โครงการการเรียนการสอนเพื่อเสริมประสบการณ์

การเรียนรู้ของเครื่องจักรในการพยากรณ์อนุกรมเวลา การประยุกต์เพื่อวิเคราะห์พื้นผิวน้ำแม่น้ำมูล

โดย

นายอานันท์ อึ้งปัญสัตวงศ์

เลขประจำตัวนิสิต 6032740723

โครงการนี้เป็นส่วนหนึ่งของการศึกษาระดับปริญญาตรี

การเรียนรู้ของเครื่องจักรในการพยากรณ์อนุกรมเวลา การประยุกต์เพื่อวิเคราะห์พื้นผิวน้ำแม่น้ำมูล

นายอานันท์ อึ้งปัญสัตวงศ์

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตร์บัณฑิต
ภาควิชาธรณีวิทยา คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

MACHINE LEARNING FOR TIME SERIES FORECASTING:

APPLICATION IN WATER SURFACE ANALYSIS OF MUN RIVER

Mister Anan Ungpansattawong

A Project Submitted in Partial Fulfillment of the Requirements

For the Degree of Bachelor of Science Program in Geology

Department of Geology, Faculty of Science, Chulalongkorn University

Academic Year 2020

หัวข้อโครงงาน	การเรียนรู้ของเครื่องจักรในการพยากรณ์อนุกรมเวลา การประยุกต์เพื่อวิเคราะห์พื้นผิวน้ำแม่น้ำมูล
โดย	นายอานันท์ อึ้งปัญสัตวงศ์
สาขาวิชา	ธรณีวิทยา
อาจารย์ที่ปรึกษาโครงงานหลัก	อาจารย์ ดร.พงศ์เทพ ทองแสง

วันที่ส่ง.....14 พ.ค. 2564.....

วันที่อนุมัติ.....7 พ.ค. 2564.....

.....

อาจารย์ที่ปรึกษาโครงการหลัก

(อาจารย์ ดร.พงศ์เทพ ทองแสง)

Project Title	MACHINE LEARNING FOR TIME SERIES FORECASTING: APPLICATION IN WATER SURFACE ANALYSIS OF MUN RIVER
By	Mister Anan Ungpansattawong
Field of study	Geology
Project Advisor	Pongthep Thongsang, Ph.D.

Submitted date.....14 May 2564.....

Approval date.....7 May 2564.....

.....

Project Advisor

(Pongthep Thongsang, Ph.D.)

อานันท์ อึ้งปัญสัตวงศ์ : การเรียนรู้ของเครื่องจักรในการพยากรณ์อนุกรมเวลา การประยุกต์เพื่อวิเคราะห์พื้นผิวน้ำแม่น้ำมูล. (MACHINE LEARNING FOR TIME SERIES FORECASTING: APPLICATION IN WATER SURFACE ANALYSIS OF MUN RIVER)

อ.ที่ปรึกาโครงงานหลัก : อาจารย์ ดร.พงศ์เทพ ทองแสง, 27 หน้า

บทคัดย่อ

ในปัจจุบันโลกกำลังประสบกับปัญหาภาวะโลกร้อน ประเทศไทยก็เป็นอีกหนึ่งประเทศที่ต้องประสบกับผลกระทบจากภาวะโลกร้อนเช่นกัน หนึ่งในปัญหาที่พบเห็นได้บ่อยครั้ง คือการที่เปลี่ยนแปลงของฤดูกาล ฤดูร้อนก็อุณหภูมิสูงมากกว่าที่เคยเป็น ฤดูฝนก็ไม่ตกตามฤดูกาล หรือเมื่อฝนตกเป็นเวลานานก็เกิดภาวะน้ำท่วมจากการระบายน้ำไม่ทัน ซึ่งในจุดนี้เองหากเราสามารถพยากรณ์ได้ว่าปริมาณน้ำที่จะพบในปีนี้มีปริมาณมากหรือน้อย คงสามารถทำให้เตรียมการรับมือได้อย่างถูกต้องและทันท่วงทีมากกว่า โดยความตั้งใจของงานงานศึกษานี้คือการสร้างเครื่องมือการพยากรณ์น้ำจากแหล่งข้อมูลที่ทุกคนสามารถเข้าถึงได้ จึงได้เลือกการศึกษาผ่านภาพถ่ายดาวเทียมขึ้นมาเป็นจุดเริ่มต้น โดยใช้หลักการที่ว่าหากปริมาณน้ำเพิ่มขึ้นและแม่น้ำไม่ได้มีลักษณะแคบลึกพื้นที่ของแม่น้ำในภาพถ่ายดาวเทียมย่อมต้องเพิ่มขึ้นด้วย จึงเกิดเป็นการศึกษาเพื่อสร้างเครื่องมือที่สามารถพยากรณ์พื้นที่ผิวแม่น้ำจากภาพถ่ายดาวเทียม โดยผลลัพธ์ที่ได้จากการศึกษานี้ก็พบว่าพื้นที่ผิวของแม่น้ำสามารถบอกถึงปริมาณน้ำที่เพิ่มขึ้นได้จริง และการพยากรณ์ก็ได้ผลลัพธ์ที่น่าพึงพอใจ

ภาควิชา ธรณีวิทยา ลายมือชื่อนิสิต .....อานันท์ อึ้งปัญสัตวงศ์.....

สาขาวิชา ธรณีวิทยา ลายมือชื่อ อ.ที่ปรึกษาหลัก.....พงศ์เทพ ทองแสง.....

ปีการศึกษา 2563

6032740723 : MAJOR GEOLOGY
KEYWORDS: MACHINE LEARNING / TIME SERIES FORECASTING
ANAN UNGPANSATTAWONG: MACHINE LEARNING FOR TIME SERIES FORECASTING: APPLICATION IN WATER SURFACE ANALYSIS OF MUN RIVER
ADVISOR: PONGTHEP THONGSANG, Ph.D., 27 pp.

Abstract

At present, the world is facing with a global warming problem. Thailand is another country that has suffered from global warming as well. One of the problems that are often seen change that is of the season's Summer temperature is higher than it used to be. The rainy season does not fall according to the season. Or when the rain became a long time, it was flooded from the drainage in time at this point, if we can predict whether the amount of water to be found this year is high or low. Therefore, the study through satellite imagery was chosen as the starting point. This study methodology is consisting of data collection, data processing 1, data processing 2, and prediction. Using the principle that if the amount of water increases and the river does not look deep, the area of the river in the satellite image will also have to increase. Therefore, the study was to create a tool that can predict the river's surface area from satellite imagery. The results obtained from this study revealed that the river's surface area could indicate an increase in water content and the forecasts yield satisfactory results.

Department: Geology
Field of study: Geology
Academic Year: 2020

Acknowledgement

First, I would like to express my appreciation to my advisor, Ph.D. Pongthep Thongsang of the Department of Geology, Faculty of Science, Chulalongkorn University for his support, suggestions, valuable ideas, coding knowledge, and report writing.

Second, I would like to thank the Department of Geology, Faculty of science, Chulalongkorn University, for supporting the study.

Third, I would like to thank Ms. Jaruphichaya Tadthai, Mr. Phakorn Intassingha and Ms. Wipaporn Nuttasin for the suggestions, knowledge sharing, and support my work.

Finally, I am thankful to my family for their support and my Geo'61 friends for being good friends and support me all along my university life.

List of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Rational

In the present day, Thailand had many effects from climate change by global warming, caused the rain not to fall in the season, which leads to drought and flood, as we can see in the news. Of course, climate change would have a lot of effects the agriculture. When there are too many water volumes, the water from the dam had to be released and flooded the agricultural area, damaged the products, and caused the loss to the farmer. If we could build a tool to predict the water volume in the future from information in the past, we might stop the losses.

To predict, we need a large amount of information (e.g., Satellite imagery). We must separate our study area from other details, and the last step is to obtain the water area to create a graph. The various processes above show the needing for time and accuracy. If we use human force in this section would cause lots of resources, both worker and time. So, we use the tool to help instead. To make the tool to calculate the process, we need to know the type of information and type of calculation. Classifying water area was required to be separated automatically by using the unsupervised classification technique, which is the k-mean method that achieved better accuracy in extracting the water-body segmentation in satellite images. (Yousefi et al., 2018; Duda, & Canty, 2010) For prediction, a linear graph could not be used for water data (Figure 1.1) because water data is time-series data. There are trends, seasonality, and cycle. (Hyndman & Athanasopoulos, 2013) Gaussian process models achieved a more realistic prediction variance in the case of noisy input. (Brahim-Belhouari & Bermak, 2004; Girard, 2002; Hu & Wang, 2015)

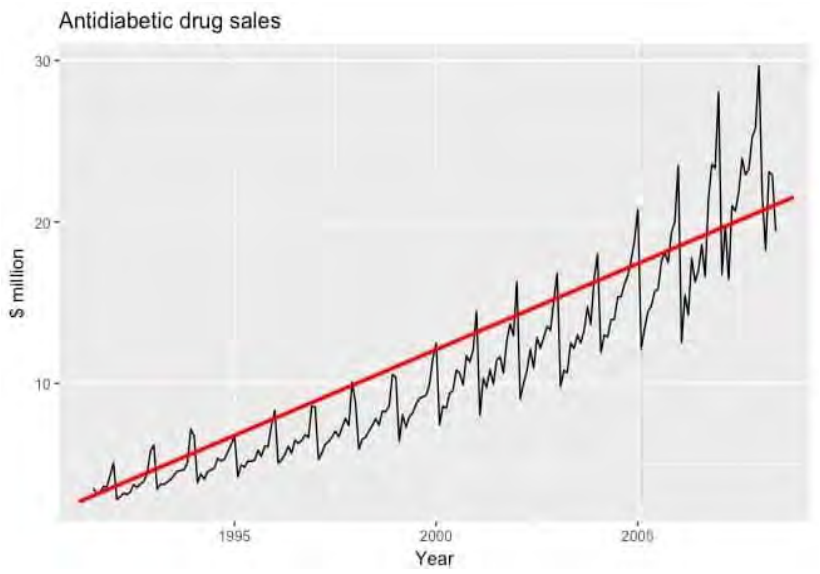

Figure 1.1 If we use linear regression to draw lines, it is not suitable for time series data with seasons.

