MAPPING GLACIER CHANGES USING CLUSTERING TECHNIQUES ON CLOUD COMPUTING INFRASTRUCTURE

V. Ayma [1, 2, *], C. Beltrán [1], P. Happ [3], G. Costa [4], R. Feitosa [3, 4]

[1] Pontifical Catholic University of Peru, 1801 University Ave., Lima, Peru
- (vaaymaq, cbeltran)@pucp.pe
[2] Peruvian Navy, 36 Marina Ave., Callao, Peru
[3] Pontifical Catholic University of Rio de Janeiro, 225 Marquês de São Vicente St., Rio de Janeiro, Brazil
- (patrick, raul)@ele.puc-rio.br
[4] Rio de Janeiro State University, 524 São Francisco Xavier St., Rio de Janeiro, Brazil
- gilson.costa@ime.uerj.br

KEY WORDS: Remote Sensing, Big Data, Cloud Computing, Glacier Changes, Clustering Techniques.

ABSTRACT:

Climate change and its effects are taking more importance nowadays; and glaciers are one of the most affected ecosystems by that, considering that the energy of Earth's surface and its temperature may be directly related to glacier temporal changes. Then, the comprehension of glaciers behaviour, by its retreating or melting critical conditions, can be achieved by the analysis of Remote Sensing data, but considering the unprecedented volumes of information currently provided by satellites sensors, we can refer to this analysis as a big data problem. Machine learning techniques have the potential to improve the analysis of this type of data; however, most current machine learning algorithms are unable to properly process such huge volumes of data. In the attempt to overcome the computational limitations related to Remote Sensing Big Data analysis, we implemented the K-Means and Expectation Maximization algorithms, as distributed clustering solutions, exploiting the capabilities of cloud computing infrastructure for processing very large datasets. The solution was developed over the InterCloud Data Mining Package, which is a suite of distributed classification methods, previously employed in hyperspectral image analysis. In this work we extended the functionalities of that package, by making it able to process multispectral images using the aforementioned clustering algorithms. To validate our proposal, we analysed the Ausangate glacier, located on the Andes Mountains, in Peru, by mapping the changes in such environment through a multi-temporal Remote Sensing analysis. Our results and conclusions are focused on the thematic accuracy and the computational performance achieved by our proposed solution. Thematic accuracy was assessed by comparing the automatically detected glacier areas by the clustering approaches against the manually selected ground truth data. We compared the computational load involved in executing the clustering processes sequentially and in a distributed fashion, using a local mode and cluster configuration over a cloud computing infrastructure.

1. INTRODUCTION

The scientific community has placed much attention nowadays on the understanding of how climate changes affect Earth's behaviour (Houghton, 2001), and the analysis of the cryosphere, comprised by sea ice, snow cover, frozen ground and glaciers extensions, plays an important role in the characterization of the Earth's climate system (Ke, 2016) (Kaser, 1990). It is expected that glaciers extents and thickness decrease as the climate warms, as glaciers' mass fluctuations correlate well with global climate changes, which provides strong evidence of the effects of global warming on the energy balance of the Earth's surface (Callegari, 2017) (IPCC, 2013). Nevertheless, analysing glacier changes is a complex, expensive and time-consuming process, mostly affected by the difficulties in consistently collecting reliable data, by the large areas involved in the analysis, and by the capacity of computational techniques to process big amounts of data (Winsvold, 2015) (Racoviteanu, 2010) (Cui, 2010).

Remote Sensing (RS) can be considered as one of the most important and efficient tools for Earth observation (EO). According the literature, much effort is placed towards comprehensive studying and monitoring of glaciers through multi-temporal analysis of remotely sensed multispectral images, which can be used for mapping glacier-covered areas and help detecting its recessions and fluctuations (Yue, 2018) (Paul, 2015) (Muhammad, 2013) (Bolch, 2011) (Kääb, 2012) (Willis, 2012) (Raup, 2007) (Barry, 2006) (Kääb, 2014). Typically, satellite imagery is used to measure glacier changes between images taken at different dates (Vignon, 2003).

