Users guide for Multi Environment Trial analysis with CloneSelector

Raul Eyzaguirre and Jens Riis-Jacobsen July 8th 2011

Table of Contents

Introduction

Genotype by environment interaction, GxE, is one of the challenges in plant breeding and other experiments with plants. Testing across multiple environments is therefore a standard procedure in plant breeding as well as other experiments. The outcome of these experiments is data sets from multiple environments, and these need to be analyzed using statistical methods that allows the scientist to draw conclusions for example related to stability of a genotype across target environments or in relation to the genotypic performance with less interference of GxE. These different types of analysis have in CloneSelector been automated as MET analysis or Multi environment trial analysis.

It should be noted that the MET analysis generates many analytical outputs; however, **the researcher must select the analytical outputs that are relevant to the particular trial or experiment, and ensure all assumptions are fulfilled**. The MET analysis will for example produce different ANOVA tables, but you must choose the one that have the right combination of fixed and random effects. The MET analysis also produces for example plots of the data or residuals, and other types of analytical tools to assess if the assumptions for a particular analysis are fulfilled. **CloneSelector aims to automate the task of carrying out the statistical analysis of a multi-environment trial, however, the researcher must interpret the outputs and ensure they are valid in relation to the particular experiment**.

Before you can do the MET analysis you must have a series of experiments in CloneSelector where you have tested at least **3 clones** in the same three **3 environments**. Some of the tests can also be done for only 2 clones in 2 environments, but most of the analysis requires at least 3 clones that have been tested in at least 3 environments, if you have more clones and more environments that obviously also work. If your data is not in CloneSelector you may still be able to use the MET analysis, but it will require a bit of manual copy/paste – and we cannot guarantee that you will not have problems when trying to analyze non-CloneSelector data.

The MET analysis tools in CloneSelector helps you:

1. Copy in an automated way the individual trial results from the Master sheets into one consolidated file that has the data from all the trials properly aligned for analysis.
2. You can then create for each trait individual sheets before analyzing the data. The purpose of these sheets is to be able to easily resolve issues with data quality, and especially missing data problems. The analysis requires at least one observation from each genotype and environment combination, and no more than 10% missing data. If this is not fulfilled you will have to delete genotypes or environments before running the analysis.
3. Once the data is prepared you can run the MET analysis and it will produce some 15 different types of statistical analysis and 5 plots for each trait.
4. You can also generate a MET summary result sheet that presents the mean values of each trait for each genotype, and also the minimum and maximum values for each trait.
5. Finally, you can generate a selection index for the traits you choose to include. The index is based on Elstons method using k-values.

How to prepare for a MET Analysis

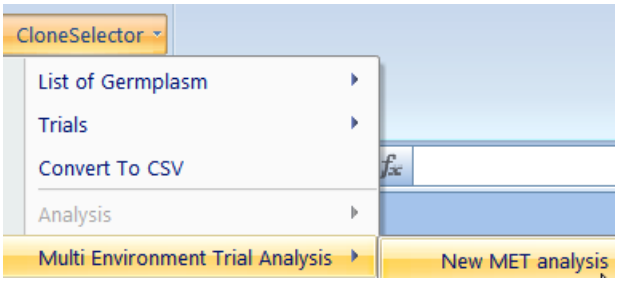
In this part you will join the data from the different experiments into one single workbook, which facilitates the subsequent analysis. **All the experiments to be entered in the analysis must be CloneSelector Trials, and you must have done the analysis of each individual experiment before you do the multi environment trial analysis**.

If your data is not in yet entered as a CloneSelector Trial you must first do so. For an explanation on how to do it look in CloneSelector users guide under **Other Functions** and **Copy Existing Data Set into CloneSelector Fieldbook**.

Open CloneSelector 2-0

Open CloneSelector2-0. Make sure you open the right version, else the MET menus does not appear. See CloneSelector Users Guide for additional information on how to open CloneSelector or similar basic information.

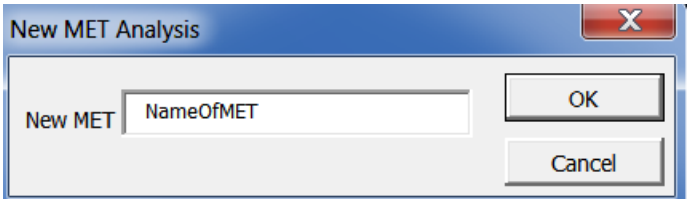
Make a new MET analysis file



Click:

CloneSelector -> Multi Environment Trial Analysis -> New MET analysis

In the New MET Analysis dialog box **enter the name of the new analysis** you want to do. Typically the same name as the fieldbooks you made for the trials or something similar.



Click **OK**

A new workbook opens with one sheet called MasterMETA and a few column names:

A	B	C	D
Environme	Genotype	Replication	First trait
env	geno	rep	AbrTrait1

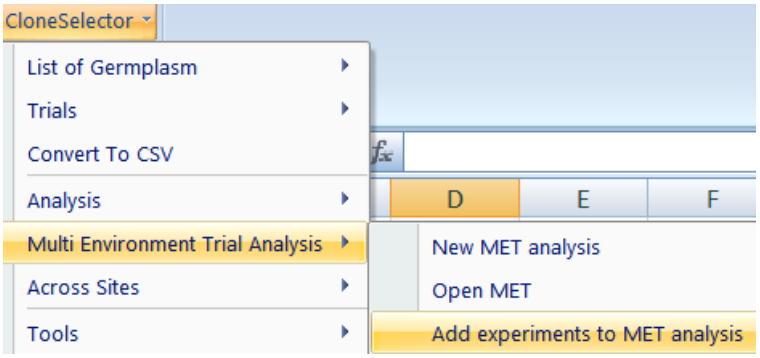
In this file you will in the next step add the Fieldbooks (or more exactly, the Master sheets) of the fieldbooks you want to analyze.

Add experiments to MET analysis from a CloneSelector trial series

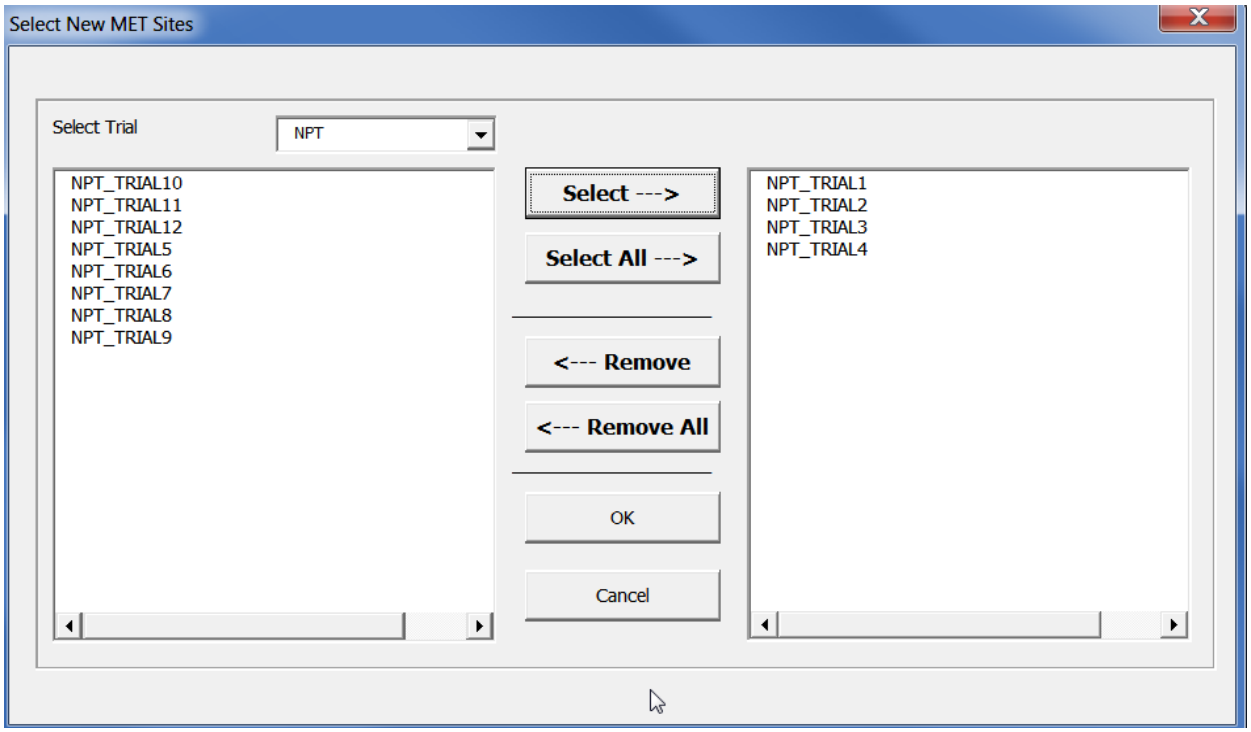
If your trial series is in CloneSelector fieldbooks you should use this procedure. If they are in another format you should do as is described in the next section about data not collected in CloneSelector.