(Hyndman & Athanasopoulos, 2013)

We are separating water from the study area by k-means clustering and building a predicting graph using gaussian process regression. We could fix the accuracy of the information and save time. Calculating by the tool instead of hiring experts could help people access easier.

To conclude, this project aims to study water areas in the study area through time and build the tool to predict future water areas helping the farmer plan their farming and reduce risk in the future.

1.2 Objective

1. To calculate water area from satellite imagery by automatic method.
2. To predict water area in the future from information in clause 1.

1.3 Study area

A part of Mun river in Ubon Ratchathani

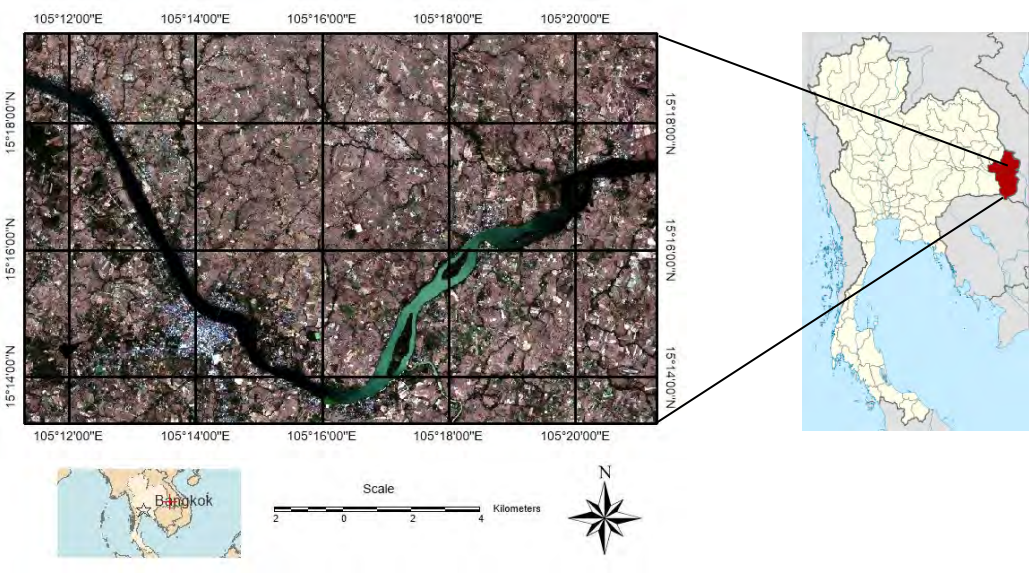

Figure 1.2 Study area

1.4 Scope of study

1. Prediction for water area in the study area.
2. The system will predict the amount of water up to 3 months in advance.

1.5 Methodology

Data collection

Download satellite images from GEE by downloading a total of 4 satellites: landsat8, sentinel 1, sentinel 2a and sentinel 2B.

Data processing

After obtaining the information, we have to separate what we want. Is the water area, where we will use k-mean to classify the water class from the image first and then use image segmentation to cleaning data.

predicting water surface bodies using gaussian process regression

Make a selection of the parameters that are appropriate for our information.

Task	Duration (From October 2020 to May 2021)							
	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
1 Literature reviews	—	—						
2 Choose the study area	—							
3 Data collection		—	—					
4 Data processing				—	—			
5 Predicting water surface bodies using gaussian process regression						—	—	
6 Report writing								—

Table 1.1 Work plan schedule

Chapter 2

Literature Reviews

2.1 Mun river

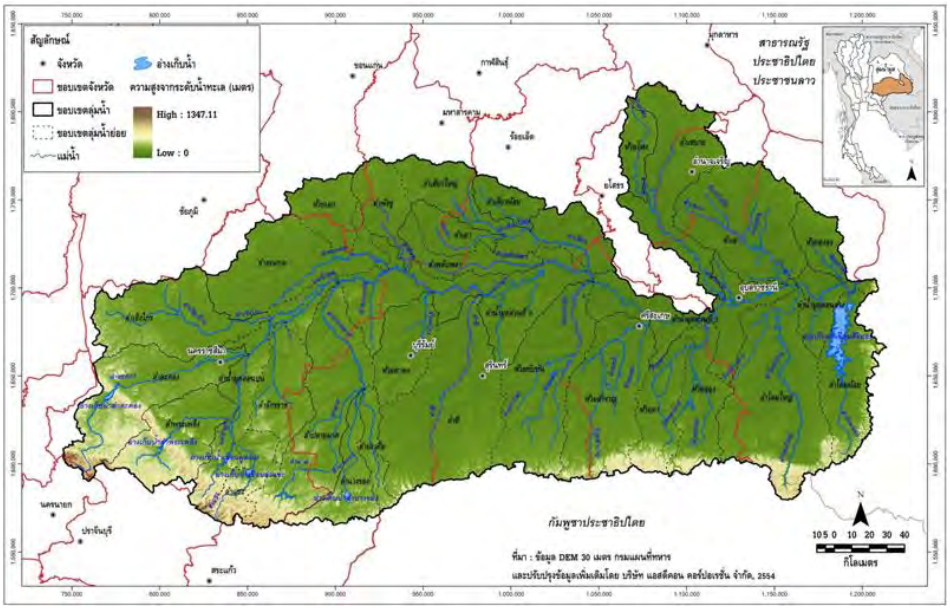

Figure 2.1 Mun River basin

(http://mekhala.dwr.go.th)

Mun River basin is located in the Northeast. It has an area of approximately 71,060 square kilometers, located between latitude 14 ° 7′ north to latitude at 16 ° 20′ north and between longitude 101 ° 17′ east to longitude at 105 ° 40. ′ East Most areas cover 10 provinces, including 118 districts, 19 sub-districts in the lower Isan region. And parts of the central northeastern region with the following territories.

North next to the Chi River and the Isan Mekong Basin.

South next to the Prachin Buri River Basin, Tonle Sap River Basin, and Cambodia Democracy.

East next to the Mekong River and Lao People's Democratic Republic.

West adjacent to the Pasak River Basin and Bang Pakong River Basin.

(Water crisis prevention center, 2005)

2.2 Satellite image

Figure 2.2 Satellite image was taken by Sentinel 2B

Due to our study area is relatively large so that we collect satellite imagery from Landsat-8, Sentinel-1, Sentinel-2A, and Sentinel-2B to cover the entire region

2.2.1 Landsat-8

Landsat 8 is an American Earth observation satellite launched on 11 February 2013. It is the eighth satellite in the Landsat network, and the seventh to successfully reach orbit. It was formerly known as the Landsat Data Continuity Mission (LDCM), and it is a joint NASA-US Geological Survey project (USGS). NASA's Goddard Space Flight Center in Greenbelt, Maryland, designed the launch vehicle, mission systems engineering, and acquisition, while the USGS developed the ground systems and would oversee mission operations. It consists of the Thermal Infrared Sensor (TIRS) and the Operational Land Imager (OLI) camera, both of which can be used to study Earth surface temperature and global warming. (Li et al., 2021)

Spectral Band	Wavelength	Resolution	Solar Irradiance
Band 1 - Coastal / Aerosol	0.433 – 0.453 μm	30 m	2031 W/(m^{2}μm)
Band 2 - Blue	0.450 – 0.515 μm	30 m	1925 W/(m^{2}μm)
Band 3 - Green	0.525 – 0.600 μm	30 m	1826 W/(m^{2}μm)
Band 4 - Red	0.630 – 0.680 μm	30 m	1574 W/(m^{2}μm)
Band 5 - Near Infrared	0.845 – 0.885 μm	30 m	955 W/(m^{2}μm)
Band 6 - Short Wavelength Infrared	1.560 – 1.660 μm	30 m	242 W/(m^{2}μm)
Band 7 - Short Wavelength Infrared	2.100 – 2.300 μm	30 m	82.5 W/(m^{2}μm)
Band 8 - Panchromatic	0.500 – 0.680 μm	15 m	1739 W/(m^{2}μm)
Band 9 - Cirrus	1.360 – 1.390 μm	30 m	361 W/(m^{2}μm)
Band 10 - Long Wavelength Infrared	10.30 – 11.30 μm	100 m	Band 10 - Long Wavelength Infrared
Band 11 - Long Wavelength Infrared	11.50 – 12.50 μm	100 m	Band 11 - Long Wavelength Infrared

