Chapter 2

Methodology

In this chapter we describe data source, selected variables, data management, determinants, outcome, study sample, path diagram and statistical methods used for the analysis of youth non-participation either at work or at school in Pattani and Songkhla Provinces of Thailand.

2.1 Data source

The study is based on population data extracted from the 2000 Population and Housing Census of Pattani and Songkhla that were provided by the National Statistical Office.

The enumeration form number 2 is one of several forms that the National Statistical Office used to collect the data that was later used in our study. Table 2.1 shows variables from the enumeration form.

Table 2.1: selected variables in the 2000 Population and Housing Census

Variable	Meaning	Variable	Meaning
AMPHOE	12 Amphoe codes in Pattani 16 Amphoe codes in Songkhla	OCCUPATION	group 1 legislators senior officials and manager group 2 professional group 3 technicians and associate professionals group 4 clerks group 5 service workers and shop and market sales worker group 6 skilled agricultural and fishery workers group 7 craft and related trades workers group 8 plant and assemblers machine operators group 9 elementary occupations group 0 such as 0110 armed forces 9970 works not classifiable by occupation of unknown 9980 do not work 9999 unknown blank
TAMBON	115 tambon codes in Pattani 127 tambon codes in Songkhla		
SEX	1 = Male 2 = Female		
AGE	0-97 years 98 = years and over 99 = unknown		
RELIGION	1 = Buddhism 2 = Islam 3 = Christianity 4 = Hinduism 5 = Chinese Confucius 6 = Others 7 = No religion 9 = unknown		
GRADE OF SCHOOL ATTENDED	0 = Not attending school 1 = Pre-elementary level 2 = Elementary level 3 = Secondary level 4 = Higher level 5 = Religious education		

The raw data were retrieved using SPSS, and subsequently stored in an SQL database. SQL commands were used to create new variables (non-participation, super-tambon, age group, gender and religion). The non-participation rates were computed as the

number of cases per 100 residents in the super-tambons according to the 2000 Thai Population and Housing census. The mapping, graphical methods and statistical analysis were performed using the R program. The flow diagram for data management is summarized in Figure 2.1.

