A DATA MINING APPROACH TO THAILAND URBANIZATION INDEX DEVELOPMENT

Vichit Lorchirachoonkul, Jirawan Jitthavech and Duangpen Teerawanviwatp,
School of Applies Statistics, National Institute of Development Administration,
vichit@as.nida.ac.th; jirawan@as.nida.ac.th and duangpen@as.nida.ac.th

ABSTRACT

The city index is typically developed from the researcher's perspective and experience but in this paper the city index is developed from the data collected from local administrations in Thailand by data mining techniques. Evaluations show that the Logistic Regression and Decision Tree models can classify the local administration into city/town municipality, district municipality and district administration with the accuracy of more than 97.9 percent. The process of urbanization can be also observed from the study of classification.

INTRODUCTION

The topic of urbanization index has been studied from various perspectives. Education, health, social services, water supply, income, air pollution and transportation are typical components is urbanization index. The Habitat proposed a City Development Index–CDI for comparing city development around the world (UNCHS, 1998). The major drawback of CDI is due to unavailability of working data.

In order to improve the efficiency of the local administration in Thailand, the local administration financial data center was set up in 1999 by the cooperation of the Ministry of Interior and the National Economic and Social Development Board. All local administration units, provincial administrations, municipalities and district administrations, enter both basic and financial data directly via the virtual private network into the database server at the center. The basic data covers 489 variables including population–age structure, number of households, number or agricultural households, geographical area under jurisdiction, number of hotels, number of gasoline stations, number of schools and students, number of hospitals and medical personnels, number of households with water supply utility, availability of infrastructures, such as fresh food market, post office, fire–station, slaughter house, telephone, radio broadcasting station, rice mill, vocational education institutes, etc. These data are suited for developing urbanization index. Beside working out the index, the research team takes the opportunity to study the process of urbanization as well.

This research is aimed at distinguishing difficult characteristics between urban and rural areas and understanding the urbanization process in Thailand by applying data mining techniques to data collected in the database server at the local administration financial data center. Two models, Logistic Regression and Decision Tree are to be developed to classify the local administration in Thailand into city/town municipality, district municipality and district administration.

LITERATURE REVIEW

Urbanization, industrialization, and demographic and socio-economic development are three interrelated processes of advancement. They have to be considered correspondingly as they are illustrated in the establishment of the City Development Index by the Habitat

of UNCHS. It employs 23 quantitative criteria and 6 qualitative ones to cover 20 most important characteristics of all excellent cities worldwide, and classifies them into 6 categories, as shown in the following table.

Categories	**Characteristics**	**Criteria**
Residence	1. Assurance of steady residence	1. Rental or ownership of residence 2. Possibility of being expelled
	2. Promotion of individual right in obtaining residence	a. Privilege in residence 3. Income in proportion to residence cost
	3. Promotion of equal opportunity in possessing land	4. Income in proportion to land value
	4. Promotion of equality in obtaining loan	5. Mortgage and non-mortgage
	5. Promotion of equality in obtaining fundamental utility and facility	6. Accessibility to water resources 7. Households connection
Social Development and Poverty Elimination	6. Equal right to achieve a healthy and safe life	8. Deceased rate of children under 5 years old 9. Crime rate
	7. Promotion of having strong and cooperative society, and encouragement of the less developed communities	b. Severity in urban area 10. The proportion of poor households
	8. Promoting equal male/female right to advance his/her residence	11. Gap between men and women
Environmental Management	9. Encouraging even geographic settlement	12. Population growth in urban area
	10. Effective management of water supply and utility system	13. Water consumption 14. Water price
	11. Pollution diminution in urban area	15. Air pollution 16. Sewage disposal 17. Refuse disposal
	12. Prevention of catastrophic loss as well as guarantee rehabilitation measures	c. Catastrophe prevention and alleviation measures toward public disaster
	13. Promotion of effective communication devoid of environmental detriment	18. Transportation time 19. Transportation system
	14. Initiating Agenda 21 in the communities as well as encouraging mechanism for creation and application of community environmental plan	d. Community environmental plan

