Weekly Journal

Leila Uy

June 28, 2021

1 Work Update

I'm trying to be more active on Teams with updates and issues I encounter because I know that Valerie and Diljot are roadblocked by quite a bit of coding issues, so I want everyone to know the progress that is happening on my end. If it gets too much, just tell me to tone it back, but that also means a lot of things I will write in my Journal will be things you might already know. This is is going to be an especially long weekly journal as a summary, I apologize.

1.1 Project Board on GitHub

This week we set up a project board on GitHub. When I get the chance, I usually update the project board with new tasks or problems I encounter. Recently, I archived all the old k-means clustering issues and created a handful of new ones. The one Jishnu and I are both working on is calling the k-means clustering function from the stats library in parallel using foreach or parallel apply. With all the problems we encounter with R, AWS, and parallelization, it helps me organize my thoughts.

1.2 aws.s3 Library

I created the simple aws.s3 tutorial for everyone so that everyone knows how to setup the connection with S3. I struggled for awhile on Sunday with getting the data from an S3 bucket into memory but I was able to solve that for importing shapefiles using:

```
shapefile <- s3read_using(FUN = readOGR,
    object = "data/shapefiles/n_c_boundary.geojson", bucket = "democluster")
```

But I was only able to get this to work with type geojson because when you use readOGR, it expects either a dsn (for shapefiles with .prj, etc.) or a geojson, but when you use s3read_using, you have to give it the key of a single file. Therefore, we might have to use geojsons for boundaries.

Now in Teams, I posted a problem I encountered when trying to stack the rasters after putting them into memory from S3. The main idea of what I was trying to accomplish is using get_bucket to get a list of all the raster keys which I would then loop s3read_using on to get the rasters into memory. The problem is I get an error when I try to stack these rasters.

Once this problem is solved, all the steps after should be similar to using local files because we do not have to use the bucket until we have to save our output. That should not be too much of a problem because writing into a bucket seems fairly easy when I made the tutorial.

1.3 Parallel K-Means

Once I got stuck using aws.s3, I decided to try implementing parallel k-means using sftp to copy files from my local drive to an EC2 instance. I currently have an m4.xlarge instance launched in EC2. If you start the instance, you can see R is installed and you will find 3 main folders which contain world clim 10m, a north carolina shapefile, and an R script. When I ran the Rscript in the instance, it worked up till running the raster to dataframe line. In order to find out why k-means was not working I ran the script locally and got an error because it was not converging after 10 iterations. Therefore, I had to define the variable nstart in kmeans, so I am wondering how you were able to get the code base working without nstart because I do not want to have to keep defining nstart for different centers. In additon, I

removed the coordinates but I do not know how to attach the coordinates back again so we can create the final cluster raster.

2 Literature Review

I am going to be honest, I wanted to focus this week and next week on getting parallel k-means clustering working with S3 and EC2. This means that I was not able to find a lot of time to actually read the articles I found [1, 2, 3, 4]. I am changing the focus of my project to a study region so although my Draft Literature Review from last week will be helpful in my final product, I will have to rewrite some paragraphs to include my shift in focus from methodology to analysis. Rather than continuing my readings on k-means/parallelization, I will devote my next few readings to possibly United States agriculture and shifting ecoregions.

How I see the breakdown of my new focus and format of my Literature Review will go:

1. California's major commodity is in agriculture so their economy is vulnerable to any agricultural changes
2. Climate change is affecting agriculture by [insert reasons e.g., IPS, growing season]
3. We need to study these agricultural shifts using ecoregions but there is no set classification method
4. With larger data sets, we also need to make the methodology of our classification quicker

Purpose: to use parallel programming and prior ecoregion classification literature to predict the impact of climate change on California's agricultural sector.

It's very similar to my already created Literature Review with the exception of the 1st topic and more in-depth research into the 2nd. Any comments or feedback of this idea in shifting the focus would be helpful!

References

[1] K M Havstad, J R Brown, R Estell, E Elias, A Rango, C Steele, David S Gutzler, Connie J Maxwell, and Connie J Maxwell. Vulnerabilities of southwestern u.s. rangeland-based animal agriculture to climate change. *Climatic change*, 148(3):371–386, 2018.

[2] David Kroodsma and Christopher Field. Carbon sequestration in california agriculture, 1980-2000. *Ecological applications*, 16(5):1975–1985, 2006.

[3] David Lobell, Angela Torney, and Christopher Field. Climate extremes in california agriculture: California second assessment: New climate change impact studies and implications for adaptation. *Climatic change*, 109, 2011.

[4] J Reilly, F Tubiello, B McCarl, D Abler, R Darwin, K Fuglie, S Hollinger, C Izaurralde, S Jagtap, J Jones, L Mearns, D Ojima, E Paul, K Paustian, S Riha, N Rosenberg, and C Rosenzweig. U.s. agriculture and climate change: New results. *Climatic change*, 57(1):43–67, 2003.