CHAPTER 3

Manuscripts

In this chapter, we present results of the studies in this thesis by means of manuscripts published in Journal. The order of manuscripts presented is set by the date of submission to the publishers as follows:

(a) Spline Interpolation of Demographic Data Revisited, published in Songklanakarin Journal of Science and Technology, volume 33(1), page 117-120, publication year 2011.

(b) District-level Variations in the Quality of Mortality Data in Thailand, published in Asia-Pacific Population Journal, volume 25(1), page 79-91, publication year 2010.

(c) Graphing Incidence Rates over Regions using R and Google Earth, published in Journal of Map And Geography Libraries, volume 7, page 211–219, publication year 2011.

(d) Analyzing National Elections of Thailand in 2005, 2007, and 2011 – Graphical Approach, published in International Journal of Business and Social Science, volume 3(19), page 70-79, publication year 2012.

Songklanakarin J. Sci. Technol.
33 (1), 117-120, Jan. - Feb. 2011

SJST SONGKLANAKARIN JOURNAL OF SCIENCE AND TECHNOLOGY
http://www.sjst.psu.ac.th

Original Article

Spline interpolation of demographic data revisited

Nittaya McNeil[1]*, Patarapan Odton[2], and Attachai Ueranantasun[1]

[1] *Faculty of Science and Technology,*
Prince of Songkla University, Pattani Campus, Pattani, 94000 Thailand.

[2] *International Health Policy Program Thailand,*
Ministry of Public Health, Bangkok, Thailand

Received 3 February 2008; Accepted 28 February 2011

Abstract

Spline functions have been suggested in demographic research for interpolating age-specific data as they have desirable smoothness optimality properties. However, difficulties arise when boundary conditions need to be satisfied. An additional problem is that age-specific demographic data functions are necessarily non-negative, requiring the interpolating spline to be monotonic non-decreasing. In this paper we describe a simple and effective alternative that circumvents these problems. We show that natural cubic splines can be used to interpolate age-specific demographic data and ensure that relevant boundary conditions on second derivatives are satisfied, thus preserving the desirable optimality property of the interpolating function without the need to increase the degree of the spline function. The method involves incorporating one or two additional strategically placed knots with values estimated from the data. We describe how the method works for selected fertility, population, and mortality data.

Keywords: natural cubic spline, monotonic, fertility, population, mortality

1. Introduction

"Spline functions (Greville, 1969) are useful for smooth the interpolation of data". A cubic spline with n knots $x_1 < x_2 < ... < x_n$ 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

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

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 for 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 following two sets of equations

$$d_2 = 0, \sum_{i=1}^{n} c_i = 0, \quad (2)$$

$$d_3 = 0, \sum_{i=1}^{n} x_i c_i = 0. \quad (3)$$

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

* Corresponding author.
Email address: nittaya@bunga.pn.psu.ac.th

degree *m* outside the knots, and has the property that for 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 and 3 are then replaced by two sets of *m*+1 equations.

As pointed out by McNeil *et al.* (1977), splines can be used in demographic research for interpolating age-specific cumulative fertility, where data are usually available at 5-year age intervals from 15 to 50 years. However, in practice fertility increases slowly from 0 at age 15 and also decays even more slowly to 0 at the other extreme, so it is desirable to impose the additional conditions that both the derivatives and the second derivatives of the spline are 0 at the extremes. To satisfy these boundary conditions, McNeil *et al.* (1977) suggested relaxing the requirement that the spline function is natural. However, although it appears that for a cubic spline the four equations in Equation 2 and 3 could be replaced by the four conditions on the first and second derivatives at the boundaries, this cannot be done because the resulting equations do not involve c_n. So the proposed solution not only fails to satisfy the smoothness optimality condition but also must be of degree 5 or higher.

A further problem is that age-specific demographic data, including fertility functions are necessarily non-negative, requiring the interpolating spline to be monotonic non-decreasing. In a recent paper investigating the use of spline functions for another demographic application involving the interpolation of Australian mortality data, Smith *et al.* (2004) demonstrated this inadequacy of the spline function. He applied the Hyman filter (Hyman, 1983), which imposes alternative constraints on the derivatives of a cubic spline, possibly sacrificing smoothness if the filter has changed first derivatives.

In this paper we describe a simple and effective alternative that circumvents both of these problems. It uses natural cubic splines to interpolate age-specific demographic data and ensures that relevant boundary conditions on second derivatives are satisfied, thus preserving the desirable optimality property of the interpolating function without the need to increase the degree of the spline function. The method involves incorporating one or two additional strategically placed knots with values estimated from the data.

In the next section we describe how the method works for situations of interest to demographers. After that we illustrate the method with three examples, (a) Italian fertility (Festy, 1970) described by McNeil *et al.* (1977), (b) the cumulative distribution by age of males from the 2000 population census of Thailand (National Statistical Office, 2000), and (c) Australian female mortality in 1901 considered by Smith *et al.* (2004).

2. Methods

The first situation involves fitting a natural cubic spline to a cumulative age-specific fertility schedule, where it is required that the function is 0 at x_1 and has first and second derivatives 0 at both x_1 and x_n. Note that if the first derivative of the required function is 0 at x_1 and x_n the second derivative must also be 0 at these points. This is because a cubic spline has continuous second derivatives.

For simplicity we shift the *x*-axis so that the first knot is located at $x_1 = 0$, and we place two additional knots at *a*, *b*, with coefficients *g*, *h*, respectively. Since the value of both the function *s*(*x*) and its derivative at *x* = 0 is 0, the linear terms in Equation 1 vanish and its functional form is thus

$$s(x) = \sum_{i=1}^{n} c_i (x - x_i)_+^3 + g(x-a)_+^3 + h(x-b)_+^3 \,. \quad (4)$$

A total of *n*+4 coefficients need to be determined. These comprise the *n*+2 coefficients in Equation 4 together with the values of the function to be determined at *a*, *b*. The linear equations determining these coefficients comprise (a) the *n*+1 equations requiring that the function have specified values at the *n*-1 knots apart from the first where its value is necessarily 0, (b) the two equations needed to ensure that the function is linear for $x > x_n$, and (c) the requirement that the derivative is 0 at x_n.

The parameters *a*, *b* may be chosen to vary the result to satisfy other requirements including smoothness and monotonicity. These values were chosen by selecting a value in between the first and second knots and a value between the last and the second last knot. If these two knots are too close to the two boundary knots, then the value of the function can become negative. In practice, the region will have to be determined to ensure that the function will be monotone non-decreasing and smooth. If the function is symmetric then the two distances, the first knot to *a* and *b* to the last knot, should be similar. If the function is skew to the right, the distance from *b* to the last knot would be slightly more than the distance from the first knot to *a*, and vice versa for a left skew function.

The second situation involves fitting a natural cubic spline to cumulative population data. In this case it is not necessary to impose a condition that the derivative of the function is 0 at *x* = 0, but the condition is desirable at the other end. We place an additional knot at *b* with coefficient *h*. The functional form is thus

$$s(x) = dx + \sum_{i=1}^{n} c_i (x - x_i)_+^3 + h(x-b)_+^3 \quad (5)$$

A total of *n*+3 coefficients need to be determined. It comprises the *n*+2 coefficients in Equation 5 together with the value of the function to be determined at *b*. The linear equations determining these coefficients have similar conditions as the first situation.

The third situation involves fitting a natural cubic spline to mortality data. In this case there are no end-point requirements, but one or more artificial knots may need to be included to ensure that the function is monotonic. In this situation, the monotonicity condition cannot easily be met by satisfying equations involving values of the derivatives similar to boundary conditions, so the values of the function at the artificial knots are estimated by trial and error. We give an example in the next section.

3. Application

The Italian cumulative age-specific fertility data (Festy, 1970) at five-year age periods are shown in Table 1.

There are $n = 8$ knots so the method described in Section 2 involves solving 12 linear equations. To fit a natural cubic spline to these data we have chosen the knots at the ages 15 and 50 years and every 5 years in between and an additional two knots at ages 16 and 48. The values were chosen by selecting the first knot in between age 15 and 20 year, the second knot in between age 45 and 50 year. To ensure that the values of the function are not negative, for this data the first knot should be between 15.62 and 20 years, while the second should be between 45 and 48.16 years. For simplicity, we chose to round the knots to the nearest whole age. Thus, 16 and 48 were chosen as the values of the additional knots.

The fitted values of the cumulative fertility at the two extra knots are 0.00029 and 2.30491, respectively. The derivative of the fitted cubic spline is shown in the left panel of Figure 1.

The right panel of Figure 1 shows the results of fitting a natural cubic spline to the age-specific cumulative Thai male population in 2000 as shown in Table 2. Since the proportion living beyond age 100 years is very small we take this value as the upper limit. In this example there are $n = 18$ knots. To fit the natural spline to these data we need only to insert one extra knot before the maximum age bracket, at about age 85 years, to make the first and second derivative of the spline function 0 at $x_n = 100$. Using Equation 5 there are 21 unknowns together with 21 equations. Solving these equations gives the value 30.8095 for the male population below age 85 years.

For the last example, we fit a natural cubic spline to the Australian 1901 female cumulative mortality at ages 1, 5, 20, 40, 60, 65, and 100 years. The data are estimated from Figure 1 of Smith *et al.* (2004). Table 3 gives these estimates.

Table 1. Cumulative age-specific fertility data.

Age (years)	Cumulative fertility rate
15	0.000
20	0.080
25	0.593
30	1.297
35	1.840
40	2.171
45	2.296
50	2.306

Table 2. Thai male population in 2000.

Age (years)	Male population in millions
0	0.000
5	2.081
10	4.570
15	7.092
20	9.536
25	12.230
30	15.018
35	17.849
40	20.623
45	23.122
50	25.195
55	26.809
60	27.968
65	28.924
70	29.721
75	30.287
100	30.936

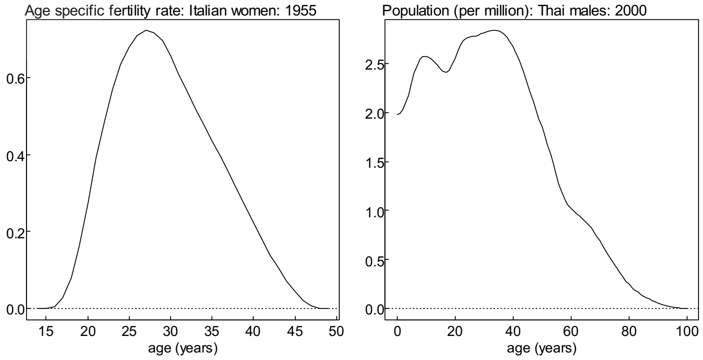


Figure 1. Natural cubic spline interpolations of fertility and population densities.

Table 3. Cumulative age-specific mortality data for Australian women in 1901.

Age (years)	Female mortality per 1000
0	0.00
1	4.75
5	6.40
20	7.75
40	11.20
60	13.85
65	14.80
100	19.15

There are 8 knots. The left panel of Figure 2 shows the natural spline fitted to the data shown in Table 3 (dotted curve), the monotonic spline interpolation given by Smith *et al.* (2004) (lighter curve), and the natural spline using the data with the extra knot (darker curve). The right panel of Figure 2 shows the corresponding density curves obtained by differentiating the cubic spline functions.

From the figure, the natural spline fluctuates in the first 20 years, which include the first five knots. The additional knot was chosen between ages 1 and 5 years and we obtained a smoother result than that obtained by Smith *et al.* (2004) by choosing the value of 2 years. The cumulative mortality at this age obtained from Figure 1 of Smith *et al.* (2004) is 5.9 deaths per 1,000. Three additional knots at ages 10, 30 and 50 years with cumulative mortality values of 6.5105, 10 and 12.4 deaths per 1,000, respectively, gave the results similar to the one given by Smith *et al.* (2004) as shown in Figure 2.