With the increasing availability of data provided by new constellations of EO satellites, the quantity of RS images available has grown exponentially, at higher spatial and temporal resolutions. Nevertheless, relatively few approaches have been proposed to cope with the challenges imposed by the analysis of such large amounts of data, a problem that can be characterized as Remote Sensing Big Data analysis (Ghamisi, 2017) (Li, 2017) (Chi, 2016).

Machine learning techniques have the potential to improve the analysis of RS data; however, most current machine learning algorithms are unable to properly process very large volumes of data (Bekkerman, 2012). This research represents an attempt to

* Corresponding author

overcome the limitations of this class of techniques, applied to Remote Sensing Big Data analysis.

In this work, the K-Means and the Expectation Maximization (EM) clustering algorithms were implemented as distributed solutions, which are able to exploit the capabilities of cloud computing infrastructure for processing large RS datasets. The solution was developed as part of the InterCloud Data Mining Package, which is a suite of distributed classification methods, previously employed in hyperspectral image analysis (Ayma, 2017) (Ayma, 2015). In this work we extended the functionalities of that package, by making it able to process multispectral images using the aforementioned clustering algorithms.

We evaluated the application of the clustering techniques for analysing the Ausangate glacier, located on the Andes Mountains, in Peru, by mapping the changes in that environment through multi-temporal RS image analysis. Our results and conclusions are focused on the thematic accuracy and the computational performance of the proposed solutions. Thematic accuracy was assessed by comparing the automatically detected glacier areas with ground truth data. Additionally, we compared the computational load involved in executing the respective processes sequentially and in a distributed fashion, using a physical local machine and cloud computing infrastructure.

The remainder of this paper is organized as follows; Section 2 presents a brief overview of the main underlying concepts of the InterCloud Data Mining Package, and its extension with the aforementioned clustering algorithm. Section 3 describes the Ausangate glacier dataset, the experimental design, and the results achieved. Finally, conclusions and direction for future works are presented in Section 4.

2. DESCRIPTION OF THE APPROACH

Cloud computing refers to the on-demand delivery of computing services, servers, compute power, storage, databases, networking, software, analytics, applications, and other IT resources through a cloud service platform over the Internet,. Cloud computing enables users to efficiently exploit a distributed infrastructure, with scaling capabilities, according to their particular needs (Srinivasan, 2014) (Buyya, 2011) (Microsoft Azure, 2019) (Amazon, 2019).

The InterCloud Data Mining Package is an open-source distributed tool, able to perform MapReduce-based processes over cloud computing infrastructure, supporting distributed execution, network communication, and fault tolerance. The package was designed to enable the execution of the classification algorithms available at the Waikato Environment for Knowledge Analysis – WEKA library (WEKA, 2019), over large volumes of RS data, through the distribution of data and processing tasks among machines connected over a network. In this work we included the K-Means and the Expectation Maximization (EM) algorithms in the package, embedded into the machine learning layer.

K-Means is an iterative algorithm that tries to find the best clusters for every sample in a dataset and it requires the definition of an initial number of clusters to iterate. To determine which sample belongs to a given class, the algorithm uses the distances from the centroids of clusters to the sample being evaluated, assigning the class as the one corresponding to the closest cluster´s centroid (Arthur, 2007). EM is a statistical approach and linearly convergent algorithm, capable of dealing with incomplete data, and commonly used for unsupervised clustering. It estimates the set of parameters in a statistical model, usually considering a Gaussian Mixture Model, computing its maximum likelihood (Kurban, 2016).

InterCloud Data Mining Package architecture is comprised of three layers, each providing different abstraction levels. The first, *distribution layer*, is responsible for the execution of the applications, the second, *machine learning layer*, allows to insert classification and clustering algorithms into an implementation of the distributed architecture, and the third, the *project definition*, is in charge of end user interaction, encompassing all the information required for the execution of the classification or clustering applications (Ayma, 2017).