Click on

CloneSelector -> Multi Environment Trial Analysis -> Add experiments to MET analysis



In the Select MET sites dialog



Select first the Season from the drop down list, and then **select trial sites** to include in analysis.

It is important to notice that the MET is to **compare the same list of germplasm across locations**, so the germplasm tested in the experiments should be the same or almost the same, else you will not be able to do the analysis. **You need data from at least 3 environments and 3 genotypes across the environments**.

Click **OK**

CloneSelector will now open each trial and copy the MasterSheet into the MasterMETA.

Wait for this to finish

If you have trials from different seasons, they may be in different folders and you will have to use the Add experiment for each folder separately, as you can only add data from one folder at the time.

Save the workbook

Add experiments to MET analysis from a trial series not made with CloneSelector

The multi environment analysis can be run on any data set that is comprised of:

- Three or more environments
- Three or more germplasm that has been tested in the same environments
- Traits scored on a numerical scale
- Trials with a randomized complete block design

The MET analysis of the individual traits will work without problems. However, in the summary sheet you may have to do a little additional work of putting in for example selection direction for the selection indexes if the traits names are not exactly as in CloneSelector.

You probably already have experience with copy/pasting a trial series into one sheet as a preparation for running a statistical analysis, and this is also what you need to do here.

The first step is to create a New MET analysis file as described above. Note: it is important that you create a new one, and avoid just changing an existing file. CloneSelector works based on some hidden information and the analysis may fail if you simply modify an old MET file.

A	B	C	D
Environme	Genotype	Replication	First trait
env	geno	rep	AbrTrait1

When you **copy/paste the data together** it is recommended that you work on a backup version of you original data, so that you do not accidentally lose any data. You should copy/paste the data into one single sheet similar to the MasterMETA sheet:

1. First column called env, and have a name or code for each environment.
2. Second column called geno, with the names of all the genotypes
3. Third column called rep, with the number of the replication
4. All subsequent columns should be with numerical data from the trials

Things that may cause problems

1. Environment and genotype names must be spelled consistently as any inconsistent spelling will mean that CloneSelector will consider them as different environments or genotypes
2. All missing values must be left as empty cells in Excel

When you have copy/pasted the data into one single Excel sheet you can **copy them into the MasterMETA sheet**. It is recommended that you paste the data using Past Special and only **Paste Values** so that you do not accidentally copy over formulas or any special formatting that may cause problems when running the MET analysis.

When you have copied over the data, revise that they are ok, and then **Save the workbook.** You are now ready to run the MET analysis as described below.

How to create a MET summary sheet and selection index

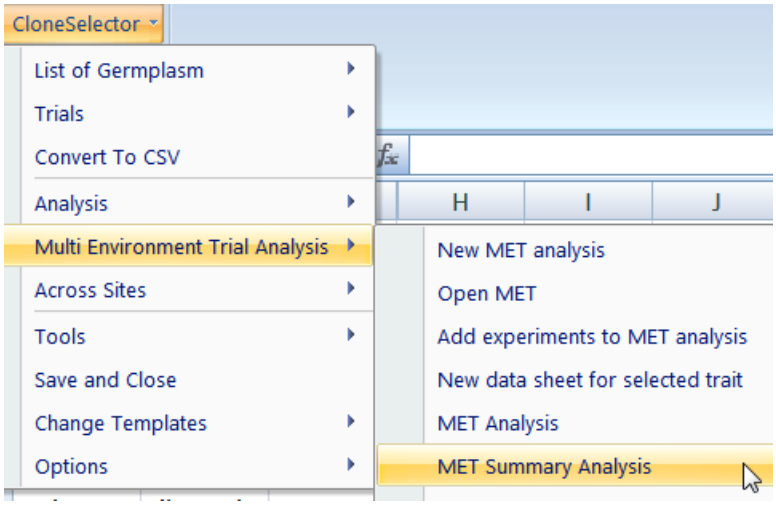
The MET summary sheet presents for each genotype the mean value of each trait across all environments, and it also includes the overall mean, and the minimum and maximum values for each trait. The summary table provides an overview of the trial results, and is also the basis for calculating the selection index.

Create MET summary analysis

To create the MET summary you must first create the MET analysis and add the trials as described above. And you must make the MasterMETA sheet the active sheet which simply means click on it. If the MET Summary Analysis menu is not available you are not on the right sheet.

Click on

CloneSelector -> Multi Environment Trial Analysis -> MET Summary Analysis.



Wait for the analysis to be carried out. It will start R and run an analysis.

Once the analysis is finished you will see a new sheet called Results, similar to the Results sheet from the individual experiments.

Entry	Name	SelectionIndex1	SelectionIndex2	RootDryMatterPer	HarvestSowingI	VirusSymptoms2	Ha
ENT	CODE	SI1	SI2	RDM	HIS	VIR2	
NumberFormat		0.00	0	0.00	0	0.0	
Lower Limit				0.00		1.0	
Upper Limit				100.00		9.0	
SelectionDirection				+	+	-	
SelectionWeight							
Entry	**Name**	Selection Index 1	Selection Index 2	Storage root Dry matter content	Harvest Soving Index	Virus symp2_1_S 1 M. bef harv	
				Units	Units	Units	
	Libertada			0.29	0.50	3.75	
	MUSG 0703-37			0.24	0.65	3.25	
	UW119 06-296			0.25	0.55	3.63	
	UW119 06-32			0.25	0.74	2.75	
	UXIPHONE 06-1			0.31	0.34	3.33	
	W119 06-39			0.29	0.67	2.92	
Overall_mean				0.27	0.58	3.27	
Min_mean				0.24	0.34	2.75	
Max_mean				0.31	0.74	3.75	

The MET summary is presented in the gray rows which have the titles of the traits, and below that you have first at table that present the means of each genotype for each trait. Further down is an additional table with the overall mean, and the smallest and largest mean value for each trait. In the top rows of the sheet you have information that is used for formatting, control of data quality, and for calculating the selection index, which is described in the next section.

Create Selection index.

