National Technical University of Athens

School of Electrical and Computer Engineering

MSc in Data Science and Machine Learning

Paddy Rice Mapping based on Multi-temporal Sentinel-1 and Sentinel-2 Data in a High Performance Data Analytics Environment

Master Thesis

KOUKOS ALKIVIAIDS-MARIOS

Supervisor: Karantzalos Konstantinos
Associate Professor NTUA

Athens, November 2020

National Technical University of Athens
School of Electrical and Computer Engineering
MSc in Data Science and Machine Learning

Paddy Rice Mapping based on Multi-temporal Sentinel-1 and Sentinel-2 Data in a High Performance Data Analytics Environment

Master Thesis

KOUKOS ALKIVIAIDS-MARIOS

Supervisor: Karantzalos Konstantinos
Associate Professor NTUA

Approved by the three-member examination committee on 12th November 2020.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
K. Karantzalos	C. Kontoes	G. Goumas
Assoc. Professor NTUA	Reearch Director NOA	Assistant Professor NTUA

Athens, November 2020

(Signature)

.....

Koukos Alkiviadis-Marios

National Technical University of Athens
School of Electrical and Computer Engineering
MSc in Data Science and Machine Learning

Abstract

Over the last years, the continuous increase of global population, together with the climate change, is expected to affect the food sector significantly. Rice is a main source of nutrition for more than half of world's population and it contributes significantly to food security, global economy and climate change. Towards the efficient rice growth and yield monitoring, the main objective of this thesis is to develop a generic and transferable model for rice crop mapping based on Sentinel-1 and Sentinel-2 imagery.

Specifically, the present thesis deals with agriculture monitoring challenges, for the purposes of food security monitoring in South Korea. South Korea's food security problems include the overproduction of rice, which consequently leads to low self-sufficiency in the production of other major crops. For this reason, the accurate and large scale mapping of the paddy rice offers valuable information for the high-level decision-making related to food security. Therefore, in order to address this problem, a big data paddy rice mapping application was developed. In this regard, a set of processes has been implemented, using the computing framework Apache Spark and the Hadoop Distributed File System (HDFS) in a High Performance Data Analytics (HPDA). The input data comprise of long time-series of Sentinel-1 and Sentinel-2 images, but also pertinent vegetation indices produced from them.

At first, a 2-step data interpolation methodology was implemented to create a robust feature space with fixed timestamps. Furthermore, an unsupervised pixel-based technique, which utilizes the K-means algorithm, was approached for creating trustworthy and close-to-reality training data. Finally, a Random Forest model was trained using the generated training data in order to classify paddy rice in every pixel of the Area of Interest. The proposed paddy rice classification method achieves an accuracy of more than 92%, from as early as the end of July, for a study area in Northwestern South Korea.

Keywords

Sentinel, Big Data, Remote Sensing, Earth Observation, Machine Learning, High Performance Data Analytics, Hadoop, HDFS, Spark, Rice Mapping, Food Security

Περίληψη

Τα τελευταία χρόνια, η συνεχόμενη αύξηση του παγκόσμιου πληθυσμού, σε συνάρτηση με την κλιματική αλλαγή, αναμένεται να επηρεάσουν σημαντικά τον επισιτιστικό τομέα. Το ρύζι αποτελεί κύριο προιόν για πολλές χώρες γης και τρέφει περισσότερο από τον μισό πληθυσμό της. Επομένως, η παρακολούθησή του συνεισφέρει σημαντικά στον έλεγχο της επισιτιστικής ασφάλειας, της παγκόσμιας οικονομίας καθώς της κλιματικής αλλαγής. Με σκοπό την αποτελεσματική παρακολούθηση της ανάπτυξης και παραγωγής του ρυζιού, ο βασικός στόχος της παρούσας διπλωματικής εργασίας είναι να αναπτύξει ένα γενικευμένο και χωρικά μεταβιβάσιμο μοντέλο, με σκοπό τη χαρτογράφηση της καλλιέργειας του ρυζιού, κάνοντας χρήση δεδομένων Sentinel-1 και Senintel-2.

Συγκεκριμένα, η παρούσα εργασία ασχολείται με προκλήσεις παρακολούθησης της γεωργίας, για τους σκοπούς της παρακολούθησης της επισιτιστικής ασφάλειας στη Νότια Κορέα. Τα προβλήματα επισιτιστικής ασφάλειας της Νότιας Κορέας, αφορούν στην υπερπαραγωγή ρυζιού, η οποία κατά συνέπεια οδηγεί σε υψηλά κόστη αποθήκευση του πλεονάσματος. Επιπλέον, αυτή η υπερπαραγωγή συνεπάγεται χαμηλή αυτάρκεια στην παραγωγή άλλων σημαντικών καλλιεργειών. Για αυτόν τον λόγο, η υψηλής ακρίβειας και μεγάλης κλίμακας χαρτογράφηση του ρυζιού προσφέρει πολύτιμες πληροφορίες για τη λήψη αποφάσεων υψηλού επιπέδου, όσον αφορά την επισιτιστική ασφάλεια. Επομένως, για να αντιμετωπιστεί αυτό το πρόβλημα, αναπτύχθηκε μια εφαρμογή μεγάλης δεδομένων για τη χαρτογράφηση ορυζώνων. Πιο αναλυτικά, έχει υλοποιηθεί ένα σύνολο διαδικασιών, που κάνει χρήση της υπολογιστικής πλατφόρμας κατανεμημένης επεξεργασίας Apache Spark και του κατανεμημένου αποθηκευτικό σύστημα Hadoop (HDFS) σε ένα υπολογιστικό σύστημα υψηλών επιδόσεων (High Performance Data Analytics - HPDA). Τα δεδομένα εισόδου αποτελούνται από μεγάλες χρονοσειρές εικόνων Sentinel-1 και Sentinel-2, καθώς και σχετικούς δείκτες βλάστησης που παράγονται από αυτές.

Αρχικά, εφαρμόστηκε μια μεθοδολογία παρεμβολής δεδομένων 2 βημάτων για τη δημιουργία ενός δυναμικού χώρου χαρακτηριστικών με σταθερό χρονικό βήμα. Επιπλέον, υλοποιήθηκε με μια μέθοδος μη επιβλεπόμενης μάθησης σε επίπεδο εικονοστοιχείου, χρησιμοποιώντας τον αλγόριθμο των (K-means), για τη δημιουργία αξιόπιστων δεδομένων εκπαίδευσης. Τέλος, εκπαιδεύτηκε ένα μοντέλο Random Forest χρησιμοποιώντας τα παραγόμενα δεδομένα εκπαίδευσης, προκειμένου να ταξινομήσει το ρύζι σε κάθε εικονοστοιχείο της περιοχής ενδιαφέροντος. Η προτεινόμενη μέθοδος ταξινόμησης ρυζιού επιτυγχάνει ακρίβεια άνω του 92%, από το τέλος Ιουλίου, για μια περιοχή μελέτης στη βορειοδυτική Νότια Κορέα.

Λέξεις Κλειδιά

Μεγάλα Δεδομένα, Sentinel, Τηλεπισκόπηση, Παρατήρηση Γης, Μηχανική Μάθηση, Υπολογιστικό Σύστημα Υψηλών Επιδόσεων , Hadoop, HDFS, Spark, Χαρτογράφηση Ρυζιού

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Konstantinos Karantzalos, for his persistent support and supervision, as well as his engagement through the learning process of this thesis. I am also equally thankful to my colleague and friend Vasileios Sitokonstantinou for his continuous guidance and technical support. I am also deeply thankful to my colleague and friend Thanasis Drivas for his valuable advice and constant help. Their contribution to this work was very crucial and the result would not be the same if is wasn't for their help. I am deeply grateful to my supervsors Dr. Charalampos (Haris) Kontoes and Dr. Ioannis Papoutsis, with whom I collaborated excellently through the entire process of this study and their contribution has been more than worth mentioning. I consider myself truly lucky to had had the opportunity to work with all the aforementioned people, who have mentored me, guided me, supported me and most importantly inspired me. I am also thankful to my friend Antonis Koutroumpas, with whom I had been working side by side through the entire duration of this study. I would also like to express my gratitude to Assistant Prof. Dr. Georgios Goumas, member of the evaluation committee.

Moreover, I am also very grateful to my beloved, Maria, for her unconditional support, love and patience, and for always being there. Finally, I would like to thank all of my friends for all the emotional support. A special appreciation should be expressed to Giannis, who is like family to me and has been nothing less than a continuous source of support and motivation for several years now. Finally, I would like to thank my family who has supported and provided me with countless opportunities, every step of the way.

This work has been supported by the EOPEN project, which has been funded from the European Union's Horizon 2020 research and innovation programme under grant agreement 776019. Author acknowledge also the Copernicus Open Access Hub and the Hellenic National Sentinel Data Mirror Site for providing free access to Sentinel-1 and Sentinel-2 images.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last years, the continuous increase of global population, together with the climate change, is expected to affect the food sector significantly[1]. Therefore, the increase in global food production is of paramount importance in the coming years. The agricultural productivity needs to significantly increase to accommodate the growing population. This, however, needs to happen under environmentally-friendly and sustainable agricultural practices to alleviate, at the same time, the adverse impacts of climate change. It is therefore understood that there is great significance in the timely, large-scale and accurate monitoring of agriculture; to provide the necessary knowledge for evidence-based decision making on food security matters. Earth Observation (EO) science and data, constitute a unique basis for addressing the aforementioned requirements. In that direction, the European Union (EU), has invested greatly in the exploitation of EO i) to monitor the compliance of farmers to the Common Agricultural Policy (CAP), ii) to develop smart farming services that will increase the productivity and decrease the costs of the farmers and iii) to accurately monitor the extent, health, growth and productivity of the agricultural land over very large areas for the purposes of food security monitoring.

EO data can be collected through multiple sources; however, satellite EO data are the ones that allow the exhaustive and large-scale monitoring, which is so important in food security monitoring systems. The Landsat satellites have been successfully used the past years for food security purposes. However, due to their rather long revisit times (15 days), their usage is considered insufficient or suboptimal for solving specific problems, such as crop classification. This is particularly true in countries that suffer from extended cloud coverage. This gap has been filled the last few years by the Sentinel missions from the Copernicus Program, which has been developed and is being operated by European Space Agency (ESA). The data provided by the Sentinels are open, have a much shorter revisit time compared to Landsat (5-6 days) and offer a EO data of high spatial resolution (at 10, 20 and 60m). Therefore, large scale agricultural monitoring for food security is now possible, with minimal cost. Moreover, the frequent revisit times combined with the high image resolution offered by these missions creates paradigm shift in the applications and services that they enable. These new applica-

tions and services can be viewed as disruptors, tapping new non-traditional markets, allowing evidenced-based decision making and retrieval of actionable information for domains such as food security. However, these new opportunities may come along with a demand of handling these Big Earth Observation Data.

1.1 Problem Statement

The present work deals with agriculture monitoring challenges, for the purposes of food security monitoring in South Korea. South Korea's food security issues engage the overproduction of rice together with the low self-sufficiency in the production of other major crops. For this reason the systematic and large scale monitoring of the paddy rice extent has been identified as key knowledge for the high-level decision-making in regard to food security. Therefore, in order to address this problem a Big Data paddy rice mapping application was created. In this regard, a pipeline of processes has been implemented, using the computing framework Apache Spark and the Hadoop Distributed File System (HDFS) in a High Performance Data Analytics (HPDA). Specifically, the pipeline consists of a Data interpolation methodology, an unsupervised technique for creating training data and finally supervised learning for paddy rice classification. The input data comprise of long time-series of Sentinel radar and optical data, but also pertinent vegetation indices produced from them.

1.2 Thesis Objectives and Contributions

Through the extended literature review, which is presented in the following section, certain challenges and gaps become evident to the large-scale crop classification state-of-the-art, for food security monitoring. Multiple studies, have used optical and SAR imagery to classify rice and thus monitor food security. Additionally, several papers have used advanced machine learning and deep learning techniques to tackle such a problem. However, to the author's knowledge, there is no study that looks to all pertinent challenges of food security monitoring, namely scalability, transferability, scarcity of ground truth information, without utilizing cloud processing services such as Google Earth Eninge (GEE). To be more specific, this study addresses the processing of Big Earth Observation data, which are required for the large scale mapping of crops. This is done through the use of distributed filing systems and distributed implementations of machine learning algorithms. Furthermore, this work addresses the ever-present scarcity of high quality ground truth information for model training; using pseudo-labeling approaches based on unsupervised classifiers. Finally, this study is one of the few to attempt to provide crop classification outputs that early in the year (i.e. July for the rice crop). This is significant when talking about food security monitoring, as timely information is of essence.

To break it down even further, the overarching contribution of this work is the development and evaluation of a novel approach for a rice classification scheme that can scale to a national

application, which combines unsupervised and supervised learning techniques. The specific contributions of this thesis are:

1. Development of a temporal interpolation method for Big Earth Observation data to fit a distributed processing engine.

2. Generation of training data for paddy rice in the regions of Seosan and Dangjin using an unsupervised technique on Big Earth Observation data, in order to fill the gap of missing or unreliable ground truth data, to further train and create a transferable and generic model.

3. Creating a transferable model with fixed time steps for paddy rice mapping across the whole country of South Korea.

4. Producing trustworthy validation data for paddy rice through photo-interpretation in the Seosan and Dangjin provinces in South Korea.

Chapter 2

Literature Review

2.1 Background on rice cultivation

Rice is ranked, worldwide, as the third-highest commodity in production, with 7.82 million tonnes in 2018 (FAOSTAT). It is harvested in over 167 million ha in multiple countries all over the world, mainly in Asia and Africa. Regional, national and global statistics about rice production can be found, for example on FAOSTAT, but in many cases this information is restricted to only national level.

Rice plants take around 3–6 months to grow from seeds to mature plants, depending on the variety and environmental conditions. The phenological stages of the plant are the germination, vegetative, reproductive, and ripening phases.[...]. Germination occurs when the first shoots and roots start to emerge from the seed and the rice plant begins to grow. During the vegetative phase, tillers and more leaves are developed, with a simultaneous increase of the plant height. The duration of this stage is typically between 55 and 85 days. The vegetative phase can be split into 3 suphases, the early vegetative, the seedling and the late vegetative phases. The early vegetative phase begins right after the seed germination. The seedling stage starts with the emergence of the first root and shoot, and end with the appearance of the first tiller. Later on, the late vegetative phase begins when first tiller appears and lasts until the maximum number of tillers is reached. Following the vegetative, comes the reproductive phase. Entering this stage, a bulging of the leaf stem that conceals the developing panicle appears on the plant, which then continues to grow. Afterwards, flowering begins a day after the panicle is fully visible and can continue for about a week. Finally, the ripening phase starts with the flowering and ends, when the rice is ready to be harvested. This stage is affected directly by the temperature and usually takes around 30 days, but it can last even twice as much, in cool temperate areas. [2]

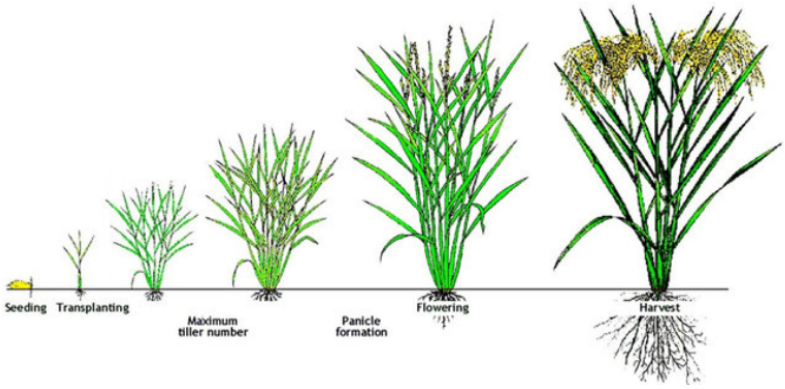


Figure 2.1: Rice life cycle [3]

2.2 Remote Sensing

Remote sensing (RS) can be defined as the technology used to acquire physical data about an object by detecting energy reflected or emitted by that object when the distance between the object and the sensor is much greater than any linear dimension of the sensor. [4] [5] The remote sensing data consists of three different types of resolution[6]:

1. Spatial resolution, which indicates the size of the smallest area from which a satellite sensor can receive information. With higher spatial resolution, the image received from a sensor contains more detailed information.