The clustering process works as follows. First, all the parameters of the clustering algorithm are defined, including two datasets, one for creating the clustering model (training dataset), and another over which the clustering process will be performed (generalization dataset). In sequence, the parameters of the clustering model and the two dataset are partitioned and distributed among the nodes of the cluster. Afterwards, the clustering model is built in each node, in such a way that the clustering model will be the same across all computing nodes. Last, once the model is created at each node, the clustering process is performed on its respective portion of the generalization dataset, and the outcomes are then returned to the master node, which finally provides the overall clustering outcome on the complete generalization dataset.

3. EXPERIMENTAL DESIGN

To assess the performance of the approach, we used a time series of satellite images covering the Ausangate glacier, located on the Andes Mountains in Cusco, Peru, geographically located at 71°13′52″W longitude and 13°47′19″S latitude. The database is composed of three images from the Sentinel-2 satellite, considering low cloud coverage and identical temporal seasonality, corresponding to the months of July, June, and May, from 2016, 2017 and 2018, respectively. The images were acquired through the Earth Explorer web service, from the United States Geological Survey (USGS, 2019).

We have considered the red, green, blue and NIR spectral bands of each image, containing 10980 x 10980 pixels at a fine resolution of 10 meters per pixel, and the SWIR spectral band, containing 5940 x 5940 pixels at a spatial resolution of 20 meters per pixel. We selected the image from July, 2016, as shown in Figure 1, for constructing our validation data through its visual analysis, considering that image as the one with the largest snow covered area from the dataset.

From image in Figure 1, we manually refined a semi-automatic segmentation of the glacier extension based on the binary mask provided by the NDSI (Normalized Difference Snow Index) image, with an empirical threshold of 0.4, as recommend in (ke, 2016). The ground truth dataset was comprised of two 2 classes, one for snow and ice covered extensions, and the other considering the rest of the image as background, as illustrated in Figure 2.

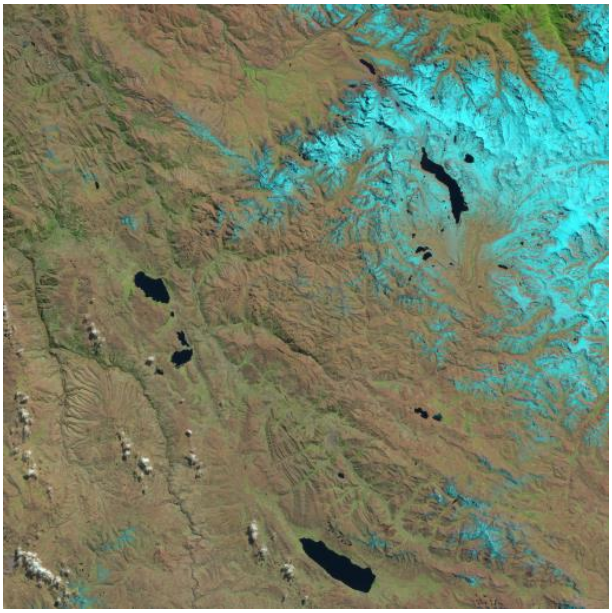

Figure 1. Figure placement and numbering

To compute the thematic accuracy of both distributed implementations, K-Means and the EM algorithms, we used a scaled version at 40% of the dataset (original image and its binary mask). For that purpose, we divided the dataset into training and generalization samples, we randomly selected 10% of the pixels in the ground truth image (for each class) as the training data, and the remaining 90% of pixels were considered as the generalization data for assessing the thematic accuracy of the clustering models created with the previously training data, as presented in Figure 3.