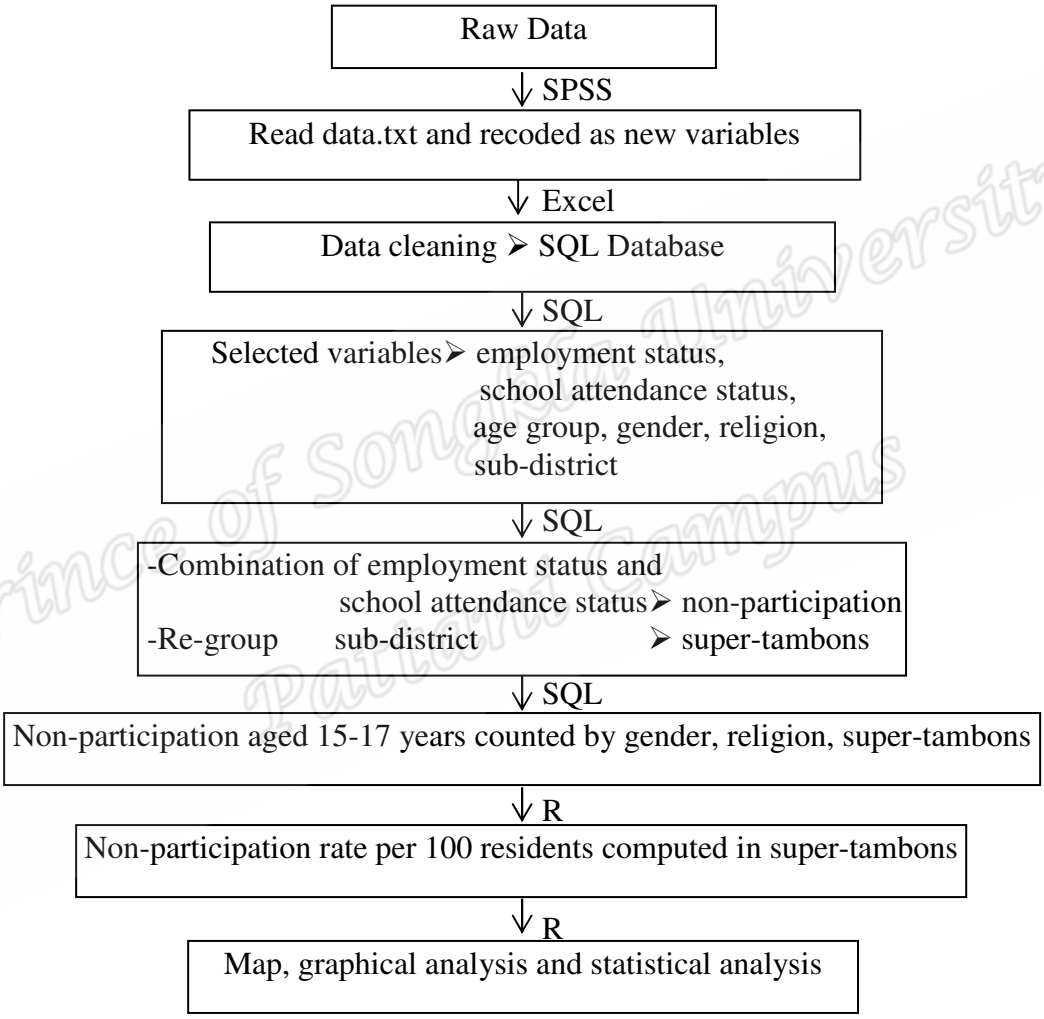


Figure 2.1: Flow diagram for data management

2.2 Selected variables

The selected variables from the raw data are gender, age, religion, and regions based on sub-districts of Pattani and Songkhla.

Gender: Male and Female

Age: The age groups were categorized as : 0-5, 6-11, 12-14, 15-17, 18-24, 25-59, 60-98, and not specified.

Religion: Religion was classified as 1=Islam (Muslims) and 2=Other religions (Non-Muslim, mainly Buddhist) or not specified.

Regions: The area comprises 34 regions in Pattani and 52 in Songkhla that are referred to as super-tambons. They were created from TAMBON.

Unemployment: For OCCUPATION, employed people were categorized by employment status (groups 1 to 9), and both the 0110 group, who were in the armed forces and the 9970 group who were not classified, or were of "unknown employment status", were placed in category 1 = "employed". Those in the 9980 group who did not work were placed in category 2 = "unemployed" and "others" were assigned to category 3 = blank/unknown.

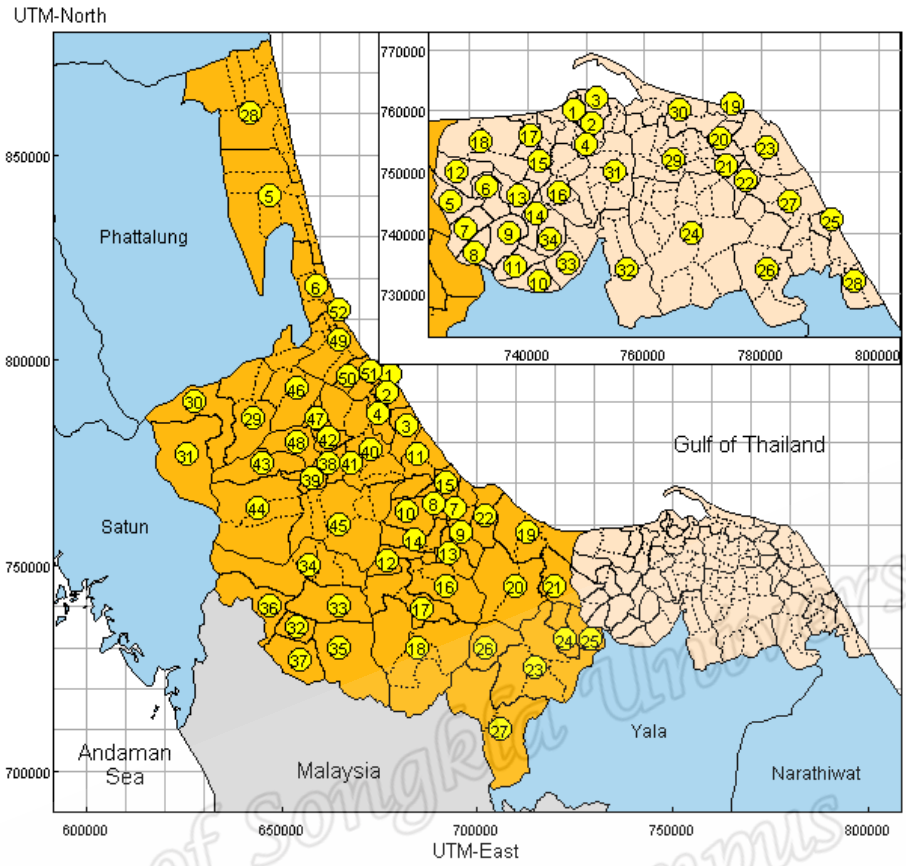
School non attendance: From GRADE OF SCHOOL ATTENDED, individuals were divided into 3 groups: 1=being educated (school attendance), 2=not being educated (school non-attendance), and 3=not specified or missing.

