

Tunki Project: Wetland Remote Sensing using Satellite Images and High Performance Computing

Alvaro Mamani Aliaga, PhD(c)¹, Yessenia Yari Ramos, PhD(c)^{1,2}, Ronald Apaza Veliz, MSc¹ and Pablo Yanyachi Aco Cardenas, PhD¹
¹Instituto Astronómico y Aeroespacial Pedro Paulet, Universidad Nacional de San Agustín, Arequipa-Perú, {amamaliani, yyari, rapazav, raulpab}@unsa.edu.pe

²Universidad Católica San Pablo, Arequipa-Perú, {ydyari}@ucsp.edu.pe

Abstract- *Wetlands are important for the world due to the benefits they offer to humanity. However, there are studies showing their decrease in number in several regions. A form of monitoring of the wetlands is necessary to evaluate their reduction, appearance and / or disappearance. The Tunki project is a project to detect wetlands in a given region using satellite images and high-performance computing. The present work, exposes its architecture and its case study: the wetlands in the Chili river basin of the Arequipa-Peru region. The results show the wetlands detected and the mosaics generated using the satellite images and HPC.*

Key word — *Satellite Images, High Performance Computing, wetlands.*

I. INTRODUCTION

The variety of geo-referential applications (GIS) were initially designed for the study of the land, which have allowed numerous areas such as marketing, agriculture, statistics, etc.

Wetlands are very important for human survival. According to the "Secretary of the Ramsar Convention on Wetlands", they are one of the most productive environments in the world and are cradles of biological diversity and water sources and primary productivity of which innumerable plant and animal species depend for subsistence.

Wetlands by their nature and depending on the season can vary in size or in some cases appear and disappear, this can be considered normal if the variation is by natural cause, however, in many cases it is for some cause external to nature, causing a change in the habitat and possible irreparable ecological impact.

It is necessary to have monitoring tools on wetlands to carry out periodic analysis on the region where the wetlands are located and subsequent decision making.

The most used technique to observe and analyze wetlands is through remote sensing of satellite images. Remote sensing consists of collecting information through different devices in a specific area, for example, the acquisition of images through satellites.

Technological advances in the field of remote sensing generate large volumes of geospatial data; geographic information systems (GIS) that generally do not have support for massive processing of large satellite images, such as the SPRING project [1] developed in Brazil and QGIS [2].

In recent years, the use of high performance computing has allowed processing large amounts of data for further

analysis through complex algorithms, generating more accurate results and obtained in less time.

A. Objectives

The Tunki project is a proposal for periodic monitoring of wetlands in a certain region using high performance computing. Satellite images, due to their large size and information, require great computational power to be processed, which is why the Tunki project uses high performance computing.

The main objective of the project is to offer a tool to detect and monitor the number of wetlands in a given region. The case study of the project is the Chili River basin, in which, using satellite images and high-performance computing, by means of the construction of a thematic map of the wetlands of said region.

B. Text organization

The present work is divided as follows: in section II the background is exposed, these concepts are important to understand the proposal. In section III the related works are exposed, showing their differences with respect to the proposal. Section IV presents the architecture designed for the Tunki project, emphasizing high performance computing. Section V shows the experiments carried out using the Tunki project, making a comparison with the current alternatives. Finally, section VI presents the conclusions and future work.

II. BACKGROUND

The important concepts for a better understanding of the proposal are presented in this section.

A. Remote Sensing, Wetlands and Satellite Images

1) Wetlands

The term wetland is the one used to refer to a specific type of ecosystem biome that is characterized by having muddy or not entirely firm territories due to a high proportion of water. According to Ramsar, "wetlands are one of the most productive environments in the world and are cradles of biological diversity and water sources and primary productivity of which innumerable plant and animal species depend for subsistence". In the Peruvian territory more than 12 thousand 200 lakes and lagoons are considered, of which thirteen wetlands were recognized.

Digital Object Identifier (DOI): <http://dx.doi.org/10.18687/LACCEI2018.1.1.307>
ISBN: 978-0-9993443-1-6
ISSN: 2414-6390

2) Remote Sensing

Remote Sensing includes a series of techniques for locating and collecting data from objects and phenomena at a distance, without physical contact with the object or phenomenon of interest. Some of its interesting applications are atmospheric pollution, monitoring of deforestation, geomorphology, detection of wetlands, natural disasters, etc. Remote sensing is being used as a tool in various scientific areas, related to the physical environment.