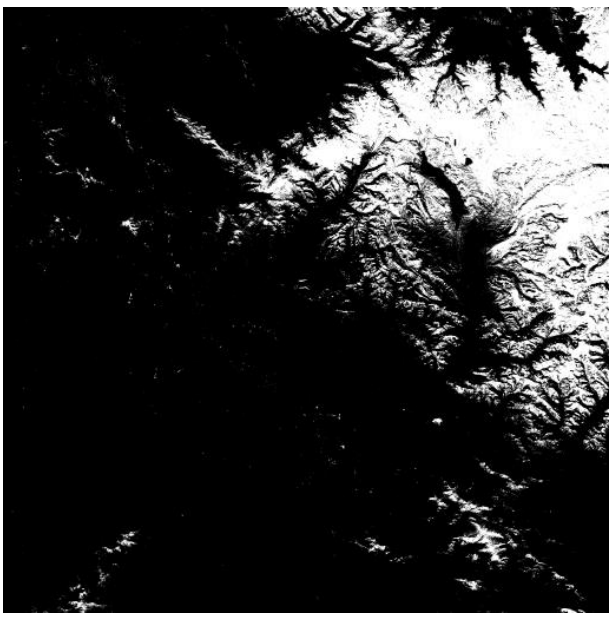

Figure 2. Ausangate Glacier ground truth dataset. In white: glacier extension; in black: rest of the scene (background).

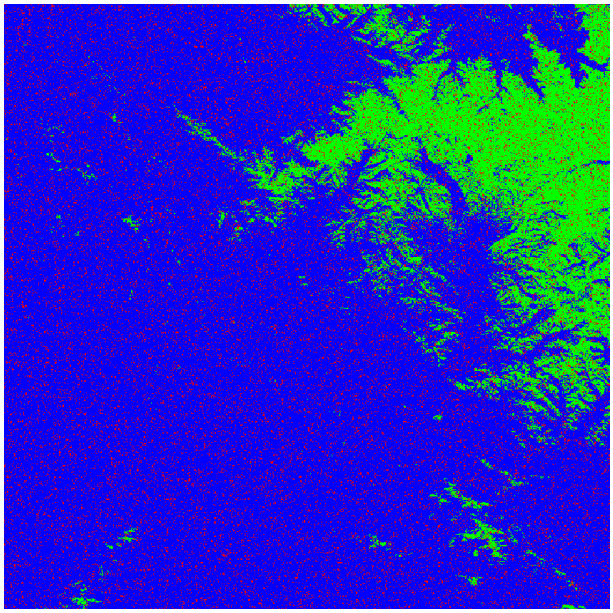

Figure 3. In blue and green, generalization samples; in red, training samples; both datasets randomly selected from ground truth data, for glacier and non-glacier areas, respectively.

To assess glacier multi-temporal changes, the algorithm that achieved the best thematic accuracy was considered for processing the other two glacier images, each equally scaled at 40% of its original size to compute their glacier extensions at each period of time; then, we tracked the evolution of the glacier extension across the years considering the outcomes provided by clustering algorithm with the best performance.

To assess computational performance, we constructed subsampled versions of the original image and from its ground truth, as shown in Table 1. Each scaled version of the original image was processed in sequential and distributed configurations, using local and cloud computing environments. Sequential processing on local mode configuration refers to the implementation of a cluster on the cloud containing 2 nodes, and the processing times achieved with this configuration were taken as baseline for speed up analysis. These sequential times are presented as well in Table 1, for both algorithms.

Image Scale (%)	Dimensions (pixels)	Data Size (Mb)
40	4392 x 4392	1342.30
50	5490 x 5490	2097.39
100	10980 x 10980	8387.14

Table 1. Scaled versions of the original image.

4. RESULTS

As conceived in the experimental design, in Figure 4 and 5 we present the outcomes for both algorithms when applied in the clustering process on the scaled version at 40% of the image in Figure 1. Thus, the thematic accuracy achieved on that image, compared against the ground truth from Figure 2, was of 92.25% and 73.66% for the K-Means and the EM algorithms, respectively, as shown in Figure 4 and Figure 5, where, from a qualitative point of view, the clustered image corresponding to the K-Means algorithm appears more similar to the ground truth than the outcome provided by the EM algorithm. In addition,

Figure 5 shows that the EM algorithm was not capable to correctly cluster glacier extensions and water surfaces into separate classes, leading to a decrease in its overall accuracy.