Categories	characteristics	Criteria
Economy Development	15. Strengthening both small and micro businesses, especially those running by women	20. Non systematic employment
	16. Promoting business cooperation between government and individuals as well as strengthening employment opportunity	e. Business cooperation between state/government and individuals 21. Production of city/town 22. Unemployment
State/governmental Administration	17. Promotion of separate administration as well as strengthening strong local administration	f. Extent of separate administrative power
	18. Encouraging community participation in the administration	g. Community participation in making all the important decisions
	19. Ensuring transparency, authenticity, and affectivity in all levels of administration	h. Emphasis of transparency and authenticity 23. Budget of local administrations
International Cooperation	20. Encouraging international cooperation	i. Participation in international cooperation

* Criteria in alphabets are qualitative viewpoints

These criteria are applied to analyse the following six aspects of city development problem:

1. Urban poverty
2. Urban human development
3. City investment potential
4. Urban environment
5. Urban governance
6. Overall quality of urban life

Hugo (1997) suggested an eight-dimensional perspective to differentiate between urban and rural. He weighed mainly on the characteristics of population in these perspectives.

1. Economic status. Activities in rural area are of primary economy and/or supplying inputs to production, while city activities are of secondary economy and technological production.
2. Professional structure. Rural population are involved in agriculture and simple agricultural industry, whereas urban population are involved in sophisticated production and services.
3. Educational opportunity. Education in rural population is inferior to urban population.
4. Accessibility to essential services and infrastructure. Rural population has less accessibility to such facilities than urban population.
5. Population. Birth rate as well as deceased rate in rural population are greater than in urban population.
6. Politics. The attitude towards politics of rural population is conservative, whereas attitude of urban population tends to be liberal.

7. Races in population. Race in urban population is diverse, while race in rural population is simple.
8. Migration. Migration in city is a totally move-in pattern, while migration in country is a completely move-away pattern.

Patamasiriwatt and colleagues(2001) have defined two decision rules for identifying characteristic of urbanization.

Rule 1 which is based on the deduction that city is the center surrounded by rural area, and the development of city is an expansion from the center to the surrounding city sprawl. Moreover, city development is a stepwise process which starts from ruralization, suburbanization, and then urbanization.

Rule 2 which is applied to independently growing and rapidly established city through impetus of some kind of economic boom. An example for this is the change of Lan Krabue District in Kamphengphet Province following the discovery of local natural resources.

Patamasiriwatt and colleagues in their research have developed three criteria under the circumstances of rule 1 for classifying local administrations into city/town, suburb, and country as follows.

Criteria 1 The minimum population density is set at 1000 people per square kilometer.

Criteria 2 The maximum proportion of agricultural population is set at 50 percent.

Criteria 3 The minimum land use feature is set at 4 establishments out of the following 9 types of facility :

1. Slaughterhouse
2. Financial bank
3. Massive shopping center
4. Governmental institutions
5. College and university
6. Intercity public transport station
7. Hospital with at least 30 in–patients capacity
8. Secondary school
9. A minimum proportion of 50 percent road length is covered with either concrete or asphalt pavement.

Urbanization which is classified as city municipality, town municipality, or district municipality satisfies all three criteria of population density, occupation, and land use feature. Suburbanization established as municipality satisfies only the criterion of population density. Ruralization established as municipality does not meet the criterion of population density.

The criteria of Paramasiriwatt and colleagues are derived from 448 sample data of municipality, which comprise 7 city municipalities, 25 town municipalities, and 416 district municipalities.

From the review of literature it may be concluded that criteria for urbanization are stipulated through the researchers' knowledge of urbanization and are confirmed by the collected data.

METHODOLOGY

The local administration financial data center, which has been established under the project of efficient local financial administration development creates the database for keeping general information data, monetary transaction, as well as arrears of local administrations. These administrations include provincial administrations, city municipalities, town municipalities, district municipalities, district administrations, Bangkok Metropolitan, and the corporate town Phathaya. There are a total of 489 variables in the database; a part of which depicts geographical environment, population and services in the area, and the other part relates variables about financial status of the local administrations. Thus, it is an advantage to make use of these information for development of urbanization index in Thailand.