The function using this method is smooth and always non-negative for the three examples we have shown. However, there is no guarantee that the function will be smooth and non-negative for other datasets. The value of the additional knots could be solved linearly using the property of natural cubic splines where the smoothness function interpolates the data by minimizing the integral of its squared second derivative. Further investigations will be needed on these issues.

Acknowledgement

We are grateful to Emeritus Prof Don McNeil for his helpful suggestions.

References

Festy, P. 1970. Evalution de la Fécondité en Europe Occidental Depuis la Guerre. Population. 25, 229-274.

Greville, T.N.E. 1969. Introduction to Spline Functions. In Theory and Applications to Spline Functions. T.N.E. Grevillea, editor. Academic Press, New York, U.S.A., pp 1-35.

Hyman, J.M. 1983. Accurate monotonicity preserving cubic interpolation, SIAM Journal on Scientific Computing. 4(4), 645-654.

McNeil, D.R., Trussell, T.J. and Turner, J.C. 1977. Spline Interpolation of Demographic Data. Demography. 14(2), 245-252.

National Statistical Office. 2000. Preliminary Report the 2000 Population and Housing Census. Statistical Data Bank and Information Dissemination Division, Bangkok. Thailand.

Smith, L., Hyndman, R.J. and Wood, S.M. 2004. Spline Interpolation for Demographic Variables: the Monotonicity Problem. Journal of Population Research. 21(1), 95-98.

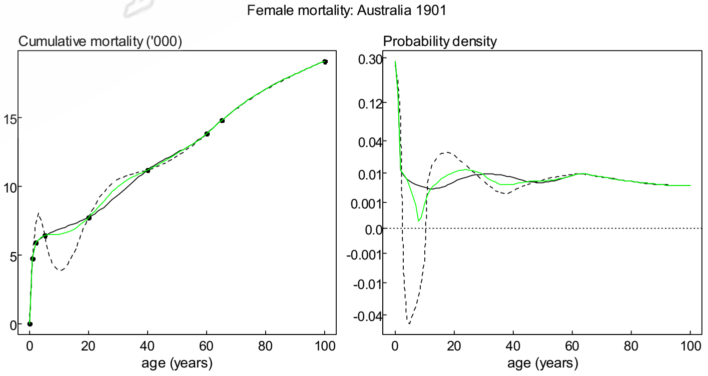


Figure 2. Spline interpolations of cumulative mortality for Australian women in 1901.

District-level Variations in the Quality of Mortality Data in Thailand

The results of the present study show that the quality of cause-of-death data varies markedly across Thailand. This study analyses the proportion of ill-defined and unknown causes of deaths occurring outside the hospital. About 75 per cent of all deaths occurred outside hospitals, but since persons who are not medically trained may not accurately diagnose the cause of death, there is a need to focus on improving these diagnoses.

Patarapan Odton, Kanitta Bundhamcharoen and Attachai Ueranantasun*

Different studies on variations of cause-specific mortality provide different policy implications and suggestions. Some findings mirror existing health care and services. Costantini and others (2000) concluded that differences in proportions of cancer patients dying at home across 13 provinces in Italy could not be explained by the known determinants, suggesting inappropriate hospital admission in the terminal phase of cancer. A study on geographical variations in breast cancer mortality in older American women by Goodwin and others (2002) suggested ways to improve the quality of breast cancer care. Some studies suggest further research in specific areas.

* Patarapan Odton and Kanitta Bundhamcharoen, Researchers, International Health Policy Programme, Ministry of Public Health, Thailand; e-mails: patarapan@ihpp.thaigov.net and kanitta@ihpp.thaigov.net; Attachai Ueranantasun, Lecturer, Department of Mathematics and Computer Science, Faculty of Sciences and Technology, Prince of Songkla University, Thailand, e-mail: attachai@gmail.com.

Vacchino (1999) constructed maps of cancer mortality that found associations of lung cancer with smoking and behaviour in women living in southern Argentinean provinces. An analysis of death rates from Parkinson's disease in Japan during 1977-1985 by Imaizumi (1995) concluded that the age-adjusted death rates from this disease were higher in the south-western than in the north-eastern parts of the country, indicating probable environmental risk factors.

In Thailand, a study by Lotrakul (2006) of suicide death rates during the period 1998-2003 found a high incidence in the upper northern region where HIV infection was high, and another study of district variation in cause-specific mortality by Faramnuayphol, Chongsuvivatwong and Pannarunothai (2008) found a clustering of liver cancer deaths in the upper north-east and chronic obstructive pulmonary disease in the upper north.

However, Mathers and others (2005) recently classified the death registration system in Thailand as low quality, with 49 per cent of total registered deaths assigned as ill-defined. Tangcharoensathien and others (2006) also evaluated the national death registration system in Thailand and identified two major problems contributing to both under-reporting of death registration and inaccuracy of cause-of-death attribution: problems in recording events or certifying deaths and problems in transferring information from death certificates to death registers. Although Thailand uses verbal autopsy studies (Choprapawan and others, 2001; Choprapawan, 2005) as an assessment tool for cause-of-death validation, the lack of regular quality control is an important issue (it is only undertaken every five years). During the period 1999-2001, about 40 per cent of the total registered deaths in Thailand were assigned as having either ill-defined or unknown causes of mortality.

Thailand introduced the national vital registration system in 1917 under the Ministry of Interior. By law, the death certificate must be issued by the district registrar within 24 hours after the death occurs. A medical death certificate is issued by health personnel (physician, nurse or medical coder) for a death occurring in a health institution, while a death notification report is issued by a local registrar (village head or health centre personnel) for a death occurring outside a health institution (Rukumnuaykit, 2006). The quality of data on cause of death in the vital registration system depends on the attendant at death and the methodology used to identify actual causes of deaths (Rukumnuaykit, 2006). Many deaths are coded by either non-medically trained staff or by health personnel who had little contact with the deceased.

The death registration system relies partially on second-hand reports from non-medical persons, unlike those of other countries, which rely primarily on medical or professionally trained coroners. As a consequence, some regions of Thailand rely more on non-medical and non-professional personnel, which may lower the quality of cause of death data at the national level.

In this study, the authors used a statistical method to estimate where low-quality mortality data occurred in Thailand. Thai mortality data from 1999 to 2001 were used to model the proportion of ill-defined causes that occurred outside hospitals over the 926 districts of Thailand.

Data and method

Gender-, age-, and cause-specific mortality data for the 926 districts in Thailand during the period 1999-2001 were obtained from the vital registration database. The database is provided by the Ministry of Interior of Thailand and is coded by the country's Bureau of Policy and Strategy, Ministry of Public Health, using the cause-of-death codes in the tenth revision of the International Classification of Diseases (ICD-10) (World Health Organization, 1992; Thailand, 2001).

Ill-defined causes are defined as deaths coded in ICD-10, issued in 1992 by the World Health Organization (WHO), as R00-R99 – "symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified" – when information was unavailable on cause of death. The proportion of ill-defined deaths is one of the indicators of (poor) quality in a national death registration system (Mathers and others, 2005).

Since populations of districts in Thailand vary substantially, with the total resident populations in 2000 ranging from a minimum of 2,088 (in King-Ko-Kut in Trat province) to a maximum of 451,447 (in Samut-Prakan city), the authors analysed the mortality incidence rates in aggregated districts called "super-districts", defined as regions comprising contiguous districts in the same province and having a total population of at least 200,000 (Lim and Choonpradub, 2007). A total of 235 super-districts were thus obtained, which varied in distribution from just 1 super-district in 14 provinces (Angthong, Singburi, Chainat, Nakhon-Nayok, Trat, Samut-Songkam, Amnat-Charoen, Mukdahan, Uthai-Thani, Phangnga, Phuket, Ranong, Krabi and Satun) to 24 in Bangkok province.

The adjusted proportions of ill-defined and unknown causes of mortality by age group and super-district were estimated using a logistic regression model. The probability P_{ij} that a reported death for age group i_{ij} and super-district j is ill-defined was thus modelled as:

$$\ln\left(\frac{P_{ij}}{1-p_{ij}}\right)=\mu+\alpha_i+\beta_j \qquad (1)$$

where μ is a constant and α_i and β_j are parameters associated with individual age groups and super-district, respectively, that sum to 0. These coefficients were estimated from the data and the adequacy of the model was assessed using statistical methods described in Venables and Ripley (2002).

To compare differences in ill-defined and unknown causes of mortality across the region of interest, illustrations by graphical method were used. Super-districts were classified into three groups, according to whether the confidence interval for the proportion was: (a) totally above the mean; (b) crossing the mean; or (c) totally below the mean. A thematic map was used to display this information using corresponding colours: (a) red; (b) orange; and (c) blue. Statistically valid conclusions can be made using this map, that is, the mortality in each red-coloured super-district is greater than the average mortality and the mortality in each blue-coloured super-district is less than the average mortality.

Statistical modelling and graphical displays used R commands (R Development Core Team, 2008).

Results

During the years 1999-2001, approximately 75 per cent of deaths among the Thai population aged less than 85 years occurred outside hospitals each year. About 42 per cent of the total deaths were certified as being of ill-defined or unknown causes. Table 1 shows the mortality rates per 100,000 population and the percentages of ill-defined or unknown causes of mortality for Bangkok and the four regions of Thailand. Residents in the northern region, both males and females, had higher mortality than those living in other regions. The central and northern regions had lower-than-average proportions of ill-defined deaths outside hospitals.

Figure 1 shows the results of applying model (1) to the proportions of ill-defined/unknown mortality outside the hospital in Bangkok by age group and super-district for males and females separately. It is clear from the graphs in the left-hand panels of figure 1 that there were high proportions of ill-defined mortality for persons aged 60 years and over (proportion of ill-defined mortality higher than average in men aged 60 years and over and in women aged 65 years and over). The residuals plots on the middle panels of figure 1 indicate that the model fit well. The observed and fitted proportions of ill-defined deaths are also plotted in the right-hand panels of figure 1.

The adjusted proportions of ill-defined mortality and their confidence intervals for each super-district are shown in the top panel of figure 2, where the horizontal dotted lines denote the average proportions: 39.8 per cent for males and 51.5 per cent for females. The solid horizontal lines correspond to the overall proportion of ill-defined mortality for males and females combined (44.1 per cent). The maps in the lower panel of figure 2 indicate that there were four super-districts in Bangkok with higher than average proportions of male mortality. For female mortality, six super-districts had higher than average proportions.

Table 1. Mortality rates and proportions of ill-defined causes of mortality for males and females aged less than 85 years in Thailand, 1999-2001

Place/region	Males		Females		Both genders	
	All causes[a]	Ill-defined (percentage)	All causes[a]	Ill-defined (percentage)	All causes[a]	Ill-defined (percentage)
Outside hospital						
Bangkok	233	39.8	130	51.5	181	44.1
Central	462	34.8	288	47.9	375	39.9
North-eastern	484	37.6	339	50.1	412	42.7
Northern	657	35.1	450	46.1	553	39.6
Southern	414	41.4	243	53.2	329	45.8
National	**475**	**36.9**	**312**	**48.9**	**393**	**41.7**
In hospital						
Bangkok	244	21.0	162	22.0	202	21.4
Central	216	18.3	138	19.6	177	18.8
North-eastern	109	19.1	66	20.0	88	19.5
Northern	180	14.5	118	16.2	149	15.2
Southern	115	15.0	67	16.9	91	15.7
National	**162**	**17.8**	**103**	**19.1**	**133**	**18.4**
All places						
Bangkok	477	30.2	292	35.2	383	32.1
Central	678	29.6	426	38.7	552	33.1
North-eastern	593	34.2	405	45.2	499	38.6
Northern	837	30.7	568	39.9	703	34.4
Southern	529	35.7	310	45.4	420	39.3
National	**637**	**32.0**	**415**	**41.6**	**526**	**35.8**

[a] deaths per 100,000 population.