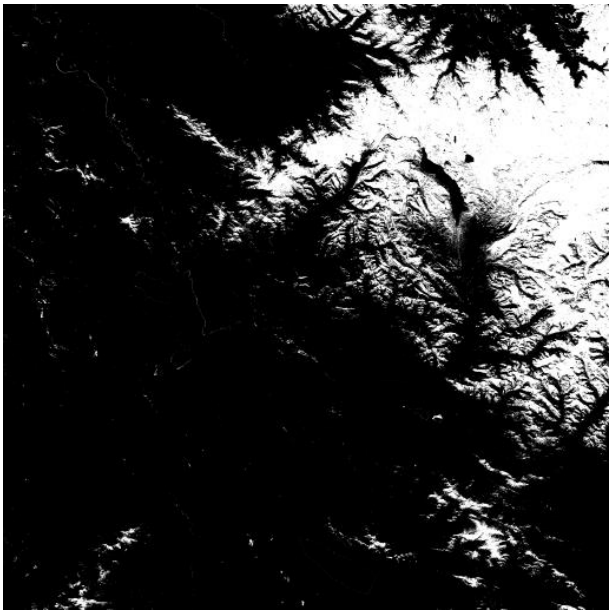

Figure 4. Ausangate Glacier clustering outcome with K-Means algorithm.

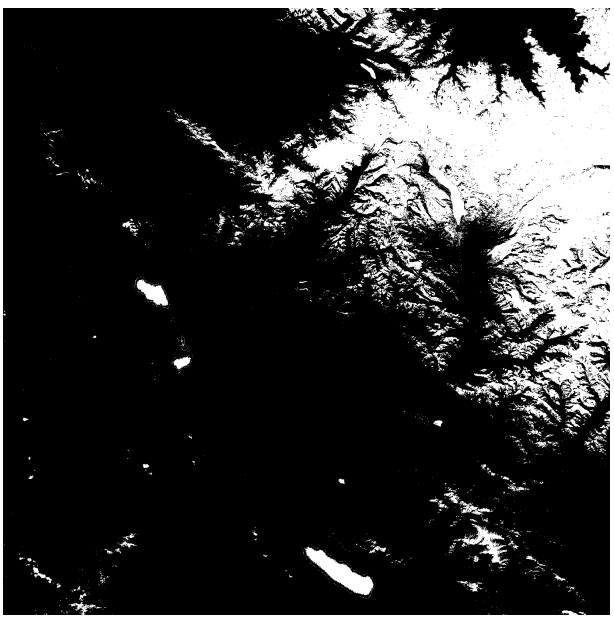

Figure 5. Ausangate Glacier clustering outcome with EM algorithm.

As shown in the previous two images, we can consider the K-Means algorithm as the one with the best performance in identifying glacier extensions; thus, we applied this clustering method to the other two images in the dataset to compute the glacier extension of the Ausangate glacier at each period of time. Figure 6 shows the evolution of the changes in extension of the Ausangate glacier, where it can be seen that its extension has suffered a severe retreat during the period, as compared to its extension at 2016, which is an major issue, considering that by the seasonality of the images, the glacier was supposed to be at its maximum expansion. This figure also presents two estimations of Ausangate glacier extensions, in green is the one provided by K-Means method, and in blue the other provided by a supervised semi-automatic segmentation. It is worth to notice that even though both approaches delivered slightly differences in their glacier extension estimations, it's important to remark that they do follow the same trend about the changes in glacier extension.

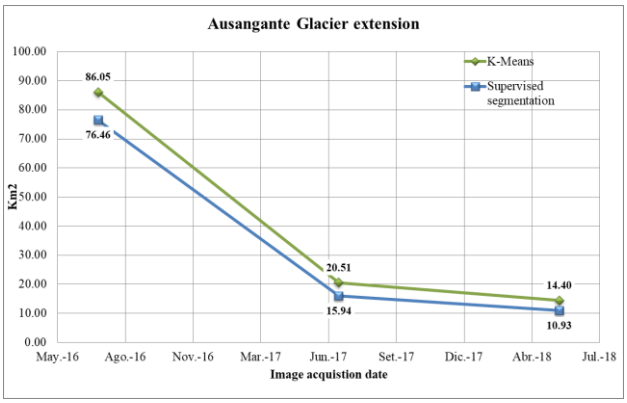

Figure 6. Ausangate Glacier clustering outcome with EM algorithm.

