GEOINFORMATICS AND PREDICTIVE MODEL FOR MALARIA RISK IN THAILAND

Miss Patcharaporn Krainara

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Bioinformatics and Computational Biology
Inter-Department of Bioinformatics and Computational Biology
Graduate School
Chulalongkorn University
Academic Year 2019

ภูมิสารสนเทศและตัวแบบการพยากรณ์ความเสี่ยงของมาลาเรียในประเทศไทย

น.ส.พัชราภรณ์ ไกรนรา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาชีวสารสนเทศศาสตร์และชีววิทยาเชิงคอมพิวเตอร์ สหสาขาวิชาชีวสารสนเทศศาสตร์และ
ชีววิทยาทางคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Thesis Title	GEOINFORMATICS AND PREDICTIVE MODEL FOR MALARIA RISK IN THAILAND
By	Miss Patcharaporn Krainara
Field of Study	Bioinformatics and Computational Biology
Thesis Advisor	Assistant Professor PONGCHAI DUMRONGROJWATTHANA, Ph.D.
Thesis Co Advisor	Associate Professor PATTARASINEE BHATTARAKOSOL, Ph.D.

Accepted by the Graduate School, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

.... Dean of the Graduate School

(Associate Professor THUMNOON NHUJAK, Ph.D.)

DISSERTATION COMMITTEE

.... Chairman

(Associate Professor TEERAPONG BUABOOCHA, Ph.D.)

.... Thesis Advisor

(Assistant Professor PONGCHAI DUMRONGROJWATTHANA, Ph.D.)

.... Thesis Co-Advisor

(Associate Professor PATTARASINEE BHATTARAKOSOL, Ph.D.)

.... Examiner

(Assistant Professor KITIPORN PLAIMAS, Ph.D.)

.... Examiner

(Assistant Professor SITTIPORN PATTARADILOKRAT, Ph.D.)

.... External Examiner

(Associate Professor Panjai Tantatsanawong, Ph.D.)

พัชราภรณ์ ไกรนรา : ภูมิสารสนเทศและตัวแบบการพยากรณ์ความเสี่ยงของมาลาเรียในประเทศไทย. (GEOINFORMATICS AND PREDICTIVE MODEL FOR MALARIA RISK IN THAILAND) อ.ที่ปรึกษาหลัก : ผศ. ดร.พงษ์ชัย ดำรงโรจน์วัฒนา, อ.ที่ปรึกษาร่วม : รศ. ดร.ภัทรสินี ภัทรโกศล

มาลาเรียเป็นโรคติดเชื้อที่เกิดจากปรสิต *Plasmodium spp.* และแพร่กระจายโดยยุงก้นปล่องเป็นพาหะ หลังจากยุงก้นปล่องเพศเมียที่ติดเชื้อกัดคนแล้วปรสิตจากต่อมน้ำลายจะถูกส่งไปยังกระแสเลือดของมนุษย์ ระยะฟักตัวของโรคอยู่ระหว่าง 10-14 วันและในกรณีที่รุนแรงอาจทำให้เกิดโรคดีซ่าน ชัก โคม่าหรือเสียชีวิตได้ งานวิจัยหลายชิ้นระบุว่าการเกิดโรคมาลาเรียต้องอาศัยปัจจัยหลายอย่างรวมถึงปัจจัยด้านสิ่งแวดล้อม ภูมิทัศน์ และภูมิอากาศ ดังนั้นหน่วยงานด้านการดูแลสุขภาพทุกแห่งจึงส่งมอบนโยบายเพื่อควบคุมปัจจัยเหล่านี้ที่มีต่อวิถีชีวิตของมนุษย์ แม้ว่ามาลาเรียจะหายไปนานหลายสิบปีแล้ว แต่ก็มีปัญหาสำคัญอย่างหนึ่งที่เกิดขึ้นกับปรสิตมาลาเรียที่ต่อต้านการใช้ยาหลายชนิด ดังนั้นจึงทำให้เกิดความซับซ้อนในกระบวนการควบคุมมาลาเรีย และเพื่อเป็นการแก้ไขปัญหาที่เกิดขึ้นนี้ การศึกษานี้ได้รวบรวมข้อมูลส่วนใหญ่จากฐานข้อมูลสำนักงานสถิติแห่งชาติและประยุกต์ใช้วิธีการทางสถิติ ร่วมกับการใช้เทคนิคการเรียนรู้ด้วยเครื่องและระบบสารสนเทศภูมิศาสตร์เพื่อกำหนดปัจจัยเสี่ยง รูปแบบความเสี่ยง และแผนที่ความเสี่ยง สำหรับการแพร่กระจายของโรคมาลาเรียในประเทศไทย ซึ่งผลจากวิธีการเหล่านี้ทำให้การศึกษานี้เสนอชุดพารามิเตอร์การติดตามใหม่ 28 พารามิเตอร์ โมเดลความเสี่ยงต้นไม้แบบลอจิสติก (LMT) และแผนที่ความเสี่ยงสำหรับการกระจายของโรคมาลาเรียที่ได้จากกลไกการเรียนรู้ของเครื่อง ค่าความแม่นยำและค่าความครบถ้วนของต้นไม้นี้คือ 0.780 และ 0.821 ตามลำดับ ซึ่งสูงกว่าการใช้แบบจำลองความเสี่ยงที่สร้างขึ้นจากปัจจัยทั่วไป ผลจากการค้นพบนี้ รัฐบาลสามารถใช้ปัจจัยที่ระบุใหม่เหล่านี้เพื่อพัฒนากลยุทธ์ที่เหมาะสมในการควบคุมโรคมาลาเรียโดยพิจารณาตามเกณฑ์ที่เหมาะสม ตัวอย่างเช่น รัฐบาลควรควบคุมคุณภาพของแม่น้ำหรือเพิ่มจำนวนครั้งในการติดตามแม่น้ำในบางพื้นที่เพื่อลดความเสี่ยงของโรคมาลาเรีย เมื่อปัจจัยเหล่านี้ได้รับการบริหารจัดการอย่างสมบูรณ์พร้อมกับการมีแผนพัฒนาที่ดี ส่งผลให้แต่ละชุมชนสามารถลดความเสี่ยงของโรคมาลาเรียได้โดยไม่ต้องกังวลเกี่ยวกับการดื้อยา ยิ่งไปกว่านั้น จากการที่แผนที่ความเสี่ยงแต่ละพื้นที่ในประเทศไทยถูกกำหนดว่ามีการกระจายของมาลาเรียในระดับสูง ปานกลาง และระดับต่ำภายใต้ปัจจัยต่าง ๆ ที่ได้จากการค้นพบนี้ ทำให้รัฐบาลสามารถกำหนดกลยุทธ์การติดตามและป้องกันโรคมาลาเรียที่เหมาะสมสำหรับแต่ละพื้นที่พร้อมกับการจัดสรรงบประมาณที่เป็นไปอย่างเหมาะสม

สาขาวิชา	ชีวสารสนเทศศาสตร์และ ชีววิทยาเชิงคอมพิวเตอร์	ลายมือชื่อนิสิต ....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

5987834720 : MAJOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

KEYWORD: Malaria, Risk factor, Risk Model

Patcharaporn Krainara : GEOINFORMATICS AND PREDICTIVE MODEL FOR MALARIA RISK IN THAILAND. Advisor: Asst. Prof. PONGCHAI DUMRONGROJWATTHANA, Ph.D. Co-advisor: Assoc. Prof. PATTARASINEE BHATTARAKOSOL, Ph.D.

Malaria is an infectious disease caused by the parasite *Plasmodium* spp., and transmitted by an anopheles mosquito as a vector. After an infected female Anopheles mosquito bites a person, a parasite from salivary glands is delivered into the human’s blood flow. The incubation period of the disease is between 10-14 days and in the severe cases, it can cause jaundice, seizures, coma, or death. Many researches indicated that the occurrence of malaria requires many factors, including environmental factors, landscape and climate. Therefore, every healthcare organization delivered policies to control these typical factors towards human’s living styles. Though malaria has been cleared for decades, there is one serious problem that there were emergences of malaria parasite that resisted various types of medicines in use. So, it causes the complication in malaria control processes. To solve such complication, this study collected data mainly from the National Statistical Office database and applied statistical method, combining with the machine learning technique and Geographic Information System to determine risk factors, risk model, and risk map for malaria distribution in Thailand. As a result of these methods, this study proposed a new set of monitoring parameters, 28 parameters, a logistic model tree (LMT) risk model, and a risk map for malaria distribution derived by the machine learning mechanism. The precision and recall of this tree are 0.780 and 0.821, respectively, which are higher than using the risk model created from the typical factors. As a result of this finding, the government can utilize these newly identified factors to develop suitable strategies to control malaria considering proper criteria. For example, the government should control the quality of rivers or even increase the amount of river monitoring in some areas to reduce the malaria risk. Once these factors are completely managed with well-developed plans, each community can reduce the risk of malaria without concern regarding drug resistance. Moreover, according to the risk map, each area in Thailand has been defined as high, medium, low distribution of malaria based on the discovery factors. In such result, the government can set suitable the malaria monitoring and protection strategies for each individual area with a proper budget to be provided.

Field of Study:	Bioinformatics and Computational Biology	Student's Signature ....
Academic Year:	2019	Advisor's Signature ....
		Co-advisor's Signature ....

ACKNOWLEDGEMENTS

The research was supported by the 90th Anniversary of Chulalongkorn University Fund. This work would not have been possible without the advice and support from the my beloved thesis advisor, Assistant Professor Pongchai Dumrongrojwatthana and co-advisor, Associate Professor Pattarasinee Bhattarakosol, along with other thesis committee members, including Associate Professor Teerapong Buaboocha, Assistant Professor Kitiporn Plaimas, Assistant Professor Sittiporn Pattaradilokrat and Associate Professor Panjai Tantatsanawong. Lastly, I would like to express my profound gratitude to my parents (Col. Pongsak Krainara, Mrs. Lamyai Krainara and Mr. Peerapat Krainara), relatives, and friends for providing both financial and emotional supports during my years of study. This accomplishment would not have been possible without them.

Patcharaporn Krainara

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1

Introduction

Malaria is an infectious disease caused by *Plasmodium* spp. parasites, which are transmitted by Anopheles mosquitoes. This disease is considered a major public health problem, especially countries in the tropical and the subtropical regions of the world; approximately 36% of the global population in over 90 countries is at risk for malaria infection. Based on global records, more than 90% of malaria cases occur in the African Region, while cases in other parts of the world account for less than or equal to 5% (World Health Organization, 2017, 2018). These data suggest that malaria is most prevalent in low-income countries.

Notably, a report from the World Health Organization (WHO) by Alanso (World Health Organization, 2017) stated that severe drug-resistant strains of malaria parasites, also known as super malaria, have appeared in the Mekong River Basin and are continuously spreading in many countries in South-east Asia. Initially, an epidemic involving Cambodia and parts of Thailand, Laos and southern Vietnam was reported. In Vietnam, one-third of patients presented with malaria that was resistant to medical treatments, while 60% of patients in Cambodia presented with malaria that did not respond to traditional drug treatments (Imwong, Tinh Tran, Nguyen, Dondorp, & White, 2017). A charitable organization for medical research named " Wellcome Trust" from London, the United Kingdom, reported that approximately seven hundred thousand people worldwide die from drug-resistant infections, including malaria, each year (Chew, 2017). Thus, if no serious remedies are implemented, the number of deaths caused by drug-resistant infections may increase to millions of people per year by 2050. Currently, the primary action to combatting drug resistance is to change medication types. Additionally, pharmaceutical researchers are attempting to develop a new drug for treating malaria, but eventually, malaria will develop resistance to this new drug, resulting in short-lived time on the market. Therefore, investing in drugs to treat infectious diseases is not as attractive as investing in medicines used to treat chronic diseases such as diabetes, high blood pressure, and heart disease that can be continuously sold and expanded in the marketplace. Considering these circumstances, we suspect that the spread of malaria will continue to increase, as stated by the Wellcome Trust, London the United Kingdom.

Since national and international transportation are very convenient and inexpensive in Association of South-east Asian Nations (ASEAN) communities, people can easily travel to any place at any time. Unfortunately, this travel also leads to an increased risk of malaria transmission. Although the Ministry of Health in Thailand reported that the malaria situation is likely to improve, the drug resistance problem still exists for *P. falciparum* and other *Plasmodium* species, similar to other countries around the world. Since the risk of contracting malaria can increase because of the conditions of the surrounding environment, people can be infected easily and repeatedly. As a consequence, the issue of drug resistance must be seriously addressed (Lertpiriyasuwat, 2019).

To prevent the spread of malaria, the United Nations High Commissioner for Refugees (UNHCR) has established 5 global action plans. The first action plan involves enhancing awareness and understanding of antimicrobial resistance. The second action plan involves strengthening surveillance and research, while the third action plan involves reducing the incidence of infection. The fourth and the fifth action plans involve rationally using and developing antimicrobial drugs and ensuring that funding to study antimicrobial resistance will be sustainable, respectively. Though the fourth and fifth plans focus on drug production, the long-term sustainability of these plans is questionable because the disease is continuously evolving. Consequently, the first three plans are highly important, as they are more consistent with epidemiological studies on the physical impacts and advantages than on studies on control through the use of genetic methods. Moreover, controlling on the physical impacts can result in lower expenses and greater efficacy in the long term than solving drug resistant problems.

Generally, physical impacts are controlled by surveillance and monitoring systems. These systems rely on either single or multiple factors, such as predicting the distribution of mosquitoes using mosquito density data or examining the relationship between land use, land cover, and mosquito distribution. Currently, no molecular control methods are being studied. In addition, there is no information linkage among authorities that work in malaria control. As a consequence, full surveillance efficiency cannot be completely obtained. Each authoritative unit manages different information; some of them maintain data on environmental factors related to malaria distribution, and some of them maintain data on healthcare, malaria patients, etc. Some of these data, such as temperature, land use and land cover, and rainfall, are considered factors that are related to malaria distribution. Nevertheless, other data, such as healthcare units, burn areas, and socio-economic status are hardly associated with the transmission of malaria. Thus, these factors might have some indirect impacts on the malaria control system. Therefore, this research emphasizes the control of physical and environmental factors, consistent with previous epidemiological studies. In this study, we consider various types of data in the analysis and modelling processes; these data include ecological, social-economic, health situation, disease, and policy data. Furthermore, we use machine learning techniques to create a risk model that supports the monitoring and control of malaria distribution factors. The expected outcome of this study is a proper strategy to eliminate malaria in Thailand. The analyzed data are secondary data obtained from the public databases of many related government agencies. Since there are large volumes of data from the various sources, these data are calculated, modelled, and interpreted to obtain the greatest benefits for further usage.

Once significant factors and the risk model are discovered, the next step to increase efficiency of malaria control policy is to implement a risk map. The risk map is used to demonstrate the risk-ranks for malaria distribution; so, the government can setup proper policy to monitor, protect, and detect the spread of malaria. In addition, there is a benefit from taking risk maps to study the spread of malaria, such as applied a mathematical modeling approach for standardized morbidity ratio (SMR) calculated by annual parasite incidence using routinely aggregated surveillance reports, environmental data and non-environmental anthropogenic data to create fine-scale spatial risk distribution maps of western Cambodia (Okami & Kohtake, 2016). Mapping multiple components of malaria risk for improved targeting of elimination interventions (Cohen et al., 2017). To study the risk maps indicate the spatial heterogeneity of malaria prevalence. The aim of this study is to analyses and map malaria risk in children under 5 years old, with the ultimate goal of identifying areas where control efforts can be targeted (Yankson, Anto, & Chipeta, 2019).

Education risk factors are another important action in understanding the spread of malaria. For examples, to study and to explore the associated risk factors of malaria transmission at the microeconomic level (households) in two rural villages of mainland Equatorial Guinea (Guerra, de Sousa, Ndong-Mabale, Berzosa, & Arez, 2018), to analyze malaria risk factors based on human and housing conditions in Kaligesing, Purworejo, Indonesia (Cahyaningrum & Sulistyawati, 2018), and to investigate the magnitude and associated factors with malaria outbreak (Tesfahunegn, Berhe, & Gebregziabher, 2019).

Objectives

1. To identify the relationship of various factors, both physical and biological, related to the malaria distribution

2. To create a risk model that affects malaria distribution by applying various analytical techniques

3. To create a risk area map for monitoring malaria distribution in Thailand

Structure of the dissertation

Chapter 2 comprises the literature review focusing on techniques for risk factor analysis and risk map. The research methodologies are elaborated in Chapter 3 which includes data collection and analysis, model derivation, and mapping classification. The results of this study are presented in Chapter 4 which are divided into 4 parts: 1) relationship of various factors, both physical and biological, related to the malaria distribution, 2) risk model that affects malaria distribution, 3) risk map for monitoring malaria distribution in Thailand, and 4) step-by-step guideline/manual to use the constructed model from this study. The discussion and conclusion are presented in Chapter 5 and Chapter 6, respectively.

Chapter 2

Literature Review

This chapter describes the review literature by discussing the symptoms severity of disease and ways to prevent from relevant agencies, including the situation of malaria from the past to the present from around the world and the situation in Thailand as well as related research in malaria studies. Moreover, the concepts in machine learning have been elaborated to show relations between machine learning mechanisms and medical diagnostics in the present world. In addition, the concepts in Geographic Information System have been to spatial data.

Background and Significance

Malaria is an infectious disease caused by the parasite *Plasmodium* spp. The *Plasmodium* life cycle involves both vertebrates and mosquitoes (Epidemiology) Department of Disease Control, Ministry of Public Health. Malaria transmission occurs when humans are bitten by *Anopheles* mosquitoes that carry *Plasmodium.* 5 species of the *Plasmodium* parasites cause malaria in humans, and 2 of these species, *P. falciparum* and *P. vivax*, pose the greatest threat. Once the *Plasmodium* parasites enter to the human body, the incubation period starts which can be ranged from 2 weeks to 2 months depending on each infected parasite species. After the incubation period, the carrier will have a high fever, fatigue, vomiting, and headache. In severe cases, it can cause jaundice, seizures, coma or death. Patients may have fever every day, or every two days. Thus, the patients must be treated properly; otherwise, the severe symptom occurs. With the severe state, the patients will have a jaundice, renal failure, cerebral malaria, and eventually it will lead to death (Thaitravelclinic) Useful traveler tips, Hospital for Tropical Diseases.

Presently, malaria is counted as a major public health problem of the world because approximately 36 percent of population around the world from over 90 countries is in the area with malaria transmission. According to the latest World malaria report, released in November 2018, there were 219 million cases of malaria in 2017, up from 217 million cases in 2016. The estimated number of malaria deaths stood at 435,000 in 2017, similar the number to the previous year (World Health Organization, 2018). Children aged less than 5 years are the most vulnerable group affected by malaria. In 2017, they accounted for 61% (266,000) of all malaria deaths worldwide. Although there were 20 million fewer cases in 2017 than in 2010 globally, the number of patients during the period 2015 to 2017 was slightly increased, despite a dip in cases in 2015. WHO African Region still bears the largest burden of malaria morbidity, with 200 million cases (92%) in 2017, followed by the WHO South-East Asia and the WHO Eastern Mediterranean (5% and 2%, respectively). In addition, almost 80% of all malaria cases globally were in 15 African countries and in India (World Health Organization, 2018). Moreover, nearly 50% of all cases globally were accounted for by Nigeria (25%), the Democratic Republic of the Congo (11%), Mozambique (5%), India (4%) and Uganda (4%). The 10 highest burden countries in Africa reported increases in cases of malaria in 2017 compared with 2016. In contrast, WHO reported that there were 3 million fewer cases in India during 2017 which was 24% decrease when comparing to cases in 2016. In addition, Rwanda has noted estimated reductions in its malaria burden, with 430,000 fewer cases in 2017 than in 2016; Ethiopia and Pakistan estimated decreases of over 240,000 cases over the same period, additionally the incidence rate of malaria declined globally between 2010 and 2017, from 72 to 59 cases per 1,000 populations at risk. Although this represents an 18% reduction over the period, the number of cases per 1,000 populations at risk has stood at 59% for the past 3 years. The WHO South-East Asia Region continued to observe its incidence rate fall from 17 cases per 1,000 population at risk in 2010 to 7 cases in 2017 (a 59% decrease) (World Health Organization, 2018). All other WHO regions recorded either little progress or an increase in incidence rate. The WHO Region of the Americas showed an increasing of the incidence rate, mainly due to increases in malaria transmission in Brazil, Nicaragua and Venezuela (Bolivarian Republic of Venezuela) (World Health Organization, 2018). In the WHO African Region, the malaria incidence rate remained at 219 cases per 1,000 population at risk for the second year in a row. Through, the *P. falciparum* is the most prevalent malaria parasite in the WHO African Region (99.7%) of estimated malaria cases in 2017, followed by in the Western Pacific (71.9%), the Eastern Mediterranean (69%), and the WHO South-East Asia (62.8%) Regions. *P. vivax* is the predominant parasite in the WHO Region of the Americas, representing 74.1% of the malaria cases.

Table 1: Estimated malaria cases between 2010-2017, with 95% upper and lower confidence interval (CI).

	Number of cases							
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Lower 95% CI	218,600	210,500	206,700	200,500	199,600	198,700	200,400	202,800
Estimated total	**238,800**	**229,100**	**226,400**	**221,000**	**217,100**	**214,200**	**216,600**	**219,000**
Upper 95% CI	285,400	273,200	271,600	266,200	259,300	257,200	259,000	262,000
Estimated *P.vivax*								
Lower 95% CI	11,440	10,390	9,190	7,040	6,040	5,530	5,960	5,720
Estimated total	**16,440**	**14,940**	**13,300**	**10,230**	**8,720**	**7,950**	**8,250**	**7,510**
Upper 95% CI	24,560	23,970	22,050	17,240	12,730	11,410	11,300	9,900

Source: World Health Organization (WHO, 2018)

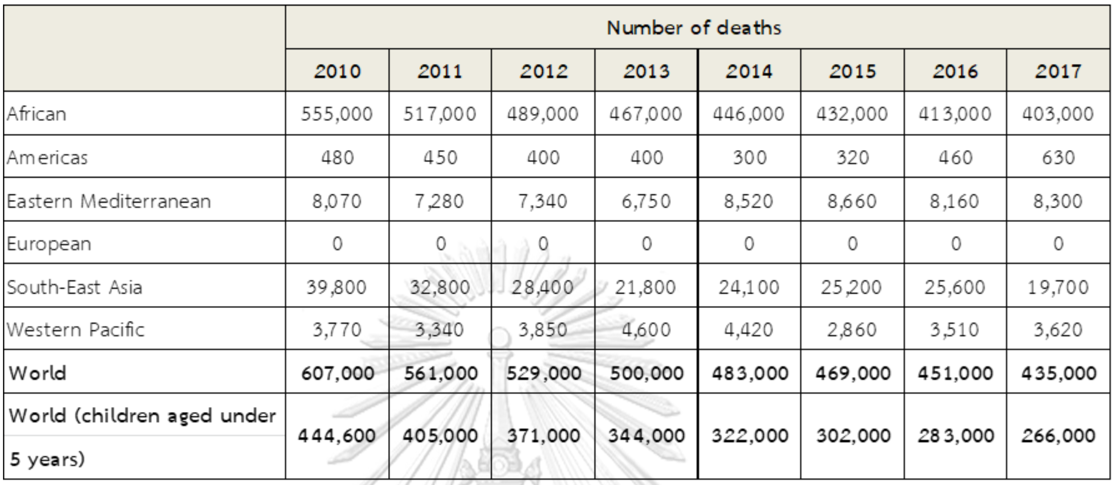
Table 2: Estimated malaria cases by WHO region in 2017, with 95% upper and lower confidence interval (CI).

	Number of cases					
	African	Americas	Eastern Mediterranean	South-East Asia	Western Pacific	World
Lower 95% CI	184,500	880	3,630	8,560	1,395	202,800
Estimated total	**200,500**	**976**	**4,410**	**11,290**	**1,857**	**219,000**
Upper 95% CI	243,600	1,128	5,560	14,840	2,399	262,000
Estimated *P.vivax*						
Lower 95% CI	19	648	1,162	2,881	330	5,720
Estimated total	**701**	**723**	**1,366**	**4,200**	**523**	**7,510**
Upper 95% CI	2,197	843	1,773	5,900	774	9,900
Proportion of *P.vivax* cases	0.30%	74.10%	31.00%	37.20%	28.10%	3.40%

Source: World Health Organization (WHO, 2018)

From Table 1 and Table 2, it can be seen that the estimated 219 million cases of malaria occurred worldwide in 2017 (95% CI: 203–262 million) compared with 239 million cases in 2010 (95% CI: 219–285 million) and 217 million cases in 2016 (95% CI: 200–259 million). Currently, the number of malaria cases tends to decrease considerably compared to in the past (Table 1). Although there were 20 million fewer cases in 2017 than in 2010, globally, the period 2015 to 2017 showed only a minimal of slightly upward change in trend, despite a dip in cases in 2015, suggesting that progress had generally stalled. The WHO African Region still bears the largest burden of malaria morbidity, with 200 million cases (92%) in 2017, followed by the WHO South-East Asia Region (5%) and the WHO Eastern Mediterranean Region (2%) (Table2). Globally, 3.4% of all estimated cases were caused by *P. vivax*, with 56% of the vivax cases being in the WHO South-East Asia Region. *P. vivax* is the predominant parasite in the WHO Region of the Americas (74%), and is responsible for 37% of cases in the WHO South-East Asia Region, and 31% in the WHO Eastern Mediterranean Region. Therefore, it can conclude that *P. vivax* cannot spread well in the hot and dry weather as in Africa region but it can grow rapidly in the hot, humidity, and rainy region, such as America, and South-East Asia.

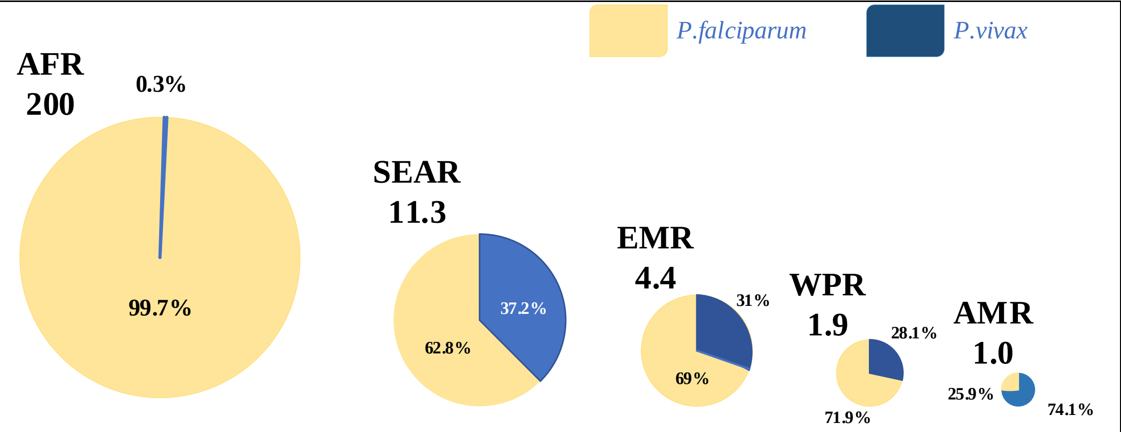
Table 3: Estimated number of malaria deaths by WHO region between 2010–2017



	Number of deaths							
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
African	555,000	517,000	489,000	467,000	446,000	432,000	413,000	403,000
Americas	480	450	400	400	300	320	460	630
Eastern Mediterranean	8,070	7,280	7,340	6,750	8,520	8,660	8,160	8,300
European	0	0	0	0	0	0	0	0
South-East Asia	39,800	32,800	28,400	21,800	24,100	25,200	25,600	19,700
Western Pacific	3,770	3,340	3,850	4,600	4,420	2,860	3,510	3,620
World	**607,000**	**561,000**	**529,000**	**500,000**	**483,000**	**469,000**	**451,000**	**435,000**
World (children aged under 5 years)	**444,600**	**405,000**	**371,000**	**344,000**	**322,000**	**302,000**	**283,000**	**266,000**

Source: World Health Organization (WHO, 2018)

From Table 3, it shows the estimated number of malaria deaths by WHO in several regions between 2010 - 2017, most of them occurs in African areas, followed by South-East Asia. Nonetheless, the areas with no deaths due to malaria are in Europe. Moreover, based on the situation of every area in the world, it is found that the deaths due to malaria tend to decrease considerably.



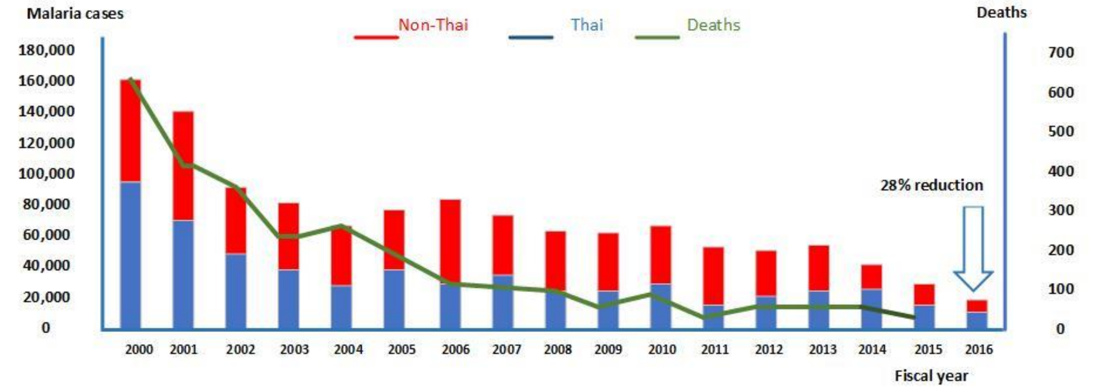
Note: AFR: WHO African Region; AMR: WHO Region of the Americas; EMR: WHO Eastern Mediterranean Region; *P. falciparum: Plasmodium falciparum; P. vivax: Plasmodium vivax*; SEAR: WHO South-East Asia Region; WHO: World Health Organization; WPR: WHO Western Pacific Region.

Source: World Health Organization (WHO, 2018)

Figure 1: Estimated malaria cases (millions) by WHO region, 2017. The area of the circles is shown as a percentage of the estimated number of cases in each region.

From Figure 1 Most areas were *P. falciparum* except in the Americas. Most were *P. vivax*. Most malaria cases in South America come from Amazon rain forest areas in northern countries, where more than half of malaria is caused by *Plasmodium vivax*, while Plasmodium falciparum malaria incidence has decreased in recent years (Recht et al., 2017).

Similar to the world record, Thailand also considers malaria as a major public health problem since over 13 million people (19% of the total population) currently at risk, most of them are situated along border areas with its neighboring countries. There is significant geographical heterogeneity in spatial distribution of disease incidence, which exemplifies the “border malaria” type , characterized by high transmission along international borders (Mercado et al., 2019). The malaria situation in Thailand was that in 2017, there were 10,965 patients, divided into 7,201 Thai people and 3,764 foreigners. There was a decrease about 70.89% from the year 2016, comparing with the same period (Detail at Day 1 Jan - 22 Dec 2017). Similar to the year 2017, the number of patients was dropped for 41.74% from the previous year. So, the total cases in 2018 were 6,641 cases or could be counted as 0.1 per thousand population's sick rate; and it can be classified as 72.44 % of Thai citizen and 27.56 % of non-Thai. According to the Ministry of Health's report in 2018, it indicated that males were a dominant group for malaria infection rather than female, with the ratio of 2:1. In addition, the working age-range of 25-44 years has the highest chance, with 32.54% from all ages, to be infected with malaria than other age ranges. Moreover, more than 50 % of the patients were famers. This report may lead to the assumption that the occupation of patients can be counted as a factor of malaria infection. In addition, according to the Asia Pacific Malaria Elimination Network (APMEN) website, since 1990s, intensive vector control along with improved targeting of at risk populations has achieved a marked reduction in Thailand's malaria incidence. Between 2000 and 2014, reported malaria cases declined from 78,561 to 37,921 cases (52%). During the same time period, deaths due to malaria decreased by 94% from 625 to 38 deaths. Thailand is two transmission peaks in Thailand from June to August and October to November, coinciding with the rainy season and a corresponding increase in density of the main vectors. The primary vectors responsible for malaria transmission are *Anopheles dirus*, *Anopheles minimus*, and *Anopheles maculatus* which become increasingly important due to deforestation.



Source: Bureau of Vector-borne Diseases BVBD (2018)

Figure 2: Trend of malaria morbidity and mortality in Thailand among Thais and non-Thais (M1, and M2) (FY 2000-2016)

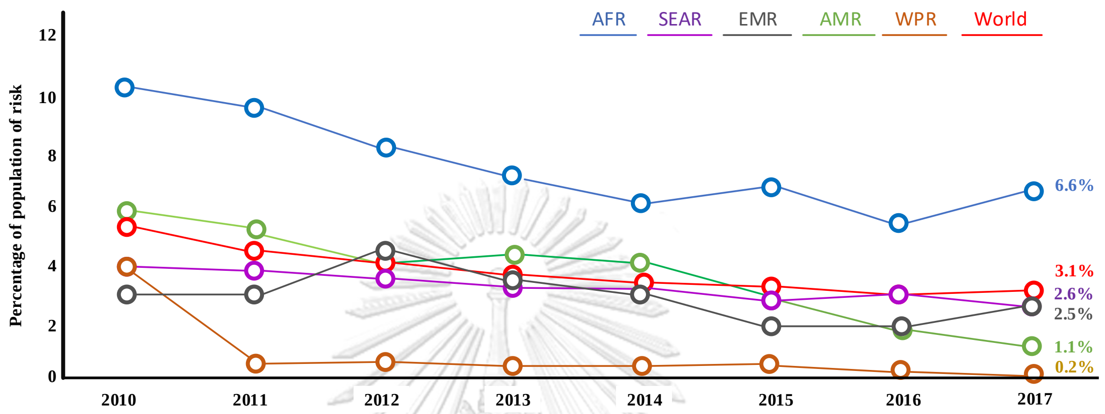
Figure 2 shows malaria cases among Thais and non-Thais. There was a reduction in malaria cases from 24,840 in 2015 to 17,153 in 2016. In addition, the Annual Parasite Incidence (API) was decreased from 0.38 to 0.28 per 1,000 populations during the same period. These cases included those who crossed the border and sought treatment at malaria posts and health facilities in Thailand. Between 2013 and 2015, mortality due to malaria has also reduced from 47 to 33 deaths. The increasing proportion of *P. vivax* malaria cases is observed with the proportion attributable to *P. vivax* steadily increasing from 58% in 2010 to 72% in 2016.

One interesting issue is that *P. vivax* is responsible for 37% of all cases within the WHO South-East Asia Region and 31% in the WHO Eastern Mediterranean Region(World Health Organization, 2018). This WHO article supports the report of Department of Disease Control, Ministry of Public Health, Thailand, in 2018, which stated the majority of the parasite were of *P. vivax* (82.14 %) and *P. falciparum* (11.93 %). About 82% of the estimated *P. vivax* malaria cases in 2017 occurred in just five countries (India, Pakistan, Ethiopia, Afghanistan, and Indonesia). Moreover, the decreasing of malaria cases, estimated in 20 countries, is more than 20% compared with 2016. Most of these changes occurred in countries with low to very low malaria burden, and in several countries the absolute difference was small (World Health Organization, 2018).

Although *P. vivax* is the main cause of malaria sickness around the world, it is reported as the predominant parasite in the WHO Region of the Americas where the spread of malaria is small. Thus, the differences among the American, Africa, South East Asia, and Eastern Mediterranean Regions should be considered, such as education system, economic system, and healthcare system, etc.

Since malaria can spread easily via mosquitoes, the vector control is the main way to prevent and reduce malaria transmission. A measure of protection will be conferred across the community, if coverage of vector control interventions within a specific area is high enough. WHO recommends protection for all people at risk of malaria with the effective control of the malaria vector. Two forms of vector control insecticide-treated mosquito nets (ITN) and indoor residual spraying (IRS) with residual insecticides are effective in a wide range of circumstances (World Health Organization, 2018).

