# Weekly Journal

Leila Uy

June 7, 2021

## 1 Work Update

A large bulk of my week was spent learning AWS, starting my Literature Review Submission, and researching more articles [1, 2, 3, 4].

### 1.1 AWS

Jishnu and I collaborated in starting our first instance on AWS using the free tier t2.micro to familiarize ourselves with instance creation and Linux commands. The t2.micro only has one processor so it is not good for parallel processing but it helped me with navigating the E2 console. For example, when I was learning about establishing a possible GUI for an instance, I discovered that the Security Groups tabs in the console act as a virtual firewall. It uses the rules to determine if it should allow traffic into and out of our instance (e.g., RStudio Servers). I created an excel sheet to help keep track of instance types we could use to do experiments on. I think m6g.xlarge, m4.xlarge, or c4.xlarge are the best starter instance types.

### 1.2 Literature Review Submission

Mrs. Dr. Maddalena showed us some techniques to use for our Literature Review Submission which I used to establish my final goal and four trends that I discovered from the articles I have read since the beginning of the session.

**Research Goal:** I want to create a parallel k-means clustering R package to help create a consistent, accurate, and efficient clustering of ecoregions using large spatial data sets. We can use this classification to preform analysis on high resolution data sets of North Carolina to determine if there is a significant shifting of ecoregions as a result of climate change.

| Article Trends | | | | |
|---|---|---|---|---|
| Source | Climate change and ecoregions | Ecoregion consistency | Parallel programming | Clustering efficency/accuracy |
| Kumar et al. | _ | X | X | _ |
| Ellis et al. | _ | X | _ | _ |
| Olson et al. | _ | X | _ | _ |
| Pathak et al. | X | _ | _ | _ |
| Hofierka et al. | _ | _ | X | _ |
| Khan et al. | _ | _ | _ | X |
| Alguliyey et al. | _ | _ | X | _ |
| Pourahmad et al. | _ | X | _ | _ |
| Global Eco. | X | _ | _ | _ |
| Hargrove et al. | _ | X | _ | X |
| Luke et al. | _ | _ | X | _ |
| Gbadamosi et al. | X | _ | _ | _ |
| Tang et al. | _ | _ | X | X |
| Jain | _ | _ | _ | X |

Table 1: Literature Review Synthesis

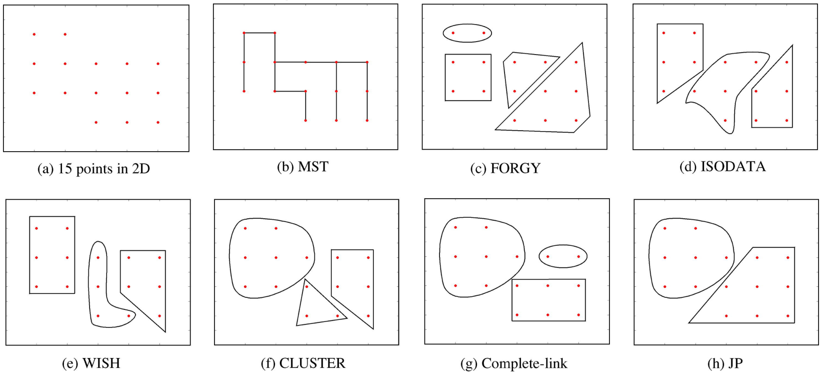
**Trends:**

- Climate change is affecting agriculture and causing shifts in eco-regions
- No consistent and replicable classification of determining eco-regions
- Parallel programming makes analysis (e.g., k-means) of big data faster
- Clustering can be made more accurate and efficient

## 2 Literature Review

60 years after k-means clustering was established, it is still one of the most popular clustering algorithms, but several problems arise from clustering and as a result, several extensions of k-means clustering are created. For example, there is the Fuzzy cmeans which allows each data point to be a member of multiple clusters using membership value. Different clustering algorithms will often result in different partitions, so it is important to choose an algorithm that suits our purpose. [2].

Figure 1: Result of several clusterings of fifteen patterns in two dimensions from Jain 2010 [2]



(a) 15 points in 2D

(b) MST

(c) FORGY

(d) ISODATA

(e) WISH

(f) CLUSTER

(g) Complete-link

(h) JP

For my research goal, I want to determine changing eco-regions using large data sets. As we progress technologically, our data grows in size and cause problems for iterative algorithms like k-means clustering which will exponentially increase in time complexity. Therefore, we want to create an efficient and consistent parallel R program. [4] An article by Tang et al. proposed a similar algorithm to Kumar by proposing an elimination of unnnecessary calculations through triangular inequality. This article set out to prove this theory using mathematical proofs and proved that extreme point distance calculations can be quickened by using Manhattan distance. [3]

## References

[1] Gbadamosi Babatunde, Adeniyi Abidemi Emmanuel, Ogundokun Roseline Oluwaseun, Oladosu Bukola Bunmi, and Anyaiwe Ehiedu Precious. Impact of climatic change on agricultural product yield using k-means and multiple linear regressions. *International Journal of Education and Management Engineering*, 9(3):16–26, 2019.

[2] Anil Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[3] Zhuo Tang, Kunkun Liu, Jinbo Xiao, Li Yang, and Zheng Xiao. A parallel k-means clustering algorithm based on redundance elimination and extreme points optimization employing mapreduce. *Concurrency and computation*, 29(20), 2017.

[4] Luke Tierney, Rossini Anthony, and Na Li. Snow: A parallel computing framework for the r system. *International Journal of Parallel Programming*, 37(1):78–90, 2008.