Figure 1. Percentage of ill-defined and unknown causes of mortality outside the hospital among males and females aged less than 85 in Bangkok, 1999-2001

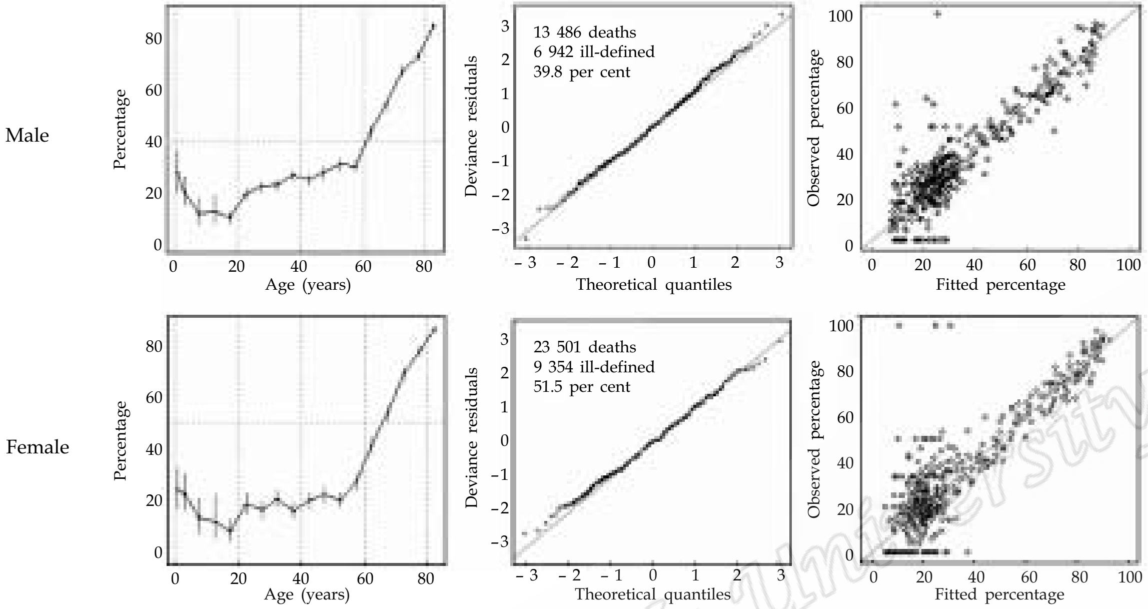
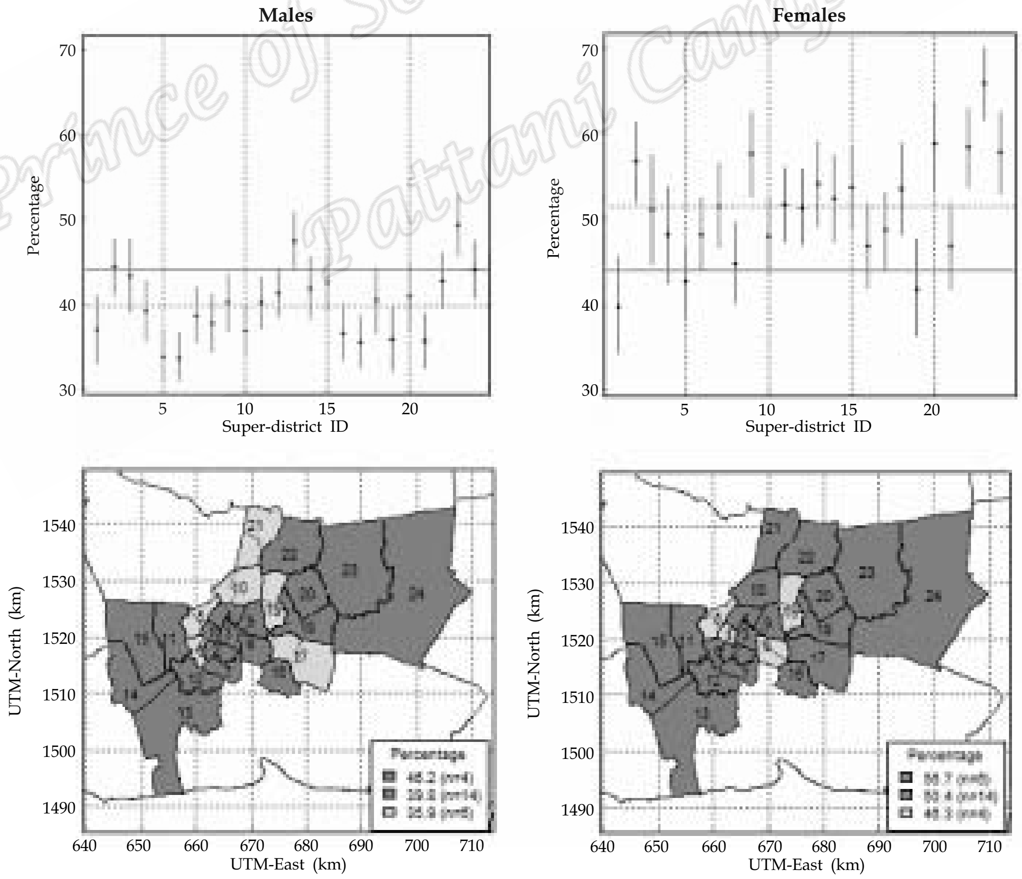


Figure 2. Age-adjusted percentage of ill-defined mortality outside the hospital for males and females aged less than 85 in Bangkok, 1999-2001



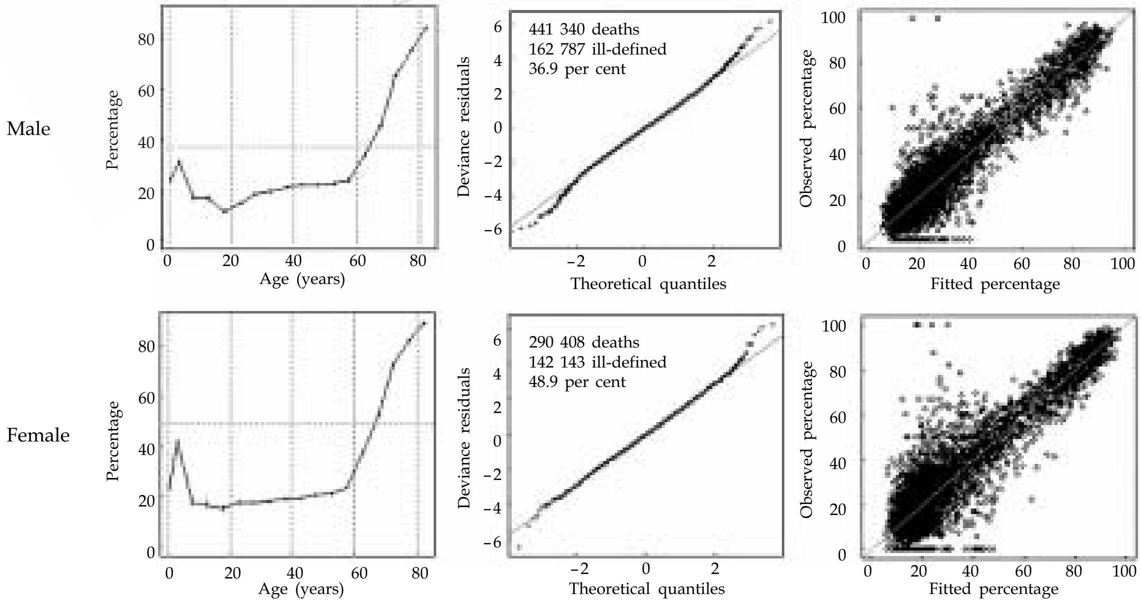
Super-districts in Bangkok

1 PhraNakhon, PomPrapSattruPhai, Samphanthawong
2 BangRak, KhlongSan, Sathorn
3 PathumWan, RatThewi
4 Dusit, PhayaThai
5 BangkokNoi, BangPlad
6 ThonBuri, BangkokYai
7 Yannawa, BangKhoLaem
8 KhlongToei, Wattana
9 HuaiKhwang, DinDaeng
10 Bangsue, Chatuchak
11 TalingChan, PhasiCharoen
12 Ratburana, ChomThong
13 Bangkhuntien, Thungkru
14 Nongkheam, BangBon
15 ThawiWattana, BangKhae
16 Prakanong, BangNa
17 Prawet, SuanLuang
18 BangKapi, Saphansung
19 LatPhrao, WangThonglang
20 BungKum, Kannayao
21 DonMuang, LakSi
22 BangKhen, Saimai
23 Minburi, KhlongSamWa
24 NongChok, LatKrabang

Abbreviation: UTM, Universal Transverse Mercator, ID, identification.

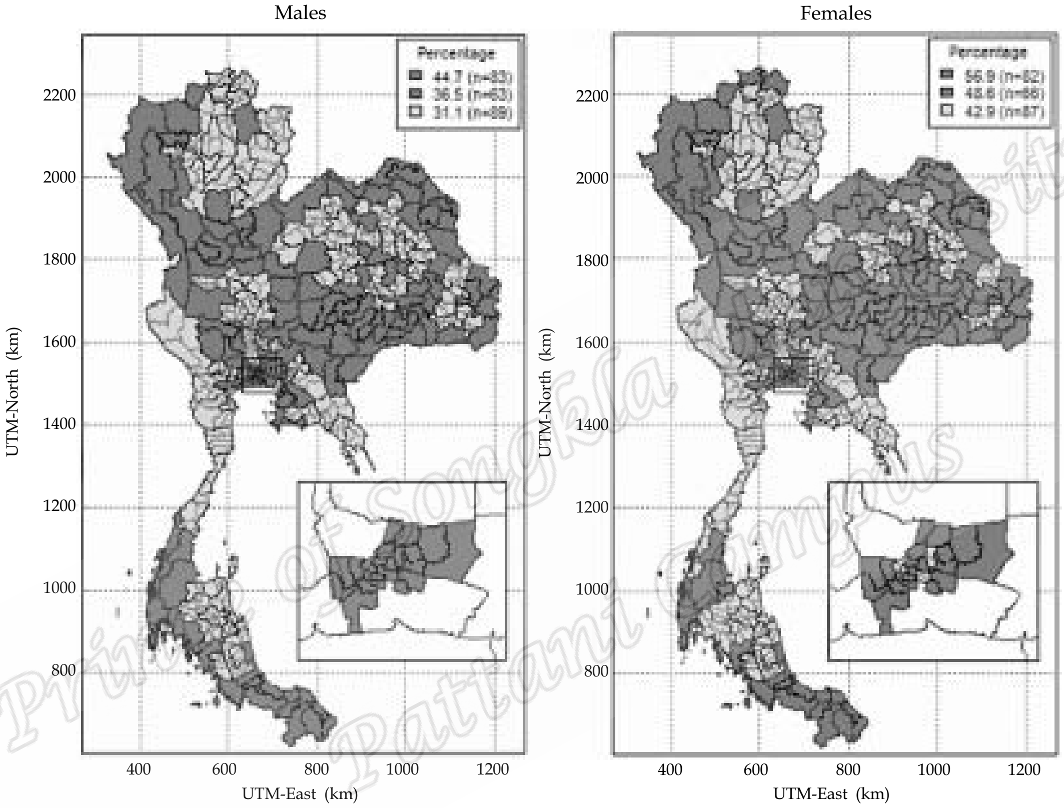
Figure 3 shows the results of applying model (1) to ill-defined mortality outside the hospital in the whole of Thailand. The graphs in the left-hand panels of figure 3 indicate a high percentage for children aged less than 5 years, older males and females aged 60 years and over.

Figure 3. Model results for the percentage of ill-defined and unknown causes of mortality outside the hospital for males and females aged 0-84 years, Thailand, 1999-2001



The age-adjusted percentage of ill-defined mortality was then estimated from the model for males and females separately. The 95 per cent confidence intervals of these percentages for the 235 super-districts were compared with the averages (36.9 per cent for males and 48.9 per cent for females) to produce the thematic map in figure 4.