3) Satellite Images

A satellite image or satellite image can be defined as the visual representation of the information captured by a sensor mounted on an artificial satellite.

There are different sources of satellite image generation, these sources are categorized by the set of satellites that generate such images, among which are: Landsat, Meteosat, NOAA, Spot and Rapideye.

In the field of remote sensing energy emanating from the earth's surface is measured using digital sensors usually mounted on ships, satellites and space platforms, these measurements are used to build spectral images. This energy emanated from the bodies (by absorption, dispersion or emission) by the effect of reflection and emission of this radiation, also known as reflectance measurement [4]. Each spectral reflectance value is recorded as a digital number, these numbers are transmitted back to the ground where by processing they are converted to shades or colors of gray to create an image that looks like a photograph [5].

The resulting satellite image consists of a set of matrices, one for each channel of the sensor, in which numbers from 0 to 255 appear. The zero indicates that no radiation arrives from that point and 255 reaches the highest value of radiation. In Fig. 1 Channel 1 of the LandSat Satellite Image of Arequipa-Peru, an example of a LandSat 7 satellite image from the province of Arequipa - Peru is shown.

For the processing of satellite images, there are different types of images according to the number of bands and the value that each pixel takes [6], such as:

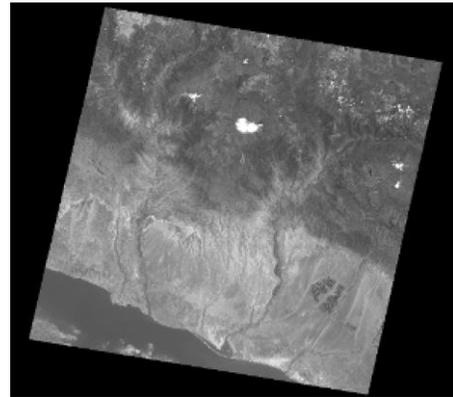


Figure 1. Channel 1 of the LandSat Satellite Image of Arequipa-Peru.

- *Panchromatic images:* These are the images captured by a sensor that measures the reflectance of energy in a large part of the electromagnetic spectrum. These images usually encompass the visible and near-infrared part of the spectrum, which are represented in black and white images.
- *Multispectral Images:* These are the images captured by the sensor in many bands, thanks to the set of detectors that it has, the different reflectance values are combined to create color images, this number of bands is measured from three to fourteen bands, depending on the sensor.
- *Hyperspectral images:* These are images captured with a hyperspectral sensor that measures the reflectance in hundreds of bands, which allows detecting very subtle characteristics among the features of the surface, especially referring to vegetation. Son imágenes captadas con un sensor hiperespectral que mide la reflectancia en cientos de bandas, lo que permite detectar características muy sutiles entre los rasgos de la superficie, especialmente referidos a la vegetación.

The following are the most commonly used Remote Sensing Bands [6]:

- Remote sensing of the reflective infrared, and visible; generally used for the recognition of vegetation, water monitoring and identification of soil and rock.
- Remote sensing in the far infrared, generally used to detect the thermal emission emitted by the bodies, used in the detection of forest fires.

- Remote sensing in the microwave zone, generally used in two ways, through passive and active remote sensing, both for microwave detection.

B. High Performance Computing

In the concepts in the domain of high performance computing, we have:

1) Apache Hadoop - HDFS

It comes from the abbreviations of Hadoop Distributed File System [7], it is a distributed, scalable and portable file system written in the Java programming language. Sub-module of the framework for processing and storage of large amounts of Hadoop data, through which a distributed file management system can be implemented on a set of low-cost computational resources [8].

2) Apache Spark

Spark [9] is a platform designed to improve the speed and performance of big data applications. Spark was born in the research laboratories of the University of Berkeley as an open source research project in 2009. However, it began to take a name from 2013, when it became a project of the Apache Software Foundation. Spark was designed to speed up the response to interactive queries and iterative algorithms from the distribution of processes through the memory of different machines. In some applications, this tool can operate up to 100 times faster than Apache Hadoop.

3) Geotrellis

Geotrellis [11] is a geographic data processing engine for high-performance applications that arises in response to the growing need for satellite map processing, reaching data processing at the terabyte level. Its characteristics include: (i) it is open source, (ii) it allows the manipulation and operation of the map data, (iii) it provides support for the deployment of Web services through REST services [10].

III. RELATED WORKS

In the literature there are several works that focus on solving different problems regarding remote sensing of wetlands.

