An Introduction to the MergeGUI Package

Xiaoyue Cheng * Dianne Cook Heike Hofmann
Department of Statistics Iowa State University

January 27, 2014

1 Introduction

The MergeGUI package intends to visualize the process of merging multiple datasets. The mergence can be directed vertically – merging the variables, or horizontally – merging the cases, or say, merge the IDs.

2 Design of GUI

Beginning with the command

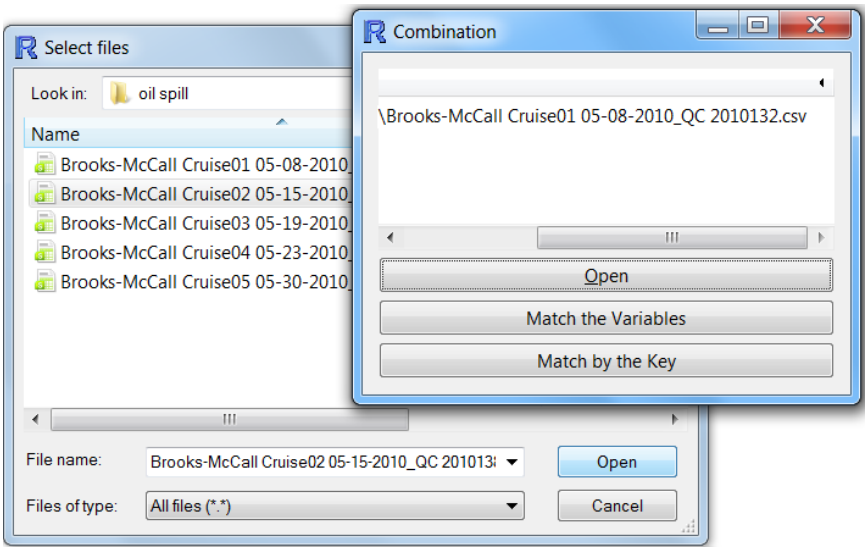
```
> library(MergeGUI)
> MergeGUI()
```

the starting interface will show up. The following instruction will give a tour in this package.

2.1 Starting Interface

The users can open files in the starting interface and select some files to do the next command. They can match the variables or match the observations by the key. At least two files need to be selected to match. If no files are specifically chosen, then all the files are used in the next step.

*Email: xycheng@iastate.edu

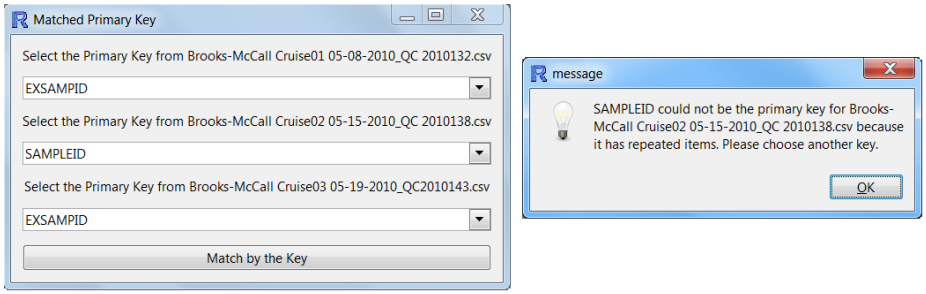


Reading the data frames from R console is also allowed. The code below could also give the starting window.

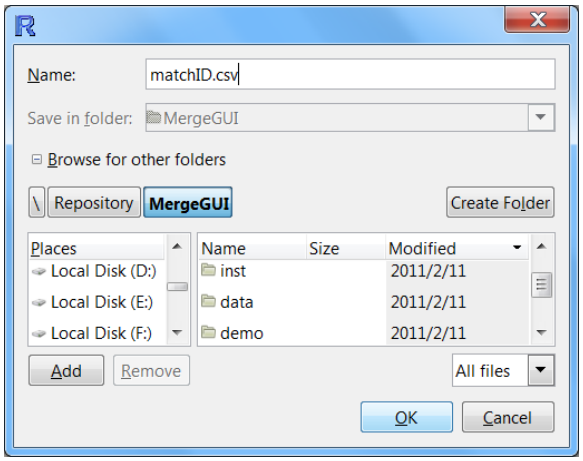
```
> data(iris)
> setosa=iris[iris$Species=="setosa",1:4]
> versicolor=iris[iris$Species=="versicolor",1:4]
> virginica=iris[iris$Species=="virginica",1:4]
> MergeGUI(setosa,versicolor,virginica)
```