2.3 Data management

Super-tambons

The data obtained from the National Statistical Office were case by case, with 595,985 cases in Pattani and 1,255,662 cases in Songkhla recorded in text files.

Some subdistricts had very low populations and so they were not suitable for statistical comparison. Therefore some of them were grouped into super-tambons so that each had a minimum total population of 1,600 persons of all ages. This reduced the number of regions in Pattani from 115 tambons to 34 super-tambons. In Songkhla, we similarly reduced the number of regions from 127 tambons to 52 super-tambons. The bold lines in Figure 2.2 show the super-tambons.



Songkhla super-tambons name

1	BoYang	14	KlongPia+TalingChan	27	KaoDaeng+BaHoi	40	NHatYai+NNamom
2	KhaoRupChang	15	NamKhao+KhunTatWai	28	Ranot	41	KhoHong
3	KoTaeo+Tungwang	16	ENatawi	29	ERattaphum	42	KlongHae+KlongUtapao
4	Phawong+Koyo	17	W-CNathawi	30	THaChamuang	43	Chalung
5	NSatingPa+Krasasin	18	SNathawi	31	KhaoPhra	44	TungTamSao+WKhk
6	SSatingPa+BangKiat	19	Tepa+PakBang+KoSaba	32	Sadao	45	EKhk+SHatyai+SNamom
7	BanNa	20	LamPlai+WangYai	33	Prik	46	KhuanNiang
8	PaChing	21	ThaMuang	34	NSadao	47	NBangklam
9	SapanMaiKaen+Sakom	22	SaKom	35	SamNakTaeo	48	ThaChang
10	NaWa	23	CSabayoi	36	PadangBesa	49	NSingHaNakorn
11	ThaMoSai	24	Pian	37	SamnakKham	50	SathingMo
12	NaThap+Chanong	25	BanNot	38	HatYai	51	HuaKhao
13	Khu+Khae	26	Khuha	39	KhuanLang	52	MuangNgam

Pattani super-tambons name

1	Sabarang+Anuru	10	Pahlo	19	PanareCity	28	MaiKaen
2	Bana	11	ThungPhala	20	Thakumcham+Bn	29	SouthYaring
3	Jabangtiko+Talubo	12	ThaRua	21	Don	30	NorthYaring
4	Rusamilae+	13	NaKet	22	Khuan+Thanum	31	NorthYarang
5	KhokPro	14	KhuanNori	23	NorthPanare	32	SouthYarang
6	Makrut+Bangkro	15	Tuyong+BangTawa	24	Mayo+TYD	33	Maelan+MungTia
7	Pabon+Changhaitok	16	CentralNongChick	25	Taluban	34	Parai
8	Saikhao	17	SouthNongChick	27	EastSaiburi		
9	Napradu	18	WestNongChick	26	WestSaiburi+Karubi		

Figure 2.2: Map of 115 tambons together with 34 super-tambons in Pattani Province, and 127 tambons together with 52 super-tambons in Songkhla Province.