Among the most relevant we have Medekisa et al. [3], this work focuses on quantifying the land cover of the African continent over a period of 15 years using Landsat satellite images and the Google Earth Engine cloud platform. The work uses remote sensing and high-performance computing but to do a generic analysis of the African content using cloud tools already defined for processing the images.

In W. Nina et al. [12] an extension of a tool for the processing of images is made using the Hadoop framework, called HIPI, this tool does not have the option for the

processing of satellite images and in [12] the proposal is made to process images in TIFF format. This solution is limited to the proposal of a tool for processing, compared to the Tunki project, which performs a periodic analysis of satellite images.

VI. ARCHITECTURE OF THE TUNKI PROJECT

Satellite images often contain a greater amount of information than the images of our day-to-day lives; for example, if a satellite image is compared with an RGB image, the latter is composed of only 3 bands (Red, Green and Blue), while a satellite image can be composed of a greater number of bands (such as red limit, infrared). close, among others). Processing these images that contain a greater amount of information make it necessary to use high performance computing, which allows us to use clusters (grouping computational resources for processing) to divide the workload into its multiple nodes; thus, reducing the response time in the processing of satellite images.

In the configuration of the Tunki project architecture for the processing of satellite images, the following tools were considered:

- Hadoop Distributed File System - HDFS
- Apache Spark
- Geotrellis

A. Hadoop Distributed File System - HDFS

The HDFS architecture responds to the communication model called master/slave, where one of the cluster nodes fulfills the role of organizer of the file system and is responsible for managing the address space of the file system (NameNode), this node is also responsible for the mapping between files and their blocks, and to coordinate the access that customers make to the data.

The rest of the nodes fulfill the role of DataNodes, and these are responsible for storing the blocks of files and provide the necessary data in response to customer requests.

For this project, this file system is being used to store the satellite images in 2 formats: (i) for the first case, the images are stored without any modification, that is, they are directly stored in a distributed way in the cluster; (ii) and for the second case the images are stored in the format required by the Geotrellis tool, that is, the images will be divided into tables (known as Tiles in English), which are portions of images that maintain their data. The pyramids of these images are stored for this case, these pyramids help to recover the images quickly for different sizes requested by the user.

One of the main reasons why this technology was used in the project is the tolerance to failures that the Hadoop framework offers, since the larger the cluster of computers on which the distributed file system is running, the greater the

probability that some node fails. HDFS has a replication system [7], this means that each block can be stored in more than one node, this to achieve high data availability even when some node stops working (since a copy the block of the necessary data will be available in another node).

In Fig. 2 the operation of the replication system in HDFS is shown, in this figure it can be seen how internally the file is considered a sequence of blocks (of equal size), and each one of said blocks is copied in multiple nodes to guarantee his persistence and availability.

The replication factor can be established by the user and determines the number of replicas that will be created for each of the blocks in the file. For this project, the value of 3 was set as a replication factor.

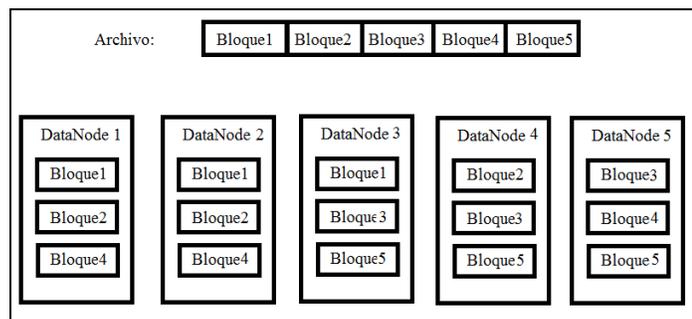


Fig. 2. Replication scheme with Factor 3.

B. Apache Spark

This framework was used as a requirement of the Geotrellis tool, since to process the immense amount of data provided by the satellite maps, a tool with a high performance is necessary, Spark is used for 2 fundamental tasks: (i) The first to store distributed form the satellite maps to the distributed file system, which will be stored in Tiles (portions of the maps but in a format that can be read by Geotrellis); (ii) the other task performed with this framework is to process the previously stored Tiles in a distributed manner; therefore the Tiles are processed in the different nodes where they are stored, so this tool will help these Tiles can be processed in a distributed way.

C. Geotrellis

In this project, Geotrellis was used as a tool to read the data stored in the distributed file system (using Spark) and then the algorithm programmed for the detection of the wetlands was applied to each Tile.