In year 2000, Thailand is made up of 7,953 local administrations, of which include 75 provincial administrations, 20 city municipalities, 76 town municipalities, 1,033 district municipalities, 6,747 district administrations, Bangkok Metropolitan, and corporate town Phathaya.

The four hundred and eighty nine variables are grouped into four types of variables.

1. Status quo of local administration
2. Details of income, classified by source and type of income.
3. Details of expenditure, grouped by monetary sources, categories of plans/projects, activities, and expenditure categories.
4. Details of properties, arrears, and cumulative funds.

The variables of status quo are further divided into many subgroups as follows.

1. General status	2. Economic status	3. Income collection
4. Finance discipline	5. Population participation	6. Handling skill development
7. Education	8. Water resources and sewage disposal	
9. Refuse disposal	10. Parks and social welfares	11.Communications
12. Public health	13.Miscellaneous affairs.	

As the database keeps too much detailed information, all the detailed descriptive variables are incorporated respectively into one typical variable. These typical variables are then modified to create more sensible variables; for examples, variable of number of population is modified into variable of population density, as well as variable of number of population per public employee. The adjustment has resulted in reducing the total number of variables to 476.

In the second step, the data file that contains the 476 variables is analyzed through Input Data Source in the SAS Enterprise Miner, to separate the variables into two groups : Interval Variables and Class Variables, and to estimate descriptive statistics of each variable. The percentage of lost data in the two groups of variables is shown in Table 1.
Three hundred and sixty nine variables which show more than 31 percent of lost data are repudiated. The remaining 107 variables are then revised by changing some into binary variables, and all income variables are incorporated into three corresponding variables, namely Collected Income, General Subsidy, and Specific Subsidy. Then these revised data are analyzed again through Input Data Source in the SAS Enterprise Miner. After the repudiation of variables that show more than 26 percent lost data, only thirty variables remain. The rectification has reduced the number of sample units to 5,121 from the total 7,867 municipalities and district administrations, which account for 65.02 percent.

Table 1 Percentage of Lost Data in the Two Groups of Variables

Percentage of Lost Data	Interval Variables	Class Variables
91 – 100	25	22
81-90	151	13
71-80	55	1
61-70	27	-
51-60	24	-
41-50	26	1
31-40	24	-
0-30	74	33
Total	406	70

The number of city/town municipality district municipality and district administration in the remaining 5,127 sample units can be summarized as follows.

1. City/Town municipality. 60 sample units are adopted from the total number of 96 administrations, which account for 62.50 percent of the total.
2. District municipality. 525 sample units are adopted from the total number of 1,033 administration which account for 50.82 percent of the total.
3. District administration. 4,536 sample units are adopted from the total number of 6,747 administrations, which account for 67.23 percent of the total.

The thirty variables employed in the analysis of the 5,121 sample units are as follows.

1. Population density per square kilometer
2. Number of population per public employee
3. Total population in the administrative area
4. Ratio of agricultural households over total households
5. Ratio of households equipped with water supply utility over total households
6. Number of population per medical person
7. General subsidy
8. Specific subsidy
9. Number of trained persons
10. Number of hotel enterprises
11. Number of gasoline stations
12. The shortest journey to town by vehicle(in minute)
13. Total collected income
14. Public road length
15. Number of hospitals
16. Degree in urbanization
17. Availability of slaughter house
18. Existence of river
19. Availability of telephone exchange
20. Availability of radio broadcasting station as well as other communicating stations
21. Availability of religious institutes
22. Availability of dam, barrage, pond and well
23. Availability of private mass transportation enterprise
24. Availability of state-owned mass transportation
25. Availability of stadiums
26. Availability of post office and telegraph station
27. Availability of fire station or fire fighting vehicle
28. Availability of rice mill
29. Availability of vocational education institution

30. Availability of fresh food market

By careful discernment of all the variable values, it is concluded that variables regarding population density, number of population per public employee, and general subsidy are three indispensable characteristics of local administrations. Thus, sample units lacking such information are repudiated. So the total 5,127 sample units are further reduced to 4,208, which account for 53.51 percent of the total. These include 46 sample units of city/town municipality, which account for 47.92 percent of the total; 406 sample units of district municipality, which account for 39.30 percent of the total; 3,756 sample units of district administration, which account for 55.67 percent of the total.