2. Spectral resolution, which is the ability of a RS system to separate the difference in reflectance of the same ground object at different wavelengths.

3. Temporal resolution, which refers to the frequency that required for imaging the same ground area at the same viewing angle by the same sensing system.

4. Radiometric resolution, which corresponds to the sensitivity to the magnitude of the electromagnetic energy of the sensor.

2.2.1 Earth Observation

Earth observation (EO) is a specific field of Remote Sensing which refers to the gathering of information about planet Earth's physical, chemical and biological systems via remote sensing technologies, usually involving satellites carrying imaging devices [7]. For purposes of agricultural monitoring, EO data has been widely used, mainly focusing on the crop type classification, the cropping intensity and the farming practices monitoring, in various scales of precision. The data generated via remote sensing are of two types depending on the source of energy; Passive and active. Passive remote sensing indicates the existence of a natural source of energy, which is the sun. On the other hand, active remote sensing exploits several controlled energy sources that beam section of the electromagnetic spectrum.

2.2.2 Multispectral remote sensing

The passive remotely sensed data are known as optical imagery. The basic characteristic of this type of remote sensing is that the imagery is acquired only during the day as it depends on the on the reflections of sunlight from objects on the earth surface. In addition, the existence of clouds is considered as another drawback. The optical satellites are set in near earth orbits. Thus, they have the potential to provide detailed data at high ground resolution. Data from these satellites is provided both free and commercial. Free data can be found at least 10m of spatial resolution. Therefore, optical satellite imagery can and is used to several fields such as land use – land cover change, crop mapping, disaster monitoring et cetera.

2.2.3 SAR remote sensing

The active remotely sensed data can be found as imagery (e.g. Synthetic Aperture Radar, SAR) or in other form such as altimeters. The process of generating data is based on the emission of radar signals towards an area of interest. As the objects in that area reflect the signals, radar instruments capture the reflected signal. The main advantages of the radar is the penetration of cloud cover and the potential to acquire images at any instance, day or night. Thus. SAR satellites are used to measure soil moisture in bare areas, water bodies' detection and land-use and land cover change. The drawback of this type of sensors is that various surfaces or objects can be significant backscatters affecting the accuracy of captured information.

2.2.4 Satellites in this study

Sentinel-1

The launch of Sentinel-1 mission took place on April 3, 2014 for Sentinel-1A and on April 25,2015 for Sentinel-1B. Sentinel-1 provide observations under any weather conditions [8] and its products are suitable for water detection, which is the key indicator for paddy rice mapping during the transplanting period. In Figure 2.3 we can see snapshot taken from space of the satellite.

Figure 2.2: Sentinel-1 satellite image. source: `https://sentinel.esa.int`

2.2.5 Sentinel-2

The launch of Sentinel-2 mission took place on June 23,2015 for Sentinel-2A and on March 7, 2017 for Sentinel-2B. As all Sentinel data, Sentinel's 2 data can be found free from several hubs. In this case, data has been downloaded automatically from the Hellenic Mirror Site. The spatial resolution of raw data is 10m, 20m and 60m. Table 2.1 presents information about the central wavelength and resolution of each band.

Sentinel-2 Bands	**Central Wavelength (μm)**	**Resolution (m)**
B01 - Coastal aerosol	0.443	60
B02 - Blue	0.490	10
B03 - Green	0.560	10
B04 - Red	0.665	10
B05 - Vegetation Red Edge	0.705	20
B06 - Vegetation Red Edge	0.740	20
B07 - Vegetation Red Edge	0.783	20
B08 - NIR	0.842	10
B8A - Vegetation Red Edge	0.865	20
B09 - Water Vapour	0.945	60
B10 - SWIR - Cirrus	1.374	60
B11 - SWIR	1.610	20
B12 - SWIR	2.190	20

Table 2.1: Sentinel-2 Bands.

All available cloud-free Sentinel-2 data were acquired from March to October of 2018.

Figure 2.3: Sentinel-2 satellite image. Source: `https://sentinel.esa.int`

Sentinels (1, 2), Landsat-8, MODIS: spectral channels, spatial resolution, revisit time, swath/FOV SAR: TSX, COSMO-SKYMED, ALOS2/PALSAR (carrier frequency, spatial resolution, revisit time, swath/FOV, polarisations)

2.3 Rice Classification in Earth Observation

Paddy rice is a major crop and staple food in many Asian countries, including South Korea. It is usually overproduced, due the long experience of local farmers on this cultivation and the absence of incentivization for the cultivation of alternative crops.This results in large storage costs for the overproduced rice, but also great dependence on imports for most of other major grains. It is therefore apparent that the accurate, timely and large-scale mapping of paddy rice in a country, such as South Korea, is of great importance towards food security monitoring.
The accurate classification of paddy rice requires the input data to capture the different phenological phases of the fields, thus demanding the exploitation of a time-series of images. Over the last years, several different approaches have been tested for EO-based rice mapping, which can be divided into four main categories; namely supervised learning, unsupervised learning, knowledge-based and phenology-based approaches, using a variety of data sources [9]. During the 80s and 90s, Landsat was the main data source for paddy rice classification. Most of these studies have explored supervised learning techniques, such as Maximum Likelihood Classification (MLC)[10] [11], or unsupervised learning techniques [12]. Unfortunately, the issue of frequent cloud coverage created restrictions and challenges that remained unsolved for the scientific sector. After 2000, improved techniques were introduced by applying new classification methods, and thanks to new data sources, such as MODIS, together with the integration of vegetation indices (VIs). The usage of vegetation indices, which arise from spectral transformation of two or more bands, in specific phenological stages of the rice crop, provided useful information for feature engineering in the classification. For example, since paddy rice is transplanted in inundated fields, water related indices can be used to identify such

areas. Respectively, indices measuring its greenness after a couple of months from seeding, can offer very valuable information about the plant's growth. Together with the use of VIs, other more sophisticated algorithms were used. For example, Neural Networks, which were combined with backscatter input data [13] [14] and Support Vector Machines (SVM) [15] [16] presented very promising results. Moreover, studies have been conducted using time series of vegetation indices and threshold-based techniques. For example [17] followed a threshold-based approach in the Mekong Delta using EVI[18] and [19] used a NDVI[18] threshold-based approach in southern China. Additionally, the capabilities of rice mapping using SAR data has been also examined in multiple studies so far[20] [21] [22], resulting to very efficient results. All the studies that were mentioned, highlight the significance of paddy rice growth phases in the rice mapping. Some of the approached include temporal variations of VIs, but the do not make use of remote sensing data to recognise the key phenological stages. In more recent studies, phenology-based approaches research has been examined extensively. Researchers, have managed to generate large scale paddy rice maps, for example in South and Southeast Asia and southern China [23] [24] using MODIS data, in northeastern Asia (Japan, North Korea, South Korea, and NE China) [25] using Landsat-8 data. Monitoring of rice from remote sensing requires SAR data at high resolution (10-30 m) and temporal resolution of 10 days [26]. These type of data became freely available after the launch of Sentinel-1A and Sentinel-1B missions, at 2014 and 2016 respectively. Specifically, for paddy rice mapping multiple studies have been conducted over the last 5 years either only Sentinel-1 data [27] or combined with data from optical sensors [28] [29] [30]. Moreover, the last couple of year Sentinel-2 images have been also used extensively for rice crop mapping and monitoring resulting also to efficient and more accurate regional and national paddy rice maps [31] [32] [33] [34]. It is apparent that rice classification and monitoring issues have been examined since many years, but the last five and thanks to the data offered by the Sentinel satellites, the scientific interest has been raised significantly resulting to more improved results, not only for rice crops, but also in agricultural monitoring in general. To reinforce this argument, figure 2.4 presents the number of publications about agricultural monitoring and remote sensing, as searched in the Scopus website (`https://www.scopus.com/`). It is worth mentioning that the number of publications in 2019 is more than double compared to that of 2014.

2.4 Big Data technologies for remote sensing

Big Data as a term was introduced by Roger Mougalas in 2005, to refer to a large set of data that cannot be managed and processed using traditional algorithmic techniques. In 2010, Eric Schmidt stated: "there were 5 exabytes of information created by the entire world between the dawn of civilization and 2003. Now that same amount is created every two days.". Currently, Big Data is defined by the "5Vs", which are also termed as the characteristics of Big Data as follows:

1. Volume: Typically, volume of data itself defines if the term "Big Data" is suitable.

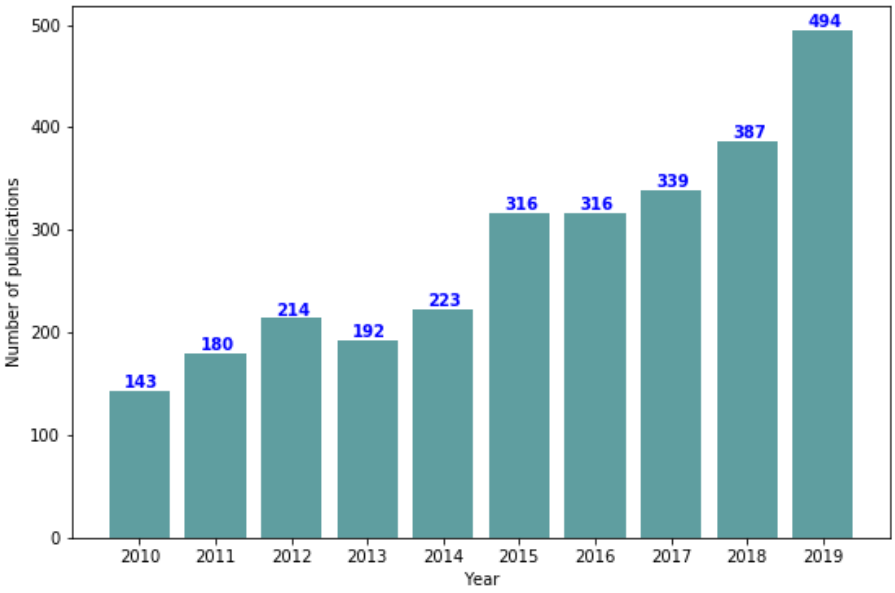


Figure 2.4: Number of publications per year for keywords crop classification and remote sensing available on Scopus.

Cloud–computing, IoT, mobile traffic etc, have been major causes for the Big Data effect.

2. Velocity: Refers to the speed at which data is being accumulated. What is prominent from the past decade is the accelerating pace of data creation/processing/storage etc.

3. Variety: Refers to the structure of the acquainted data. From that perspective, they may be:

   - Structured: Traditional structured data, that can be stored in a relational database.
   - Semi–Structured: Data organized in variant informal structures, (i.e. Log files, JSON files etc).
   - Unstructured: Include all other types of data, i.e. e-mails, images, voicemails etc. It is estimated that more than 80% of data generated today are unstructured.

4. Veracity: Since data can be collected from various sources, there is a growing need for the evaluation of data gathered, before using it for business/research purposes. Most indicative such metrics are quality, integrity, credibility and accuracy of the data.

5. Value: Another metric, which relates the data to its actual contribution in decision making and/or the solution suggested, is Value.

The Sentinel mission created new opportunities in the remote sensing and earth observation, and therefore in the food security monitoring using these resources. However, together with these new opportunities comes the need of handling and processing Big EO Earth. In parallel, massive amounts of satellite images are becoming available that can be used for the creation of value-added Earth Observation (EO) products. One challenge is to extract knowledge from the raw satellite data in an automated way. An additional technical challenge involves the effective management of the extracted information, to allow fast and accurate decisions of spatio-temporal nature in a real operational scenario. Remote sensing big data are attracting more and more attention from commercial applications to academic fields. This argument is reinforced by an increase in the number of publications relevant to the big data and remote sensing subjects over the last years, as shown in Figure 2.5. The cloud-based platform Google Earth Engine (GEE), which combines a tremendous amount of satellite imagery and geospatial datasets with planetary-scale analysis capabilities [35], has offered a lot of potential to the scientific sector. As far as the agricultural monitoring is concerned, GEE has been widely used for multiple purposed such as covering areas around vegetation monitoring [36] [37] [38] and cropland mapping [39] [40] [41], among others [42]. Moreover, research about the paddy rice crop mapping, which is the subject of study in this thesis, has been conducted [25] [34] [43] [44] . However, as helpful as these tools may be, they do have some restrictions. The most frequent problems found when running an program in the GEE environment are time limits and memory and storage. Memory limit problems normally arise when running some commands on big images. Secondly, restrictions exists in saving results in Google Drive or Google Cloud where the standard free space is 15 Gb storage [45]. Moreover, models created on the GEE platform are also limited to be used only inside the platform. Another solution to manipulate Big Data is High Performance Computers (HPC). Nowadays, HPCs have become abudant and large-scale computing is available. Furthermore, a variety of tools have been created in order to process large-scale data such as GeoSpark [46] and HADOOP [47]. The drawback of using these resources is that it requires specific technical expertise.

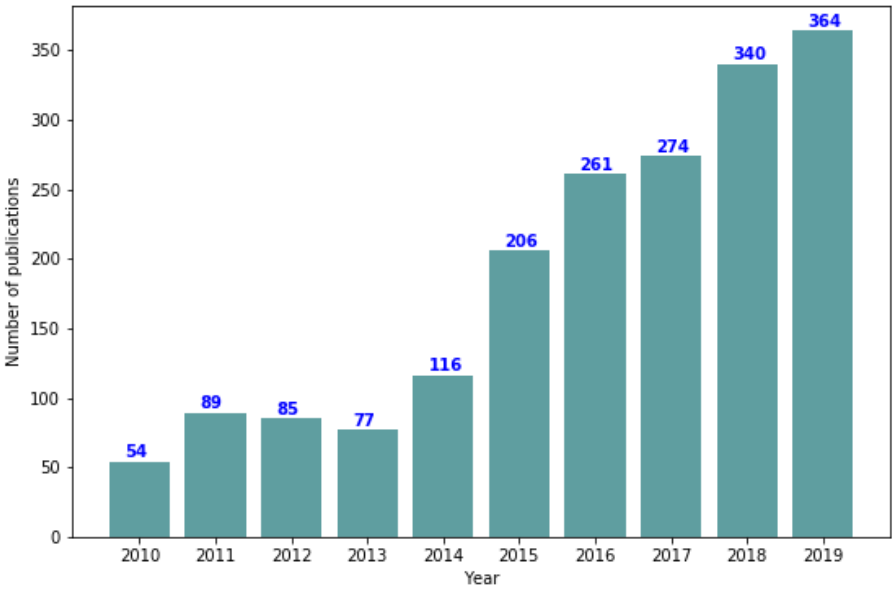


Figure 2.5: Number of publications per year for keywords big data and remote sensing available on Scopus.

Chapter 3

Study Area, Datasets and Data Analytics Framework

3.1 Area of Study

Methods and results that will be presented refer to the region of South Korea, and more specifically a wide area containing mainly two adjacent cities, Dangjin and Seosan, which are located at the northwestern end of South Chungcheong Province of the country. Dangjin and Seosan are in the temperate monsoon and continental climate zones, as is the entire country. Based on meteorological data for the last years, Seosan has a humid subtropical climate/humid continental climate with annual mean temperature of 11.8 C and annual precipitation of 1,285 mm and mean humidity of 74.1%. The total area is 741.2 km^2, consisting of 261.51km^2 cultivated land from which 78% are paddy rice fields [48]. At the same time, Dangjin's annual precipitation reaches 1,158.7 mm and annual mean temperature is 11.4 C. The total area here is 664.13 km^2, while the cultivated land extends to 244.29 km^2 with paddy rice area consisting 83.5% of it [48]. The two regions are recorded among the highest rice producers in the country [49]. The planting and transplanting processes take place in May, while the harvesting starts from September and lasts until the end of October.