2.2 The Matching-ID Interface

Under some situations, we have several files containing the same subjects or experimental units, but different treatments or recording times. Then one of the desirable tasks is to combine those files by the subject ID's. In the 'Combination' window, the last button is 'Match by the Key'. Clicking it we'll see the window 'Matched Primary Key'. With each file we select, one primary key is needed. A warning message will pop up if the variable we selected is not a proper key, for example if some of its records are replicated.



Upon clicking the button 'Match by the Key', we can save the merged data into a new csv file.

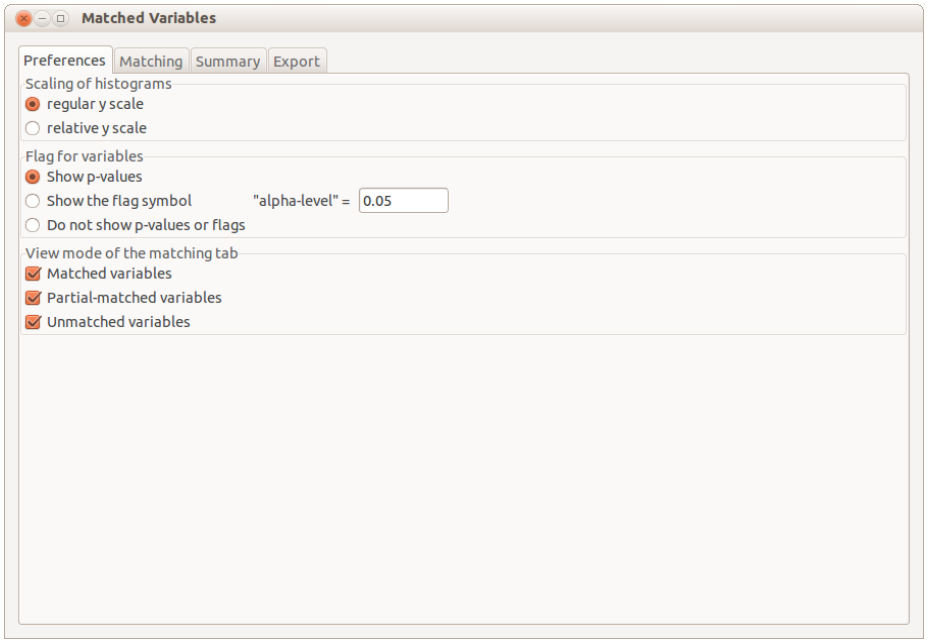


2.3 The Matching-Variable Interface

This part allows users to match the data from the different experimental units that have the same variables. In the 'Combination' window, the second button is 'Match the Variables'. Clicking it will open the window 'Matched Variables'. This window has four tabs. We can activate one page by clicking the tab title.

2.3.1 Perferences Tab

In the preferences tab, users can choose (1) whether to free the y-scales for different data files when facetting the plots by file; (2) whether to display the numerical p-values or the flag symbols in the summary tab; (3) which part of the variables to show. All the changes will be presented in the Summary tab.