The data of 4,208 sample units are randomly divided into two unequal portions. The portion containing 70 percent sample units is analyzed by applying the SAS Enterprise Miner in order to develop the models through Logistic Regression method and Decision Tree method. The other portion of 30 percent sample units is reserved for validation of the models.

RESULTS OF ANALYSIS

The results of the analysis through Logistic Regression are as follows.

Probability of urbanization = P(urban) = $\frac{EXP(ARG1)}{1+EXP(ARG1)}$

Probability of suburbanization = P(suburban) = $\frac{EXP(ARG2)}{1+EXP(ARG2)}$ – P(urban)

Probability of ruralization = P(rural) = $1-\frac{EXP(ARG2)}{1+EXP(ARG2)}$

When ARG1 = –10.6519 + 0.669 * POP_DEN – 0.160 * POP_PUB_EMP
+ 0.4063 * GEN_SUBSIDY + 0.8252 * VOC_ED
+ 4.0562 *FIRE_FIGHTING_VEC + 1.4219 * POST_OFFICE
+ 1.2952 * MARKET

and ARG2 = ARG1 + 7.8776

POP_DEN	= Population density thousand persons per square kilometer
POP_PER_PUB_EMP	= Number of population per 10 public employees
GEN_SUBSIDY	= General subsidy, million Baht
VOC_ED	= Availability of vocational education institution
FIRE_FIGHTING_VEC	= Availability of fire station or fire fighting vehicle
POST_OFFICE	= Availability of post office and telegraph station
MARKET	= Availability of fresh food market

Classification of a sample unit is judged by the greatest value of probability. If the greatest value of probability of the sample unit is probability of urbanization, that unit will be classified as city/town municipality; if it is probability of suburbanization, that administration unit will be classified as district municipality; and if it is probability of ruralization, that administration unit will be classified as district administration.

This model is capable of making correct classification of 2,888 sample units out of a total of 2,946 training sample units, which account for 98.03 percent correctness. The accuracy

of the Logistic Regression model is tested by applying it with the 30 percent validation sample data, which consist of 1,262 sample units. It has proved to be capable of making 97.86 percent correct classification of the local administrations, which is comparable to the accuracy (98.03 percent correctness) obtained from the training data. The capability of the model in classifying local administrations in the training and validation data is shown in Table 2.

Table 2 The Number and Percentage of Local Administrations in the Training and Validation Data Classified by the Logistic Regression Model

Class of Local Administration	City/Town Municipality	District Municipality	District Administration	Total
Training Data				
City/Town Municipality	20 (60.61%)	13 (39.39%)	0 (0.0%)	33
District Municipality	7 (2.51%)	250 (89.61%)	22 (7.89%)	279
District Administration	0 (0.0%)	16 (0.61%)	2,618 (99.39%)	2,634
Total	27	279	2,640	2,946
Validation Data				
City/Town Municipality	10 (76.92%)	3 (23.08%)	0 (0.0%)	13
District Municipality	3 (2.36%)	119 (93.70%)	5 (3.94%)	127
District Administration	0 (0.0%)	16 (1.43%)	1,106 (98.57%)	1,122
Total	13	138	1,111	1,262

The Logistic Regression model depends on seven variables mentioned above. The first three variables, population density per square kilometer, number of population per public employee, and general subsidy are continuous variables; the other four variables, availability of vocational education institution, availability of fire station or fire fighting vehicles, availability of post office and telegraph station, and availability of fresh food market, are binary variables. In the fictitious representatives of local administration the average values in the corresponding set are used for continuous variables, and median values are used for binary variables except in the case of district administration, the typical values are set accordingly but in the subset of district administration with the probability of ruralization less than 0.9. The typical values of the fictitious city/town municipality, district municipality and district administration are shown in Table 3 together with the associated probabilities of urbanization, suburbanization and ruralization. In the typical district administration which is in the rural area, there is fresh–food market and post office, but absence of fire station and vocational education institute. The population density is not dense, only 161.68 persons per square kilometer. The public service is rare ; one public employee serves 182.46 customers. As the area becomes more urbanization, moving from ruralization to suburbanization, the fire station or fire fighting vehicle is established. A move–in pattern of migration takes place increasing the population density to 1,428.364 persons per square kilometer, the number of public employees also increases correspondingly, reducing the ratio of population per public employee to 111.55 and the general subsidy from the central government increases to 2.5794 million Baht per annum. In the urban area which is city/town municipality, all basic physical infrastructures exist,