For assessing the computational performance of both distributed implementations, and in order to analyse the computational load involved in executing each algorithm on a cloud computing infrastructure, we used the scaled versions of the original image, as presented in Table 1, and we registered the execution times in local mode for both algorithms. We performed the experiments with the extended version of the InterCloud Data Mining Package deployed over the Amazon Web Service (AWS) cloud computing platform. Table 2 shows the processing times obtained after processing the scaled images on a cluster with 2 nodes (local mode configuration), each node in the cluster correspond to a m3.xlarge type machine, containing an Intel Xeon E-5-2670 v2 processor operating at 2.5GHz with 4 physical cores (8 logical cores), 15 GB of RAM and 2 disks of 40 GB.

Image Scale (%)	Dimensions (pixels)	Data Size (Mb)	K-Means processing times (s)	EM processing times (s)
40	4392 x 4392	1342.30	410.52	2158.96
50	5490 x 5490	2097.39	1854.87	10314.23
100	10980 x 10980	8387.14	5365.78	34748.14

Table 2. Processing times for K-Means and EM distributed algorithms on local mode configuration.

Taking the local mode processing times as baseline for assessing the speed up achieved by the distributed algorithms, we used a cluster configuration with 5, 10 and 20 nodes, each node with the same characteristics as the ones used in the local mode configurations. Figure 7 and Figure 8 shows the speedup achieved with the distributed clustering version of the K-Means and the EM algorithms, respectively, and using the three scaled datasets from Table 2.

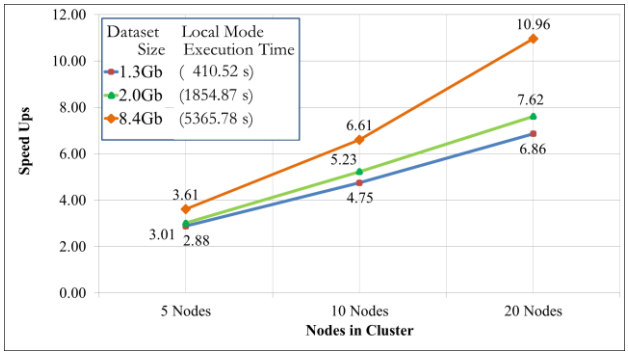

Figure 7. Speedups for the K-Means distributed clustering process.

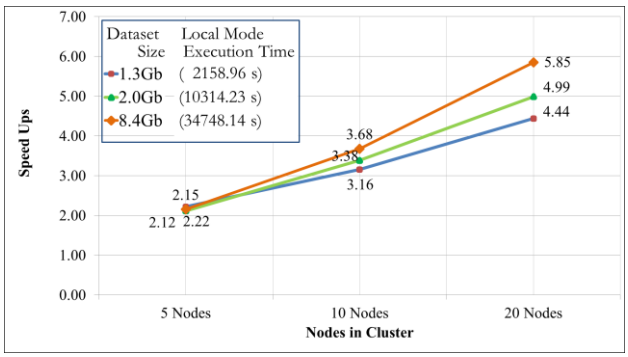

Figure 8. Speedups for the EM distributed clustering process.

Regarding the K-Means algorithm, for the case of the 1.3 Gb dataset size, the speedups were 2.88, 4.75, and 6.86 for 5, 10, and 20 nodes respectively; and for the larger dataset, of 8.40 Gb size, the speedups were 3.61, 6.61, and 10.96 for the same configurations. For the case of the EM algorithm, the speedups achieved were not as high as the K-Means algorithm, thus, for the case of the larger dataset, the speedups achieved were 2.15, 3.68, and 5.85 for 5, 10 and 20 nodes respectively. From the figures, it is worth to notice that for smaller size datasets, as the number of nodes increases, the efficiency of the distributed solution tends to decrease; on the counterpart, for bigger size datasets, the cluster achieves better speedups as the number of nodes was increased.