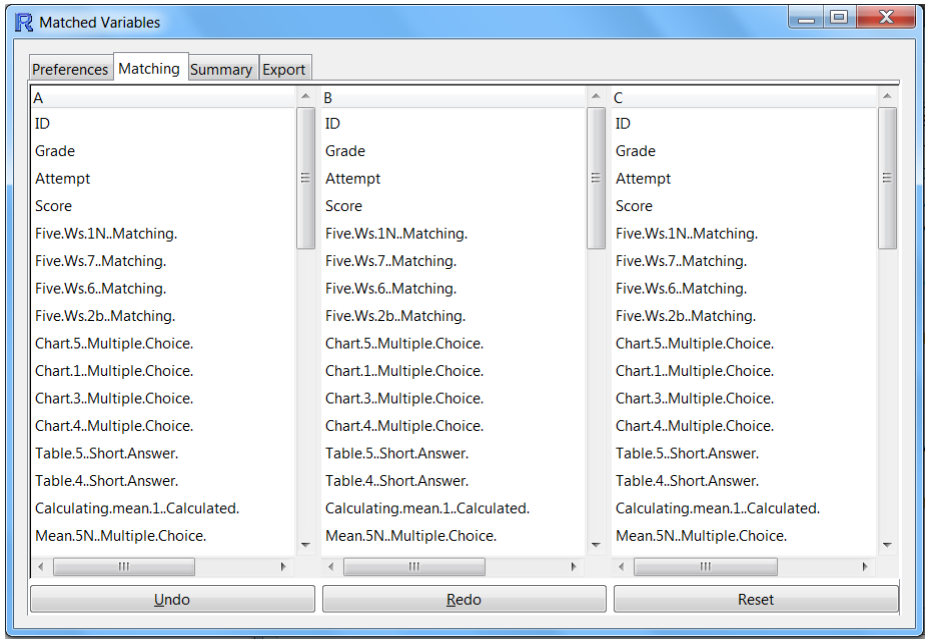


2.3.2 Checking Tab

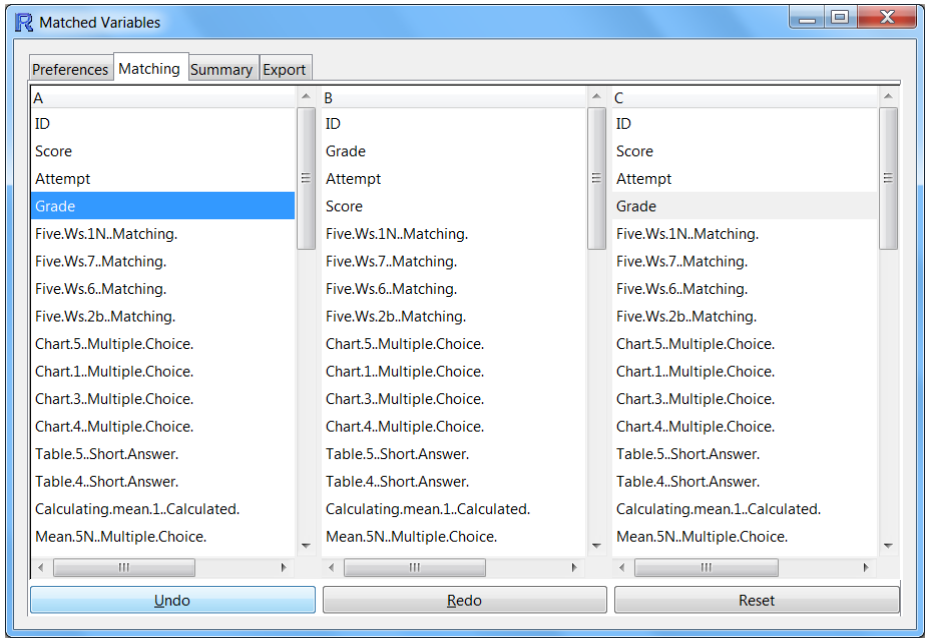
The 'Checking' page shows the automatic variable matching results and allows the user to change the variable matches. The variables present in each file are listed in separate panes. The order that the variable names are listed shows the variable matching.

The functionality is as follows:

