The user's guide of the type 2 MAD analysis pipeline package
(V1.0, revised in March 2013)

Contents

1. MAD analysis pipeline package

Package components

The MAD pipeline program package includes five programs:

(1) *transpose_MAD_data.pl*: This Perl program converts the MAD data from one table format (Table 2) to another table format (Table 1) which is used in the *MAD_analysis_Step1a.sas* program.

(2) *MAD_analysis_Step1a.sas*: This SAS program performs ANOVA of individual MAD experiments, in which two separate ANOVA for plot and subplot controls are performed. The output from this SAS program is used as input for subsequent analysis using Perl program, *MAD_analysis_Step1b.pl*.

(3) *MAD_analysis_Step1b.pl*: This program summarizes the ANOVA results from the SAS program *MAD_analysis_Step1a.sas*, adjusts the observations of test genotypes and controls, and estimates the RE of different adjustment methods. Ultimately, a data file with values adjusted by the most appropriate method is exported for further analysis. If the same plot and subplot controls are used in multiple experiments, adjusted values of test genotypes and controls can be further used as input of the second SAS program, *MAD_analysis_Step2a.sas*, for joint ANOVA over multiple environments.

(4) *MAD_analysis_Step2a.sas*: This SAS program performs the joint ANOVA of MAD data over multiple environments if the same plot and subplot controls are used in all experiments.

(5) *MAD_analysis_Step2b.pl*: This program calculates the correct F values and performs the significance test based on the ANOVA results obtained from the SAS program *MAD_analysis_Step2a.sas*.

Installation and requirements

The pipeline package does not need installation. Unzip the downloaded MAD pipeline package file to the any folder of your choice. The following software tools and Perl modules are perquisites for successful use of this pipeline package:

(1) SAS software: It is commercial statistical analysis software.

(2) Perl: The Perl software can be freely downloaded from http://www.activestate.com/activeperl/downloads.

(3) Perl module "Statistics::Distributions": It can be downloaded from http://search.cpan.org/~mikek/Statistics-Distributions-1.02/Distributions.pm

(4) Perl module "Statistics::Regression": It can be downloaded from http://search.cpan.org/~iawelch/Statistics-Regression-0.53/Regression.pm

2. Data preparation

The SAS program *MAD_analysis_Step1a.sas* accepts only a tab-separated text file with 12 columns and a head line (Table 1):

Table 1. Raw data format 1 for the MAD analysis pipeline

Record	Plot	Row	Columm	Cp	Csp	Entry	Year	Location	Genotype	Trait	Value
75	4575	2	6	0	0	1	2009	MD	CN18973	OIL	46
75	4575	2	6	0	0	1	2009	MD	CN18973	IOD	191
75	4575	2	6	0	0	1	2009	MD	CN18973	PAL	4.6
75	4575	2	6	0	0	1	2009	MD	CN18973	STE	3.8
75	4575	2	6	0	0	1	2009	MD	CN18973	OLE	20
75	4575	2	6	0	0	1	2009	MD	CN18973	LIO	15
75	4575	2	6	0	0	1	2009	MD	CN18973	LIN	57
75	4575	2	6	0	0	1	2009	MD	CN18973	YIELD	.
169	4669	4	7	0	0	2	2009	MD	CN18979	OIL	47
169	4669	4	7	0	0	2	2009	MD	CN18979	IOD	191
169	4669	4	7	0	0	2	2009	MD	CN18979	PAL	5
169	4669	4	7	0	0	2	2009	MD	CN18979	STE	3.7
169	4669	4	7	0	0	2	2009	MD	CN18979	OLE	20
169	4669	4	7	0	0	2	2009	MD	CN18979	LIO	15
169	4669	4	7	0	0	2	2009	MD	CN18979	LIN	57
169	4669	4	7	0	0	2	2009	MD	CN18979	YIELD	.
261	4761	6	8	0	0	3	2009	MD	CN18980	OIL	44
261	4761	6	8	0	0	3	2009	MD	CN18980	IOD	193
261	4761	6	8	0	0	3	2009	MD	CN18980	PAL	5.4
261	4761	6	8	0	0	3	2009	MD	CN18980	STE	3.1
261	4761	6	8	0	0	3	2009	MD	CN18980	OLE	19
261	4761	6	8	0	0	3	2009	MD	CN18980	LIO	14
261	4761	6	8	0	0	3	2009	MD	CN18980	LIN	58
261	4761	6	8	0	0	3	2009	MD	CN18980	YIELD	.

The raw data file must have 12 columns with a header line. The columns 3-6 and 8-12 (highlighted in yellow) are used for data analysis. Values in other columns are not used for analysis and thus can be any values but the columns are required as a place holder.