5. CONCLUSIONS

We proposed an extension of the InterCloud Data Mining Package, which was devised for performing distributed clustering processes on large Remote Sensing datasets, more specifically when working on large multispectral images. Our approach provides a robust, flexible and scalable solution operating over a cloud computing environment, allowing end-users to perform Remote Sensing Big Data analysis such distributed infrastructures.

From our experimental analysis it can be identified that Ausangate glacier has suffered a considerable retreat in its extension within the last 3 years, decreasing its surface extension from 2016 on more than 70% and 80% in 2017 and 2018 respectively. In order to assess if these is a constant behaviour on Peruvian glaciers, it would be important to perform a complete analysis over the entire "Cordillera Blanca", so as to assess if the Peruvian glaciers are indeed melting and retreating, then, decision makers could get a better comprehension on the effects that climate changes and global warming are producing in our ecosystems.

Considering the experiments we conducted for clustering large scale remote sensing datasets using the K-Means and EM algorithms, we can observe that K-Means algorithm had a better performance than the EM algorithm, in both, the thematic accuracy and the computational performance, by achieving higher accuracies and higher speedups when deployed over the distributed infrastructures. It is also important to notice that, as in previous works, when working with bigger datasets, higher speedups could be achieved.

Finally, our work is an initial attempt in order to analyse the glacier changes on the Peruvian Andes, we based our study on one of the biggest glaciers in the country, but it will be important to analyse the complete ecosystem. In this sense, the outcomes that we achieved with our clustering approach are motivating, in terms of continuing our research in the field, as to evaluate more clustering techniques, or to validate a hierarchical clustering approach under this type of distributed infrastructures.

REFERENCES

Amazon, 2019. What is cloud computing?. Amazon Web Services. From: https://aws.amazon.com/what-is-cloud-computing/?nc1=h_ls.

Arthur, D., Vassilvitskii, S., 2007. K-Means++: The advantages of careful seeding. *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*.

Ayma, V., Costa, G., Happ, P., Feitosa, R., Ferreira, R., Oliveira, D., Plaza, A, 2017. A new cloud computing architecture for the classification of remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2), 409-416.

Ayma, V., Ferreira, R., Happ, P., Oliveira, D., Feitosa, R., Costa, G., Plaza, A., Gamba, P., 2015. Classification Algorithms for Big Data Analysis, A Map Reduce Approach. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3/W2, 17-21.

Barry, R., 2006. The status of research on glaciers and global glacier recession: a review. *Progress in Physical Geography: Earth and Environment*, 30, 285-306.

Bekkerman, R., Bilenko, M., Langford, J., 2012. Scaling up Machine Learning: Parallel and Distributed Approaches. Cambridge University Press.

Bolch, T., Pieczonka, T., Benn, D., 2011. Multi-decadal mass loss of glaciers in the Everest area (Nepal Himalaya) derived from stereo imagery. The Cryosphere, 5, 349-358.

Buyya, R., Broberg, J., Goscinski, A., 2011. Cloud Computing: Principles and Paradigms. New Jersey, John Wiley & Sons.

Callegari, M., Marin, C., Notarnicola, C., 2017. Multi-temporal and multi-source alpine glacier cover classification. *9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*.

Chi, M., Plaza, A., Benediktsson, J., Sun, Z., Shen, J., Zhu, Y., 2016. Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104(11), 2207-2219.

Cui, D., Wu, Y., & Zhang, Q., 2010. Massive spatial data processing model based on cloud computing model. Third International Joint Conference on Computational Science and Optimization.

Ghamisi, P., Yokoya, N., Li, J., Liao, W., Liu, S., Plaza, J., Rasti, B., Plaza, A., 2017. Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 5(4),37-78.

Houghton, J., Ding, Y., Griggs, D., Noguer, M., van der Linden, P., Dai, X., Maskell, K., Johnson, C., 2001. Climate Change 2001: The Scientific Basis. Cambridge, Cambridge University Press.