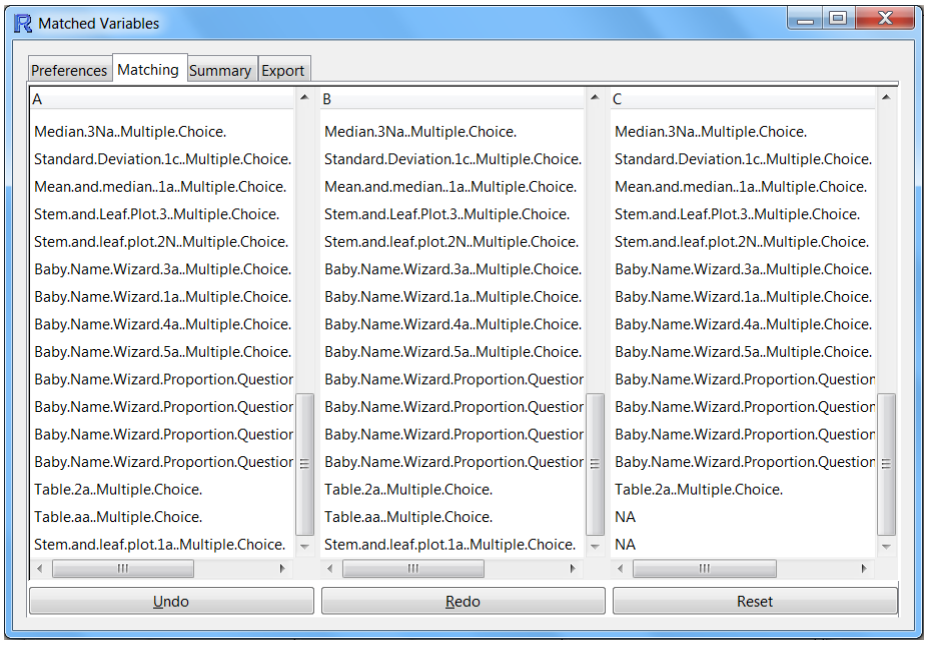
1. Variables are automatically matched by variable name in an initial pass through the data. These matched variables from the different files are aligned along a row.



2. Switch which variables are matched by switching the order of the variables in one file's list. The re-matching is achieved by clicking on two variable names in the same file list. Here the variable Score in studentsB.csv should be matched the variable Grade in the other two files.



3. Unmatched variables are listed at he bottom of the list. Here the last two variables in studentsA.csv, studentsB.csv have no equivalent in studentsC.csv.

