**MEGAN’S GUIDE TO USING THE MAD PIPELINE PACKAGE**

**(for single year analyses)**

**SUMMARY OF STEPS:**

1. Download and install software needed for MAD pipeline.

2. Check the data and configure the data file.

3. Transform the data.

4. Run SAS program 1a.

5. Run Perl program 1b.

**NOTE FOR HELEN: Connect to Cedar and go to the folder “MAD\_PIPELINE\_WITH\_MEGAN” and enter the following:**

export PERL5LIB=$PERL5LIB:"/home/hmb959/project/hmb959/MAD\_PIPELINE\_WITH\_MEGAN/Statistics"

To change directories, use cd. To see what’s listed in a directory use ls.

**Then skip to step 2.**

**STEP 1: DOWNLOAD AND INSTALL SOFTWARE NEEDED FOR MAD PIPELINE.**

*MAD Pipeline Package:*

* Can get a copy of this from Megan (who got it from Gaofeng). It is also now stored on Jade.
* Unpackage the compressed file to get each of the individual files.
* Need to put a copy of the perl programs onto westgrid, and need to keep the SAS programs somewhere that SAS can access them.
* Need to also install two perl modules. This part can be complicated, but these are the steps that I used:
  + First, create a directory where you want the packages to be stored. For me, it is a directory called BREEDING\_STATISTICS.
  + Download the modules for Statistics::Distributions and Statistics::Regression from <https://metacpan.org/pod/Statistics::Distributions> and <https://metacpan.org/pod/Statistics::Regression>. Can download the software in westgrid using the wget command and the link to the software.

$ wget https://www.cpan.org/modules/by-module/Statistics/Statistics-Regression-0.53.tar.gz

$wget http://www.ring.gr.jp/archives/lang/perl/CPAN/authors/id/M/MI/MIKEK/Statistics-Distributions-1.02.tar.gz

*Decompress and unpackage*

$tar xvzf Statistics-Distributions-1.02.tar.gz

$tar xvzf Statistics-Regression-0.53.tar.gz

*Prepare the software.*

* Go into each of the folders where the Makefile.pl program is and enter the following:

$perl Makefile.PL PREFIX=/home/housem/project/housem/BREEDING\_STATISTICS/Statistics-Distributions-1.02

$perl Makefile.PL PREFIX=/home/housem/project/housem/BREEDING\_STATISTICS/Statistics-Regression-0.53

* Enter the following within each of the directories:

$make test && make install

* Not sure if it’s needed, but I changed the permissions for the executable files:

$chmod 555 Distributions.pm

$chmod 555 Regression.pm

NOTE: Read permissions = 4 and execute permissions = 1, so we enter 5 for user, group, and everyone.

* Now you need to add the path to the executable files to the list of paths that perl will search:

Move both executable files to one folder (I called mine “Statistics”) and make sure they are still executable (i.e. can use chmod again). Enter the following so that perl knows where to look for the modules:

Export PERL5LIB=$PERL5LIB:”/home/housem/project/housem/BREEDING\_STATISTICS/Statistics”

(but insert the path to where *your* files are)

\*\*For Helen, this will be:

export PERL5LIB=$PERL5LIB:"/home/hmb959/project/hmb959/MAD\_PIPELINE\_WITH\_MEGAN/Statistics"

NOTE: Installation of the modules should be done slightly differently now. Perl seems to only look in a single location, so move both executable files to one directory and point perl that way (using the export command above).

<https://docs.computecanada.ca/wiki/Perl>  
  
To install the module, I did the following:  
  
module load perl  
  
cpan install Statistics::Regression  
  
{If it is the first time you install perl, it will ask you some questions as you can see from the documentation above}  
  
As for the second:  
  
1. Download and unpack the code:  
  
wget <http://www.ring.gr.jp/archives/lang/perl/CPAN/authors/id/M/MI/MIKEK/Statistics-Distributions-1.02.tar.gz>  
tar -xvf Statistics-Distributions-1.02.tar.gz  
  
2. Compile and install the package:  
  
cd Statistics-Distributions-1.02  
perl Makefile.PL  
make  
make test  
make install

\*\*can check my .bashrc file to see what lines of code were added. This might help anyone who needs to install these modules in the future.

**STEP 2: CHECK THE DATA AND CONFIGURE THE DATA FILE**

* Generally the data received is not formatted for the MAD pipeline. There are usually unnecessary columns that can complicate things, headers might not be formatted to be identified by the pipeline, and there can be outliers or information entered into cells that the pipeline won’t understand.
* First, make sure the necessary columns are present and the unnecessary columns are removed.
  + Columns required are:
    1. Record (there needs to be a record column for the SAS program, but the information within it isn’t used –so anything can be put in here)
    2. **Plot** (this is the plot # from the field test)
    3. **Row** (the row numbers of the whole plots, starting from 1)
    4. **Column** (the column numbers of whole plots starting from 1)
    5. **Cp** (plot control code. 1 represents the plot control cultivar and 0 represents test plots. Only 0 and 1 can be in this column).
    6. **Csp** (subplot control code. 1 represents the first subplot control, 2 represents the second subplot control cultivar and 0 represents the remaining test genotypes. Only 0, 1, and 2 can be in this column).
    7. Entry (Needs to be a column for this but the data in it doesn’t technically matter for the software to run. The entry is the unique value associated with each genotype within a test though).
    8. **Year** (year the test was run)
    9. **Location** (location where the test was run; this can be the full location name or the location code)
    10. **Genotype** (the name of the genotype used)
    11. **Trait** (name of the traits – NOTE: In the original file you can have columns with the actual trait name, i.e. YIELD with the values entered in the cells and when the file is transformed with the perl program it was adjust the cells so there is a column called TRAIT and another called VALUE)
    12. **Value** (the actual data value corresponding to the plot for each trait)