Fewer people at risk of malaria are being protected by indoor residual spraying (IRS). Globally, IRS protection declined from a peak of 5% in 2010 to 3% in 2017, with decreases seen across all WHO regions (World Health Organization, 2018). In the WHO African Region, IRS coverage dropped from 80 million people at risk in 2010, to a low point of 51 million in 2016 before rising to 64 million in 2017. In other WHO regions, the number of people protected with IRS in 2017 was 1.5 million in the Americas, 7.5 million in the Eastern Mediterranean, 41 million in South-East Asia, and 1.5 million in the Western Pacific. The declines in IRS coverage occur as countries change or rotate insecticides (changing to more expensive chemicals), and as operational strategies change (e.g. decreasing at-risk populations in malaria elimination countries).



Note: AFR: WHO African Region; AMR: WHO Region of the Americas; EMR: WHO Eastern Mediterranean Region; IRS: indoor residual spraying; NMP: national malaria program; SEAR: WHO South-East Asia Region; WHO: World Health Organization; WPR: WHO Western Pacific Region.

Source: World Health Organization (WHO, 2018)

Figure 3: Percentage of the population at risk protected by indoor residual spraying by WHO region, 2010–2017

The percentage of the population at risk protected by IRS declined globally from a peak of 5% in 2010 to 3% in 2017, with decreases seen in all five WHO regions for which data were analyzed (Figure. 3). The number of people protected in 2010 was 180 million globally, but, by 2017 this had reduced to about 116 million. In the WHO African Region, the percentage of the population at risk protected by IRS decreased from 10.1% (80 million) in 2010 to 51 million (5.4%) in 2016, before rising to 64 million (6.6%) in 2017. Most of these increases in the period 2016–2017 were reported in Burundi, Ethiopia, Ghana, Kenya, Mozambique, Uganda, the United

Republic of Tanzania and Zambia. In other WHO regions, the number of people protected with IRS was 41 million in South-East Asia, 7.5 million in the Eastern Mediterranean, and 1.5 million in both the Americas and the Western Pacific. However, in most countries, IRS implementation is focused and is targeted at a much smaller population at risk; NMP reports show that, among the targeted population, operational coverage is substantially higher than what is shown in Figure 3.

In 2019, the study of using insecticide-treated bed nets (ITNs) was performed in the highlands of Western Kenya (Essendi et al., 2019). This research compared families with and without malaria (families use ITNs). Additionally, the paper also focused on the use of ITNs associated with the lower level of clinical malaria episodes to identify risk factors for malaria infection. As the consequence, the information about local malaria transmission and higher effectiveness of malaria control measurement were suggested. As same as the study in Kenya, the Regional Artemisinin Resistance Initiative (RAI) project of Myanmar studied the barriers in distributing, ownership, and utilizing of ITNs among the high-risk migrant communities in the RAI area(World Health Organization, 2018). This study suggested that the map related to the high-risk migrant communities must be drawn, including the continuity distribution of the ITNs (Linn et al., 2019). This mapping concept is like the research of (Noé et al., 2018) that recommended maps to determine the pattern and stability of malaria hotspots in Bangladesh with the end goal of informing intervention planning for elimination.

Half of people at risk of malaria in Africa are sleeping under an ITN: in 2017, 50% of the population was protected by this intervention, an increase from 29% in 2010. Moreover, the percentage of the population with access to an ITN increased from 33% in 2010 to 56% in 2017 (World Health Organization, 2017). However, coverage has improved only marginally since 2015 and has been at a standstill since

2016. Households with at least one ITN for every two people doubled to 40% between 2010 and 2017. However, this shows a slightly increase over the past 3 years. It remains far from the target of universal coverage.

Another prevention method is the use of IRS as stated earlier. Various researches have been performed in the recent year, such as that by (Eskenazi et al., 2019), which studied the importance of IRS in minimizing insecticide exposure in the Vhembe District of Limpopo, South Africa. Furthermore, the efficacy of bendiocarb (FICAM WP 80) spray was investigated on different wall surfaces to disclose the impact on malaria vectors (Lo et al., 2019). Besides, the impact on entomological outcomes of combining IRS and long-lasting ITN (LLINs) was compared with the implementation of either IRS or LLINs in Adami Tullu district, south-central Ethiopia (Kenea et al., 2019). Even though ITNs and IRS are the common use items suggested by WHO, these methods are not applied to the pregnant women and infants. For these special groups, WHO has recommended the intermittent preventive treatment for the pregnant women (IPTp), and the intermittent preventive treatment for the infants (IPTi) (World Health Organization, 2018). Furthermore, a policy on seasonal malaria chemoprevention for malaria control in the highly seasonal transmission areas should be implemented for children under 5 years of age (World Health Organization, 2018).

The protected women in areas of moderate and high malaria transmission in Africa, WHO recommends “intermittent preventive treatment in pregnancy” (IPTp) with the antimalarial drug sulfadoxine-pyrimethamine (World Health Organization, 2018). Referring to the report of IPTp coverage levels in 2017, an estimated 22% of eligible pregnant women among 33 African countries received the recommended three or more doses of IPTp (World Health Organization, 2017). In 2017, 15.7 million children in 12 countries in Africa’s Sahel sub-region were protected through seasonal malaria chemoprevention (SMC) programs. However, about 13.6 million children who could have benefited from this intervention were not covered, mainly due to a lack of funding.

There are various studies and preventive measures that aimed to solve the spread of malaria in the past. For examples, an evolutionary-epidemiological modeling framework was designed under the context of drug resistant evolution; this model provides a better treatment’s outcome based on multiple-first line therapies (MFT) (Boni, Smith, & Laxminarayan, 2008), the study on the role of chemoprophylaxis treatment for malaria infection and distribution in the risk area, including pregnant women indicated that this method is high efficient than using vaccine for malaria prevention (Greenwood, 2010), the stochastic modeling was applied to simulate the effectiveness of using long lasting insecticide treated net (LLIN) and indoor residual spraying (IRS) (Stuckey et al., 2014). At present, there are a variety of studies such as predictions of mosquito distribution areas using density of *Anopheles* mosquitoes (Adde et al., 2016). The study in the relationship of land use and the distribution of mosquitoes was also determined (Baeza, Santos-Vega, Dobson, & Pascual, 2017). In addition, there is a study of biomolecules that can identify new transcripts and 179 *vir* like genes, including 3018 non-coding RNAs (Zhu et al., 2016). Moreover, the study of genetic variability of *P. falciparum* histidine-rich proteins 2 and 3 in Central America was performed (Fontecha, Pinto, Escobar, Matamoros, & Ortiz, 2019). The study of effectiveness in the accelerating elimination with long-lasting systemic insecticides for contemporary vector control in 3 different areas was performed; the results showed that about 85% coverage of anti-malaria in mass drug administration campaigns (Selvaraj, Suresh, Wenger, Bever, & Gerardin, 2019).

In Thailand, there is a long history of malaria researches. For example, the combining between a gametocytocidal drug and an artesunate-mefloquine for gametocyte clearance time is much effective than using only artesunate-mefloquine in the malaria transmission area (Tangpukdee et al., 2008). A research found out that the first-line treatment and a three-day course of artesunate-mefloquine combination provided a high efficiency and tolerability in the treatment with highly multidrug resistant falciparum malaria (Congpuong, Bualombai, Banmairuroi, & Na-Bangchang, 2010). The understanding of demographic variables and all related variables in the community-level on individual malaria occurrence highly supports the intervention strategic specifically planning of each location along the Thai-Myanmar border (Lawpoolsri et al., 2010). The comparative detection of *P. vivax* and *P. falciparum* DNA in saliva and urine samples from symptomatic malaria patients in a low endemic area was characterized (Buppan, Putaporntip, Pattanawong, Seethamchai, & Jongwutiwes, 2010). The biting pattern of malaria vector was studied along the Thailand-Myanmar border to find the malaria control method by the Faculty of Agriculture, Kasetsart University (Kwansomboon et al., 2017).

The research focuses on Thailand where real-time malaria surveillance is crucial because malaria is re-emerging, developing, and resisting to pharmaceuticals in the region (Ocampo, Chunara, & Brownstein, 2013). Moreover, there is a study that describes the pattern and epidemiological profile of malaria recurrence in a hypoendemic area of western Thailand and identified factors associated with having multiple malaria episodes performed by (Lawpoolsri et al., 2019). In addition, measure human population movement, associated predictors of travel, and human population movement correlates of self-reported malaria among people living within malaria hotspots are identified (Saita et al., 2019). Besides, the potential challenges associated with the goal and future strategies for malaria elimination in the Greater

Mekong Sub-region must be continuity supported (Kaehler et al., 2019). At present, the government, by the Ministry of Public Health, has a policy for all provinces to expedite the implementation of the malaria eradication strategy Thailand from 2017 – 2026. This policy focuses on the development of malaria monitoring system; increasing the detection for the malaria infection in both patients with and without symptoms, monitoring the malaria vector control, promoting the malaria protection, and monitoring the emergence of drug resistant parasite. To achieve the goal, there are 5 main procedures including, 1) actively search and screen for malaria patients, 2) increase the access to healthcare services, 3) monitor disease in emigrant population and workers, 4) promote the use of insecticide coated net and provide indoor chemical spraying, and 5) create the cooperation network among concerned stakeholders.

Current malaria situation of Thailand shows downward trend in the transmission rate. Nevertheless, the report of Ministry of Health in 2018 stated that there were emergences of malaria parasite that resisted various types of medication in use. If the drug resistant strains are transmitted to other areas, it causes the complication on the malaria control process. Furthermore, it may affect the wellbeing, health, economic and social aspect of people, especially at the family and community level of society, which finally has impact to the budget of healthcare system of both government and individuals.

Based on the reviewing contents mentioned previously, the main issue for malaria transmission monitoring and tracking at the present depended only on single aspect of factor related with malaria transmission. Although of the objective of this study is to identify factors and their relationships that influence to the distribution of malaria, the factors that directly related to the patients will not be taken in the consideration, such as their health information, drug resistance, and changing of medical treatment etc. So, this study mainly concerns in the quantitative factors, such as ecological data (e.g. forest area data as shown in Figure 4, water sources as shown in Figure 5), socio-economics data (e.g. land use size as shown in Figure 6, population movement), and disease information (number of patient). The expected outcome of this study is a proper strategy to eliminate malaria in Thailand. The analyzed data are secondary data obtained from public databases of many related government agencies. Since there is a large volume of data from various sources, these data are calculated, modelled, and interpreted to obtain the highest benefits for further usage.

Figure 4: Malaria risk areas (Bureau of Risk Communication and Health Behavior Development, Department of Disease control)

Source: http://kanchanapisek.or.th/

Figure 5: Streams in the forest are the breeding ground of Anopheles mosquitoes.

Source: https://indohun.org/news/disease-emergence-and-economic-evaluation-of-altered-landscapes-deal-colaboration-research-in-deforestation-and-zoonosis-disease/

Figure 6: Deforestation is basically converting forest land to non-forest, such as plantations, agriculture, settlements, and so on. As is well known, forests are often associated with malaria transmission dynamics.

Machine Learning Techniques

To create an efficiency malaria distribution model for surveillance, the machine learning techniques are applied. The concept of machine learning allows machine to learn from either sample data sets or ad hoc data sets. Once learned, knowledges includes the interpretation for understanding, skills, or experiences can be stored in the knowledge base with one knowledge substitution model, such as rules, functions, etc. The machine learning can be categorized in 3 main sub-mechanisms: Supervised learning, unsupervised learning, and Reinforcement learning. These sub-mechanisms are individually applied in different conditions as the prediction tools using large amount of data to analyze to gain the predicted result. Therefore, the prediction model of the malaria distribution can be obtained with high accuracy when the machine learning is applied to derive the model.

Supervised learning refers to a mechanism that a sample data set and the results in which the "Supervisor" needs to be entered. The goal is to create general rules that can connect inbound and outbound data. Therefore, as show in Figure 7, the training data which the results are already known in advance is the input data of the supervised learning algorithm to build the predictive model. Then, the new data set in which the model has never seen before is entered; the outcome is the prediction of an expected event.

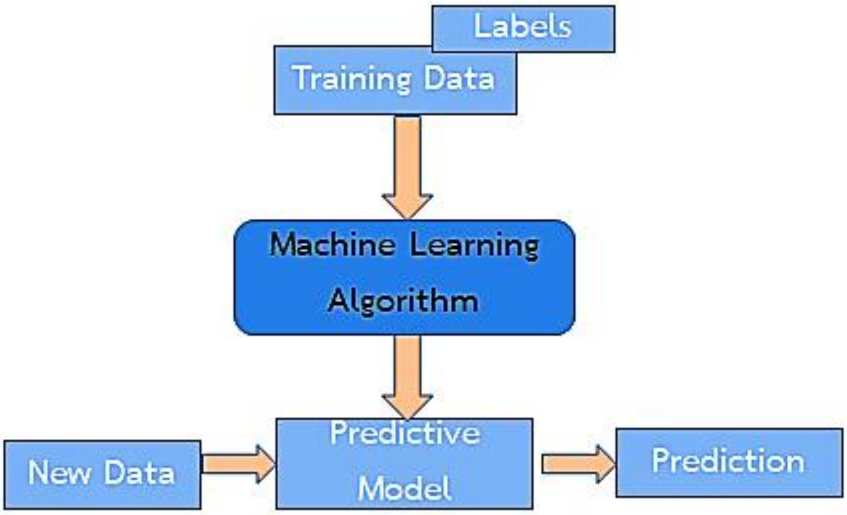


Figure 7: Basic forms of supervised learning mechanism (Raschka, 2015).

Supervised learning can be identified into two models: classification model, and regression model. Considering the classification model, the input data is divided into many classes and the learner, or the machine, must create a suitable model that is capable to assign categories to the new data which has never been seen before. Generally, the classification model can be derived based on the input that is provided by supervisors; examples of data classification include common characteristics such as recognition of people faces. For this classification, images of faces for the learning process are labelled. Then, with a new image, the prediction result of the process will be the name of the person's face. So, it is clear that the classification model can predict a person from the specific picture using the identified class; in this case, the person's name determines the class.

The second model of the supervised learning is the statistical regression. This model is used to determine the relationship between two variables or more. Although there are various models for finding the relationships among variables, various conditions to create the prediction models must be concerned. Nevertheless, the most popular method is the linear regression model (Raschka, 2015).

Unsupervised learning is the learning algorithm that concerns only the input data sets without any results provided in advance as supervise learning. The machine will learn all characteristics of the input data and classify them into various independent groups according to some criteria found after finishing the learning process. Thus, the outcome is the predicted model that cannot be predicted in the early state, not as same as the instructional learning. In addition, the obtained model is the most suitable model where the common properties are defined. Unsupervised learning can be identified into two models: clustering model and dimensionality reduction. Figure 8 shows the sample of cluster model where data are classified into 3 groups.

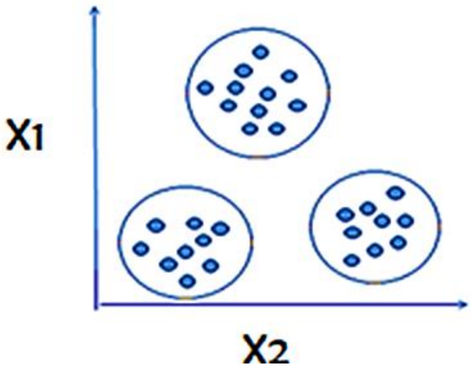


Figure 8: A sample of the clustering model (Raschka, 2015).

The second model of the unsupervised learning is the dimensionality reduction. This method considers all available features with their correlations and tries to reduce the correlated features. Since the complexity of the training is depended on the number of focused features, hence, reduction of the redundant features help reducing the complication of the training process. Therefore, the objective of the dimensionality reduction is to reduce the redundant features using correlated values and finally the principle variable set can be obtained and the classification procedure is simplified. Figure 9 shows a sample of the dimensionality reduction method.

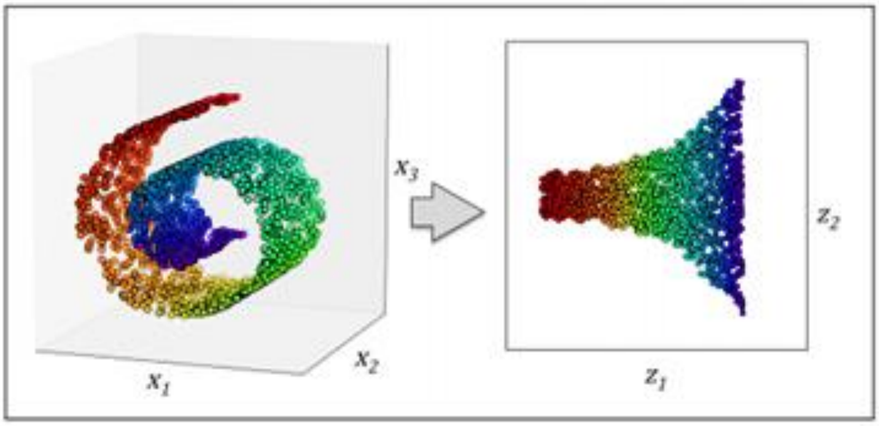


Figure 9: The sample of the dimensionality reduction model (Raschka, 2015).

Reinforcement learning is the learning process performed under the ever-changing environment. This method is dissimilar to the supervised learning in the point that there is no answer has been provided in advance but the reinforcement finds the solution from the learning based on its experiences. The mechanism of the reinforcement starts from the initial state and various outputs are derived for decision; all these actions are responded in the Environment box. The user must decide whether the output must be rewarded in the Agent box. Then, the process finishes; otherwise, each model keeps learning. Once every model finishes the learning process, the selected model is the one that provides the best solution where the maximum reward is granted. This tool is an Artificial Intelligence (AI) because it will learn and change according to the environment. For example, in the case of news around the world, the AI system named "Alpha Go" developed by Google's Deep Mind Company. This software was developed and studied the learning styles from experts around the world. Thus, Alpha Go develops the individual playing strategies and also practices millions of playing styles to increases its abilities to beat the highest world-class skilled expert in playing with Alpha Go.

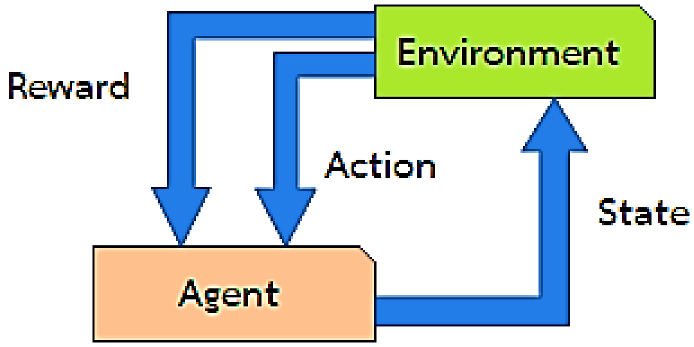


Figure 10: Reinforcement learning process. Credit (Raschka, 2015).

A way to construct a model to analyze risk factors that affect the distribution of malaria is the use of machine learning. In order to construct the machine learning model to predict the malaria distribution, the distribution data and other related historical data are divided into 2 parts: training datasets, and testing datasets. A large amount of the training data is required to obtain the completely and thoroughly trained model, which leads to a high accuracy predictive model. To validate the model's accuracy, the testing data set is used with the validation metrics, such as False Positive, and False Negative values in the confusion matrix; otherwise, precision and recall values can be considered. There are many machine learning methods for constructing predictive models as listed below.

1. Linear Regression is a linear approach to model the relationship between one or more independent variables or input (x) and dependent variable or outputs (y). The relationship is formed by calculating the coefficients of x in the equation ($\boldsymbol{\beta}$), which is similar to assign weights to the relationship between x and y such as $y = \beta_0 + \beta_1 x$. The goal of the algorithm is to find the regression coefficients of x, which are β_0 andβ_1, that gives the best-fitted line from the training dataset by minimizing errors from each data point to the line, in order to use the best-fitted line to predict y for a given set of x. The common methods for fitting a regression line are ordinary least squares and gradient descent optimization. Before using these techniques, two or more variables those correlated to each other must be split out from the same dataset; this includes removing the noise and useless data.

2. Logistic Regression is a technique to model a binary dependent variable or a dependent variable with two possible values, such as pass/fail, where the two values are labeled as “0” and “1” to classify the data into two groups. The output from logistic regression is the probability of the data belonging to each group. This technique is useful when we want to explain the predicted data.

3. Linear Discriminant Analysis (LDA) is a technique to classify two or more groups of data. LDA consists of statistical data such as mean and covariance parameters for each group. The prediction is made by calculating differences of the statistical data between groups. By assuming that the data has Gaussian distribution, the outlier must be removed from the datasets for the higher accuracy classification.

4. Classification and Regression Trees is a technique that can be used for classification or regression predictive modeling problem. The representation for the CART model is a binary tree (each node has at most two children). The tree data structure consists of two parts: nodes (input), and links (conditions to separate data inside a node into two groups). The bottom-most node is called the leaf node (output from the model). A new data is filtered through the conditions along the tree and lands in one of a leaf node, and the output value for that leaf node is the prediction obtained by the model. The decision tree model provides high accuracy for many types of prediction problems, and there is no need to pre-process the data used for constructing the model.

5. Naïve Bayes is a simple technique for constructing classifiers. The goal of any probabilistic classifier is, with features $x_0, \dots, x_n$ and classes$c_0, \dots, c_k$, to determine the probability of the features occurring in each class, and to return the most likely class. Therefore, for each class, $P(c_i |x_0 , \dots, x_n)$ can be calculated by Bayes rule. In order to simplify its computation to the Naïve Bayes classifier, we make the assumptions that $x_0, \dots, x_n$ are conditionally independent given to c_i. Although the assumption is most likely not true, the Naïve Bayes classifier performs well in most situations.

6. K-Nearest Neighbors (KNN) is a simple, but efficient, machine learning algorithm used for classification and regression. The input consists of the k closest

training datasets in the feature space. In KNN regression, the output is the property value (average value of its k nearest neighbors) for the object. In KNN classification, the output is a class membership. The KNN algorithm directly searches through all the neighbors or the training datasets by calculating distances between the testing data and all the training data in order to identify its nearest neighbors, which is the classification output. The distance between two data points is calculated by a distance function; the Euclidean distance is the most commonly used as the distance function. KNN is a memory intensive algorithm since it memorizes all the training datasets (n). When it comes to query a new point to find the nearest$k$, the query time will be expensive with $O_{(n)}$ running time. The training datasets can be changed or improved frequently, in order to increase the model's accuracy.

7. Learning Vector Quantization is a technique arises from overcoming the weakness of KNN. A downside of KNN is that the researcher needs to hang on to the entire training datasets, in order to classify a new point. The LVQ is an artificial neural network algorithm that lets the researcher choose how many training data points to hang onto and learns exactly what the properties of those data points are. The representation for LVQ is a collection of codebook vectors. The codebook vectors start with a pool of random codebook vectors, and then are modified by learning from each training data points. Predictions are made using the LVQ codebook vectors in the same way as KNN by searches through all codebook vectors for the most common class value (the nearest neighbor).

8. Support Vector Machines (SVM) is one of the most popular machine learning algorithms. SVM is a discriminative classifier (separation of classes) defined by a separating hyperplane. For a given labeled training datasets (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new data point. In two dimensional spaces, the hyperplane is a line dividing a plane into two parts

where each class lay in either side. The output of SVM is coefficients or weights that define an optimal hyperplane, which is the hyperplane that maximizes the margin (distance between the hyperplane and the closest data point for a given coefficient vector) between the two classes. The support vectors are the data points that lie closest to the hyperplane, and to support the separation of classes. SVM is the most efficient classifier used to solve real-world problems.

9. Bagging and Random Forest is one of the popular and powerful machine learning algorithms. It is a type of Bootstrap Aggregate or bagging algorithm. Bootstrap is an efficient statistical method for estimating a quantity, such as mean and standard deviation, from a data sample. For example, to estimate big mean, we can divide all samples into sub-samples, calculate mean of each sub-sample, and estimate big mean by averaging all of our collected means. For Bagging, Bagging is the application of Bootstrap procedure to a high-variance machine learning algorithm, typically decision tree in CART (classification and regression tree). For example, we had 5 bagged decision trees that made the following class predictions for each input sample: blue, blue, red, blue, and red; we would take the most frequent class (mode) and predict blue. Random Forests are an improvement over bagged decision trees. A problem with decision trees like CART is that they are greedy (use greedy algorithm to minimize error). In CART, when selecting a split point, the learning algorithm is allowed to look through all variables and all variable values in order to select the most optimal split-point. The random forest algorithm changes this procedure so that the learning algorithm is limited to a random sample of features of which to search. The output of the random forest is the mode of the sub-trees.

10. Boosting and AdaBoost Boosting is a method that creates a strong classifier from a number of weak classifiers. This is done by building a model from the training data, then creating a second model that attempts to correct the errors

from the first model. Models are added until the training set is predicted perfectly or a maximum number of models are added.

AdaBoost was the first really successful boosting algorithm developed for binary classification. Modern boosting methods built on AdaBoost such as stochastic gradient boosting machines. The most suited and most common algorithm used with AdaBoost is the decision trees. Initially, all observations are given equal weights. While creating the next model, higher weights are assigned to the data points which were predicted incorrectly (predictions are made on the whole dataset). Weights can be determined using the error value (differences between the prediction and the actual values), the higher the error. In other words, at the end of every model prediction, we end up boosting the weights of the misclassified data points so that the next model does a better job on them. The process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached. Note that the observations must not contain outliers.

Machine learning has been applied to various fields such as medical, industrial, and information security. In the medical area, the development and testing of a machine-learning-based system that predicts the risk of hypoxaemia and provides explanations of the risk factors in real time during general anaesthesia were reported (Lundberg et al., 2018). In addition, a live-primary-cell phenotypic-biomarker assay with single-cell resolution, and its validation with prostate cancer and breast cancer tissue samples for the prediction of post-surgical adverse pathology using machine learning was proposed in the year 2018 (Manak et al., 2018).

In 2019, a machine learning technique named as a Hidden Markov Model (HMM) was explored its potential role to validate the performance of the Framingham Diabetes Risk Scoring Model (FDRSM) as a well-respected prognostic model (Perveen, Shahbaz, Keshavjee, & Guergachi, 2019). Furthermore, the electronic health record (EHR) data at Stanford Health Care were used as a part of an implementation of a machine learning classifier to identify potential Familial hypercholesterolemia patients (Banda et al., 2019). Besides, the machine learning mechanisms are deployed for end-to-end drug discovery and development (Ekins et al., 2019).

Since malaria is an important issue for most countries in the world, so malaria parasite detection is a significant process in the malaria detection in every risk community. Unfortunately, the conventional detection method is expensive and time consuming as mentioned in (Olugboja & Wang, 2017); therefore, the use of stained blood smear images for malaria parasite detection cooperated with the machine learning classifier was proposed to replace the microscopic malaria diagnosis. Later, the different approaches according to techniques used for imaging, image preprocessing, parasite detection and cell segmentation, feature computation, and automatic cell classification were drawn (Poostchi, Silamut, Maude, Jaeger, & Thoma, 2018). In addition, an implementation of a new image processing system was introduced using machine learning algorithm to learn, detect and determine the types of infected cells according to its features for detection and quantification of *Plasmodium* parasites in a blood smear slide (Kunwar, Shrestha, & Shikhrakar, 2018). Based on the high benefits of using machine learning in malaria researches have obviously been seen, the use of mid-infrared spectroscopy and supervised machine-learning were deployed to identify vertebrate blood meals in the malaria vector, *Anopheles arabiensis* (Mwanga et al., 2019).

K-Means algorithm

K-Means algorithm is extremely easy to implement and very efficient computational method. Those are the main reasons that explain why this method is popular. Unfortunately, this method is not suitable to identify classes when dealing with groups that do not have the spherical distribution shape.

The K-Means algorithm aims to find and group data points into classes where the similarities between data are identified. In the terms of the algorithm, this similarity is understood as the opposite of the distance between data points. The closer the data points are the more similar and more likely to belong to the same cluster they will be. Following Figure11 shows a brief description of the standard k clustering algorithm.

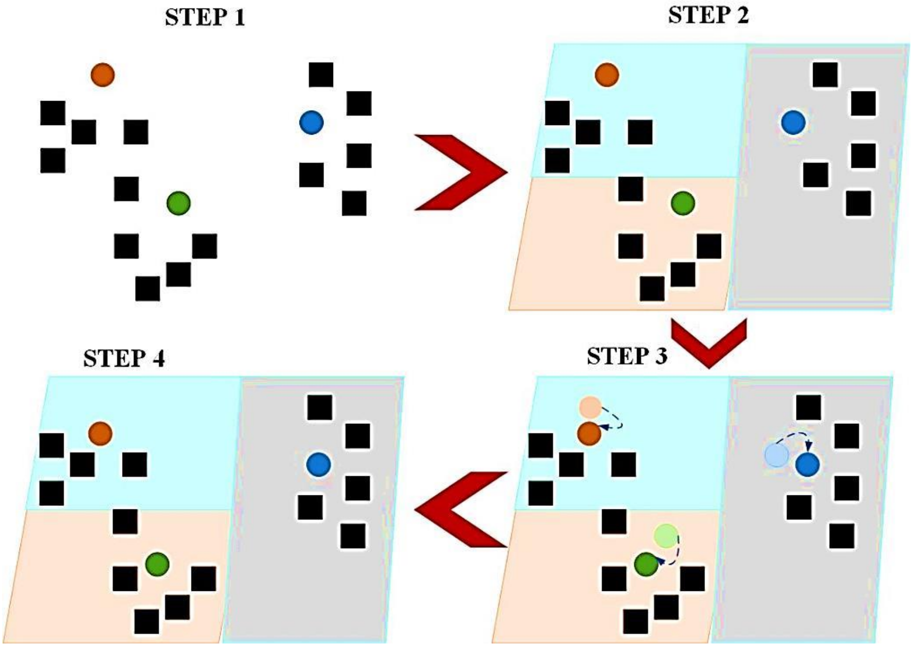


Figure 11: The K-Means algorithm

Figure 11 demonstrates the *K*-Means algorithm that classifies objects into *k*-different groups and it can be described as follow.

Step 1. Set number of *k* cluster

Step 2. Determine the centroid coordinate

Step 3. Determine the distance of each object to the centroids

Step 4. Group the object based on minimum distance

Then, the *k*-means algorithm will repeat step 2 to step 4 until convergence.

Geographic Information Systems and Risk map

The process of creating a risk map using the principles of geographic information systems: Geographic Information System or GIS is the working process on spatial information using a computer system. This manages spatial database in relation to the position on the map and maps in the system by inputting geographic data such as images maps, satellite imagery, numbers, distance, then analyze those data by computer programs. The results are often highly accurate can be applied in many aspects. GIS has benefits in many fields especially in environmental management, city planning, and utilities management by calculating the area to be used from the map image, such as distance measurement, or point determination on the map.

Geographic Information System or GIS has been applied to various fields, such as medical, economy, transportation, land use planning, and emergency and disaster management. In medical area, GIS had been applied since 2004 for developing an internet GIS map to monitor situations of illness and socio-economic for many researchers. This software was an open source with small cost of maintenance; it can also serve the specific needs (Maclachlan, Abernathy, & Jerrett, 2004). Then, (Nykiforuk & Flaman, 2009) had concluded the usage of GIS in health-related researches and also critically examine the issues strengths, and challenges inherent to those approaches from the lenses of health promotion and public health.

The GIS is not only supports the community illness protection process as mentioned above, but the uses of GIS also help solving public health and treatment of health problems in different geographic areas (Fradelos et al., 2014). For example, the spatial analysis of hospitals in the city of Mashhad uses arc GIS software and network analyst model (Bazargan, 2018); in Australia, the Australian Cancer Atlas had been developed with the mission for spatial modelling, visualization, and reporting of cancer estimation; the statistical methods are applied to incorporate changes in geographical patterns from a large volume of patients data (Duncan, Cramb, Aitken, Mengersen, & Baade, 2019). In addition, there is a usage of GPS data loggers to describe the impact of spatial-temporal movement patterns on malaria control in a high-transmission area of Northern Zambia (Hast et al., 2019). On the other hand, GPS tracking was implemented in an application for health and aging research under the objectives of daily lives evaluation (Fillekes, Giannouli, Kim, Zijlstra, & Weibel, 2019) as similar to the project of Ministry of Health and Population (MOHP), Cairo, Egypt (Ibrahim Ramzi & Abdl-Latif El-Bedawi, 2019).

Currently, geographic information systems for importing storing, analyze display data in spatial form can be used to link spatial data and descriptive data, to support decisions solving problems related to disease prevention and control planning. (Srivastava et al., 2003) has adopted a geographic information system to help managing the database of a malaria monitoring system in India and to develop models that can help planning and developing of the malaria control system , including identify areas with high patients and identify the areas under malaria risk. (Martin, Curtis, Fraser, & Sharp, 2002) has developed a geographic information system for malaria research, monitoring, surveillance and control in South Africa for a Medical Council. (Qayum, Arya, Kumar, & Lynn, 2015) has study socio-economic, epidemiological and geographic features based on GIS-integrated mapping to identify malarial hotspots. To study malaria hotspots, defined as areas where transmission intensity exceeds the common level, turn out to be extra mentioned as transmission declines. Targeting hotspots may accelerate mark downs in transmission and may want to be pivotal for malaria elimination (Platt, Obala, MacIntyre, Otsyula, & Meara, 2018). In addition, a provincial network in Uganda had studied the uses of remote detecting systems and spatial autocorrelation models to distinguish and organize the most productive Anopheles larval environments for controlling them (Tokarz & Novak, 2018).

Therefore, from the above reason, applying GIS to malaria studying is significantly important since the risk of the disease can be explained in various aspects, such as their basic characteristics, risk factors that contribute to the occurrence, and epidemic of the disease. Additionally, this GIS's implementation helps malaria management planning in each individual area. In such reason, the outcome from this research is the analytical model that can be an important tool to help predicting the distribution of malaria with higher accuracy and reliability. Furthermore, the risk maps can be drawn to better identify the coordinates of the epidemic area which the related organizations can use it for monitoring malaria distribution effectively. Moreover, this map helps planning to improve personal's malaria protection. Consequently, the result of this study serves as the basis for developing a risk model to monitor other diseases in the future.

Chapter 3

Research methodology

This chapter describes the experimental and analytical methods that are applied in this research. The first part of this chapter draws the experimental process started from collecting data until analyzing them. In the analysis phase, there are two significant parts: factor classification and model derivation where a risk model was derived. Details of each part are described as follow.

Experimental methods

This part provides details of the data collection methods and the data analysis methods those are important to this experiment.

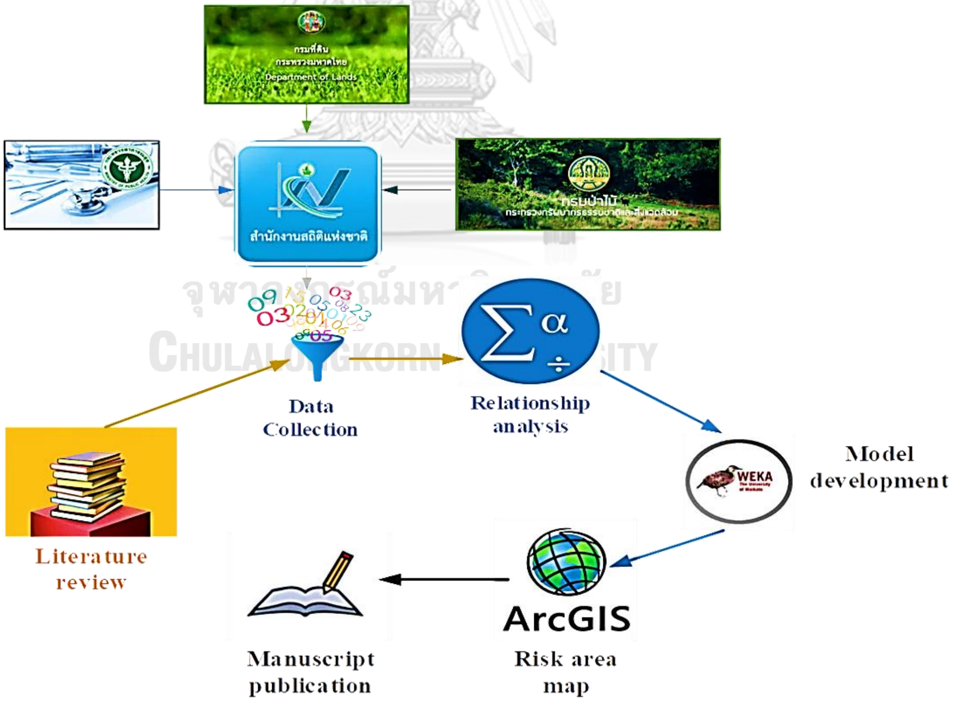


Figure 12: Experimental and analytical methods

Based on Figure 12, the experimental and analytical methods of this research are displayed. The first step of this research started from the literature review to obtain variables related to the distribution of malaria. Then, data of required variables were collected from the public database, filtered and cleaned due to the characteristics of data. Afterwards, step 3, the cleaned data were analyzed based on statistical methods to identify relationships among them using SPSSv22.0 (Chulalongkorn University License). The results from the statistical analysis will pass to step 4 to create a risk model using the Weka (v3.8) before drawing the risk map in step 5 using ArcGISv10.4 (Chulalongkorn University License). Lastly, the manuscript contains summarization of the results is written and published.