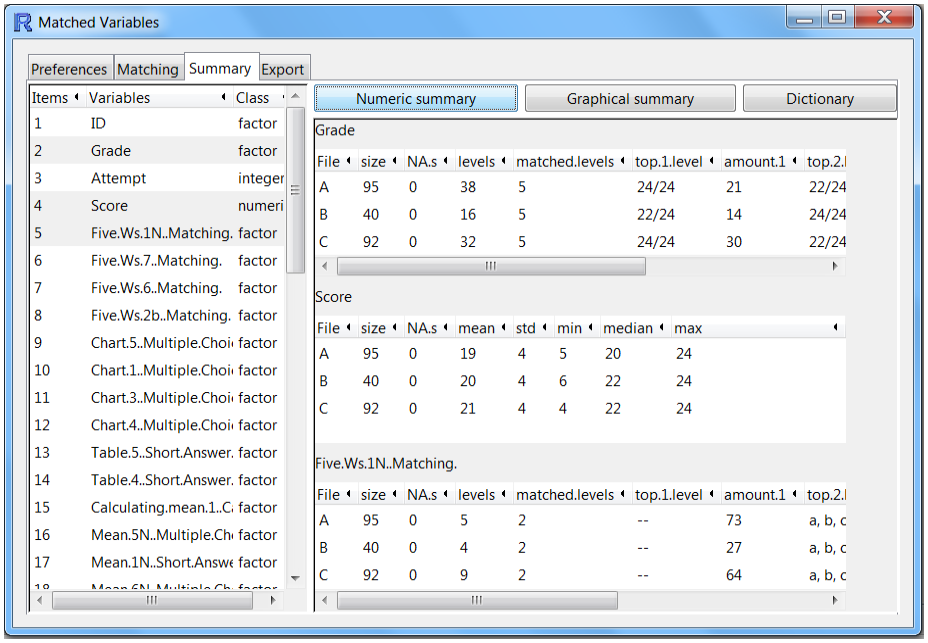


4. It is possible to undo, redo, or reset the matching.

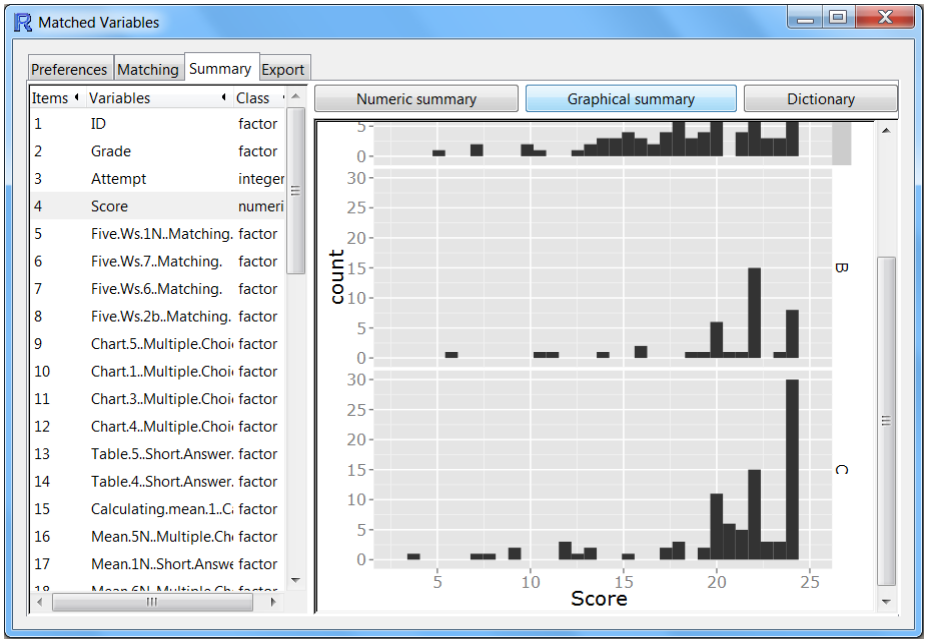
2.3.3 Summary

The 'Summary' page, allows the user to check the values of the variables numerically, or graphically and set up a data dictionary for the categorical variables.

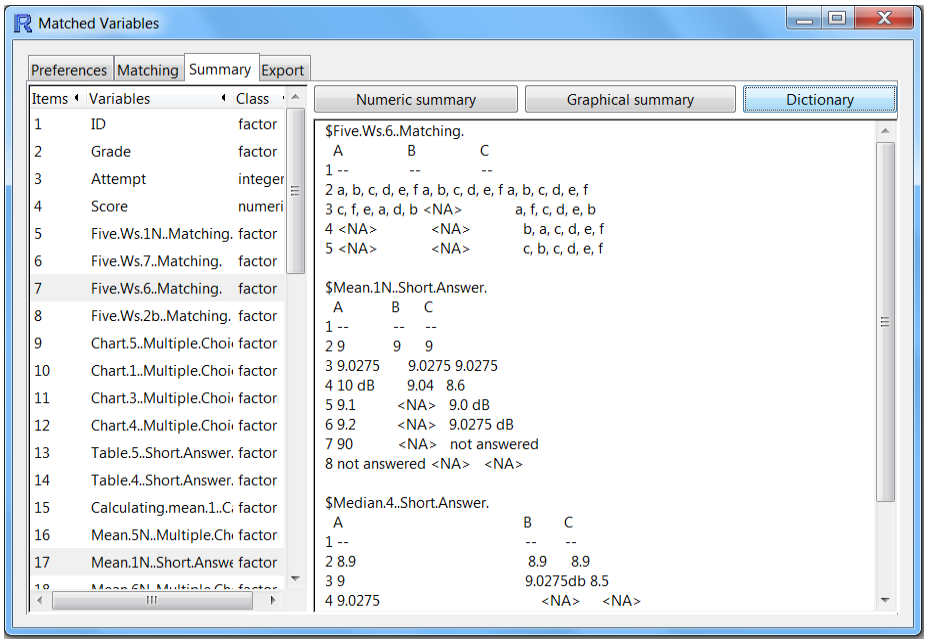
Select the variables that you want to summarize and then click 'Numeric summary' shows numerical information about these variables by file in the right panel. Here the variables Grade and Score have been selected for summary. Grade is categorical because it shows the student score/total score. Factor summaries are used, number of observations, levels, and the three most frequent levels. Score is a numeric variable so the summary statistics are used, mean, standard deviation, minimum, median and maximum, along with number of observations and number of missing values.