The programmed algorithm allows us to extract the normalized index of vegetation and the normalized index of water, which were used to determine when a pixel could be considered as part of a wetland, later this tool also helps us to filter the pixels that were previously labeled as part of a wetland through a verification of connectivity of the pixels and

a distinction of those conglomerates of pixels that do not comply with the appropriate dimensions.

D. Wetland Detection

To detect wetlands, we will proceed to make use of the technique called threshold to pre-process and then restrict again the data in order to result in the detection of a wetland. As a first step, the first restriction will be given by the remote sensing of the bodies of water in a multispectral image, this will be achieved with the "Normalized Differenced Water Index" (NDWI) [13]. As a second step and more important restriction will be given by the amount of water contained in a body of water, this will have to overcome an extension of more than 50 Ha, considering the maximum flood. This will give us a high precision in the detection of wetlands.

When using the NDWI, the large amount of data is pre-processed obtaining values lower than 1, which show how close to the water are the pixels of a multispectral image, however due to the large amount of data necessary to apply the technique of threshold, to make a cut in the image, this value will be called "NDWI trust" that provides relevant information and ensures similarity to water.

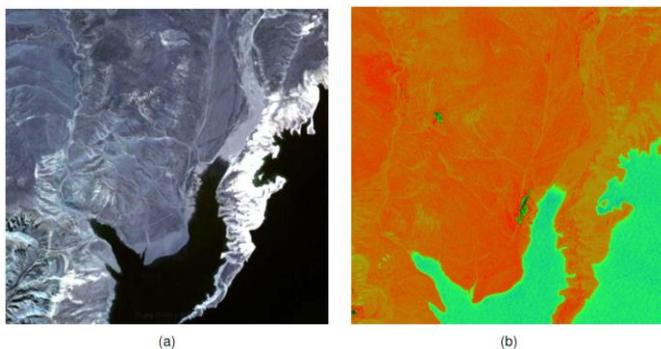


Fig. 1. Part (a), original image. Part (b), image generated using NDWI.

In Fig. 3, in part (a) we can see a satellite image and in part (b) we can see the image generated using the NDWI. The NDWI index can be determined with a certain degree of confidence [14], the degree of confidence is determined according to Table 1.

TABLE 1
CONFIDENCE OF THE NDWI

Low	NDWI \geq 0.02
Medium	NDWI \geq 0.06
High	NDWI \geq 0.09

The tests for the confidence levels provide approximate results about the wetlands, however, these data are not determinant for the detection of a wetland because they present multiple errors. For example, in Fig. 4 you can see the level of confidence and its influence on the data, the blue color shows the similarity to a body of water. Therefore, the "high" confidence level provides greater precision when detecting a wetland.

V. EXPERIMENTS

As a case study of the Tunki project experiments, the Chili River Account was established in the city of Arequipa-Peru. As you can see in Fig. 6 there are three wetlands in Arequipa currently recognized by Ramsar:

- Laguna del Indio - Dyke of the Spanish
- Bofedales and Laguna de Salinas
- Lagunas de Mejía National Sanctuary

The database with which the experiments were conducted was provided by the National Water Authority, ANA, an institution of the Peruvian state, which had acquired RapidEye satellite images.

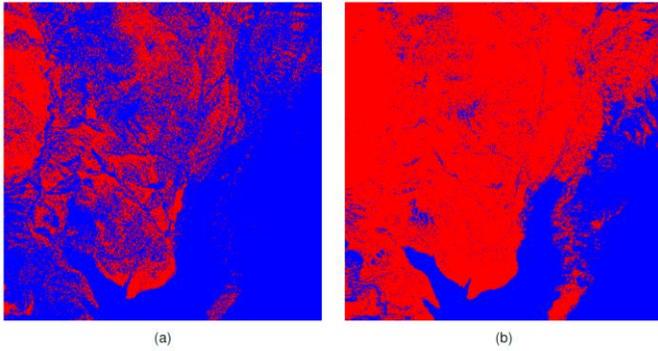


Fig. 2 Part (a), image for NDWI with confidence 0.06. Part (b), image for NDWI with confidence 0.09.

Once you have detected regions that represent water using the NDWI, it is necessary to have a criterion to determine if it is a wetland or not. For that, an evaluation of the pixels neighboring a pixel representing water is carried out. For example, in Fig. 5 a set of pixels is shown, in which the pixel of the medium (x, y) is the pixel representing water, to validate the water region a comparison must be made with its 4 adjacent pixels, in this way a better filter will be made to show regions that represent a wetland.