Data collection

Based on the results from various researches, many factors have impacts to the distribution and transmission of malaria, especially the demographic factors. Although the demographic factors are counted as simple variables, such as the quantity of household, and the range of population density, etc., these factors are significantly related to the distribution and transmission of malaria. Moreover, some researches indicated that the effect from the populace migration also exists and avoidable. Besides the demographic factors, the factors of socio-economic also have relation with the distribution of malaria as same as the demographic factors. Examples of the socio-economic factors are such as average monthly income, debt, money spent per month, and the proportion of lower income people. In addition, according to the literature reviews, the factors that should be studied for the distribution of malaria are housing information, the access to healthcare, weather information, land usage, transportation, water sources, and the access to information from the public database of relevant organizations during the years 2006-2018.

In Thailand, there are many organizations that store information related to malaria, such as the National Statistical Office (NSO), Department of Forestry, Department of Land Development, and Ministry of Public Health. Among these organizations, the most significant data center is NSO, a division of the Ministry of Digital Economy and Society, because this organization has mission as a central government division to operate statistical data in monitoring, evaluating, and supporting the implementation of government policies. The statistical data are obtained from every government department, including some specific surveys by the internal staffs. Therefore, data in this research is relied on the data from NSO. According to the types of data from NSO, the collection method can be classified into two types: the census method and the survey method.

Still, some data from NSO must be considered in details. Thus, the information from the Department of Forestry, Ministry of Natural Resources and environment, are included to complete the data context from NSO. The responsibilities of the Department of Forestry are conserving and managing of forest resources in the country. Based on their responsibilities, databases for public information services are available such as a database of the forest research, a database of plant types in the forest, a database of insects in the forest, and a database of biodiversity.

Based on some malaria researches, the data from the Land Development Department, Ministry of Agriculture and Cooperatives, are needed because this department is responsible for studying and researching soil and land to support strategic planning and developing of the land usages. In addition, knowledges about soil, water, plants, and fertilizer either general or obtained from the internal researches that related to land development are inspected and advised to farmers and government works. Besides, any activities that are determined by laws must be activated. In this research, the provincial land use data is used for analysis. Besides to external environmental factors, the number of provincial malaria cases has been included in the website of the Ministry of Public Health, which is responsible for health promotion, prevention, control and treatment of diseases danger rehabilitation of populations.

According to the use of public databases from various government agents, problems, such as duplication of data, missing data, and data that does not match, usually occur. Therefore, all collected data must pass the data filtering and cleaning process before sending to the analyzing procedure. In the data filtering and cleaning processes, the researcher chose the data from the National Statistical Office (NSO) as the dominant database because NSO has responsibility on every organization, both government and private sections, in Thailand. As a consequence of being a data center from all sections around Thailand, most required data factors can be retrieved easily from this organization. Nevertheless, some information from other organizations may be added to clarify the existing data from NSO such as the use of land from the Land Development Department or the forest area from the Royal Forest Department etc. Not only unclear data existed in the NSO databases, but also some missing data of some period are encountered. These missing data might occur during the data elicitation process at the sources or the real data value might be zero.

Data analysis

Define variables

After filtering and screening all data, the variables of these data are defined and grouped in some certain conditions. The first condition of the data grouping is based on the total number of patients within 13 years, from 2006 to 2018. The variable that stores the number of patients each year is named as NoPatient; and the group of patients is named as GrPatient. These two variables are dependent variables of the study.

According to the data of NoPatient within 13 years, the data in GrPatient can be classified into 2 groups: a high distribution group (HDG), and a low distribution group (LDG). Therefore, to classify the HDG and LDG, the NoPatient in each year from 2006 to 2018 will be sorted and the provinces within the rank of 1-10 were recorded and are marked as the HDG. The result from this grouping is 21 provinces in the HDG, and 56 provinces in the LDG.

Since the research focuses on factors that affect to the distribution of malaria patients in Thailand, the independent variables that cause changes of values of the NoPatient and the GrPatient must be determined. So, from the previous section, there are factors, such as demographic factor, the socio-economic factor, the sanitation factor, etc. Each factor contains many variables. For example, the variables in the demographic factor are a variable stores the number of houses, named as NoHouse, a variable stores the number of population density, named as NoPopulation, and a variable stores the number of migrated persons, named as NoMigrate, etc.

Referring to multiple factors mentioned previously, there are 71 main variables can be extracted from these factors. Moreover, each main variable was continuously record for 11 years, from 2007 to 2017, as needed. So, instead of considering only main variables, all sub-variables that store data for each year are extracted. For example, the main variable is “NoPopulation”, its sub-variables for 11 years could be NoPopulation07, NoPopulation08, …, and NoPopulation17. As a consequence of using this rule, there are 294 variables for entire 11 years which can be grouped to 71 main variables, such as NoPopulation, AverageIncome, AgriculturalArea, ForestArea, NoHouse, PovertyRatio, Migration, HouseMaterial, NoHospitality, etc. Though, using all 294 sub-variables of 71 main variables will be too complicated. Therefore, the mean value of each sub-variable that occurred more than two times within 11 years was calculated and be used as the representative of the main variable. For example, the mean value of NoPopulation07, NoPopulation08, …, and NoPopulation17 was calculated and this mean was used as the representative of the NoPopulation for the entire 11 years.

FactorClassification

Factor analysis is the technique used to reduce the number of factors by finding the structure of relationship between each factors, and find new factor called component. The component is the way to incorporate all variable that related with each other into the same component. The factors are on different component are the factors that do not have any correlation or very low correlation. In which the factor that high relation is factor 1, therefore, factor1 is selected and analyzed in the next step. The process of applying factor analysis is shown in Figure 13.

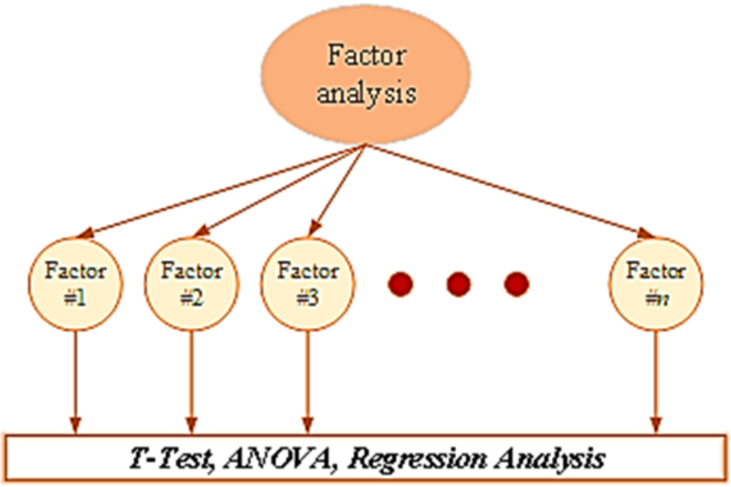


Figure 13: Relationship analysis

As the fact that 71 independent variables are a large number of variables, the complication in modelling analysis can be occurred. Thus, the similarity among these variables can be indicated using Factor analysis method. Factor analysis is a method of grouping variables that are related or have the same qualities as one group by studying the structure of relationship between variables, and creating new variables, called elements which are collective of variables that are highly correlated within the same group. Therefore, the relation between variables in different elements is too small to be accepted. Moreover, only some variables in the same element can be chosen to be analyzed as needed, without using every variable in the same element. As a consequence, the number of variables that is used in the analysis process will be small and not complicated while the reliable results can be obtained as much as using all variables. Thus, it can be said that this technique helps to reduce the number of variables in the analysis process, but is still able to maintain the value and meaning of the information of the original data completely. The factor score of each group of variables can be used as a factor in various statistical analyses, such as regression and correlation analysis, analysis of variance - ANOVA, Discriminant analysis, etc. In addition, variable grouping can also prevent multicollinearity problems for regression analysis (Menke, 2018). Therefore, this method studies the structure of relationship between variables, and creating new variables, called as elements which are collective of variables that are highly correlated within the same group. On the other hand, the relation between variables in different elements is too small to be accepted. The consequence of separating variables to elements, only some variables in the same element can be chosen to be analyzed as needed. As a result, the number of variables in the analysis process will be small and not too complicated whilst the reliable results can be obtained as same as using all variables.

Analyzed using regression analysis is performed to study the influence of factors on the spread of malaria which is likelihood or risk of malaria distribution from various factors. Multiple linear regression analysis is a statistical method to analyze the relationship between independent variables and a dependent variable based on the assumption that the relationship between dependent variables and the measured variables are linear. The result of regression analysis indicates whether there is a relationship between an independent variable and the dependent variable in a linear form or not. Then, a linear regression model that shows the relationship between dependent variables and measured variables will be calculated. This linear regression model is used to predict the value of the dependent variable when the values of all independent variables in the regression model are known. This model is used for studying the risk of malaria distribution from various factors (Pandis, 2016).

The result of the regression analysis is a linear regression model that shows the relationship between various independent variables (x) with dependent variable (y), the y variable is the number of malaria patients and the x variable is the factors related with malaria including environmental factor (such as the amount of rain, temperature), socio-economic factors (such as household income in the study area). If we know the value of various independent variables, we can predict the number of malaria patients in that area using linear regression generated. The independent variables used in the analysis process must have supporting theoretical evidences or related research reports.

Using 95% confident level to run Factor analysis on SPSSv.22 of Chulalongkorn University license, there are 9 factors to be categorized. These factors are applied as independent variables (x_i) in the regression analysis method where the dependent variable is the number of patients in each province(y_i). y_i variable is the number of malaria patients and the x_i variable is the factor (*i*) related with malaria including environmental factor (such as the amount of rain, temperature), socio-economic factors (such as household income in the study area). Thus, knowing the value of various independent variables, prediction of the number of malaria patients in an area using the generated linear regression model,$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_9 x_9$ should be performed.

To generate a regression model by assuming to be linear,$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_9 x_9$, the result from running regression analysis indicated that there is at least one independent factor that has effect to the number of patients. However, the value of R^2 is close to zero and the standard error of estimation is higher than 500 points. Therefore, there is a possibility that the regression model among the number of patients and factors is in the non-linear form. As a consequence, the machine learning mechanism is applied to derive a suitable relationship model among all these variables.

Variable determination

As a result of the regression analysis mentioned above, it indicated that the regression model based on classified factors was not a proper solution. Thus, the machine learning method using Weka 3.8, a freeware, is applied to determine a proper relation model of the malaria distribution. Nonetheless, instead of running all 9 factors, only Factor_1 obtained from Factor analysis is considered because variables in this category have the highest correlation value. As a consequence, all 28 variables in Factor_1 were applied to Weka 3.8; these variables are listed in Table 4 below.

Table 4: List of variables that are applied to machine learning mechanism.

Name of variable	Description
Physical	
NoHouse	Number of houses
NoHouseTapW	Number of houses with tap water
NoHouseNoElec	Number of houses without electricity usage
NoHouseWithElec	Number of houses with electricity usage
NoPoverty	Number of poverty
NoGovHealthCareUnt	Number of Government's healthcare unit
NoPatientBed	Number of patients' beds in the government hospital or healthcare unit
NoPrvHospital	Number of private hospitals
NoOfHealth	Number of healthcare officers
ElecFromGov	Electricity from government section
NoMobile	Number of mobile phones
NoComputer	Number of computers
NoInternet	Number of Internet
NoRegVehicle	Number of registered vehicles
NoCommSources	Number of communication sources
NoRiver	Number of rivers
NoCanal	Number of canals
NoSW	shallow well
NoAW	Number of artesian well
NoIrrC	Irrigation Canal
NoStreet	Number of streets

Table 4: List of variables that are applied to machine learning mechanism.

Name of variable	Description
Biological	
NoPopulation	Number of population density in a province
NoFamilyFamer	Number of famers' families
NoNonMigrate	Number of non-migrated persons
NoMigrate	Number of migrated persons
ProvArea	Provincial area
ProvAgr	Provincial agricultural area
ProvNonAgr	Provincial non-agricultural area

Model derivation

This section describes the model derivation method using 28 variables through Weka 3.8. Since the previous result of the running linear regression showed that the linear model was not a suitable model for predicting the distribution of malaria, this model was not selected when running machine learning process. Machine learning technique is used to predict complex data with high accuracy measuring by false positive and false negative values, or using precision and recall values. Table 5 shows types of various machine learning mechanisms with their characteristics.

Table 5: Types of machine learning mechanism.

Support Vector Machines	
Pros	Cons
- Can model complex, nonlinear relationships	- Need to select a good kernel function - Model parameters are difficult to

Table 5: Types of machine learning mechanism.

- Robust to noise (because they maximize margins) - Performs similarly to logistic regression when linear separation - Performs well with non-linear boundary depending on the kernel used - Handle high dimensional data well	interpret - Sometimes numerical stability problems - Requires significant memory and processing power - Susceptible to overfitting/training issues depending on kernel
K-Nearest Neighbors	
Pros	**Cons**
- Simple - Powerful - No training involved ("lazy") - Naturally handles multiclass classification and regression	- Expensive and slow to predict new instances - Must define a meaningful distance function - Performs poorly on high-dimensionality datasets
Linear regression	**(all features must exist)**
Pros	**Cons**
- Very fast (runs in constant time) - Easy to understand the model - Less prone to overfitting	- Unable to model complex relationships - Unable to capture nonlinear relationships without first transforming the inputs

Table 5: Types of machine learning mechanism.

Logistic Regression	Not all features are required to complete the estimation.
Pros	Cons
- low variance - provides probabilities for outcomes - works well with diagonal (feature) decision boundaries	- high bias
Decision Trees	Not all features are required to complete the estimation.
Pros	Cons
- easy to interpret visually when the trees only contain several levels - Can easily handle qualitative (categorical) features - Works well with decision boundaries parallel to the feature axis - Fast - Robust to noise and missing values - Accurate	- prone to overfitting - possible issues with diagonal decision boundaries - Complex trees are hard to interpret - Duplication within the same sub-tree is possible
Naive Bayes	all features must exist
Pros	Cons
- Computationally fast - Simple to implement - Works well with high dimensions	- Relies on independence assumption and will perform badly if this assumption is not met

Table 5: Types of machine learning mechanism.

Random Forest	all features must exist
Pros	Cons
Decorrelates trees (relative to bagged trees) important when dealing with multiple features which may be correlated reduced variance (relative to regular trees)	Not as easy to visually interpret

Although there are various classification techniques, such as K-Nearest Neighbors, Logistic Regression, Naive Bayes Classifier, Decision Trees, and Support Vector Machines, the suitable classification method is depended on the characteristics of the collected data. Hence, the suitable classification models are in the category of the Decision Trees. One interesting sub-category of the decision trees named as Logistic Model Tree, LMT. LMT is a classification model introduced by Landwehr, et.al. in the year 2005 (Landwehr, Hall, & Frank, 2005). LMT is a classification model with an associated supervised training algorithm that combines logistic regression (LR) and decision tree learning. Logistic model trees are based on the earlier idea of a model tree: a decision tree that has linear regression models at its leaves to provide a piecewise linear regression model

This model is a combination of the logistic regression model and the decision tree. Furthermore, this model fits the data with two independent groups. As a result of using logistic regression in deriving the LMT, the explicit class probability estimation can be performed as it is one significant advantage from the use of logistic regression. In addition, this model avoids confounding effects by analyzing the association of all variables together (Sperandei, 2014). Table 6 shows the algorithm of LMT (Landwehr et al., 2005).

Table 6 The algorithm of LMT (Landwehr et al., 2005)

```
LMT(examples){
    root = new Node()
    alpha = getCARTAlpha(examples)
    root.buildTree(examples, null)
    root.CARTprune(alpha)
}
buildTree(examples, initialLinearModels) {
   numIterations =
      CV_Iterations(examples,initialLinearModels)
    initLogitBoost(initialLinearModels)
    linearModels = copyOf(initialLinearModels)
    for i = 1...numIterations
        logitBoostIteration(linearModels,examples)
    split = findSplit(examples)
    localExamples = split.splitExamples(examples)
    sons = new Nodes[split.numSubsets()]
```

```
      for s = 1…sons.length
          sons.buildTree(localExamples[s],nodeModels)
}
CV_Iterations(examples,initialLinearModels) {
     for fold = 1…5
          initLogitBoost(initialLinearModels)
          //split into training/test set
          train = trainCV(fold)
          test = testCV(fold)
           linearModels = copyOf(initialLinearModels)
           for i = 1…200
               logitBoostIteration(linearModels,train)
               logErrors[i] += error(test)
       numIterations = findBestIteration(logErrors)
       return numIterations
}
```

The purpose of this research is to develop a model for predicting the distribution of malaria, but the results of the data analysis show that the data is unbalanced according to the number of data in one group is much larger than another. So, the researcher has solved the problem by balancing these data sets using the technique of random sampling technique (Synthetic Minority Over – Sampling Technique: SMOTE) and develop the model using decision tree techniques.

The unbalancing of the target groups has effects on the validity of the prediction equation. Therefore, it is the problem that every researcher must be concerned as the data imbalances are often encountered in the real world. Consider the medical information such as information on variables related to illness, the number of patients admitted to the hospital is high, but the number of patients diagnosed with cancer is small compared to the total number of patients. Besides, some diseases may be difficult to find in some area such as African swine fever (ASF) is difficult to find in ASIA but it commonly spreads in Africa. Thus, studying this disease in ASIA using machine learning and data mining with unbalancing data sets before classification, the classification result will be less accurate.

K-Means Clustering for mapping with ArcGIS 10.4

Although the density of patients is classified in two levels, high and low densities, the level of risk factors' distribution is also divided to 2 for clearly controllable and manageable. These levels are high level, Intermediate level, and low level of distributions. Once levels of risk factors are considered to be 2, the benefit of surveillance and control in the distribution of the budget thoroughly can be obtained. Nevertheless, the criteria to classify values of all 28 risk-factors to be 2 levels will be performed by machine learning mechanism, namely K-Means Clustering algorithm.

The processes of K-Means Clustering algorithm are listed below.

1. Before starting the K-Means Clustering algorithm, the number of clusters K must be specified. In this research, K value is 2.

2. The clusters must be shifted among each other; then, K data points are randomly selected for initiating centroids without replacement.
3. Repeat step#2 till there is no alteration to the centroids; the data points in the assigned clusters are not altered.
4. Each data point is assigned to a cluster when the sum of the square distance between data points and a centroid is the smallest value; the sum of the square distances between data points and all centroids must be computed and compared.
5. Identify the centroid for each cluster by computing the mean of the all data points that belong within the cluster.

The result from K-Means Clustering algorithm is used to derive a malaria spreading map of 2-levels according to the 28 risk-factors, drawing by ArcGIS program licensing by Chulalongkorn University. Like the risk-factor mapping, 2-level distribution map of patients around Thailand was drawn. Then, comparison between the risk-factor map and patients' distribution map is performed to determine the similarity and efficiency of all 28 risk-factors from the LMT model.

To achieve an efficiency managent in the malaria monitoring and protecting procedures, separated areas to be only 2 levels is not a proper solution because some provinces in the high-risk group might not need a high caring for all times as same as others in its group. Therefore, the risk area is considered to be 3 distribution's classes: high, medium, and low. In such case, the K-Mean Clustering technique, with K = 3, is applied to create the risk map of malaria distribution based on 28 factors. The following processes listed below are the risk-map creation process.

Step 1. Using K-Mean Clustering technique using K = 3, the set of low distribution of malaria obtained from the K-Mean method was 100% similar to the set of low distribution based on the number of patients. Though, to confirm the correctness of the sets of medium and high risk provinces, the observations of number of patients in every provience are performed; the results of this process indicated that there were some inconsistencies among number of patients and the defined ranks obtained from the K-Means Clustering algorithm. As a consequence, step 2 continues.

Step 2. Merges 2 sets of high and medium risk areas to be just one set, run K-Mean Clustering technique using K = 2 to separate this new set into 2 new subsets, namely high and medium groups. Two new groups are verified their suitabilities using LMT technique; the results from the LMT technique indicates that the new separations are suitable because their TPRs were high.

Step 3. Draw the 3-level risk map using provinces classified in step 1 and 2.

With these 6 factors, all provinces are arranged as same as the provinces that are arranged by the number of patients. The results are elaborated in Chapter 4.

Chapter 4

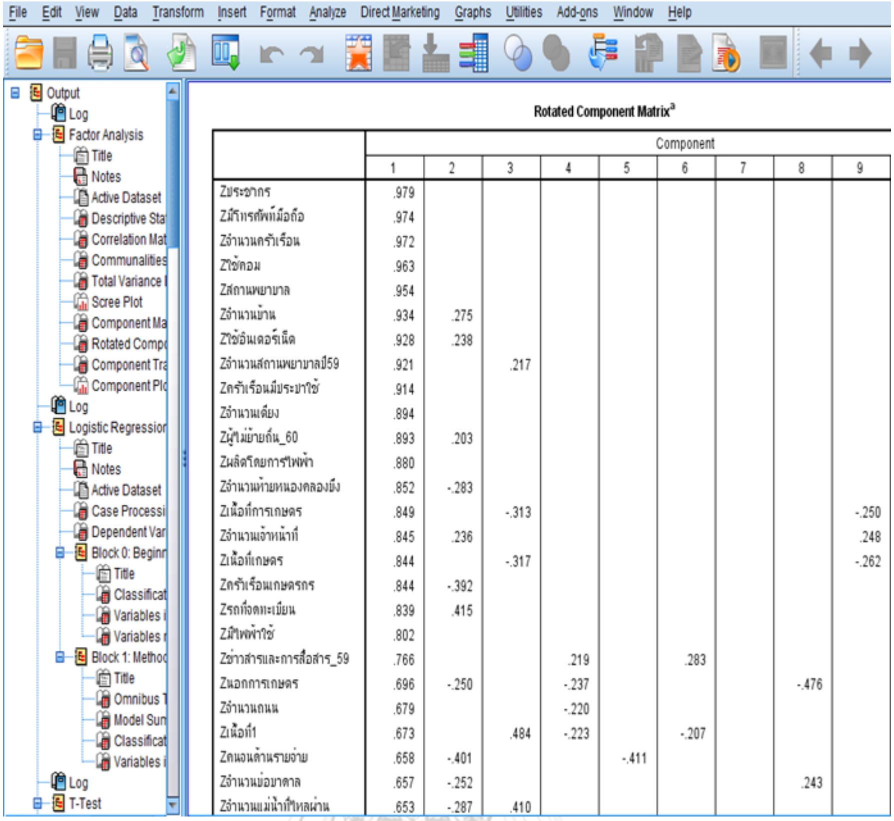
Results

This chapter describes the analysis results using a machine learning method which is considered as a risk model of malaria distribution with 28 new risk-factors related to malaria distribution. In addition, the results are analyzed using K-Mean Clustering technique to divide the level of surveillance at the local level by displaying the map images using the ArcGIS program.

Analyze the correlation between various factors

Based on the Factor analysis method for 294 independent variables with 71 non-repeated variables, there are 9 groups of factors can be classified, as shown in the rotated component matrix in Figure 14. The rotated component matrix is used to reduce the number of factors on which the variables under investigation have high loadings. In the Figur14, the variables are sorted and the items that have the highest loading (not considering whether the correlation is positive or negative) from factor 1 (28 variables in this analysis) are listed first. The variables are sorted from the one with the highest weight or loading (i.e., Number of population density in a province, with a loading of 0.979) to the one with the lowest loading from that first factor 0.653.

Once the variables are classified to 9 overlapping factors, the regression analysis based on variables in Factor 1 as independent variables and the number of patients each year is the dependent variable is calculated and discovered that they have a high Standard Deviation and the relationship is nonlinear. Therefore, all variables in factor 1 are passed to the machine learning procedures to form a suitable risk model.



Rotated Component Matrix[a]

	Component								
	1	2	3	4	5	6	7	8	9
Zประชากร	.979								
Zมีโทรศัพท์มือถือ	.974								
Zจำนวนครัวเรือน	.972								
Zใช้คอม	.963								
Zสถานพยาบาล	.954								
Zจำนวนบ้าน	.934	.275							
Zใช้อินเตอร์เน็ต	.928	.238							
Zจำนวนสถานพยาบาลปี59	.921		.217						
Zครัวเรือนมีประปาใช้	.914								
Zจำนวนเตียง	.894								
Zผู้ไม่ย้ายถิ่น_60	.893	.203							
Zผลิตโดยการไฟฟ้า	.880								
Zจำนวนห้วยหนองคลองบึง	.852	-.283							
Zเนื้อที่การเกษตร	.849		-.313						-.250
Zจำนวนเจ้าหน้าที่	.845	.236							.248
Zเนื้อที่เกษตร	.844		-.317						-.262
Zครัวเรือนเกษตรกร	.844	-.392							
Zรถที่จดทะเบียน	.839	.415							
Zมีไฟฟ้าใช้	.802								
Zข่าวสารและการสื่อสาร_59	.766			.219		.283			
Zนอกการเกษตร	.696	-.250		-.237				-.476	
Zจำนวนถนน	.679			-.220					
Zเนื้อที่1	.673		.484	-.223		-.207			
Zคนจนด้านรายจ่าย	.658	-.401			-.411				
Zจำนวนบ่อบาดาล	.657	-.252						.243	
Zจำนวนแม่น้ำที่ไหลผ่าน	.653	-.287	.410						

Figure 14: Rotated component matrix

Table 7: List of variables that are applied to machine learning mechanism.

Name of variable	
Physical	**Description**
NoHouse	Number of houses
NoHouseTapW	Number of houses with tap water
NoHouseNoElec	Number of houses without electricity usage
NoHouseWithElec	Number of houses with electricity usage
NoPoverty	Number of poverty
NoGovHealthCareUnt	Number of Government's healthcare unit
NoPatientBed	Number of patients' beds in the government hospital or healthcare unit

Name of variable	
Physical	**Description**
NoPrvHospital	Number of private hospitals
NoOfHealth	Number of healthcare officers
ElecFromGov	Electricity from government section
NoMobile	Number of mobile phones
NoComputer	Number of computers
NoInternet	Number of Internet
NoRegVehicle	Number of registered vehicles
NoCommSources	Number of communication sources
NoRiver	Number of rivers
NoCanal	Number of canals
NoSW	Shallow well
NoAW	Number of artesian well
NoIrrC	Irrigation Canal
NoStreet	Number of streets
Biological	**Description**
NoPopulation	Number of population density in a province
NoFamilyFamer	Number of famers' families
NoNonMigrate	Number of non-migrated persons
NoMigrate	Number of migrated persons
ProvArea	Provincial area
ProvAgr	Provincial agricultural area
ProvNonAgr	Provincial non-agricultural area

Malaria Distribution Risk Model

Based on 28 variables from the previous Chapter, a risk model can be derived using a machine learning method and a LMT model is obtained. In order to indicate the suitability of the obtained risk model, the indicators to identify the correctness of prediction are obtained from the confusion matrix below.

Table 8: Confusion matrix

Result from prediction	Real situation	
	True	False
True	True Positive (TP)	False Positive (FP)
False	False Negative (FN)	True Negative (TN)

Using the confusion matrix, this research used the values of correctly classified instances (*CCIs*), incorrectly classified instances (*ICIs*), TP rates (*TPRs*), FR rates (*FPRs*), precision, and recall. Each index was calculated as follows.

The total number of sample classes: $N = TP + FP + FN + TN$

The number of correctly classified classes: $n_{correct} = TP + TN$

The number of incorrectly classified classes: $n_{incorrect} = FP + FN$

Each performance index was be calculated as follows.

Percentage of Correctly Classified Instances (CCI)

Percentage of Incorrectly Classified Instances (ICI)

$$CCI = \frac{n_{correct}}{N} \times 100$$

$$ICI = \frac{n_{incorrect}}{n} \times 100$$

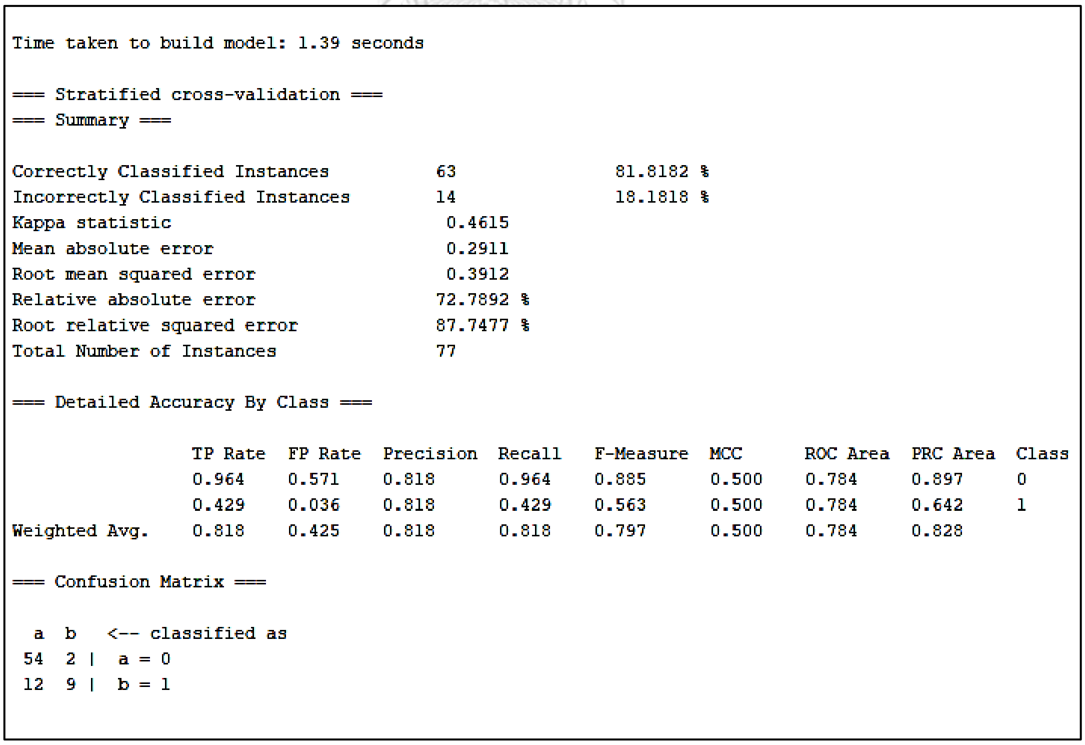
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = TPR = \frac{TP}{TP + FN}$$

Once the data from 28 variables are input to the LMT method using Weka3.8, the results of all indexes are displayed in Table 9.

Table 9: Performance Indexes of LMT for malaria distribution in Thailand (CCI = 81.8182%, ICI = 18.1818%)

Class	TPR	FPR	Precision	Recall
Low distribution (0)	0.964	0.571	0.818	0.964
High distribution (1)	0.429	0.036	0.818	0.429
Weighted Avg.	0.818	0.425	0.818	0.818



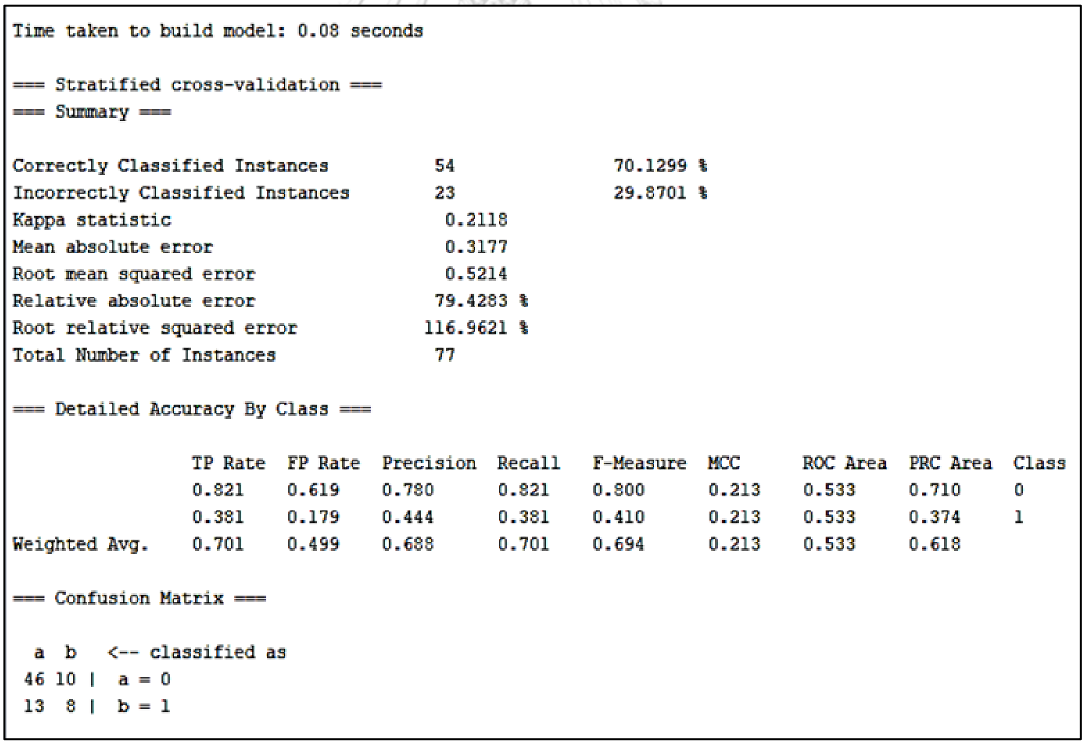
```
Time taken to build model: 1.39 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          63               81.8182 %
Incorrectly Classified Instances        14               18.1818 %
Kappa statistic                          0.4615
Mean absolute error                      0.2911
Root mean squared error                  0.3912
Relative absolute error                 72.7892 %
Root relative squared error             87.7477 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.964    0.571    0.818      0.964    0.885      0.500    0.784     0.897     0
                 0.429    0.036    0.818      0.429    0.563      0.500    0.784     0.642     1
Weighted Avg.    0.818    0.425    0.818      0.818    0.797      0.500    0.784     0.828

=== Confusion Matrix ===

  a  b   <-- classified as
 54  2 |  a = 0
 12  9 |  b = 1
```

Figure 15: Performance indexes of LMT for malaria distribution in Thailand

Since LMT is a type of decision trees classifiers where its leaf nodes are in the form of logistic regression model, the process to compare between basic decision tree (J48) and LMT is executed. The performance indexes of the basic decision tree are drawn in Table 10 below.

Table 10: Performance Indexes of basic decision tree (J48) for malaria distribution in Thailand (CCI = 70.1299%, ICI = 29.8701%)

Class	TPR	FPR	Precision	Recall
Low distribution (0)	0.821	0.619	0.780	0.821
High distribution (1)	0.381	0.179	0.444	0.381
Weighted Avg.	0.701	0.499	0.688	0.701

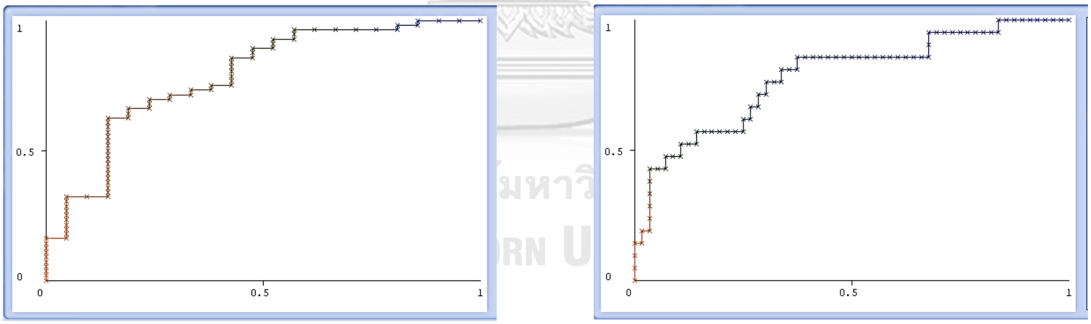


```
Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          54               70.1299 %
Incorrectly Classified Instances        23               29.8701 %
Kappa statistic                          0.2118
Mean absolute error                      0.3177
Root mean squared error                  0.5214
Relative absolute error                 79.4283 %
Root relative squared error            116.9621 %
Total Number of Instances               77

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.821    0.619    0.780      0.821    0.800      0.213    0.533     0.710     0
                 0.381    0.179    0.444      0.381    0.410      0.213    0.533     0.374     1
Weighted Avg.    0.701    0.499    0.688      0.701    0.694      0.213    0.533     0.618

=== Confusion Matrix ===

  a  b   <-- classified as
 46 10 |  a = 0
 13  8 |  b = 1
```

Figure 16: Performance Indexes of basic decision tree (J48) for Malaria distribution in Thailand

Based on Table 9 and Table 10 above, it is clear that LMT is more suitable model than basic decision tree (J48). Therefore, it can be concluded that the spread of malaria can be demonstrated in the form of Logistic Model Tree with the risk variables of 28 items.