Row: the row numbers (index or coordinate) of whole plots starting from 1.
Column: the column numbers (index or coordinate) of whole plots starting from 1.
Cp: plot control code. 1 represents plot control cultivar and 0 represents test plots. Only 0 and 1 are allowed in this column.

Csp: subplot control code. 1 represents the first subplot control cultivar, 2 represents the second subplot control cultivar and 0 represents the remaining test genotypes. Only 0, 1 and 2 are allowed in this column.

Year: years of experiments.

Location: locations of experiments.

Genotype: genotype names of all plot control, subplot controls and test genotypes.

Trait: names of traits.

Value: observed values. The dot (".") presents any missing value.

Data from different experiments and traits are integrated into one file. Ordering data is not required as the pipeline package will process all data simultaneously.

An alternative data format (Table 2) is also acceptable but data conversion using the Perl program "*transpose_MAD_data.pl*" is required.

Table 2: Raw data format 2 for the MAD analysis pipeline

Record	Plot	Row	Columm	Cp	Csp	Entry	Year	Location	Genotype	OIL	IOD	LIO	LIN	YIELD
75	4575	2	6	0	0	1	2009	MD	CN18973	45.51	191.2	14.93	56.61	.
169	4669	4	7	0	0	2	2009	MD	CN18979	47.35	191.27	14.83	56.82	.
261	4761	6	8	0	0	3	2009	MD	CN18980	44.09	193.32	13.95	58.35	.
269	4769	6	7	0	0	4	2009	MD	CN18981	.	191.18	12.77	57.4	.
229	4729	5	6	0	0	5	2009	MD	CN18982	40.54	185.8	19.52	51.56	.
26	4526	1	6	0	0	6	2009	MD	CN18983	39.36	184.22	15.26	53.54	.
292	4792	6	2	0	0	7	2009	MD	CN18986	40.73	186.72	18.48	52.41	.
95	4595	2	2	0	0	8	2009	MD	CN18987	38.68	187.12	18.1	52.84	.
346	4846	7	10	0	0	9	2009	MD	CN18988	39.51	193.04	17.56	56.28	.
289	4789	6	3	0	0	10	2009	MD	CN18989	43.55	193.75	12.69	59.96	.
146	4646	3	10	0	0	11	2009	MD	CN18991	41.39	177.14	18.76	47.62	.
412	4912	9	3	0	0	12	2009	MD	CN18993	43.29	174.54	12.42	49.97	.
37	4537	1	8	0	0	13	2009	MD	CN18994	44.34	186.03	15.57	53.51	.
257	4757	6	9	0	0	14	2009	MD	CN18997	40.81	180.49	16.52	49.92	.
459	4959	10	9	0	0	15	2009	MD	CN18998	40.05	184.01	18.44	51.89	.
406	4906	9	2	0	0	16	2009	MD	CN19001	41.47	183.83	15.29	53.1	.
466	4966	10	7	0	0	17	2009	MD	CN19003	46.56	187.82	17.02	54.3	.
375	4875	8	6	0	0	18	2009	MD	CN19004	45.59	194.64	15.18	58.23	.
39	4539	1	8	0	0	19	2009	MD	CN19005	46.47	193.39	17.1	56.73	.
370	4870	8	7	0	0	20	2009	MD	CN19007	38.77	177.41	13.07	50.96	.

3. Usage of MAD pipeline package

The flowchart of MAD data analysis is shown in Figure 1.

Step 1: Data conversion if necessary

If you have a data file in a format shown in Table 2, it needs to be transformed as shown below.

Usage:
```
perl transpose_MAD_data.pl -i MAD_design_phynotypic_data
```

For example:
```
perl transpose_MAD_data.pl -i test_MDA_data.txt
```

A converted file test_MDA_data.txt_converted.txt will be generated.