Introduction to selection index in CloneSelector

A selection index is a way to combine multiple traits into a single value that can then be used to compare the different germplasm. The selection index value can often not be interpreted in itself, but is used to rank the germplasm, typically, the largest index value indicate the best germplasm. Defining a selection index contains a serious of problems such as:

1. The different traits are measured on different scales and may have different variance, which may influence the calculation of the index so that e.g. a trait with large measurement values or a huge variance may dominate the index value.
2. For some traits a large value is desirable (for example yield) and for others a small value is desirable (for example in many disease scores).
3. Not all traits may be of equal importance to the breeding objectives.
4. Some traits have a large covariance and are not independent of each other, for example yield is an element in multiple traits, and if for example several traits based on yield are included in an index, then this will favor yield over other uncorrelated traits.

5. Some traits may be of minor importance, and are only measured to avoid major problems with the material. In this case the breeder may for example accept all material that is average and above, but not be interested in including the trait as a weight beyond “good enough”.

Due to these challenges, a selection index ranking should only be considered an input in the selection process, but the breeder will in many cases want to refine the ranking based on a broader analysis.

Typically the calculation of a selection index involves some degree of transformation of the observations, so that the problems related to the use of different scales and differences in variance are reduced. Examples include: Divide with LSD value, subtract minimum value or k-value from observed values, and use this transformed valued in the index calculation.

In some selection indexes it is possible to include selection weights so that not all traits are given equal importance in the index. The possibility for including selection weights will depend on whether the selection value calculated for each trait is multiplied together to get the compound selection value of the germplasm, or whether the values for each trait are added up. If the values are multiplied together, a selection weight will not change the ranking, however, if the values are added up it is simple to include selection weights for individual traits.

We did a small comparative study using three different selection indexes which is included in Annex A of the present document. The conclusion of this small study was that the three indexes that were being tested all gave similar rankings, and even the inclusion of selection weights only moderately changed the ranking. Ideally a more comprehensive literature study and testing would have been undertaken, however, time has not permitted that. Instead we have included in CloneSelector a first option which is Elstons index using k-values. Compared to the index using minimum values it has the advantage of ranking all germplasm, were as the minimum value index will leave all germplasm with the smallest value in any trait, also as the last in the overall index, i.e. they are tied for last place with a 0 index value. The LSD index is not included as we are currently not calculating LSD values in the summary table.

In conclusion: This version of CloneSelector includes a selection index based on Elton’s k-value method (Elston 1963). The index is calculated as:

Selection index = ((X1-K1)*selection direction1 * selection weight1) *

((X2-K2)* selection direction2*selection weight2) and so on for each trait.

K= (n*Ximinimum-Ximaximum)/(n-1) for positive selection direction and K= (n*Ximaximum - Ximinimum)/(n-1) for negative selection direction, and n = number of germplasm.

It is important to stress that Selection Weight does not change the ranking, but it can be used to scale the selection index values if desired, however, **in CloneSelector the most important function of selection weight is that if it is not larger than cero, the trait is ignored in the selection index**.

The two different ways of calculating k-values solve the problem of different selection directions. For a positive selection direction k is a little smaller than the minimum value, and for a negative selection

direction it is a little larger than the maximum value. Subtracting the k from the observed value and multiplying by selection direction (+1 or -1) creates in both cases a rescaled value that reflects the desired selection direction.

This is a first attempt to introduce selection indexes for sweetpotato and potato, and feed back is requested.

How to generate the selection index and ranking

Note: As explained above, CloneSelector only includes traits with a Selection Weight larger than cero. If all selection index values are -1 and the Rank 1, then you have not put in any selection weights!

To create the Selection index you must first create the Results sheet of the MET summary analysis as described in the previous section.

On the Result sheet in the MET analysis:

Enter 1 in the traits you want to include in the selection index.

SelectionDirection				+	+	-
SelectionWeight				1.00		1.0
Entry	Name	Selection Index 1	Selection Index 2	Storage root Dry matter content	Harvest Soving Index	Virus symp2_1_9 1 M. bef harv

In the example above both Storage Rood dry matter content and virus symptoms have Selection weight 1 and will be included in the index, but Harvest sowing index is left blank and will not be considered.

For the traits included in selection index revise that the selection direction is correct

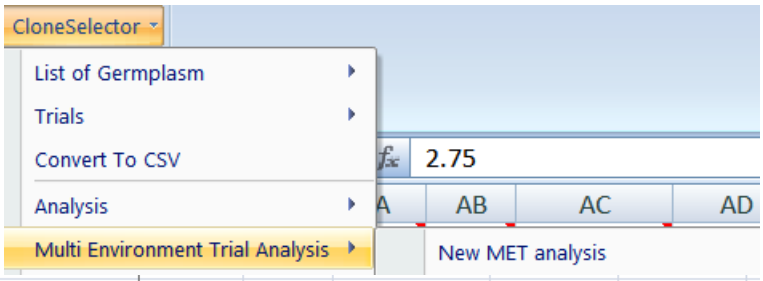
For **positive selection direction** (large value is desirable) enter **+** as Selection direction

For **negative selection direction** (small value is desirable) enter **–** as selection direction.

After entering selection weights and having revised the selection directions:

Click on

CloneSelector -> Multi Environment Trial Analysis -> Selection Index



On the Result sheet you will now see in column **AA the Selection index value** and in **column AB the**

SelectionDirection				+	+	-
SelectionWeight				1.00		1.0
Entry	Name	Selection Index 1	Selection Index 2	Storage root Dry matter content	Harvest Soving Index	Virus symp2_1_9 1 M. bef harv
				Units	Units	Units
	Libertada	0.01	6	0.28	0.61	4.17
	MUSG 0703-37	0.01	5	0.23	0.60	3.58
	UW119 06-296	0.02	4	0.26	0.56	3.92
	UW119 06-32	0.05	3	0.26	0.77	3.25
	UXIPHONE 06-1	0.07	2	0.31	0.52	3.64
	W119 06-39	0.07	1	0.29	0.69	3.44

ranking. In the example below you can observe how the last germplasm has the second best score on both traits in selection index, but achieves the highest overall score. The one ranked 2nd has the best score in one trait and is third in the other. Similarly the 3rd ranked has a 1st and 3rd ranking. As such the ranking produced is probably quite similar to what you would have expected.

If we for example include the Harvest sowing index in the selection index, we can see that UXIPHONE 06-1 has the lowest score and by including the harvest sowing index, it is also demoted from 2nd to 3rd in the overall index.

You can enter new Selection Weights to try different combinations of traits in the selection index, and recalculate it as many times as you want.

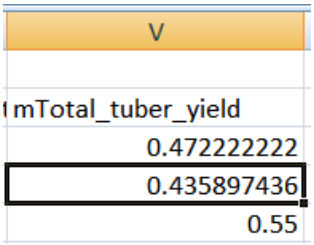
How to do MET analysis of a trait

To do this step you must first have consolidated the experiments in a MasterMETA sheet as described above.

It is a two step process where you first create a New data sheet for the selected trait, and then run the MET analysis. The reason for creating the additional data sheet is to be able to easily delete for example locations or germplasm with too many missing values. **The analysis requires all germplasm to have at least one observation in each experiment**. Also it **allows a maximum of 10% missing values**.

The analysis will generate a frequency table that will allow you to easily see if you have problems with missing values.

Create a New data sheet for selected trait

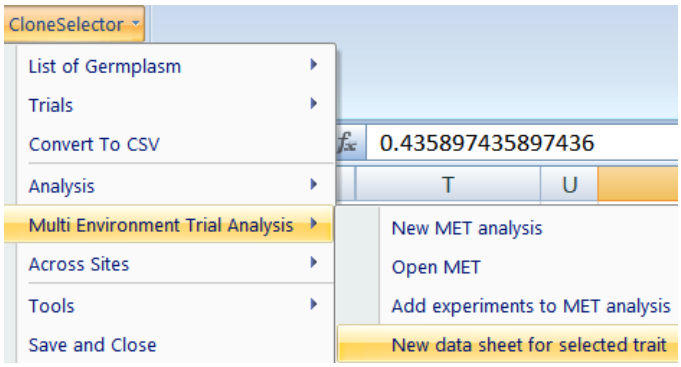


Select the trait by Clicking in a cell in the column on the MasterMETA where the trait is you want to analyze. It does not matter which row you click in, only the column matters.