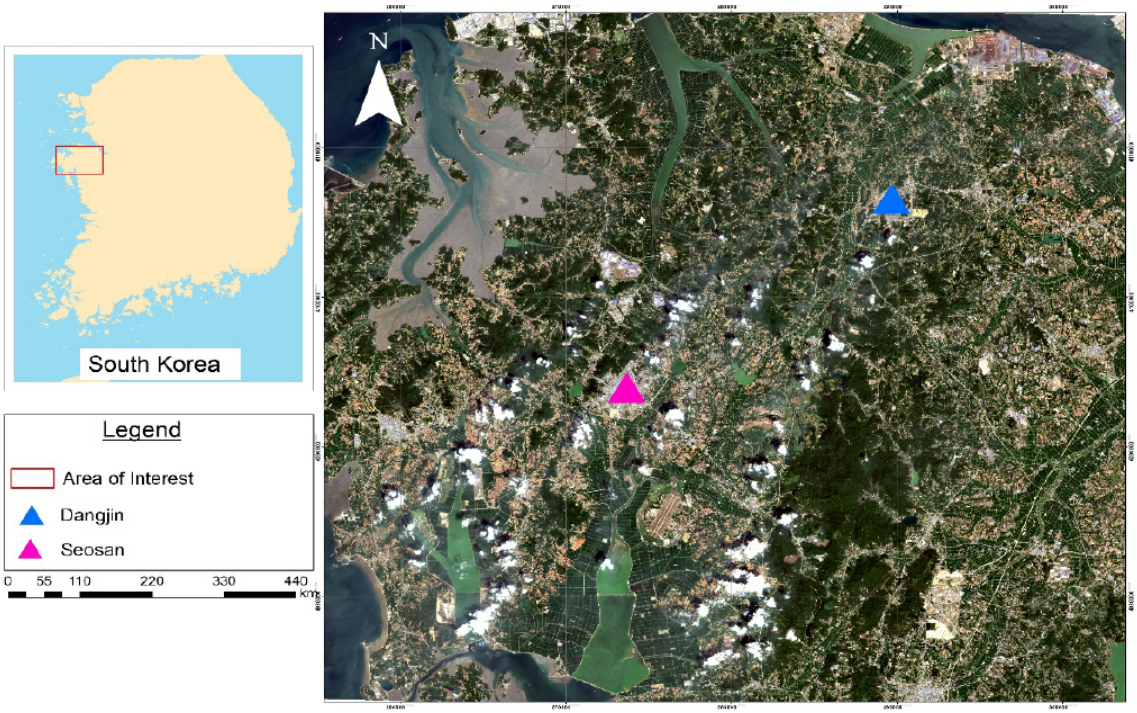


Figure 3.1: Study Area located in the cities of Seosan and Dangjin in South Korea

3.2 Data

The data used involved Sentinel-1 and Sentinel-2 imagery retrieved from the Hellenic Mirror Site, and were made available from the National Observatory of Athens as part of the EOPEN project of the European Space Agency (ESA).

3.2.1 Sentinel Data

The data acquired, dated from 01/03/2018 to 31/10/2018. Specifically, for Sentinel-1, data were acquired for one date inside each 10-day window of each month from March to October. Subsequently, the Level 1 Ground Range Detected products in Ground Range Detected High Resolution (GRDH) format and in Interferometric Wide (IW) swath mode were used. There are two types of incoming data; dual-polarized vertical transmission with Single copolarization along with vertical transmit/vertical receive (VV) and Dual-band cross-polarization along with vertical transmit/horizontal receive (VH) bands in a spatial resolution of 10 m and a swath of 250 km. As the raw data does not have the required format, a preprocess pipeline has been build up. It consists of several modules such as clipping to the area of interest, radiometric calibration, speckle filtering using Lee filter, terrain correction using Shuttle Radar Topography Mission (SRTM) 10-m and conversion of back-scatter coefficient (σ^0) in decibels (dB).

For Sentinel-2 data, since the images are prone to clouds, cloud free ones were selected for

the year of inspection (e.g. 2018), from March to October in order to capture the entire crop period of the paddy rice cultivation. Figure 3.2 presents the acquisitions dates for both Sentinel-1 (red lines) and Sentinel-2 (blue lines) products. The 60m bands were excluded from the final dataset while the 10m (B02, B03, B04 and B08) and 20m (B05, B06, B07, B8A, B11 and B12) bands were included. Moreover, an atmospheric correction was applied to the downloaded data using the sen2cor software [50]. In detail, images were transformed from Top-Of-Atmosphere (TOA) Level 1C products, to Bottom-Of-Atmosphere (BOA) Level 2A products. The total amount of the retrieved images is 230, which includes 23 different acquisitions for each of the 10 used bands.

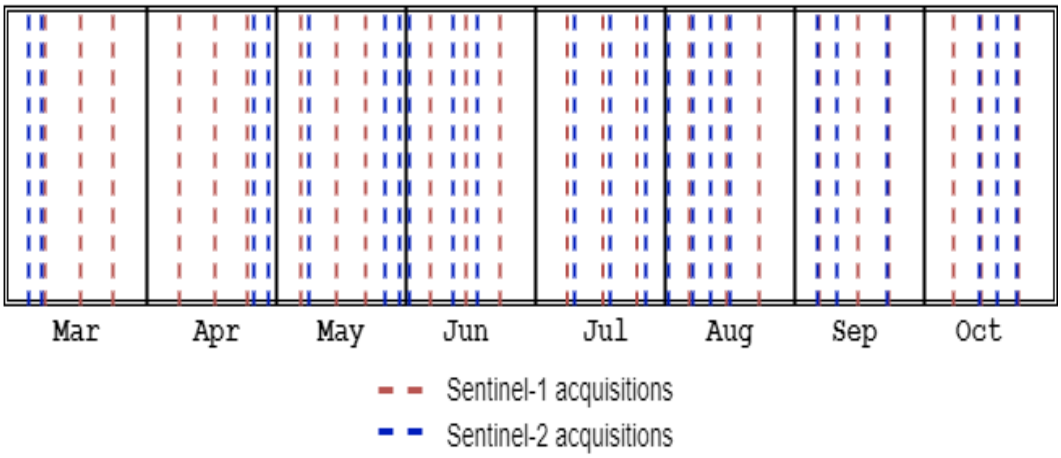


Figure 3.2: Dates that data was acquired from Sentinel-1 and Sentinel-2 satellites

3.2.2 Volume of Data

Products of Sentinel-2 imagery include 10 bands that were mentioned before, as well as three vegetation indices, as described at chapter 4.1. Sentinel-2 images were acquired for 23 different dates across the rice cultivation period resulting to a very large dataset of 93GB volume in a .csv format. Likewise, Sentinel-1 data, which include 24 VV and 24 VH backscatter, has a volume of 15GB in a .csv format, creating thus, a total of 107GB remote sensing data. It is obvious that data of those volumes cannot be processed by conventional computational machines. There is a necessity for technologies that can cope with such large data, and thus, Hadoop Apache File System (HDFS) of the High Performance Data Analytics (HPDA) was used to store it.

3.2.3 High Performance Data Analytics (HPDA)

High performance computing (HPC) has bee widely utilized to satisfy the needs of Big data analytics for many years. The exponential increase in data, however, creates new requirements in high performance computing in order to handle massive amounts of data. High Performance Data Analytics is a term that was conceived to explain the meeting of high per-

formance computing and big data analytics. HPDA is the process of investigating extremely large datasets to extract enlightening information, using parallel processing of HPC to run powerful analytic software. In this study , the powerful system for data analytics workflows Cray-Urika-GX from the High-Performance Computing Center Stuttgart (HLRS) was used. The system comes with state-of-the-art frameworks and tools in the data science domain such as Apache Hadoop and Apache Spark to cater the needs of analytics experts. Specifically, the system consists of 41 compute nodes. Each of the nodes is equipped with 2 Intel BDW 18-Core, 2.1 GHz processors, has memory of 512G and disk capacity of 2TB. There are a number of resource/workload management tools installed on the Urika-GX system, including Mesos, Marathon and YARN. These tools enable management of analytic workloads, dynamically allocate system resources to applications as needed, and provide the flexibility of running multiple jobs across the cluster concurrently. Specifically, Apache Mesos acts as the primary resource manager on the Urika-GX platform . It is a cluster manager that provides efficient resource isolation and sharing across distributed applications and/or frameworks. It lies between the application layer and the operating system and simplifies the process of managing applications in large-scale cluster environments, while optimizing resource utilization [51]. Finally, for the connection to the HLRS system, the ssh protocol was used.

Number of compute nodes	41
Processor compute nodes	2 x Intel® BDW 18-Core, 2.1 GHz
Memory/node	512 GB (DDR4 2400)
Disk capacity per node	2 TB HDD; Intel DC P3608 (1.6TB, MLC) SSD
Lustre filesystem	Sonexion 900: 240 TB, 4.0 GB/s
Software Stack	SPARK, Hadoop, Cray Graph Engine

Table 3.1: Specifications og Cray-Urika-GX computer of High-Performance Computer Center of Stuttgard

3.2.4 Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS)[1] is a distributed file system designed to run on commodity hardware. HDFS has a master/worker architecture. The master server, namely the NameNode, executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The workers, namely the DataNodes, manage storage attached to the nodes that they run on. They are responsible for reading and writing requests from the file system's clients, as well as block creation, deletion, and replication when instructed by the NameNode [52]. In figure 3.3 is shown the architecture of the file system.

HDFS is very suitable for applications of large data. Apart from csv format, Apache Hadoop suports a columnar storage format, called parquet, which is a more efficient way of storing and handling data, and was also used for this study. First of all, by converting the

[1] `https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html`

data to parquet format its size is reduced from 107GB to 30GB, equivalent to 72% decrease. Moreover, this format maximizes the effectiveness of querying data, which is very crucial for the computational complexity of the experiments that are presented on the next chapters. Indicatively, the total time of parsing a parquet file, compared to a csv file that contains the same information, can be more that 30 times faster.

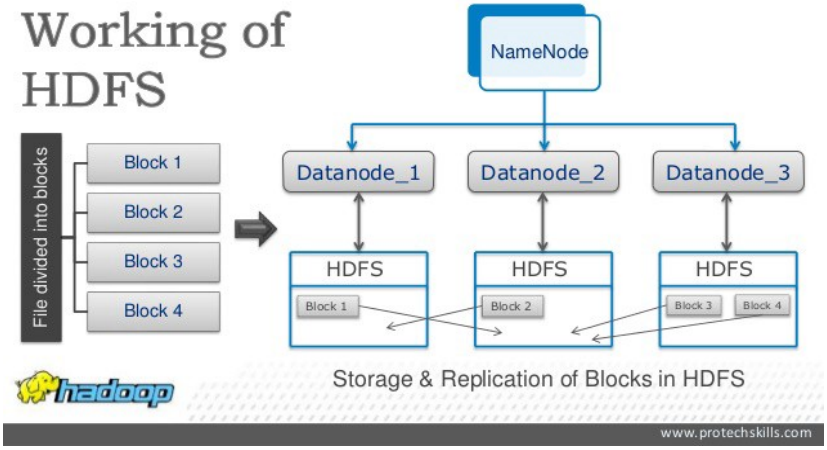


Figure 3.3: HDFS architecture

3.2.5 Validation Data

A labeled dataset has been created for the area of interest to choose the optimal clustering for pseudo-labels (see Section 4.3) and ii) evaluate the accuracy of supervised classification. The rice paddy area for the validation datasets was extracted from the level-three land cover map, as acquired from the Korean Ministry of Environment (KME)[2]. The level-3 land cover maps are 3-m wide for linear elements and 100 m2 for plane elements, thus the spatial resolution is comparable to the one of the Sentinel data.The maps are digitized using VHR satellite imagery (KOMPSAT-2 and IKONOS) and aerial ortho-photos [53]. KME produces these land cover maps for each county, but they are updated only every few years, while being subject to significant errors. The year of inspection for this study is 2018, for which a validation rice map has been created for the Seosan/Dangjin region. In this case, the latest available level-3 land cover map was for 2015. For this reason, the past land cover map of 2015 was updated to be representative for 2018. This was done using a change detection method, as published in [54], to remove instances that have been classified as rice in 2015, but do not represent rice fields in 2018. Then, a photointerpretation was conducted based on the time-series of Sentinel-1 and Sentinel-2 data to create the labeled datasets. Labels for approximately 5% of the pixels of the area of interest were created, for both the rice and the non-rice classes. More specifically, random sub-regions were generated and within those regions rice and non-rice pixels were selected, ensuring that the validation dataset is representative and therefore the classification metrics are trustworthy. Finally, 1,380,643 non-rice and 1,281,934 were

[2] `https://eng.me.go.kr/eng/web/main.do`

selected. For the photo-interpretation, both radar and optical satellite data were used along with NDVI and NDWI (see section 4.2.3). In Seosan the the flooding period starts in end of April/start of May. Sentinel's -1 sensitivity to water, in combination with its weather independent capacity helped in identifying flooded plots before transplanting stage begins. In addition, observation of Sentinel-2 natural color RGB time series from April to October contributed in visual monitoring of whether the specific crop follows the different phenological stages of rice (vegetative, reproductive and ripening). Finally, the analysis of NDVI and NDWI (see section 4.2.3), multi-temporal profiles of the potential rice crop led to the determination of the labelling. Specifically, NDVI values start rising after the transplanting period and then decrease after reaching their maximum in flowering stage, whereas the NDWI values are higher before the transplanting period and they decrease afterwards. For the non-rice class, pixels that represent forest and cities were selected. Moreover, pixels of water, and mainly sea areas near the coast or lake areas, were selected as non-rice. Under many of those environments, it seems that plants are growing and the spectral signatures of these plants are usually very similar to the corresponding spectral signatures of rice.

Figure 3.4: The validation dataset. Purple color represent the rice while red color the non-rice class.

Chapter 4

Methodology

4.1 Algorithmic architecture

The architecture of the proposed methodology consists of three sections, data preprocessing, creation of training data and rice classification. At first, raw data are retrieved from Sentinel-1 and Sentinel-2 satellite sensors and then processed using Python scripts in order to create a large feature space consisting of a time-series of Sentinel-1 backscatters, Sentinel-2 bands as well as Sentinel-2 VIs for each pixel. Then they are passed as input in an unsupervised model (k-means), which is used to extract the land, water and rice classes of the region. Afterwards, a 2-step interpolation method is applied to generate time-series of fixed time step. The interpolation to create a fixed time-step is used to make the model transferable to other regions. S-1 and S-2 acquisitions over different areas have different dates of pass. Having a methodology that translates the scattered, in time, acquisitions to a fixed temporal grid, allows for potentially applying the model to other regions; now that one-to-one matching among the features is possible. Finally, using the created rice map a Random Forest classifier is trained, to generate the final model for paddy rice pixel classification.

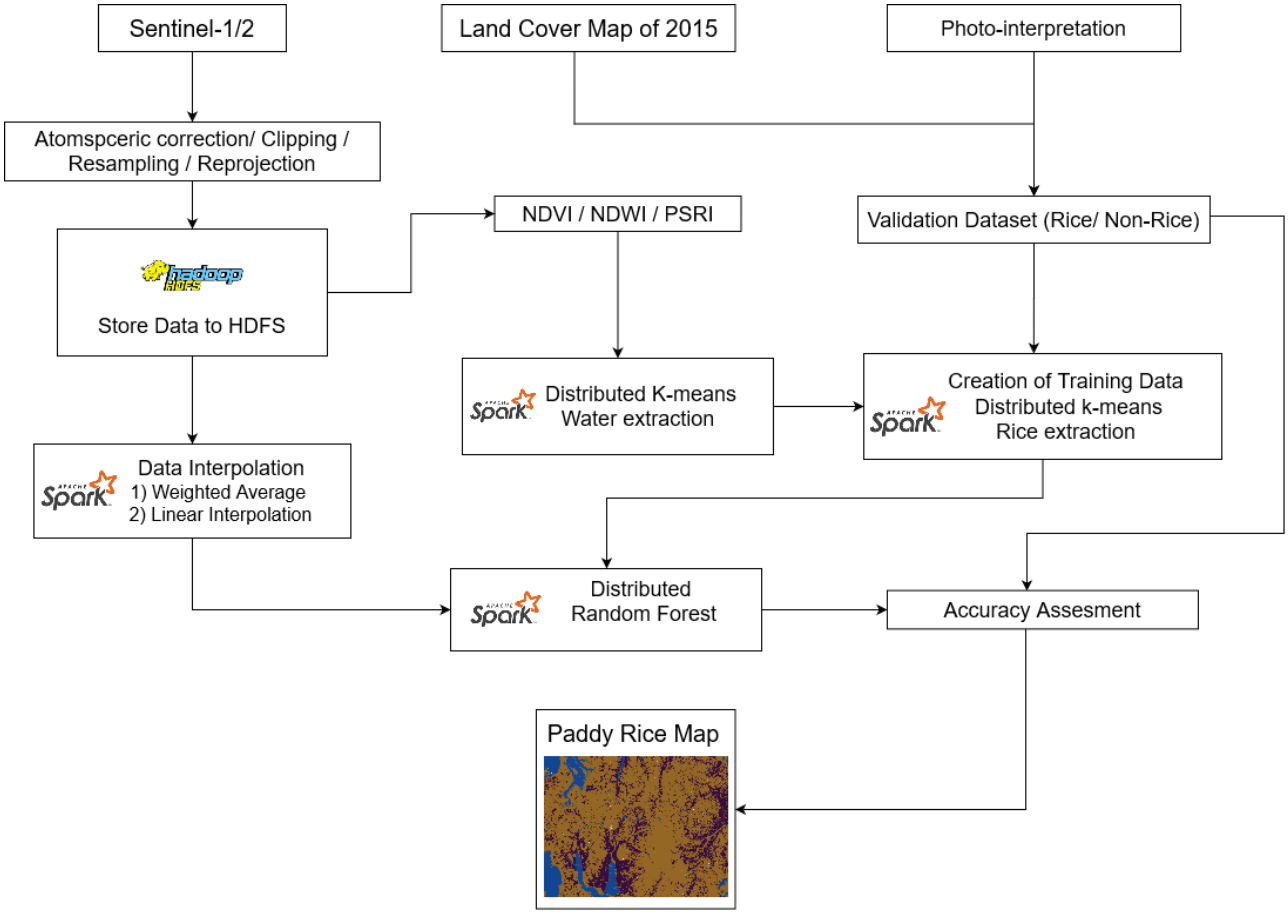


Figure 4.1: Workflow of the proposed methodology

4.2 Data Preprocessing

4.2.1 Data Transformation

Sentinel data are offered freely via various data hubs. For this study, data was retrieved through the Hellenic Mirror Site. At the time the raw data was discovered and downloaded, a preprocess pipeline was triggered automatically. This process consists of the atmospheric correction applied to Top-Of-Atmosphere (TOA) Level-1C orthoimage products in order to create an orthoimage Bottom-Of-Atmosphere (BOA) corrected reflectance series of products and was based on the Sen2Cor software. In addition, Sen2cor produces a scene classification map indicating cloud, snow and other probabilities for each pixel. Figure 4.2 presents the result of the scene classification map of Sen2cor for an image acquired in the 2nd of June. The corrected images were then converted to TIFF format, resampled at 10m and reprojected to EPSG 3857. The final step of this pipeline consists of the generation of Level-3 products; vegetation indices and binary cloud masks.

