DETERMINING THE OPTIMAL NUMBER OF GRID CELLS

Kenneth Annan

INTRODUCTION

In this project, I tend to advise the farmer on the optimal number of grid cells and then compute a yield estimate for each cell. The data for the project includes samples for the field for the years 2013, 2015, 2016, 2017, and 2018. There are four main variables in each year. However, the Timestamp Column will help me decide on the number of data set to use. I found the Timestamp harvest interval to be less than 1-week (7 days) after screening the data, thus, this indicates that I can use all five data sets, but for clarity, I dropped the 2018 data set. The variables that my analysis will focus on are Latitude, Longitude, and Yield. I will then merge the data by grid cell and compute a normalized yield estimate and standard deviation for each cell across years. Lastly, I will use these estimates to classify cells as having High, Average, or Low yields; and as having Stable, Average, or Unstable yields. I finally plot my outcome or results based on the classification of the yields of the normalized mean and the normalized Standard deviation.

METHODOLOGY

Statistical distribution of the 2013 data set

	n	mean	sd	median	min	max	range	se
X	21612	10806.5	6239.0	10806.5	1.0	21612.0	21611.0	42.4
Longitude	21612	551.9	266.3	553.8	4.7	1007.0	1002.4	1.8
Latitude	21612	334.8	176.6	351.0	0.0	621.2	621.2	1.2
IsoTime*	21612	10795.5	6232.8	10795.5	1.0	21590.0	21589.0	42.4
Yield	21612	55.6	14.6	57.5	0.0	480.7	480.7	0.1
SWATHWIDTH	21612	5.0	0.0	5.0	5.0	5.0	0.0	0.0
Swaths	21612	7.0	0.0	7.0	7.0	7.0	0.0	0.0
DISTANCE	21612	5.7	0.6	5.8	0.1	8.4	8.3	0.0
Heading	21612	248.8	99.4	180.6	0.0	360.0	360.0	0.7
WetMass	21612	3339.2	878.8	3449.8	0.0	28844.7	28844.7	6.0
Moisture	21612	11.6	1.2	11.4	0.0	30.0	30.0	0.0

Step 1

I divide the field into grid cells with 20 rows for the Latitude and 6 columns for the Longitude with grid cells that are 100m wide and 20m long making a total of 120 rows in total. In this step, I calculate the grid cell means using the aggregate function. The aggregation is done by the grid cell name with the function mean.

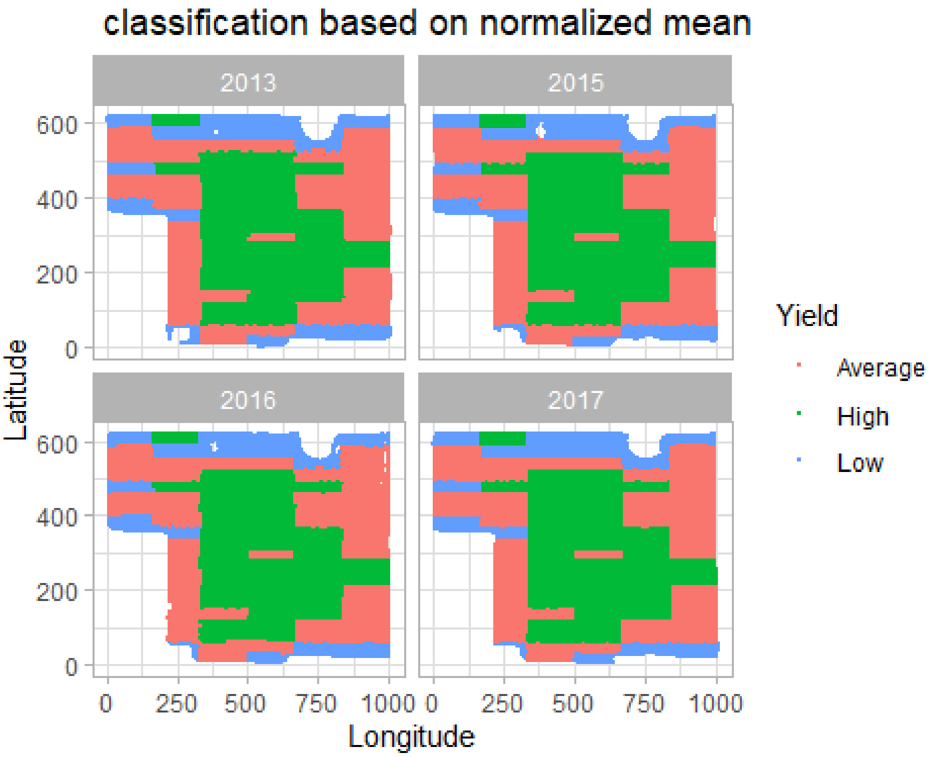
Also, I convert the data to a common scale since there are very different means. I normalize yield using option 2(the Z score method) and compare the normalization of the grid cell estimates with the normalization of the yield sample values. Option 2(the Z score method) follows the assumption that the data is from a normal distribution and that there is the need to calculate for skewness and kurtosis to check the assumptions.

