**Work Files Instruction**

Zongqi Yu

**Part1: Test Files**

Within the zip file, there are three folders – “***PureReal***”, “***SimResult***”, and ***“ErrorDistribution***”. The following is an instruction on how to use the files in them.

1. **PureReal**

This folder is composed of all real data. The data comes from three years: 2013-2014, 2014-2015, and 2016-2017. However, 2016-2017 data are not tested. Depend on the naming of the files, the files can be categorized into two types: full matrix data, and randomly punched out matrix data.

Full matrix data are always named using the word “***Original***”, for instance: “***RPunch1314\_Original.csv***”.

Randomly punched out matrix data is named as “***RPunch1314\_25\_r1.csv***”, for example. There are several components in the naming. First, “RPunch” means that the missing items are randomly punched out. “1314” represent the year the dataset is from. “25” means that 25% of the dataset has been punched out. “r1” means run 1, and there are totally three runs for each missing percentage level (25%, 35%, 45%).

Something worth noting is that, in the randomly punched out test files, the matrix entries of “0” are the **OBSERVED** and the matrix entries with numbers are **UNOBSERVED**.

1. **SimResult**

This folder is composed of simulated data using WOFOST R version from cycle 2013-2014 and 2014-2015. The meaning of each wheat line can be referred using the picture below. In testing, these 18 lines of simulated lines are usually attached at the end of the matrix. More information about the simulation is in the WOFOST simulation R file “**WofostSim.Rmd**”.

A close-up of a notebook

Description automatically generated

1. **Error Distribution**

Files in this folder records all error from every one of the missing entries. The values recorded is . The files are categorized in three types: RPunch (randomly punched out), ColumnMiss (consecutive punched out in one column), Asym (asymmetrically punched out the entries higher than one standard deviation from the mean). For instance, “***KNN\_1314\_45\_doc.csv***” means “using KNN method”, “using the data from 13-14 cycle”, “with 45% of data missing”. For files in the “**Asym**” folder, “asym1” and “asym2” means merely “part1” and “part2”.

**Part2: Test Results**

The excel sheets’ name are very straight-forward about what the worksheet is about. However, here are more information to note:

1. In the randomly punched out test sheet, the error metrics is the **corrected** one: **sum(abs(1-pred/obs))** and divided by the number of missing entries.
2. In the simulated data attached test sheet, the error metrics is the old one: **(pred-obs)^2** and divided by the number of missing entries. This is because the simulated data we obtained are too poor, so that we **did not** use corrected error metric to run this test again.
3. In the Error Distribution sheet, the result is derived from recorded error metrics in **1-pred/obs, with no absolute sign**.
4. The source records for the statistics **haven’t divided by the number of missing entries yet**.

**Part3: Test Codes**

All of the test codes ended by .ipynb should be opened using **google colab**. Codes ends with .Rmd should be opened using **R**.

**random\_punch\_updated\_error\_metrics.ipynb**: includes the code for SVT, Soft Impute, Avg, and KNN methods for data missing by randomly punched out.

**column\_miss\_updated\_error\_metrics.ipynb**: Includes the code for SVT, Soft Impute, Avg, and KNN methods for data missing consecutively in a column.

**Simple\_Method\_with\_error\_calculation.ipynb**: Simple Method for data missing consecutively in a column.

**Python-IBCF.ipynb**: IBCF method for data missing in both consecutively in a column or randomly punched out.

**R-IBCF.Rmd**: This is used to collect the randomly selected items for punching out. However, in our records, all the error data are calculated **based on the Python method**.