Table 2.1 Spectral bands for the Landsat 8 sensors (Irons et al., 2012)

2.2.2 Sentinel-1

The Sentinel-1 project is a joint European Commission (EC) and European Space Agency (ESA) initiative called the European Radar Observatory for Copernicus (ESA). Copernicus is a European initiative for the introduction of environmental and security-related information services. It is based on data from Earth Observation satellites as well as information from the ground. (European Space Agency, 2014)

Name	Description	Min*	Max*	Resolution	Units	Wavelength
HH	Single co-polarization, horizontal transmit/horizontal receive	-50	1	10 m		5.405GHz
HV	Dual-band cross-polarization, horizontal transmit/vertical receive	-50	1	10 m		5.405GHz
VV	Single co-polarization, vertical transmit/vertical receive	-50	1	10 m		5.405GHz
VH	Dual-band cross-polarization, vertical transmit/horizontal receive	-50	1	10 m		5.405GHz
angle	Approximate viewing incidence angle	0	90	-1 m	Degrees	

* = Values are estimated

Table 2.2 Spectral bands for the Sentinel-1 sensors (European Space Agency, 2014)

2.2.3 Sentinel-2A

Sentinel-2A is an optical imaging satellite launched by the European Space Agency in 2015. It is the first Sentinel-2 satellite launched as part of the Copernicus Program of the European Space Agency. With 13 spectral bands, the satellite carries a wide-swath high-resolution multispectral imager. It will collect data on the ground to help services including forest monitoring, land cover change identification, and natural disaster management. (Justice, 2015)

2.2.4 Sentinel-2B

Sentinel-2B is an optical imaging satellite launched by the European Space Agency on March 7, 2017. It is the second Sentinel-2 satellite launched as part of the Copernicus Program by the European Space Agency, and its orbit will be 180 degrees apart from Sentinel-2A. With 13 spectral bands, the satellite carries a wide-swath high-resolution multispectral imager. It will provide data for agriculture and forestry, among other things, enabling crop yield predictions. (European Space Agency, 2017)

Sentinel-2 bands	Sentinel-2A		Sentinel-2B		
	Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	Spatial resolution (m)
Band 1 – Coastal aerosol	442.7	21	442.2	21	60
Band 2 – Blue	492.4	66	492.1	66	10
Band 3 – Green	559.8	36	559.0	36	10
Band 4 – Red	664.6	31	664.9	31	10
Band 5 – Vegetation red edge	704.1	15	703.8	16	20
Band 6 – Vegetation red edge	740.5	15	739.1	15	20
Band 7 – Vegetation red edge	782.8	20	779.7	20	20
Band 8 – NIR	832.8	106	832.9	106	10
Band 8A – Narrow NIR	864.7	21	864.0	22	20
Band 9 – Water vapour	945.1	20	943.2	21	60
Band 10 – SWIR – Cirrus	1373.5	31	1376.9	30	60
Band 11 – SWIR	1613.7	91	1610.4	94	20
Band 12 – SWIR	2202.4	175	2185.7	185	20

Table 2.3 Spectral bands for the Sentinel-2 sensors (European Space Agency, 2017)

2.3 K-means clustering

K-means clustering is one of the methods of unsupervised classification. That we choose to use in this work because it can extract environmental information well (Usman, 2013). With the following work processes.

Step 0: Import input data. We can notice that there are five clusters so how ML classify the dataset.

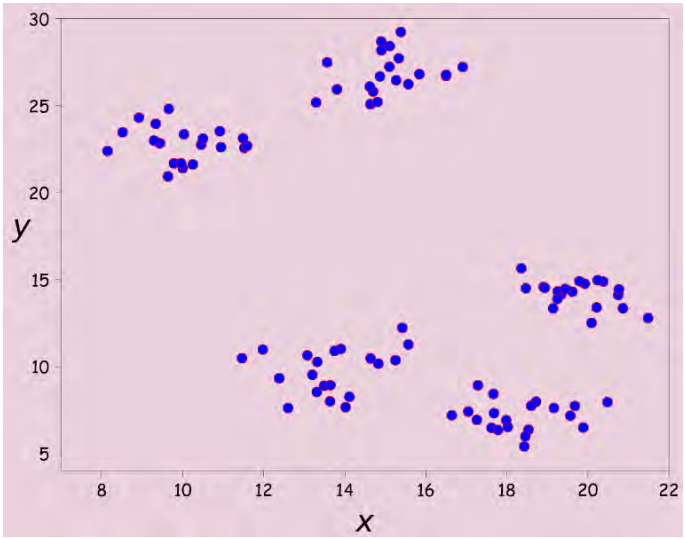

Figure 2.3 K-means clustering step 0

Step 1: Create Voronoi diagramme (the color segments) based on random centriods. We notice that some observed data are located more than one Voronoi segment.

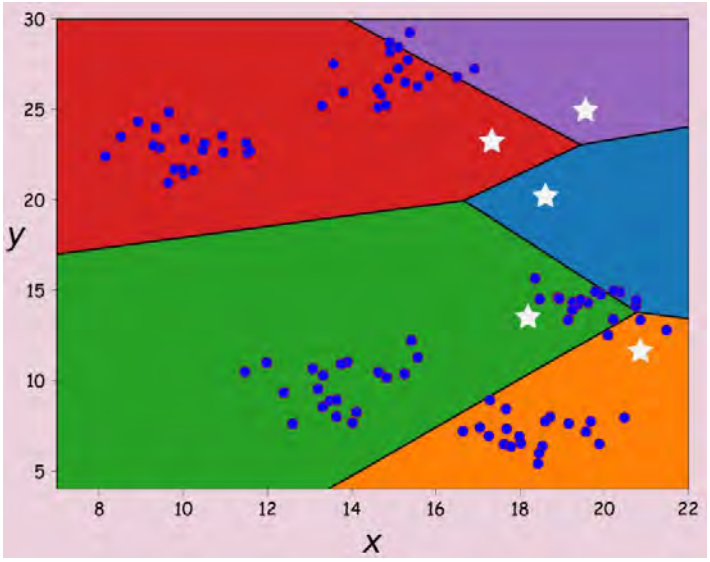

Figure 2.4 K-means clustering step 1

Step 2: Calculate the Euclidean distance. In each Voronoi diagram or one segment, each pair of centroid and data point will measure the Euclidean distance. Calculating the mean of Euclidean distances and the mean distance will be used to update the centroid location.

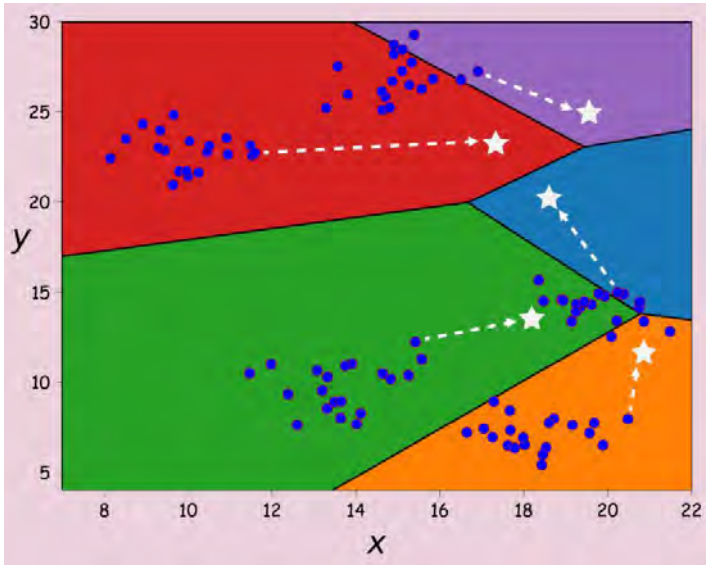

Figure 2.5 K-means clustering step 2

Step 3: Update centroid positions. These migrations are based on the means of Euclidean distances in each Voronoi segment.