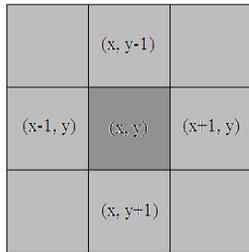


Fig. 3 Representation of pixels of an image.



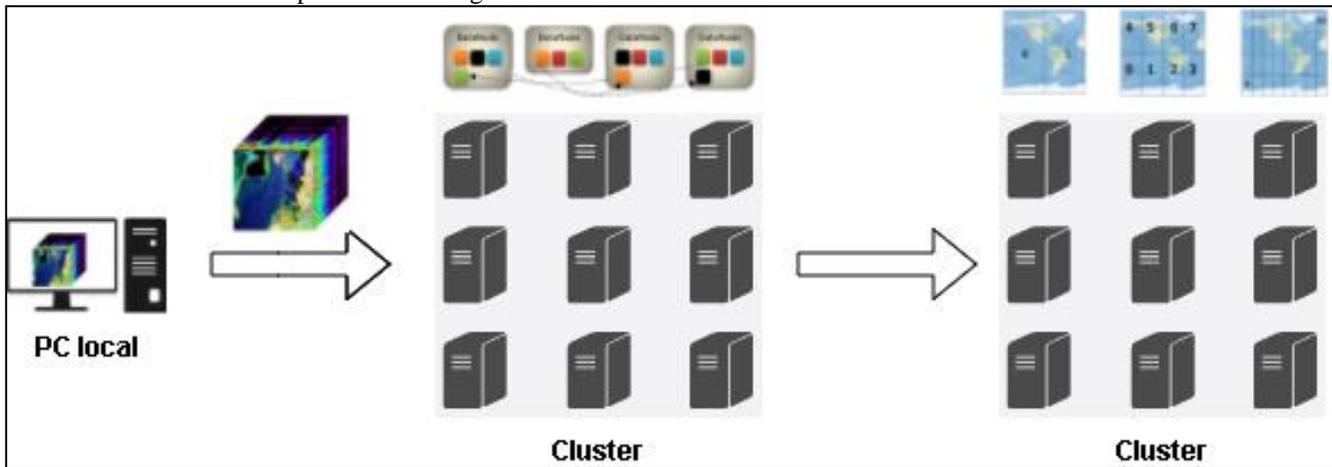
Fig. 4 Wetlands of the Chili-Arequipa River Basin, recognized by Ramsar

A. RapidEye Satellite images

The RapidEye satellites are the first commercial satellites to include the Red-Edge band, which is particularly sensitive to changes in chlorophyll content. Studies show that this band can help monitor plant health, in the identification of plant species and crops, and contribute to the measurement of protein and nitrogen content in biomass.

Fig. 5 Process of ingesting the images in the Cluster.

The characteristics of the pixels of the images are:



- Pixel depth: 12 bit.
- Pixel size (orthorectified): 5 m.

The characteristics of the Bands are:

- Band 0 (Blue): 440 - 510 nm
- Band 1 (Green): 520 - 590 nm
- Band 2 (Network): 630 - 685 nm
- Band 3 (Red Edge): 690 - 730 nm
- Band 4 (Near IR): 760 - 850 nm

The processed images were 47 RapidEye images which are detailed in the following Table 2:

TABLE 2
SATELLITE IMAGES ANA - 2010

Month	Number of images
August	11
September	7
October	9
November	20
Total	47

B. Cluster Inkari – UNSA

In the Tunki project, the High-Performance Computer of the National University of San Agustín was configured and used, which has the following characteristics:

- 1 Head Node SGI
2x AMD-Opteron processors, twelve-core, 2.4GHz.
64 GB of RAM memory DDR3 1600 MHz
4 Discs of 2TB SATA 7200 RPM
- 36x Server Nodes (9x SGI), each Node server includes:
2x AMD Opteron twelve-core processors, 2.4GHz
64 GB of RAM memory DDR3 1600 MHz
3 SATA 2TB Disks 7200 RPM

Of which, 12 server nodes were configured with the tools and frameworks for the use of the Tunki project.

C. Ingesting images in the cluster

For the processing of the satellite images it is necessary that the images are in the distributed system of the Tunki project ingested.