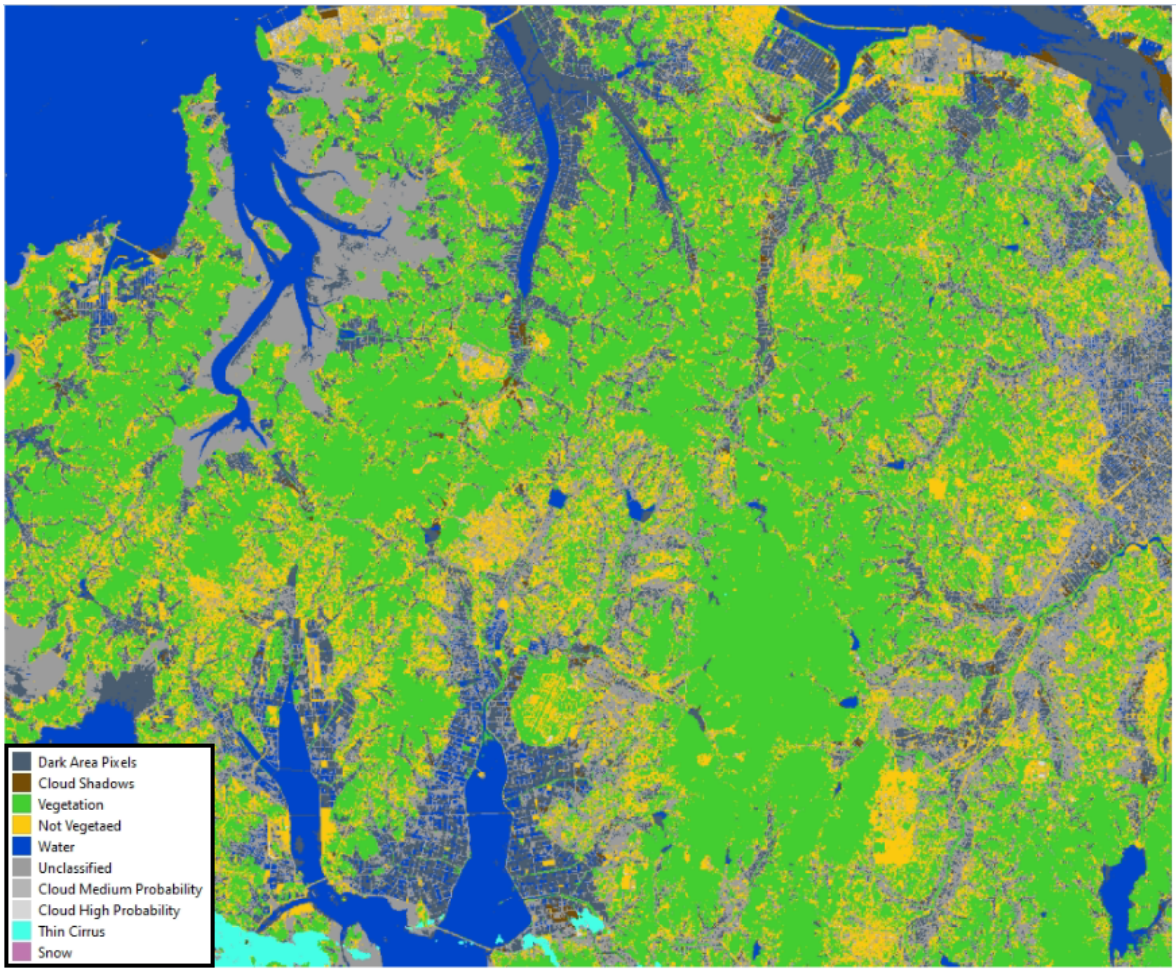


Figure 4.2: Scene Classification product from sen2cor software, of the 2nd of June.

4.2.2 Data Interpolation

The dates of data acquired by the Sentinels differ from region to region. Therefore, in order to be able create a transferable model, a feature space with fixed time stamps for each band and index was necessary. In order to avoid temporal gaps between acquisitions, due also to prior image selection after cloud coverage with a threshold of 65%, this approach lies on the creation of robust dekadal time-series, indicating the 5th, 15th and 25th day of each month. For this purpose, a pipeline of two interpolation methods was applied. Initially, the method examines a 10-day window temporal space of the acquisition dates and applying weighted average interpolation for each pixel, for the dates that fall within, constructs part of the dekadal time-series. More specifically, for constructing features for the 5th of each month, acquisitions between 1 and 10 of that month were considered, relevantly for the features of the 15th of each month, acquisitions between 11 and 20 of that month were inspected, while for the features that generate the 25th of each month, we examined the acquisitions from 21 to the end of each month. Since many images have been dropped, due to cloud coverage, from Sentinel-2 data, it is very possible we have no image information within one of these specific

time intervals. Consequently, linear interpolation is applied to form the remaining dekadal dates, as well as the remaining null values of cloudy pixels. An example of the interpolation pipeline is presented in Figure 4.3. The first column corresponds to the real data acquired from the Sentinel-2 satellites, in the listed dates, the second to the results of the weighted average method and the third to the final outcome, after interpolating the data linearly to fill the missing dates. Furthermore, Figures 4.4 and 4.5 present an example of the data interpolation pipeline. At Figure 4.4 can be observed an example of the weighted interpolation method, between 2 NDVI images that fall within the same 10-day time interval. Respectively, Figure 4.5 indicates the outcome of the second part of interpolation, where there are no data for a specific time interval, namely from 20 to 31 of August. Indicatively, NDVI image for that date is presented, which is constructed through linear interpolation between the previous and the next available acquisition. In section 4.5, is described in detail the implementation and the tools that were used for this task.

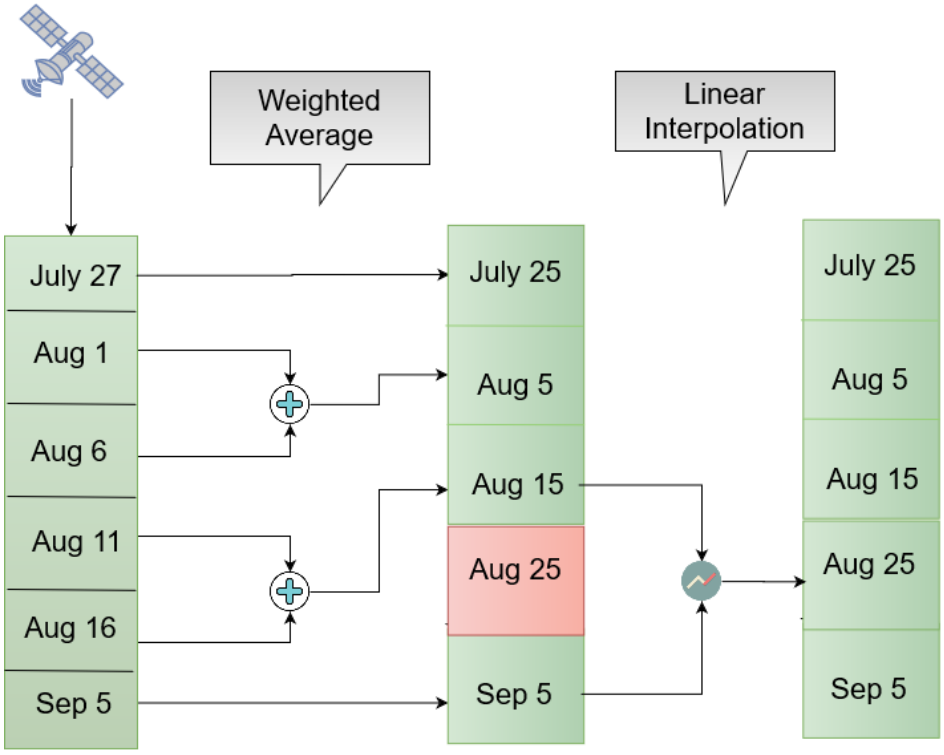
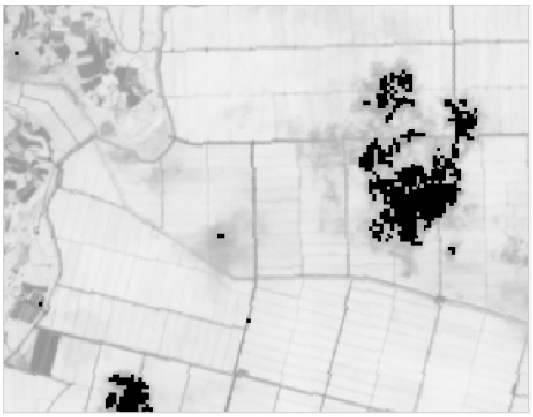


Figure 4.3: An example of the interpolation method. EXPLANATION!

4.2.3 Extration of Vegetation Indices

For crop monitoring and classification, there are multiple indices that have been extensively used, including the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Water Index (NDWI) [55] [56], which were also used here. Also, the usage of Plant Senescence Reflectance Index (PSRI) was considered valuable for this study, being particu-



(a) August 11

(b) August 16

(c) Weighted Average Result

Figure 4.4: Weighted Average Interpolation example. (a) refers to the NDVI image from 11th of August, (b) to the NDVI image from the 16th of August and (c) in the final product from the weighted average method, which indicates the 15th of August. The black pixels of the figures (a) and (b) correspond to nan values due to clouds.

(a) August 15

(b) September 5

(c) Missing date, August 25

Figure 4.5: Linear Interpolation example. (a) refers to the constructed NDVI image for 15th of August, (b) to the NDVI image from the 5th of September and (c) in the final product from the linear interpolation which created a new image on the 25th of August.

larly sensitive to the senescence phase of the rice cultivation. NDVI is a vegetation index widely used for its ability to measure the greenness of plants, their chlorophyll content and their vigor. This is possible and well proven as chlorophyll strongly absorbs the visible light (0.4-0.7 µm) and strongly reflects the NIR (0.7-1.1 µm) light [57]. NDVI formula is presented in equation (4.1), and in formula (4.2) for Sentinel-2 bands:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{4.1}$$

$$NDVI = \frac{B08 - B04}{B08 + B04} \tag{4.2}$$

NDWI is an index in regard to plant water content. It uses the NIR and Green parts of the electromagnetic spectrum in order to provide a normalized index of the NIR reflectance of plants' chlorophyll varies along with its water -or otherwise- dry matter content [57]. NDWI formula is presented in equation (4.3), and in formula (4.4) for Sentinel-2 bands:

$$NDWI = \frac{Green - NIR}{Green + NIR} \tag{4.3}$$

$$NDWI = \frac{B03 - B08}{B03 + B08} \tag{4.4}$$

PSRI has been proposed to determine the stage of leaf senescence with a high degree of precision, because it is sensitive to retention of carotenoid [58]. PSRI formula is presented in equation (4.5), and in formula (4.6) for Sentinel-2 bands:

$$PSRI = \frac{Red - Blue}{NIR} \tag{4.5}$$

$$PSRI = \frac{B04 - B02}{B06} \tag{4.6}$$

4.3 Creation of Training Data

As described in section 3.2.5, the only available ground truth information for training a rice classification algorithm were some out of date land cover maps, provided by Korea Statistics. These land cover maps, which are ilot based, are updated every several years through sampled based statistics. In the case of Seosan/Dangjin the latest available land cover map was from 2015. After carefully inspecting the land cover map and comparing it with the 2018 reality, which is the year of inspection for this study, certain issues became apparent. First of all, as expected, approximately 15% of the rice fields of 2015 land cover map, were no longer rice fields in 2018. In the same manner, some new rice fields in 2018 were not included in the land cover map.

Inspecting, the parcel boundaries of land cover map it became obvious that they have been subject to serious delineation errors. So much so that did not make sense to correct them. Therefore, a pixel based approach seemed to be the only alternative. Additionally, and given

that this study ultimately aspires to an upscale of its application to a national extent, a pixel based approach makes more sense, as quality parcel boundaries are a rarity.
Now, from a remote sensing point of view, rice classification is not that challenging. It is a binary classification problem for a crop class of distinctive characteristics. Therefore, and as supported by the literature, unsupervised classification algorithms can suffice for such a problem. However, unsupervised classifiers such as k-means are algorithmically exhaustive, implying an immense computational complexity when talking about a pixel based nationwide application. However, the high accuracy results of an unsupervised method for only a small area like Seosan/Dangjin can create quality labels that can then be used to train a supervised classifier. This way, we end up with a trained model that only needs to be applied to the rest of the country to provide rice maps. It is understood that the complexity of a nationwide application is now significantly reduced and the problem is now manageable.

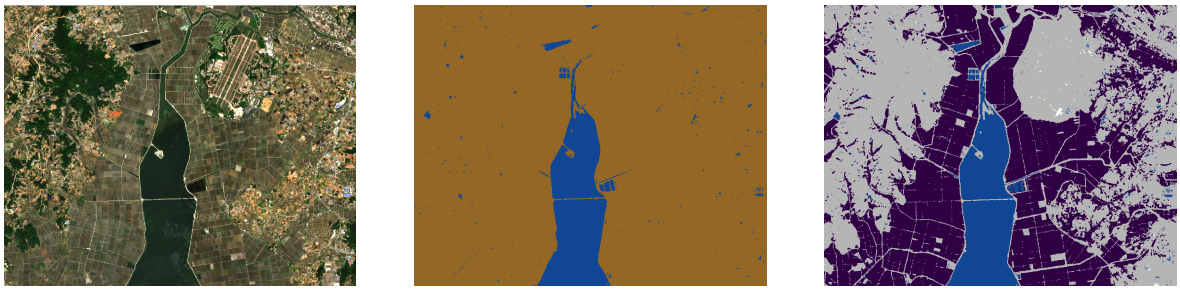
4.3.1 K-means

K-means is one of the simplest and most popular unsupervised machine learning clustering algorithms [59] [60] [61] [62]. The algorithm aims to cluster the data into k categories, where k is user-specified variable. At first, k different clusters are created, using either k randomly generated candidates or generated by sophisticated algorithms as kmeans++ [63]. Each cluster is characterised by a signature of length equal to the number of features, which is called centroid of the cluster. Then, iteratively, every instance is assigned to a cluster based on the minimum distance between the centroid and the features of the instance; the most common metric that is used the Euclidean distance. The algorithm keeps iterating until the assignment of data points to clusters remains unchanged. The outputs of the algorithm are the centroids of the k clusters, as well as the labels for each sample, to be then used as training data. The computational complexity of the algorithm is $O(ndki)$, where n is the number of d-dimensional vectors, k the number of clusters and i the number of maximum iterations.