the move–in pattern of migration becomes more evident increasing the population density to 3,819.80 persons per square kilometer, the number of public employees increases further, reducing the ratio of population per public employee to 58.22 and the subsidy from the central government increases to 9.8535 million Baht per annum.

Taber 3 Typical Values of Representatives of City/town Municipality, District Municipality and District Administration

Variable	City/town Municipality	District Municipality	District Administration
POP_DEN, thousand persons/sq.km	3.8198	1.4284	.1617
POP_PER_PUB_EMP persons/ten public employees	58.2152	111.5503	182.462
GEN_SUBSIDY, million Baht	9.8535	2.5794	1.2179
VOC_ED	1	0	0
FIRE_FIGHTING_VEC	1	1	0
POST_OFFICE	1	1	1
MARKET	1	1	1
P(urban)	0.9292	0.0251	0
P(suburban)	0.0708	0.9604	0.0852
P(rural)	0.0000	0.0145	0.9148

The rate of change in the probability of urbanization with respect to continuous variable x_j may be written as

$$\frac{\partial P(\text{urban})}{\partial x_j} = \frac{a_j * P(\text{urban})}{1+EXP(ARG1)}$$

where a_j is the coefficient of the continuous variable x_j in ARG1.

The probability of suburbanization is equal to

$$P(\text{suburban}) = \frac{EXP(ARG2)}{1+EXP(ARG2)} - P(\text{urban})$$

where $$ARG2 = ARG1 + b_2$$

and b_2 is the change in intercept of Logistic Regression when moving downward from urbanization classification to suburbanization classification.

The rate of change in the probability of suburbanization and of ruralization with respect to continuous variable x_j may be shown as

$$\frac{\partial P(\text{suburban})}{\partial x_j} = \frac{a_j * EXP(ARG2)}{(1+EXP(ARG2))^2} - \frac{\partial P(\text{urban})}{\partial x_j}$$

$$\frac{\partial P(\text{rural})}{\partial x_j} = -\frac{a_j * EXP(ARG2)}{(1+EXP(ARG2))^2}$$

It should be noticed that all coefficients of binary variables in ARG1 are positive. The absence of any basic infrastructure, i.e, changing in binary variable from 1 to 0. ($\Delta x_k = -1$), will decrease the probability of urbanization and vice versa. The incremental change in the probability of urbanization, suburbanization and ruralization with respect to binary variable x_k may be shown in general as

$$\Delta P\left(\text{urban}\left|\Delta x_k = -1\right.\right) = P\left(\text{urban}\right) * P\left(\text{urban}\left|\Delta x_k = -1\right.\right) * \frac{\left(1 - \text{EXP}(a_k)\right)}{\text{EXP(ARG1)}}$$

$$\Delta P\left(\text{urban}\left|\Delta x_k = 1\right.\right) = P\left(\text{urban}\right) * P\left(\text{urban}\left|\Delta x_k = 1\right.\right) * \frac{\left(\text{EXP}(a_k) - 1\right)}{\text{EXP(ARG1}+a_k)}$$

$$\Delta P\left(\text{suburban}\left|\Delta x_k = -1\right.\right) = \Delta P\left(\text{urban}\left|\Delta x_k = -1\right.\right) * \frac{\text{EXP}(b_2) - 1}{\text{EXP}(b_2)} * \frac{\text{EXP}\left(b_2\right) - \text{EXP}\left(2*\text{ARG2} - a_k\right)}{\left(1+\text{EXP}\left(\text{ARG2}\right)\right)\left(1+\text{EXP}\left(\text{ARG2} - a_k\right)\right)}$$