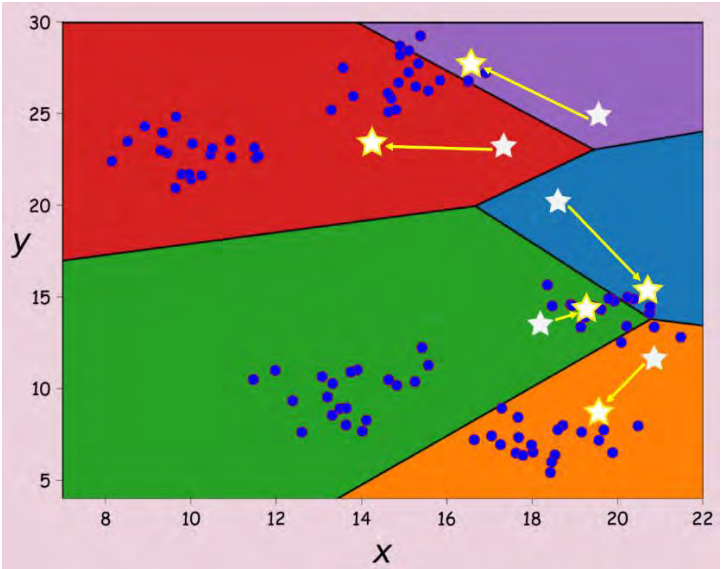

Figure 2.6 K-means clustering step 3

Step 4: Update the Voronoi diagram based on new iterated centroids. We recognize that the Voronoi segments are updated with the following of the migrated centroids. We need some iterations to fit the observed data into exactly one Voronoi segment: step 1 to step 4 will be iterated until it reaches the convergence.

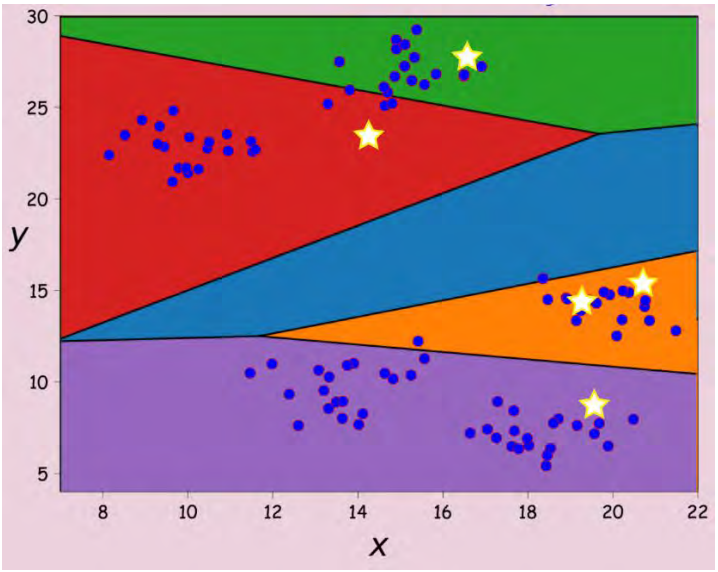

Figure 2.7 K-means clustering step 4

Step 5: Reach the convergence. In each segment, the centroid will migrate to the mean of Euclidean distance. The convergence here refers to the mean of the euclidean distance does not change.

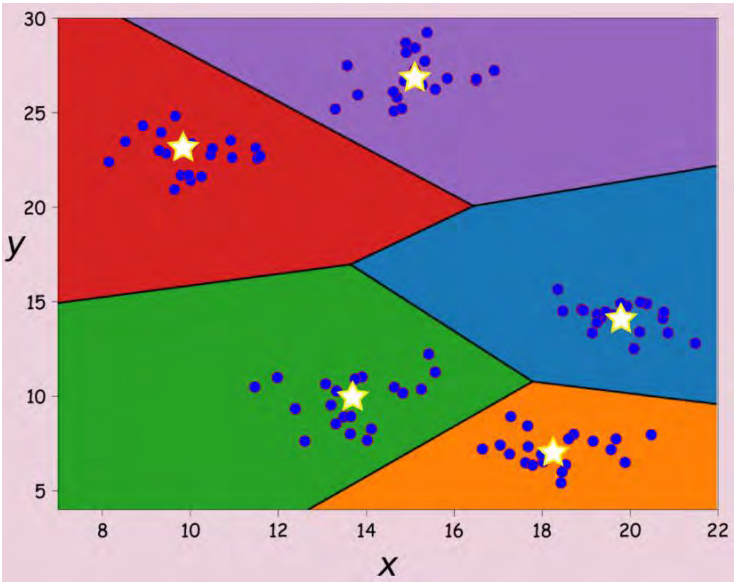

Figure 2.8 K-means clustering step 5

2.4 Image segmentation

Image segmentation is the method of partitioning a digital image into several segments in digital image processing and computer vision (sets of pixels, also known as image objects). Segmentation aims to make an image more meaningful and easier to interpret by simplifying and/or changing its representation. Objects and boundaries (lines, curves, etc.) in images are usually located using image segmentation. Image segmentation, to put it another way, is the method of assigning a label to each pixel in an image such that pixels with the same label have similar characteristics. (Thongsang et al., 2020; Shapiro & Stockman, 2001)

Image segmentation produces either a set of segments that cover the entire image or a set of contours derived from the image (see edge detection). In terms of some characteristic or computed property, such as color, intensity, or texture, each pixel in a region is identical. When it comes to the same trait, adjacent regions vary significantly (s). When image segmentation is applied to a stack of images, as is common in medical imaging, the resulting contours can be used to create 3D reconstructions using interpolation algorithms like marching cubes. (Shapiro & Stockman, 2001; Zachow et al., 2007)

2.5 Gaussian process regression (GPR)

Gaussian process regression (GPR) is a Bayesian approach to the regression of waves in the field of machine learning. GPR can provide various advantages, work well on small datasets and measure predictive uncertainty. (Sit, 2019)

Holding all the training data and assume that the uncertainty or noise would adopt a multivariate Gaussian distribution when predicting new data points from old ones. Making predictions from the training examples, resulted from the relationship provided by that distribution. The Gaussian distribution, aka the normal distribution, has a bell-shaped curve, with the mean being the most likely point and the likelihood gradually decreasing as steping away from it. Correlations between multiple variables can be defined using the multivariate Gaussian. The Gaussian distribution occurs naturally in many situations and is the default noise model in many machine learning applications. (Winterstein, 2016)

The probability distribution from which the measures are taken has a significant impact on the direction that the point takes. The kernel is a Gaussian normal distribution. Fundamentally altering the random direction that the point walks along by modifying the kernel. Here are some examples of time walks using three different kernel functions (Bilogur, 2018)

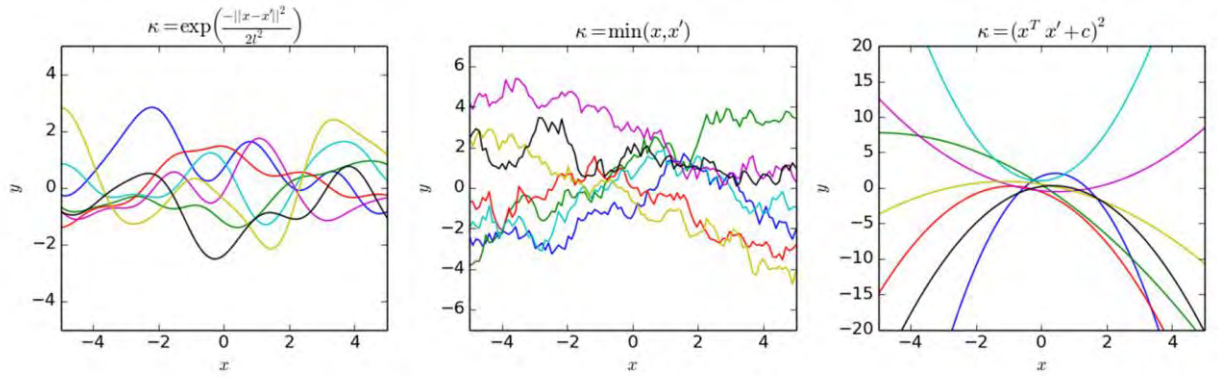

Figure 2.9 GPR with different kernel functions

(www.kaggle.com)

2.6 Tools

2.6.1 Google Earth Engine

Earth Engine is a forum for academic, non-profit, industry, and government users to conduct scientific research and visualization of geospatial datasets. Earth Engine hosts satellite imagery and stores it in a public data repository of historical earth photos dating back over four decades. The photos are then made available for global-scale data mining after being ingested regularly. Earth Engine also offers APIs and other resources that make it possible to analyze massive datasets.

This is not the same as Google Earth. By interacting with a virtual globe, Google Earth allows you to fly, discover, and learn about the planet. Satellite imagery, charts, terrain, 3D structures, and much more are all available. Earth Engine, on the other hand, is a geospatial knowledge analysis tool. Among the many potential assessments, you can look at forest and water coverage, land use reform, and the health of agricultural fields, to name a few. Although the two tools use some of the same data, Earth Engine only has access to a portion of Google Earth's imagery and data.