4.3.2 Pseudo-labeling

The lack of reliable ground truth information in general for crop classification problems, and specifically here the rice classification model, has driven us to the generation of labeled data; using only a limited amount of photo-interpreted ground truth information (section 3.2.5), in order to automatically assign meaningful labels to the k-means cluster. For this reason, an unsupervised approach is introduced using the k-Means algorithm. The proposed methodology resulted in labeling each pixel of the whole image with water, rice or other. As mentioned in section 4.3.1, the computational complexity of k-means is linear with relevant to the number of features, the number of instances, the numbers of iterations and the number of clusters. Since the number of pixels of the area of interest is in the scale of tens of millions, each feature that is used increases linearly the computational complexity, and even for the HPDA environment this can be computational and time consuming. Therefore, only NDVI, NDWI and PSRI

were selected as input data. Furthermore, experiments proved that acquisitions until June did not generate a satisfying result (see section 5.1.2), while acquisitions from March to early July resulted to much more better results. The creation of training data was designed to be applied only to the Seosan-Dangjin region, and thus there was no need of applying the temporal interpolation methodology. Instead, a simple linear interpolation was performed in order to fill the null values from cloudy pixels. Moreover, the pixels of the validation dataset have been excluded from the data fed to the unsupervised method, to avoid any bias. The pipeline begins with an execution of k-means with only two clusters in order to separate land from water pixels. Afterwards, a second-level clustering on the land mask was executed, this time for multiple k values (5-15). For each of the different k-means, a rice cluster emerged and the clustering with the highest precision and recall combination of the rice cluster is selected.



(a) RGB image from 2 of June (b) Water Mask (c) Rice Mask

Figure 4.6: Example of the resulted maps of the clustering methodology

Label assignment on clusters

K-means algorithm results', however, are not labeled as rice or water or whatever the label of interest may be. Consequently, a methodology that assigns a meaningful label to each cluster was necessary. Often, this assignment is achieved through exploratory analysis of the results. On the contrary, in this study an automated technique was approached. For the water label, a small water area was selected in the area of interest and then the mean values for that area's pixels were calculated for each index and each date of the feature space. Likewise, for the rice label means of the validated rice pixels were also estimated. This way, average time-series signatures of water and rice labels were created. Thus, the identification of water and rice categories on each K-Means execution is based on the Mean Squared Error (MSE) that each cluster has against the aforementioned time-series signatures, respectively.

4.4 Supervised Classification

4.4.1 Random Forest

Random Forest is a supervised learning algorithm [64]. Specifically, it is an ensemble classifier which is based on multiple different individual decision trees and each one is trained on a

subset of the original data created using the bootstrap method. Finally, for the classification each decision tree exports a result for the dominant class of the input and predictions are made through a majority voting mechanism of all trees [64]. The number of trees and the number of features used by each tree for a split is defined by the user. A default value is usually the square root of the total number of features, and a such limitation can reduce the computational complexity of the algorithm significantly. As a result, the Random Forest algorithm can handle high dimensional data and use a large number of trees in the ensemble.

4.4.2 Paddy Rice Classification

The Random Forest algorithm was used in this thesis, since it has been efficiently used over the last years for paddy rice mapping, as mentioned in section 2.3. It is important to mention here that climate conditions are not similar across the country of South Korea and thus the cultivation practices differ from place to place. This study aims to create a generalized and geographically transferable model which could be applied over any area of the country. Thus, and since the majority of rice fields are transplanted no later than early-June, the input data passed to the model consists of images with dates from early-June and further. Moreover, to be able to identify paddy rice regions in different times of the year, from July until the end of the cultivation period, multiple models were trained with incrementally larger feature spaces. These incrementally larger feature spaces refer to any new Sentinel-1 and Sentinel-2 acquisitions that accumulate each month and amend the previous set of features. For each one of these models, a grid search hyperparameter optimization of depth and number of trees parameters, which is explained in detail in section 5.1.2, is also applied to determine the best set of parameters for each model. Finally, from the available training data, 20% was used for training the Random Forest algorithm; the pixels of the validation dataset have also excluded from the selection of the training data.

4.5 Implementation

In section 3.5, it was mentioned that for the purposes of this study the Hadoop Distributed File System was used to store and handle the large amount of the available data. However, Hadoop data require a compatible processing engine. Spark is a data processing framework that can perform processing tasks fast on huge data sets, handle distributed files and distribute the data processing tasks across multiple computational nodes. So in this study, PySpark, which is a Python API for Spark, was used to implement the methods and techniques that were described, apart from the retrieval and proccessing of the satellite data which was made in pure Python. Pyspark offers two data structures, the Resilient Distributed Datasets (RDDs) and the DataFrame. An RDD is the fundamental data structure of Spark and represents an immutable distributed collections of data elements, partitioned across nodes. On the other hand, PySpark DataFrame is a distributed collection of data organized into named columns. Conceptually, is equivalent to a data frame in Python but designed to support big data along

with data science applications. As far as Machine Learning in Spark is concerned, there exist two separate libraries, the Spark MLlib which is based on RDDs and the Spark ML which is based on DataFrames. The primary API of Spark is the DataFrame-based API, while the RRD-based one has been in maintenance mode. To that reason, and because DataFrames are more user friendly structures with many capabilities, utilization of them was preferred instead of RDDs. For the algorithms of K-means and Random Forest, the corresponding functions of ML library were used. On the other hand, in PySpark there exists any function or method to perform even a simple linear interpolation. To that reason, the interpolation method was implemented upon DataFrames from scratch. Specifically, three different methods were implemented. The first fills the nan values of each time-series feature, using linear interpolation throughout time. The second conducts the weighted average methodology as described in section 4.2.2, and respectively the third one implements the linear interpolation to generate features for the missing dates, as described also in the same section.

4.6 Quantitative evaluation metrics

In the metrics that are presented below, TP the refers to True Positive instances or the number of correctly classified rice pixels, FP to the False Positive instances or the number of non-rice pixels classified as rice, TN to the True Negative or the number of correctly classified non-rice pixels and FN to the False Negative instances or the number of rice pixels classified as non-rice.

- **Precision**
  Refers to the ratio of correctly classified pixels for a given class to the total number of pixels predicted to belong to that class.

$$Precision = \frac{TP}{TP + FP} \tag{4.7}$$

- **Recall**
  Refers to the ratio of correctly classified pixels over the total number of pixels for a ground truth class.

$$Recall = \frac{TP}{TP + FN} \tag{4.8}$$

- **F1-score**
  Refers to the harmonic mean of precision and recall [65].

$$F_1 score = \frac{2}{2 \cdot TP + FP + FN} \tag{4.9}$$

  or

$$F_1 score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4.10}$$

Chapter 5

Experimental Results

This chapter explains the outcomes of the different experiments performed based on the methods defined in Section 5. This chapter is also divided into two sections. The first evaluates the results of each componenet of the proposed methodology, while the second evaluate the model in terms of computational complexity.

5.1 Performance of proposed methodology

The evaluation of both the psedo-labeling and random forest classification results was assessed by the created validation data, as described in Section 3.2.5.

5.1.1 Evaluating the developed Interpolation method

At first, in order to evaluate the interpolation method on the time-series data, Figure 5.1, 5.2. and 5.3 present the interpolated values along with the original NDVI, NDWI and PSRI values. As we can see, the interpolation is rather successful, not only for filling appropriately any missing values, but also for effectively removing any obvious outliers. Outliers may be present due to unmasked cloudy pixels, as explained in previous sections. This can be particularly seen in the acquisition of the 6th of August for NDVI and NDWI plots. The black dot, which corresponds to the value obtained directly by the bands of S-2, has significantly lower value than we would normally expect, and the interpolated values correct this anomaly.

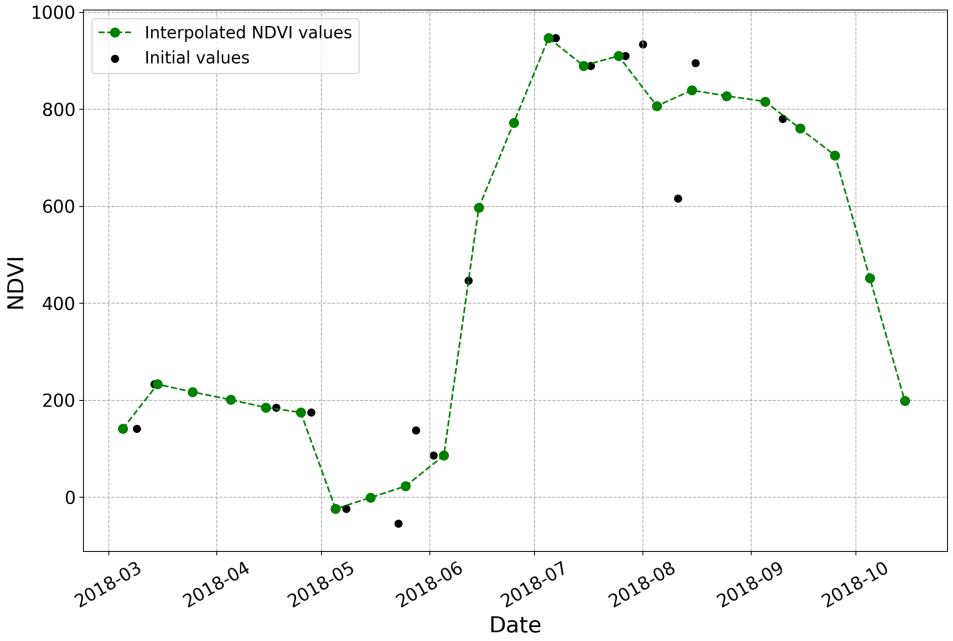


Figure 5.1: Interpolated NDVI values of a random rice pixel.

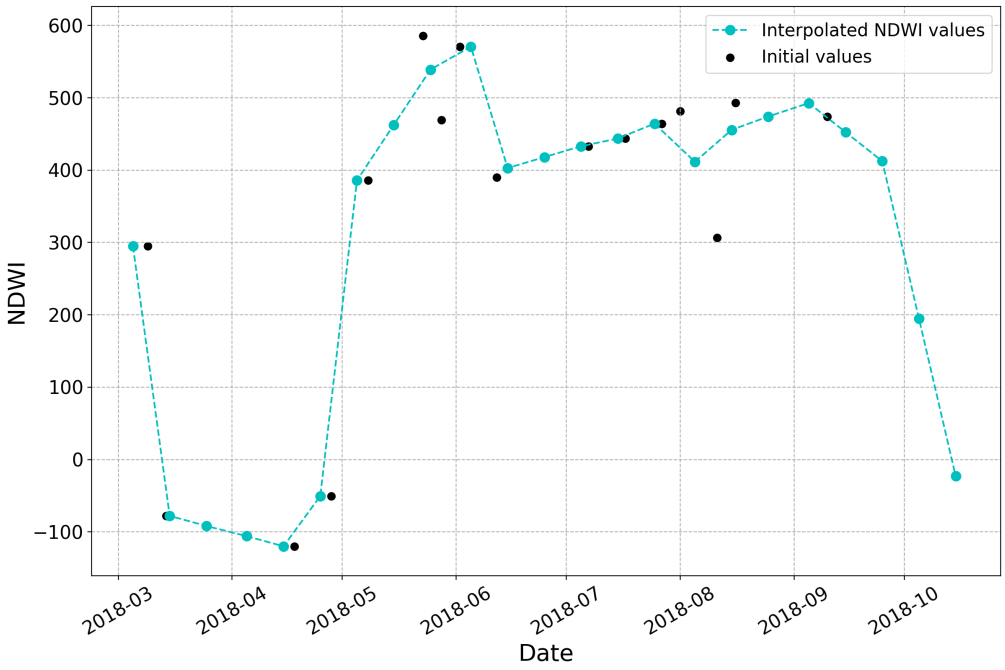


Figure 5.2: Interpolated NDWI values of a random rice pixel.

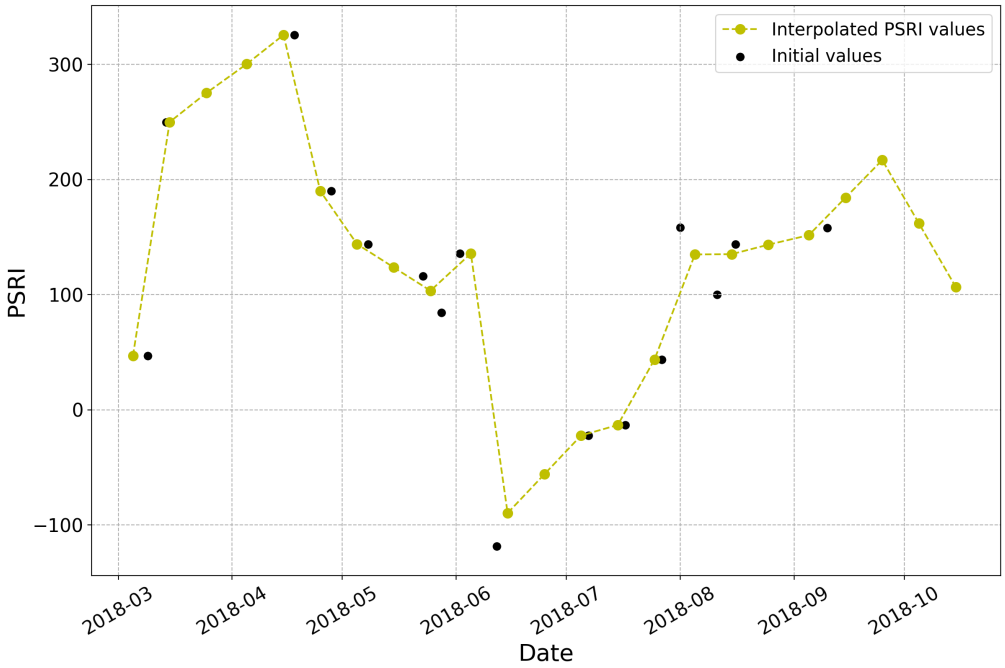


Figure 5.3: Interpolated PSRI values of a random rice pixel.

5.1.2 Pseudo-labeling evaluation

In this section are presented the results of the pseudo-labeling methodology. Table 5.1 presents the results obtained from K-means runs in different time instances throughout the cultivation period. It can be observed that for the first two runs, late-May and late-June the performance, particularly for recall, is less than optimal. In early-July, however, the recall substantially increases and for this reason this run constitutes the training data of choice. In order to have a representative training dataset that incorporates as much of the plurality of the reality, a high recall is necessary.

Rice			
Date	**Precision**	**Recall**	**F1-score**
Late-May	97.07	90.44	93.64
Late-June	98.14	89.26	93.49
Early-July	97.96	92.02	94.90

Table 5.1: Metrics for clustering in Seosan-Dangjin region for different time periods.

Table 5.2 presents the precision, recall and f1-score for every different run of k-means algorithm for the feature space of choice, namely the one that includes acquisitions until early-July. We can see that the best result, seen in Table 5.1, arose from the run with 7 clusters. High recall is necessary in order to have a representative training dataset that includes as many different

paddy rice examples as possible, while high precision is also important since we do want to keep the noise in the created training data in low levels. So, from Table 5.2 the acceptable created labels would be those of 6, 7, 8 and 10 clusters, from which the one with the best recall score (i.e. of 7 clusters) is selected as the training data.

Clusters	**Precision**	**Recall**	**F1-score**
4	79.08	99.68	88.19
5	80.77	98.98	88.95
6	97.80	90.88	94.21
7	97.96	92.02	94.90
8	97.99	91.07	94.40
9	98.48	79.15	87.77
10	98.37	88.19	93.00
11	98.44	82.25	89.62
12	98.57	78.83	87.60
13	98.64	77.58	86.65

Table 5.2: Metrics for all different clusterings in Seosan-Dangjin region for acquisitions until 5 of July.

Finally, Table 5.3 shows the precision, recall and f1-score of the clustering of choice, namely the run of early-July, for both rice and non-rice classes.