Determinants

The determinants considered were demographic factors (gender, religion) and a geographic factor (super-tambons).

Outcome

The outcome, a binary variable created from "unemployment" and "school non-attendance", used the "population at risk" as those not at work and not studying. To categorize all aged individuals we referred to the variable "school attendance" and "employment status" to be coded as either 1: "not at work and not at study" or 0: "at work or study or both". Table 2.2 shows cross tabulation of occupation and school attended.

Table 2.2: Cross tabulation of unemployment and school attendance for youth

School attended	Occupation		
	Employed	Unemployed	Not stated
Attending school	A	B	C
Not attending school	D	F	G
Not stated	E	H	I

Category 0: the "at work or study or both" group included:

A: persons who were attending school and were employed

B: persons who were attending school and had not been employed

C: persons who were attending school and had not stated their employment status

D: persons who were not attending school but were employed

E: persons who had not stated their school attendance status but were employed

Category 1: the "at risk" or "not at work and not at study" group included:

F: persons who were not attending school and were not employed

G: persons who were not attending school and had not stated their employment status

H: persons who had not stated their school attendance status and were not employed

I: persons who had not stated either their school attendance or their employment status

Table 2.3 shows the number of cases, of being "at work or at study" and of being "neither at work nor at study". The latter defines being "at risk" of being neither employed nor at school.

Table 2.3: Number of cases in "either school attendance or employment" and in neither, in Pattani and Songkhla, from Population and Housing Census, Thailand, 2000.

Pattani

Age Group	work or study						not at work and not at study				
	A	**B**	**C**	**D**	**E**	**Total**	**F**	**G**	**H**	**I**	**Total**
0-5	0	3	8745	4	13	**8765**	5	4281	1	63114	**67401**
6-11	0	23	75048	10	13	**75094**	4	3459	2	480	**3945**
12-14	75	18272	13757	775	5	**32884**	3390	1260	71	115	**4836**
15-17	297	20025	1181	5249	22	**26774**	7669	400	155	24	**8248**
18-24	753	15986	831	36771	192	**54533**	18599	1010	228	43	**19880**
25-59	493	937	66	45709	153846	**201051**	10351	670	22878	1560	**35459**
60-98	0	21	50	32	27964	**28067**	8	3	27422	1010	**28443**
Unknown	0	37	63	19	257	**376**	5	10	84	130	**229**
Total	**1618**	**55304**	**99741**	**88569**	**182312**	**427544**	**40031**	**11093**	**50841**	**66476**	**168441**

Songkhla

0-5	2	9	16671	11	51	**16744**	2	4406	16	102880	**107304**
6-11	2	83	126863	36	39	**127023**	14	4445	15	1090	**5564**
12-14	167	36066	25737	1531	28	**63529**	2931	1575	217	301	**5024**
15-17	715	46617	3186	10469	85	**61072**	8115	905	455	74	**9549**
18-24	2015	46076	2838	78812	537	**130278**	29884	2687	994	167	**33732**
25-59	1361	2134	167	107420	378502	**489584**	18685	2165	51008	6741	**78599**
60-98	9	82	105	84	57723	**58003**	19	11	61706	3655	**65391**
Unknown	30	554	551	275	1343	**2753**	81	35	557	840	**1513**
Total	**4301**	**131621**	**176118**	**198638**	**438308**	**948986**	**59731**	**16229**	**114968**	**115748**	**306676**

2.4 Study sample

Figure 2.3 shows the study sample and sample size in this study. We excluded persons who did not state their age and persons aged less than 15 years and greater than 17 years, giving a total study sample of 35,022 persons in Pattani and 70,621 in Songkhla. We focused on 15-17 year olds because neither studying nor working at this age is critical for a young person's future and persons of this age are normally expected to be studying at grades 9-12 or high school or vocational education.