In significant method to identify the accuracy of the model diagnostic is the use of the Area Under Curve analysis as known in the name of the AUC analysis. This method can classify binary classes; for example, AUC is used to differentiate the area of the HDG, and the LDG. The AUC is the area under the ROC curve derived to elaborate the relationship between the TPR (y-axist) and FPR (x-axist). Then, the AUC value is calculated; if the value of the derived AUC is close to 1, the accuracy of the classifier is acceptable. Thus, the accuracy of the logistic model tree can be validated using the AUC curve, as shown in Figure 17; and it has the AUC value is 0.784 for both groups: HDG and LDG.



(a) Low distribution: AUC=0.784

(b) High distribution: AUC=0.784

Figure 17: AUC curve of malaria distribution for both groups

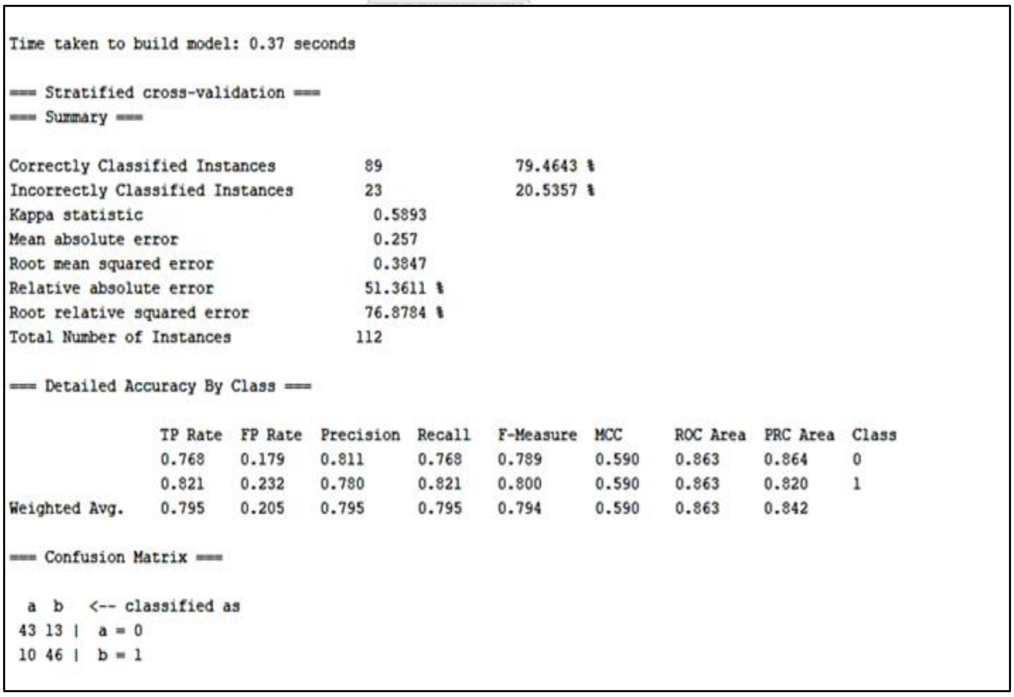
According to the result from AUC curve test, it can be concluded that the LMT is the suitable risk of malaria distribution which consists of 28 main variables.

Nevertheless, the results from analysis by LMT method indicated that the classified data set is imbalanced. This imbalance is a significant problem because the

TPR value is less than 0.429 (class1); this caused by the minority data has not been classified efficiently. Thus, resolving the problem of imbalanced data by means of additional data synthesis (Synthetic Minority Oversampling TEchnique: SMOTE), which is a special sampling technique of random sampling. instead of random sampling using the same data, SMOTE will synthesize the new from the existing one. The results of running LMT method with the new balance data set is displayed in Table11. below.

Table 11: Performance Indexes of LMT for Malaria distribution in Thailand (Synthetic Minority Oversampling Technique: SMOTE, CCI = 79.4643%, ICI = 20.5357%)

Class	TPR	FPR	Precision	Recall
Low distribution (0)	0.768	0.179	0.811	0.768
High distribution (1)	0.821	0.232	0.780	0.821
Weighted Avg.	0.795	0.205	0.795	0.795



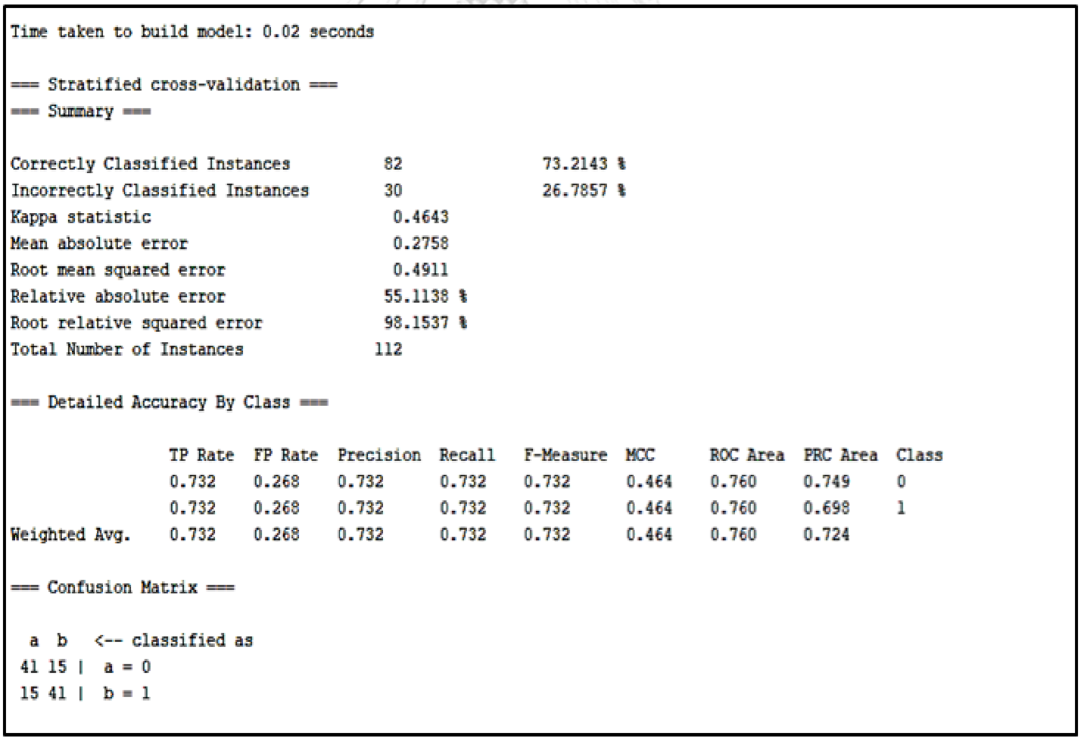
```
Time taken to build model: 0.37 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          89               79.4643 %
Incorrectly Classified Instances        23               20.5357 %
Kappa statistic                          0.5893
Mean absolute error                      0.257
Root mean squared error                  0.3847
Relative absolute error                 51.3611 %
Root relative squared error             76.8784 %
Total Number of Instances              112

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.768    0.179    0.811      0.768    0.789      0.590    0.863     0.864     0
                 0.821    0.232    0.780      0.821    0.800      0.590    0.863     0.820     1
Weighted Avg.    0.795    0.205    0.795      0.795    0.794      0.590    0.863     0.842

=== Confusion Matrix ===

  a  b   <-- classified as
 43 13 |  a = 0
 10 46 |  b = 1
```

Figure 18: Performance Indexes of LMT for Malaria distribution in Thailand (Synthetic Minority Oversampling Technique: SMOTE)

Since LMT is a type of decision trees classifiers, the process to comparison between basic decision tree (J48) and LMT is executed. The performance indexes of the basic decision tree are drawn in Table12 below.

Table 12: Performance Indexes of basic decision tree (J48) for malaria distribution in Thailand (Synthetic Minority Oversampling Technique: SMOTE, CCI = 73.2143%, ICI = 26.7857%)

Class	TPR	FPR	Precision	Recall
Low distribution (0)	0.732	0.268	0.732	0.732
High distribution (1)	0.732	0.268	0.732	0.732
Weighted Avg.	0.732	0.268	0.732	0.732

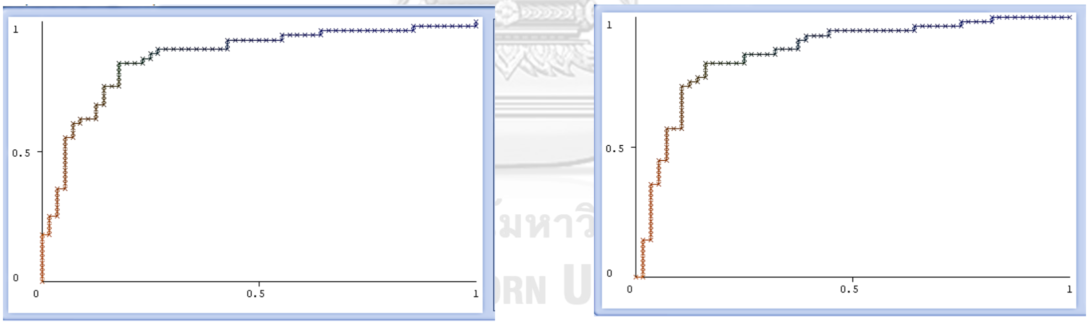


```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          82               73.2143 %
Incorrectly Classified Instances        30               26.7857 %
Kappa statistic                          0.4643
Mean absolute error                      0.2758
Root mean squared error                  0.4911
Relative absolute error                 55.1138 %
Root relative squared error             98.1537 %
Total Number of Instances              112

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.732    0.268    0.732      0.732    0.732      0.464    0.760     0.749     0
                 0.732    0.268    0.732      0.732    0.732      0.464    0.760     0.698     1
Weighted Avg.    0.732    0.268    0.732      0.732    0.732      0.464    0.760     0.724

=== Confusion Matrix ===

  a  b   <-- classified as
 41 15 |  a = 0
 15 41 |  b = 1
```

Figure 19: Performance Indexes of basic decision tree (J48) for Malaria distribution in Thailand (Synthetic Minority Oversampling Technique: SMOTE)

Although data were passing through the SMOTE and based on Table 11 and Table 12. above, it is clear that LMT is the suitable model than the basic decision tree (J48). Therefore, it can conclude that the spread of malaria can be demonstrated in the form of Logistic Model Tree with the risk variables of 28 items. Additionally, the accuracy in predicting the spreading of malaria, both Low and high distributions, is much better than the data without passing the SMOTE.

The AUC is the area under the ROC curve derived to elaborate the relationship between the TPR (y-axist) and FPR (x-axist). Then, the AUC value is calculated; if the value of the derived AUC is close to 1, the accuracy of the classifier is acceptable. Thus, the accuracy of the logistic model tree can be validated using the AUC curve as shown in Figure20 ; and it has the AUC value is 0.863 for both groups: HDG and and LDG.



(a) Low distribution: AUC=0.863

(b) High distribution: AUC=0.863

Figure 20: AUC curve of malaria distribution for both groups

According to the result from AUC curve test, it can conclude that the logistic model tree is the suitable risk of malaria distribution which consists of 28 main variables.

Ranking Classification

Based on K-Mean Clustering and the defined priority factors in Chapter 3, 77 provinces in Thailand were separated into 2 different groups: HDG and LDG. Nonetheless, the provinces in two groups were re-arranged and managed to be 3 different risk areas: high, medium, and low risk areas. The criteria to classify the provinces in each group is based on the results from K-Mean Clustering mixing with the result from the refining state of the map. Therefore, members of high risk area are the HDG members from both K-Mean Clustering and the refining state of the map. Like member of high risk area, the member of low risk area provinces are LDG members from both K-Mean Clustering and the refining state of the map. Then, the provinces that exclude from both criteria will be the members of the medium risk area. Table 13 is the list of all provinces in 3 risk-areas. Examples of high risk area are such as Chanthaburi, Chumphon, Tak etc. while examples of medium risk area are such as Chiang Rai, Chiang Mai, Si Sa Ket etc. The remaining provinces that are arranged as the low risk area are such as Krabi, Kamphaeng Phet, Cha Choeng Sao.

Table 13: Ranking Classification Risk Area

High risk area	Medium risk area	Low risk area
1. Chanthaburi	1. Chiang Rai	1. Krabi
2. Chumphon	2. Chiang Mai	2. Kamphaeng Phet
3. Tak	3. Si Sa Ket	3. Cha Choeng Sao
4. Narathiwat	4. Song Khla	4. Chai Nat
5. Prachuap Khiri Khan	5. Surat Thani	5. Trang
6. Prachin Buri	6. Ubon Ratchathani	6. Trat
7. Pattani		7. Nakhon Nayok
8. Phang Nga		8. Nakhon Pathom
9. Mae Hong Son		9. Nakhon Phanom

High risk area	Medium risk area	Low risk area
10. Yala		10. Nonthaburi
11. Ranong		11. Nan
12. Ratchaburi		12. Bueng Kan
13. Lampang		13. Pathum Thani
14. Lamphun		14. PhraNakhonSiAyuttaya
15. Kanchanaburi		15. Phayao
		16. Phatthalung
		17. Phichit
		18. Phetchaburi
		19. Phrae
		20. Phuket
		21. Maha Sarakham
		22. Mukdahan
		23. Yasothon
		24. Rayong
		25. Lop Buri
		26. Loei
		27. Satun
		28. Samut Prakan
		29. Samut Songkhram
		30. Samut Sakhon
		31. Sa Kaew
		32. Saraburi
		33. Sing Buri
		34. Sukhothai

High risk area	Medium risk area	Low risk area
		35. Suphan Buri 36. Nong Khai 37. Nongbua Lumphoo 38. Ang Thong 39. Umnad Chareun 40. Uttaradit 41. Uthai Thani 42. Bangkok 43. Kalasin 44. Khon Kaen 45. Chon Buri 46. Chaiyaphum 47. Nakhon Ratchasima 48. Nakhon Si Thammarat 49. Nakhon Sawan 50. Buri Ram 51. Phitsanulok 52. Phetchabun 53. Roi Et 54. Sakon Nakhon 55. Surin 56. Udon Thani

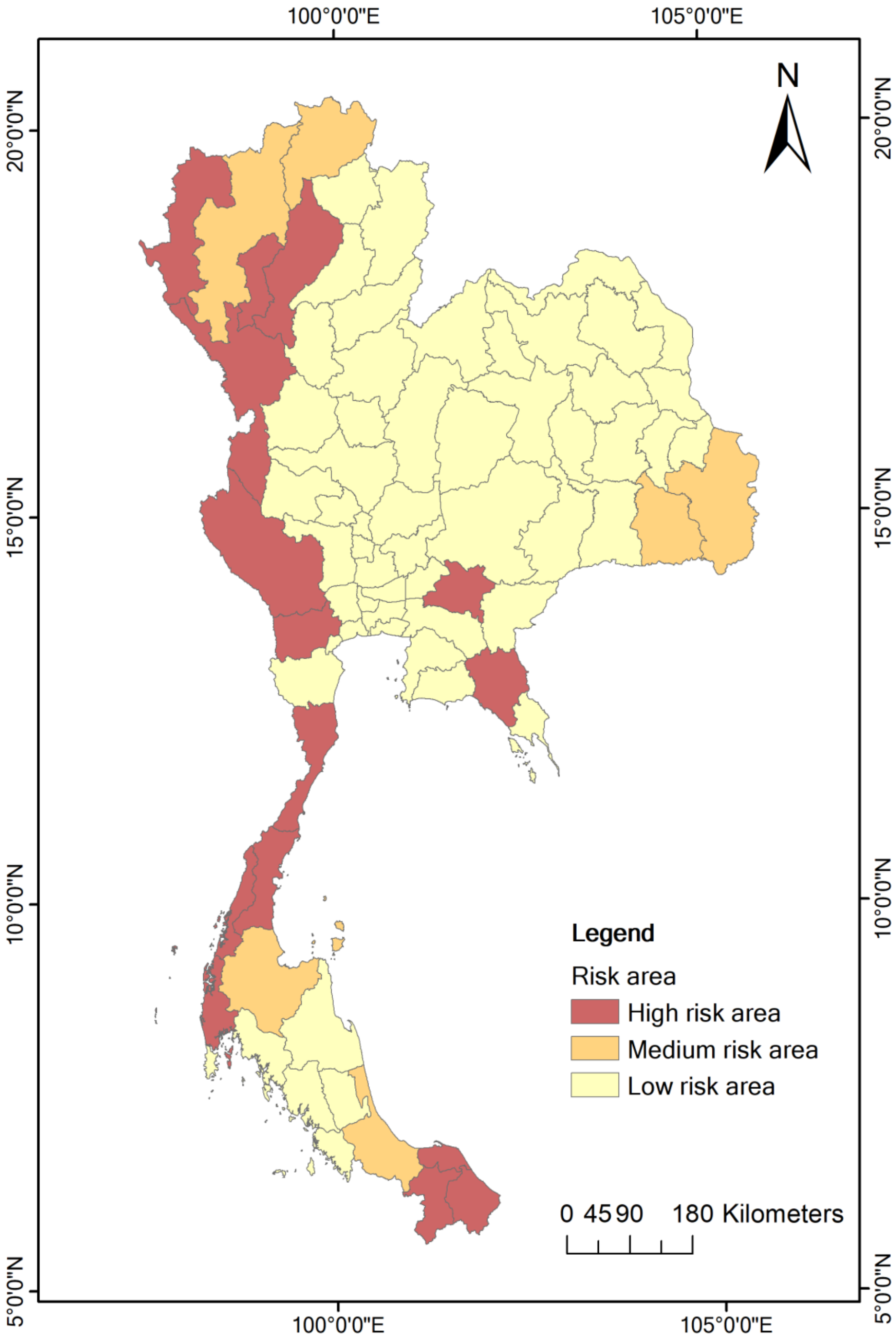


Figure 21: Map of malaria risk factors classified by province

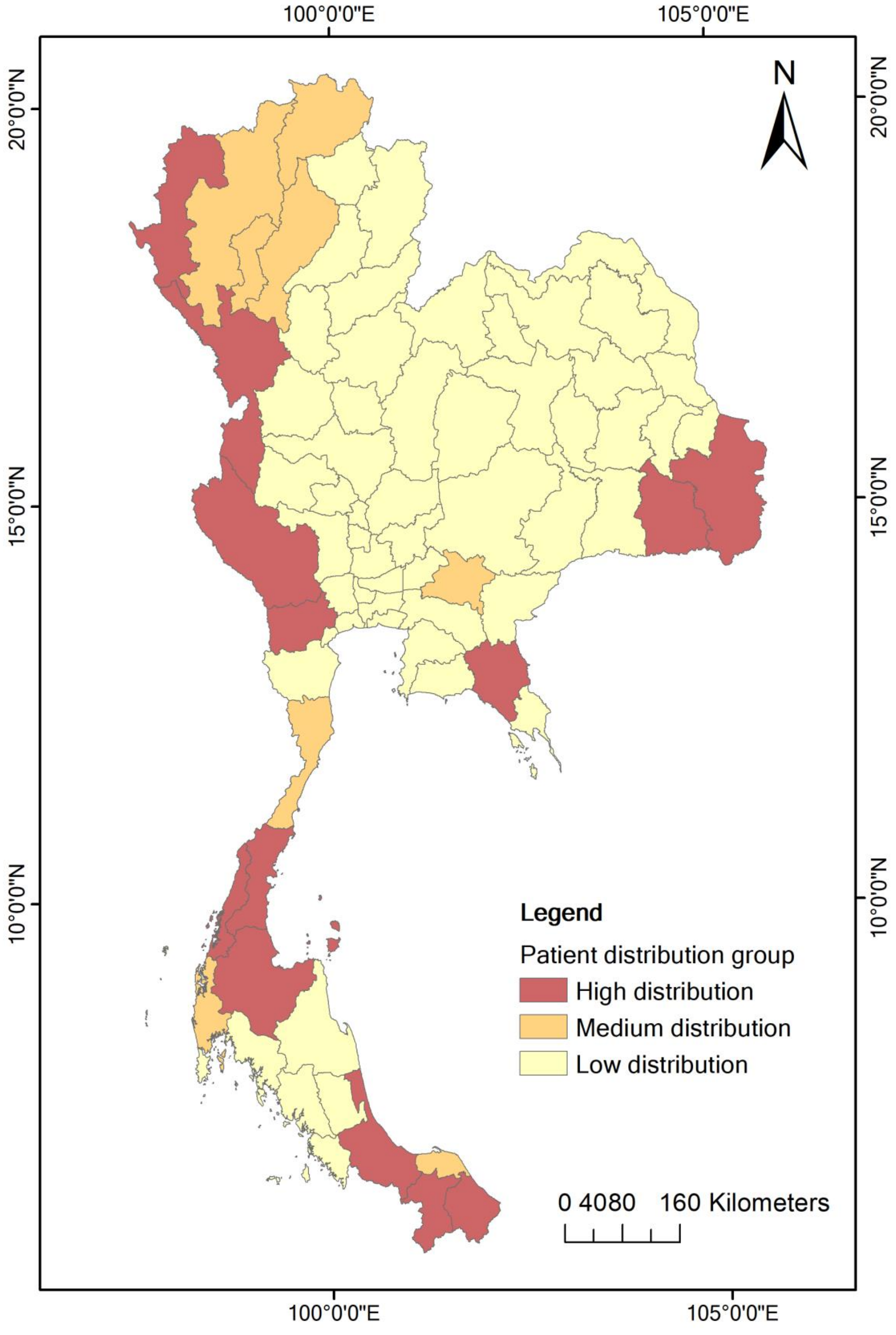


Figure 22: Map of patient distribution group classified by province

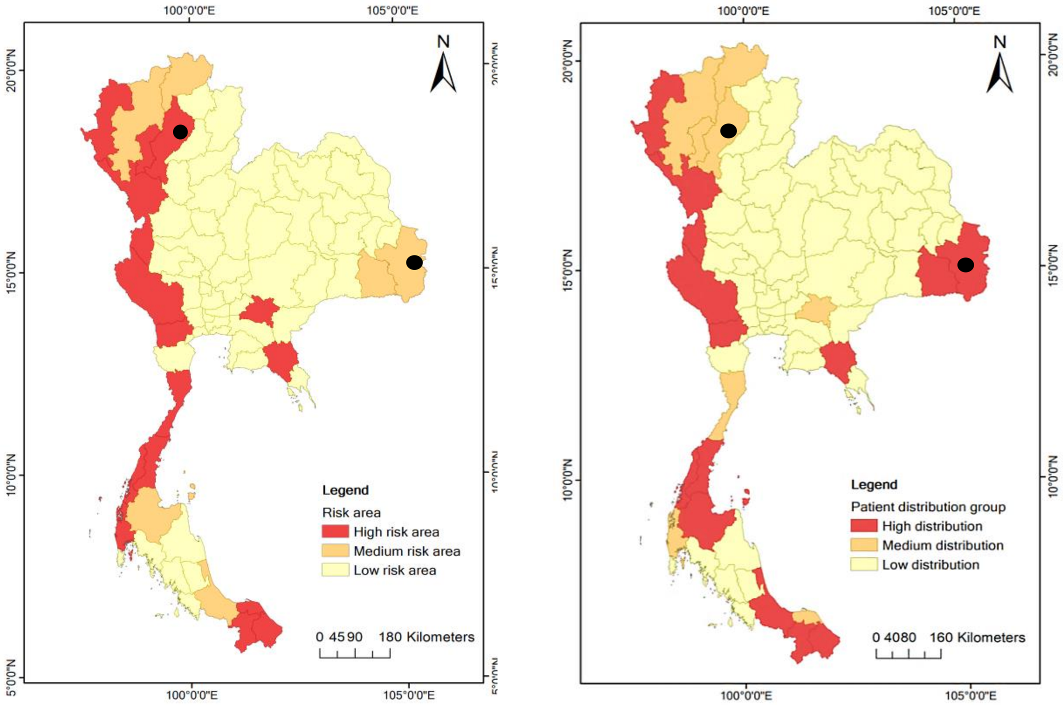


Figure 23: Comparison between the map of malaria risk factors and patient distribution group

Figure 23 demonstrations the malaria risk factors and patient distribution group maps whereas the similar results are displayed. Nevertheless, there are some slightly differences in high-risk and moderate-risk groups. For example, Ubon Ratchathani is a high spread of patients though malaria risk factors are found at a moderate level; this might be a result from the rotation of foreign workers where it is one of the causes of the high number of patients.

As a consequence of the comparison above, the result indicates that the obtained risk model with 28 risk-factors is highly accurate when comparing with the true values of malaria distribution. Besides, the risk map depending on each risk-factor is drawn to display the distribution of each factor. So, there are 28 maps to be drawn, as shown in Figure 24 to Figure 51