Click on

CloneSelector -> Multi Environment Trial Analysis -> New data sheet for selected trait

Note this will copy over the column where the active cell was, so make sure you have clicked in the right column.



You will now have a new sheet called METD_NameOfTrait. This sheet has 4 columns

Environment, Genotype and Replications which are all factors of the experiment, and the column D is the trait. In Row one it has the abbreviation for the trait (missing in this example, but look on your own sheet!) and D2 must be called y, as this is used in the R script so don't change this.

If you have problems with missing values you can delete either locations or germplasm on this sheet. To delete a germplasm you should first sort the data according to the B column (taking care not to include first two rows), and then delete the whole block of rows where the germplasm is. Similarly, you can delete a location by first sorting by environment and then delete as needed.

	A	B	C	D
1	Environme	Genotype	Replication	
2	env	geno	rep	y
3	Kibirichia 2	x392657.8	1	0.472222
4	Kibirichia 2	Tigoni	1	0.435897
5	Kibirichia 2	x393385.3	1	0.55
6	Kibirichia 2	x385524.9	1	0.475
7	Kibirichia 2	x391391.9	1	0.325
8	Kibirichia 2	x392617.5	1	0.225
9	Kibirichia 2	x393371.5	1	0.675
10	Kibirichia 2	Dutch	1	0.45
11	Kibirichia 2	Tigoni	2	0.75
12	Kibirichia 2	x393371.5	2	0.525
13	Kibirichia 2	x393385.3	2	0.55
14	Kibirichia 2	x385524.9	2	0.575
15	Kibirichia 2	x392657.8	2	0.625
16	Kibirichia 2	Dutch	2	0.5

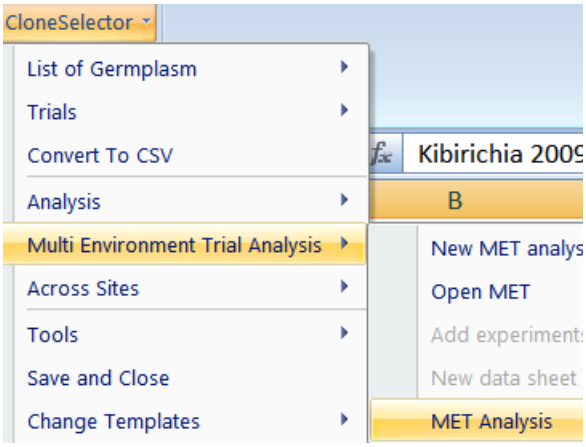
Run the MET analysis

To run the MET analysis you must be on a data sheet i.e. a sheet named METD_NameOfTrait.

Click on

CloneSelector -> Multi Environment Trial Analysis -> MET Analysis

This will open R and run a script to do the MET analysis. After a short while a new sheet will appear with the analysis:



MET Analysis

MET Information

	Trait:	Total_tuber_yield			
	Number of genotypes:	8			
	Number of environmen	4			

Number of replicates:

	Kibirichia	Kisima 20(	Limuru, Ti	Timau 2009-2010
Dutch	4	4	4	2
Tigoni	4	4	4	4
x385524.9	4	4	4	4
x391391.96	4	4	4	3
x392617.54	4	4	4	4
x392657.8	4	4	4	4
x393371.58	4	4	4	4
x393385.39	4	4	3	4

Analysis OK with estimated missing values.

The analysis includes a table called Number of replicates, and below this table is a message in red that tells you if the analysis was run. In the example above the analysis was run with estimated missing values.

If you have any ceros in the frequency table you must delete either germplasm or location. The MET analysis requires at least one observation for each germplasm/location combination.

Interpreting the MET analysis

Basic information about the trials

The output starts with some basic information about the trails. In this case, the trait analyzed was total tuber yield, and the information comes from trails conducted in 4 environments with 8 genotypes and 4 replications in each environment. It is quite common to have missing values and you can see this in the **Number of replicates** table. For these data, there were tree missing values at environment **Timau**, two of them with the Dutch and one with the x391391.96 genotype. It is important to know that you must define a missing value in CloneSelector by leaving the cell empty. To run the analysis **you need at least one value at each genotype by environment combination**, and in the case of missing values, CloneSelector will estimate them if no more than 10% are missing.

If there are zero frequencies in at least one of the genotype by environment combinations, or if the number of missing values is greater than 10%, you will only get this basic information and the analysis will not be done. You will have to delete some genotypes and/or environments to be able to carry out the analysis.

	A	B	C	D	E	F	G
1	***MET Analysis***						
2							
3	***MET Information***						
4							
5		Trait:	Total_tuber_yield				
6							
7		Number of genotypes:	8				
8							
9							
10		Number of environmer	4				
11							
12							
13	***Number of replicates:***						
14			Kibirichia	Kisima	Limuru	Timau	
15		Dutch	4	4	4	2	
16		Tigoni	4	4	4	4	
17		x385524.9	4	4	4	4	
18		x391391.96	4	4	4	3	
19		x392617.54	4	4	4	4	
20		x392657.8	4	4	4	4	
21		x393371.58	4	4	4	4	
22		x393385.39	4	4	4	4	
23							
24		***Analysis OK with estimated missing values.***					

Descriptive statistics

The first results are in a table with name Overall statistics. Here you have the minimum (Min.), first quartile (1st Qu.), the median (Median), the mean (Mean), the third quartile (3rd Qu.), the maximum (Max.), and the number of missing values (NA's). A quartile is a value that divides the ordered observations in quarters, so the first quartile for instance is the value that divides the data in one quarter below and tree quarters above. The median is the second quartile. Both mean and median, gives you an idea about the central tendency of the data; in the context of this example the average total yield per hectare.

The minimum and maximum are good indicators of extreme values and you may consider if they are too extreme to be true. However, the analysis of the individual trials include control of data quality and you should detect improbable values when analyzing each experiment and resolve the issue at that point.

	A	B	C	D
26	***Overall statistics***			
27		Min.	4	
28		1st Qu.	18	
29		Median	25	
30		Mean	28.52	
31		3rd Qu.	38	
32		Max.	63	
33		NA's	3	

Below this table there are tables with the means by genotype, by environment and for all the genotype by environment combinations. These means are computed with the observed data plus the predicted data if there are missing values. Hence, if you have a balanced data set, the mean of the means of any of the three tables below will be equal to the overall mean (28.52 in the example), but if you have missing values, they may be a little bit different.

