Final Project, STAT 600 2020

Overview

We will continue the analysis started with the midterm project. We will use the three data sets from the midterm and two additional data sets for the years 2013 and 2018. We will divide each data set into grid cells, then compute a yield estimate for each cell. We will then merge the data by grid cell and compute a normalized yield estimate and standard deviation for each cell across years. We will use these estimates to classify cells as having High, Average or Low yields, and as having Stable, Average or Unstable yields.

Grid Cell Size

I've reviewed the midterm proposals, and I've decided on a grid cell size. I've also measured the farm equipment, and it's my opinion that grid cells 100 m wide (`Longitude`) are best for management purposes. Thus, the field will be divided into 6 columns. You may determine column number as you wish, but I've used the code

```
harvest.dat$Col <- ceiling(6*harvest.dat$Longitude/600)
```

If you wish, you may repeat the calculations from the midterm given this new constraint on grid cell width to determine an optimal cell length (`Latitide`), given an assumed difference of 10%. You may also refer to the your midterm submission. My preference is to have 20 rows, so for the sake of this discussion we'll assume the field is to be divided into a grid with 20 rows and 6 columns with grid cells that are 100m wide and 20m long.

Data sets

I've added two additional data files, from the years 2013 and 2018. Before we merge the data we must process the data. We will require two operations before we merge the data sets.

Data screening

Before we merge the five data sets, we will screen to determine if the data were uniformly sampled. The data files uploaded to D2L in the Final Project directory include a data column `TimeStamp`. This column contains date-time values in the form

```
Y-m-d H:M:S
```

where `Y` is year, `m` is month (`1-12`), `d` is day (`1-31`), `H` is hour (`0-23`), `M` is minutes (`0-59`) and `S` is seconds (`0-59`). Process the text in this column to determine the harvest interval for each data set; that is, the difference in days from the earliest time stamp to the latest time stamp. We will use a data set if the harvest interval is less than 1 week (7 days).

You may process the text in this column at your discretion; you might consider parsing the text manually, or you might covert the text to a DateTime instance (see `?DateTimeClasses`).

Normalization

We have three different crops, and these may have very different means, so we need convert the data to a common scale. Denote the i^{th} `Yield` observation for Year j as y_{ij}, we normalize yield by one of the following methods, in each case holding j constant and iterating over i only within years. If we assume 20 rows and 6 columns, then $y_{ij} = \{y_{1j}, y_{2j}, \ldots, y_{N_i j}\}$ where $N_i = 120$. Similarly, we would denote the successive yield estimates for grid cell i as $y_{ij} = \left\{y_{i1}, y_{i2}, \ldots, y_{iN_j}\right\}$ where $N_j = 5$.

Note that we do not have an index for the yield samples within each cell. You may, but are not required, compare normalization of the grid cell estimates with normalization of the yield sample values.

You may choose a normalization method at your discretion. We've listed some possible normalization formula below. You are not required to implement all three, but you must use some method to convert yield values across different crops to a common scale. You may choose to compare the different methods; they have different statistical properties and may lead to different conclusions.

Option 1. Rank

Replace y_{ij} with $r_{ij} = \text{rank}(y_{ij})$. Determine ranks independently for $j = 1, 2, \ldots, N_j$ for years $\{2013, 2015, \ldots, 2018\}$

Option 2. Z-score

Calculate

$$\overline{y}_{.j} = \frac{\sum_{i=1}^{N_i} y_{ij}}{N_i}$$

and

$$s^2_{.j} = \frac{\sum_{i=1}^{N_i} (y_{ij} - \overline{y}_{.j})^2}{N_i - 1}$$

where N_i are the number of `Yield` values for year j. Replace y_{ij} with

$$z_{ij} = \frac{y_{ij} - \overline{y}_{.j}}{s_{.j}}$$

.

Calculate $\overline{y}_{.j}$ and $s^2_{.j}$ independently for $j = 1, 2, \ldots, N_j$ for years {2013, 2015, ..., 2018}. Note that this method makes use of the first (mean) and second moments (variance). It would be best practice to check for skewness or kurtosis of these data.

Option 3. Percent

Replace $y_i j$ with

$$100 \times \frac{y_{ij}}{\overline{y}_{.j}}$$

Calculate $\overline{y}_{.j}$ independently for $j = 1, 2, \ldots, N_j$ for years {2013, 2015, ..., 2018}. Note that this method assume the arithmetic mean is a reasonable estimate of central tendency. It would be best practice to check for skewness or kurtosis of these data.

Merged Data

Merge the data sets by grid cell number. For each grid cell i, calculate a normalized mean and a standard deviation over the N_j estimates. That is, if n_{ij} is the normalized yield value for grid cell i in year j, then calculate

$$\overline{n}_{i.} = \frac{\sum_{j=1}^{N_j} n_{ij}}{N_j}$$

and

$$s_{i.}^2 = \frac{\sum_{j=1}^{N_j} (n_{ij} - \overline{n}_{i.})^2}{N_j - 1}$$

Classification

Classify each grid cell according to the following criteria.

If the mean normalized score for a grid cell is in the largest 25% percent of all cells, classify this as a `High` yielding cell. If the mean normalized score is in the smallest 25%, classify this cell as `Low` yielding. Otherwise, classify the cell as `Average` yield.

Similarly, if the standard deviation of the normalized scores for a grid cell is in the largest 25% percent of all cells, classify this as an `Unstable` yielding cell. If the standard deviation of the normalized scores is in the smallest 25%, classify this cell as `Stable` yielding. Otherwise, classify the cell as `Average` yield.

We do not want a table wth 120 rows. Instead, produce a plot with either `Column` or `Longitude` as the independent variable and `Row` or `Latitude` as the dependent variable. Plot using points, and use different point color or style to distinguish the grid cells class. Produce one graph to illustrate the classification by normalized mean, and one for the classification based on standard deviation of normalized means.