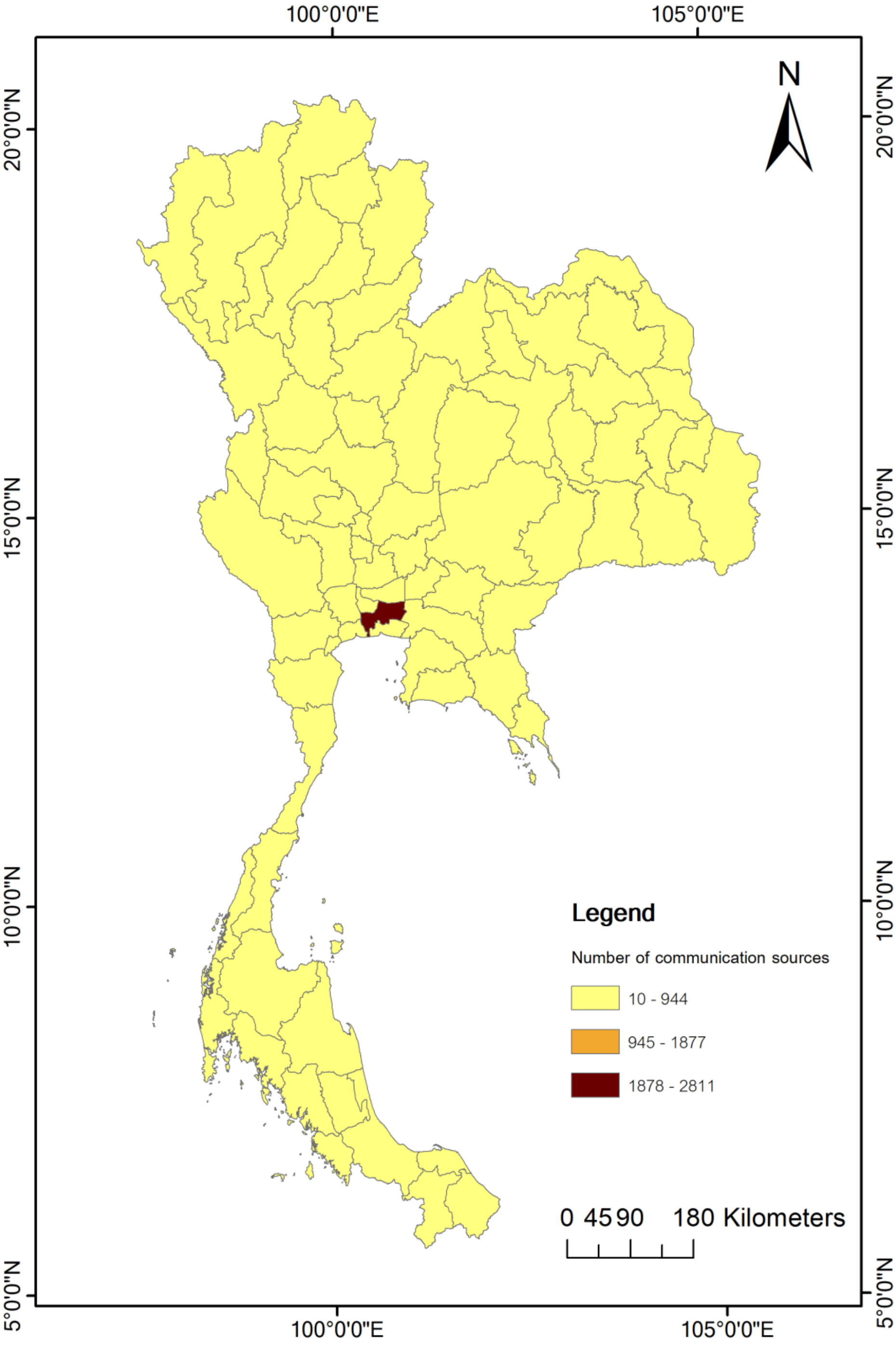


Figure 24: Map displays the number of communication sources

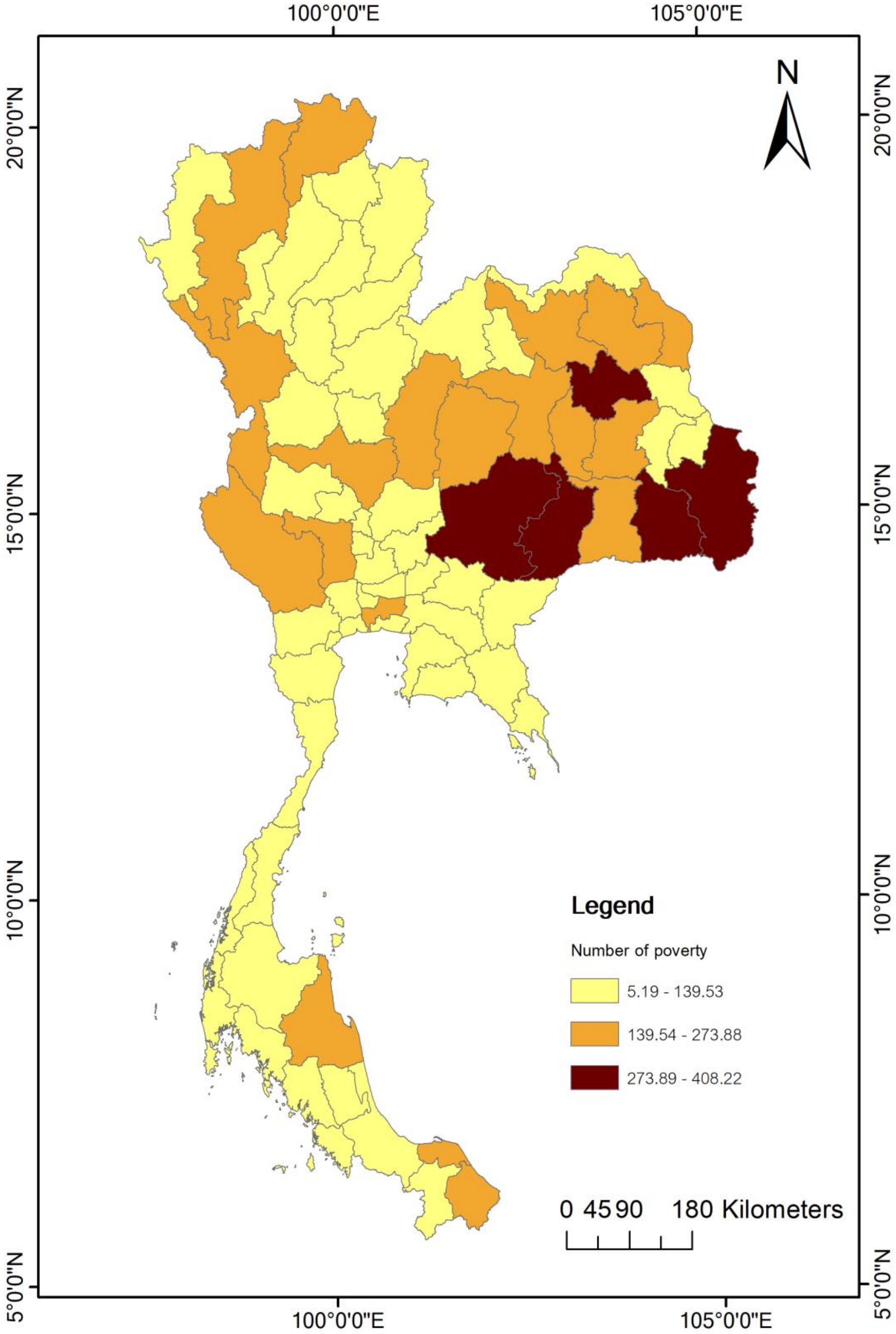


Figure 25: Map displays the number of poverty

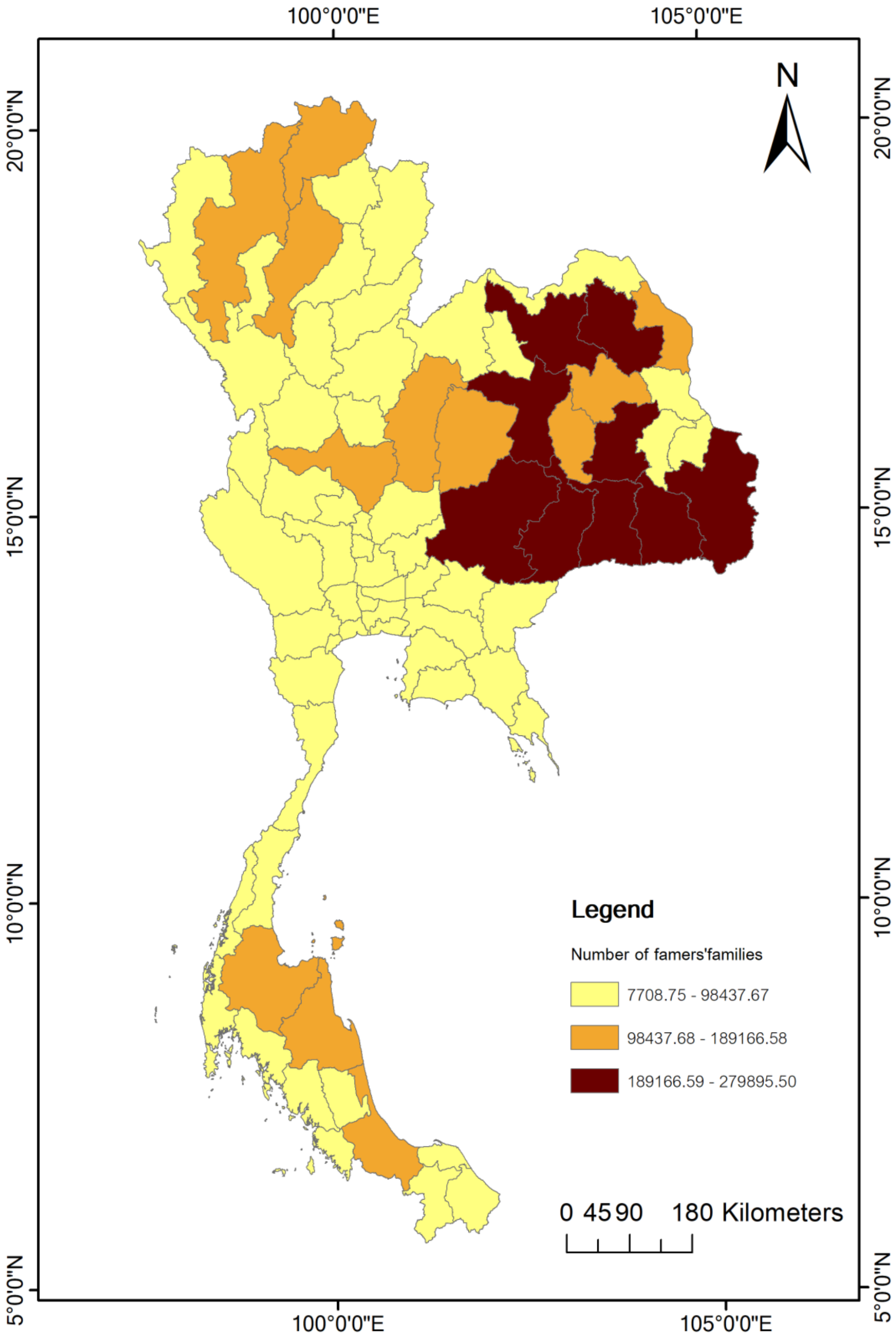


Figure 26: Map displays the number of famers'families

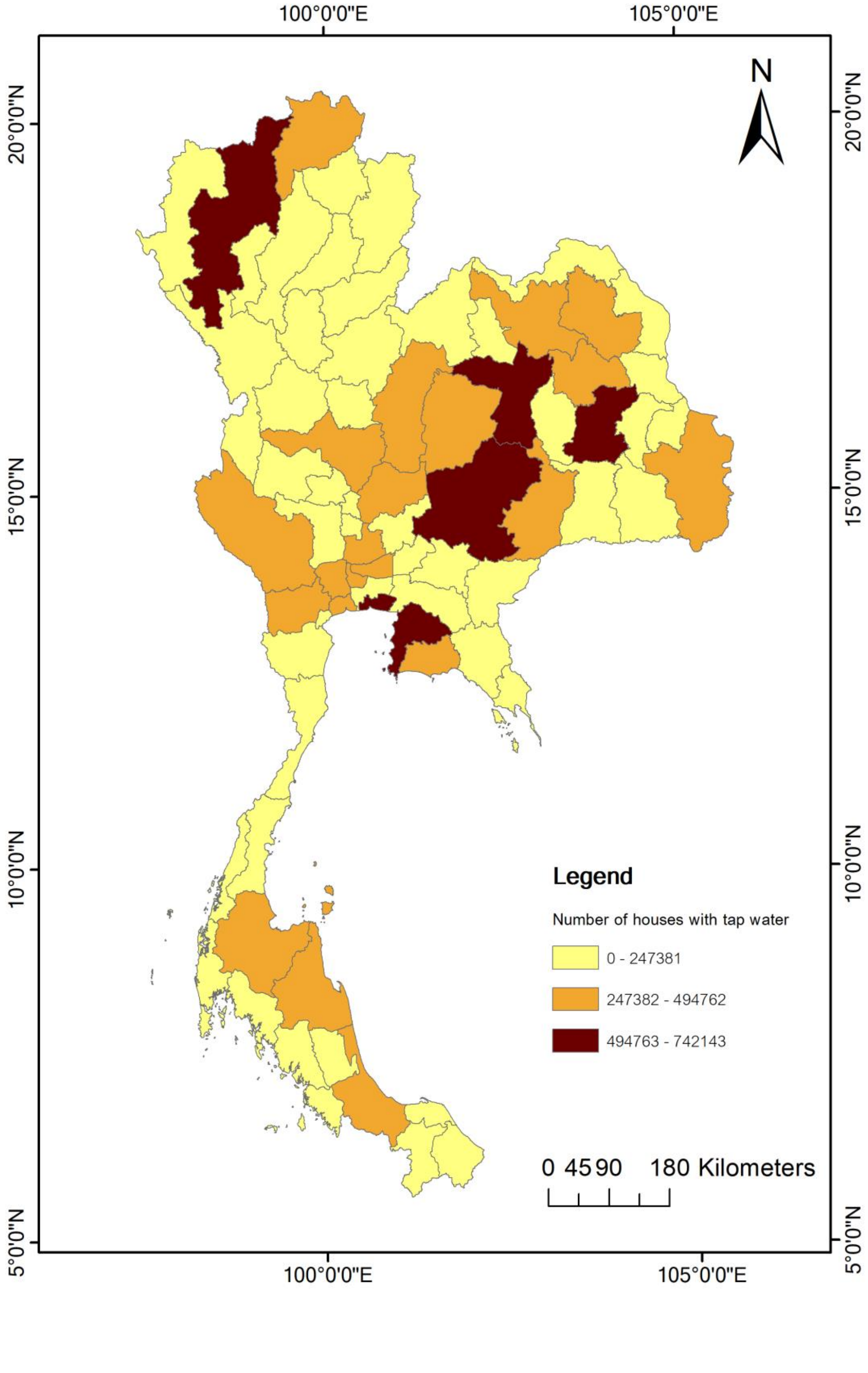


Figure 27: Map displays the number of houses with tap water

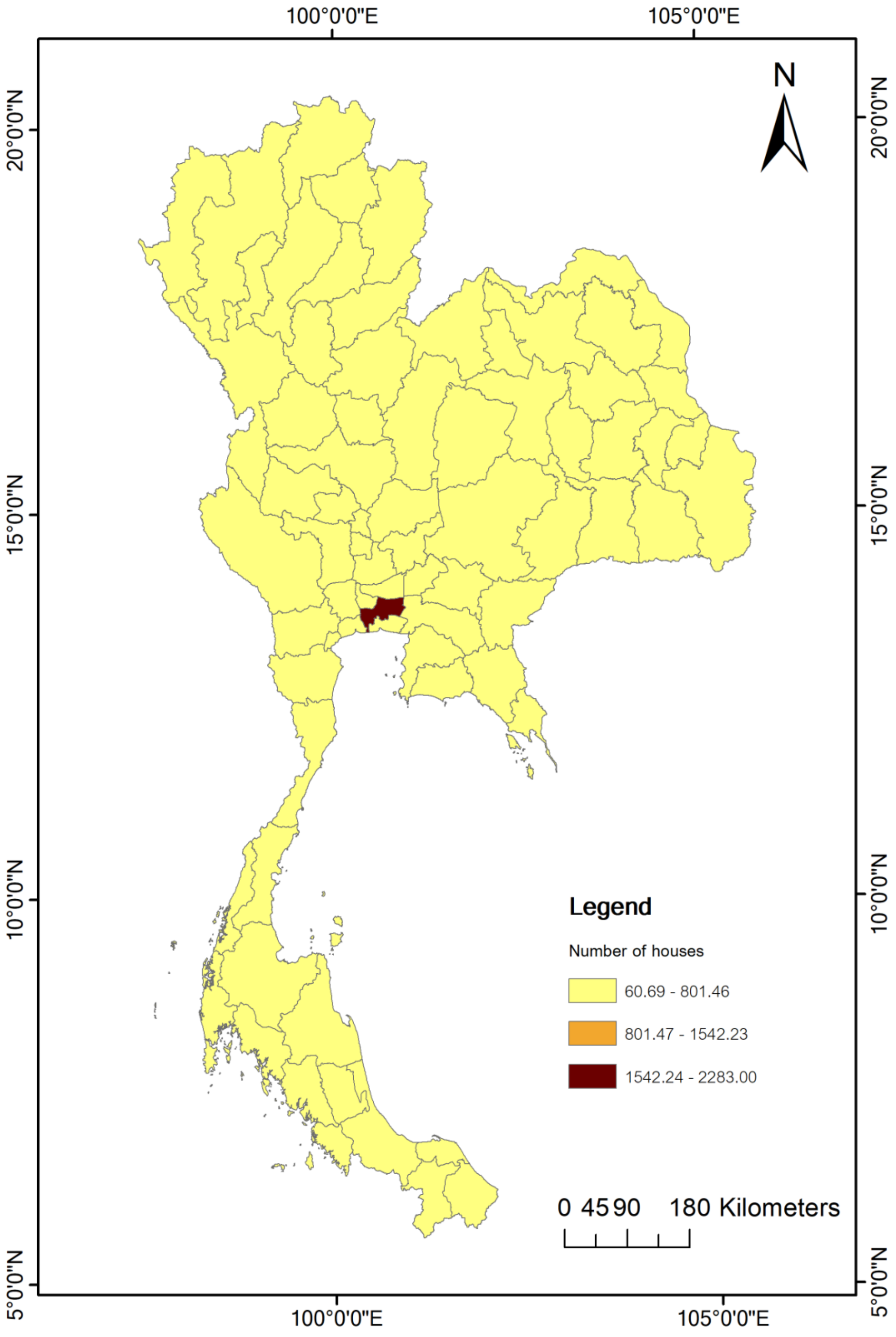


Figure 28: Map displays the number of houses

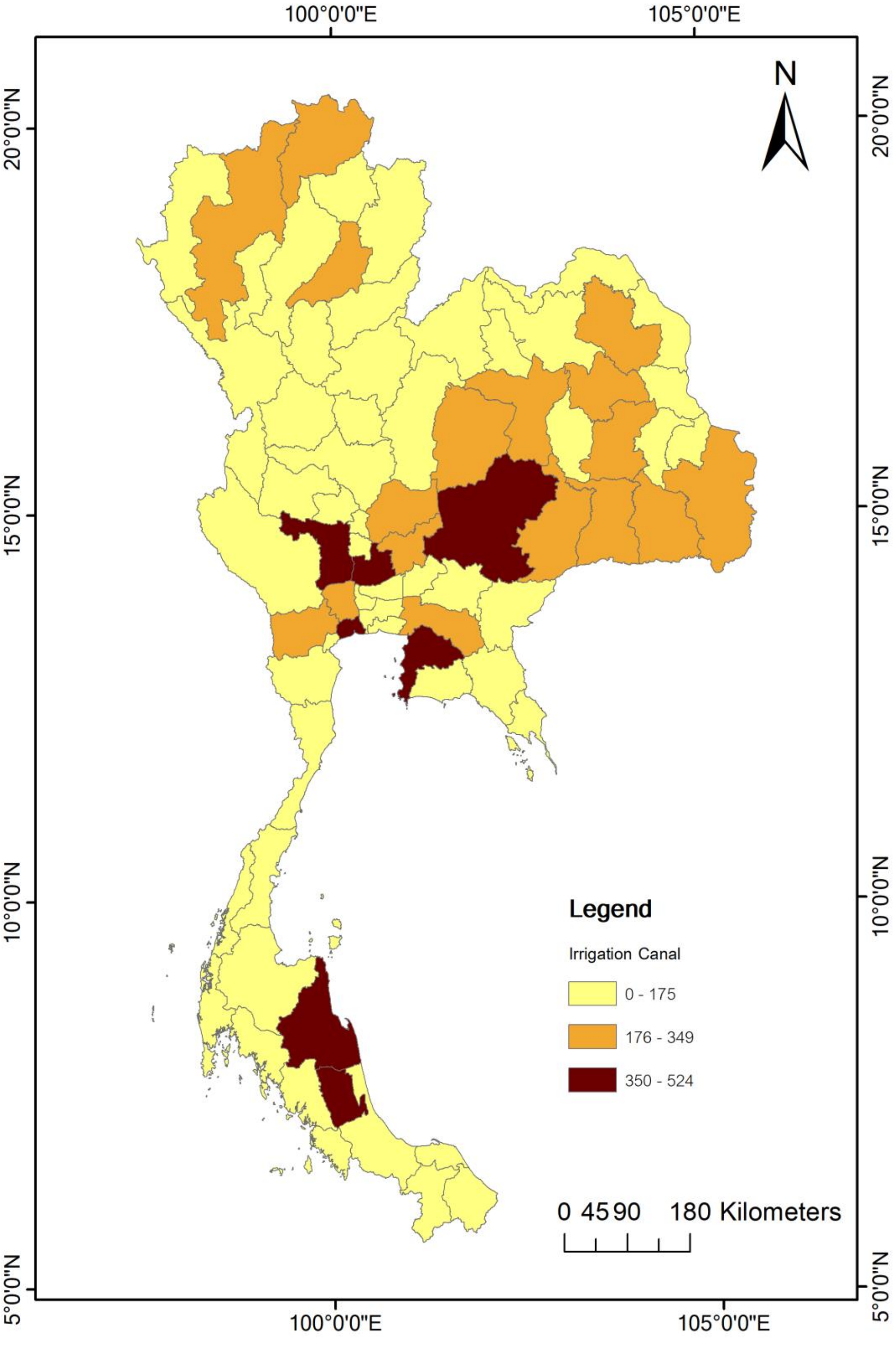


Figure 29: Map displays the Irrigation Canal

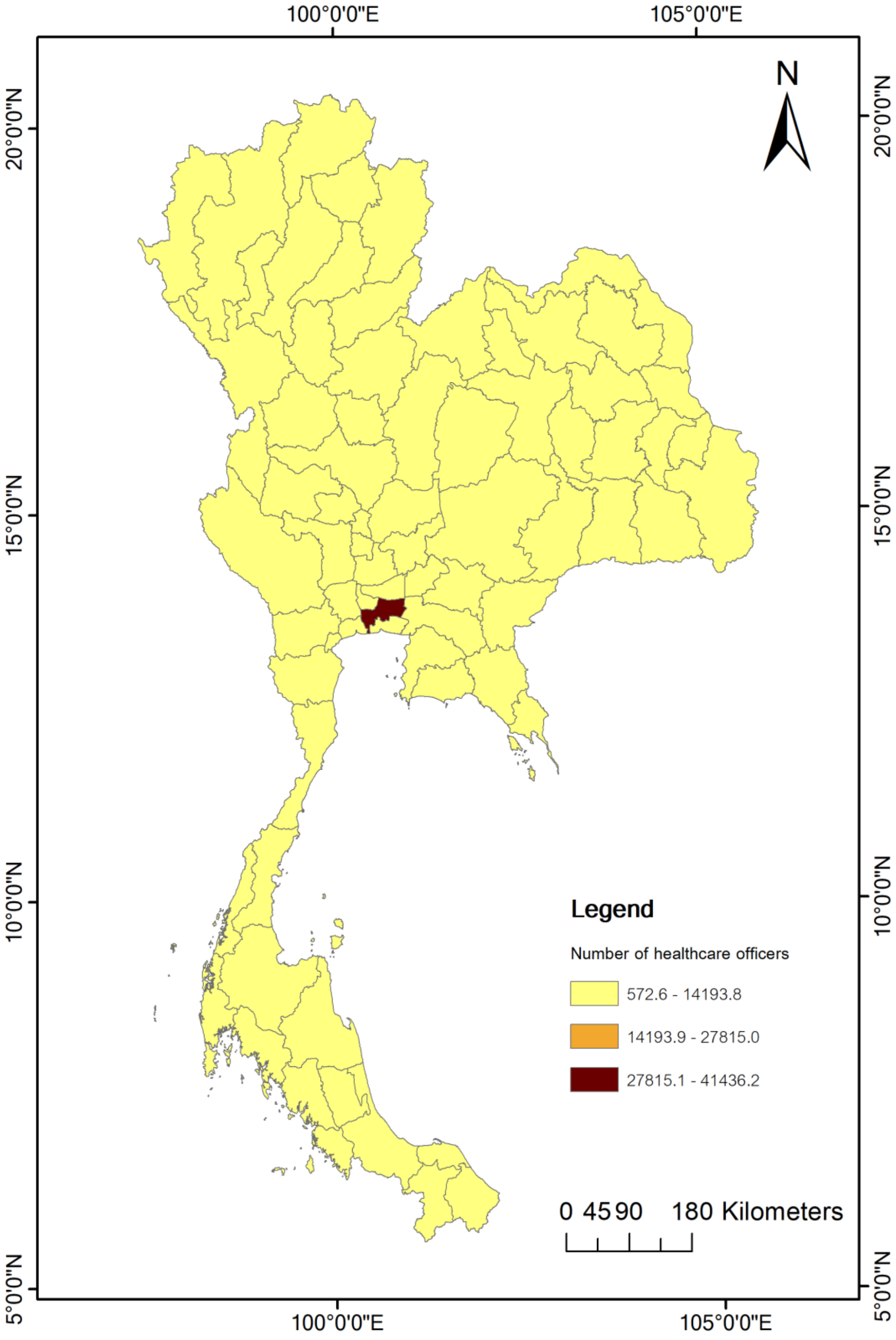


Figure 30: Map displays the number of healthcare officers

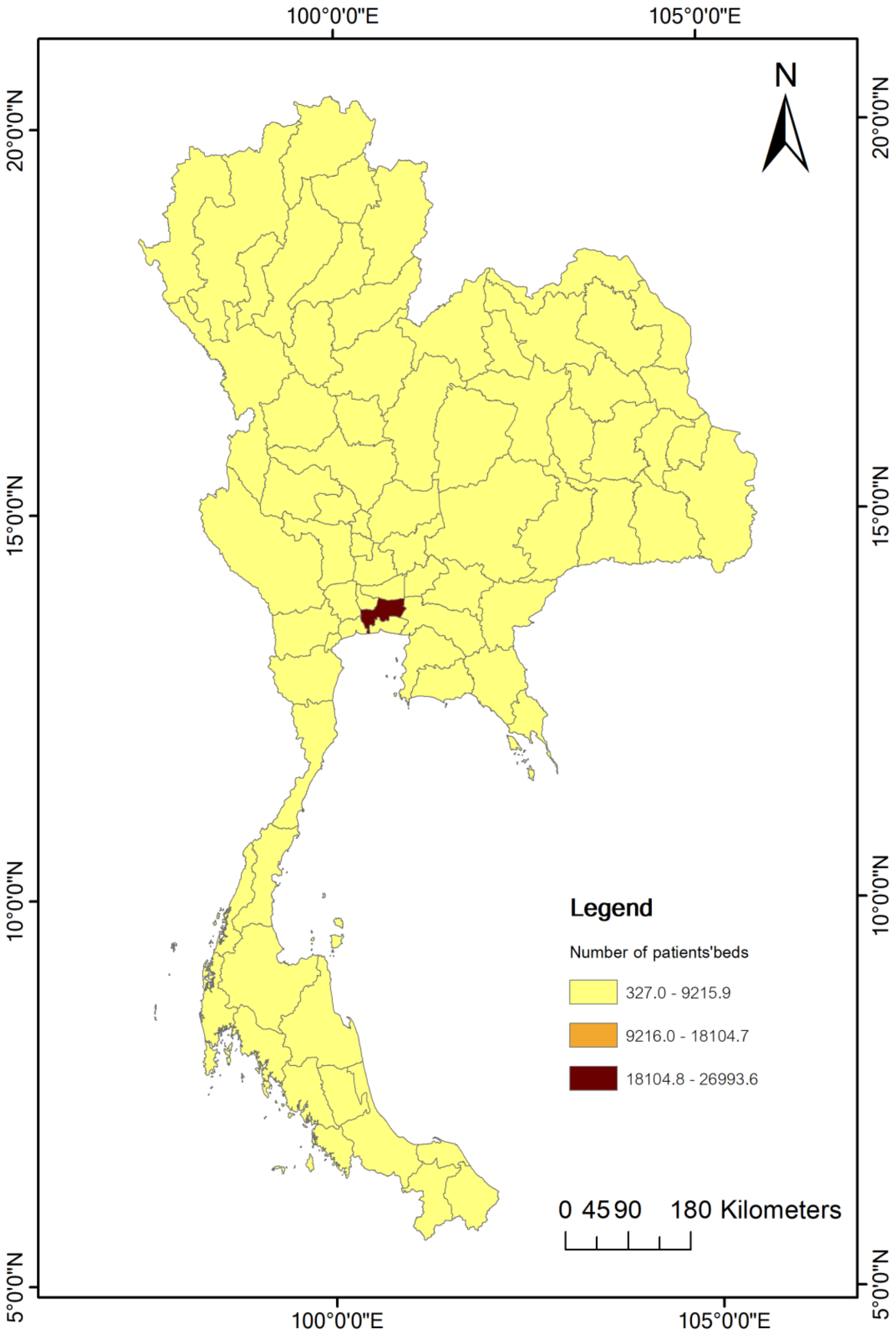


Figure 31: Map displays the number of patients 'beds

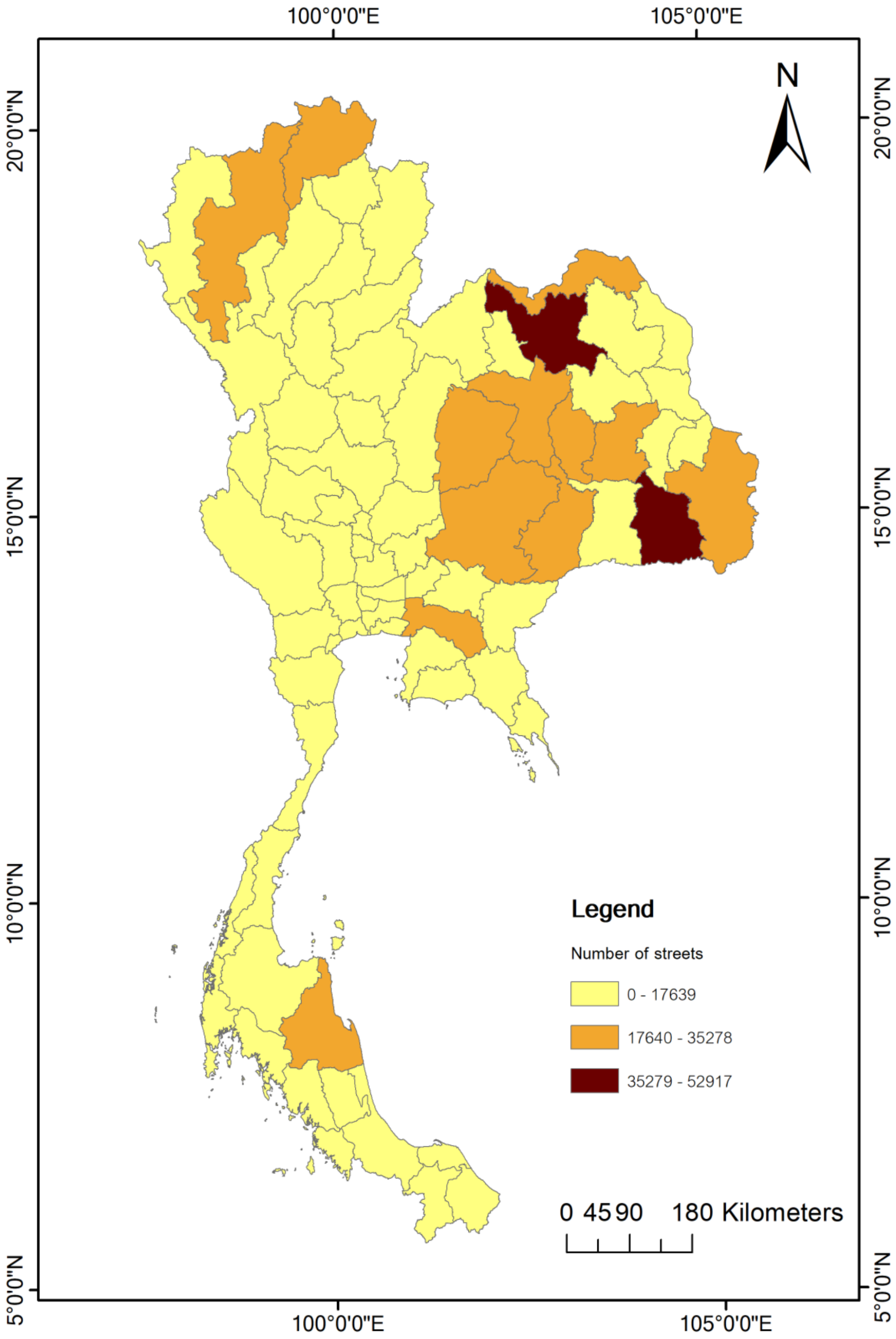


Figure 32: Map displays the number of streets

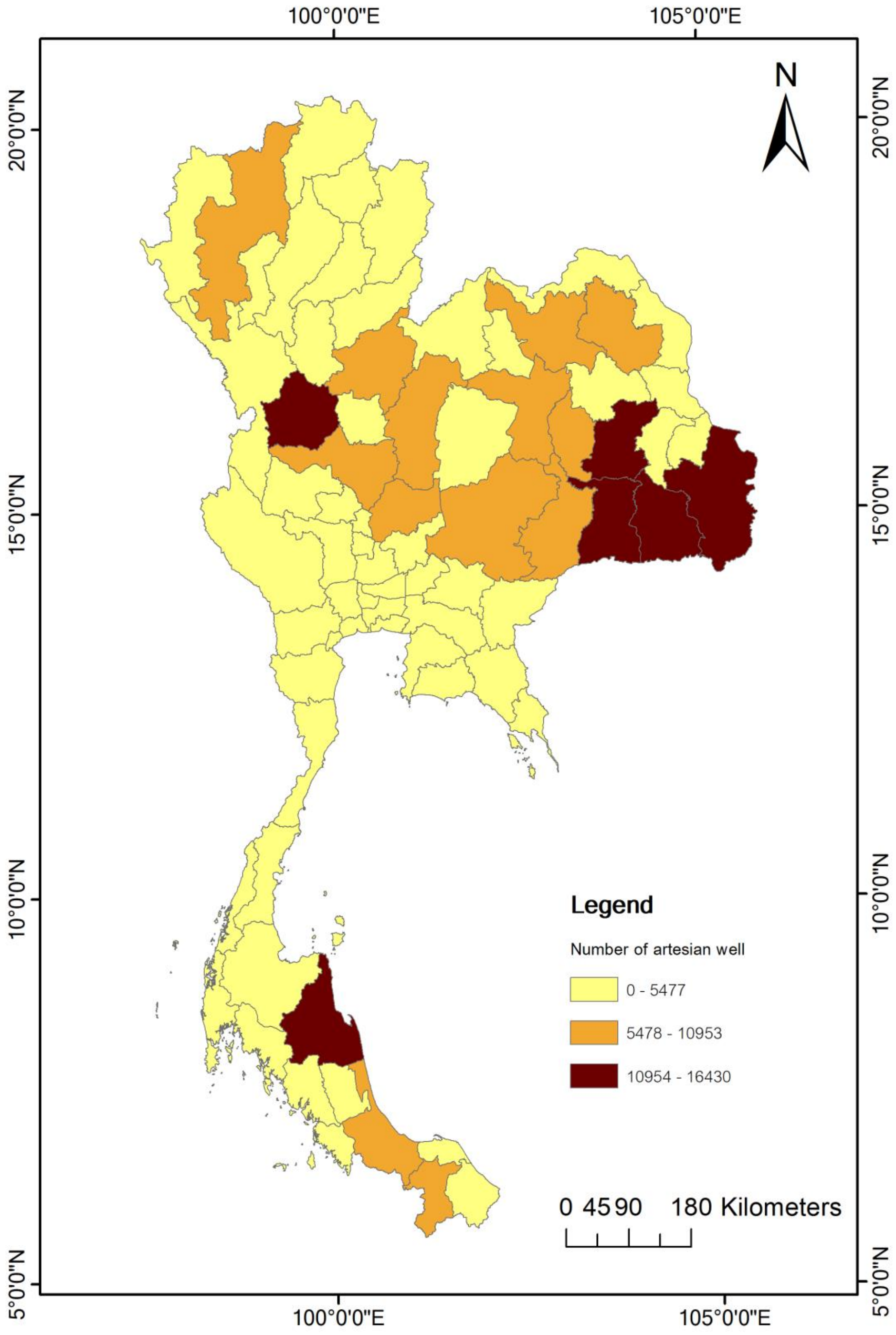


Figure 33: Map displays the number of artesian well

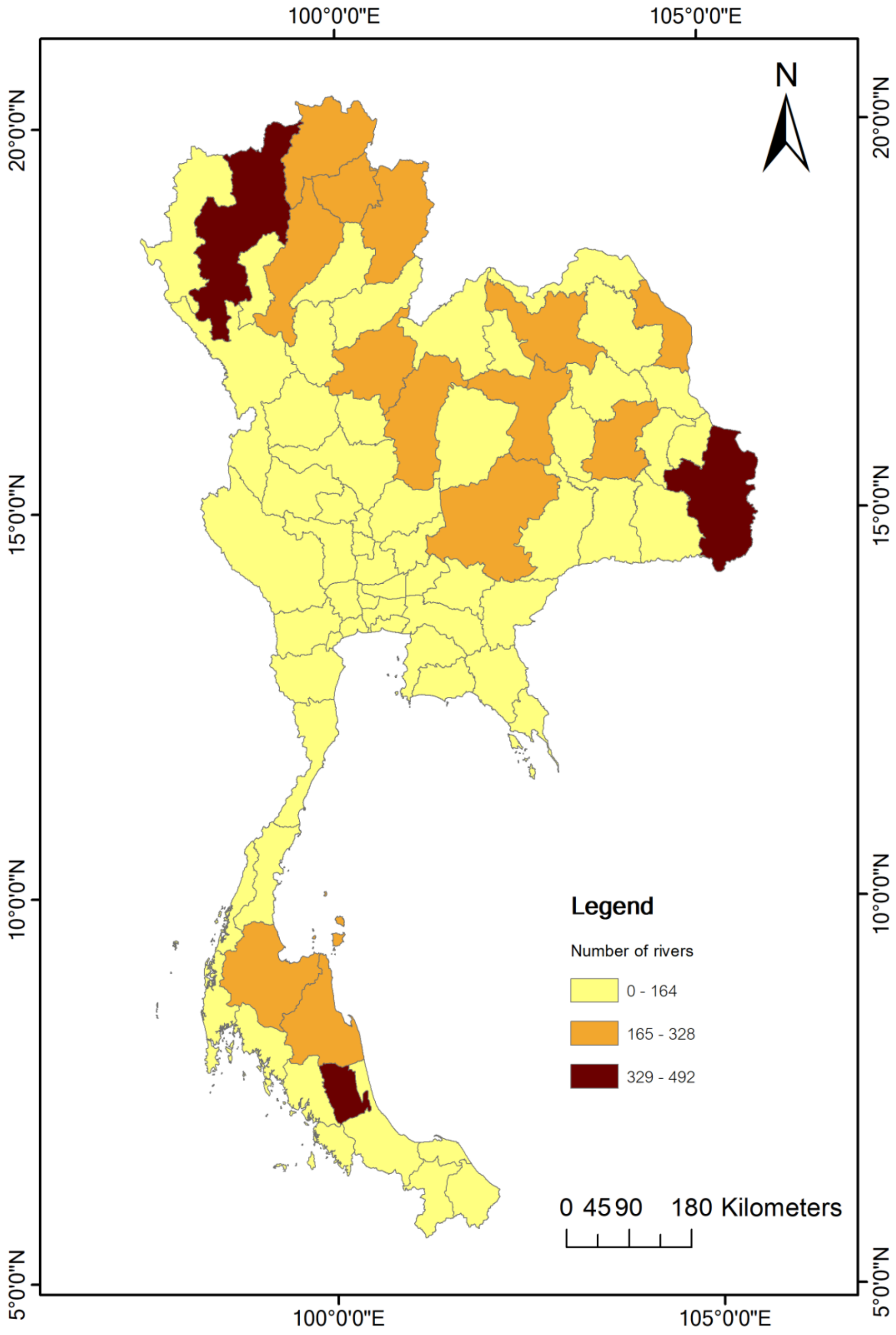


Figure 34: Map displays the number of rivers

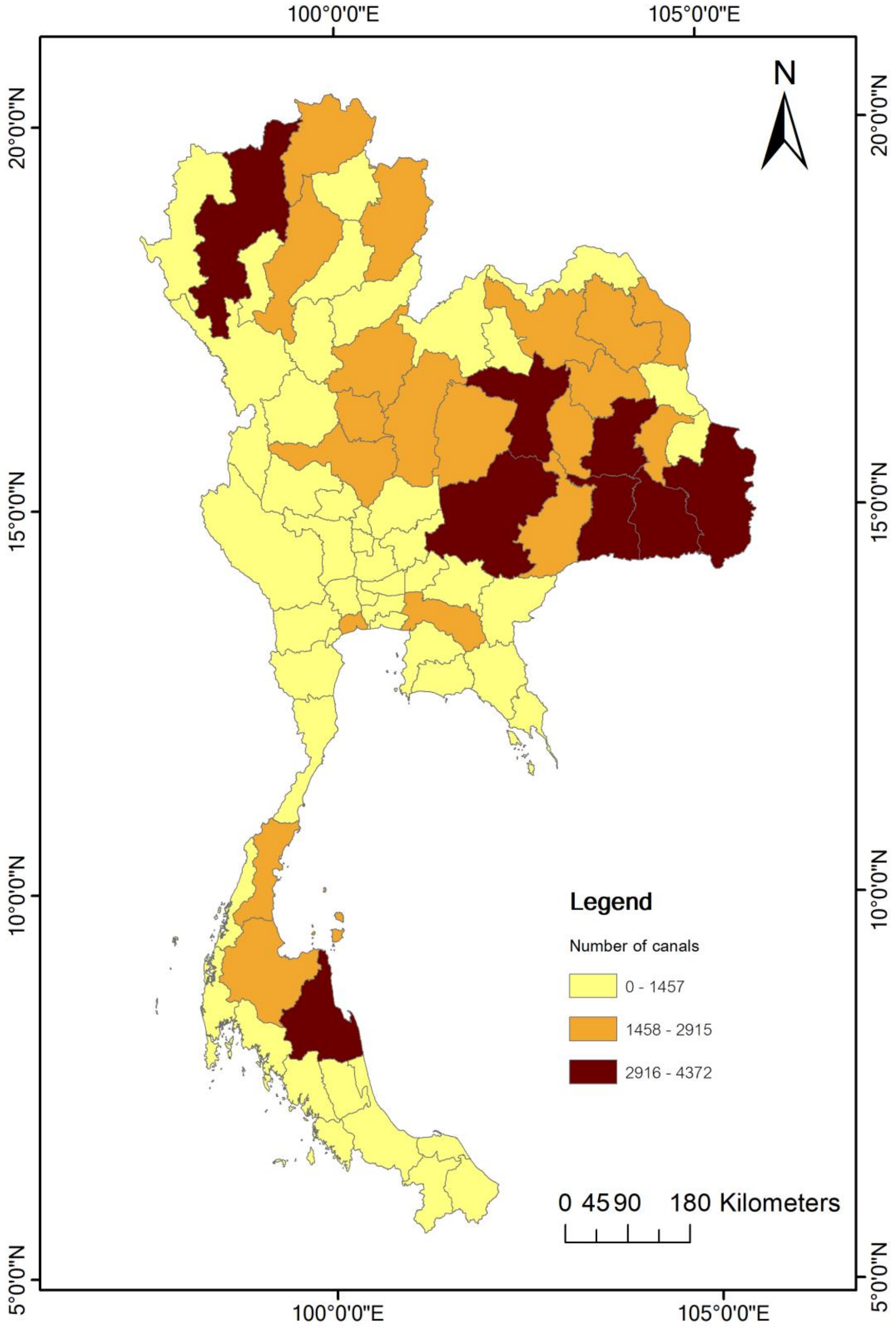


Figure 35: Map displays the number of canals

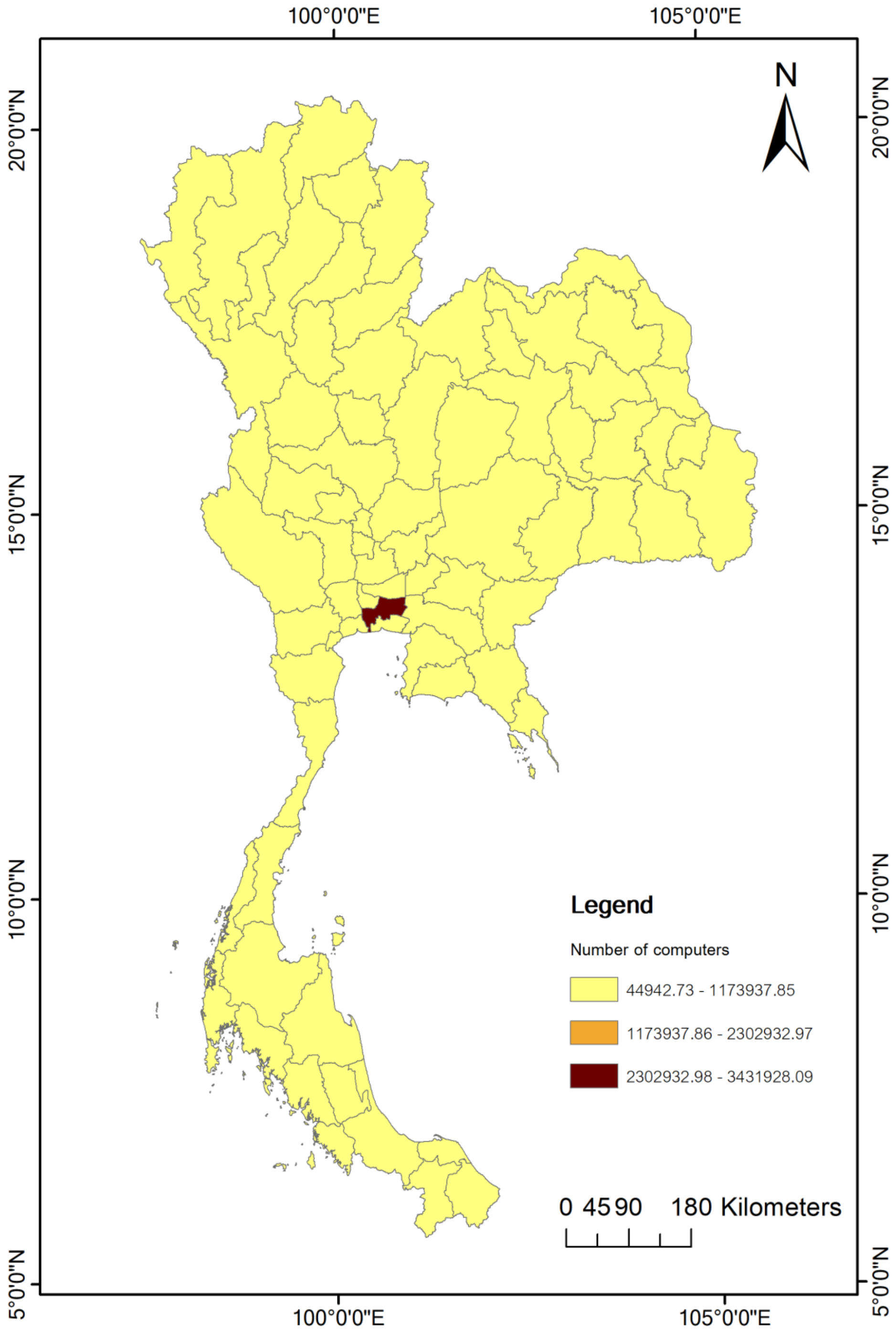


Figure 36: Map displays the number of computers

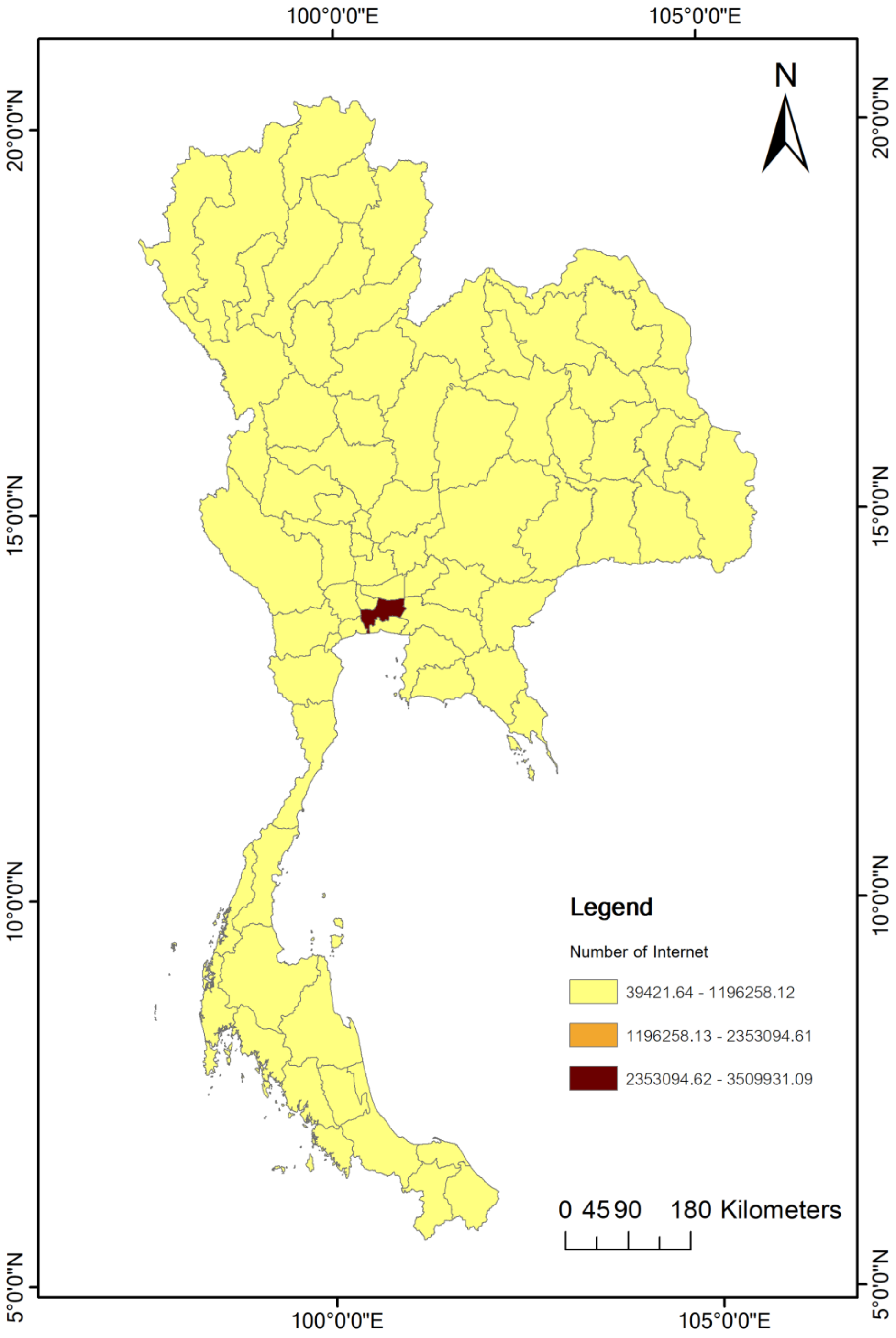


Figure 37: Map displays the number of Internet

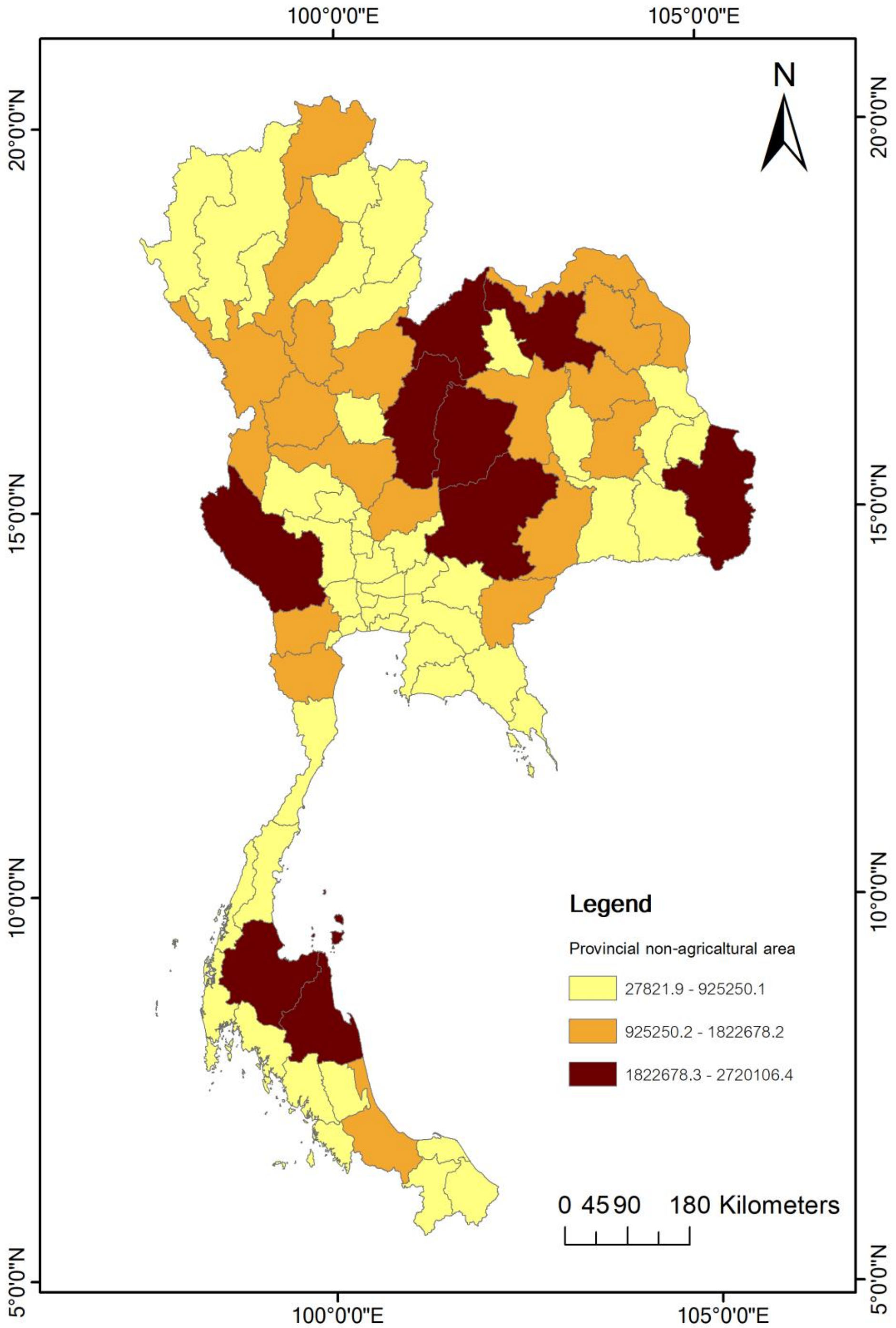


Figure 38: Map displays the Provincial non-agricultural area

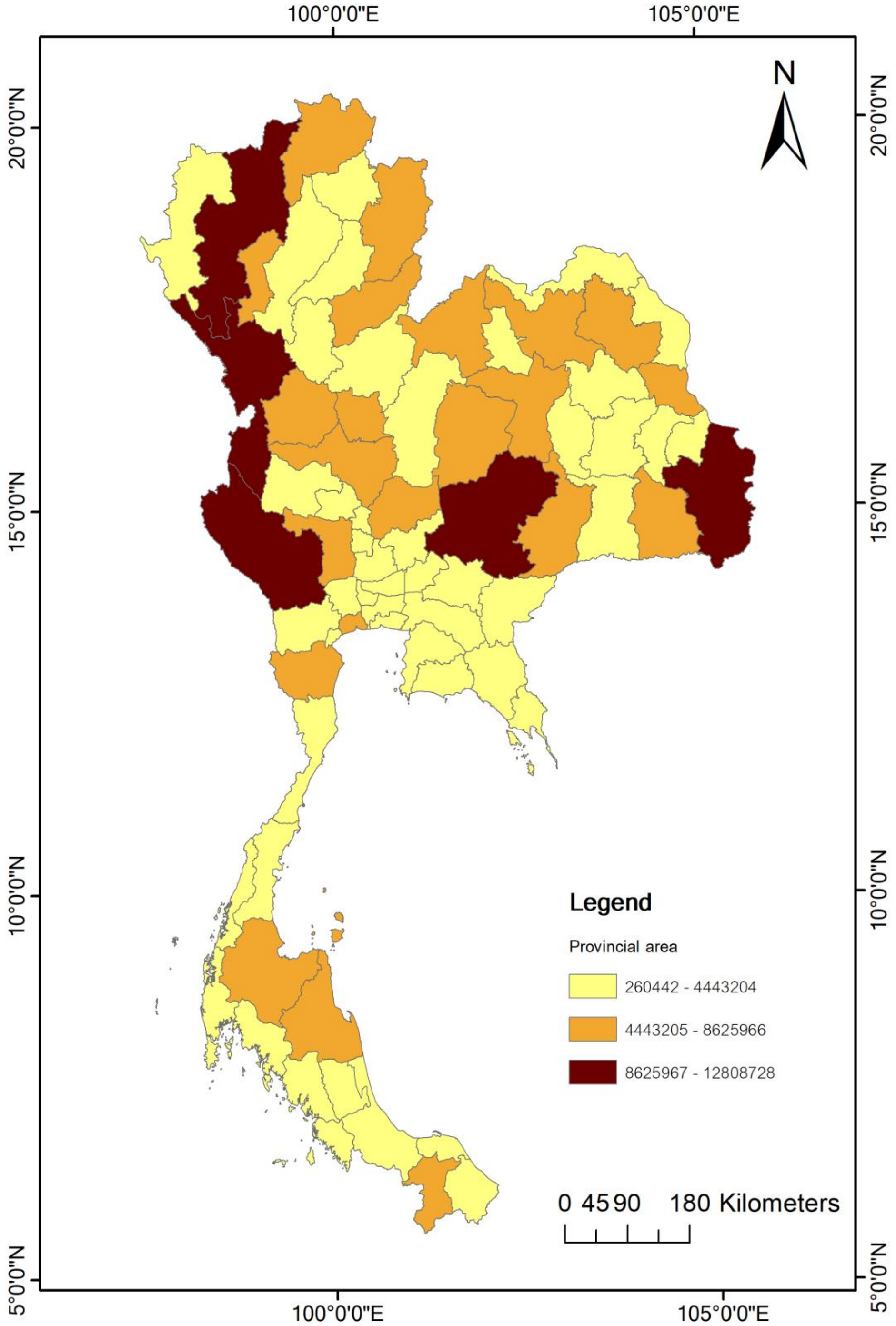


Figure 39: Map displays the Provincial area

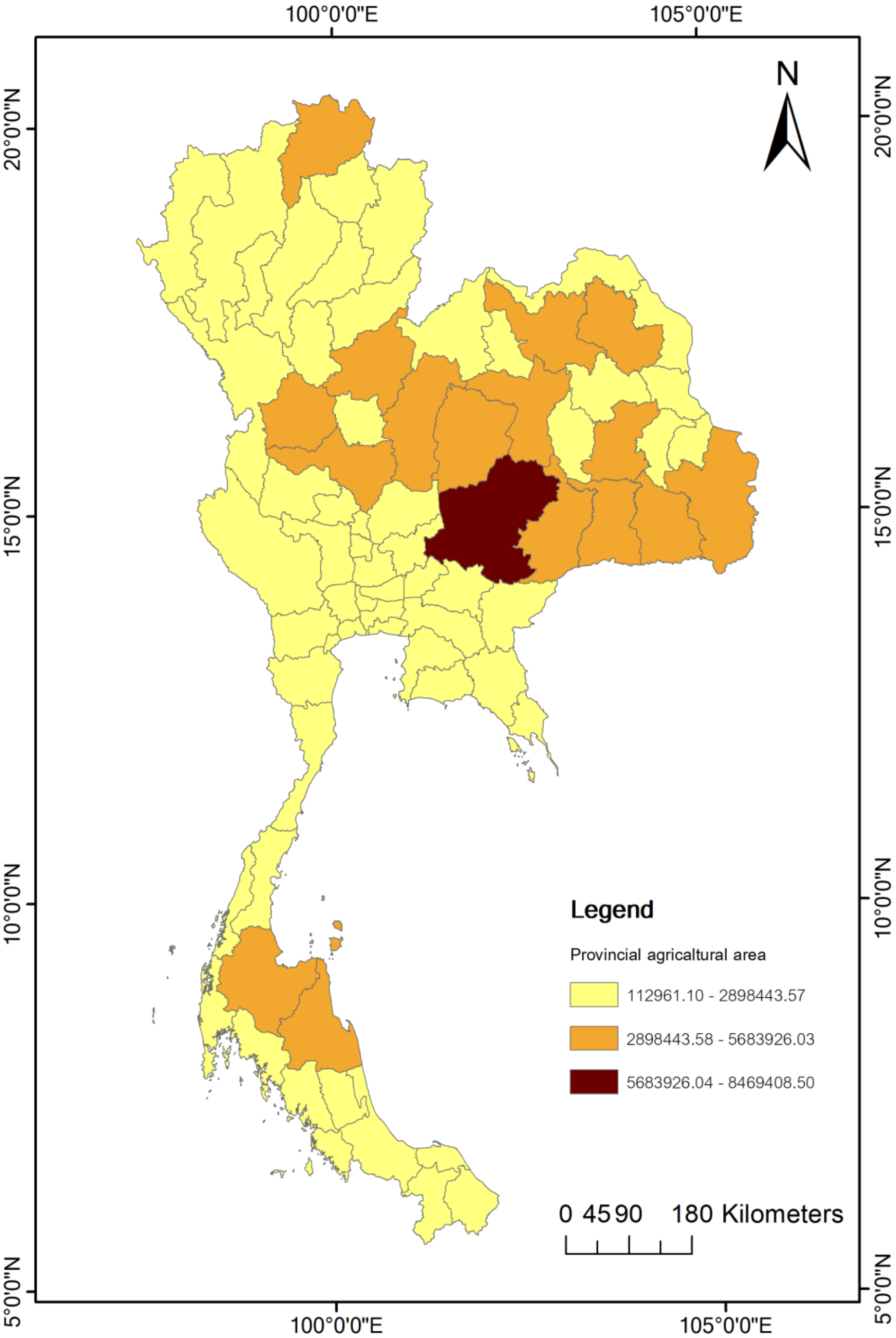


Figure 40: Map displays the Provincial agricultural area

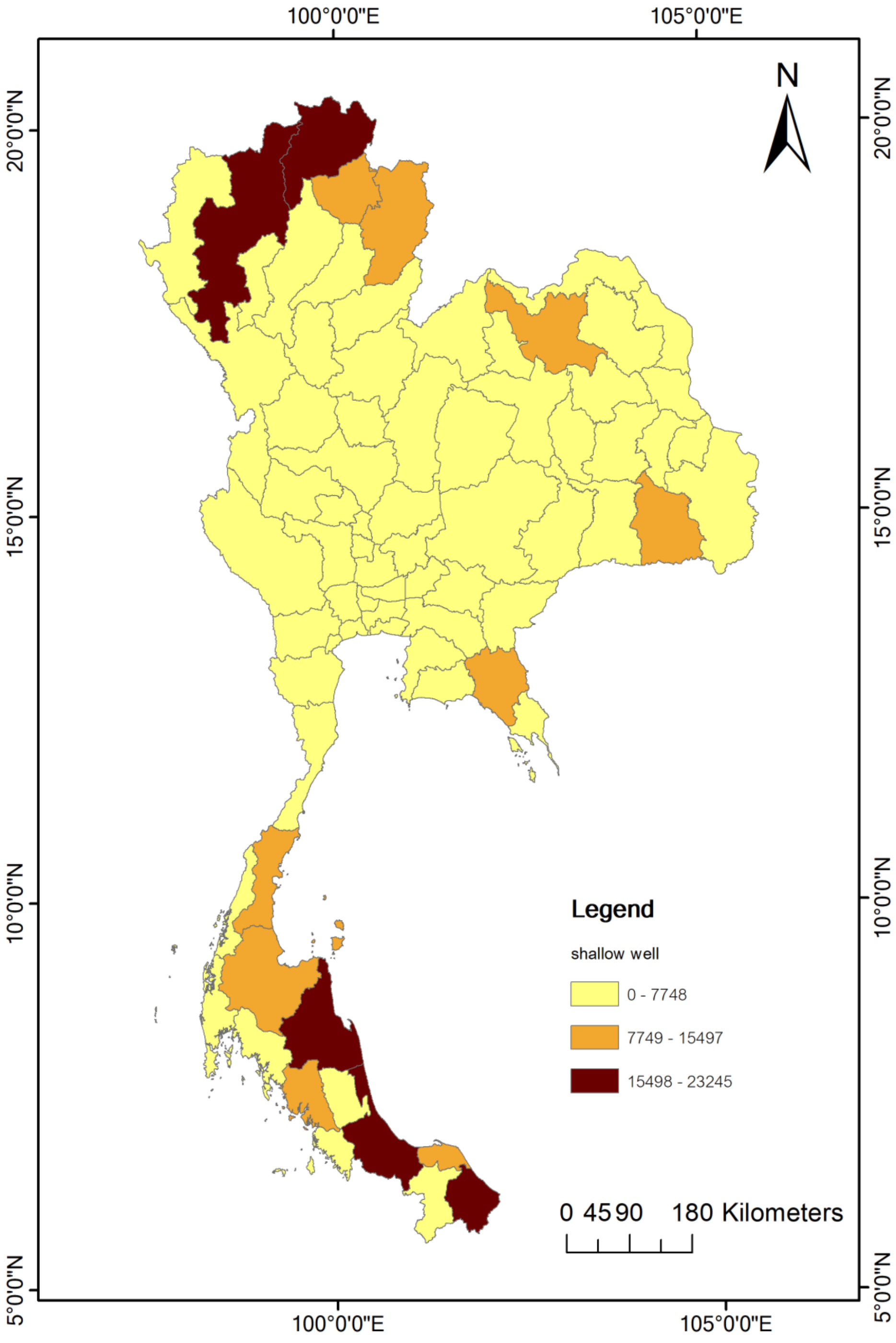


Figure 41: Map displays the shallow well

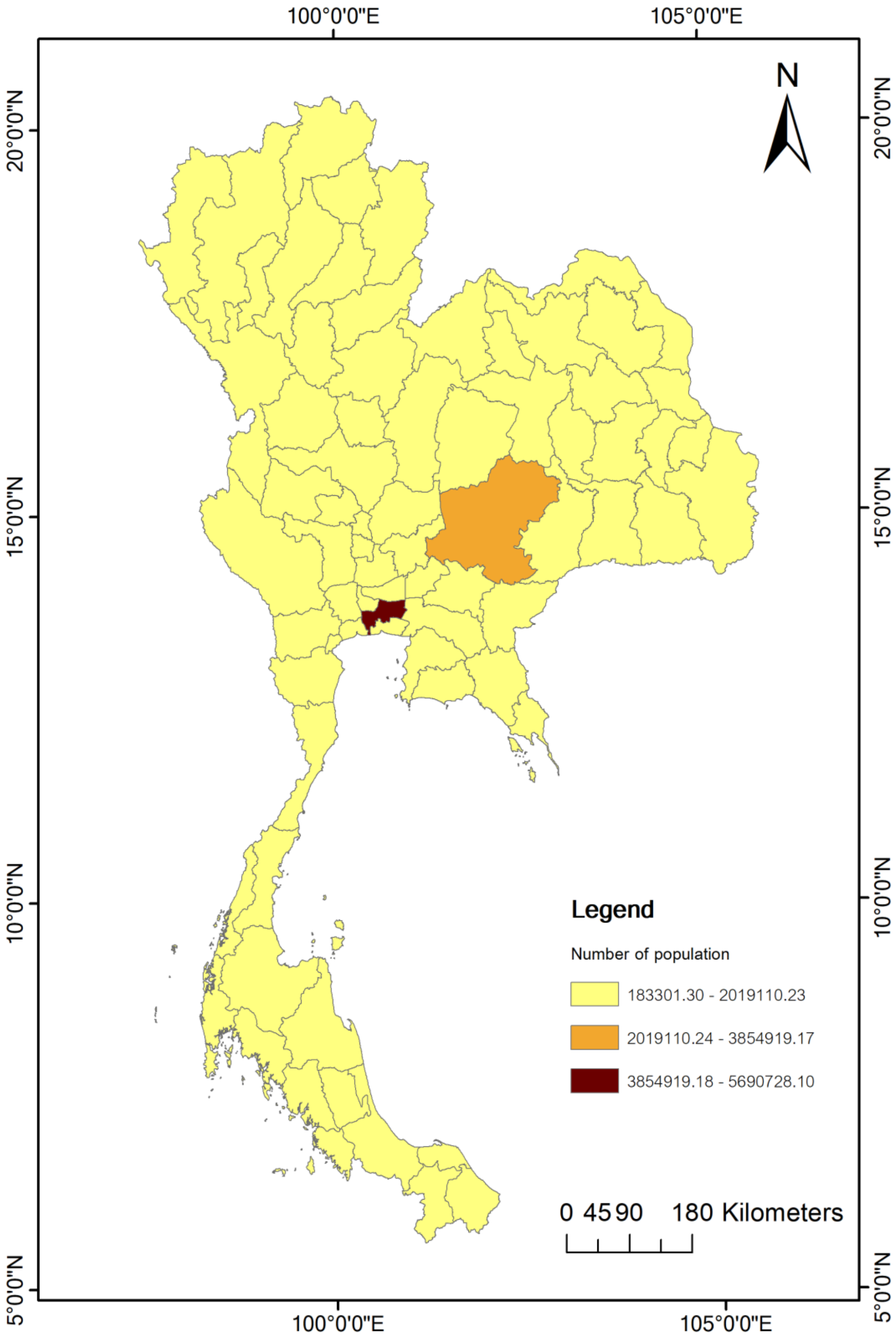


Figure 42: Map displays the number of population

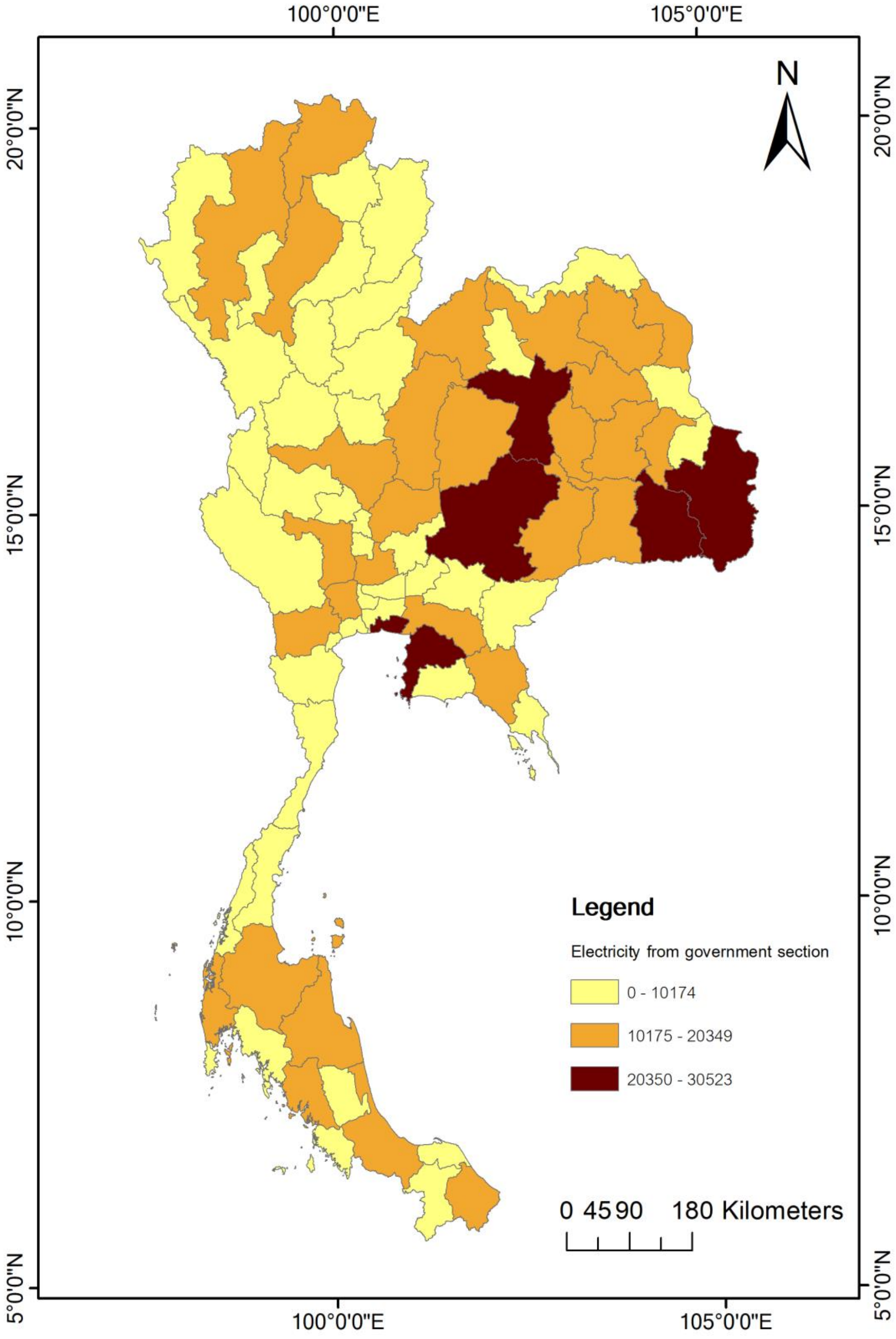


Figure 43: Map displays the electricity from government section

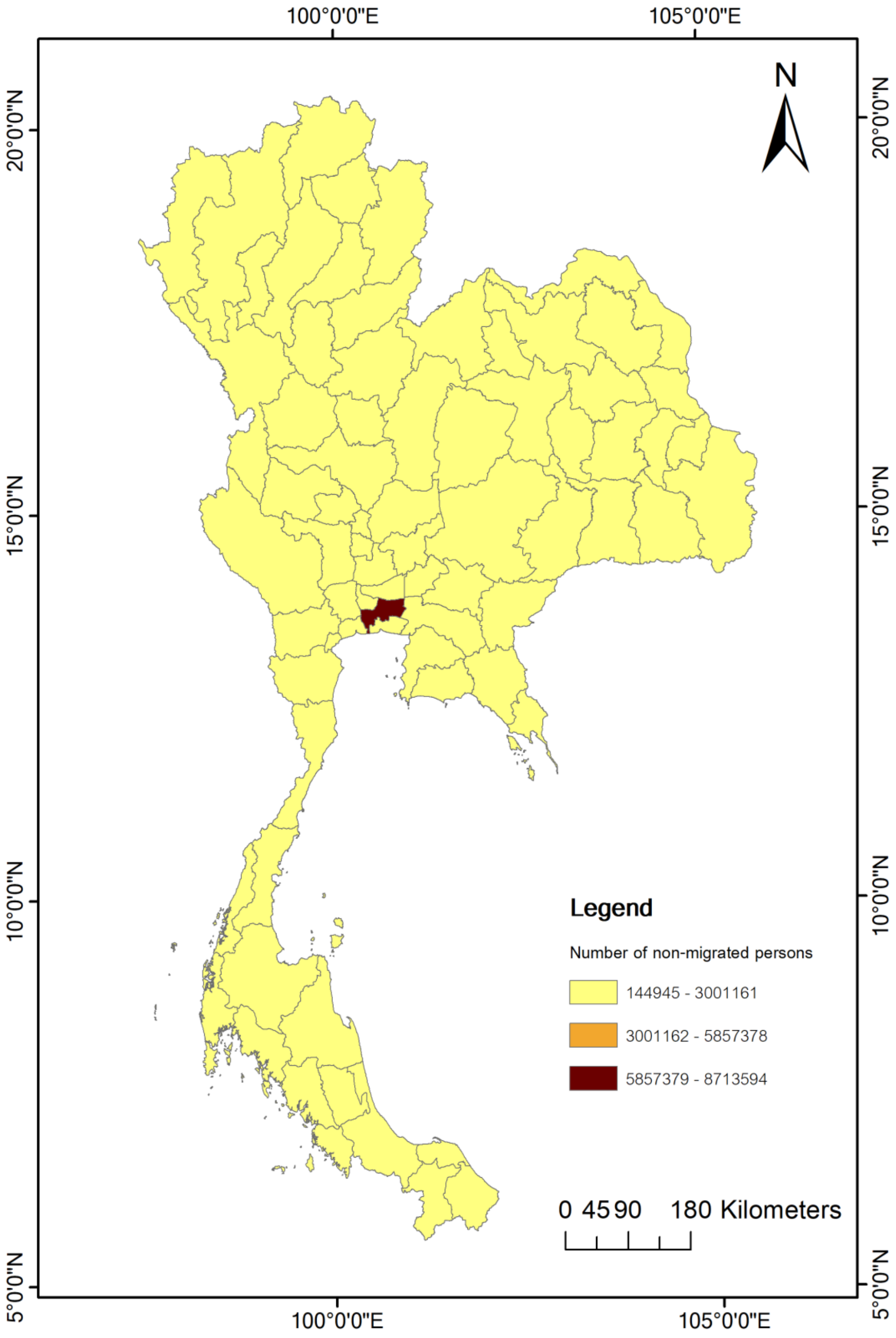


Figure 44: Map displays the number of non-migrated persons

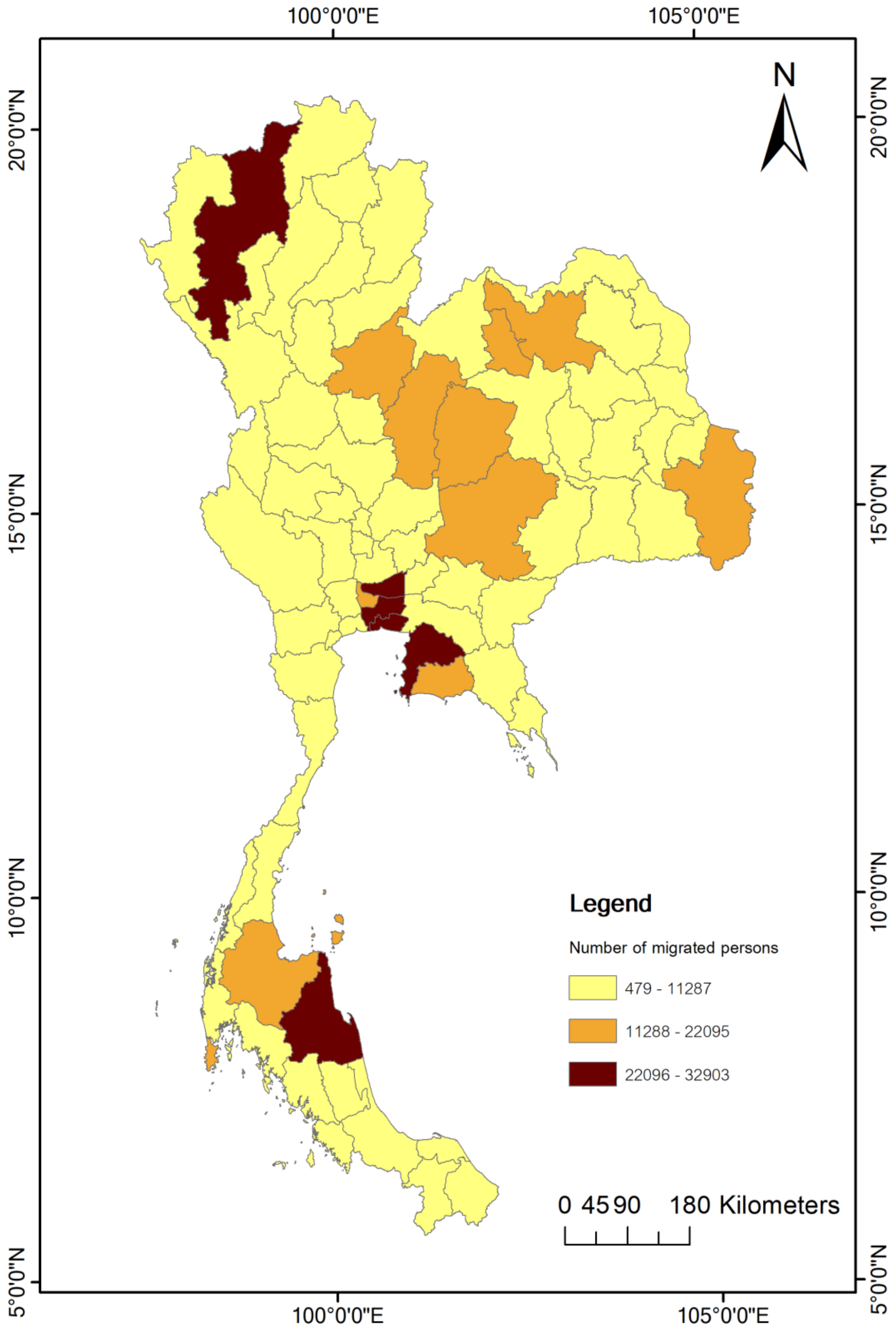


Figure 45: Map displays the number of migrated persons

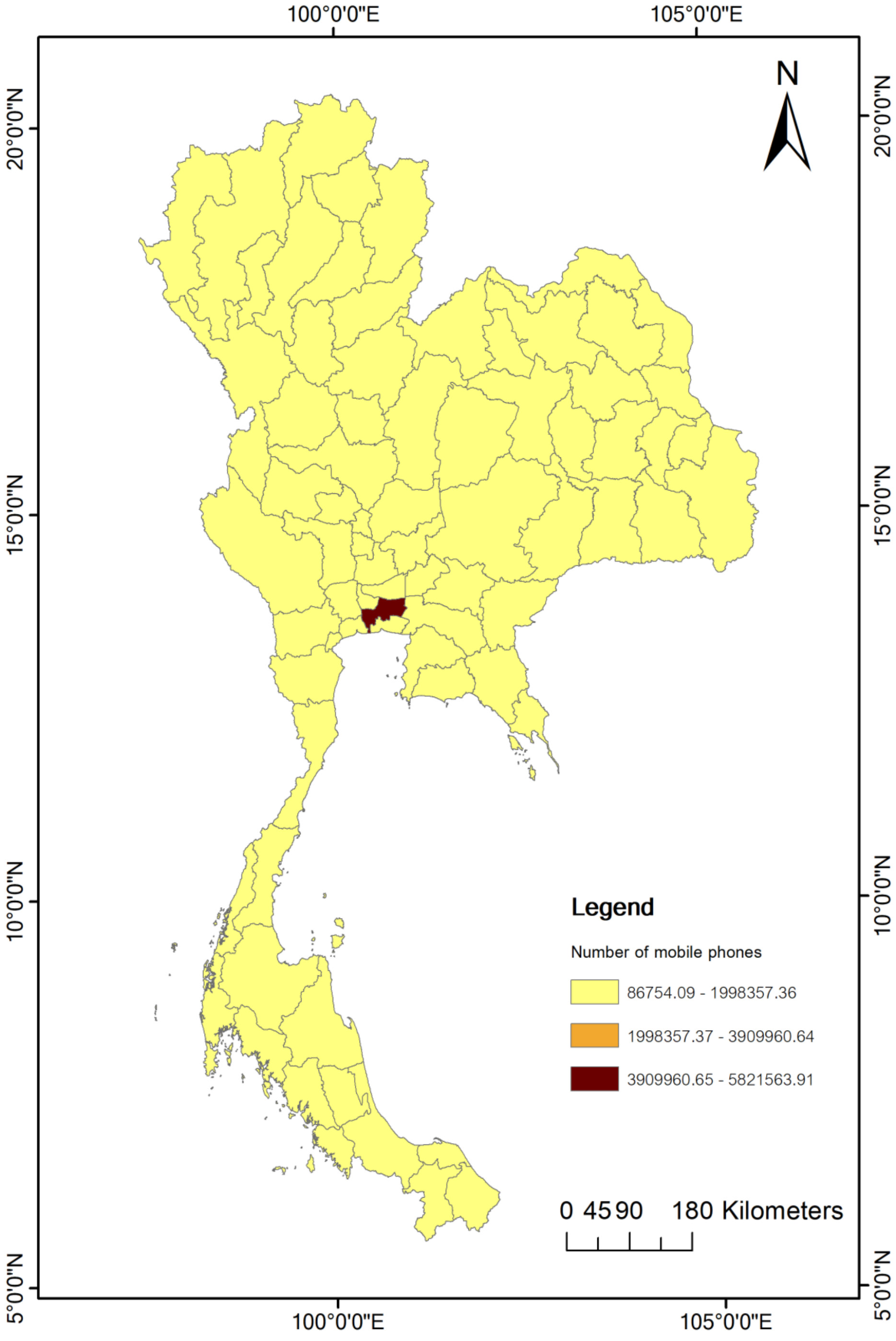


Figure 46: Map displays the number of mobile phones

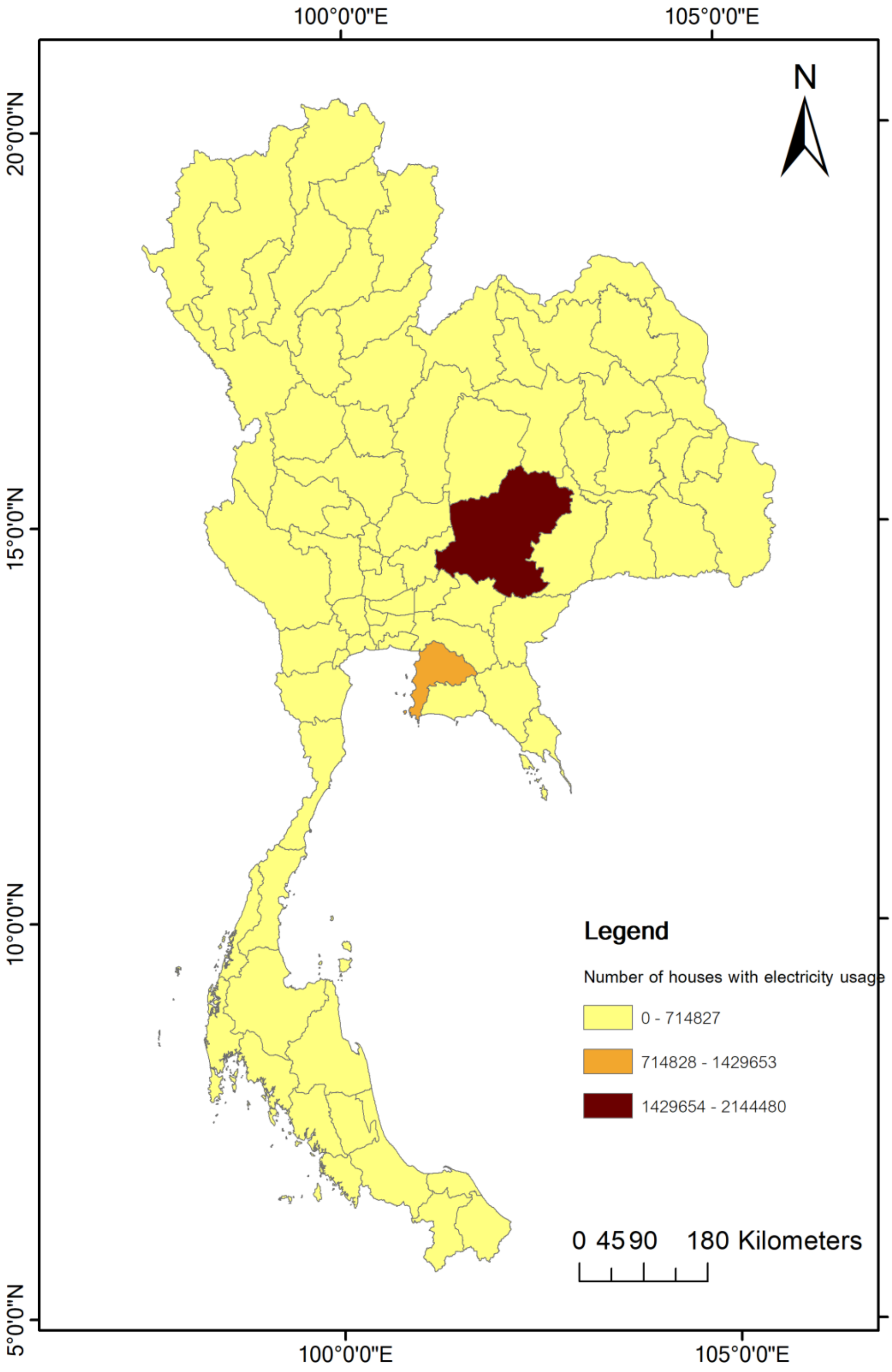


Figure 47: Map displays the number of houses with electricity usage

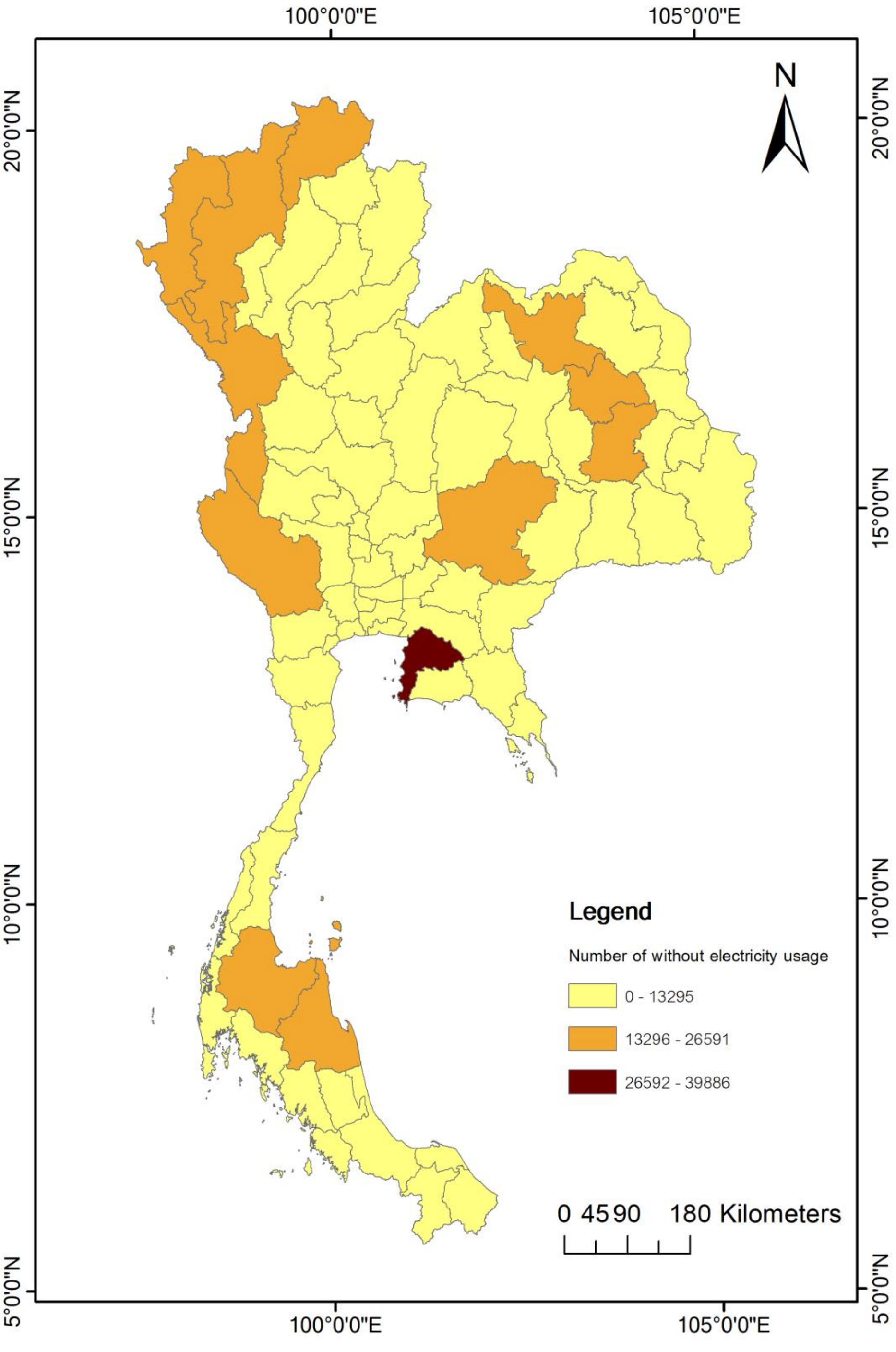


Figure 48: Map displays the number of houses without electricity usage

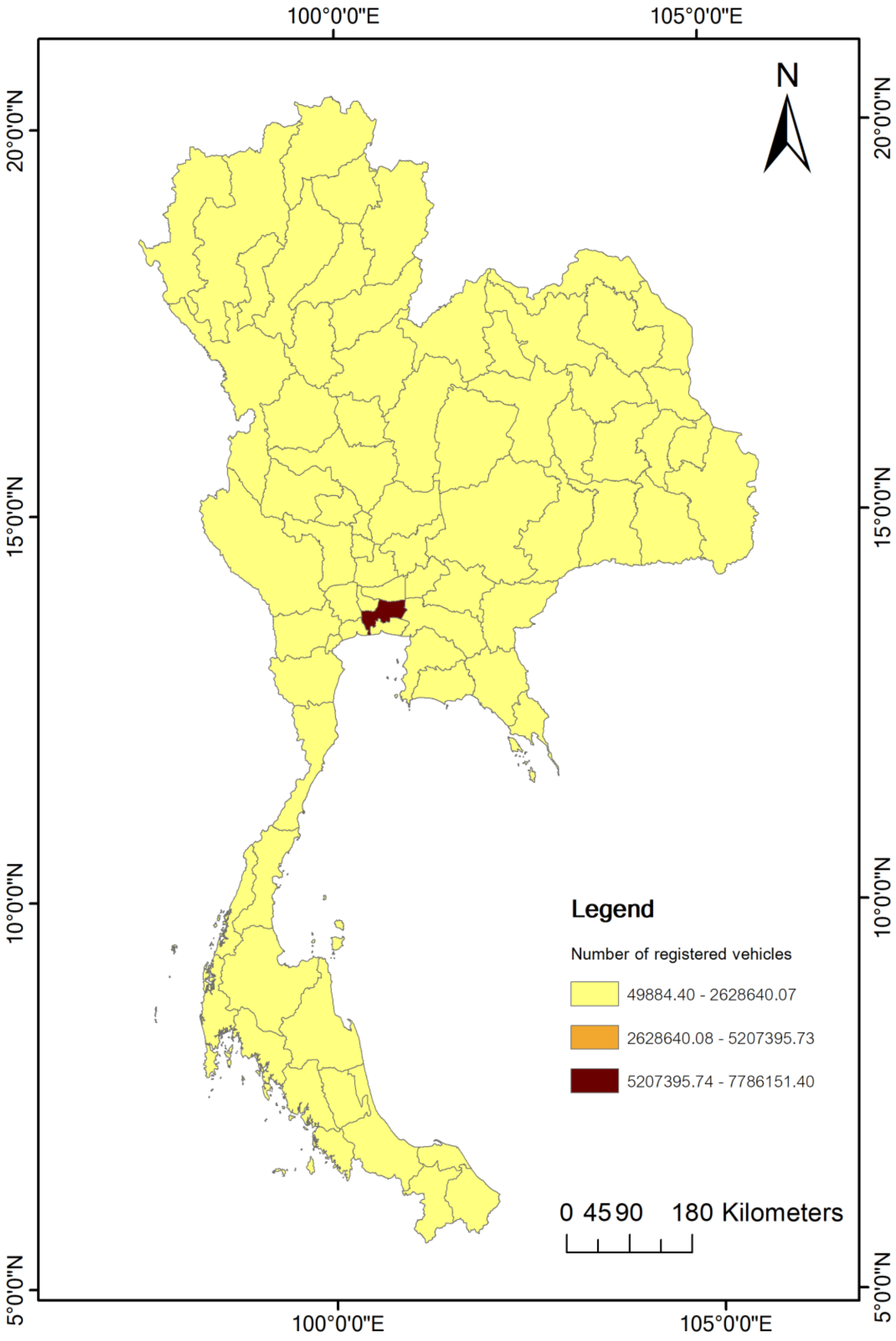


Figure 49: Map displays the number of registered vehicles

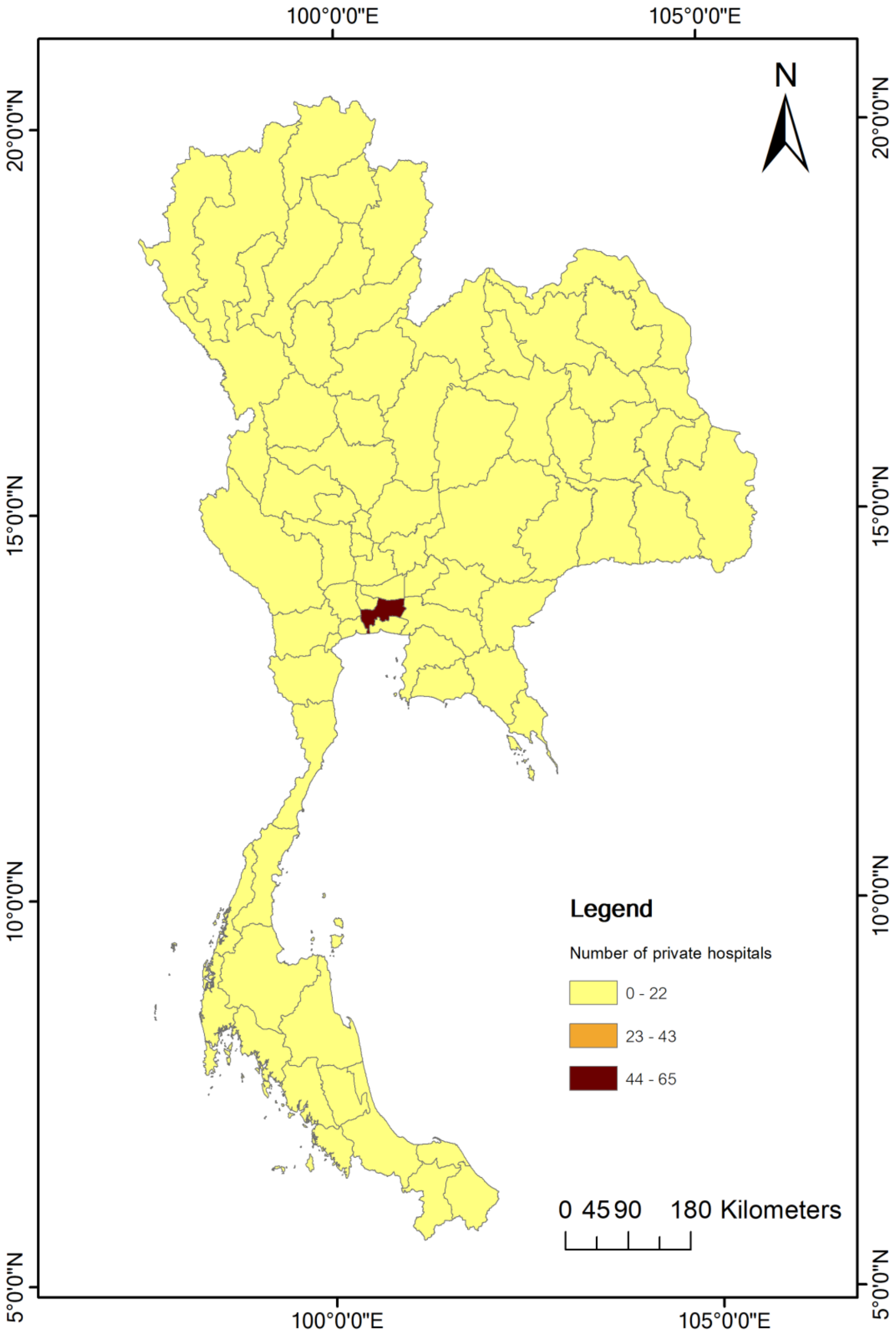


Figure 50: Map displays the number of private hospitals

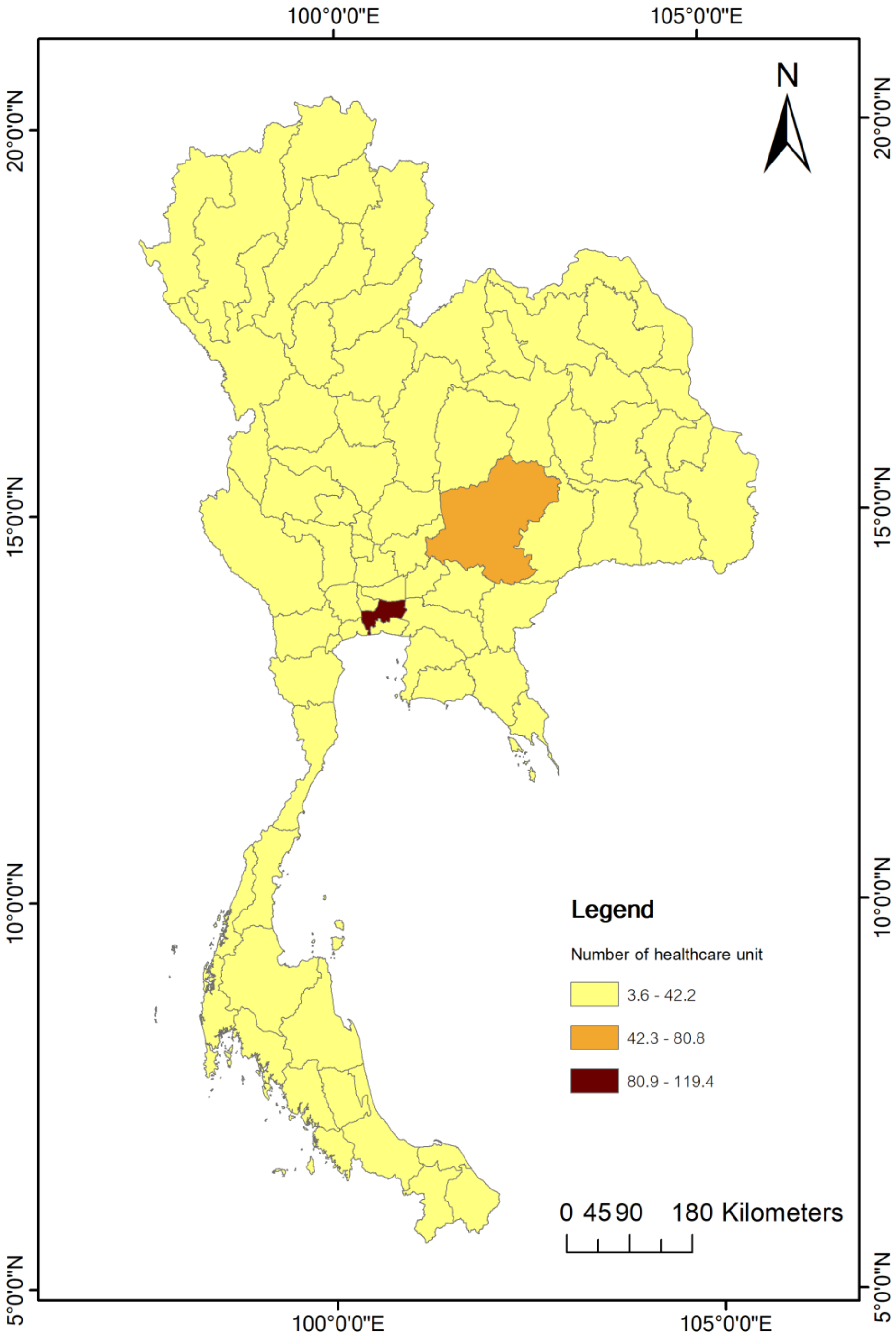


Figure 51: Map displays the number of healthcare unit

Chapter 5

Discussion

Malaria is a significant disease that harms people to death. Therefore, there are studies to detect and protect the spreading of the disease. The objective of the studies can be classified into categories. First category focuses on the study for factors that have direct impact to malaria transmission among people; this leads to the protection plan for each community area. These preventive factors mainly concerned to social and demographic such as occupation, household size, house-wall types, drinking water sources, irrigation areas, economic plants, and the nets usages, etc.

The second category focuses on the study for factors that have indirect impacts to malaria distribution which can lead to government's protection strategies. However, Thai government's still focuses on the small details of the community prevention rather than indirect factors that can power up the spreading of the disease, such as electricity usage, communication technology, rivers, canal and artesian well, etc. Therefore, in this research, these indirect factors are counted as parameters of the risk model. So, the awareness in malaria detection and protection can be completely controlled.

As the fact that many researches had identified factors related to the distribution this ailment, factors such as forests, vegetation cover, temperature, rainfall and humidity have impact to sustainability of this symptom (Amadi et al., 2018; Kar, Kumar, Singh, Carlton, & Nanda, 2014; Le, Kumar, Ruiz, Mbogo, & Muturi, 2019; Nath & Mwchahary, 2012; Ssempiira et al., 2018; Thomson et al., 2017). In addition, the changes of the land usages also affect to the malaria spreading (Paul, Kangalawe, & Mboera, 2018; Stefani et al., 2013). Although various factors are determined and monitored in malaria protection system, the spreading of this disease still exists, including the resistant drug occurs. Consequently, the number of malaria illness is increase and difficult to be managed. In order to solve such situation, this research had proposed new 28 variables that affect the existing of malaria. Moreover, a risk model of malaria distribution based on these variables was derived in the form of the LMT model with high accuracy rate.

In order to interpret the obtained LMT model, the significant variables that significantly affect to the spreading of malaria were taken into account; and a fundamental logistic regression model was created as shown in Table 14.

Table 14: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	NoIrrC	-1.975	.762	6.714	1	.010	.139	.031	.618
	ProvArea	1.173	.549	4.562	1	.033	3.233	1.102	9.486
	ProvAgr	-1.662	.771	4.644	1	.031	.190	.042	.860
	NoSw	1.570	.600	6.857	1	.009	4.808	1.484	15.577
	ElecFromGov	1.432	.859	2.780	1	.095	4.185	.778	22.518