	A	B	C	D	E	F	G
34	***Means by genotype***						
35		Dutch	20.99572				
36		Tigoni	32.0625				
37		x385524.9	34.75				
38		x391391.96	22.71747				
39		x392617.54	22.39375				
40		x392657.8	28.34375				
41		x393371.58	29.5625				
42		x393385.39	36.125				
43							
44	***Means by environment***						
45		Kibirichia	22.375				
46		Kisima	19.9375				
47		Limuru	40.8375				
48		Timau	30.32534				
49							
50	***Interaction means***						
51			Kibirichia	Kisima	Limuru	Timau	
52		Dutch	22.25	16.5	21	24.23288	
53		Tigoni	25.75	21	36.25	45.25	
54		x385524.9	24.75	23.75	38.5	52	
55		x391391.96	17.5	14.5	43.25	15.61986	
56		x392617.54	14.5	17.25	37.825	20	
57		x392657.8	24.75	22.25	48.375	18	
58		x393371.58	24.5	16.25	52.75	24.75	
59		x393385.39	25	28	48.75	42.75	

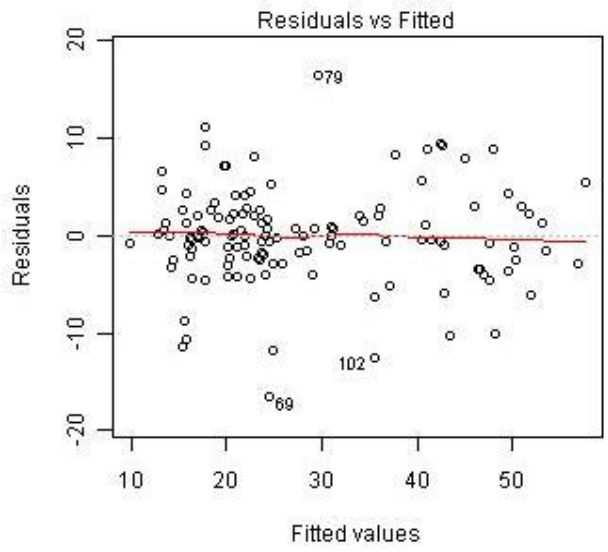
Finally, there is a table with the interaction effects. This matrix is computed by subtracting from each interaction mean its corresponding genotype mean and environment mean, and then adding the overall mean.

	A	B	C	D	E	F	G
61	***Interaction effects***						
62			Kibirichia	Kisima	Limuru	Timau	
63		Dutch	7.248116	3.935616	-12.4644	1.280652	
64		Tigoni	-0.31866	-2.63116	-8.28116	11.23099	
65		x385524.9	-4.00616	-2.56866	-8.71866	15.29349	
66		x391391.96	0.77637	0.21387	8.06387	-9.05411	
67		x392617.54	-1.89991	3.287586	2.962586	-4.35026	
68		x392657.8	2.400086	2.337586	7.562586	-12.3003	
69		x393371.58	0.931336	-4.88116	10.71884	-6.76901	
70		x393385.39	-5.13116	0.306336	0.156336	4.668493	

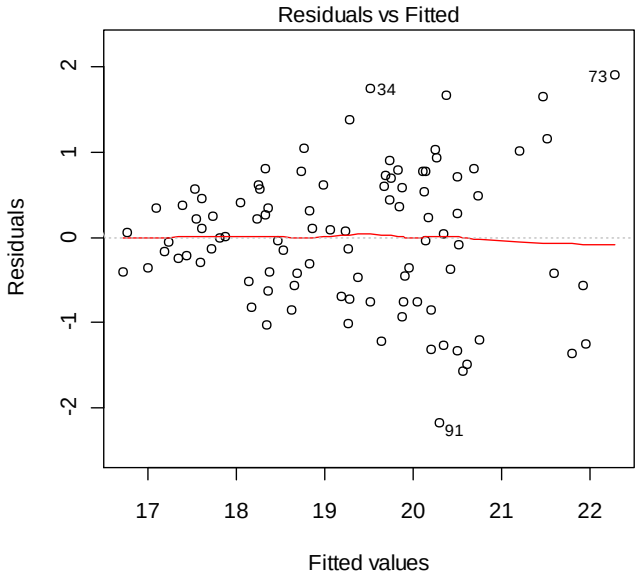
Model checking

Two important assumptions in the statistical analysis of the linear model proposed are that the data have a normal distribution, and that the variance of this normal distribution is the same for the different genotypes and environments. These assumptions can be checked by some plots and statistical tests. Here, two plots of residuals, and the tests of Bartlett and Shapiro-Wilk are shown.

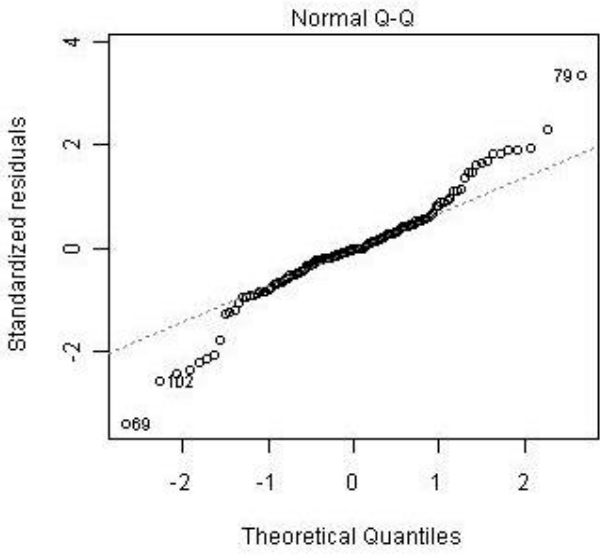
The plot of residuals versus fitted values shown below is a diagnostic plot for homogeneity of variances. It is suppose that the dispersion of the dots in the vertical axes is the same for the different fitted values. A funnel shape for instance would indicate that the variance increases (or decreases) with the fitted values implying that the variance is not independent of the mean response. The picture below does not show any suspicious pattern.



Below you can see an example with simulated data to illustrate the case when the variance is an increasing function of the mean.



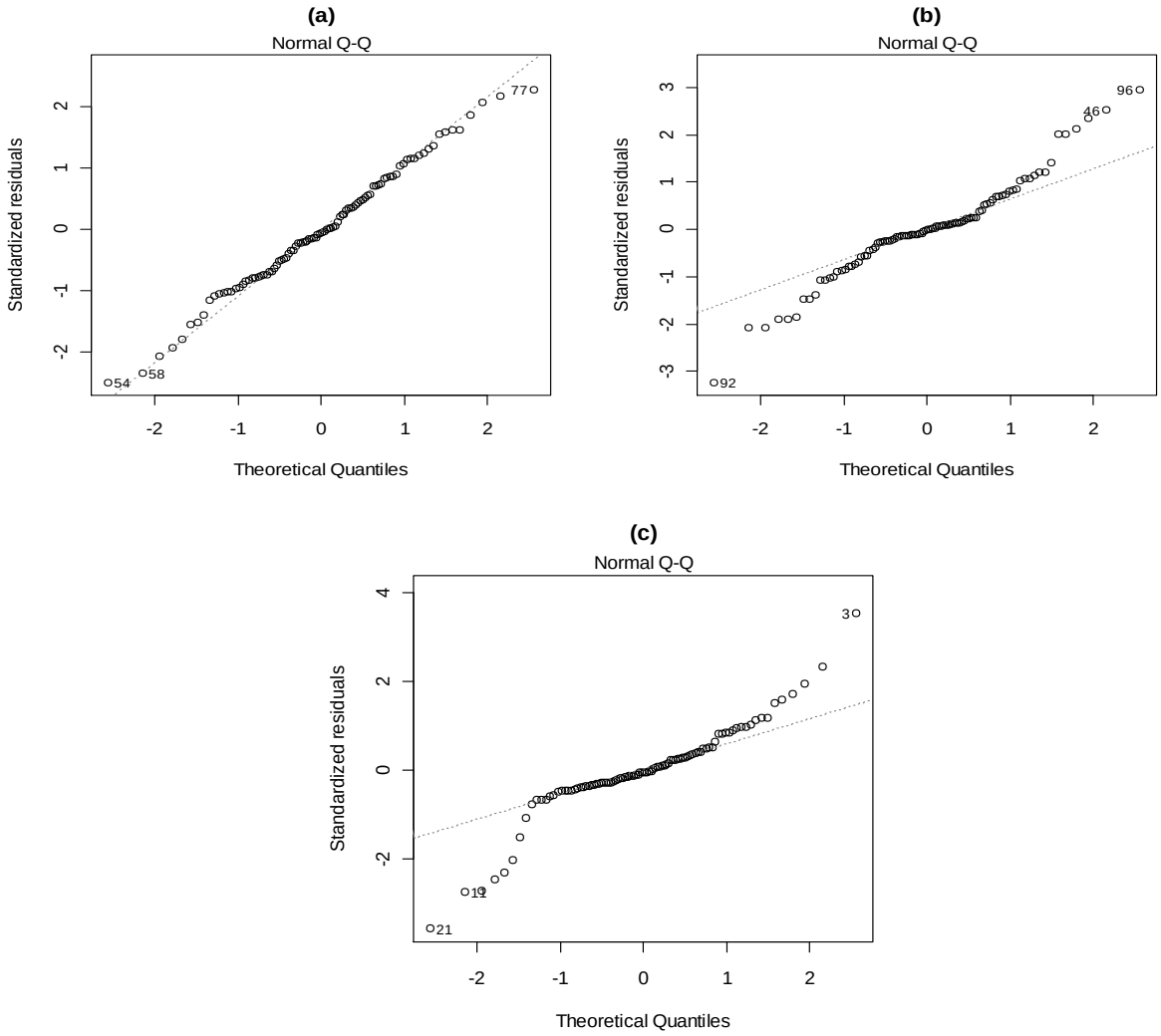
The second graph is a Normal Q-Q plot (quantile-quantile plot), and it is useful to check the normality assumption. It plots the standardized residuals versus the normal theoretical quantiles. The normal theoretical quantiles should follow the dashed line in the graph, so departure from this line is an indicator of lack of normality. Since sample information will never be perfectly normal, it is usual to see some departure, especially in the tales, so the dots tend to have an S shape. In order to properly judge if the departure is (or is not) big enough to suggest a normality problem, some experience is needed, and formal statistical tests are also important. The graph below shows some important deviation in the tales.