$$\Delta P\left(\text{suburban}\left|\Delta x_k = 1\right.\right) = \Delta P\left(\text{urban}\left|\Delta x_k = 1\right.\right) * \frac{\text{EXP}(b_2) - 1}{\text{EXP}(b_2)} * \frac{\text{EXP}\left(b_2\right) - \text{EXP}\left(2*\text{ARG2} + a_k\right)}{\left(1+\text{EXP}\left(\text{ARG2}\right)\right)\left(1+\text{EXP}\left(\text{ARG2} + a_k\right)\right)}$$

$$\Delta P\left(\text{rural}\left|\Delta x_k = -1\right.\right) = \frac{\text{EXP}(\text{ARG2} - a_k)\left(\text{EXP}\left(a_k\right) - 1\right)}{\left(1+\text{EXP}\left(\text{ARG2}\right)\right)\left(1+\text{EXP}(\text{ARG2} - a_k)\right)}$$

$$\Delta P\left(\text{rural}\left|\Delta x_k = 1\right.\right) = \frac{\text{EXP}(\text{ARG2})\left(1 - \text{EXP}\left(a_k\right)\right)}{\left(1+\text{EXP}\left(\text{ARG2}\right)\right)\left(1+\text{EXP}(\text{ARG2}+a_k)\right)}$$

where a_k is the coefficient the binary variable x_k in ARG1 or ARG2 depending on the presence or absence of the corresponding physical infrastructure.

The rates of change in probabilities with respect to continuous variables and changes in probabilities with respect to binary variables in the case of the fictitious city/town municipality, district municipality and district administration in Table 3 are shown in Table 4. Among the continuous variables, the rates of change in the probability of urbanization with respect to the population density at the typical city/town municipality, district municipality is highest. It suggests that encouragement of move–in migration may be the effective strategy to accelerate the urbanization process in the typical city/town municipality and district municipality. The same strategy can be applied in the district administration to effectively increase the probability of suburbanization.

Among the binary variables, the existence of fire station or fire fighting vehicle is crucial to the probability of urbanization. In case of city/town municipality in the absence of the facility, the probability of urbanization will decrease 0.7441 or 80.08 percent, and the probability of suburbanization will decrease 0.7424 or 10.51 times. In case of district municipality, if the fire station or fire fighting vehicle is absent, the probability to become suburbanization will decrease 0.4205 or 43.78 percent, and the probability of ruralization will increase 0.4452 or 30.66 times. In case of district administration, the presence of the fire station or fire fighting vehicle will increase, the probability of suburbanization 0.7560 or 8.87 times.

Table 4 Rates of Change in Probabilities

Variables	P(urban)	P(suburban)	P(rural)
City/Town Municipality			
POP_DEN	0.04440	–0.0440	0.0000
POP_PUB_EMP	–0.0011	0.0011	0.0000
GEN_SUBSIDY	0.0267	–0.0267	0.0000
VOC_ED $(\Delta x_k = -1)$	–0.0774	–0.0773	0.0000
FIRE_FIGHTING_VEC $(\Delta x_k = -1)$	–0.7441	–0.7424	0.0016
POST_OFFICE $(\Delta x_k = -1)$	–0.1693	–0.1692	0.0001
MARKET $(\Delta x_k = -1)$	–0.14698	–0.1468	0.0001
District Municipality			
POP_DEN	0.01637	–0.00679	–0.00957
POP_PUB_EMP	–0.00039	0.00016	0.00023
GEN_SUBSIDY	0.00994	–0.00413	–0.00581
VOC_ED $(\Delta x_k = 1)$	0.0304	–0.0223	–0.0081
FIRE_FIGHTING_VEC $(\Delta x_k = -1)$	–0.0246	–0.4205	0.4452
POST_OFFICE $(\Delta x_k = -1)$	–0.0189	–0.241	0.0430
MARKET $(\Delta x_k = -1)$	–0.0181	–0.0184	0.0365
District Administration			
POP_DEN	0.0000	0.0521	–0.0521
POP_PUB_EMP	0.0000	–0.0012	0.0012
GEN_SUBSIDY	0.0000	0.0317	–0.0317
VOC_ED $(\Delta x_k = 1)$	0.0005	0.0901	–0.0901
FIRE_FIGHTING_VEC $(\Delta x_k = 1)$	0.0020	0.7560	–0.7580
POST_OFFICE $(\Delta x_k = -1)$	–0.0003	–0.0632	0.0632
MARKET $(\Delta x_k = -1)$	–0.0003	–0.0603	0.0603