Class	**Precision**	**Recall**	**F1-score**	**Num of Clusters**
Non-Rice	93.91%	97.81%	95.82%	7
Rice	97.96%	92.02%	94.90%	7

Table 5.3: Metrics for clustering in Seosan-Dangjin region compared to the validation dataset.

In Figure 5.4, we have the visual representation of the optimal clustering result. It is apparent that even this unsupervised classification method represent the reality rather closely.

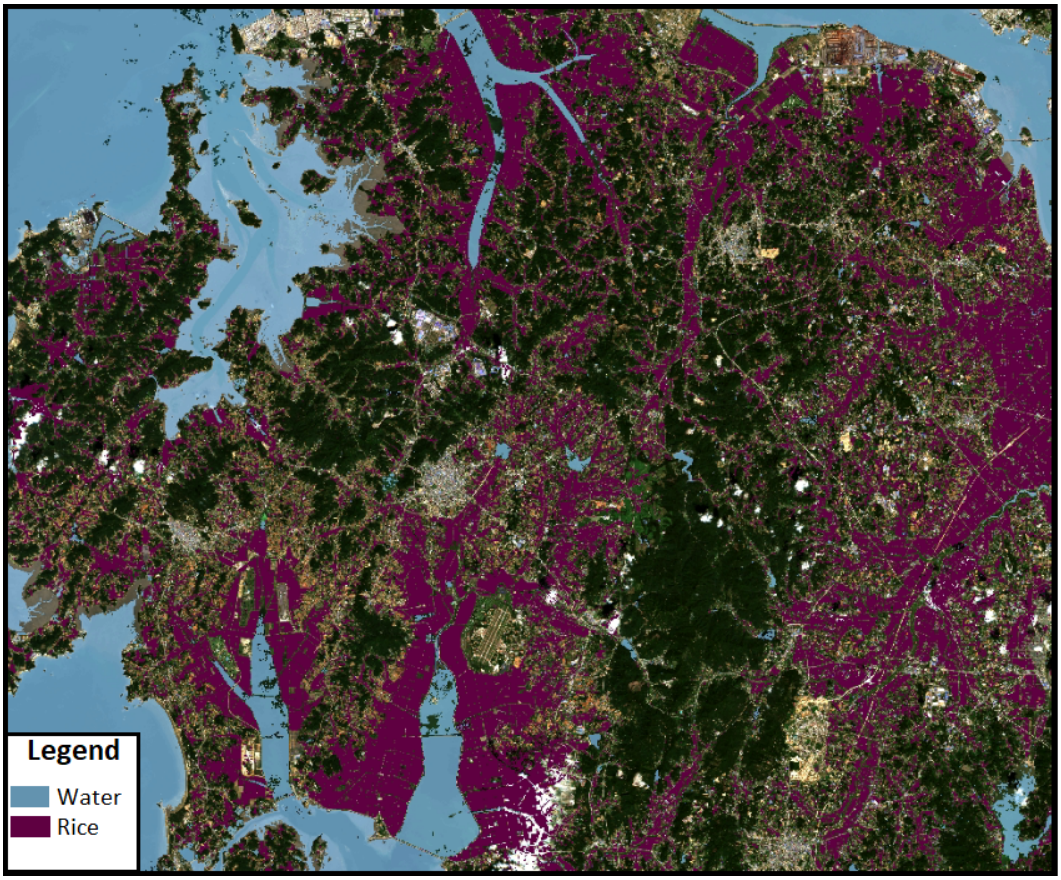


Figure 5.4: Result of the clustering methodology upon an RGB image of the area of interest. Rice is represented with purple color while water with blue.

In Figures 5.5 and 5.6, we present indicative sub-regions of the study area in order to further evaluate qualitatively the resulted map. In particular, cases that the developed model fail to perform optimally and lead to misclassification errors are presented. In Figure 5.5 the algorithm misclassifies water to rice. This is actually the most common mistake that has been noticed. Respectively, in Figure 5.6 a similar case is presented in which the clustering methodology does not manage to capture all the rice areas completely.Such errors are usually observed in rice parcels located in the mountains and having irregular shapes, which consequently leads to more noisy spectral signatures satellite data observations.

(a) Zoomed RGB image (2nd of June) (b) Misclassification

Figure 5.5: An example of over-estimation for the rice class in a Seosan sub-region.

(a) Zoomed RGB image (2nd of June) (b) Misclassification

Figure 5.6: An example of under-estimation of the rice class, on the 2nd of June in a Seosan sub-region. Blue color represents the water, purple the rice from the clustering and green the rice of the validation dataset

5.1.3 RF Hyperparameter Optimization

The High Perfomance Data Analytics (HPDA) offers the potential to train a model in much shorter time than a conventional machine. Therefore it is possible to train the Random Forest model with multiple parameterizations in a efficient amount of time. Thus, different models were trained with multiple combinations of depth and number of trees parameters and with input the interpolated data from June to July. Specifically, the range of values for depth was from 3 to 20 while for the number of trees parameter the values 15, 35, 50, 100 and 400 were tried. As Figure 5.7 implies, as depth parameter is increased, the performance of the model is lessened, while the tree parameter does not have too much effect on the accuracy of the model. The difference of the models is not significant in terms of absolute values (only 1% decrease), but since the input image consists of 42 million pixels, 1% reduction

could be translated in loss of valuable information. In general, it is understood, that the random forest algorithm using pseudo-labeled training data can accurately map paddy rice pixels.

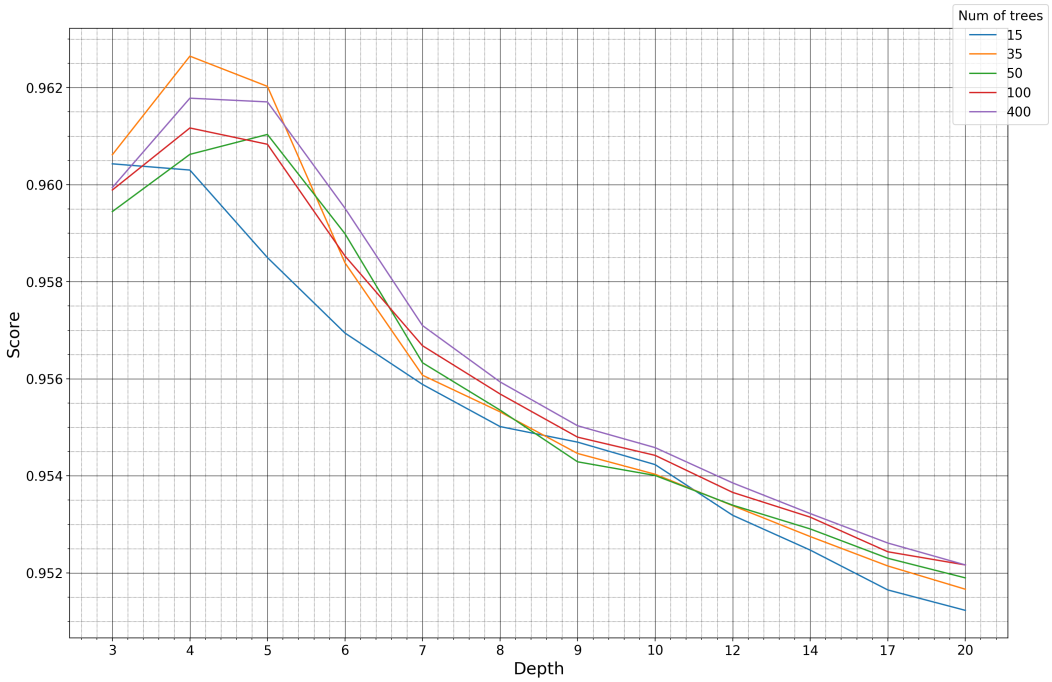


Figure 5.7: F1 score for varying values of the parameters of the number of trees and depth.

Figures 5.8 and 5.9 present precision and recall scores for the models that achieve the highest scores. The parameterization that resulted to the highest combined score, of both recall and precision, is 4 for the depth and 35 for the number of trees parameter. Based on 5.7, the recall of 35 trees and depth of 4 is by far the best. On the other hand, the precision scores for the different depths and trees, as shown in Figure 5.6, are comparable. For this reason, the aforementioned of the two parameters is considered to be the optimal one.

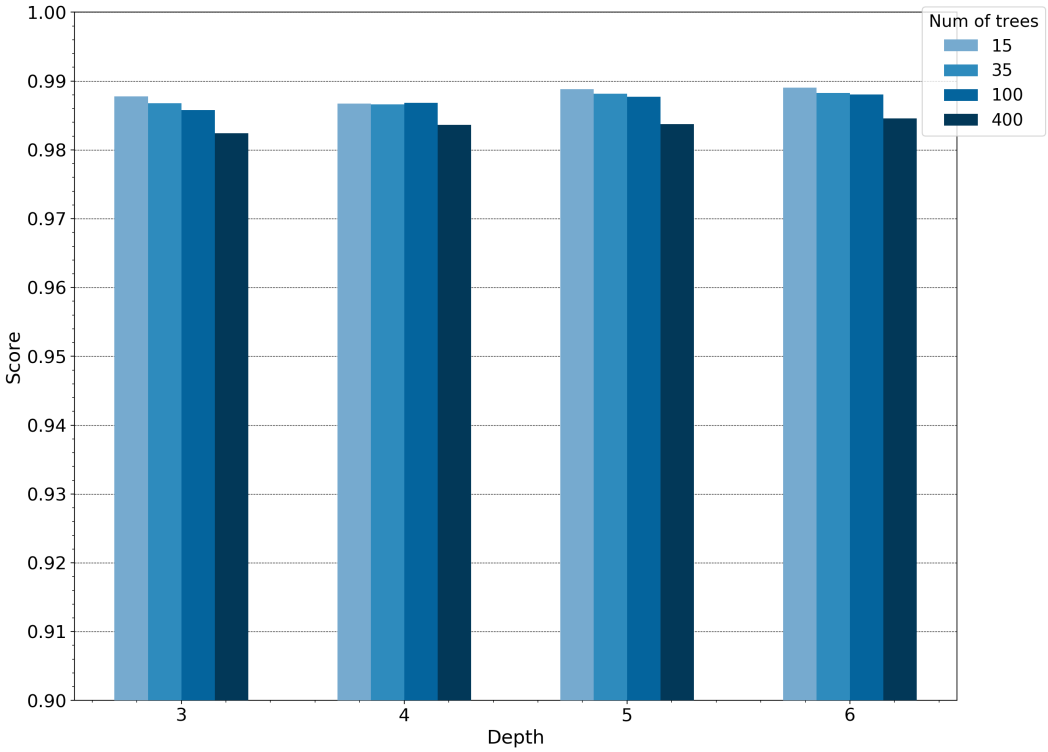


Figure 5.8: Random forest precision of varying number of trees and depth.

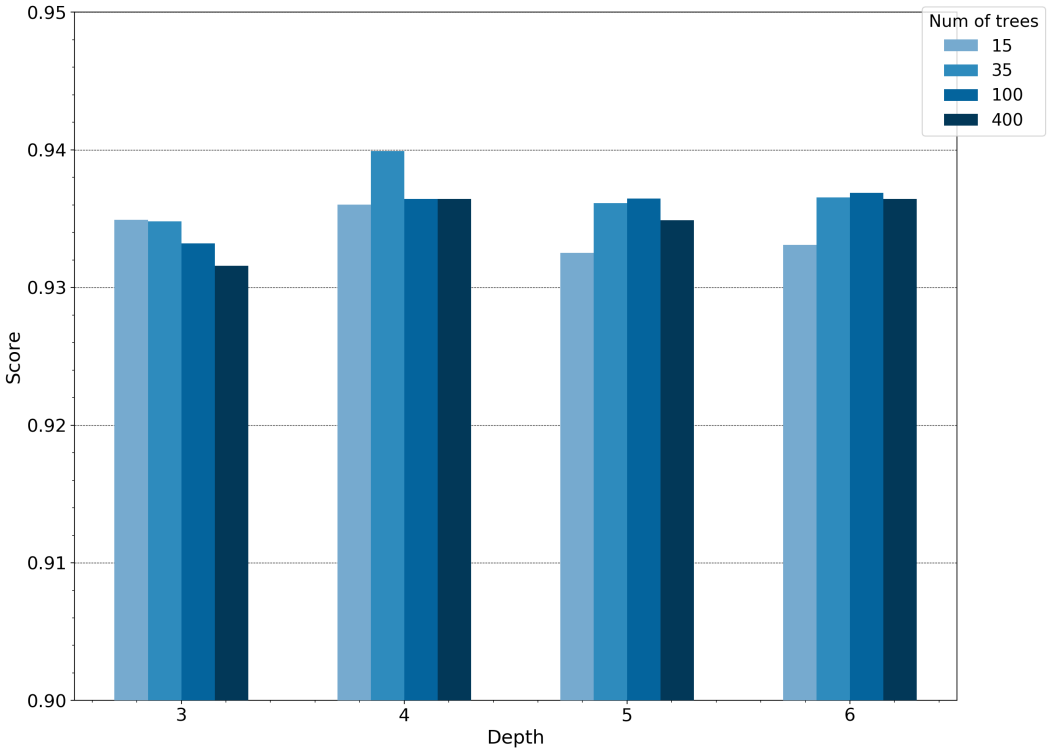


Figure 5.9: Random forest recall of varying number of trees and depth.

Table 5.4, present the precision, recall and f1-score of both rice and non-rice classes for the parameterization of choice.

Class	Precision	Recall	F1-score
Non-Rice	95.36%	99.00%	97.14%
Rice	98.64%	93.79%	96.15%

Table 5.4: Precision, Recall and F1-score for rice and non-rice classes, for the final RF model.

In Figures 5.10 and 5.11, we can observe the evolution of the performance of rice classification. In Figure 5.10, we see the evolution of f1-score throughout time, while in Figure 5.11 we see the visual representation of those results. Looking at 5.11, it is seen that even from July on wards near-optimal results are achieved. This is particularly important in a study like this, working towards food security monitoring where timely and accurate information is of the essence.

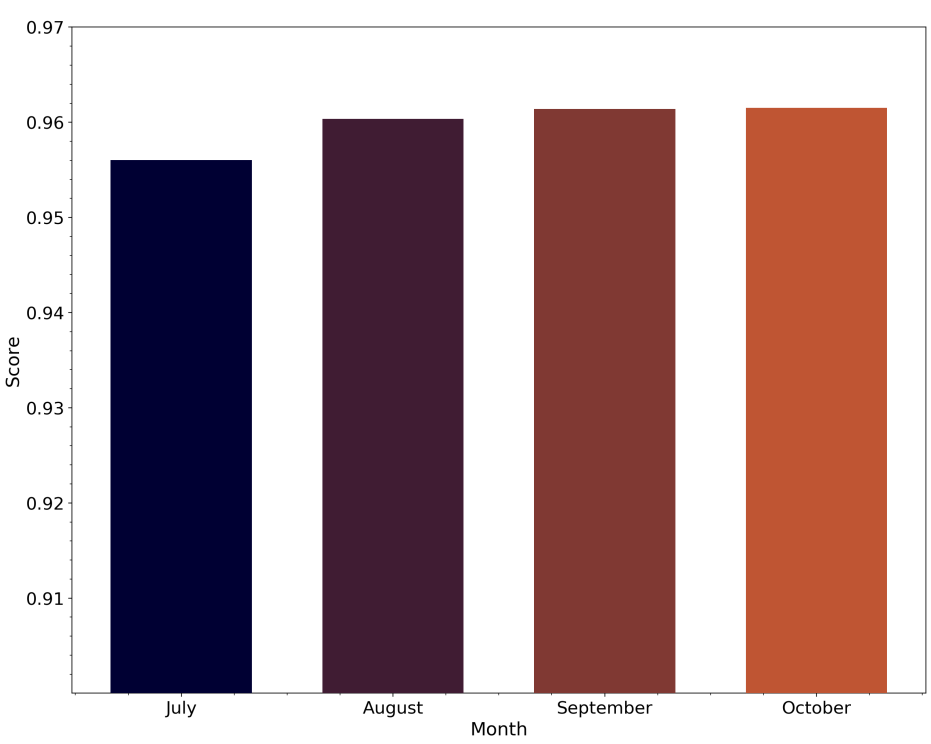


Figure 5.10: Evolution of random forest f1-scores throughout time.