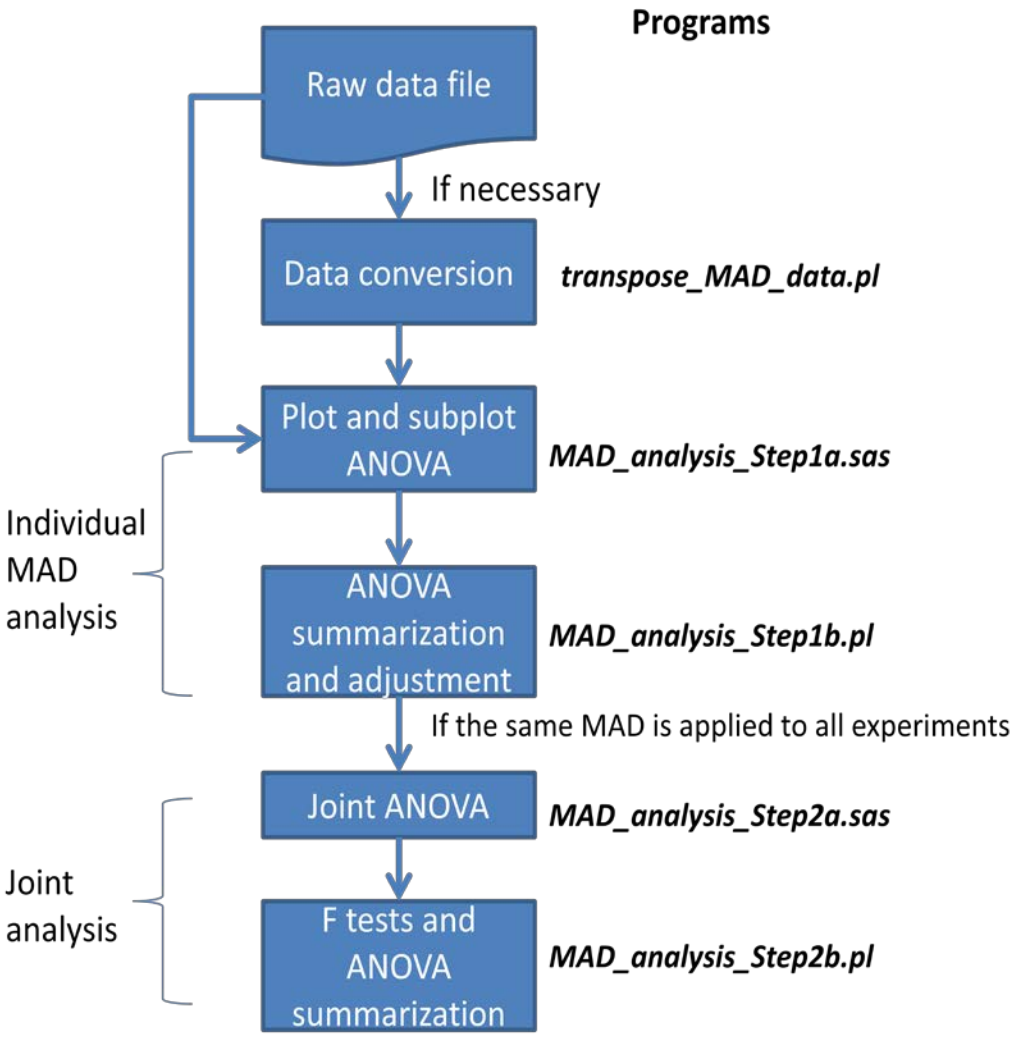

Figure 1. Flowchart of MAD data analysis using the pipeline programs

Step 2: ANOVA of plot and subplot controls
Before you run the SAS program "*MAD_analysis_Step1a.sas*", you need to change the input data file name to your MAD data file name in the SAS program. For example,

```
data MAD.mad_phenotypic_data;
    infile 'test_MDA_data.txt_converted.txt' dlm='09'x LRECL=1000;
    length Genotype $30  Whole_plot $10;
    input Record Plot Row Column Cp Csp Entry Year $ Location $ Genotype $ Trait $ Value;
    Whole_plot = '.';
    if Csp > 0 or Cp = 1 then Whole_plot = cat (Row, '+', Column);
    if _n_ > 1;
run;
```

The part highlighted in red color in the above SAS module must have correct file path and file name.

Ths program will produce the following two files for the downstream analysis using *MAD_analysis_Step1b.pl* in Step 3.

(1) unadjusted_subplot_anova_stats.txt
(2) unadjusted_plot_anova_stats.txt

Step 3:Summarization of ANOVA results and adjustment of test genotypes

The Perl program *MAD_analysis_Step1b.pl* summarizes the ANOVA results from the SAS program *MAD_analysis_Step1a.sas*, calculates necessary statistics for adjustment of observations of test genotypes and controls, and estimates the RE of different adjustment methods for selection of the most appropriate adjustment method.

Usage:

```
perl MAD_analysis_Step1b.pl -i converted_MAD_design_phynotypic_data
```

For example:

```
perl MAD_analysis_Step1b.pl -i converted_test_MDA_data.txt
```

This program reads two files, "unadjusted_subplot_anova_stats.txt" and "unadjusted_plot_anova_stats.txt", generated in Step 2. The names of these two files cannot be altered. Place these two files in the same folder with the raw or converted MAD phenotypic data file.