THE DECISION TREE

The SAS Enterprise Miner has created the Decision Tree to classify different types of local administrations from the same sample data of 4,208 units, which have been employed in the establishment of the Logistic Regression Model. The Decision Tree model applies only four variables : availability of fire station or fire fighting vehicle, general subsidy, population density per square kilometer and number of population per public employee, accompanying with the following seven decision rules :

Rule 1 : If neither fire station nor fire fighting vehicle is available, and population density is less than 714.6675 persons per square kilometer, the sample unit will be classified as district administration.

Rule 2 : If neither fire station nor fire fighting vehicle is available, population density is not less than 714.6675 persons per square kilometer, and general subsidy is less than 1.530 million Baht, the sample unit will be classified as district administration.

Rule 3 : If neither fire station nor fire fighting vehicle is available, population density is not less than 714.6675 persons per square kilometer, and general subsidy is not less than 1.530 million Baht, the sample unit will be classified as district municipality.

Rule 4 : If fire station or fire fighting vehicle is available, general subsidy is less than 1.6178 million Baht, and ratio of population per public employee is less than 69.3380, the sample unit will be classified as district municipality.
Rule 5 : If fire station or fire fighting vehicle is available, general subsidy is less than 1.6178 million Baht, but ratio of population per public employee is not less than 69.338, the sample unit will be classified as district administration.
Rule 6 : If fire station or fire fighting vehicle is available, general subsidy is not less than 1.6178 million Baht but not more than 6.0153 million Baht, the sample unit will be classified as district municipality.
Rule 7 : If fire station or fire fighting vehicle is available, general subsidy is not less than 6.0153 million Baht, the sample will be classified as city/town municipality.

These seven decision rules are capable of making correct classification of 2,897 sample units out of the total 2,946 training sample units, which account for 98.34 percent correctness. There are only 49 sample units being misclassified, which account for 1.66 percent. When these 7 decision rules are applied for validation in the reserved 1,262 validation sample units, they are proved to be capable of classifying 1,243 sample units correctly, which account for 98.49 percent correctness. Classification details of local administrations in the training and validation data by the Decision Tree model through the 7 decision rules are shown in Table 5.

SUMMARY

The characteristics of urbanization, suburbanization, and ruralization have been established through data analysis of general circumstances of respective city/town municipality, district municipality, and district administration in Thailand by data mining techniques with stepwise method in the developing process of the Logistic Regression model, and with Gini Reduction in the Decision Tree model. Variables are selected from the thirty refined variables through statistical significance criteria of data analysis in reference to general circumstances of local administrations. The correctness of the two established model in classifying the local administrations is shown in Table 6.

The overall accuracy of both models is comparable, 97.98 to 98.38 percent correctness. However, if the respective type of administrations is considered separately, the Logistic Regression model shows better correction in city/town municipality classification than the Decision Tree model; on the contrary, the Decision Tree model is slightly superior to the Logistic Regression model in classifying district municipality. In classification of district administration, the capability of both models is similar.

Seven variables at 10 percent significance level are being employed in the Logistic Regression model: population density per square kilometer, number of population per public employee, general subsidy, availability of vocational education institution, availability of fire station or fire fighting vehicle, availability of post office and telegraph station and availability of fresh food market.

The Decision Tree model is created by applying only four variables, also at 10 percent significance level, which are availability of fire station or fire fighting vehicles, general subsidy, population density per square kilometer and number of population per public employee. These four variables are formulated into 7 decision rules with defined threshold values of the variables to facilitate the classification.