These two plots label the three most extreme residuals, so you can have a view of them in the context of the assumptions. In this example, the observations 69, 79 and 102 are the three with most extreme residuals. The data points 69 and 79 correspond to environment Timau with genotype X393371.58. The

yield observed at the four replications under these conditions were 8 (observation 69), 46 (observation 79), 27 and 18. Hence, observation 69 is too low and observation 79 is too high under the normality assumption and the estimated variability.

Below you can see some Normal Q-Q plots for different situations with simulated data. The first one (a) corresponds to a data that fits well to the normal distribution, the second one (b) correspond to a data with extreme values under the normality assumption (similar to the case of our data example), and the third one (c) was the result of replacing some of the values of case (a) with zeros.



In order to have a better view about the validity of the assumptions, two formal statistical tests are included, to complement the information given by the graphs.

The Bartlett test is a test for homogeneity of variances. In the example, there are 4 environments and 8 genotypes, so you can think in 32 different treatments. The null hypothesis in the Bartlett test is that the variance is the same in these 32 different treatments. The Bartlett test statistic (82.48029 in the example) is compared with a chi-square distribution whit $k-1$ degrees of freedom, where k is the

number of different treatments (32 – 1 in the example). In the example, the probability value is quite low (1.47E-06) indicating strong evidence against the null hypothesis.

The Shapiro-Wilk test is a test for normality where the null hypothesis is that the sample came from a normally distributed population. The results shown are the test statistic W (0.961275 in the example) and its correspondent probability value (0.001034 in the example). Again, the probability value in this example shows evidence against the null hypothesis.

	A	B	C	D	E
98	***Bartlett test***				
99		Bartlett test of homogeneity of variances			
100		standardized residuals by treatments			
101		Bartlett's K-squared	82.48029		
102		df	31		
103		p-value	1.47E-06		
104					
105	***Shapiro-Wilk test***				
106		Shapiro-Wilk normality test			
107		standardized residuals			
108		W =	0.961275		
109		p-value	0.001034		

A couple of words must be said in order to properly interpret and understood these results. Firstly, strictly speaking, there is no data that follow the assumptions perfectly, and secondly, the ANOVA procedure is more or less robust to departures from the assumptions, so it is not completely discredited if we observe some evidence against them. Hence, experience is important here to see the whole picture, by observing the graphs and the statistical tests, and perhaps doing some additional analysis. In order to help you interpret these results, we could define the following scenarios:

- If there are no problems with the assumptions, you can follow with the statistical analysis and you can be quite sure about the validity of the results.
- If there are minor problems with the assumptions, perhaps nothing really strange in the graphs, but some evidence in the statistical tests, you can follow with the statistical analysis and you can be more or less sure about the validity of the results.
- If there are major problems with the assumptions, maybe the graphs show here and some further analysis could help you identify a problem with the data. Maybe it was the case that experimental control was not good in some environment or that there was a brush or infection problem that affected some genotypes in an environment. Then the researcher (who knows what happened on the field) could decide for instance to exclude a complete environment, or a block in an environment, or a genotype. If the problem is more structural (perhaps an asymmetrical distribution with some vitamin or mineral concentration) some practitioners would try to use a transformation for the data.
- If there are major problems with the assumptions, but it is not possible to identify a cause and solve the problem, you must interpret the results of the analysis just as a reference, having in mind that they are not 100% valid.

Consider the following examples. If the Bartlett test indicates a serious problem with the homogeneity of variances assumption, further inspection could indicate that it is exclusively due to one environment with a high standard deviation, and hence you could try to figure it out what happened with that environment. If you observe a funnel shape in the residuals versus fitted values plot, it indicates a relation between the mean and the variance, and then a transformation for the data could be useful. A normality problem could be the result of some extreme values, the unnoticed presence of zeros in the data, or the presence of some asymmetry in the distribution of the data. In each case, with maybe some additional information, you should be able to make some sensible decision on how to interpret the results of the experiments.

Analysis of variance and variance components

The statistical model for this kind of experiments is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{k(j)} + \varepsilon_{ijk}$$

where y_{ijk} is the observed response with genotype i, environment j, and block k nested in environment j, α_i is the effect for genotype i, β_j is the effect for environment j, $(\alpha\beta)_{ij}$ is the interaction effect between genotype i and environment j, and $\gamma_{k(j)}$ is the effect of the block k which is into environment j. The assumption in this model is that the error terms, ε_{ijk}, are independent random variables with normal distribution with mean 0 and constant variance.

The factors in this model, namely genotypes, environments and blocks, can be assumed to correspond to fixed or random effects. We say that a factor has a fixed effect when the interest of the researcher is in the specific levels of the factor tested in the experiment. A fixed effect is an unknown constant (parameter) that we want to estimate. On the other hand, we say that a factor has a random effect when the interest of the researcher is in the whole population, hypothetical or real, of possible levels to test. In this case, the researcher selects a random sample of levels for the experiment. A random effect is a random variable, and then it is not something to estimate. Instead, we would like to estimate the parameters that describe the probability distribution of these random effects. Typically we assume that the random effects have a normal distribution with some unknown variance, and the interest is to estimate the variance for each random effect (this is the variance components estimation). Sometimes however, we would like to predict the random effects, and then we can do it by means of the Best Linear Unbiased Predictors, BLUPs for short.

Some of the factors in the model may correspond to fixed effects and some to random effects. If this is the case, the model is called mixed effects model.

In the model defined above, we assume that the blocks correspond to a random effect. The analysis of variance tables in the CloneSelector MET analysis correspond to three different situations: Genotypes and environments fixed, genotypes fixed and environments random, and genotypes and environments random.