	NoMigrate	-1.792	.743	5.811	1	.016	.167	.039	.715
	NoHouseNoElec	1.765	.662	7.115	1	.008	5.839	1.597	21.354
	Constant	-1.770	.466	14.406	1	.000	.170		

From Table14, the fundamental model of logistic regression for 7 main significant variables can be written as

$$\log(Odds) = \beta_0 + \beta_1 NoIrrC + \cdots + \beta_7 NoHouseNoElec$$

The $Odds$ refers to the odd ratio (OR) which is used to indicate the effect of independent variables. Thus, the value of Exp(B) in Table 14 can be interpreted as an effect from a variable in the table. For example, the Exp(B) of NoSw, or number of shallow well, is 4.808 means if the value of NoSw is increase one unit, the number of patients will be increase up to 5 times.

In addition, it is worth noting that the fire area variable and the number of times to put out the fire are two new variables that have been added to the model and cause the model's accuracy slightly raises up.

After obtaining the risk map based on 28 risk factors, this map was compared with the malaria distribution map using the patient numbers. The comparing results show that 8 out of 13 provinces of high risk area is similar to the high distribution of the patients, 2 out of 8 provinces in the medium risk area are the same as the medium distribution of the patients. Lastly, all provinces in the low risk area are the same as provinces in the low distribution of the patients.

Future works

The data in this study are public sources that can be easily accessed. However, there are some data that do not currently exist such as the mosquitoes' repository. Therefore, many research topics can be setup for further studies according to the mosquitoes' information, such as the number of mosquitoes, the spread of mosquitoes, etc. As a result, the malaria distribution model or risk model will be implemented with better accuracy or better TPR value.

Chapter 6

Conclusions

Malaria is a serious disease that can harm human' lives if the patients do not have a good care after infection. The infection causes by Anopheles mosquitoes that transmit this disease to human. Therefore, most researches focused to protect the infection parts, such as the use of nets, and the forest area where mosquitoes can grow, etc. As a result of these protections, the number of malaria patients is decreasing; unfortunately, the number of malaria drug resistant is increasing. So, in the long term, there is a possibility that the number of malaria sicknesses can be increase because of the increasing of drug resistant. This research proposed new 28 variables that most of them have hardly been mentioned by any existing researches. These 28 variables have been proved that they are clearly related to the spreading of malaria more than other existing variables. The risk model obtained from these new variables is in the LMT form with the precision and the recall values are 0.780 and 0.821, respectively. Furthermore, the true positive rate is the same as the recall value while the false positive rate is as small as 0.232. As a result of this finding, the government can consider these revealed factors and state suitable strategies to control them under proper criteria. For example, the government should control the quality of rivers or even increase the number of monitoring rivers to some area for reducing the malaria risk. Once these factors are completely managed under well-thought-out plans, each community can reduce the risk of malaria distribution without concerning of the patients' drug resistant.

In addition, displaying the results in the form of a risk map will give an image of the areas with various risk factors for the benefit of effective surveillance and control. Moreover, according to the risk map, each area in Thailand has been defined as high, medium, low distribution of malaria based on the discovery factors. In such result, the government can set suitable the malaria monitoring and protection strategies for each individual area with a proper budget to be provided.

References

Adde A, Roux E, Mangeas M, Dessay N, Nacher M, Dusfour I, Girod R, Briolant S (2016) Dynamical Mapping of Anopheles darlingi Densities in a Residual Malaria Transmission Area of French Guiana by Using Remote Sensing and Meteorological Data. PLOS ONE 11(10):e0164685-e0164685. https://doi:10.1371/journal.pone.0164685

Amadi JA, Olago DO, Ong'amo GO, Oriaso SO, Nanyingi M, Nyamongo IK, Estambale BBA (2018) Sensitivity of vegetation to climate variability and its implications for malaria risk in Baringo, Kenya. PloS one 13(7):e0199357-e0199357. https://doi:10.1371/journal.pone.0199357

Baeza A, Santos-Vega M, Dobson AP, Pascual M (2017) The rise and fall of malaria under land-use change in frontier regions. Nature Ecology &Amp; Evolution 1:0108. https://doi:10.1038/s41559-017-0108 https://www.nature.com/articles/s41559-017-0108#supplementary-information

Banda JM, Sarraju A, Abbasi F, Parizo J, Pariani M, Ison H, Briskin E, Wand H, Dubois S, Jung K, Myers SA, Rader DJ, Leader JB, Murray MF, Myers KD, Wilemon K, Shah NH, Knowles JW (2019) Finding missed cases of familial hypercholesterolemia in health systems using machine learning. npj Digital Medicine 2(1):23. https://doi:10.1038/s41746-019-0101-5

Bazargan M (2018) A case study on accessibility of medical and healthcare facilities in Mashhad using GIS. SAUES Journal 1(1):39-48. https://doi:10.22034/saues.2018.01.05