Figure 4. Percentage of ill-defined mortality outside the hospital for males and females aged less than 85 in Thailand, 1999-2001



Abbreviation: UTM, Universal Transverse Mercator.

For male mortality, 83 super-districts with a group mean of 44.7 per cent had higher than average percentages of ill-defined home deaths, while 82 super-districts with a group mean of 56.9 per cent had higher than average percentages of ill-defined home deaths among females.

The highest proportion of ill-defined home deaths for males occurred in a super-district in Ubon Ratchathani containing NaChaluai, NamYun and Buntharik districts. For females, a super-district in Pattani containing Panare, Mayo, ThungYangDaeng, SaiBuri, MaiKaen, Yaring and KaPho districts had the highest proportion. The five super-districts with highest proportions of ill-defined home deaths for males and females are listed in table 2.

Table 2. Super-districts with the highest percentages of ill-defined mortality outside the hospital for males and females, Thailand, 1999-2001

Rank	Super-district (province)	Percentage (95% CI)	
Males			
1	NaChaluai, NamYun, Buntharik (Ubon Ratchathani)	66.8	(64.0,69.5)
2	Panare, Mayo, ThungYangDaeng, SaiBuri, MaiKaen, Yaring, KaPho (Pattani)	65.9	(63.6,68.1)
3	PlaPak, ThatPhanom, RenuNakhon, NaKae, WangYang (Nakhon Phanom)	65.0	(63.1,66.9)
4	MuangLoei, ChiangKhan, PakChom, DanSai, NaHaeo, PhuRua, ThaLi (Loei)	62.8	(60.7,64.8)
5	SiMuangMai, Khemarat, KutKhaoPun, PhoSai, NaTan (Ubon Ratchathani)	59.6	(57.1,62.1)
Females			
1	Panare, Mayo, ThungYangDaeng, SaiBuri, MaiKaen, Yaring, KaPho (Pattani)	79.1	(76.8,81.2)
2	NaChaluai, NamYun, Buntharik (Ubon Ratchathani)	78.7	(75.7,81.3)
3	PlaPak, ThatPhanom, RenuNakhon, NaKae, WangYang (Nakhon Phanom)	77.7	(75.8,79.5)
4	MuangLoei, ChiangKhan, PakChom, DanSai, NaHaeo, PhuRua, ThaLi (Loei)	76.6	(74.3,78.9)
5	MuangKamphaengPhet, KoSamPiNakhon (Kamphaeng Phet)	72.7	(69.9,75.3)

Abbreviation: CI, confidence interval.

Conclusion and discussion

The variation of ill-defined mortality across super-districts in Thailand is reasonably fitted to the logistic regression model, as the results show. A number of contributing factors which may relate to the variation were considered in terms of age, gender and residence. The proportion of deaths in old age in the female population was higher than that of men, which seems to contribute to a higher percentage of ill-defined causes of death among females, as indicated in table 1.

Senility (ICD-10: R54) is the major ill-defined cause of mortality and accounts for 58 per cent (50 per cent for males and 66 per cent for females) of the total number of ill-defined deaths. As deaths at older age accounted for more than half of total deaths, and as senility is a major cause of mortality in older age in the vital statistics, high proportions of ill-defined mortality in older age are observed in the left-hand panels of figure 3.

People in rural areas tend to have less access to health institutions, which affects the reporting on their cause of death. The super-districts with the highest percentages of ill-defined deaths outside the hospital, as shown in table 2, appear to have a high proportion (about 80-100 per cent) of the population living in rural areas (Thailand, 2002). Bangkok, however, which should have better cause diagnosis due to its concentration of hospitals and physicians, showed a high proportion of ill-defined deaths. This may relate to complicated illness conditions, such as co-morbidity, and conditions requiring advanced investigation.

The results of the present study show that the quality of cause-of-death data varies markedly across Thailand. While Choprapawan (2005) pointed out the need to focus on cause-of-death determination for hospital deaths, this study analyses the proportion of ill-defined and unknown causes of deaths occurring outside the hospital. About 75 per cent of all deaths occurred outside hospitals, but since persons who are not medically trained may not accurately diagnose the cause of death, there is a need to focus on improving these diagnoses.

To do this, Tangcharoensathien and others (2006) recommended that priority be given to strengthening mortality statistics, which could improve cause-of-death attribution through the verification of cause of death using the verbal assessment algorithm and the review of health personnel medical records for home deaths, in collaboration with the district registrar. They suggested that the quality of cause-of-death data on deaths outside the hospital could be further improved by training more than 70,000 village heads in Thailand, but this could be very costly. The modelling and mapping approach used by the authors is a useful preliminary tool for enabling public health researchers to plan investigations in specific areas, possibly only in one third of the 926 districts of Thailand.

Acknowledgement

This study was supported by the Health Information System Development Office, under the Thailand Burden of Disease Project. The authors are indebted to Prof. Don McNeil, Patarapan Odton's supervisor, for his help on this study. Thanks are also due to Sara Curran, Laragh Gollogly and Tomas E. Blair for their guidance on writing this research publication.

References

Choprapawan, C. (2005). Report on quality of cause of death data for in-hospital death 2003. Nonthaburi, Thailand: Health Information System Development Office.

Choprapawan, C., and others (2001). Report on validation of cause of death in Thailand. Thailand, Ministry of Public Health, Bureau of Planning and Strategy.

Costantini, M., and others (2000). Geographical variations of place of death among Italian communities suggest an inappropriate hospital use in the terminal phase of cancer disease. *Public Health*, vol. 114, No. 1, pp. 15-20.

Faramnuayphol, P., V. Chongsuvivatwong and S. Pannarunothai (2008). Geographical variation of mortality in Thailand. *Journal of the Medical Association of Thailand*, vol. 91, No. 9, pp. 1455-1460.

Goodwin, J.S., and others (2002). Geographic variations in breast cancer survival among older women: implications for quality of breast cancer care. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 57, No. 6, pp. M401-M406.

Imaizumi, Y. (1995). Geographical variations in mortality from Parkinson’s disease in Japan, 1977-1985. *Acta Neurologica Scandinavica*, vol. 91, No. 5, pp. 311-316.

Khan M. (2005). Suicide prevention and developing countries. *Journal of the Royal Society of Medicine*, vol. 98, No. 10, pp. 459-463.

Lim, A., and C. Choonpradub (2007). A statistical method for forecasting demographic time series counts, with application to HIV/AIDS and other infectious disease mortality in southern Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health*, vol. 38, No. 6, pp. 1029-1040.

Lotrakul, M. (2006). Suicide in Thailand during the period 1998-2003. *Psychiatry and Clinical Neurosciences*, vol. 60, No. 1, pp. 90-95.

Mathers, C.D., and others (2005). The dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the World Health Organization*, vol. 83, No. 3, pp. 171-177.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available from: http://cran.r-project.org/doc/manuals/refman.pdf.

Rukumnuaykit, P. (2006). Mortality and causes of death in Thailand: evidence from the survey of population change and death registration. *Asia-Pacific Population Journal*, vol. 21, No. 2, pp. 67-84.

Tangcharoensathien, V., and others (2006). A critical assessment of mortality statistics in Thailand: potential for improvements. *Bulletin of the World Health Organization*, vol. 84, No. 3, pp. 233-238.

Thailand, Ministry of Public Health, Bureau of Planning and Strategy (2001). Public health statistics A.D. 2000.

______, National Statistical Office, Statistical Data Bank and Information Dissemination Division (2002). The 2000 Population and Housing Census. Bangkok.

Vacchino, M.N. (1999). Poisson regression in mapping cancer mortality, *Environmental Research*, vol. 81, No. 1, pp. 1-17.

Venables, W.N., and B.D. Ripley (2002). *Modern Applied Statistics with S*. New York: Springer-Verlag.

World Health Organization (1992). *International Statistical Classification of Diseases and Related Health Problems*, tenth revision, Geneva.

Journal of Map And Geography Libraries, 7:211–219, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 1542-0353 print / 1542-0361 online
DOI: 10.1080/15420353.2011.566843

Graphing Incidence Rates over Regions using R and Google Earth

SUMPUNT KHONGMARK, METTA KUNING,
and ATTACHAI UERANANTASUN
Prince of Songkla University, Pattani, Thailand

A method for modeling and graphically comparing incidence rates of adverse events that vary over geographical regions is shown and discussed. To achieve our goal, we used a statistical model to compare incidence rates and showed how informative three-dimensional graphs of such incidence rates can be created dynamically using R and interactively controlled using Google Earth with Keyhole Markup Language (KML). These methods are applied to the terrorism events in regions of Southern Thailand that occurred from 2004 to 2009.

KEYWORDS *Geographical mapping, three-dimensional plots, dynamic KML, statistical comparison, confidence intervals, R software, Google Earth*

INTRODUCTION

Displaying a quantity that varies by geographic region presents a cartographic challenge because two spatial coordinates are already needed to specify latitude and longitude, so a place must be found to show the statistical variation. Ideally, this variation should include not just the data variation but also the extent to which differences between values at different locations are real, or simply a result of chance variation. A method is thus needed for displaying both the data and the reality of differences on a graphic information system (GIS) map. The purpose of this paper is to provide an appropriate

We would like to thank our colleagues Don McNeil, Somporn Chuai-aree, and Patarapan Odton for their help and suggestions.

Address correspondence to Sumpunt Khongmark, Department of Communication Sciences, Prince of Songkla University, Pattani Campus, 181 Chareonpradit Road, Rusamilae, Pattani, Thailand 94000. E-mail: ksumpunt@bunga.pn.psu.ac.th

method of doing this using software that is relatively easy to use and freely available.

Our method involves first fitting an appropriate linear statistical model, and then combining the graphs produced by this model as screen overlays in a GIS with dynamic three-dimensional histograms produced by Google Earth software, where colors are used to distinguish real differences. An analysis and model-fitting part can be achieved by using R programming language, widely used in statistics research due to its structure. Users with little programming background can easily start to write their own codes in R. In addition, it is powerful enough to provide a broad range of analyzing commands, and it can work with sizeable relational databases. In terms of data visualization, R programming can produce a graphic output using a series of graphics functions from its library. These outputs can be presented as graphs and images and transferred for use in Google Earth.

All software used to develop the system is freely downloadable from the Internet. The system uses Keyhole Markup Language (KML) and the basic Google Earth program. The KML source code is created dynamically using R (R Development Core Team 2010). We also used R to fit the statistical model and to create Figures 1 and 2 (below).

In the following sections, the data and statistical method are briefly described. Further sections describe how a GIS map can be created by using KML source code in basic Google Earth software. In the Discussion section we share some extensions of the method to more general dynamic graphics applications.

ILLUSTRATIVE DATA: VICTIMS OF VIOLENCE IN SOUTHERN THAILAND

For our model, we considered incidence rates per hundred thousand population of terrorism events classified by gender, age-group (<25, 25–44, or 45 or more), district of residence (15 levels as shown in Table 1), and year (six levels from 2004 to 2009 inclusive). Each adverse outcome corresponds to a civilian victim suffering injury or death as a result of a defined violent terrorism event in the target area. In this case, the target area is defined as all districts 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 were retrieved from a database maintained by the Deep South Coordination Centre, Thailand (http://medipe2.psu.ac.th/∼dscc/). The population denominators were obtained from the 2000 population and housing census of Thailand.