This program will export all necessary results for ANOVA and adjustment of test genotypes:

(1) MAD_ANOVA_result_summary.txt: Summary of complete MAD ANOVA for each trait in individual experiments.
(2) MAD_control_total_means.txt: Total means of plot controls for all traits and experiments.
(3) MAD_adjusted_all_data_suggested.txt: Values of test genotypes plus controls adjusted by the most appropriate adjustment method. This file has the same data format as the raw phenotypic data used in Step 2, and can be also used in the next step for joint ANOVA if the same design with the same plot and subplot controls is used in all experiments.
(4) MAD_adjusted_all_data_Method1.txt: Adjusted values by Method 1.
(5) MAD_adjusted_all_data_Method3.txt: Adjusted values by Method 3.
(6) MAD_adjusted_all_data_Method13.txt: Adjusted values by Method 1+3.
(7) MAD_adjusted_data_by_three_methods.txt: Adjusted values by all three methods, Method 1, Method 3 and Method1+3.

(8) MAD_adjusted_data_for_genotypes_at_multi_locations.txt: Adjusted values by the most appropriate method are arranged by experiments in columns. This file is usually what users want to obtain in MAD data analysis.
(9) MAD_adjusted_subplot_controls_data.txt: Adjusted values for subplot controls (using the most appropriate adjustment method).
(10)MAD_adjusted_RE.txt: Relative efficiency (RE) values for different adjustment methods.

Step 4: Joint ANOVA only if the same design is applied to all experiments

The SAS program "*MAD_analysis_Step2a.sas*" is used to perform joint ANOVA only if the same design with the same plot and subplot controls is used in all experiments.

The output file from Step 3, "MAD_adjusted_all_data_suggested.txt", must be located in the working directory of SAS. No program customization is necessary for this program. The program will automatically read this file and output three ANOVA result files if data from multiple years and locations are available:

(1) multi_envirom_anova_stat_by_loc.txt
(2) multi_envirom_anova_stat_by_year.txt
(3) multi_envirom_anova_stat_YL.txt

Step 5: Summarization of joint ANOVA

The Perl program *MAD_analysis_Step2b.pl* is to summarize the ANOVA results from the SAS program *MAD_analysis_Step2a.sas*. It is worth noting that in the PROC GLM, all effects are considered fixed even when the "RAMDOM" statement is used. The PROC GLM is not able to choose suitable MS terms for the F tests in the mixed model. Thus, this program is to calculate the correct F values and perform the significance test.

The program has the following assumptions for effect model of ANOVA:

(1) For joint analysis of multiple years and locations, two effect models are used (see Table 9 in the paper of You et al. 2013):
    a. Fixed model: all effects of Year, Location and Genotype are fixed.
    b. Mixed model: the Location and Genotype effects are fixed and the Year effect is random.
(2) For joint analysis of multiple locations in one year, the Genotype and Location effects are fixed (see Table 8 in the paper of You et al. 2013).
(3) For joint analysis of multiple years in one year, the Genotype and Year effects are fixed (see Table 8 in the paper of You et al. 2013), but 'Year' is replaced by 'Location'.

The program will automatically read the following three files from the working directory where the Perl program is running. Thus, these files must be placed to the working directory:

(1) multi_envirom_anova_stat_by_loc.txt
(2) multi_envirom_anova_stat_by_year.txt
(3) multi_envirom_anova_stat_YL.txt

Usage:

```
perl MAD_analysis_Step2b.pl
```

This program will generate one ANOVA summary file "multi_envirom_anova_summary.txt".

4. **References**

- Lin CS, Poushinsky G (1983) A modified augmented design for an early stage of plant selection involving a large number of test lines without replication. Biometrics 39(3):553-561
- Lin CS, Poushinsky G (1985) A modified augmented design (type 2) for rectangular plots. Canadian J Plant Sci 65(3):743-749
- Lin CS, Poushinsky G, Jui PY (1983) Simulation study of three adjustment methods for the modified augmented design and comparison with the balanced lattice square design Soil variation, statistical models. J Agri Sci 100(3):527-534
- Lin CS, Voldeng HD (1989) Efficiency of Type 2 modified augmented designs in soybean variety trials. Agronomy J 81(3):512-517
- You FM, Duguid SD, Thambugala D, Cloutier S (2013) Statistical analysis and field evaluation of the type 2 modified augmented design (MAD) in phenotyping of flax (*Linum usitatissimum*) germplasms in multiple environments. Australia J Crop Sci.

5. Citation

If you use this pipeline package for your study, please cite the following paper:
You FM, Duguid SD, Thambugala D, Cloutier S (2013) Statistical analysis and field evaluation of the type 2 modified augmented design in phenotyping of flax germplasms in multiple environments. Australia J Crop Sci.

6. Contact information

For the latest version of the pipeline package, test data sets or any questions, please contact:

Dr. Frank M. You
Agricultural and Agri-Food Canada, Winnipeg, R3T 2M9, Canada
Frank.you@agr.gc.ca