Table 5 The Number of Local Administrations in the Training Data Classified by the Decision Tree Model Through the 7 Decision Rules

	City/Town Municipality	District Municipality	District Administration	Total
Training Data				
Rule 1	0	8	2,492	2,500
Rule 2	0	1	42	43
Rule 3	1	8	2	11
Rule 4	0	4	1	5
Rule 5	0	2	84	86
Rule 6	13	248	13	274
Rule 7	19	8	0	27
Total	33	279	2,634	2,946
Validation Data				
Rule 1	0	1	1,048	1,049
Rule 2	0	0	20	20
Rule 3	0	4	1	5
Rule 4	0	5	0	5
Rule 5	0	0	44	44
Rule 6	5	114	9	128
Rule 7	8	3	0	11
Total	13	127	1,122	1,262

Table 6 Comparison of Classification Accuracy in the Logistic Regression and Decision Models

Class of Local Administrations	Training Data			Validation Data			Total		
	Correct	Incorrect	Total	Correct	Incorrect	Total	Correct	Incorrect	Total
Logistic Regression Model									
City/Town	20	13	33	10	3	13	30	16	46
Municipality	(60.61%)	(39.39%)		(76.92%)	(23.08%)		(65.22%)	(34.78%)	
District	250	29	279	119	8	127	369	37	406
Municipality	(89.61%)	(10.39%)		(93.70%)	(6.30%)		(90.89%)	(9.11%)	
District	2618	16	2634	1106	16	1122	3724	32	3756
Administration	(99.39%)	(0.61%)		(98.57%)	(1.43%)		(99.15%)	(0.85%)	
Total	2888	58	2946	1235	27	1262	4123	85	4208
	(98.03%)	(1.97%)		(97.86%)	(2.14%)		(97.97%)	(2.02%)	
Decision Tree Model									
City/Town	19	14	33	8	5	13	27	19	46
Municipality	(57.58%)	(42.42%)		(61.54%)	(38.46%)		(58.70%)	(41.30%)	
District	260	19	279	123	4	127	383	23	406
Municipality	(93.19%)	(6.81%)		(96.85%)	(3.15%)		(94.33%)	(5.67%)	
District	2618	16	2634	1112	10	1122	3730	26	3756
Administration	(99.39%)	(0.61%)		(99.11%)	(0.89%)		(99.31%)	(0.69%)	
Total	2897	49	2946	1243	19	1262	4140	68	4208
	(98.34%)	(1.66%)		(98.49%)	(1.51%)		(98.38%)	(1.62%)	

Investigation of all the misclassified cases shows that almost all the misclassified city/town municipalities are small town municipalities; they are more like suburbanization than urbanization. They become town municipalities simply because they are the sites of

provincial town hall. For cases of misclassified district municipalities, most of them are rather large district municipalities; they are more similar to urbanization than sub–urbanization.

Moreover, the urbanization probability of local administrations can serve as a leading indicator to point out promising places that can become urbanization. If a district municipality is in the class of suburbanization, its urbanization probability value is less than 0.5. If its urbanization probability value comes close to 0.5, it indicates that this district municipality is eventually qualified to become town municipality.

The analysis also shows the process of urbanization quite clearly. For example, when the district administration becomes more urbanization, the fire station will be constructed, changing the local administration to district municipality and when the district municipality becomes more urbanization the vocational education institute will be established, changing the district municipality to town municipality.

BIBIOGRAPHY

Fay, Marianne and Charlotte Opal (2000) **"Urbanization without Growth : A not so uncommon Phenomenon"** prepared for the Summer Seminar at the World Bank.

http://www.un.org

http://www.unchs.org

Hugo, G (1977) **Rethinking the ASGC: Conceptual and Practical Issues** Monograph Series 3, National Key Centre for the Social Application of Geographic Information Systems, University of Adelaide.

Morrison, Peter A. ed. (1989) **Population movements : Their Forms and Functions in Urbanization and Development** Liege, Belgium : Ordina.

Patamasiriwatt and colleagues (2001) **Efficiency Indicator of Local Administration Performance**, Research Report submit tea to National Economic and Social Development Board. (in Thai).