I further calculate the yields of the normalized mean and the standard deviation of the normalized mean. The skewness value for all years is close to zero and the kurtosis value for all the years is also close to 3. Thus, checking for skewness and kurtosis of these data, the data is from a normal distribution and that the Z score has no flaw when using it.

step 2

In this step, I classify the yields of the normalized mean and the normalized Standard deviation. I classify a yield as classifying this as a "High yielding cell" if the mean normalized score for a grid cell is in the largest 25 percent of all cells, "Low yielding" if the mean normalized score is in the smallest 25 percent, and "Average yield" Otherwise.

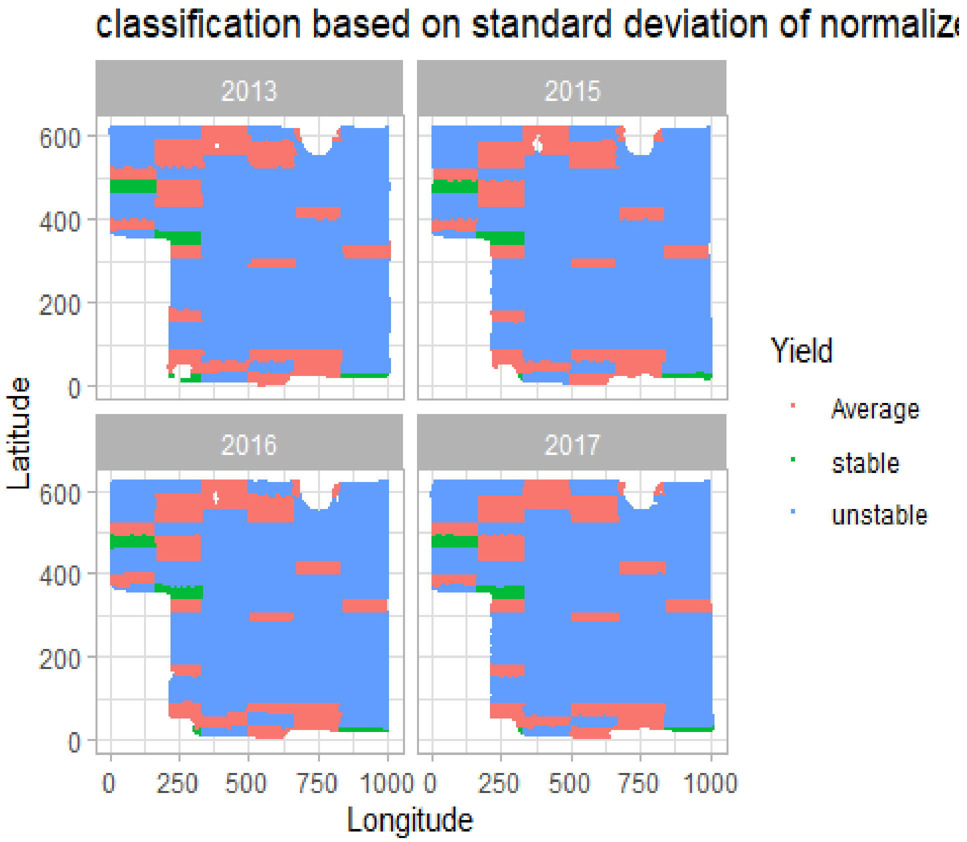
I repeat the same process for the normalized Standard deviation. But in this I classify a yield as "stable" if the standard deviation of the normalized scores for a grid cell is in the largest 25 percent of all cells, "unstable" it is in the smallest 25 percent, and "Average yield" Otherwise. The results of this process have produced the graph below.



The above graph shows the normalized mean yield for all the years. The plots indicate yield as low, high, and the average for all five years. The distinguishing factor in the plot is that the 2015 data file has a drainage line as shown from the graph.

step 3

In this last step, I produce a graph to illustrate the classification based on a standard deviation of normalized means. The graph is shown below.



The above graph shows the standard deviation of the normalized mean for all the years. The plots indicate yield as stable, unstable, and average for all five years. The distinguishing factor in the plot is that the 2015 data file has a drainage line as shown from the graph.

CONCLUSION

What I have learned from the project is that the harvest interval for all the years is less than 7 days. An attempt to reduce the constraints to 3 days will limit the data set that needs to use for this project. It can also be seen that since all the data files use the same row identifiers of 120 rows, this makes the plot for all the years to be somewhat the same. Thus, using the different optimal number of grid cells will change the results of this project.