As part of the Google Cloud public data initiative, the Earth Engine team collaborated closely with Google Cloud to add the Landsat and Sentinel-2 collections to Google Cloud Storage. Accessing data directly from Cloud providers like Google Compute Engine or Google Cloud Machine Learning is much simpler and more effective thanks to the Google Cloud collections. The Earth Engine Code Editor and API do not have access to these Cloud collections; instead, they use the Earth Engine data catalog directly. (Google, 2017)

2.6.2 Spyder

Spyder is a free and open-source scientific environment developed by and for scientists, developers, and data analysts, and written in Python for Python. It combines a robust programming tool's advanced editing, review, debugging, and profiling capability with a science package's data discovery, immersive execution, deep inspection, and stunning visualization capabilities. (Spyder, 2020)

2.6.3 Google Colaboratory

Google Research's Colaboratory, or "Colab" for short, is a product. Colab is a web-based Python editor that allows everyone to write and run arbitrary Python code. It's particularly useful for machine learning, data analysis, and education. Colab is a hosted Jupyter notebook service that doesn't need any configuration and offers free access to computing tools, including GPUs. (Google, 2014)

2.6.4 Amazon Web Services

AWS is a cloud networking company. Storage on the internet, as well as management over EC2 servers and client computers When compared to on-site storage, it is less expensive and more secure.

Chapter 3
Methodology

This study is consisting of data collection, data processing 1, data processing 2, and prediction. The details of each method are below.

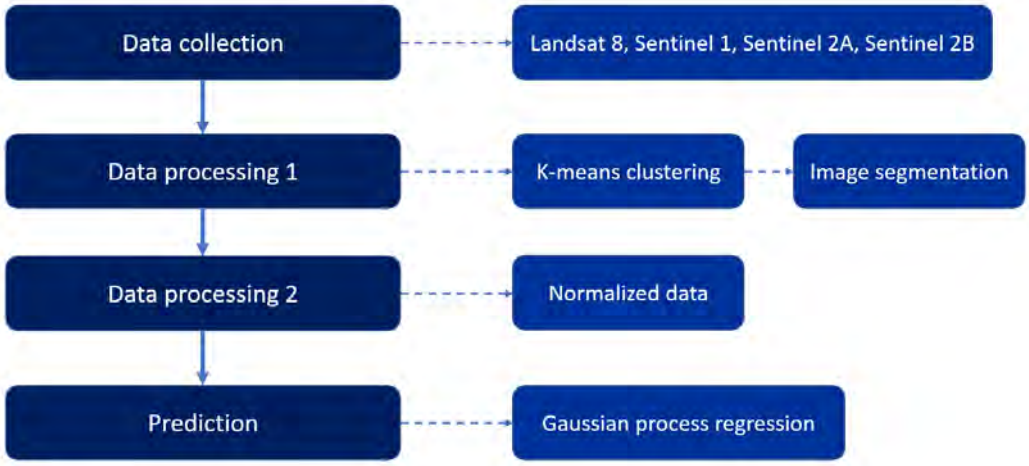

Figure 3.1 Four steps in methodology

3.1 Data collection

The satellite imagery is downloaded from the Google Earth Engine (GEE) as it has a device that allows us to download specifically selected area. The needed satellite images are not photos of one day, but photos of an entire month that comes together to form a single photograph that can represent the whole month by GEE supporting tools.

The satellite imagery including images from Landsat-8, Sentinel-1, Sentinel-2A, and Sentinel-2B satellite, which may not have been able to obtain the complete monthly required satellite images. Since most of the satellites we use have a passive electronic system, passive sensors detect reflected electromagnetic radiation from a source such as the water surface that reflected the sun. If the cloud is low, GEE has tools to exclude it. But this tool will not work in a rainy season with lots of clouds condition.

The rain satellite images results are from Sentinel-1 satellite with electronic system, active electronic sensors, active sensors both emit a pulse of energy and detect the reflected energy.

This project, the study period is in 2018 to 2020, for 36 months, counting January 2018 as the 1st month until December 2020, the 36th month.

3.2 Data processing (k-means clustering and image segmentation)

There are two steps in data processing, k-means clustering and image segmentation, which are sequential. In other words, in one picture, k-means clustering must find the class of water first, then use image segmentation to select the main river only.

After receiving satellite images of each month, to increase the amount of data, the band of each satellite image was separated again. Each satellite uses bands as follows.

Landsat-8 used bands: 5 and 6

Sentinel-1 used bands: HH and VV

Sentinel-2A used bands: 6, 7, 8, 8A, 9 and 11

Sentinel-2B used bands: 6, 7, 8, 8A, 9 and 11

To separate the water classes from the satellite images, using k-means clustering to assist in our clustering first. The suitable class number for our study area is 3 classes: agriculture, urban and water.

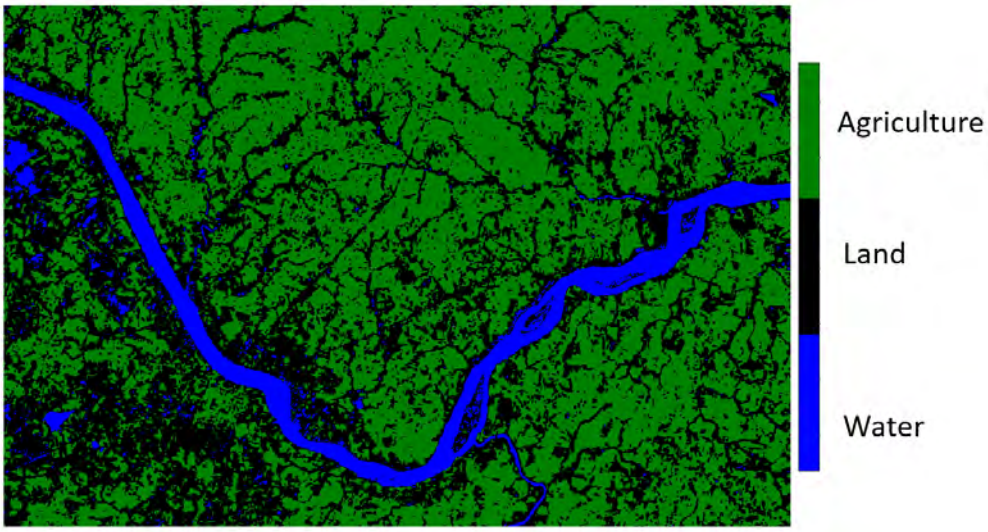

Figure 3.2 k-means classification

The next step, after splitting image classes are done, counting the image to used image segmentation, with the principle of eliminating the non-continuity data, leaving only the continuity data. The Mun river was segmented.

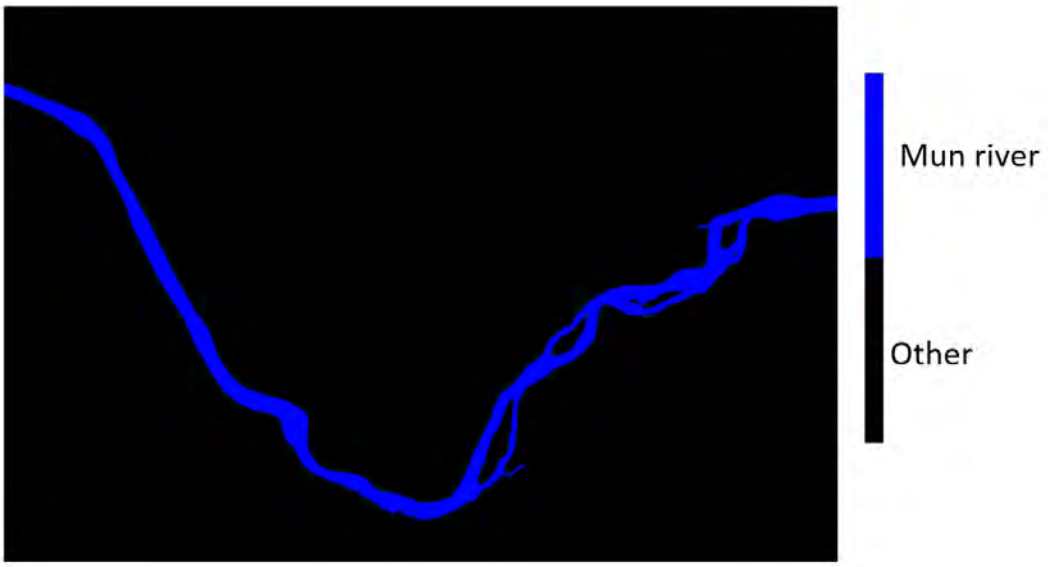

Figure 3.3 image segmentation

At this stage, inspecting the damaged image are done. If the image segmentation has been obtained other data than the main river (e.g., clouds), this information must be omitted.

