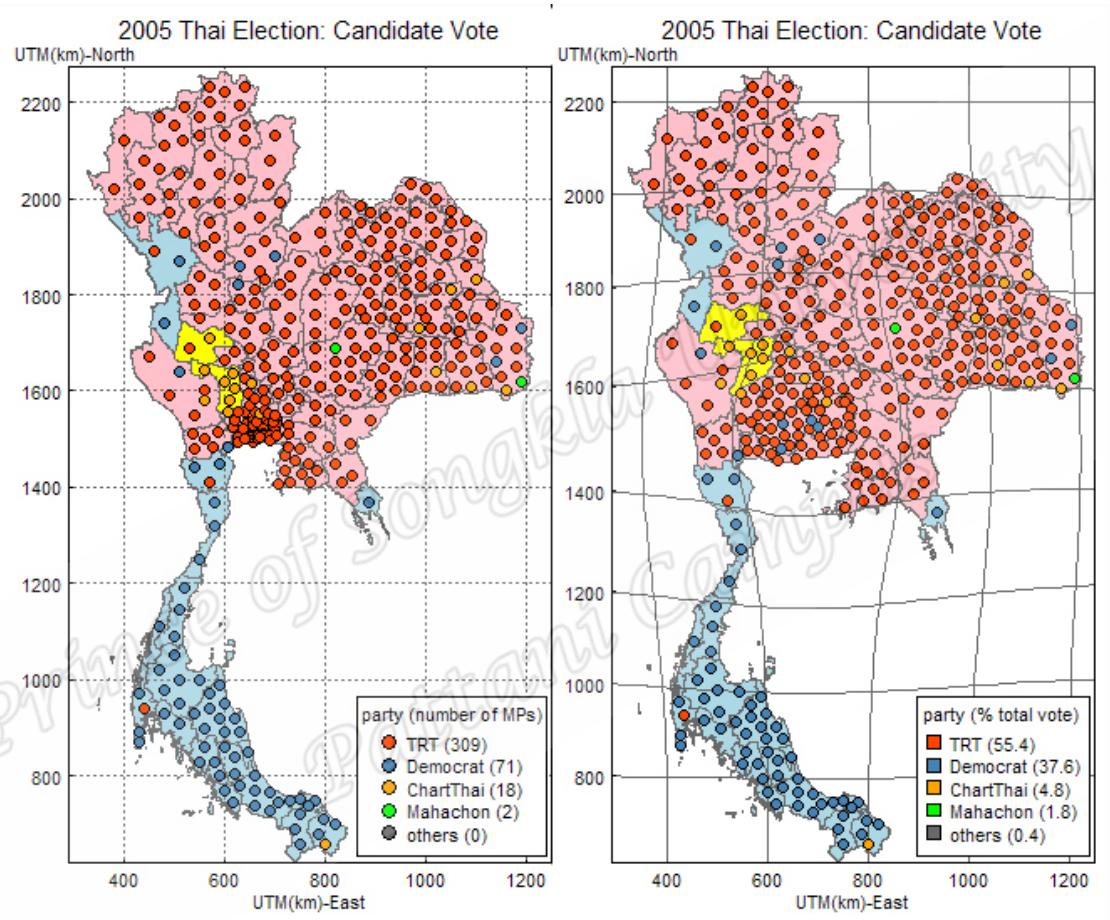


Figure 2.6: The comparison between a normal- scaled map and a variable-scaled map for voting results of Thai election in 2005

2.4.2 Vote swings

We use a simple proportional calculation to show the change of votes between each party in all three elections. The assumption is based on a concept of "one party's loss equalling other parties' gain". Thus, the excess percentage of votes gained by the

party from another in one election over the other election is proportionally distributed as follows:

$$Proportion\ gain\ of\ Party\ A = Total\ gain\ of\ party\ A\ x\ \frac{Total\ loss\ of\ party\ B}{Total\ loss\ of\ all\ parties} \quad (2.5)$$

We demonstrate an example of this calculation in Figure 2.7, showing percentage changes for each party from 2005 to 2007. In this case, the diagonal terms are the minimum values of party's percentages of candidate votes for 2005 and 2007. Democrats gained 5.2% of votes in 2007 over 2005 and the fraction of this percentage is computed from the 18.9% loss of Thai Rak Thai, later becoming People's Power, and 28% loss of all parties in 2007 over 2005. We use a similar calculation to achieve all percentage distributions. With a matrix arrangement being maintained, colored bubbles are then replaced the numerals and the size of the bubbles are compatible with the percentage of each term. Only four majority parties, in terms of percentages of votes, are illustrated, while the rest parties are included in "others".

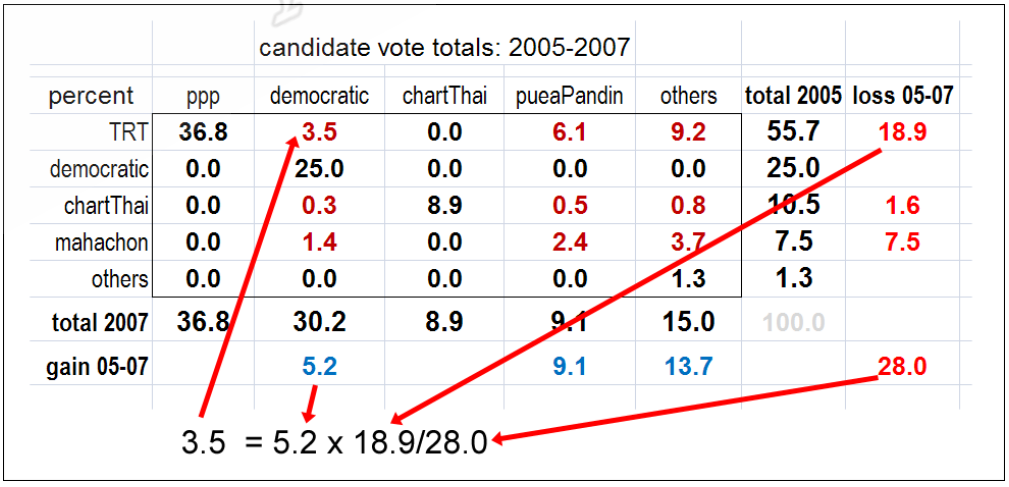


Figure 2.7: A simple calculation for the proportion of vote swings between two elections