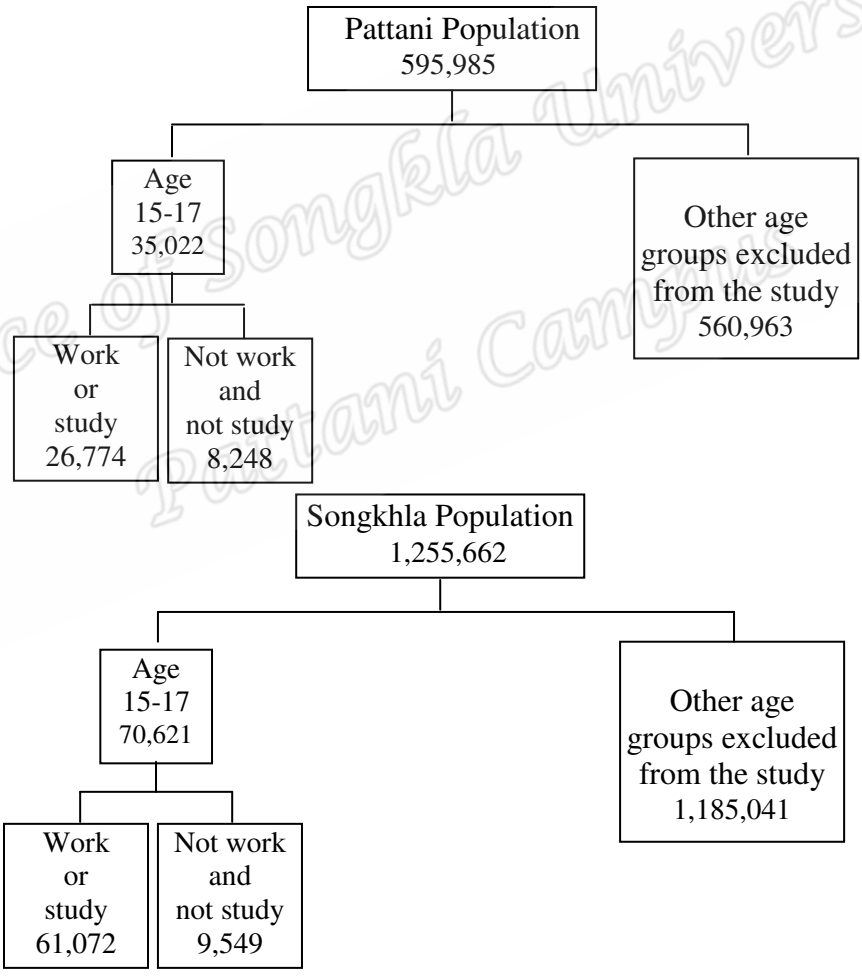


Figure 2.3: Study sample

2.5 Path diagrams

Figure 2.4 shows the path diagram of the study variables. The determinants considered were demographic factors including gender and religion, and the geographic factor. The outcome was a binary variable of either at work or at school.

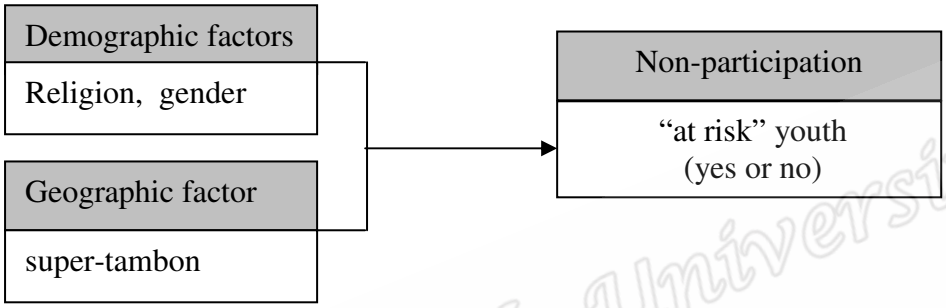


Figure 2.4: Path diagram for variables considered

2.6 Statistical methods

For the age group of 15-17 years, participation rates were identified, for all four demographic groups (two groups for religion and two for gender), for 34 super-tambons in Pattani and 52 super-tambons in Songkhla.

In our preliminary data analysis we compared the prevalence of non-participation within statistical regions by plotting these proportions separately for each combination of gender and religion using an area plot.