Because the overall victim incidence rates for Muslims were very much lower than those for other residents of the terrorist target area, we restricted the study to non-Muslim victims. For purposes of analysis, we first aggregated the thirty-seven districts within the target area into twenty-three regions with

TABLE 1 Regions used in analysis of victim violence incidence in Deep South of Thailand

Province	RegionID: Districts	Non-Muslim Pop
Songkla	1: Chana/TePa	62,621
	2: SabaYoi/NaTawi	62,236
Pattani	3: Pattani City	41,122
	4: KoPho/MaeLan	34,812
	5: NongChik/Mayo/ThungYangDang/Kapo/Yaring/Yarang	19,878
	6: Panare/Saiburi/MaiKaen	19,717
Yala	7: Yala City	75,291
	8: Betong/ThanTo	36,706
	9: BannangSata/KrongPinang/Yaha/Kabang/Raman	17,845
Naratiwat	10: Naratiwat City	31,950
	11: TakBai	15,376
	12: Bacho/YinGo/Rueso/Rangae/SaSikon/Chanae	23,560
	13: Sukirin/Waeng	11,624
	14: SungaiPadi/Cho-airong	13,563
	15: SungaiKolok	23,323

populations ranging from 54,039 to 154,634. However, some of these regions contained fewer than 5,000 non-Muslim residents, so to provide acceptably stable estimates of incidence rates, we further aggregated the regions with smaller non-Muslim populations and thus reduced the number of regions to fifteen as listed in Table 1.

STATISTICAL MODEL

A regression model for incidence rates was established using the four factors noted above. To remove skewedness and thus satisfy statistical assumptions, the incidence rates were log-transformed. There were 109 zeroes in the 540 cells in the contingency table of victim counts; the incidence rates in these cells were inflated by replacing these zeros by 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, chap. 6).

Figure 1 gives confidence interval plots for the incidence rates with respect to gender-region after adjusting for age-group and year. The incidence rates for females was substantially lower than those for males in all regions except Naratiwat City and Sukirin/Waeng.

GIS THEMATIC MAP

Figure 2 shows a simple thematic map of the target area containing a pair of colored square boxes for the two sexes in each of the fifteen regions. The

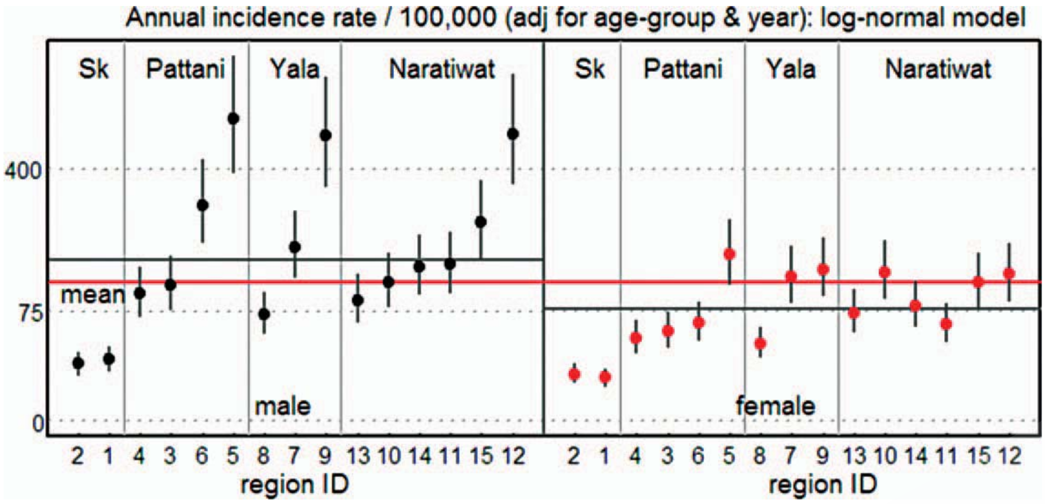


FIGURE 1 Estimates of annual incidence rates per 100,000 population of injury to civilian non-Muslim residents of the terrorism target area in Southern Thailand, classified by gender and region after adjusting for year (2004–2009) and age group (<15, 15–44, 45+). The horizontal red line denotes the overall 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.

color codes are based on the locations of the confidence intervals in Figure 1 with respect to the overall mean incidence rate. A box is colored red, blue, or green according to whether its corresponding confidence interval is entirely above the mean, entirely below the mean, or crosses the mean.

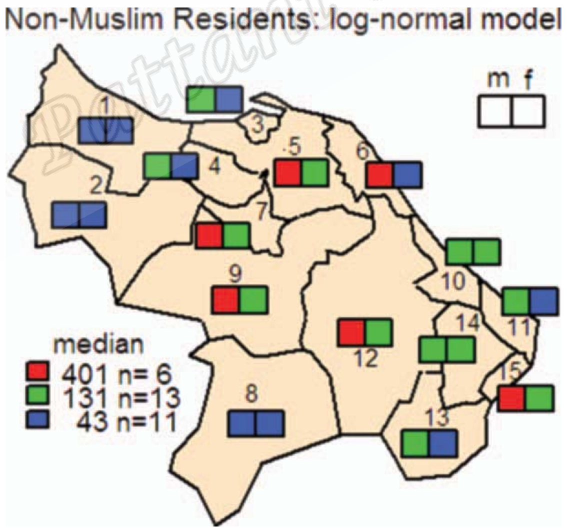


FIGURE 2 Simple thematic map of the 15 regions in the Southern Thailand terrorism target area, using three colors to classify the terrorism risk for males (left boxes) and females (right boxes) as above average (red), below average (blue), or not evidently different from average (green). The numbers in the legend denote the median incidence rates per 100,000 populations in each of the three groups.

Although colors of the boxes on the thematic map are useful for classifying the risk with respect to the overall mean, they say little about the magnitude of the risk. However, thematic maps can be improved to incorporate these magnitudes by replacing the colored boxes with colored bar charts. This alternative has the drawback of taking up valuable space on the map. In the next section we consider how to augment the thematic plot by moving into three dimensions.

INTERACTIVE GIS GRAPHICS SYSTEM

In this section we describe how to create an interactive graphics system for displaying and controlling the position of graphs and maps similar to those shown in Figures 1 and 2. This system also creates three-dimensional histograms as shown in Figure 3. (See this sample of our work and how it can be interacted with a viewer at http://scitech.sat.psu.ac.th/GEDSV/).

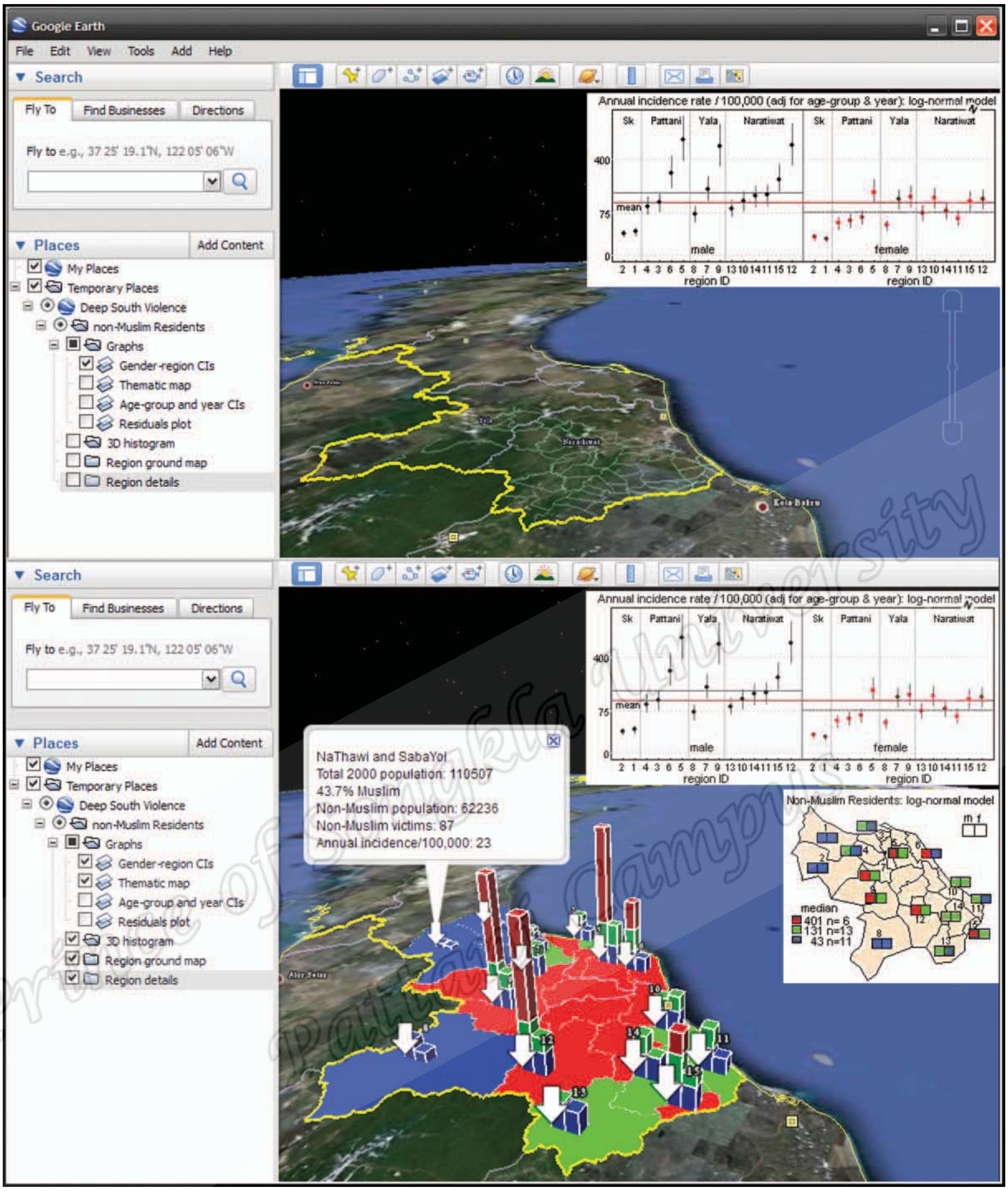
KML is an extended development of XML, a readable text file used by Google Earth for manipulating and displaying geographic information. Although R does not yet have a built-in library or function to create a KML source code file, it can handle text files and perform database functions, so additional software is not essential. However, a library for R that can facilitate KML source code creation is freely available on the Internet (XML for S-PLUS Core Team 2009).

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.

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 1 and 2, 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



☑ My Places
☑ Temporary Places
⊙ Deep South Violence
⊙ non-Muslim Residents
■ Graphs
☑ Gender-region CIs
□ Thematic map
□ Age-group and year CIs
□ Residuals plot
□ 3D histogram
□ Region ground map
□ Region details

☑ My Places
☑ Temporary Places
⊙ Deep South Violence
⊙ non-Muslim Residents
■ Graphs
☑ Gender-region CIs
☑ Thematic map
□ Age-group and year CIs
□ Residuals plot
☑ 3D histogram
☑ Region ground map
☑ Region details

NaThawi and SabaYoi
Total 2000 population: 110507
43.7% Muslim
Non-Muslim population: 62236
Non-Muslim victims: 87
Annual incidence/100,000: 23

FIGURE 3 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 1 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, and corresponding color-coded extruded polygons with ground overlays and pop-up boxes giving further details about region characteristics.

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 preprocessed 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 Web site (http://code.google.com/intl/th/apis/kml/).

DISCUSSION

Presenting experimental results in digital formats to the public without having to install a complicated tool is a goal of some scientists and statisticians. Google Earth is freely available for downloading both as a stand-alone program and as a plug-in for a Web browser in almost any computer and operating system. Google Earth also provides customizability and interactiveness. A layer can be added or removed, depending on the viewer’s request, while the ability to zoom, rotate, and tilt a global map aids viewers by adjusting an angle to compare spatial data of their own preference. Additionally, spatial outcomes from an experiment or analysis can be imported to Google Earth in the form of KML files and made available to the public via the Internet. Apart from our study, Wood et al. (2007) have demonstrated an application for using the Google Earth–displayed tag approach of data sets and embedding KML files in a server.