IPCC Working Group I Technical Support Unit., 2013. Climate Change 2013: The Physical Basis. USA, Cambridge University Press.

Kääb, A., Berthier, E., Nuth, C., Gardelle, J., Arnaud, Y., 2012. Contrasting patterns of early twenty-first-century glacier mass change in the Himalayas. *Nature*, 488, 495-498.

Kääb, A., Bolch, T., Casey, K., Heid, T., Kargel, J., Leonard, G., Paul, F., Raup, B., 2014. Glacier Mapping and Monitoring Using Multispectral Data in Global Land Ice Measurements from Space (Springer Praxis Books). Berlin, Springer.

Kaser, G., Ames, A., Zamora, M., 1990. Glacier fluctuations and climate in the Cordillera Blanca, Peru. *Annals of Glaciology.*

Ke, L., Ding, X., Zhang, L., Shum, C., Hwang, C., Luo, Y., 2016. Remote sensing of glacier distribution and change over the Qinghai-Tibet Plateau. *4th International Workshop on Earth Observation and Remote Sensing Applications.*

Kurban, H., Jenne, M., Dalkilic, M., 2016. EM*: An EM algorithm for big data. *IEEE International Conference on Data Science and Advanced Analytics.*

Li, J., Benediktsson, J., Zhang, B., Yang, T., Plaza, A., 2017. Spatial technology and social media in Remote Sensing: A survey. *Proceedings of the IEEE*, 105(10), 1855-1864.

Microsoft Azure, 2019. What is cloud computing?. Microsoft From: https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/

Muhammad, S., Gul, C., Javed, A., Muneer, J., Muhammad, M., 2013. Comparison of glacier change detection using pixel based and object based classification techniques. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium.*

Paul, F., Bolch, T., Kääb, A., Nagler, T., Nuth, C., Scharrer, K., Shepherd, A., Strozzi, T., Ticconi, F., Bhambri, R., Berthier, E., Bevan, S., Gourmelen, N., Heid, T., Jeong, S., Kunz, M., Rune, T., Luckman, A., Van Niel, T., 2015. The glaciers climate change initiative: Methods for creating glacier area, elevation change and velocity products. *Remote Sensing of Environment*, 162(1), 408-426.

Racoviteanu, A., Paul, F., Raup, B., Khalsa, S., Armstrong, R., 2010. Challenges and recommendations in mapping of glacier parameters from space: results of the 2008 Global Land Ice Measurements from Space (GLIMS) workshop, Boulder, Colorado, USA. Annals of Glaciology.

Raup, B., Racoviteanu, A., Singh, S., Helm, C., Armstrong, R., Arnaud, Y., 2007. The GLIMS geospatial glacier database: A new tool for studying glacier change. *Global and Planetary Change*, 56(1-2), 101-110.

Srinivasan, A., 2014. Cloud Computing: A Practical Approach for Learning and Implementation. New Delhi, India: Dorling Kindersley.

United States Geological Survey, 2019. USGS - EarthExplorer - Home. USGS. From: https://earthexplorer.usgs.gov/

Vignon, F., Arnaud, Y., Kaser, G., 2003. Quantification of glacier volume change using topographic and ASTER DEMs. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium.*

WEKA, 2019. Weka 3: Data Mining Software in Java. MLG at University of Waikato. From: https://www.cs.waikato.ac.nz/ml/weka/

Willis, M., Melkonian, A., Pritchard, M., Rivera, A., 2012. Ice loss from the Southern Patagonian Ice Field, South America, between 2000 and 2012. *Geophysical Research Letters*, 39(17).

Winsvold, S., Kääb, A., Nuth, C., 2015. Regional glacier mapping from time-series of Landsat type data. *8th International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multi-Temp).*

Yue, L., Shen, H., Yu, W., Zhang, L., 2018. Monitoring of Historical Glacier Recession in Yulong Mountain by the Integration of Multisource Remote Sensing Data. *IEEE Journal of Selected Topics in Applied Earth Observations and RemoteSensing*, 11(2), 388-400.