**NOTE: bolded column names are those where the data is actually used.**

* Make sure the headers are formatted to start with a capital letter (i.e. Record instead of record)
* Check for outliers
  + Can use max and min usually to identify any values that should be removed.
* Empty cells need to have “.” for SAS to understand that they are empty cells.

**STEP 3: TRANSFORM THE DATA**

Generally, we start with the data formatted in the following way, with headers for the different traits and their values following in the cells below:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Record** | **Plot** | **Row** | **Column** | **Cp** | **Csp** | **Entry** | **Year** | **Location** | **Genotype** | **Yield** | **Height** |
| 1 | 101 | 1 | 1 | 0 | 0 | 1 | 2019 | Preston | Bethune | 24.5 | 123 |
| 1 | 102 | 1 | 1 | 0 | 0 | 1 | 2019 | Preston | Bethune | 26.4 | 118 |
| 2 | 103 | 1 | 1 | 0 | 0 | 2 | 2019 | Preston | Sorrel | 22.7 | 120 |
| 2 | 104 | 1 | 1 | 0 | 0 | 2 | 2019 | Preston | Sorrel | 24.1 | 122 |
| 3 | 105 | 1 | 1 | 0 | 0 | 3 | 2019 | Preston | Royal | 18.2 | 90 |
| 3 | 106 | 1 | 1 | 0 | 0 | 3 | 2019 | Preston | Royal | 19.8 | 92 |
| 4 | 107 | 1 | 1 | 0 | 0 | 4 | 2019 | Preston | RE1 | 16.2 | 75 |

The way it needs to be set up is as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Record** | **Plot** | **Row** | **Column** | **Cp** | **Csp** | **Entry** | **Year** | **Location** | **Genotype** | **TRAIT** | **VALUE** |
| 1 | 101 | 1 | 1 | 0 | 0 | 1 | 2019 | Preston | Bethune | Yield | 24.5 |
| 1 | 101 | 1 | 1 | 0 | 0 | 1 | 2019 | Preston | Bethune | Height | 123 |
| 1 | 102 | 1 | 1 | 0 | 0 | 1 | 2019 | Preston | Bethune | Yield | 26.4 |
| 1 | 102 | 1 | 1 | 0 | 0 | 1 | 2019 | Preston | Bethune | Height | 118 |
| 2 | 103 | 1 | 1 | 0 | 0 | 2 | 2019 | Preston | Sorrel | Yield | 22.7 |
| 2 | 103 | 1 | 1 | 0 | 0 | 2 | 2019 | Preston | Sorrel | Height | 120 |
| 2 | 104 | 1 | 1 | 0 | 0 | 2 | 2019 | Preston | Sorrel | Yield | 24.1 |

To transform the data, make sure that the original .txt file has been copied into Cedar (can do this using WinSCP, which can be downloaded using the Software Centre) and that the perl program is in the same location (just makes it simpler). Then type:

$perl transpose\_MAD\_data.pl –i *name\_of\_your\_file.txt*

If it worked, you will see a message saying that it worked and you will now have a file with the added extension\_converted.txt.

**STEP 4: RUN SAS PROGRAM 1a**

This step will run ANOVA on the plot and subplot controls. The only thing that should be adjusted is the line in the SAS code that indicates the location and name of the data file. **Update this with the correct path and file name** and then just press run. Check the destination of the output files in the code as well. Usually need to adjust this so that the output is saved to the correct location (look at the bottom of the code for the 2 purple lines with file paths).

Should see that several files are output. There should be two called unadjusted\_subplot\_anova\_stats.txt and unadjusted\_plot\_anova\_stats.txt. Move these onto cedar so that the next perl program can be run using these files as input.

**STEP 5: RUN PERL PROGRAM 1b**

This step will summarize the ANOVA results from the SAS program and will calculate the necessary statistics for adjustment of the observations of test genotypes and controls, and it will estimate the RE of different adjustment methods for selection of the most appropriate adjustment method. Following this step you should have a file containing the adjusted data (if it was determined that adjustment is the best method).

* Move the 2 necessary output files from SAS onto Cedar.
* Make sure that the transformed raw data is also located with the two SAS output files.
* Run the perl program for step 1b:

$perl MADPipeline\_Step1b.pl –i name\_of\_your\_file.txt\_converted.txt

That should be the end of it! There will be a number of different files but one of them, the one called MAD\_adjusted\_data\_by\_three\_methods will indicate what the best method was for each data point. The file MAD\_adjusted\_all\_data\_suggested will contain the actual adjusted data. This is the data that should be used moving forwards for further analysis.