Dodsworth (2008, 67) shows examples of using Google Earth with historical maps, but some applications include purchased software such as ArcGIS or Excel to create KML files. This is not a practical choice when purchasing propriety software is not an option. Alternatively, R is developed under GNU general public license and can be freely used. This means that using R combined with Google Earth is a virtually free approach opening possibilities for conveying research with less, or no, cost involved.

Open-source software and free software have become common in GIS research for some time now. GRASS GIS (http://grass.fbk.eu/) is among the most preferred tools due to its open-source nature and functionalities, and its combination with R has been successfully demonstrated (Bivand and Neteler 2000). Although this combination is effective, it demands a high learning curve and extra programming skill to develop interactive and widely accessed maps compared with using the combination R and Google Earth.

CONCLUSION

The prototype system we have developed takes advantage of ongoing developments in computing and communications technology to provide an improved paradigm for statistical graphics, following pioneering work by Sandvik (2008). The ability to combine screen overlays of statistical graphs with an interactive geographical information system having the ability to range over the earth's surface and zoom into local areas provides a very powerful tool for the scientist. The terrorism victim application clearly shows the importance of looking at both the statistical graphs and the raw data at the same time. Viewing the confidence intervals alone is very useful for making valid statistical comparisons of risk rates, but transforming these rates to remove skewedness, and thus satisfy statistical assumptions, can obscure their real magnitudes. Showing the untransformed histograms on a map together with the statistical graphs restores the balance, giving the viewer a more complete and informative picture.

This system can be developed much further, not only for assisting social scientists to gain a better understanding of terrorism study, but also for other studies with a similar data structure, such as providing informative graphs of death rates by age-group, gender, district of residence and year, with the object of identifying environmental factors associated with mortality (Odton 2010). In studies wherein region can be separated from other factors that include interactions, it is of interest to graph incidence rates for these factors after adjusting for region, and such graphs can still be created as ground overlays using KML, even though they are not connected to a particular geographical location. The method can also be applied to outcomes other than incidence rates, such as means and proportions. A further application of current interest is the pattern of global temperature changes on the earth's surface in recent decades. There are several GIS software systems that have been used to analyze data and display outcomes on a map. However, a combination of R and Google Earth has some key benefits compared with other software.

REFERENCES

Bivand, R. S., and Markus Neteler. 2000. Open source geocomputation: Using the R data analysis language integrated with GRASS GIS and PostgreSQL data base systems. Paper presented at the 5th Conference on GeoComputation, University of Greenwich, U.K., August 23–25.

Dodsworth, E. 2008. Historical mapping using Google Earth. *Cartographic Perspectives* 61: 63–69.

Murrell, P. 2006. *R Graphics*. Boca Raton, FL: Chapman and Hall/CRC, Taylor and Francis Group.

Odton, P. 2010. Trend and levels of liver cancer mortality in the upper north-eastern region of Thailand: A case study of using Google Earth in public health. Paper presented at the annual National Statistical Conference in Chiang Mai, Thailand, May 27–28.

R Development Core Team. 2010. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org (accessed September 10, 2010).

Sandvik, B. 2008. "Thematic mapping at Googleplex." *Thematic mapping blog*. http://blog.thematicmapping.org/2008/12/thematic-mapping-at-googleplex.html (last modified December 13, 2008).

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. New York: Springer.

Wood, J., J. Dykes, A. Slingsby, and K. Clarke. 2007. Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. *IEEE Transactions on Visualization and Computer Graphics* 13(6): 1176–1183.

XML for S-PLUS Core Team. 2010. XML: Tools for parsing and generating XML within R and S-Plus. In *The Comprehensive R Archive Network*, http://cran.r-project.org/web/packages/XML/index.html (accessed December 14, 2010).

International Journal of Business and Social Science Vol. 3 No. 19; October 2012

Analyzing National Elections of Thailand in 2005, 2007, and 2011 – Graphical Approach

Attachai Ueranantasun
Department of Mathematics and Computer Science
Faculty of Science and Technology
Prince of Songkla University, Pattnai Campus
Pattani, Thailand

Abstract

This study examines recent Thailand's national election on the last three occasions – 2005, 2007, and 2011. The voting results are analyzed and thematic maps, using a variable scale to clarify dense data plots, for the whole of Thailand are produced to display the majority vote in each province and coloured points representing winning candidates. The graphical results show that the number of winning candidates coincides with the percentage of majority votes in almost all provinces. The results from the 2007 and 2011elections illustrate that there is a clear split between the North-East and the South-West, indicating the main competition of two major parties. The vote swings are estimated from 2005 to 2007 and 2007 to 2011, using a simple statistical analysis to distribute gain-loss proportion in the form of bubble plots. It shows that minor parties, gained benefit from the political turmoil in 2007, but lost their significant votes in 2011.

Keywords: Thailand's National Election, Voting Analysis, Statistical Graphics, Thematic Map

1. Introduction

Since transforming into a democratic system in 1932, there have been 25 general elections in Thailand. The latest election was held on 3 July 2011. The last two constitutions of Thailand, of 2007 and 2009 (Asian Legal Information Institute, 2012), specify that voting in Thailand is compulsory for every Thai citizen aged 18 years of age or more. Moreover, the 1997 Constitution introduced a different election system from the past. Members of Parliament are determined by two forms of voting systems: party list vote and constituency vote. For each election, a voter is given two types of ballots for both voting systems. In the party list voting system, a party provides a list of candidates and a party with more than 5% of the total vote is considered an eligible party. The numbers of MPs for each eligible party are calculated using the proportion of total votes that the party receives. A constituency or district voting system is more straightforward. Voters can cast a vote to select directly a preferred candidate available in their electoral area. The introduction of the party list voting system is thought to add stability to the Thai political system and grown interest from Thai citizens in the election (Surarit, 1996). Remarkably, it is found that more voters turned out after the application of 1997 Constitution. There were less than 65% of turnout voters for general elections before 1997 (King Prajadhipok's Institute, 2010), whereas all elections after the introduction of 1997 Constitution showed a growth the treatment in percentage - 69.94% (2001), 72.56% (2005), 74.52% (2007) and 75.03% (2011), respectively (The Election Commission of Thailand, 2011).

A few studies concentrate on election results, but they mostly discuss the result of one election (Phatharathananuntha, 2008; Schafferer, 2009). In this study, we discuss the results of more than one election and present a simple but informative way to display results.

2. Thailand Elections from 2005 To 2011

There have been three official national elections between 2005 and 2011. According to the Election Commission of Thailand (2011), the numbers of MPs were different for these three elections. In the 2005 election, the total number of MPs was 500, divided into 400 for constituency seats and 100 for party list seats.

The overall number was reduced to 480 in 2007, (400 constituency seats and 80 party list seats). The number of MPs changed again in 2011, to 500 (375 constituency seats and 125 party list seats). For both the 2005 and 2011 elections, there was one seat per constituency, while the seats in one electoral district, in 2007, ranged from one to three, depending on the population in the area. In 2005 and 2007, there were 76 provinces in Thailand. Bueng Kan, the most northern province of the North-East region, was established in early 2011 by dividing Nong Khai into two provinces.

During the period, the coup d'état, the rallies, and the dissolution of political parties by the Constitutional Tribunal resulted in a change of political stability including parties involving in elections. These changes are displayed in Figure 1.

3. *Methodology*

Since the election result data are area-bound and there now 77 provinces in Thailand, the demonstration of election results as tables and numbers makes comparison difficult. However, the graphical displays can provide understanding of election outcomes in terms of trends in votes and changes in the political situation.

3.1 Election data

The election data used in this study are retrieved from official election results maintained by The Election Commission of Thailand. These data cover three elections for the House of Representatives of Thailand in 2005, 2007, and, the most recent election in 2011. We consider only the results from constituency votes to focus on a directly elected numbers of members of parliament. Due to differences between electoral districts, numbers of constituencies and thus MPs in each election, we aggregate votes at provincial, regional, and federal levels instead of districts. All data are processed and all graphics are produced using the R statistical and graphical package, which is an open source software environment capable of analyzing cumulative data. Graphical presentations in this presentation are divided into two parts – thematic maps and bubble charts of swing votes.

3.2 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 votes in a province is selected. An individual color is assigned to all parties with elected candidates. Acknowledging a party's background and changes, we aim to continue coloring consistency for each corresponding party throughout all three elections. The color of a winning party for each province is thus filled in a matching area on the map of Thailand. The following process another adding points representing elected candidates in the corresponding province. Since the data are categorical, colored points of the same size are appropriate.

Figure 2 shows maps of Thailand with themes of majority votes in provinces and colored points of winning candidates in each province. The legends in the map sets also show percentages of total votes and numbers of elected MPs for the majority parties, while parties without a winning candidate are grouped as "others".

The maps exhibit the voting results as intended. However, there is a problem of displaying results for Bangkok and surrounding regions. Bangkok is the capital city of Thailand, and the country's administrative and economic centers are spread around Greater Bangkok, consisting of Bangkok and adjacent provinces – Nakhon Pathom, Nonthaburi, Pathum Thani, Samut Prakan, and Samut Sakhon. Greater Bangkok covers 6758 km^2 with more than 60 MPs representing the area for last three elections, comparing to Nakhon Ratchasima, the largest province of 20,494 km^2 in area, which accommodates around 15 representatives (Department of Provincial Administration, 2011). Consequently, it becomes clear in the maps that the plotted points, in Greater Bangkok, are crowded and almost undistinguished. One solution to this problem is to insert an inset map, magnifying around the area. However, an inset map requires additional space.

To solve this problem, we employ a variable-scale map to stretch the area of Greater Bangkok. Variable-scale mapping a technique used to focus on a certain area of the map by using a large-scale map at the focused area and a small-scale map for others. It can be implemented by decreasing the scale radically in a linear pattern from the center of the area to the end of perimeter (Fairbairn & Taylor, 1995).

This technique has been further adapted to use for many applications, includes, a navigation system for a small display device (Harrie, Sarjakoski, & Lehto, 2002) and a web-based map display for a mobile (Yamamoto, Ozeki, & Takahashi, 2009; Haunert & Sering, 2011).

In this study, 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. 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:

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

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 follows:

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

Consequently the new Cartesian coordinates are:

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

The radial scaling for three distance levels is shown in Figure 3.

3.4 Vote swings

Since the numbers of constituent MPs vary between provinces, regions, and the nation in all three elections, it is more preferable to compare the votes in percentage than in numbers. In this step, we estimate the percentages of the total vote transferred from one party to another from 2005 to 2007 and from 2007 to 2011. To achieve this outcome, 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:

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

We demonstrate an example of this calculation in Figure 4, which shows percentage changes for each party from 2005 to 2007 for the whole Thailand. In this case, the diagonal terms are the minimum values of party's percentages of candidate votes for 2005 and 2007. The Democrat Party gained 5.2% of votes in 2007 over 2005 and the fraction of this percentage is computed from the 18.9% loss of the Thai Rak Thai party, 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 sizes of the bubbles are compatible with the percentages of each term. Only four majority parties, in terms of percentages of votes, are illustrated, while the remaining parties are included in "others".

3.5 Framework of Analysis

After producing thematic maps and vote swing graphics, we use these depictions to analyze the outcomes of the three elections by comparing voting results, both the numbers of MPs and percentages of votes, for the whole nation, regions, and particular provinces. The analysis also focuses on the trends and changes in popularity of participating parties, that won at least one seat in an election. We use thematic maps to link between parties and provinces as well as regions, whereas vote swings are employed to understand the patterns of exchanges of votes between parties.

4. Results

4.1 Thematic maps

The colored maps in Figure 5 present the majority votes and winning candidates in each province for the whole Thailand for the three elections in 2005, 2007, and 2011.