The first table corresponds to the case with genotypes and environments fixed. This is the classical analysis of variance table where the total variability is decomposed into its different sources: Genotypes (G), environments (E), blocks or repetitions nested into the environments (R:E), the interaction between genotypes and environments (GxE), and the residual variability (Residuals). The variability due to interaction is decomposed further in two parts: which can be explained by a simple linear regression and which can not be. In this regression approach, for each genotype, a simple linear regression of its individual mean value on the mean of all genotypes for each environment is fitted and vice versa (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966; Shukla 1972).Therefore, this decomposition is done in two ways: 1) the interaction means are explained with a simple linear regression over the means of the genotypes and 2) the interaction means are explained with a simple linear regression over the means of the environments.

In the example you can see that the interaction has 21 degrees of freedom. In the first decomposition the interaction variability is split in what can be explained by a simple linear regression over the means of the environments (Het.Regr.G) and what can not be explained by this regression (Dev.Regr.G). Since it is done for each genotype, and we have 8 genotypes, Het.Regr.G has 8 – 1 = 7 degrees of freedom, and the remaining 14 go to Dev.Regr.G. In the second decomposition the interaction variability is split in what can be explained by a simple linear regression over the means of the genotypes (Het.Regr.E) and what can not be explained by this regression (Dev.Regr.E). Since it is done for each environment, and we have 4 environments, Het.Regr.E has 4 – 1 = 3 degrees of freedom, and the remaining 18 go to Dev.Regr.E. The probability values in both cases (0.719032 for Het.Regr.G and 0.188403 for Het.Regr.E) indicate that regression does not have a significant value to explain the interaction.

At the end of this ANOVA table you can see the coefficient of variation (CV = 21.19587% in the example). The CV is a measure of the variability do not explained by the model, or what can be attributable to uncontrolled sources of variation. This number is useful to compare with other experiments with the same crop and under similar conditions. If the CV is quite higher that usual, it can be consider as a warning that perhaps something went wrong with the experiment.

	A	B	C	D	E	F	G	H	I
111	***ANOVA: G fixed, E fixed, B random.***								
112			Df	Sum Sq	Mean Sq	F value	Pr(>F)	LSD5	
113		G	7	3807.175	543.8821	14.50525	4.32E-12	4.307541	
114		E	3	8521.888	2840.629	32.40779	4.91E-06	5.099675	
115		R:E	12	1051.832	87.65267	2.337683	0.012552	#N/A	
116		GxE	21	5609.795	267.1331	7.124397	3.8E-11	8.615082	
117		- Het.Regr.G	7	1354.953	193.5648	0.6369	0.719032	#N/A	
118		- Dev.Regr.G	14	4254.842	303.9173	8.105424	1.73E-10	#N/A	
119		- Het.Regr.E	3	1279.346	426.4485	1.772581	0.188403	#N/A	
120		- Dev.Regr.E	18	4330.449	240.5805	6.416244	1.75E-09	#N/A	
121		Residuals	81	3037.139	37.49554	#N/A	#N/A	#N/A	
122									
123		CV	21.19587						

The second table corresponds to the case with genotypes fixed and environments random and the third to the case with both, genotypes and environments, as random. The decomposition of the interaction with the regression technique is only shown in the first table. There are some differences between these

three tables in the F values and their corresponding probability values since the F tests depend on the type of effects in the model, random or fixed.

A least significant value at 5% significance level (LSD5) is included for the fixed effects. We can interpret these values in the following way: if all the levels of a factor had the same population mean, there is a 5% chance of getting a difference greater than the LSD5 between the sample means of any couple of levels of the factor. Hence, if any two levels have means that differ in more than the LSD5, it is usual to say that their difference is significant at the 5% level.

125	*ANOVA: G fixed, E random, B random.*							
126			Df	Sum Sq	Mean Sq	F value	Pr(>F)	LSD5
127		G	7	3807.175	543.8821	2.035997	0.098092	12.01715
128		E	3	8521.888	2840.629	32.40779	4.91E-06	#N/A
129		R:E	12	1051.832	87.65267	2.337683	0.012552	#N/A
130		GxE	21	5609.795	267.1331	7.124397	3.8E-11	#N/A
131		Residuals	81	3037.139	37.49554	#N/A	#N/A	#N/A

133	*ANOVA: G random, E random, B random.*						
134			Df	Sum Sq	Mean Sq	F value	Pr(>F)
135		G	7	3807.175	543.8821	2.035997	0.098092
136		E	3	8521.888	2840.629	8.112289	0.000354
137		R:E	12	1051.832	87.65267	2.337683	0.012552
138		GxE	21	5609.795	267.1331	7.124397	3.8E-11
139		Residuals	81	3037.139	37.49554	#N/A	#N/A

The three ANOVA tables shown above are computed by least squares, a technique that consists in estimating the parameters in the model in such a way that the residual mean square is minimized. If there are missing values in the data, they are estimated (also by the least squares technique) before doing the ANOVA, since ANOVA only works well with balanced data. For each estimated missing value, one degree of freedom is subtracted from the residual.

If all the factors are considered as random, we can estimate the model by restricted maximum likelihood (REML). This technique does not need balanced data to estimate the model, and hence can be applied with the observed data only, without including the estimated missing values. The table below shows the variance components estimation when all the factors are considered as random. Therefore, this table is computed by REML and the missing values are just ignored. In addition, a heritability estimate is included here since it is computed from the variance values. The formula for this heritability estimate is

$$H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GE}^2}{e} + \frac{\sigma_\varepsilon^2}{er}}$$

where σ_G^2 is variance component due to genotypes, σ_{GE}^2 is the variance component due to genotype by environment interactions, σ_ε^2 is the variance component due to the experimental error, *e* is the number of environments and *r* is the number of replications or blocks. Of course, the heritability estimate is computed with the estimated variance components.

	A	B	C	D	E
141	*Variance components estimation for random effects*				
142		Groups	Variance	Std.Dev	
143		G	16.86738	4.106992	
144		E	78.92911	8.884206	
145		R:E	6.260679	2.502135	
146		GxE	58.09596	7.622071	
147		Residuals	37.39044	6.114773	
148					
149		Heritab	50.00962		
150					

Stability analysis

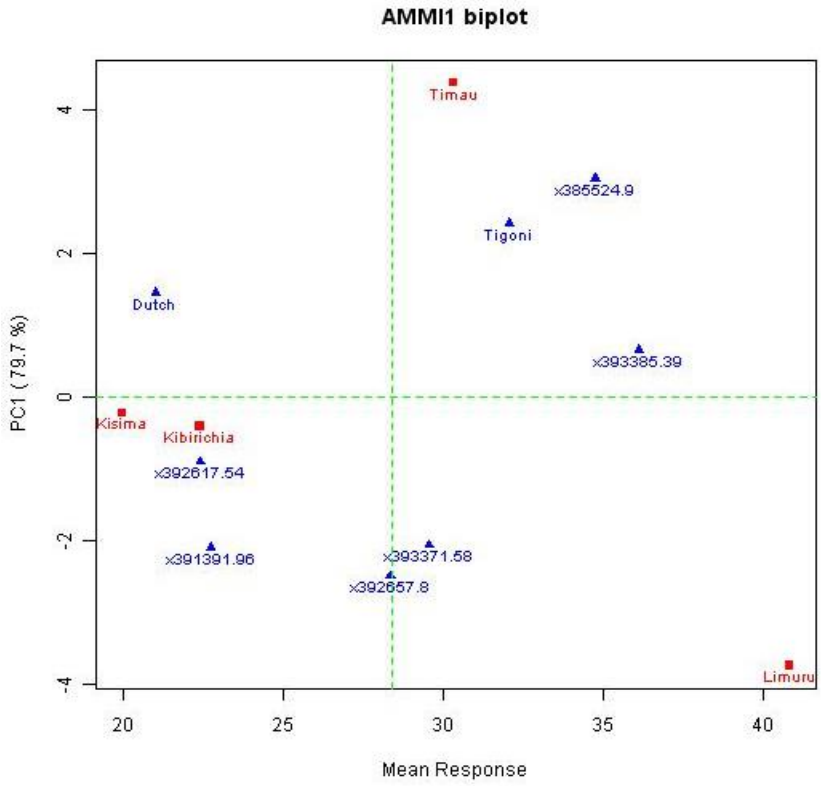
The next tables show some stability measures derived from three different approaches: linear regression (Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966; Shukla 1972), the Additive Main effect Multiplicative Interaction model known as AMMI (Gollob, 1968) and the genotypic stability analysis of Tai (Tai, 1971). We include a table for the stability analysis for genotypes and another one for the stability analysis for environments, but since the first one must be of principal interest, the comments below will be based on that table.