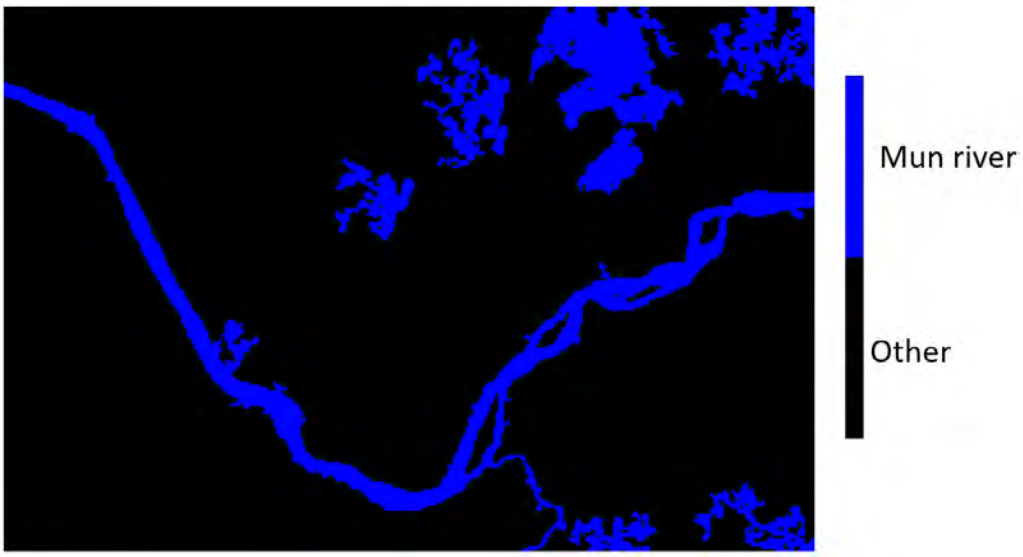

Figure 3.4 image segmentation (not taken)

By doing the image segmentation, counting the pixels of the Mun river from the pixel count of the remaining water class is available.

3.3 Data processing 2 (Normalized data)

After collected all the information, the information will show as figure 3.5

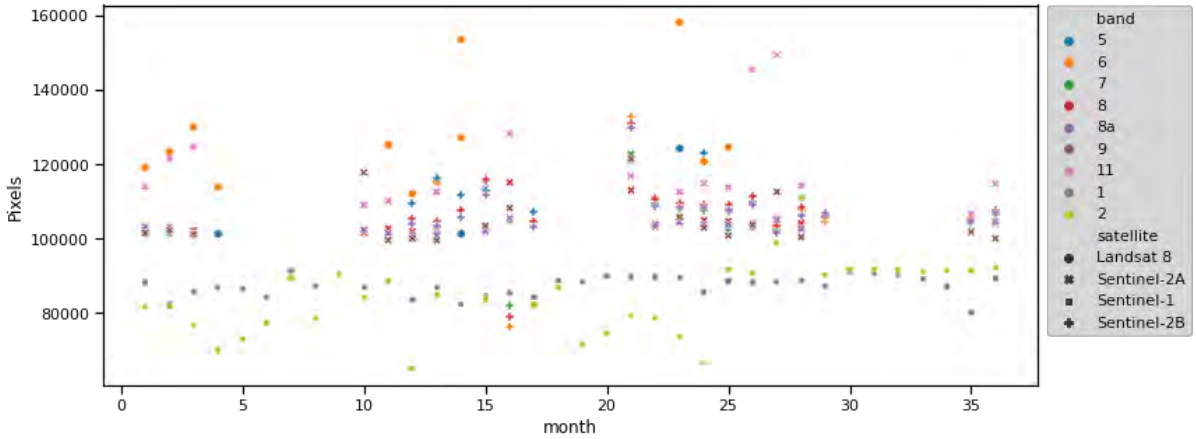

Figure 3.5 Relationship between water area and month

In this data set, there is a cohesion of the data of each satellite. Therefore, normalize the data of each band and each satellite need to be done. In the range of 0 to 1 equal to every band using the formula to normalize the data.

$$x' = \left[\left[\frac{x - \min(x)}{\max(x) - \min(x)}\right] * (highest_{value} - lowest_{value})\right] + lowest_{value}$$

Set $highest_{value} = 1$, $lowest_{value} = 0$

After the normalized is complete, the information will appear as figure ... .

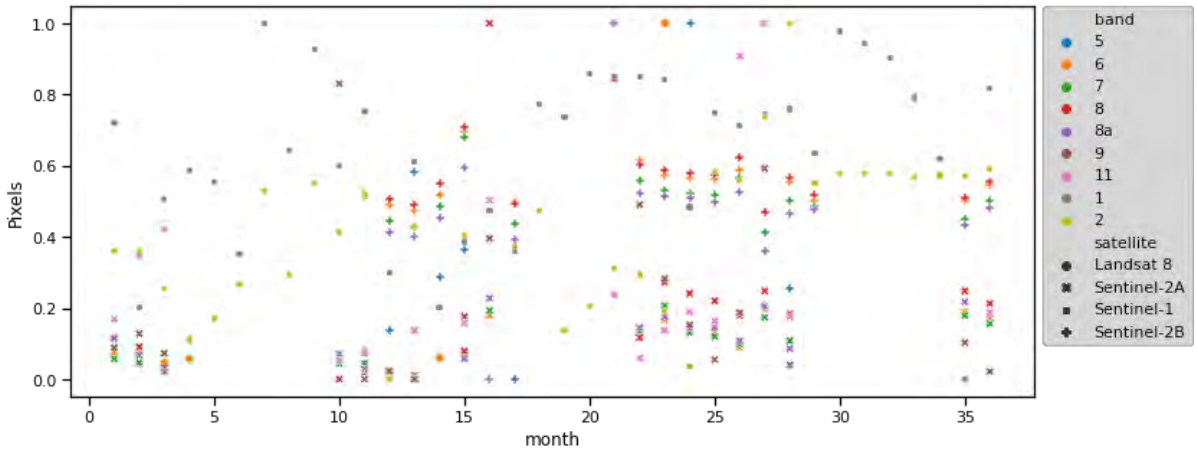

Figure 3.6 Relationship between water area and month after normalized

Once obtained the data, the next step is to make a prediction model, which the validity of the model is considered, by using the value Root Mean Square Error (RMSE) with the equation as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Error)^2}$$

Set Error = actual – predicted

Therefore, if the model is very good, it will have an RMSE value approaching 0, since the RMSE value can be difficult to compare, convert the RMSE to percentage for understandable data. 0% means the model is good, there are no errors. But if the percentage is 100%, the model is invalid with too many mistakes.

3.4 Prediction

Before using the data in GPR another prediction model is tried.