As shown in Figure 8, the user through his local machine can upload the satellite images to the system, later the images are in the distributed HDFS file system, this step guarantees the availability of the images in the system, because When there is a possible failure of a cluster node, HDFS has up to 3 replicas of each data. Once the images are stored in the distributed system, they can already be ingested. To ingest the images, we use a Geotrellis module called "geotrellis-spark-etl", this module allows us to ingest the satellite images in several distributed file systems, such as Accumulo, Cassandra, S3, and HDFS. This step should not be confused with storage, since

the ingestion allows us to have the images in the format required by Geotrellis, in this way it is possible to easily obtain a certain raster that corresponds to a portion of an image ingested with different sizes.

D. Making requests to the system

Once the images are stored and ingested, it is possible to make requests to the distributed system. In Fig. 8 the request of a user is observed.

The procedure to answer this request is to identify the parameters that were sent in the URL, the parameters sent are 3, where the first 2 will be 2 identifiers of the map and the third will be the level of size (zoom) required, later it will be obtained the raster that corresponds to the values of the request. To the obtained raster, the operations necessary for the detection of the wetland are applied, these operations will result in a new raster in which all those pixels that do not belong to a wetland will be omitted. Next, the new raster is assigned the same coordinates as the original raster, this in order to overlap the original raster on any base map. Finally this raster is converted into a png format image and sent to the user in response to his request.

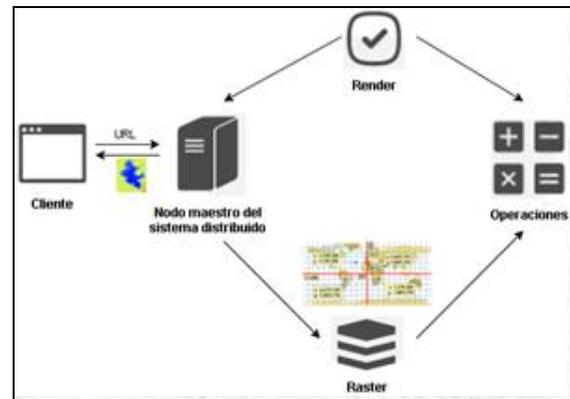


Fig. 6 Process of response to a request made by the user

D. Analysis of the results

For the experiments, the satellite images with the characteristics previously described were used. The initial objective was to identify the wetland in the correct area of the map. The chosen region was in the Arequipa region: Chili River Basin.

From the database provided, the experiments were carried out to make the detection of wetlands in the Arequipa region: Chili River Basin.

As shown in Fig. 6, this region has 3 wetlands established by Ramsar. As you can see, the wetland shown in Fig. 9 belongs to the "Laguna del Indio - Dique de los Españoles" wetland, you can see three presentations, part (a) shows a view generated on a map from the images satelital where you can see the wetland itself, in part (b) you can see the wetland in its

entirety, leaving aside the map, in this view you can get to make measurements of the wetland, finally the part (c) shows both the map and the wetland detected. These images belong to the month of August of the database provided.

In Fig. 10 the wetland region called "Bofedales and Lagunas de Salinas" is shown. The images that cover this wetland, correspond to the images of the month of September. As it is observed, the generated mosaic is not shown correctly, this is because the values of the minimum and maximum pixels are outside the established range. We can also consider that in this area there is generally cloudiness in a good part of the year, which is why the pixels of the image are not those that would show a wetland.

In Fig. 11 the wetland region called "Lagunas de Mejía National Sanctuary" is shown. In this region, multiple pixels were detected that meet the condition of the NDWI, it is assumed that this is due to the fact that there is a large number of crops in the area, which contain a high water index, which causes the algorithm to not work properly.

The mosaic generated by joining all the satellite images provided is shown in Fig. 12, as can be seen by assembling all the images of all the months the complete mosaic of the department of Arequipa was built. It must be taken into consideration that a set of satellites will not always show the complete image of a region at a certain moment.