The background colors in each province in Figure 5 can be interpreted as the indication of popularity of parties in elections. In 2005 there was effecting a race between two parties. Thai Rak Thai party received a majority of votes in nearly all provinces in the upper part of the country, while the Democrat party won in all provinces in the lower part of the country. The only exception was Chart Thai party gained popularity in two provinces in the Central region. When considering the popularity in the whole country, Thai Rak Thai won 55.7% and this figure was greater than the votes of all other parties combined. In 2007 the Thai Rak Thai party became People's Power (PPP) and lost 19 of its previously won provinces (mainly in the Central region) to the Democrat (14), Chart Thai (4), and Pracharaj (1) parties. The overall percentage for the PPP parties was also reduced by approximately 19% and only 6.6% over those of the Democrat party. The situation the PPP, becoming Puea Thai, was then changed again in 2011, when it regained a higher percentage of votes overall.

In each province, super-imposed point show the numbers of elected MPs. For all three elections, overall associations between popularity and numbers of winning MPs were consistent. In 2005 and 2007, the party that secured the most votes in the province also had the largest number of elected MPs. This trend was as well followed in 2007 except for a few provinces, where the majority of votes did not match the number of elected MPs. There even were a few provinces where all MPs did not come from the most popular party. In such cases, differences between the popularity and the numbers of winning candidates can reflect close competition between parties in the area. This is confirmed by the smaller margins of voting percentages in 2007 to those of 2005 and 2011.

4.2 Bubble charts of vote swings

In this section, we study the percentages of votes in Thailand as the allocations of loss-gain votes for the major parties in each pair of elections were shown in Figure 6.

Figure 6 shows the transferred votes for the whole country in a pair of elections. Between 2005 and 2007, TRT or PPP was the party which lost the most vote percentages, for the most of them, to other small parties and Pueapandin, respectively. Chart Thai was another party marginally losing votes. Democrat was the only continuing major party that gained the votes, mainly from PPP. Pueapandin was the new party in the 2007 election. It is interesting to see that Pueapandin benefited more from the fall of PPP than the disappearance of Mahachon. Small parties also gained higher shares of votes in 2007 mainly from PPP. Comparing 2011 to 2007, Puea Thai or PPP gained back a number of votes from Pueapandin and minor parties, while Democrats also gained slightly more votes from other parties. From 2007 to 2011, the newly established party like Poomjaithai did not gain votes from Puea Thai, but chiefly from discontinued Pueapandin and other parties. Chart Thai Pattana still experienced a downfall as its votes were taken away by others. Other small parties enjoyed a rise of votes in 2007, but faced considerably lower percentages in 2011.

4.3 Discussion

The landslide victory of TRT in both numbers of MPs and popularity in 2005 made history of which one party was able to set up an autonomous cabinet. Yet the 2006 court dismissal of the party altered the situation. Even though its successor, PPP, had a majority of seats and overall votes, it lost a number of MPs and popularity to Democrat and the parting factions in 2007. The founding of Pueapandin, Ruamjaithai Chartpattana, Pracharaj, and Machimathipatai was the answer to escape a deadlock for some TRT members. It was a successful attempt as they could gain advantage in winning seats in the Parliament. Pueapandin was the most successful separated faction from TRT as it won 17 seats and 12 of them were from North-East region. Ruamjaithai Chartpattana and Pracharaj enjoyed a provincial level of achievement in Nakhon Ratchasima and Sakaew, respectively. Additionally, Machimathipatai realized its elected MPs, from the upper part of the country. When there was dissolution for PPP in late 2008, most of the members, supported by members from Pracharaj and some members of Pueapandin, formed the new party Puea Thai.

Like its predecessor, it won a majority of votes and MPs in the 2011 election, partially from votes from Pueapandin. A combination of Pueapandin and Ruamjaithai Chartpattana did not appear to be profitable as its new formation, Chartpattana Pueapandin, was constrained with a smaller numbers of MPs, mainly in its previously won province, Nakhon Ratchasima. The court dismissal of PPP also contributed to two break-away parties, Poomjaithai and Palangchon. Poomjaithai seemed to be the more successful of the two as it won in a majority of votes in five provinces. The reason for this might be from the combination with banned Machimathipatai. For Palangchon, it was determined to be a provincial party and it was successful of doing that in Chonburi.

Chart Thai was an interesting phenomenon in Thai politics. In 2007, its addition of Mahachon, who won only two seats in the North-East region in the previous election, did not affect much on the overall percentages of votes and hence to popularity. Nevertheless, the party still won more MPs, mainly in Central region, and became third in terms of quantity of seats in 2007. However, as its popularity still declined in the 2011 election in almost all regions, the numbers of its elected MPs decreased significantly. This strangely coincided with the 2008 dismissal of Chart Thai, which later was reformed as Chart Thai Pattana. It is a well known fact that the party has a stronghold in Central region, partially in Suphanburi, where it has maintained success in all elections.

The Democrats were the only party that gained higher percentages of votes in every region at the latest elections. However, this increasing achievement was marginal compared to the percentages of Puea Thai in North and North-East regions. The numbers of constituency seats in these two regions were high enough to limit the Democrats to second place overall.

Fewer parties winning at least one constituency seats can be seen in the three elections. The numbers of these particular parties were four, seven, and seven in 2005, 2007, and 2011, respectively, while the average of seat-winning parties from 1983 to 1996 was 12 (Hicken, 2002). The dominance in the Southern region confirms the Democrats status as the southern party (Pongsudhirak, 2005), while Puea Thai established itself as a northern and north-eastern party. The Democrats recent win in Bangkok and some provinces in the western part of Thailand, divides the nation into two parts in terms of political preferences. The North-East and South-West divisions set the country towards a two-party political system. The other minor parties then tend to become more provincial or local parties. It is suggested that these tendencies can be accredited to a better political conscientiousness among Thai voters (The Asia Foundation, 2011).

5. *Conclusion*

We have shown that the employment of graphics instead of mere numeric data can be useful to analyze election results. As geographically bound data, the election results are suitable to be displayed in the form of maps. A variable-scale map provides an appropriate method for creating maps of electoral vote percentage results where districts vary substantially in area, and these maps can are enhanced using thematic bubble charts. The vote swing is also included in the study by using a simple calculation and simple bubble display to see transferred percentages among parties. Using the aforementioned tools, Thailand's election results from 2005 to 2007 can be summarized as follows:

- When comparing the majority of the votes and the number of MPs in all provinces, we found that most of them are consistent.
- The analysis also suggests that the competition was closer in 2007 and 2011 than in 2005 in North-East and Central regions.
- The North-East and South-West partition shows the development of two-party system in Thailand.

Further studies are needed to determine the most appropriate in the area of using geographical analysis of election data. Election results can also be presented using other types of maps, for example, a cartogram (Gastner, Shalizi, & Newman, 2005). It is also worthwhile to evaluate different statistical models that identify candidates and take into account demographic voting patterns within individual electorates for using in analysis of vote percentage swings from one election to the next.

References

Asian Legal Information Institute. (2012). *The Constitution of the Kingdom of Thailand 2007*. [Online] Available: http://www.asianlii.org/th/legis/const/2007/index.html (September 14, 2012)

Asian Legal Information Institute. (2012). *The Constitution of the Kingdom of Thailand 1997*. [Online] Available: http://www.asianlii.org/th/legis/const/1997/index.html (September 14, 2012)

Department of Provincial Administration . (2011). Administration Data. [Online]Available: http://www.dopa.go.th/padmic/jungwad76/jungwad76.htm (October 12, 2011)

Fairbairn, D. & Taylor, G. (1995). Developing a Variable-Scale Map Projection for Urban Areas. *Computers & Geosciences*, 21(9), 1053-1064.

Gastner, M. T., Shalizi, C. R., & Newman, M. E. J. (2005). Maps and Cartograms of The 2004 US Presidential Election Results. *Advances in Complex Systems*, *8*(1), 117-123.

Harrie, L., Sarjakoski, L. T., & Lehto, L. (2002). A Mapping Function for Variable-Scale Maps in Small-Display Cartography. *Journal of Geospatial Engineering*, 4(2), 111-123.

Haunert, J.H. & Sering, L. (2011). Drawing Road Networks with Focus Regions. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2555-2562.

Hicken, A. D. (2002). *The Market for Votes in Thailand*. Presented at the conference, Trading Political Rights: The Comparative Politics of Vote Buying, Massachusetts Institute of Technology, Massachusetts, USA.

King Prajadhipok's Institute (2010). *Politics and Administration Database*. [Online]Available: http://www.kpi.ac.th/wiki/index.php (October 27, 2011)

Phatharathananuntha, S. (2008). The Thai Rak Thai party and elections in North-eastern Thailand. *Journal of Contemporary Asia*, 38(1), 106-123.

Pongsudhirak, T. (2005). Thai Politics after the 6 February 2005 General Election. *Trends in Southeast Asia Series*. *6*(2005), 1-8.

Schafferer, C. (2009). The Parliamentary Election in Thailand, December 2007. *Electoral Studies*, 28(1), 167-170.

Surarit, Pontat (1996). *Improvement of the Constitution and the Electoral Law: A Case Study of the Electoral System* (Master's Thesis). Retrieved from ThaiLIB Digital Collection.

The Asia Foundation (2011). *2010 National Survey of the Thai Electorate: Exploring National Consensus and Color Polarization*. [Online]Available: www.asiafoundation.org/publications (October 12, 2011)

The Election Commission of Thailand (2011). *Information, Statistics and the Result of the House of Representative Elections in Many Years*. [Online]Available: http://www.ect.go.th/newweb/th/election/ (November 11, 2011)

Yamamoto, D., Ozeki, S., & Takahashi, N. (2009). *Wired Fisheye Lens: A Motion-based Improved Fisheye Interface for Mobile Web Map Services*. Presented at the conference, Web and Wireless Geographical Information Systems, Maynooth, Ireland.

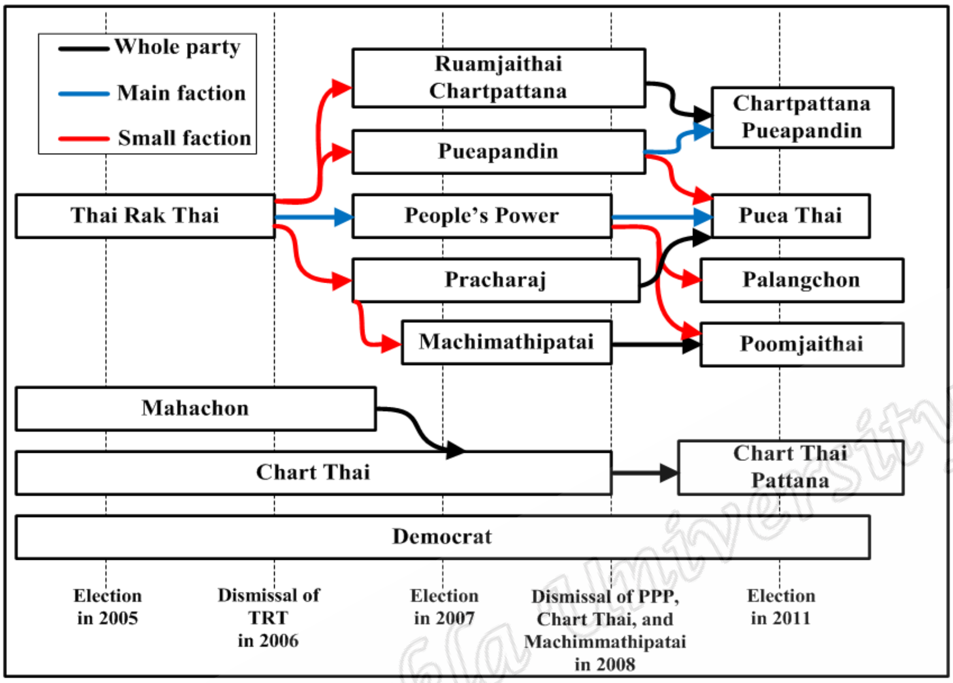


Figure 1: Timeline of changes in main political parties in Thailand from 2005-2011