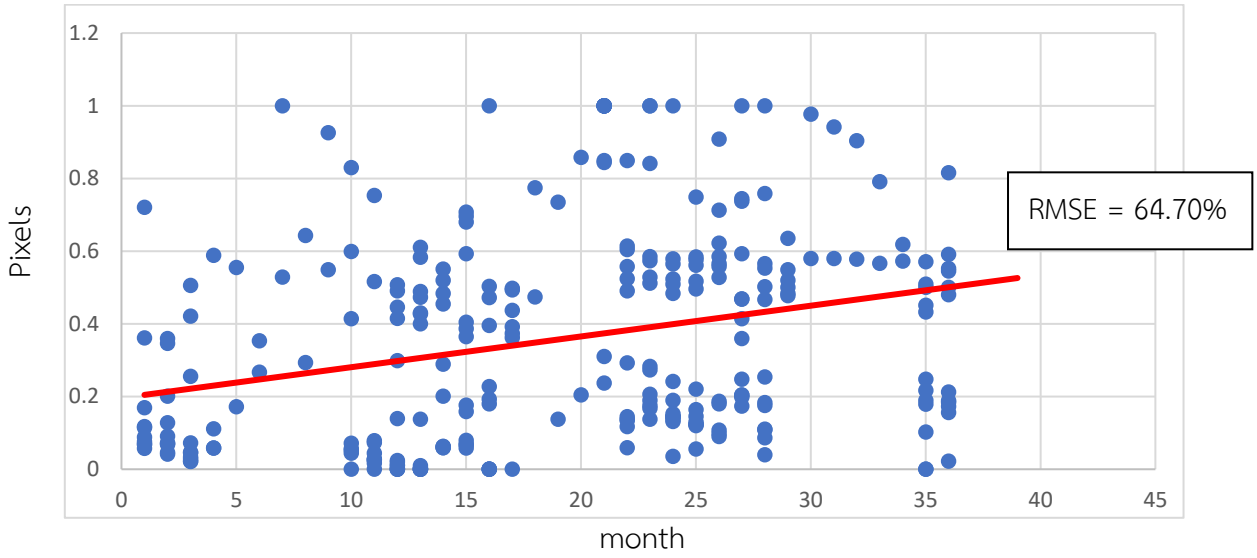

Figure 3.7 Prediction model with linear regression

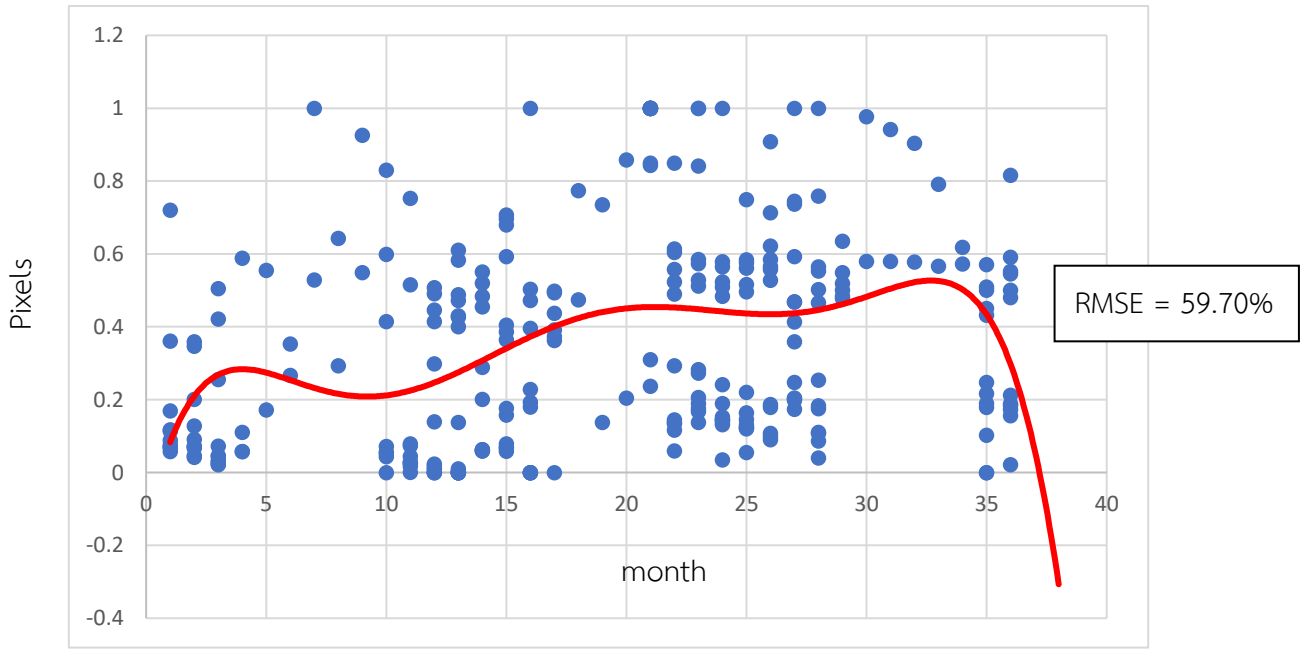

Figure 3.8 Prediction model with polynomial regression

From the showed results, it was found out to be unreliable as the high RMSE values. Another interesting point is both prediction models provide different future trends, the future linear prediction tends to be higher while the future polynomial prediction tends to be down.

This project is done by using GPR from PyMC3 package (Salvatier et al., 2016). The principle of operation is that GPR will create a function based on the collected information. The generated function will try to pass close to the point and it will be smoothed by the kernel function, the exponential kernel function for this study.

In this model, the GPR created 1,000 lines, where area with a lot of overlapping lines appear in dark red.

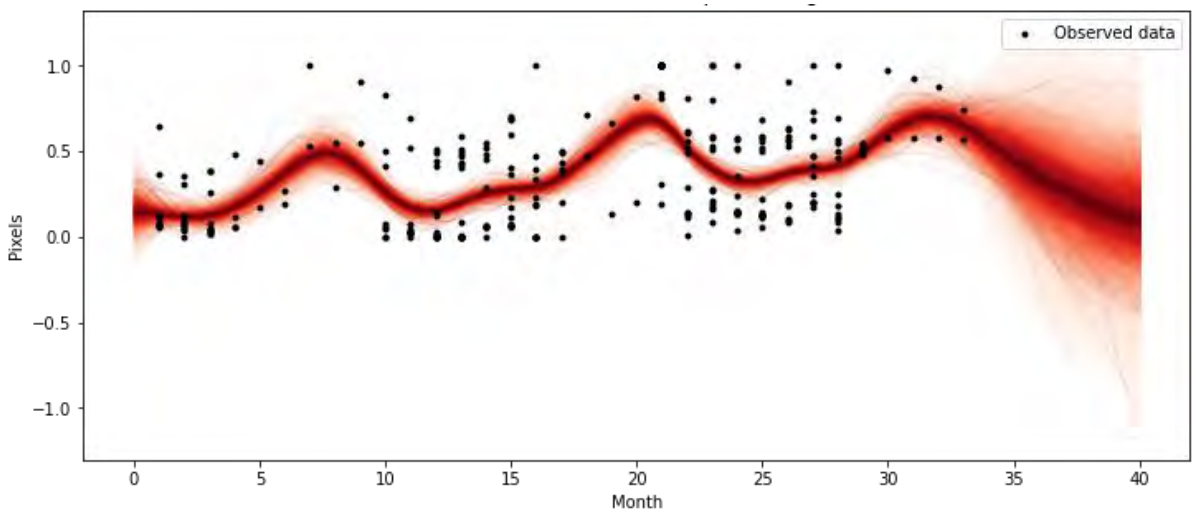

Figure 3.9 Prediction model with GPR

Chapter 4

Results

When randomized lines enough in the GPR, the resulting lines are arranged in a normal distribution, and the mean result becomes the line with the lowest RMSE of all lines. So, this mean line is chosen to represent all 1000 lines.

The result in the figure 4.1, showing the mean lines, it also shows the width of 1 SD or the confidence bar. The narrowband means that the model is high confidential. On the other hand, if the bar is very wide, the model is less confidential.

From the figure, the input point is only at x = 33, so after that period, there is no data the confidence bar will get bigger.

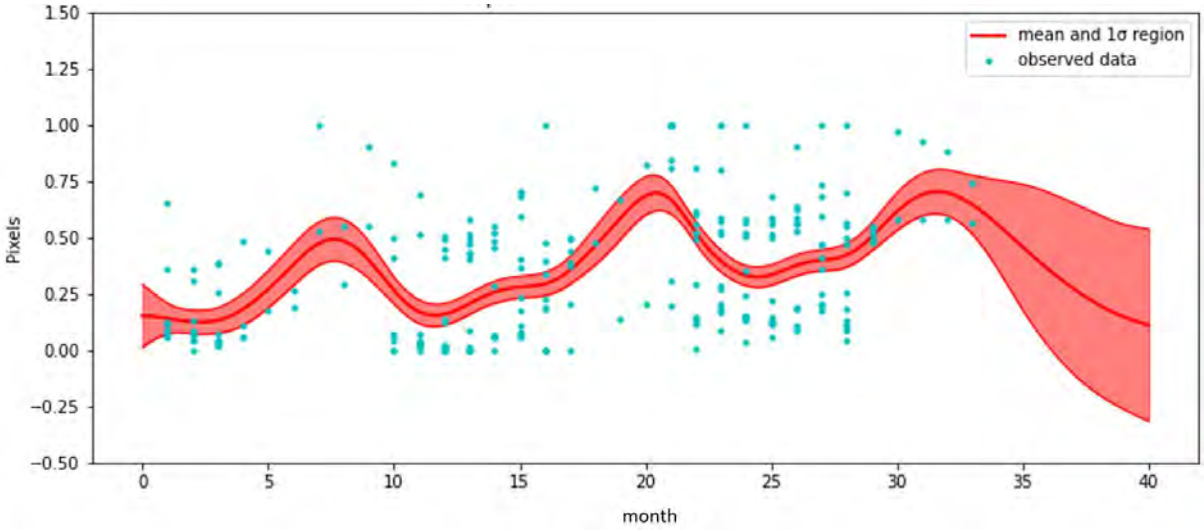

figure 4.1 Predictive mean and 1**σ** interval

Cheaper 5

Discussion and Conclusion

Discussion

The primary objective of this work is to forecasting surface water bodies from given satellite imagery. In this test, we exclude part of the dataset, month 30th - 33rd, from the training data. See figure 5.1; the blue dots indicate observed data points, and the black dots indicate the forecasting data. Without seeing that part of the data, GPR will predict models that might relate and unrelate to observed data points. Here, we require several iterations for tuning parameters. The best parameters will yield the best fit of the observed data, and the best fit model should fit the unseen observed data.

The result (figure 5.1) shown trained model has RMSE 26.20%, and the predicted model has RMSE 28.60%. The trained model demonstrates a convincing result, which can use to apply in observation data. Note that we use the mean value from the generated GPR model to measure the RMSE.

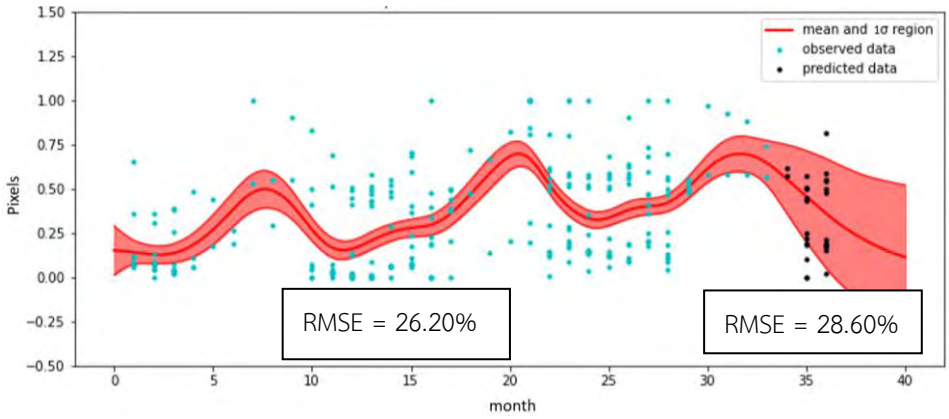

Figure 5.1 Predictive mean and 1**σ** interval show predictive data.