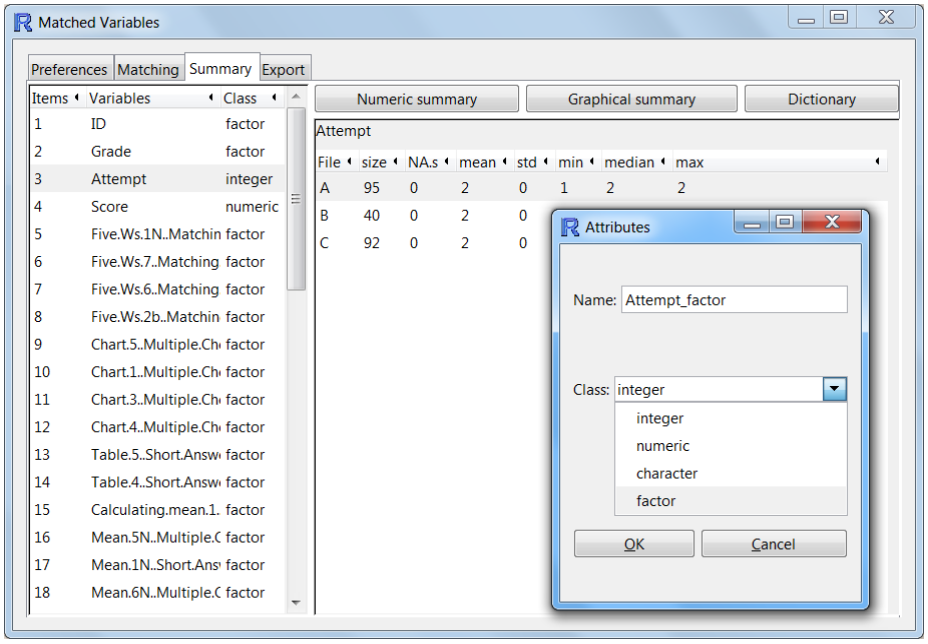
Select a single variable and click 'Graphical summary' to get plots of the variable, facetted by file. For the graphical summary, histogram or barchart will be shown if a single variable is selected. A scatterplot will be drawn if two numeric or two factor varaibles are chosen. Side-by-side boxplots will be presented when one numeric and one factor varaibles are selected. A parallel coordinate plot is shown when all the variables selected are numeric and there are more than two variables. If more than two variables are chosen but the classes of the variables are mixed, i.e. some are numeric, some are factor or character, then histograms and barcharts will be drawn individually.



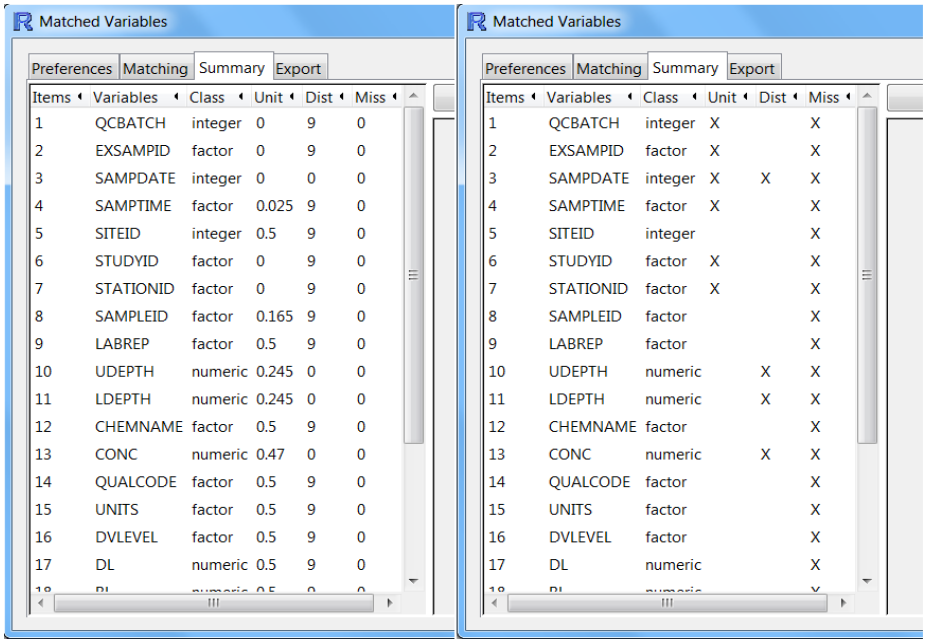
Select several factor variables and click 'Dictionary' to see all levels of the factors. The same levels shown in all files will be listed in front of other levels.