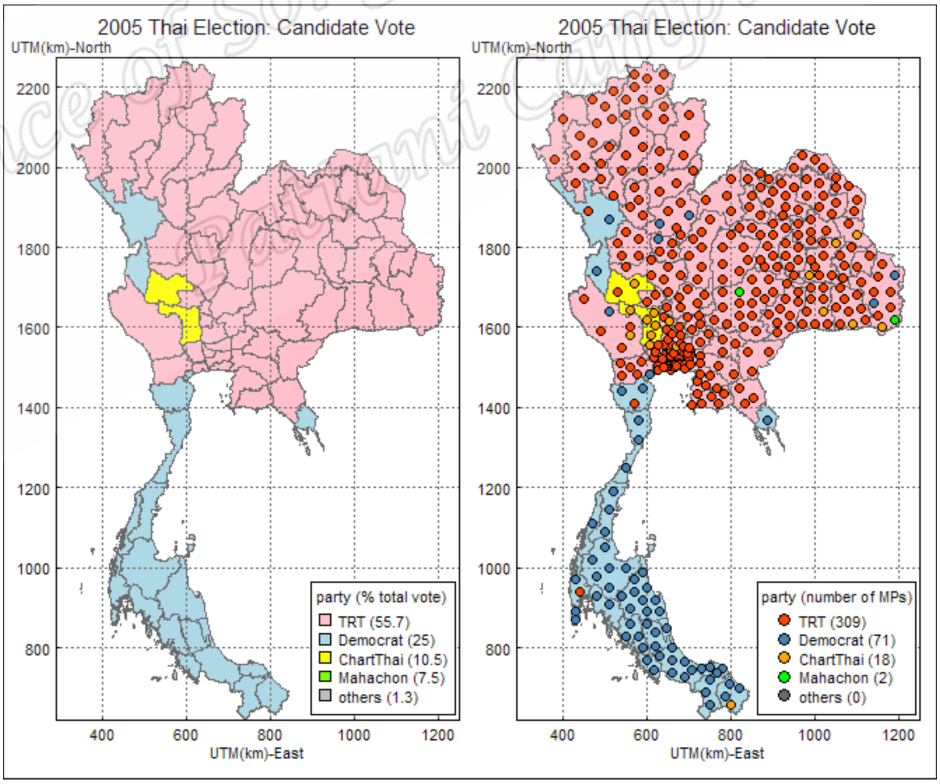


Figure 2: Maps of Thailand with the coloring themes for parties winning percentages of votes, and colored dots of parties winning MPs in each province.

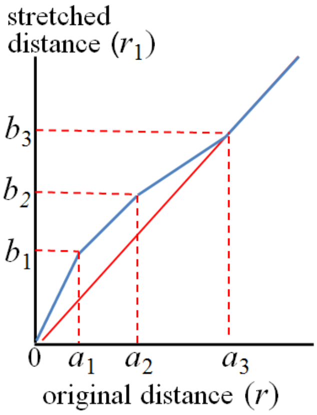
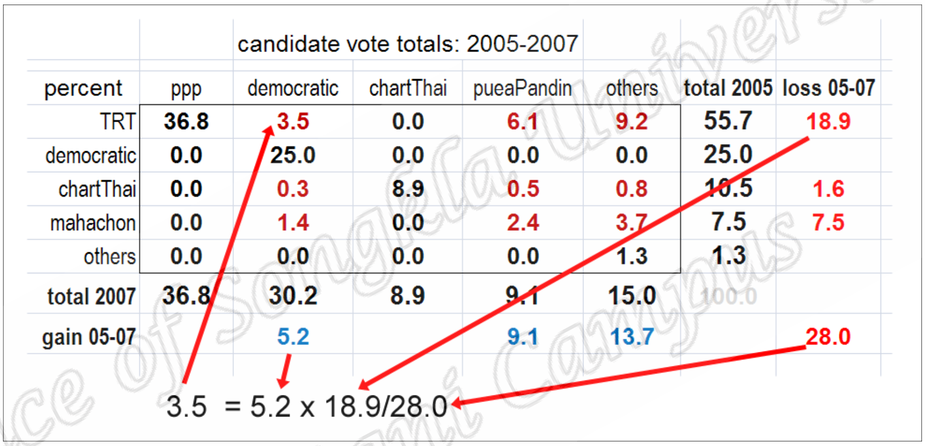


Figure 3: Radial scaling used in creating variable-scale map.



candidate vote totals: 2005-2007

percent	ppp	democratic	chartThai	pueaPandin	others	total 2005	loss 05-07
TRT	36.8	3.5	0.0	6.1	9.2	55.7	18.9
democratic	0.0	25.0	0.0	0.0	0.0	25.0	
chartThai	0.0	0.3	8.9	0.5	0.8	10.5	1.6
mahachon	0.0	1.4	0.0	2.4	3.7	7.5	7.5
others	0.0	0.0	0.0	0.0	1.3	1.3	
total 2007	36.8	30.2	8.9	9.1	15.0	100.0	
gain 05-07		5.2		9.1	13.7		28.0

3.5 = 5.2 x 18.9/28.0

Figure 4: A simple calculation for the proportion of votes swings between two elections

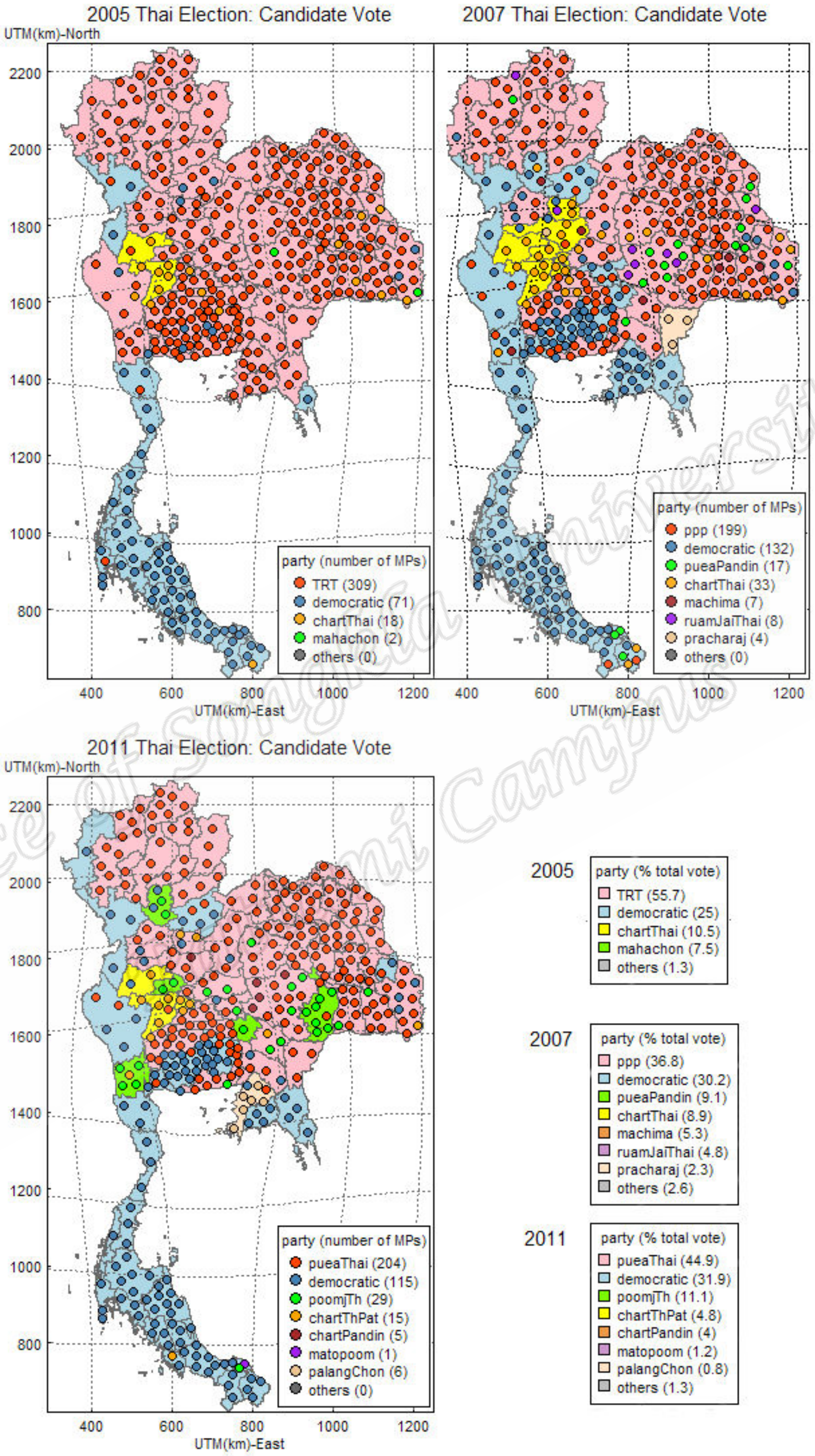
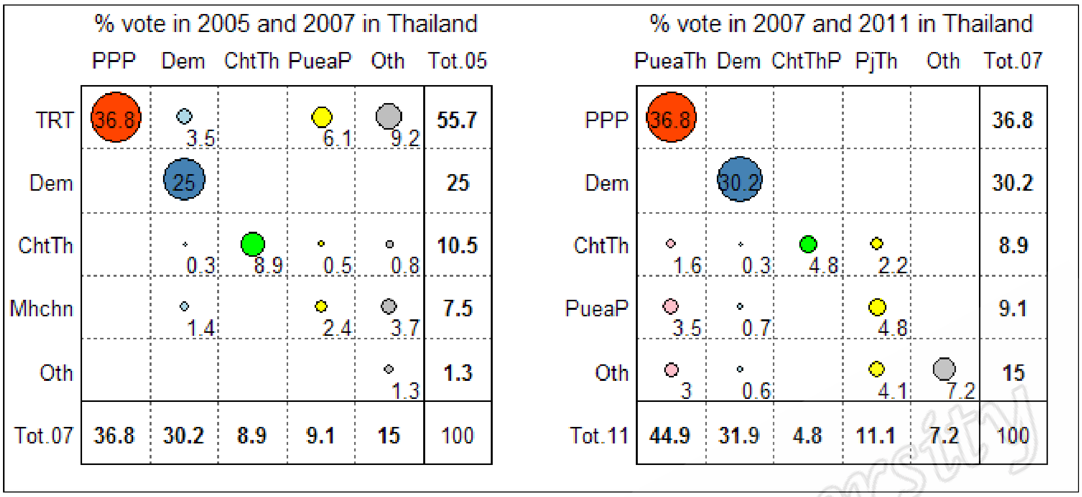


Figure 5: Representation of the numbers of MPs in a form of colored dots, comparing majority votes using background colors for each province in 2005, 2007, and 2011 elections



% vote in 2005 and 2007 in Thailand

	PPP	Dem	ChtTh	PueaP	Oth	Tot.05
TRT	36.8	3.5		6.1	9.2	**55.7**
Dem		25				**25**
ChtTh		0.3	8.9	0.5	0.8	**10.5**
Mhchn		1.4		2.4	3.7	**7.5**
Oth					1.3	**1.3**
Tot.07	**36.8**	**30.2**	**8.9**	**9.1**	**15**	100

% vote in 2007 and 2011 in Thailand

	PueaTh	Dem	ChtThP	PjTh	Oth	Tot.07
PPP	36.8					**36.8**
Dem		30.2				**30.2**
ChtTh	1.6	0.3	4.8	2.2		**8.9**
PueaP	3.5	0.7		4.8		**9.1**
Oth	3	0.6		4.1	7.2	**15**
Tot.11	**44.9**	**31.9**	**4.8**	**11.1**	**7.2**	100

Figure 6: The transferred percentages of votes between participating parties in Thailand in a pair of 2005 and 2007 elections, and in a pair of 2007 and 2011elections