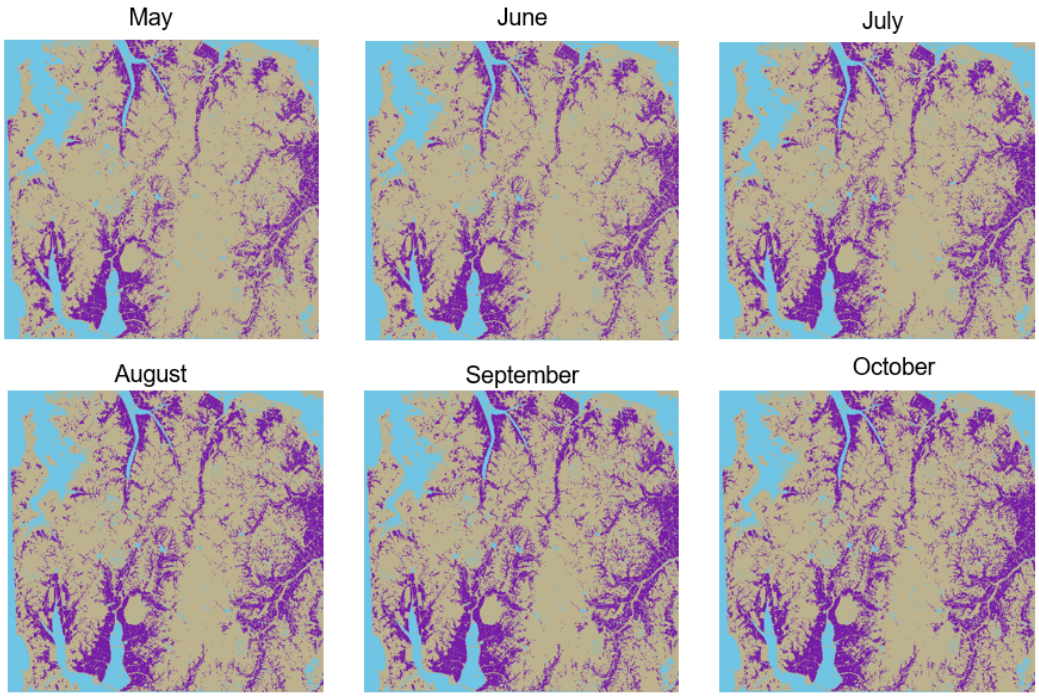


Figure 5.11: Evolution map of random forest results

5.1.4 Feature importance

From the results above, it is obvious that RF algorithm successfully copes with the problem of paddy rice mapping. In order to understand the significance of each feature to the decisions of the RF algorithm, the features importances are presented in this section. At first, the importance of each feature was calculated using the corresponding method of the PySpark random forest model, and then they were grouped per feature (Figure 5.14) and per date (Figure 5.13). Based on Figure 5.12, the most important features comprise of the NRI and NDVI of the acquisitions of June 5 and 25 and July 5 and 15. Thus, becomes apparent that NIR observations are really important in differentiating rice against anything else. As we see the 4 most important features come from acquisitions as early as the 5 of July, which is supported by the satisfactory results early in the year as seen in Figure 5.11, However, the NDVI of July 15 is 5th in rank, explaining the jump in near optimal performance at the end of July. This can be additionally supported by the aggregated importances of each date of the month in Figure 5.13.

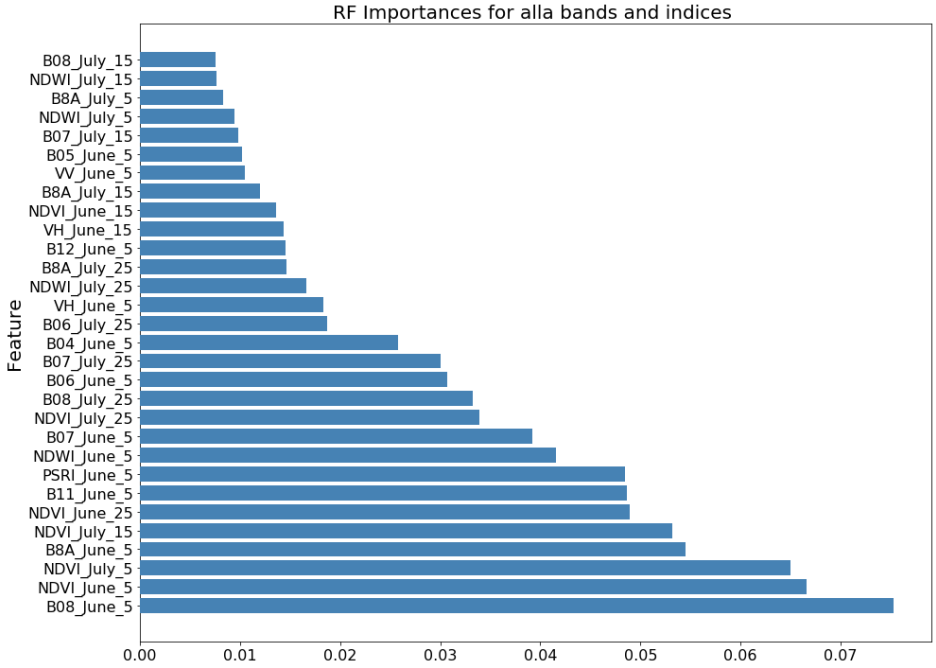


Figure 5.12: The top 20 random forest feature importances

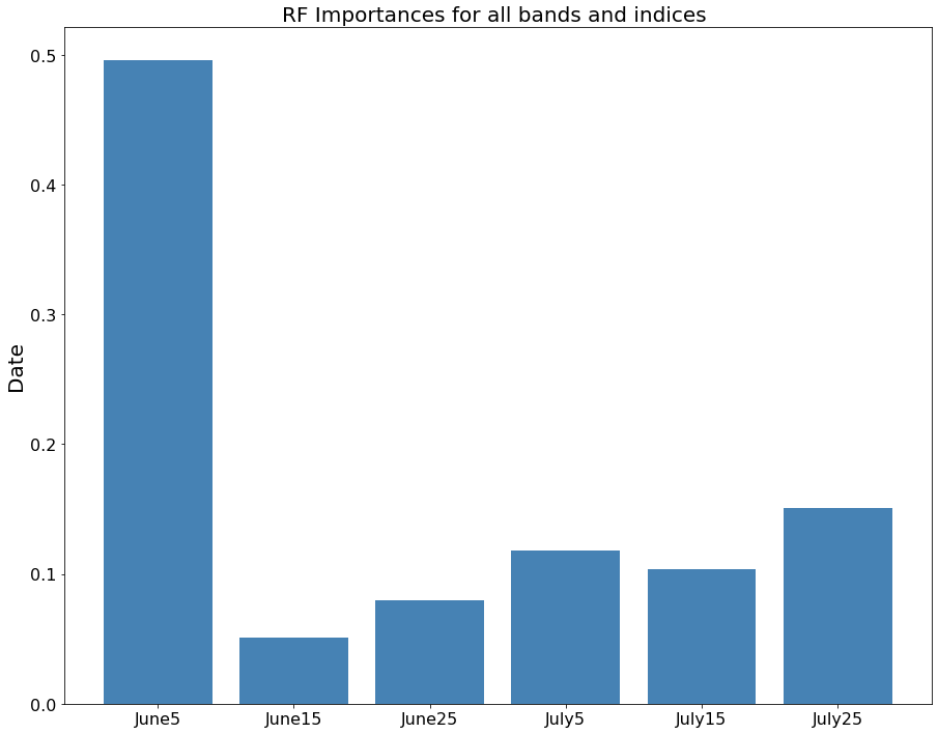


Figure 5.13: Random forest feature importances aggregated by acquisition date.

As expected, inspecting 5.14, NDVI is by far the most important feature with respect to the spectral information. The same holds true, for the NIR band B08. Other important spectral

features include the Red-Edge bands B06, B07 and B8A. Finally, NDWI is also another important feature, which can be explained by the importance of water content sensitive indices for the classification of an inundated cultivation, such as rice.

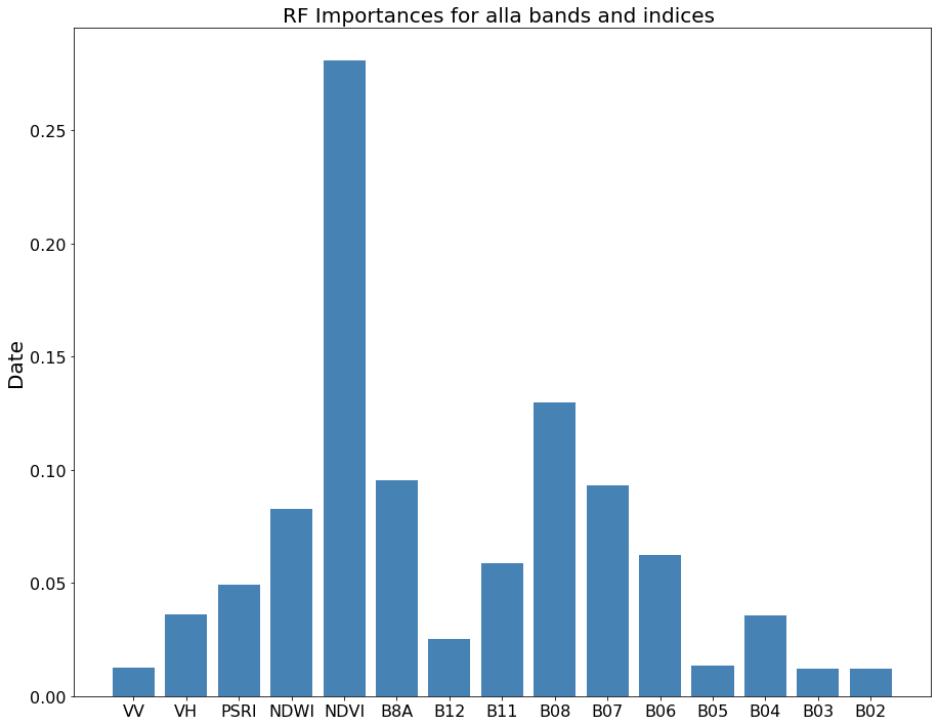


Figure 5.14: Random forest feature importances aggregated by acquisition type of feature (bands, indices, back-scatters)

5.2 Computational Complexity

Another major achievement of this study is the scalability of the underlying methodology. The processing is done at a per pixel basis, which results in very large feature spaces, even for small areas, such as the one that was presented indicatively in this study. Even a single region like Dangjin/Seosan can amount to more than a 100 of GB, which requires to be batch processed. Therefore, the need for distributed filing systems and distributed machine learning is a given, both for the complexity of the problem even in small scales and of course for its upscale to national applications. In this section, is presented the computational complexity of the supervised classification procedure. Specifically, Figure 5.15 presents the total time of training process, which includes reading files from HDFS, training the Distributed Random Forest with the best combination of parameters (see Section 5.1.2) and saving the results back to HDFS. It obvious that the total time is reduced exponential as more system nodes are used for the processing, which reaches a plateau at the level of 10 nodes. Moreover, moving from 1 node to 10, we can observe a decrease of over 5 times.

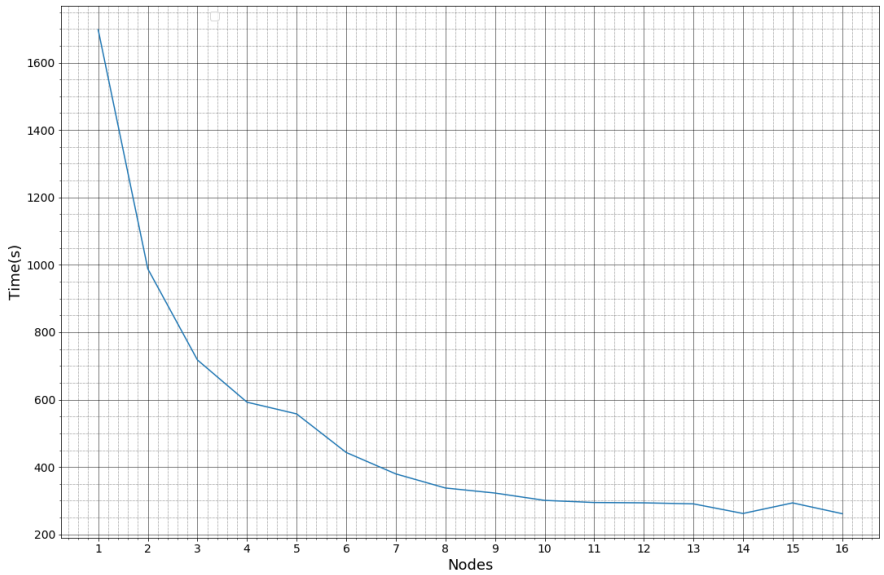


Figure 5.15: Computational complexity of training the Random Forest model, using varying number of system nodes.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The overall objective of this thesis was to develop a transferable model for paddy rice crop mapping using multitemporal Senintel-1 and Sentinel-2 images, while utilizing a High Performance Data Analytics environment together with the Hadoop Distributed File System (HDFS) and the PySpark processing tool. The feasibility of this work was evaluated through a case study in the region of Seosan and Dangjin cities of South Korea.

Scarcity of ground truth data is one of the most common problems in remote sensing. For example, in South Korea, and specifically in the area of study, the existing rice map is of the year 2015 and with several mislabeled pixels. In this work, as described in section 3.2.5, a representative validation dataset was created in order to assess the results of the rice classification as reliably as possible.

Moreover, using a small amount of ground truth information and an unsupervised technique, trustworthy training data for paddy rice crop was successfully created for the region of Seosan-Dangjin in South Korea. The results indicated that to generate sufficiently accurate training data, acquisitions from March until the start of July are much better than earlier dates. Therefore, we can conclude that the unsupervised methodology can be used to successfully generate labeled data, using only a small amount of ground truth information to create spectral signatures of the rice crop, in order to correctly indicate the rice label.

Another important contribution of this thesis is the creation of large-scale feature spaces with fixed time steps, in order to train and create transferable models for pixel-based paddy rice classification. With the proposed methodology, close-to-reality images were generated for the 5th, 15th and 25th of each month, building that way a robust feature space. Moreover, possible inaccurate values of pixels in specific acquisitions (e.g. cloudy pixels that were not masked from sen2cor software), could be corrected through the weighed average interpolation method, as shown in section 5.1.

The results also demonstrated that the distributed Random Forest in the HPDA environment is appropriate for large scale pixel-based paddy rice classification using time-series of big EO

data. Multiple combinations of the depth and number of trees parameters of the Random Forest algorithm were tested, which indicated that low values at the depth parameter showcase better performance, while the best scores were achieved with the RF model of 35 different trees of 4 depth.
Finally, section 5.2 indicated the significance of utilizing a High Performance Data Analytics environment. Training the Random Forest model in multiple nodes can reduce the computational complexity more than 5 times, compared to a single node.
To summarize, taking everything mentioned above into consideration, we can conclude that the regional application presented in this thesis, offers a lot of potential for further study, which could result to a scaling up for the entire country, providing a useful indicator about the rice extend.

6.2 Discussion and Future Work

The results presented in section 5, showcased the potential of distributed methods and algorithms, using data derived from Sentinel-1 and Sentinel-2 imagery, in a High Performance Data Analytics environment, to successfully map paddy rice across a large region in South Korea. Useful baseline approaches, with the methodological potential of upscaling, were developed, that have been, nonetheless, applied to a single area of interest. The importance of the results, however, can be extrapolated to the envisaged large scale application of those methods, highlighting their impact to real case food security monitoring scenarios.
The developed interpolation method resulted in a representative and also improved, new feature space with fixed timestamps. Therefore, it is feasible to easily transfer any model trained in a specific region and time, in different ones across the country; for the same or other years. Since this study was limited to the Seosan-Danjin area of South Korea, the following step would be to apply the final Random Forest model of section 5.1.3 to other regions of South Korea, in different geographic areas, with different climatic conditions, in order to evaluate its generalization capability. This can be done also for multiple years. [54]
Furthermore, the outcomes of section 5.1.2 show that the automated unsupervised technique generated accurate and trustworthy labeled data, for paddy rice, in the area of interest. As a next step, we could test this methodology to other crops similar to rice, that have unique spectral signatures, or even rice, in areas that there are no reference data. Producing a really small amount of reference data at such areas, using photo-interpretation, can be enough to successfully create training data using the proposed unsupervised methodology. Moreover, it would be very interesting to expand this concept to a multi-class problem. This could be achieved by conducted multiple different experiments for each one of the different classes, in a one-vs-all fusion, creating by this way labels for each one them. Subsequently, by merging the results of each experiment, a training dataset for multi-class problems would be generated.