By double clicking one variable, the attributes window will pop up. The user can edit the variable name or change the class of the variable.

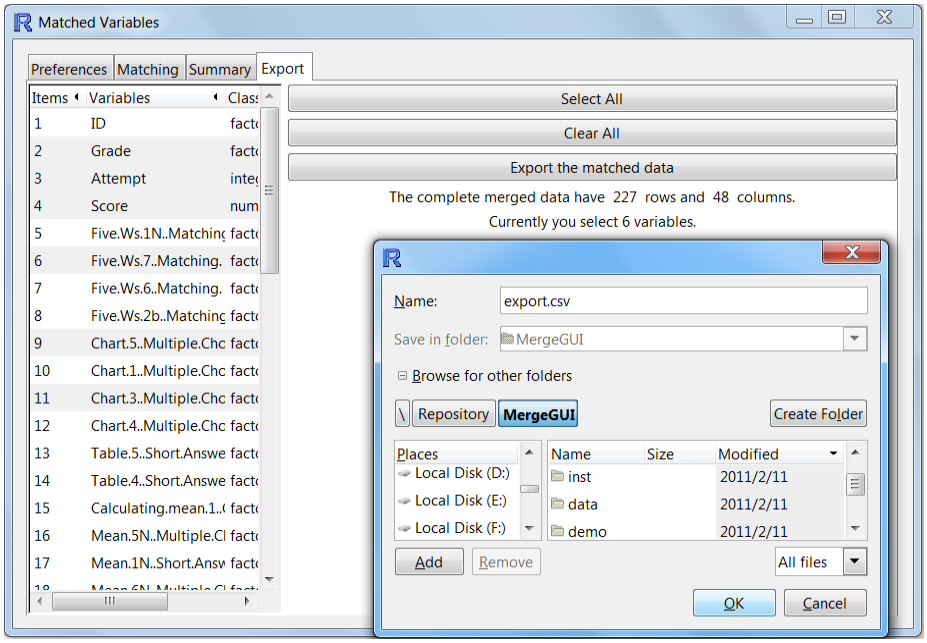


When the Preference tab sets 'Show p-values' or 'Show the flag symbols', the variable list in the Summary tab looks as follows. All the flag symbols correspond to the p-values which are less than 0.05. For each variable, the column 'Unit' gives the misclassification rates for each variable if the user wants to know whether any variable could distinguish the sources correctly. The misclassification rate is calculated through the classification tree (the rpart package). The column 'Dist' compute the p-values of the Kolmogorov-Smirnov tests between different sources. It is used to check whether any variable has different distributions for different sources. 'Miss' gives the p-values of the Chi-square tests for the counts of missing and non-missing values between different files for each variable. So the user could know the pattern of missing values among the sources.



2.3.4 Export

This tab provide the interface for saving the matched data into csv file. The user could select all or none variables by click the two buttons or choose several varaibles by Ctrl+Click. The export button will export the merged data and the numeric summaries of the selected variables into two csv files.


Matched Variables
Preferences
Matching
Summary
Export
Items
Variables
Class
1 ID
2 Grade
3 Attempt
4 Score
5 Five.Ws.1N..Matchin
6 Five.Ws.7..Matching.
7 Five.Ws.6..Matching.
8 Five.Ws.2b..Matchin
9 Chart.5..Multiple.Chc
10 Chart.1..Multiple.Chc
11 Chart.3..Multiple.Chc
12 Chart.4..Multiple.Chc
13 Table.5..Short.Answe
14 Table.4..Short.Answe
15 Calculating.mean.1..
16 Mean.5N..Multiple.Cl
17 Mean.1N..Short.Answ
Select All
Clear All
Export the matched data
The complete merged data have 227 rows and 48 columns.
Currently you select 6 variables.
Name:
export.csv
Save in folder:
MergeGUI
Browse for other folders
Repository
MergeGUI
Create Folder
Places
Local Disk (D:)
Local Disk (E:)
Local Disk (F:)
Name
Size
Modified
inst
2011/2/11
data
2011/2/11
demo
2011/2/11
Add
Remove
All files
OK
Cancel