Boni MF, Smith DL, Laxminarayan R (2008) Benefits of using multiple first-line therapies against malaria. Proceedings of the National Academy of Sciences 105(37):14216-14221. https://doi:10.1073/pnas.0804628105

Buppan P, Putaporntip C, Pattanawong U, Seethamchai S, Jongwutiwes S (2010) Comparative detection of Plasmodium vivax and Plasmodium falciparum DNA in saliva and urine samples from symptomatic malaria patients in a low endemic area. Malaria Journal 9(1):72. https://doi:10.1186/1475-2875-9-72

Cahyaningrum P, Sulistyawati S (2018) Malaria Risk Factors in Kaligesing, Purworejo District, Central Java Province, Indonesia: A Case-control Study. J Prev Med Public Health 51(3):148-153. https://doi:10.3961/jpmph.18.036

Chew M. (2017). Retrieved from https://www.straitstimes.com/asia/se-asia/super-malaria-spreading-through-se-asia-poses-global-threat

Cohen J, Menach A, Pothin E, Eisele T, Gething P, Welkhoff P, Moonen B, Schapira A, Smith D (2017) Mapping multiple components of malaria risk for improved targeting of elimination interventions. Malaria Journal 16:459. https://doi:10.1186/s12936-017-2106-3

Congpuong K, Bualombai P, Banmairuroi V, Na-Bangchang K (2010) Compliance with a three-day course of artesunate-mefloquine combination and baseline anti-malarial treatment in an area of Thailand with highly multidrug resistant falciparum malaria. Malaria Journal 9(1):43. https://doi:10.1186/1475-2875-9-43

Duncan EW, Cramb SM, Aitken JF, Mengersen KL, Baade PD (2019) Development of the Australian Cancer Atlas: spatial modelling, visualisation, and reporting of estimates. International Journal of Health Geographics 18(1):21. https://doi:10.1186/s12942-019-0185-9

Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, Hickey AJ, Clark AM (2019) Exploiting machine learning for end-to-end drug discovery and development. Nature Materials 18(5):435-441. https://doi:10.1038/s41563-019-0338-z

Epidemiology. Malaria. Retrieved from http://www.boe.moph.go.th/fact/Malaria.htm

Eskenazi B, Levine DI, Rauch S, Obida M, Crause M, Bornman R, Chevrier J (2019) A community-based education programme to reduce insecticide exposure from indoor residual spraying in Limpopo, South Africa. Malaria Journal 18(1):199. https://doi:10.1186/s12936-019-2828-5

Essendi WM, Vardo-Zalik AM, Lo E, Machani MG, Zhou G, Githeko AK, Yan G, Afrane YA (2019) Epidemiological risk factors for clinical malaria infection in the highlands of Western Kenya. Malaria Journal 18(1):211. https://doi:10.1186/s12936-019-2845-4

Fillekes MP, Giannouli E, Kim E-K, Zijlstra W, Weibel R (2019) Towards a comprehensive set of GPS-based indicators reflecting the multidimensional nature of daily mobility for applications in health and aging research. International Journal of Health Geographics 18(1):17. https://doi:10.1186/s12942-019-0181-0

Fontecha G, Pinto A, Escobar D, Matamoros G, Ortiz B (2019) Genetic variability of Plasmodium falciparum histidine-rich proteins 2 and 3 in Central America. Malaria Journal 18(1):31-31. https://doi:10.1186/s12936-019-2668-3

Fradelos EC, Papathanasiou IV, Mitsi D, Tsaras K, Kleisiaris CF, Kourkouta L (2014) Health Based Geographic Information Systems (GIS) and their Applications. Acta informatica medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH 22(6):402-405. https://doi:10.5455/aim.2014.22.402-405

Greenwood B (2010) Anti-malarial drugs and the prevention of malaria in the population of malaria endemic areas. Malaria Journal 9(3):S2. https://doi:10.1186/1475-2875-9-S3-S2

Guerra M, de Sousa B, Ndong-Mabale N, Berzosa P, Arez AP (2018) Malaria determining risk factors at the household level in two rural villages of mainland Equatorial Guinea. Malaria Journal 17(1):203. https://doi:10.1186/s12936-018-2354-x

Hast M, Searle KM, Chaponda M, Lupiya J, Lubinda J, Sikalima J, Kobayashi T, Shields T, Mulenga M, Lessler J, Moss WJ, for t, Central Africa International Centers of Excellence for Malaria R (2019) The use of GPS data loggers to describe the impact of spatio-temporal movement patterns on malaria control in a high-transmission area of northern Zambia. International Journal of Health Geographics 18(1):19. https://doi:10.1186/s12942-019-0183-y

Ibrahim Ramzi A, Abdl-Latif El-Bedawi M (2019) Towards integration of remote sensing and GIS to manage primary health care centers. Applied Computing and Informatics 15(2):109-113. https://doi:https://doi.org/10.1016/j.aci.2017.12.001

Imwong M, Tinh Tran H, Nguyen T-N, Dondorp A, White N (2017) Spread of a single multidrug resistant malaria parasite lineage (PfPailin) to Vietnam. The Lancet Infectious Diseases 17:1022-1023. https://doi:10.1016/S1473-3099(17)30524-8

Kaehler N, Adhikari B, Cheah PY, von Seidlein L, Day NPJ, Paris DH, Tanner M, Pell C (2019) Prospects and strategies for malaria elimination in the Greater Mekong Sub-region: a qualitative study. Malaria Journal 18(1):203. https://doi:10.1186/s12936-019-2835-6

Kar NP, Kumar A, Singh OP, Carlton JM, Nanda N (2014) A review of malaria transmission dynamics in forest ecosystems. Parasites & vectors 7:265-265. https://doi:10.1186/1756-3305-7-265

Kenea O, Balkew M, Tekie H, Deressa W, Loha E, Lindtjørn B, Overgaard HJ (2019) Impact of combining indoor residual spraying and long-lasting insecticidal nets on Anopheles arabiensis in Ethiopia: results from a cluster randomized controlled trial. Malaria Journal 18(1):182. https://doi:10.1186/s12936-019-2811-1

Kunwar S, Shrestha M, Shikhrakar R (2018) Malaria Detection Using Image Processing and Machine Learning.