Bibliography

[1] Marijn van der Velde, Linda See, Liangzhi You, Juraj Balkovič, Steffen Fritz, Nikolay Khabarov, Michael Obersteiner, and Stanley Wood. «Affordable nutrient solutions for improved food security as evidenced by crop trials». In: *PloS one* 8.4 (2013), e60075.

[2] Ricepedia. *Growth phases*. 2015. URL: `http://ricepedia.org/rice-as-a-plant/growth-phases`.

[3] Andrew Nelson, Tri Setiyono, Arnel B Rala, Emma D Quicho, Jeny V Raviz, Prosperidad J Abonete, Aileen A Maunahan, Cornelia A Garcia, Hannah Zarah M Bhatti, Lorena S Villano, et al. «Towards an operational SAR-based rice monitoring system in Asia: Examples from 13 demonstration sites across Asia in the RIICE project». In: *Remote Sensing* 6.11 (2014), pp. 10773–10812.

[4] C. F. Curtis Barrett E. C. *Introduction to Environmental Remote Sensing*. Macmillan.

[5] PM Teillet, RP Gauthier, A Chichagov, and G Fedosejevs. «Towards integrated earth sensing: The role of in situ sensing». In: *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences* 34.1 (2002), pp. 249–254.

[6] Mostafa Mosleh. «Use of GIS and Remote Sensing in Mapping Rice Areas and Forecasting Its Production at Large Geographical Extent». PhD thesis. University of Calgary, Calgary, 2015.

[7] EU SCIENCE HUB. *Earth observation*. 2016. URL: `https://ec.europa.eu/jrc/en/research-topic/earth-observation`.

[8] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. «GMES Sentinel-1 mission». In: *Remote Sensing of Environment* 120 (2012), pp. 9–24.

[9] Jinwei Dong and Xiangming Xiao. «Evolution of regional to global paddy rice mapping methods: A review». In: *ISPRS Journal of Photogrammetry and Remote Sensing* 119 (2016), pp. 214–227.

[10] KR McCloy, FR Smith, and MR Robinson. «Monitoring rice areas using Landsat MSS data». In: *International Journal of remote sensing* 8.5 (1987), pp. 741–749.

[11] SB Tennakoon, VVN Murty, and A Eiumnoh. «Estimation of cropped area and grain yield of rice using remote sensing data». In: *International Journal of Remote Sensing* 13.3 (1992), pp. 427–439.

[12] Hongling Fang. «Rice crop area estimation of an administrative division in China using remote sensing data». In: *International Journal of Remote Sensing* 19.17 (1998), pp. 3411–3419.

[13] Yun Shao, Xiangtao Fan, Hao Liu, Jianhua Xiao, S Ross, B1 Brisco, R Brown, and G Staples. «Rice monitoring and production estimation using multitemporal RADARSAT». In: *Remote sensing of Environment* 76.3 (2001), pp. 310–325.

[14] C Chen and H McNairn. «A neural network integrated approach for rice crop monitoring». In: *International Journal of Remote Sensing* 27.7 (2006), pp. 1367–1393.

[15] Yuan Zhang, Cuizhen Wang, Jiaping Wu, Jiaguo Qi, and William A Salas. «Mapping paddy rice with multitemporal ALOS/PALSAR imagery in southeast China». In: *International journal of Remote sensing* 30.23 (2009), pp. 6301–6315.

[16] Chue-Poh Tan, Jun-Yi Koay, Ka-Sing Lim, Hong-Tat Ewe, and Hean-Teik Chuah. «Classification of multi-temporal SAR images for rice crops using combined entropy decomposition and support vector machine technique». In: *Progress In Electromagnetics Research* 71 (2007), pp. 19–39.

[17] Nguyen-Thanh Son, Chi-Farn Chen, Cheng-Ru Chen, Huynh-Ngoc Duc, and Ly-Yu Chang. «A phenology-based classification of time-series MODIS data for rice crop monitoring in Mekong Delta, Vietnam». In: *Remote Sensing* 6.1 (2014), pp. 135–156.

[18] Alfredo Huete, Kamel Didan, Tomoaki Miura, E Patricia Rodriguez, Xiang Gao, and Laerte G Ferreira. «Overview of the radiometric and biophysical performance of the MODIS vegetation indices». In: *Remote sensing of environment* 83.1-2 (2002), pp. 195–213.

[19] Jinsong Chen, Jianxi Huang, and Jinxing Hu. «Mapping rice planting areas in southern China using the China Environment Satellite data». In: *Mathematical and Computer Modelling* 54.3-4 (2011), pp. 1037–1043.

[20] Alexandre Bouvet and Thuy Le Toan. «Use of ENVISAT/ASAR wide-swath data for timely rice fields mapping in the Mekong River Delta». In: *Remote Sensing of Environment* 115.4 (2011), pp. 1090–1101.

[21] Duy Ba Nguyen, Kersten Clauss, Senmao Cao, Vahid Naeimi, Claudia Kuenzer, and Wolfgang Wagner. «Mapping rice seasonality in the Mekong Delta with multi-year Envisat ASAR WSM data». In: *Remote Sensing* 7.12 (2015), pp. 15868–15893.

[22] Alexandre Bouvet, Thuy Le Toan, and Nguyen Lam-Dao. «Monitoring of the rice cropping system in the Mekong Delta using ENVISAT/ASAR dual polarization data». In: *IEEE transactions on geoscience and remote sensing* 47.2 (2009), pp. 517–526.

[23] Xiangming Xiao, Stephen Boles, Jiyuan Liu, Steve Zhuang Dafang an d Frolking, Changsheng Li, William Salas, and Berrien Moore III. «Mapping paddy rice agriculture in southern China using multi-temporal MODIS images». In: *Remote sensing of environment* 95.4 (2005), pp. 480–492.

[24] Xiangming Xiao, Stephen Boles, Steve Frolking, Changsheng Li, Jagadeesh Y Babu, William Salas, and Berrien Moore III. «Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images». In: *Remote Sensing of Environment* 100.1 (2006), pp. 95–113.

[25] Jinwei Dong, Xiangming Xiao, Michael A Menarguez, Geli Zhang, Yuanwei Qin, David Thau, Chandrashekhar Biradar, and Berrien Moore III. «Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine». In: *Remote sensing of environment* 185 (2016), pp. 142–154.

[26] Thi Hoa Phan. «Rice monitoring using radar remote sensing». PhD thesis. Université Paul Sabatier - Toulouse III, Hydrology, 2018.

[27] Duy Ba Nguyen, Alexander Gruber, and Wolfgang Wagner. «Mapping rice extent and cropping scheme in the Mekong Delta using Sentinel-1A data». In: *Remote Sensing Letters* 7.12 (2016), pp. 1209–1218.

[28] Lamin R Mansaray, Weijiao Huang, Dongdong Zhang, Jingfeng Huang, and Jun Li. «Mapping rice fields in urban Shanghai, southeast China, using Sentinel-1A and Landsat 8 datasets». In: *Remote Sensing* 9.3 (2017), p. 257.

[29] Nathan Torbick, Diya Chowdhury, William Salas, and Jiaguo Qi. «Monitoring rice agriculture across myanmar using time series Sentinel-1 assisted by Landsat-8 and PALSAR-2». In: *Remote Sensing* 9.2 (2017), p. 119.

[30] Alex O Onojeghuo, George A Blackburn, Qunming Wang, Peter M Atkinson, Daniel Kindred, and Yuxin Miao. «Mapping paddy rice fields by applying machine learning algorithms to multi-temporal Sentinel-1A and Landsat data». In: *International journal of remote sensing* 39.4 (2018), pp. 1042–1067.

[31] Yaotong Cai, Hui Lin, and Meng Zhang. «Mapping paddy rice by the object-based random forest method using time series Sentinel-1/Sentinel-2 data». In: *Advances in Space Research* 64.11 (2019), pp. 2233–2244.

[32] Nguyen-Thanh Son, Chi-Farn Chen, Cheng-Ru Chen, and Horng-Yuh Guo. «Classification of multitemporal Sentinel-2 data for field-level monitoring of rice cropping practices in Taiwan». In: *Advances in Space Research* (2020).

[33] Weichun Zhang, Hongbin Liu, Wei Wu, Linqing Zhan, and Jing Wei. «Mapping Rice Paddy Based on Machine Learning with Sentinel-2 Multi-Temporal Data: Model Comparison and Transferability». In: *Remote Sensing* 12.10 (2020), p. 1620.

[34] Luo Liu, Xiangming Xiao, Yuanwei Qin, Jie Wang, Xinliang Xu, Yueming Hu, and Zhi Qiao. «Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine». In: *Remote Sensing of Environment* 239 (2020), p. 111624.

[35] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. «Google Earth Engine: Planetary-scale geospatial analysis for everyone». In: *Remote sensing of Environment* 202 (2017), pp. 18–27.

[36] Yu Hsin Tsai, Douglas Stow, Hsiang Ling Chen, Rebecca Lewison, Li An, and Lei Shi. «Mapping vegetation and land use types in Fanjingshan National Nature Reserve using Google Earth engine». In: *Remote Sensing* 10.6 (2018), p. 927.

[37] Nathaniel P Robinson, Brady W Allred, Matthew O Jones, Alvaro Moreno, John S Kimball, David E Naugle, Tyler A Erickson, and Andrew D Richardson. «A dynamic Landsat derived normalized difference vegetation index (NDVI) product for the conterminous United States». In: *Remote Sensing* 9.8 (2017), p. 863.

[38] Dimosthenis Traganos, Bharat Aggarwal, Dimitris Poursanidis, Konstantinos Topouzelis, Nektarios Chrysoulakis, and Peter Reinartz. «Towards global-scale seagrass mapping and monitoring using Sentinel-2 on Google Earth Engine: The case study of the aegean and ionian seas». In: *Remote Sensing* 10.8 (2018), p. 1227.

[39] Andrii Shelestov, Mykola Lavreniuk, Nataliia Kussul, Alexei Novikov, and Sergii Skakun. «Exploring Google Earth Engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping». In: *frontiers in Earth Science* 5 (2017), p. 17.

[40] Jun Xiong, Prasad S Thenkabail, Murali K Gumma, Pardhasaradhi Teluguntla, Justin Poehnelt, Russell G Congalton, Kamini Yadav, and David Thau. «Automated cropland mapping of continental Africa using Google Earth Engine cloud computing». In: *ISPRS Journal of Photogrammetry and Remote Sensing* 126 (2017), pp. 225–244.

[41] Masoud Mahdianpari, Bahram Salehi, Fariba Mohammadimanesh, Saeid Homayouni, and Eric Gill. «The first wetland inventory map of newfoundland at a spatial resolution of 10 m using sentinel-1 and sentinel-2 data on the google earth engine cloud computing platform». In: *Remote Sensing* 11.1 (2019), p. 43.

[42] Onisimo Mutanga and Lalit Kumar. *Google Earth Engine Applications*. 2019.

[43] Dipankar Mandal, Vineet Kumar, Avik Bhattacharya, Yalamanchili Subrahmanyeswara Rao, Paul Siqueira, and Soumen Bera. «Sen4Rice: A processing chain for differentiating early and late transplanted rice using time-series Sentinel-1 SAR data with Google Earth engine». In: *IEEE Geoscience and Remote Sensing Letters* 15.12 (2018), pp. 1947–1951.

[44] Xin Zhang, Bingfang Wu, Guillermo E Ponce-Campos, Miao Zhang, Sheng Chang, and Fuyou Tian. «Mapping up-to-date paddy rice extent at 10 m resolution in China through the integration of optical and synthetic aperture radar images». In: *Remote Sensing* 10.8 (2018), p. 1200.

[45] José A Navarro. «First Experiences with Google Earth Engine.» In: *GISTAM*. 2017, pp. 250–255.

[46] Jia Yu, Jinxuan Wu, and Mohamed Sarwat. «Geospark: A cluster computing framework for processing large-scale spatial data». In: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2015, pp. 1–4.

[47] Randall T Whitman, Michael B Park, Sarah M Ambrose, and Erik G Hoel. «Spatial indexing and analytics on Hadoop». In: *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2014, pp. 73–82.

[48] Korean Statistical Information Service. *Statistical Database, Cultivated Land Area Paddy Field and Upland by Si(City) and Gun(County)*. 2019. URL: `https://kosis.kr/eng/index/index.do`.

[49] S. Park, J. Im, S. Park, C. Yoo, H. Han, and J. Rhee. «Classification and Mapping of Paddy Rice by Combining Landsat and SAR Time Series Data». In: *Remote Sensing* (2018).

[50] Uwe Muller-Wilm, Jerome Louis, Rudolf Richter, Ferran Gascon, and Marc Niezette. «Sentinel-2 level 2A prototype processor: Architecture, algorithms and first results». In: *Proceedings of the ESA Living Planet Symposium, Edinburgh, UK*. 2013, pp. 9–13.

[51] Hewlett Packard Enterprise. *Urika®-GX Analytic Applications Guide - S-3015*. URL: `https://https://pubs.cray.com/`.

[52] Apache Hadoop. *HDFS Architecture*. URL: `https://hadoop.apache.org/`.

[53] Hyun-Woo Jo, Sujong Lee, Eunbeen Park, Chul-Hee Lim, Cholho Song, Halim Lee, Youngjin Ko, Sungeun Cha, Hoonjoo Yoon, and Woo-Kyun Lee. «Deep Learning Applications on Multitemporal SAR (Sentinel-1) Image Classification Using Confined Labeled Data: The Case of Detecting Rice Paddy in South Korea». In: *IEEE Transactions on Geoscience and Remote Sensing* (2020).

[54] Vasileios Sitokonstantinou, Thanassis Drivas, Alkiviadis Koukos, Ioannis Papoutsis, and Charalampos Kontoes. «SCALABLE DISTRIBUTED RANDOM FOREST CLASSIFICATION FOR PADDY RICE MAPPING». In: *The 40th Asian Conference on Remote Sensing (ACRS 2019)*. 2019.

[55] Vasileios Sitokonstantinou, Antonios Koutroumpas, Thanassis Drivas, Alkiviadis Koukos, Vassilia Karathanassi, Haris Kontoes, and Ioannis Papoutsis. «A Sentinel based agriculture monitoring scheme for the control of the CAP and food security». In: *Eighth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2020)*. Vol. 11524. International Society for Optics and Photonics. 2020, p. 1152407.

[56] Vasileios Sitokonstantinou, Ioannis Papoutsis, Charalampos Kontoes, Alberto Lafarga Arnal, Ana Pilar Armesto Andrés, and José Angel Garraza Zurbano. «Scalable parcel-based crop identification scheme using sentinel-2 data time-series for the monitoring of the common agricultural policy». In: *Remote Sensing* 10.6 (2018), p. 911.

[57] Andrés Viña, Anatoly A Gitelson, Anthony L Nguy-Robertson, and Yi Peng. «Comparison of different vegetation indices for the remote assessment of green leaf area index of crops». In: *Remote Sensing of Environment* 115.12 (2011), pp. 3468–3478.

[58] Mark N Merzlyak, Anatoly A Gitelson, Olga B Chivkunova, and Victor YU Rakitin. «Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening». In: *Physiologia plantarum* 106.1 (1999), pp. 135–141.

[59] Stuart Lloyd. «Least squares quantization in PCM». In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[60] Hugo Steinhaus. «Sur la division des corp materiels en parties». In: *Bull. Acad. Polon. Sci* 1.804 (1956), p. 801.

[61] Geoffrey H Ball and David J Hall. *ISODATA, a novel method of data analysis and pattern classification*. Tech. rep. Stanford research inst Menlo Park CA, 1965.

[62] James MacQueen et al. «Some methods for classification and analysis of multivariate observations». In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[63] David Arthur and Sergei Vassilvitskii. *k-means++: The Advantages of Careful Seeding*. Technical Report 2006-13. Stanford InfoLab, June 2006. URL: `http://ilpubs.stanford.edu:8090/778/`.

[64] Leo Breiman. «Random forests». In: *Machine learning* 45.1 (2001), pp. 5–32.

[65] Marina Sokolova and Guy Lapalme. «A systematic analysis of performance measures for classification tasks». In: *Information processing & management* 45.4 (2009), pp. 427–437.