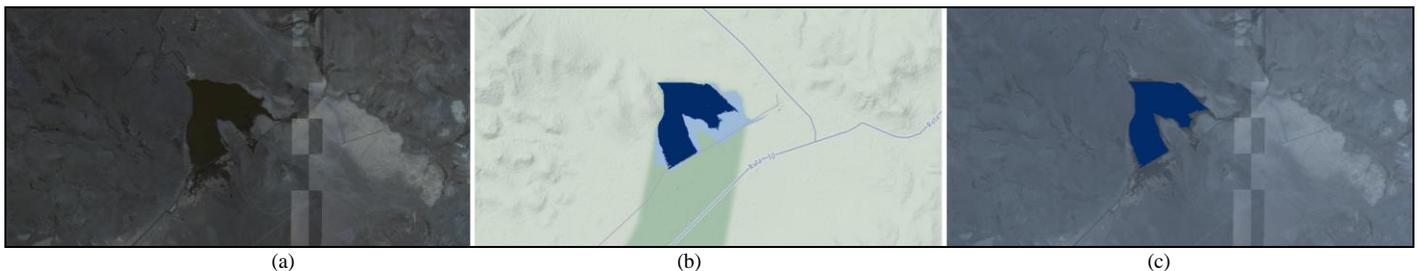


Fig. 7. Processing result. Laguna del Indio Wetland – Dique de los Españoles. In part (a) only the base map generated is shown. In Part (b) only the wetland is shown and in Part (c) it is shown in the base map and the wetland.

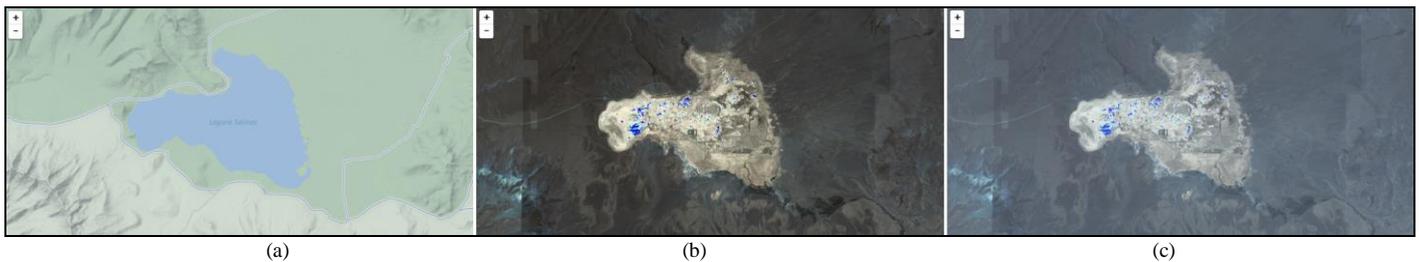


Fig. 8. Processing result. Wetland region Bofedales and Lagunas de Salinas. Part (a) shows the region where the wetland is displayed. Part (b) shows the generated mosaic. Part (c) shows the wetland region and the mosaic.

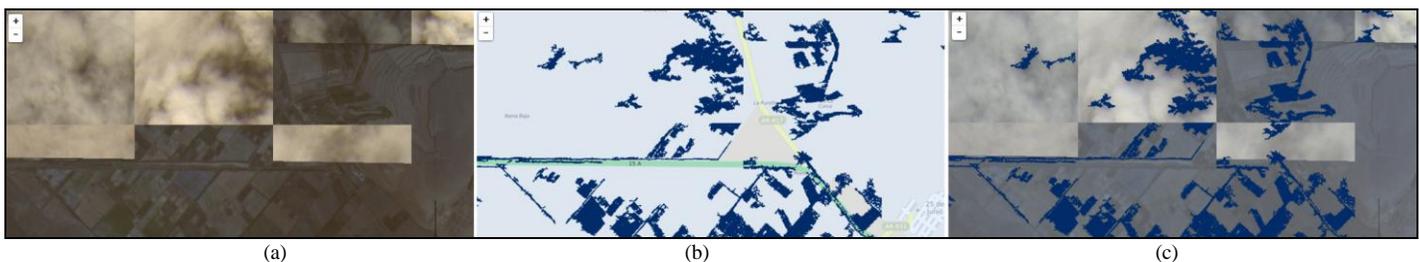


Fig. 9. Processing result. Wetland region National Sanctuary Lagunas de Mejía. Part (a) shows the generated mosaic. Part (b) shows the wetland region. Part (c) shows the mosaic and the wetland region.

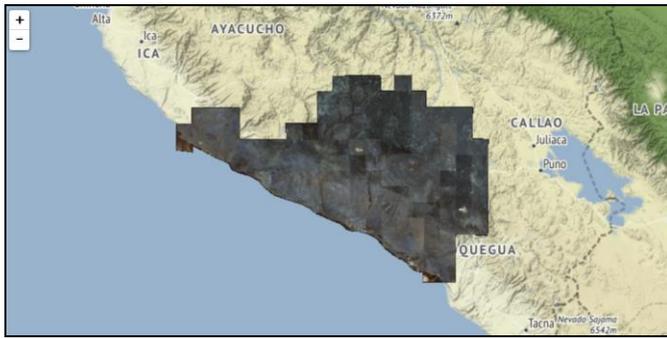


Fig. 10. Mosaic generated from the satellite images provided.