To compare the created model with the real field data. The daily water level graph of the Mun River from the Royal Irrigation Department was compared with the created model (2018-2019 data).

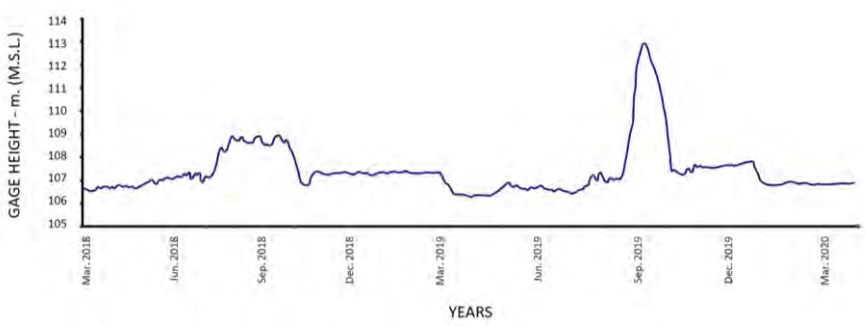

Figure 5.2 Daily mean gage height year 2018-2019

By comparison, it can be found that the data of the Royal Irrigation Department, the water level during the rainy season of 2019 is higher than that of 2018, while our model (figure 5.2), the height of the rainy season in 2018 is lower than the height during the rainy season of 2019 as well.

In figure 5.3 the altitudes of the year 2019 and year 2020 are similar. Although the water height data from the water from the Royal Irrigation Department was not public yet, it can be compared from the news in 2019 that Ubon Ratchathani province had a big flood and in 2020 also. Allowing us to tell that the information we get from satellite imagery is highly accurate.

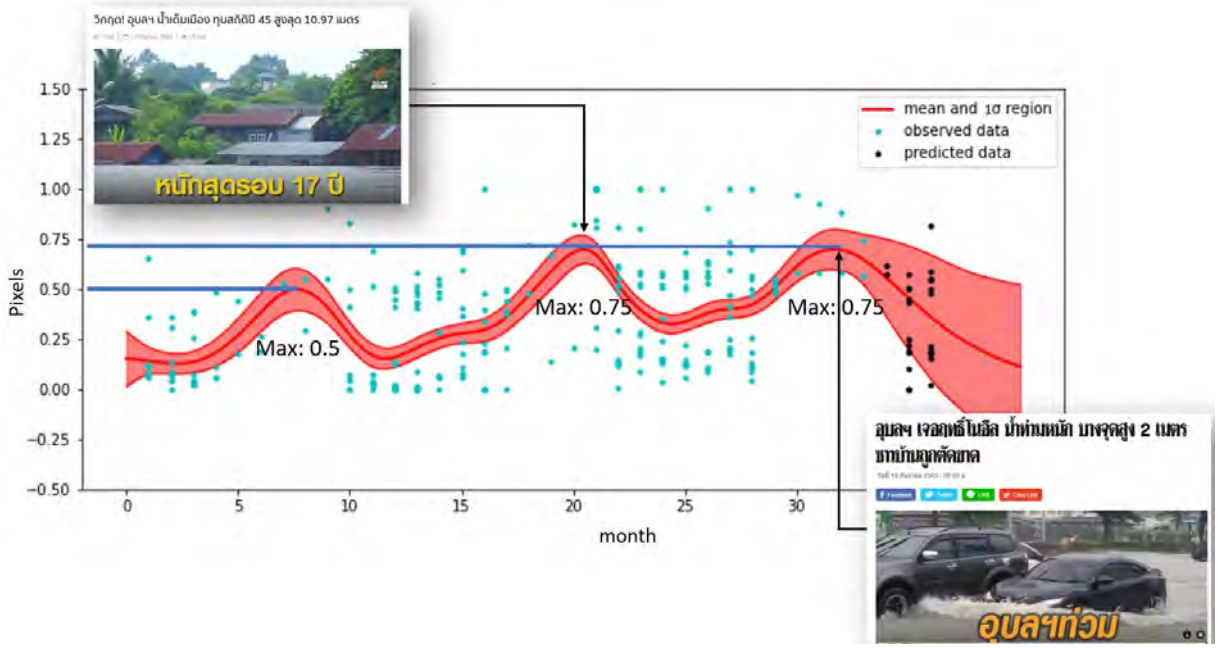

Figure 5.3 Predictive mean and 1σ interval relate with news

Conclusion

1. Gaussian process regression can create practical prediction model according to RMSE value.

2.Gaussian process regression can relate with daily mean gage height graph from royal irrigation department of Thailand.

References

Bilogur, A., 2018. Gaussian process regression and classification. [online] Available at: <https://www.kaggle.com/residentmario/gaussian-process-regression-and-classification> [Accessed 6 February 2021]

Brahim-Belhouari, S., Bermak, A., 2004. Gaussian process for nonstationary time series prediction. ComputationalStatistics & Data Analysis 47 (2004), pp. 705 – 712.

Duda, T., Canty, M., 2010. Unsupervised classification of satellite imagery: Choosing a good algorithm. International Journal of Remote Sensing, Volume 23, pp. 2193-2212.

European Space Agency, 2014. Sentinel-1 Overview. [online] Available at: <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/overview> [Accessed 2 February 2021]

European Space Agency, 2017. Sentinel-2B Launch Preparations Off to a Flying Start. [online] Available at: <http://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Sentinel-2B_launch_preparations_off_to_a_flying_start> [Accessed 2 February 2021]

European Space Agency, 2017. MultiSpectral Instrument (MSI). [online] Available at: <https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/msi-instrument> [Accessed 2 February 2021]

Girard, A., Rasmussen, C.E., Candela, J.Q., Murray-Smith, R., 2002. Gaussian Process Priors With Uncertain Inputs Application to Multiple-Step Ahead Time Series Forecasting. Advances in Neural Information Processing Systems 15 (NIPS 2002).

Google, 2014. Colaboratory. [online] Available at: <https://research.google.com/colaboratory/faq.html> [Accessed 10 February 2021]

Google, 2017. Google Earth Engine. [online] Available at: <https://earthengine.google.com/faq/> [Accessed 10 February 2021]

Hu, J., Wang, J., 2015. Short-term wind speed prediction using empirical wavelet transform and Gaussian process regression. Energy Volume 93, Part 2, 15 December 2015, pp. 1456-1466.

Hyndman, R.J., Athanasopoulos, G., 2013. Forecasting: Principles and Practice.

Irons, J.R., Dwyer, J.L., Barsic, J.A., 2012. The next Landsat satellite: The Landsat Data Continuity Mission. Remote Sensing of Environment 122, Pages 11-21

Justice, A., 2015. Sentinel-2A: Satellite blasts off to provide new, improved view of Earth. [online] Available at: <https://www.ibtimes.co.uk/sentinel-2a-satellite-blasts-off-provide-new-improved-view-earth-1507541> [Accessed 2 February 2021]

Li, R.Y.M., Chau, K.W., Li, H.C.Y., Zeng, Y., 2021. Remote Sensing, Heat Island Effect and Housing Price Prediction via AutoML. Advances in Intelligent Systems and Computing book series, volume 1213, pp 113-118

Salvatier, J., Wiecki, T.V., Fonnesbeck, C., 2016. Probabilistic programming in Python using PyMC3. PeerJ Computer Science 2: e55

Shapiro, L.G., Stockman, G.C., 2001. Computer Vision, pp 279–325

Sit, H., 2019. Quick Start to Gaussian Process Regression. [online] Available at: <https https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319> [Accessed 6 February 2021]

Spyder, 2020. Overview. [online] Available at: <https://www.spyder-ide.org> [Accessed 10 February 2021]

Thongsang, P., Hu, H., Zhou, H.W., Lau, A., 2020. Imaging Enhancement in Angle-Domain Common-Image-Gathers Using the Connected-Component Labeling Method. Pure and Applied Geophysics volume 177, pages4897–4912

Usman, B., 2013. Satellite Imagery Land Cover Classification using K-Means Clustering Algorithm Computer Vision for Environmental Information Extraction. Elixir Comp. Sci. & Engg. 63 (2013) 18671-18675

Water crisis prevention center, 2005. Mun River Basin [online] Available at: <http://mekhala.dwr.go.th/knowledge-basin-mun.php> [Accessed 3 January 2021]

Winterstein, D., 2016. A Simple Intro to Gaussian Processes. [online] Available at: <http://platypusinnovation.blogspot.com/2016/05/a-simple-intro-to-gaussian-processes.html> [Accessed 6 February 2021]

Yousefi, P., Jalab, H.A., Ibrahim, R.W., Noor, N.F.M., Ayub, M.N., Gani, A., 2018. Water-Body Segmentation in Satellite Imagery Applying Modified Kernel K-Means. Vol 31 No 2 (2018): Malaysian Journal of Computer

Zachow, S., Zilske, M., Hege, H.C., 2007. 3D reconstruction of individual anatomy from medical image data: Segmentation and geometry processing. ZIB-Report (07-41)