The prevalence of the adverse outcome may be modelled using logistic regression, which provides a method for modelling the association between a binary outcome and multiple determinants (see, for example, Kleinbaum and Klein, 2002). The simplest model takes the form

$$y_{ij} = \alpha_i + \beta_j, \tag{2.1}$$

where

$$y_{ij} = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right), \tag{2.2}$$

and p_{ij} denotes the probability of an adverse outcome in region *i* and religion-gender group *j*, where *j* takes values 1 for Muslim males, 2 for Muslim females, 3 for non-Muslim males and 4 for non-Muslim females. The terms α_i and β_j thus represent effects associated with region *i* and demographic group *j*, where the demographic effects are scaled to have mean 0 to avoid overparametrization. Equation (2.1) can be inverted to give the probability of the adverse outcome

$$p_{ij} = \frac{1}{1+\exp(-\alpha_i - \beta_j)}. \tag{2.3}$$

The logistic regression model was fitted to the counts in cells defined by combinations of demographic group and region, and the adequacy of the model was assessed by comparing the residual deviance with the number of degrees of freedom, and also by examining the linearity in the plot of deviance residuals against normal quantiles (Venables and Ripley 2002, Chapter 7).

To allow for possible interactions between region and demographic group, model (2.1) may be extended to a more general multiplicative model of the form

$$y_{ij} = \alpha_i + \gamma_i \beta_j. \tag{2.4}$$

In this model the demographic parameters are scaled to have unit variance as well as mean 0. The additional parameters γ_i provide a measure of the disparity in the adverse event rate between the different demographic groups in region *i*. Thus if region *i* has $\gamma_i = 0$, it means that there is no difference in the school adverse event rates between

demographic groups in this region, whereas if γ_i is large in magnitude there is a high disparity between these groups. We call this the *disparity index.*

Model (2.4) is non-linear and thus cannot be fitted simply using regression. However, Theil (1983) showed that the least squares estimates of the β_j parameters in model (2.4) are the elements of the eigenvector of the matrix $Y_c^T Y_c$ corresponding to its largest eigenvalue, where Y_c is the matrix with elements $y_{ij} - \bar{y}_i$ and Y^T denotes the transpose of *Y*. The corresponding least squares estimates of the γ_i parameters are then expressed in terms of the eigenvectors giving us this definition of the *disparity index*:

$$\gamma_i = \sum_{j=1}^{4} \beta_j \left(y_{ij} - \bar{y}_i \right) . \qquad (2.5)$$

Since the vector β_j is scaled to have mean 0 and standard deviation 1, it has only two free parameters. If these parameters are regarded as fixed, the model (2.4) can be fitted using standard linear regression, which provides both estimates and standard errors for the remaining parameters. The number of such parameters is thus 2*n* where *n* is the number of regions.

Model (2.4) thus contains a pair of parameters (α_i, γ_i) for each region, where α_i is the proportion of non-participating subjects and γ_i is the disparity index measuring the extent to which different demographic groups have different non-participation rates. This model has been used extensively for mortality forecasting in population science, where it is known as the Lee-Carter model (see, for example, Lee and Carter

1992, Booth et al 2002). In this research y_{ij} is the logarithm of the mortality rate in a population where the indexes *i* and *j* refer to age group and year, respectively.

To allow for values of p_{ij} equal to 0 or 1, a small constant *d* is added to the numerator and denominator in Equation (2.2) before log-transforming. Having estimated the vector β_j in model (2.4) by least squares we then used logistic regression to estimate the parameters (α_i, γ_i) for each region. Sum contrasts (Venables and Ripley 2002, Chapter 6) were used instead of the standard treatment contrasts, so that the standard error for each α_i parameter provided a confidence interval for the difference between the non-participation rates in region *i* and the overall mean participation rate in the study area.

The final results are displayed as a scatter plot of the estimated parameters (α_i, γ_i) for the regions. This plot shows the pattern of both non-participation and disparity in the study area. Using the standard errors estimated from the logistic regression model, the regions may then be classified into groups according to whether the confidence intervals for the non-participation rates exceed, contain, or fall below the overall mean, and according to whether the confidence intervals for the disparity indexes exceed zero, contain zero, or fall below zero.