VI. CONCLUSIONS AND FUTURE WORK

In this article, the Tunki project was presented, a wetland remote sensing project using satellite images and high-performance computing. The databases used and analyzed were provided by the National Water Authority, ANA, these images correspond to the region of Arequipa in the months of: August, September, October and November of the year 2010.

While, more information has the satellite image, this image will be larger. The processing of many images requires a large computational capacity. The images used in the project were RapidEye. These images are large because they include the Red-Edge band, which is particularly sensitive to changes in chlorophyll content.

The technique used for the detection of wetlands is based on the "Normalized Difference Water Index" (NDWI). The technique used detects wetlands in images without noise, this means that the images must be clear and have no distortion.

As future work is intended to expand the number of images to be analyzed, using the images provided by CONIDA of the Peruvian Satellite PeruSat1. On the other hand, it is proposed to improve the wetland detection algorithm, both in its precision and in its execution.

ACKNOWLEDGMENTS

The Tunki project team thanks Fondecyt - Peru (National Fund for Scientific, Technological and Technological Innovation) for the project's grant.

The Tunki project team was worked at the "Instituto Astronómico y Aeroespacial Pedro Paulet", Universidad Nacional de San Agustín.

The computational experiments developed in the Tunki project were carried out in the Computational High Performance Laboratory of the National University of San Agustín. The Tunki team appreciates the availability of the

Cluster "Inkari hosted at the IAAPP-UNSA, to meet the project's objectives.

REFERENCES

- [1] Camara G, Souza RCM, Freitas UM and Garrido J, "SPRING: Integrating remote sensing and GIS by object-oriented data modelling", *Computers & Graphics*, 20: (3) 395-403, 1996.
- [2] QGIS Development Team, 2009. QGIS Geographic Information System. Open Source Geospatial Foundation. URL <http://qgis.osgeo.org>.
- [3] Midekisa A, Holl F, Savory DJ, Andrade-Pacheco R, Gething PW, Bennett A, et al. "Mapping land cover change over continental Africa using Landsat and Google Earth Engine cloud computing". *PLoS ONE* 12(9), 2017.
- [4] John A Richards and JA Richards. "Remote sensing digital image analysis an Introduction", volume 5. Springer, 2013.
- [5] S Oprisescu and M Dumitrescu. "On the regularization of segmented satellite images". In *Signals, Circuits and Systems*, 2005. ISSCS 2005. International Symposium on, volume 1, pages 83–86. IEEE, 2005.
- [6] Ernesto Gómez Vargas, Nelson Obregón Neira, and Diego Fernando Rocha Arango. "Métodos de segmentación de nubes en imágenes satelitales". *Tecnura*, 17(36):96–110, 2013.
- [7] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. "The Hadoop Distributed File System". In *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. IEEE Computer Society, Washington, DC, USA, 1-10.
- [8] Tom White. 2009. "Hadoop: The Definitive Guide" (1st ed.). O'Reilly Media, Inc..
- [9] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. "Apache Spark: a unified engine for big data processing". *Commun. ACM* 59, 11 (October 2016), 56-65.
- [10] Rohit Khare and Richard N. Taylor. 2004. Extending the Representational State Transfer (REST) Architectural Style for Decentralized Systems. In *Proceedings of the 26th International Conference on Software Engineering (ICSE '04)*. IEEE Computer Society, Washington, DC, USA, 428-437.
- [11] GeoTrellis. Available at <http://geotrellis.io/>, 2018
- [12] W. Nina, R. Cruz, J. Serrano, J. Cuba, Y. Huaynacho, A. Mamani-Aliaga, Y. Yari and P. Yanyachi, "A new approach to the massive processing of satellite images," 2015 Latin American Computing Conference (CLEI), Arequipa, 2015, pp. 1-6.
- [13] Bo-cai Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space", *Remote Sensing of Environment*, Volume 58, Issue 3, 1996, 257-266.
- [14] McFeeters, S.K. Using the Normalized Difference Water Index (NDWI) within a Geographic Information System to Detect Swimming Pools for Mosquito Abatement: A Practical Approach. *Remote Sens.* 2013, 5, 3544-3561