Kwansomboon N, Chaumeau V, Kittiphanakun P, Cerqueira D, Corbel V, Chareonviriyaphap T (2017) Vector bionomics and malaria transmission along the Thailand-Myanmar border: a baseline entomological survey. Journal of Vector Ecology 42(1):84-93. https://doi:doi:10.1111/jvec.12242

Landwehr N, Hall M, Frank E (2005) Logistic Model Trees. Machine Learning 59(1):161-205. https://doi:10.1007/s10994-005-0466-3

Lawpoolsri S, Chavez IF, Yimsamran S, Puangsa-art S, Thanyavanich N, Maneeboonyang W, Chaimungkun W, Singhasivanon P, Maguire JH, Hungerford LL (2010) The impact of human reservoir of malaria at a community-level on individual malaria occurrence in a low malaria transmission setting along the Thai-Myanmar border. Malaria Journal 9(1):143. https://doi:10.1186/1475-2875-9-143

Lawpoolsri S, Sattabongkot J, Sirichaisinthop J, Cui L, Kiattibutr K, Rachaphaew N, Suk-uam K, Khamsiriwatchara A, Kaewkungwal J (2019) Epidemiological profiles of recurrent malaria episodes in an endemic area along the Thailand-Myanmar border: a prospective cohort study. Malaria Journal 18(1):124. https://doi:10.1186/s12936-019-2763-5

Le PVV, Kumar P, Ruiz MO, Mbogo C, Muturi EJ (2019) Predicting the direct and indirect impacts of climate change on malaria in coastal Kenya. PloS one 14(2):e0211258-e0211258. https://doi:10.1371/journal.pone.0211258

Lertpiriyasuwat C. (2019).   Retrieved from Retrieved from https://www.matichon.co.th/local/quality-life/news_1594405

Linn SY, Maung TM, Tripathy JP, Shewade HD, Oo SM, Linn Z, Thi A (2019) Barriers in distribution, ownership and utilization of insecticide-treated mosquito nets among migrant population in Myanmar, 2016: a mixed methods study. Malaria Journal 18(1):172. https://doi:10.1186/s12936-019-2800-4

Lo C, Dia AK, Dia I, Niang EHA, Konaté L, Faye O (2019) Evaluation of the residual efficacy of indoor residual spraying with bendiocarb (FICAM WP 80) in six health districts in Senegal. Malaria Journal 18(1):198. https://doi:10.1186/s12936-019-2829-4

Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, Lee S-I (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering 2(10):749-760. https://doi:10.1038/s41551-018-0304-0

Maclachlan JC, Abernathy T, Jerrett M (2004) Developing an internet GIS for public health. Journal of Spatial Science 49(2):121-127. https://doi:10.1080/14498596.2004.9635027

Manak MS, Varsanik JS, Hogan BJ, Whitfield MJ, Su WR, Joshi N, Steinke N, Min A, Berger D, Saphirstein RJ, Dixit G, Meyyappan T, Chu H-M, Knopf KB, Albala DM, Sant GR, Chander AC (2018) Live-cell phenotypic-biomarker microfluidic assay for the risk stratification of cancer patients via machine learning. Nature Biomedical Engineering 2(10):761-772. https://doi:10.1038/s41551-018-0285-z

Martin C, Curtis B, Fraser C, Sharp B (2002) The use of a GIS-based malaria information system for malaria research and control in South Africa. Health & Place 8(4):227-236. https://doi:https://doi.org/10.1016/S1353-8292(02)00008-4

Menke W. (2018). Chapter 10 - Factor Analysis. In W. Menke (Ed.), *Geophysical Data Analysis (Fourth Edition)* (pp. 207-222): Academic Press.

Mercado CEG, Lawpoolsri S, Sudathip P, Kaewkungwal J, Khamsiriwatchara A, Pan-ngum W, Yimsamran S, Lawawirojwong S, Ho K, Ekapirat N, Maude RR, Wiladphaingern J, Carrara VI, Day NPJ, Dondorp AM, Maude RJ (2019) Spatiotemporal epidemiology, environmental correlates, and demography of malaria in Tak Province, Thailand (2012–2015). Malaria Journal 18(1):240. https://doi:10.1186/s12936-019-2871-2

Mwanga EP, Mapua SA, Siria DJ, Ngowo HS, Nangacha F, Mgando J, Baldini F, González Jiménez M, Ferguson HM, Wynne K, Selvaraj P, Babayan SA, Okumu FO (2019) Using mid-infrared spectroscopy and supervised machine-learning to identify vertebrate blood meals in the malaria vector, Anopheles arabiensis. Malaria Journal 18(1):187. https://doi:10.1186/s12936-019-2822-y

Nath D, Mwchahary D (2012) Malaria Prevalence in Forest and Nonforest Areas of Kokrajhar District of Assam. ISRN Public Health 2012. https://doi:10.5402/2012/142037

Noé A, Zaman SI, Rahman M, Saha AK, Aktaruzzaman MM, Maude RJ (2018) Mapping the stability of malaria hotspots in Bangladesh from 2013 to 2016. Malaria Journal 17(1):259. https://doi:10.1186/s12936-018-2405-3

Nykiforuk CIJ, Flaman LM (2009) Geographic Information Systems (GIS) for Health Promotion and Public Health: A Review. Health Promotion Practice 12(1):63-73. https://doi:10.1177/1524839909334624

Ocampo AJ, Chunara R, Brownstein JS (2013) Using search queries for malaria surveillance, Thailand. Malaria Journal 12(1):390. https://doi:10.1186/1475-2875-12-390

Okami S, Kohtake N (2016) Fine-Scale Mapping by Spatial Risk Distribution Modeling for Regional Malaria Endemicity and Its Implications under the Low-to-Moderate Transmission Setting in Western Cambodia. PLOS ONE 11(7):e0158737. https://doi:10.1371/journal.pone.0158737

Olugboja A, Wang Z. (2017, 9-12 July 2017). *Malaria parasite detection using different machine learning classifier.* Paper presented at the 2017 International Conference on Machine Learning and Cybernetics (ICMLC).

Pandis N (2016) Multiple linear regression analysis. American Journal of Orthodontics and Dentofacial Orthopedics 149(4):581. https://doi:https://doi.org/10.1016/j.ajodo.2016.01.012

Paul P, Kangalawe RYM, Mboera LEG (2018) Land-use patterns and their implication on malaria transmission in Kilosa District, Tanzania. Tropical diseases, travel medicine and vaccines 4:6-6. https://doi:10.1186/s40794-018-0066-4

Perveen S, Shahbaz M, Keshavjee K, Guergachi A (2019) Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique. Scientific Reports 9(1):13805. https://doi:10.1038/s41598-019-49563-6

Platt A, Obala AA, MacIntyre C, Otsyula B, Meara WPO (2018) Dynamic malaria hotspots in an open cohort in western Kenya. Scientific Reports 8(1):647. https://doi:10.1038/s41598-017-13801-6

Poostchi M, Silamut K, Maude RJ, Jaeger S, Thoma G (2018) Image analysis and machine learning for detecting malaria. Translational Research 194:36-55. https://doi:https://doi.org/10.1016/j.trsl.2017.12.004

Qayum A, Arya R, Kumar P, Lynn AM (2015) Socio-economic, epidemiological and geographic features based on GIS-integrated mapping to identify malarial hotspots. Malaria Journal 14(1):192. https://doi:10.1186/s12936-015-0685-4

Raschka S (2015) *Python Machine Learning*. Packt Publishing, Birmingham, UK

Recht J, Siqueira AM, Monteiro WM, Herrera SM, Herrera S, Lacerda MVG (2017) Malaria in Brazil, Colombia, Peru and Venezuela: current challenges in malaria control and elimination. Malaria Journal 16(1):273. https://doi:10.1186/s12936-017-1925-6

Saita S, Pan-ngum W, Phuanukoonnon S, Sriwichai P, Silawan T, White LJ, Parker DM (2019) Human population movement and behavioural patterns in malaria hotspots on the Thai–Myanmar border: implications for malaria elimination. Malaria Journal 18(1):64. https://doi:10.1186/s12936-019-2704-3

Selvaraj P, Suresh J, Wenger EA, Bever CA, Gerardin J (2019) Reducing malaria burden and accelerating elimination with long-lasting systemic insecticides: a modelling study of three potential use cases. Malaria Journal 18(1):307. https://doi:10.1186/s12936-019-2942-4

Sperandei S (2014) Understanding logistic regression analysis. Biochemia medica 24(1):12-18. https://doi:10.11613/BM.2014.003

Srivastava A, Nagpal BN, Saxena R, Eapen A, Ravindran KJ, Subbarao SK, Rajamanikam C, Palanisamy M, Kalra NL, Appavoo NC (2003) GIS based malaria information management system for urban malaria scheme in India. Computer Methods and Programs in Biomedicine 71(1):63-75. https://doi:https://doi.org/10.1016/S0169-2607(02)00056-1

Ssempiira J, Kissa J, Nambuusi B, Mukooyo E, Opigo J, Makumbi F, Kasasa S, Vounatsou P (2018) Interactions between climatic changes and intervention effects on malaria spatio-temporal dynamics in Uganda. Parasite epidemiology and control 3(3):e00070-e00070. https://doi:10.1016/j.parepi.2018.e00070

Stefani A, Dusfour I, Corrêa APSA, Cruz MCB, Dessay N, Galardo AKR, Galardo CD, Girod R, Gomes MSM, Gurgel H, Lima ACF, Moreno ES, Musset L, Nacher M, Soares ACS, Carme B, Roux E (2013) Land cover, land use and malaria in the Amazon: a systematic literature review of studies using remotely sensed data. Malaria Journal 12(1):192. https://doi:10.1186/1475-2875-12-192

Stuckey EM, Stevenson J, Galactionova K, Baidjoe AY, Bousema T, Odongo W, Kariuki S, Drakeley C, Smith TA, Cox J, Chitnis N (2014) Modeling the Cost Effectiveness of Malaria Control Interventions in the Highlands of Western Kenya. PLOS ONE 9(10):e107700. https://doi:10.1371/journal.pone.0107700

Tangpukdee N, Krudsood S, Srivilairit S, Phophak N, Chonsawat P, Yanpanich W, Kano S, Wilairatana P (2008) Gametocyte clearance in uncomplicated and severe Plasmodium falciparum malaria after artesunate-mefloquine treatment in Thailand. The Korean journal of parasitology 46(2):65-70. https://doi:10.3347/kjp.2008.46.2.65

Tesfahunegn A, Berhe G, Gebregziabher E (2019) Risk factors associated with malaria outbreak in Laelay Adyabo district northern Ethiopia, 2017: case-control study design. BMC public health19(1):484-484.https://doi:10.1186/s12889-019-6798-x

Thaitravelclinic. Retrieved from https://www.thaitravelclinic.com/blog/th/category/all-about-malaria

Thomson MC, Ukawuba I, Hershey CL, Bennett A, Ceccato P, Lyon B, Dinku T (2017) Using Rainfall and Temperature Data in the Evaluation of National Malaria Control Programs in Africa. The American journal of tropical medicine and hygiene 97(3_Suppl):32-45. https://doi:10.4269/ajtmh.16-0696

Tokarz R, Novak RJ (2018) Spatial–temporal distribution of Anopheles larval habitats in Uganda using GIS/remote sensing technologies. Malaria Journal 17(1):420. https://doi:10.1186/s12936-018-2567-z

UNHCR UNHCfR. THE GLOBAL MALARIA ACTION PLAN. Retrieved from Retrieved from https://www.unhcr.org/4afac5629.pdf

World Health Organization. (2017). World Malaria Report Retrieved from https://www.who.int/malaria/publications/world-malaria-report-2018/report/en/

World Health Organization. (2018). World Malaria Report Retrieved from https://www.who.int/malaria/publications/world-malaria-report-2018/report/en/

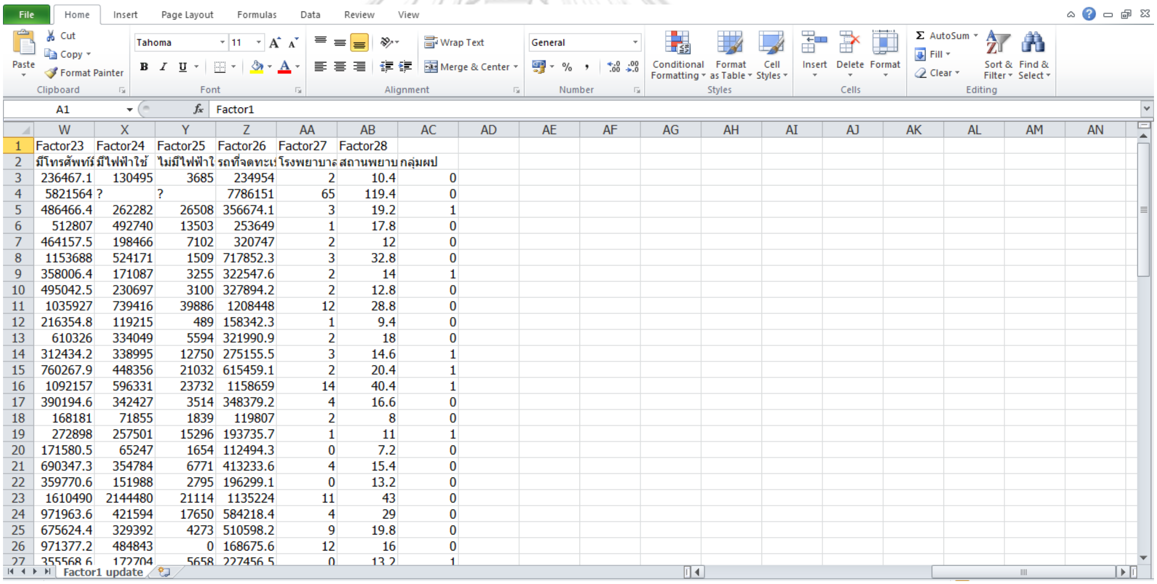
Yankson R, Anto EA, Chipeta MG (2019) Geostatistical analysis and mapping of malaria risk in children under 5 using point-referenced prevalence data in Ghana. Malaria Journal 18(1):67. https://doi:10.1186/s12936-019-2709-y

Zhu L, Mok S, Imwong M, Jaidee A, Russell B, Nosten F, Day NP, White NJ, Preiser PR, Bozdech Z (2016) New insights into the Plasmodium vivax transcriptome using RNA-Seq. Scientific Reports 6:20498. https://doi:10.1038/srep20498 https://www.nature.com/articles/srep20498#supplementary-information

Appendix A

Pre-processing – Data preparation

Most of the time, the data wouldn't be perfect, and we would need to do pre-processing before applying machine learning algorithms on it. Doing pre-processing is easy in Weka. You can simply click the "Open file" button and load your file as certain file types: Arff, CSV, C4.5, binary, LIBSVM, XRFF. You can also load SQL db file via the URL; then, you can apply filters to it. However, we won't need to do pre-processing for this post since we'll use a dataset that Weka provides for us. Figure A.1 shows data samples from 28 variables obtained from Factor Analysis, which is highly correlated with the malaria distribution.



A1 Factor1

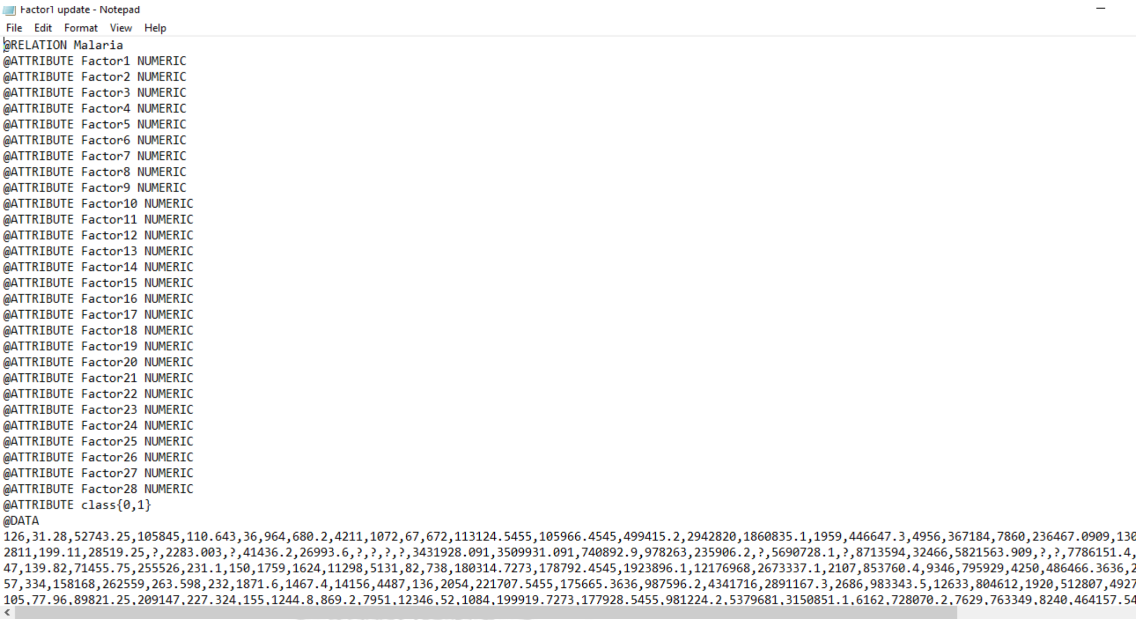
	W	X	Y	Z	AA	AB	AC
1	Factor23	Factor24	Factor25	Factor26	Factor27	Factor28	
2	มีโทรศัพท์	มีไฟฟ้าใช้	ไม่มีไฟฟ้าใ	รถที่จดทะเ	โรงพยาบาล	สถานพยาบ	กลุ่มผป
3	236467.1	130495	3685	234954	2	10.4	0
4	5821564	?	?	7786151	65	119.4	0
5	486466.4	262282	26508	356674.1	3	19.2	1
6	512807	492740	13503	253649	1	17.8	0
7	464157.5	198466	7102	320747	2	12	0
8	1153688	524171	1509	717852.3	3	32.8	0
9	358006.4	171087	3255	322547.6	2	14	1
10	495042.5	230697	3100	327894.2	2	12.8	0
11	1035927	739416	39886	1208448	12	28.8	0
12	216354.8	119215	489	158342.3	1	9.4	0
13	610326	334049	5594	321990.9	2	18	0
14	312434.2	338995	12750	275155.5	3	14.6	1
15	760267.9	448356	21032	615459.1	2	20.4	1
16	1092157	596331	23732	1158659	14	40.4	1
17	390194.6	342427	3514	348379.2	4	16.6	0
18	168181	71855	1839	119807	2	8	0
19	272898	257501	15296	193735.7	1	11	1
20	171580.5	65247	1654	112494.3	0	7.2	0
21	690347.3	354784	6771	413233.6	4	15.4	0
22	359770.6	151988	2795	196299.1	0	13.2	0
23	1610490	2144480	21114	1135224	11	43	0
24	971963.6	421594	17650	584218.4	4	29	0
25	675624.4	329392	4273	510598.2	9	19.8	0
26	971377.2	484843	0	168675.6	12	16	0
27	355568.6	172704	5658	227456.5	0	13.2	1

Figure A.1. Dataset 28 Variations of malaria distribution

If your data type is in xls format like in Figure 1, you have to convert the file. The following processes illustrate the conversion of the file format:

1. Convert your **.xls** to **.csv** format
2. Open your CSV file in any text editor and first add **@RELATION** database_name to the first row of the CSV file

3. Add attributes using the following definition: **@ATTRIBUTE** attr_name attr_type. If attr_type is numeric, it must be defined as **REAL** Sample images are displayed in Figure A.2.
4. Then, add a **@DATA** tag just above the data rows, and save the file with **.arff** extension.



```
@RELATION Malaria
@ATTRIBUTE Factor1 NUMERIC
@ATTRIBUTE Factor2 NUMERIC
@ATTRIBUTE Factor3 NUMERIC
@ATTRIBUTE Factor4 NUMERIC
@ATTRIBUTE Factor5 NUMERIC
@ATTRIBUTE Factor6 NUMERIC
@ATTRIBUTE Factor7 NUMERIC
@ATTRIBUTE Factor8 NUMERIC
@ATTRIBUTE Factor9 NUMERIC
@ATTRIBUTE Factor10 NUMERIC
@ATTRIBUTE Factor11 NUMERIC
@ATTRIBUTE Factor12 NUMERIC
@ATTRIBUTE Factor13 NUMERIC
@ATTRIBUTE Factor14 NUMERIC
@ATTRIBUTE Factor15 NUMERIC
@ATTRIBUTE Factor16 NUMERIC
@ATTRIBUTE Factor17 NUMERIC
@ATTRIBUTE Factor18 NUMERIC
@ATTRIBUTE Factor19 NUMERIC
@ATTRIBUTE Factor20 NUMERIC
@ATTRIBUTE Factor21 NUMERIC
@ATTRIBUTE Factor22 NUMERIC
@ATTRIBUTE Factor23 NUMERIC
@ATTRIBUTE Factor24 NUMERIC
@ATTRIBUTE Factor25 NUMERIC
@ATTRIBUTE Factor26 NUMERIC
@ATTRIBUTE Factor27 NUMERIC
@ATTRIBUTE Factor28 NUMERIC
@ATTRIBUTE class{0,1}
@DATA
126,31.28,52743.25,105845,110.643,36,964,680.2,4211,1072,67,672,113124.5455,105966.4545,499415.2,2942820,1860835.1,1959,446647.3,4956,367184,7860,236467.0909,13
2811,199.11,28519.25,?,2283.003,?,41436.2,26993.6,?,?,?,?,3431928.091,3509931.091,740892.9,978263,235906.2,?,5690728.1,?,8713594,32466,5821563.909,?,?,7786151.4,
47,139.82,71455.75,255526,231.1,150,1759,1624,11298,5131,82,738,180314.7273,178792.4545,1923896.1,12176968,2673337.1,2107,853760.4,9346,795929,4250,486466.3636,2
57,334,158168,262559,263.598,232,1871.6,1467.4,14156,4487,136,2054,221707.5455,175665.3636,987596.2,4341716,2891167.3,2686,983343.5,12633,804612,1920,512807,4927
105,77.96,89821.25,209147,227.324,155,1244.8,869.2,7951,12346,52,1084,199919.7273,177928.5455,981224.2,5379681,3150851.1,6162,728070.2,7629,763349,8240,464157.54
```

Figure A.2. Show file data .arff

Model derivation process by Machine Learning Software: Weka

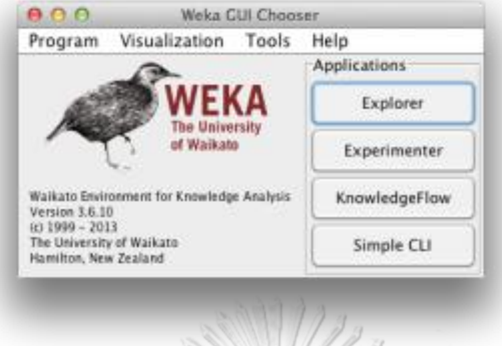
In order to gain an efficient risk model from 28 factors, Weka is deployed to derive this model by the following steps.

1. Download Weka and Install: see the installation method from Weka download.

2. Start Weka: double clicking on the weka.jar file**.** The screen in Figure A.3 will be displayed. 4 buttons in the screen are described below.

1. ***Simple CLI*** is a simple command line interface provided to run Weka functions directly.
2. ***Explorer*** is an environment to discover the data.
3. ***Experimenter*** is an environment to make experiments and statistical tests between learning schemes.

4. ***Knowledge Flow*** is a Java-Beans based interface for tuning and machine learning experiments.



Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Figure A.3. The Weka GUI Chooser

Analysis process for model driven

Click the "***Explorer***" button to call sub-features shown in Figure A.4.

Load the data

Click the "***Open file***" button from the **Preprocess** section and load data file with the**.arff** format from the local file system.

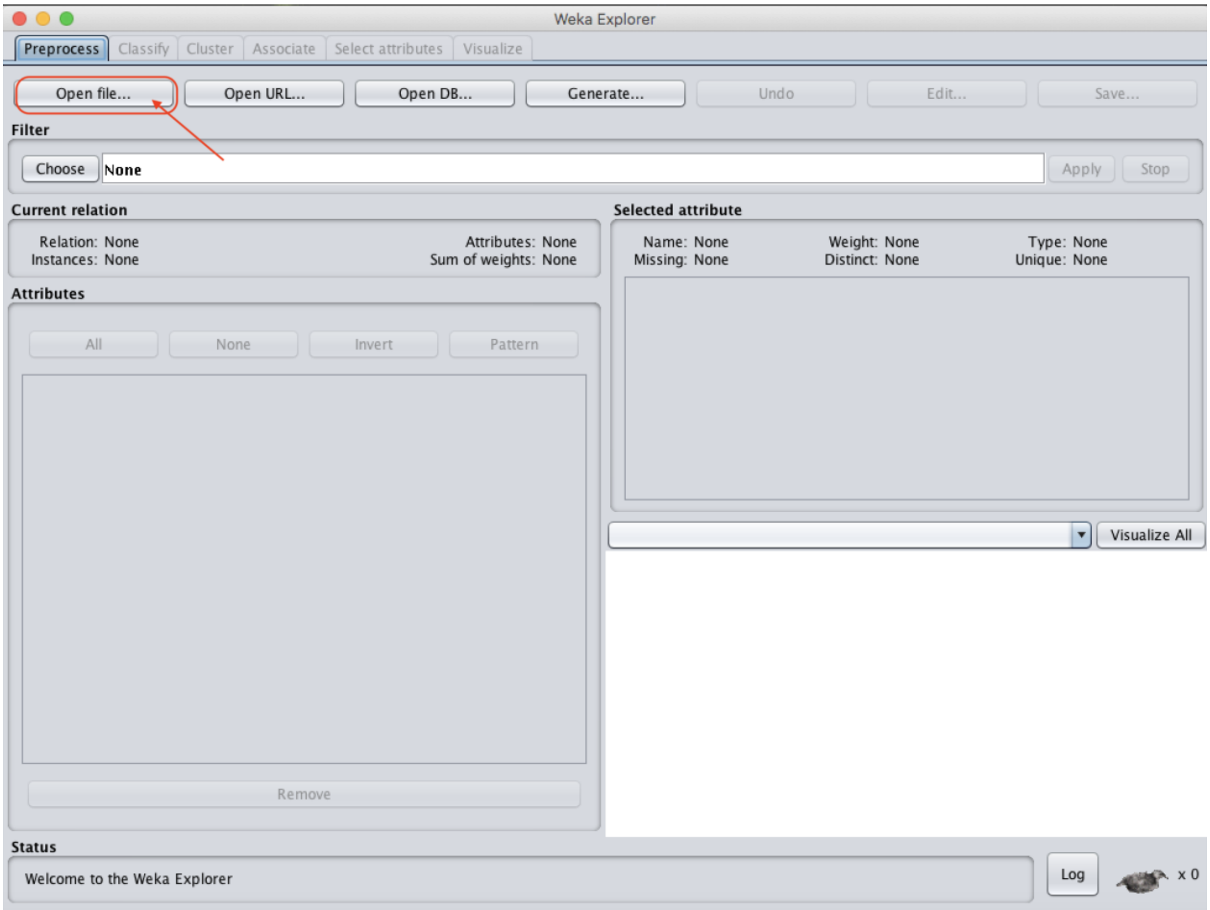
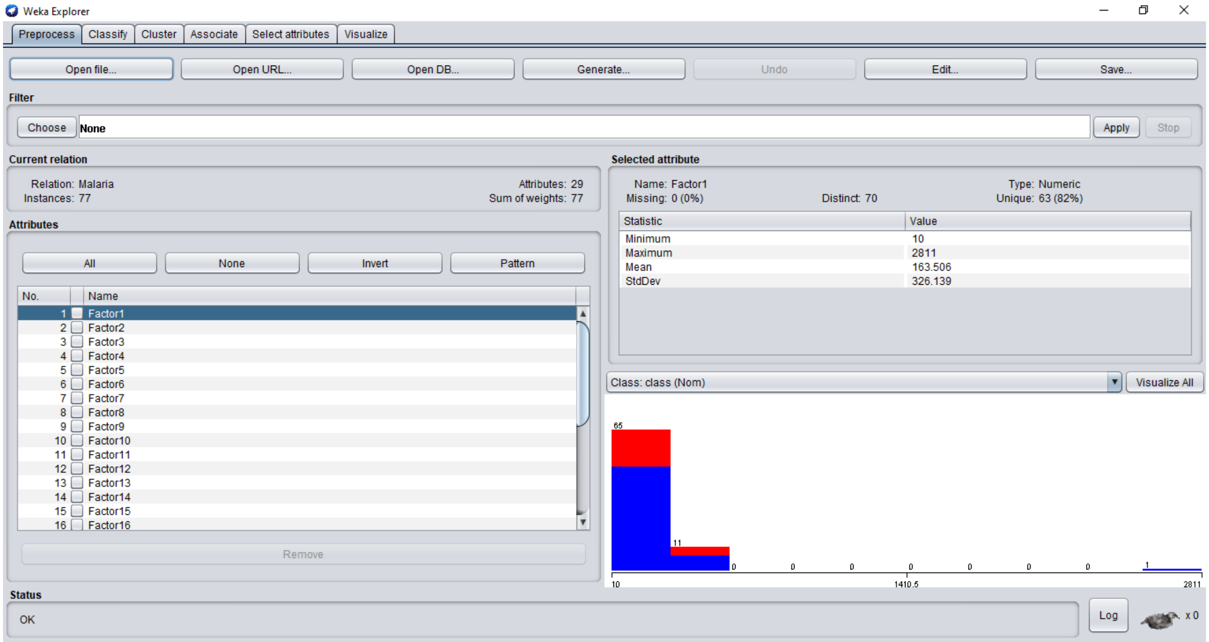


Figure A.4. Pre-process in weka

3. Open the data/ Dataset

Click the "*Open file...*" button to open a data set and double click on the "*data*" directory.

Select the "*data.arff*" file to load the dataset, as shown in Figure A.5.



Filter: None

Relation: Malaria
Instances: 77
Attributes: 29
Sum of weights: 77

Name: Factor1
Missing: 0 (0%)
Distinct: 70
Type: Numeric
Unique: 63 (82%)

Statistic	Value
Minimum	10
Maximum	2811
Mean	163.506
StdDev	326.139

No.		Name
1	□	Factor1
2	□	Factor2
3	□	Factor3
4	□	Factor4
5	□	Factor5
6	□	Factor6
7	□	Factor7
8	□	Factor8
9	□	Factor9
10	□	Factor10
11	□	Factor11
12	□	Factor12
13	□	Factor13
14	□	Factor14
15	□	Factor15
16	□	Factor16

Class: class (Nom)

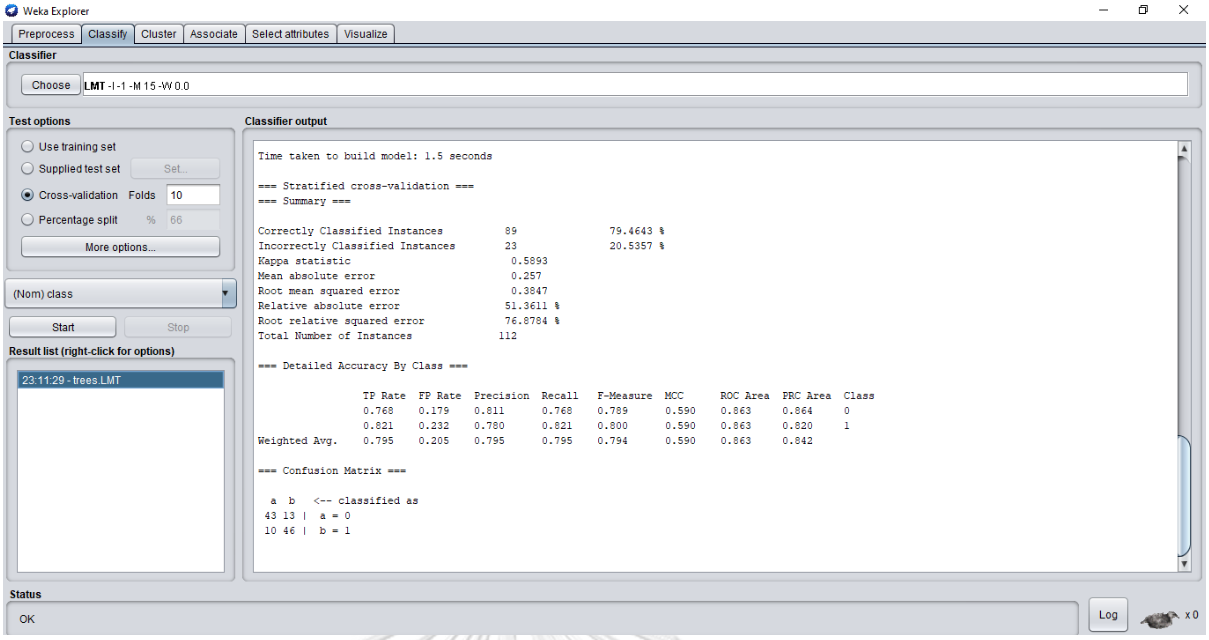
OK

Figure A.5. Weka Explorer Interface with the dataset loaded

4. Select and Run an Algorithm

Now, the dataset was loaded, choose a machine learning algorithm to model the problem and make predictions.

Click the "*Choose*" button in the **Classifier** section and click on "trees" and click on the "LMT" algorithm. Click the "*Start*" button to run this algorithm. The result is presented in Figure A.6.



Classifier: LMT -I -1 -M 15 -W 0.0

● Cross-validation Folds 10

23:11:29 - trees.LMT

Status: OK

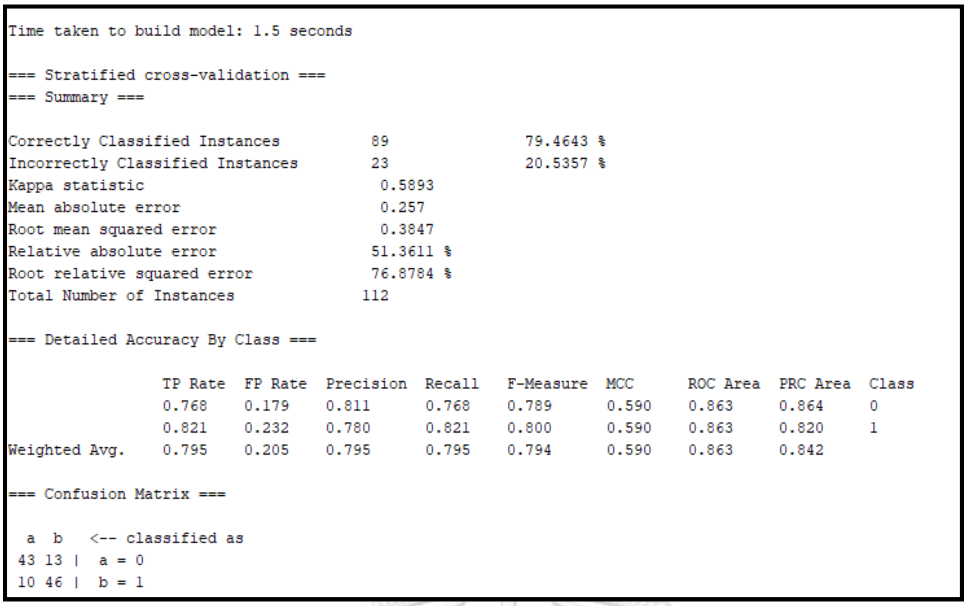
```
Time taken to build model: 1.5 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          89               79.4643 %
Incorrectly Classified Instances        23               20.5357 %
Kappa statistic                          0.5893
Mean absolute error                      0.257
Root mean squared error                  0.3847
Relative absolute error                 51.3611 %
Root relative squared error             76.8784 %
Total Number of Instances              112

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.768    0.179    0.811      0.768    0.789      0.590    0.863     0.864     0
                 0.821    0.232    0.780      0.821    0.800      0.590    0.863     0.820     1
Weighted Avg.    0.795    0.205    0.795      0.795    0.794      0.590    0.863     0.842

=== Confusion Matrix ===

  a  b   <-- classified as
 43 13 |  a = 0
 10 46 |  b = 1
```

Figure A.6. Weka Results for the LMT algorithm on the dataset

You will also note that the test options select Cross Validation by default with 10 folds. This means that the dataset is split into 10 parts: the first 9 are used to train the algorithm, and the 10th is used to assess the algorithm. This process is repeated, allowing each of the 10 parts of the split dataset a chance to be the held-out test set.

5. Review Results

After running the LMT algorithm, you can note the results in the **Classifier output** section, as show in Figure A.7. The algorithm was run with 10-fold cross-validation: this means it was given an opportunity to make a prediction for each instance of the dataset (with different training folds) and the presented result is a summary of those predictions.



```
Time taken to build model: 1.5 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          89               79.4643 %
Incorrectly Classified Instances        23               20.5357 %
Kappa statistic                          0.5893
Mean absolute error                      0.257
Root mean squared error                  0.3847
Relative absolute error                 51.3611 %
Root relative squared error             76.8784 %
Total Number of Instances              112

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.768    0.179    0.811      0.768    0.789      0.590    0.863     0.864     0
                 0.821    0.232    0.780      0.821    0.800      0.590    0.863     0.820     1
Weighted Avg.    0.795    0.205    0.795      0.795    0.794      0.590    0.863     0.842

=== Confusion Matrix ===

  a  b   <-- classified as
 43 13 |  a = 0
 10 46 |  b = 1
```

Figure A.7. Results of the LMT algorithm on the dataset in Weka

REFERENCES

VITA

NAME	Patcharaporn Krainara
DATE OF BIRTH	11 March
PLACE OF BIRTH	Nakhon Si Thammarat
INSTITUTIONS ATTENDED	Bachelor’s degree in statistics from Faculty of Science and Technology, Thammasat University, following by a Master’s degree in Statistics Information Technology from Faculty of Commerce and Accountancy, Chulalongkorn University.
HOME ADDRESS	89 Borommaratchachonnani Rd, Arun Amarin, Bangkok Noi, Bangkok 10700