The first 5 columns of the tables (Slope, SE, MS-Dev, MS-Entry, MS-GxE) are results from the linear regression approach. The Slope should be the most important statistic here, because it is used as a stability parameter. According to Finlay and Wilkinson (1963), genotypes with slope around 1 have average stability over all environments, genotypes with slope greater than 1 have below average stability (they are very sensitive to changes in the environments), and hence are suitable for high-yielding environments, and genotypes with slope less than 1 have above average stability (they are very insensitive to changes in the environments), and hence, they could be suitable for low-yielding environments. SE is the standard deviation of the slope, MS-Dev is the residual mean square of the linear regression, MS-Entry is the variance of the genotype across environments and MS-GxE is the variance of the interaction effects for the fixed genotype. (Perhaps Wolfgang or Robert could write something else about the interpretation and importance for breeding purposes of these measures derived from the regression analysis)

Columns 6 and 7 have the value of the first two principal components for the AMMI analysis. AMMI is intended for balanced designs with fixed effects. Finally, columns 8 and 9 have the value of the α and λ stability statistics of Tai. (Maybe Robert could go further in the alpha and lambda explanation)

	A	B	C	D	E	F	G	H	I	J	K
163	***Stability analysis for genotypes***										
164											
165			Slope	SE	MS-Dev	MS-Entry	MS-GxE	PC1	PC2	alpha	lambda
166		Dutch	0.138087	0.225925	13.59296	10.75464	75.0084	1.449591	3.172449	-0.88936	0.848838
167		Tigoni	0.78526	0.596413	94.72852	117.8906	67.24582	2.420604	-0.29632	-0.22158	7.683597
168		x385524.9	0.875638	0.785266	164.2174	177.5417	110.8511	3.046479	-1.20657	-0.12832	13.34218
169		x391391.96	1.286789	0.485895	62.87398	188.9032	49.21713	-2.10058	-0.08224	0.295921	5.082024
170		x392617.54	1.045425	0.279372	20.78501	110.8743	14.03985	-0.90052	0.081432	0.046872	1.688684
171		x392657.8	1.135689	0.635068	107.4053	186.0977	73.23791	-2.50365	0.795156	0.140009	8.723506
172		x393371.58	1.585705	0.420212	47.02438	254.5573	61.80201	-2.06667	-1.21221	0.604353	3.703908
173		x393385.39	1.147407	0.282331	21.22769	131.0208	16.08065	0.654746	-1.2517	0.1521	1.717889
174											
175	***Stability analysis for environments***										
176											
177			Slope	SE	MS-Dev	MS-Entry	MS-GxE	PC1	PC2		
178		Kibirichia	0.52946	0.192984	8.86191	17.125	15.12217	-0.40925	2.282943		
179		Kisima	0.665941	0.172928	7.115672	21.17411	9.892573	-0.22323	1.570114		
180		Limuru	0.81602	0.61376	89.63548	99.46571	77.98102	-3.74675	-2.23459		
181		Timau	1.988579	0.560756	74.82233	198.5555	97.35406	4.379231	-1.61847		
182											

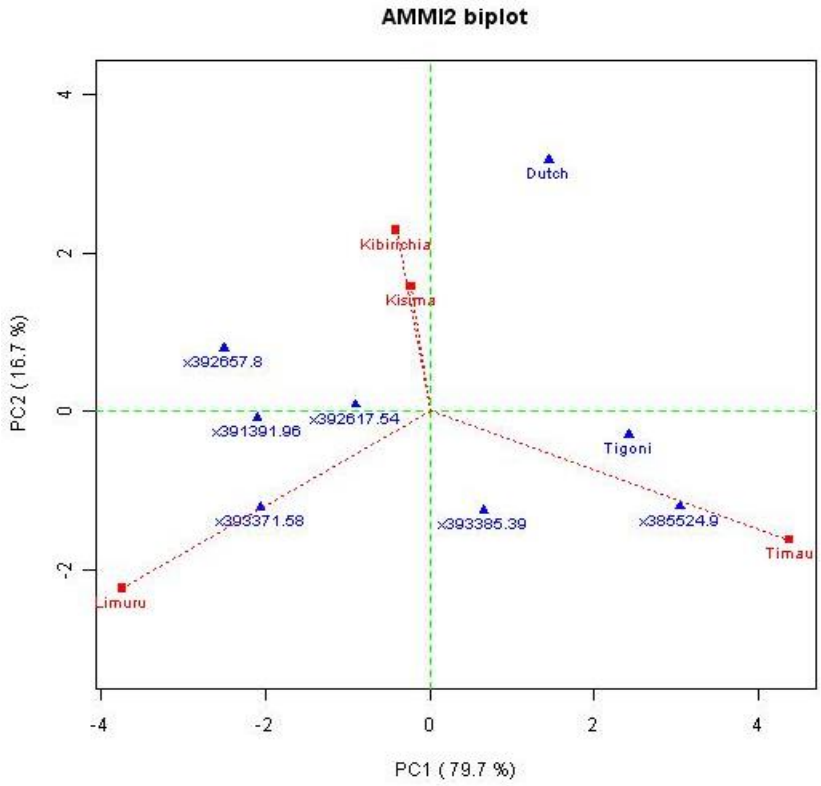
The more promising characteristic of AMMI analysis is that it allows a two dimensional graph, named biplot, where genotypes and environments are plotted on the same axes so that inter relationships can be visualized. The AMMI1 biplot shown below is a graph of the PC1 values versus the genotype and environment means.



In AMMI1 biplot, the usual interpretation is that the displacements along the abscissa indicate differences in main (additive) effects, whereas displacements along the ordinate indicate differences in

interaction effects. Genotypes that group together (e.g. 393371.58 and 392657.8) have similar mean values and, if the interaction structure explained for the PC1 is high as in this example, should have similar variation along the environments. Environments which group together (e.g. Kisima and Kibirichia) have similar mean values and again, if the interaction structure explained for the PC1 is high, should influence the genotypes in the same way.

The AMMI2 biplot is a graph of the PC1 values versus the PC2 values and helps in visual interpretation of the interaction patterns and identify genotypes or environments that exhibit low, medium or high levels of interaction effects. The PC1 coordinate explains in this example 79.7% of the interaction structure, and the PC2 explains 16.7%. Both together explain 96.4% which is quite a lot. Some care must be taking when interpreting this graph if the first two components explain a low amount of the interaction structure (I would say less than 70%). A good discussion about the validity and limitations of biplot analysis can be found in Yang et al. (2009).

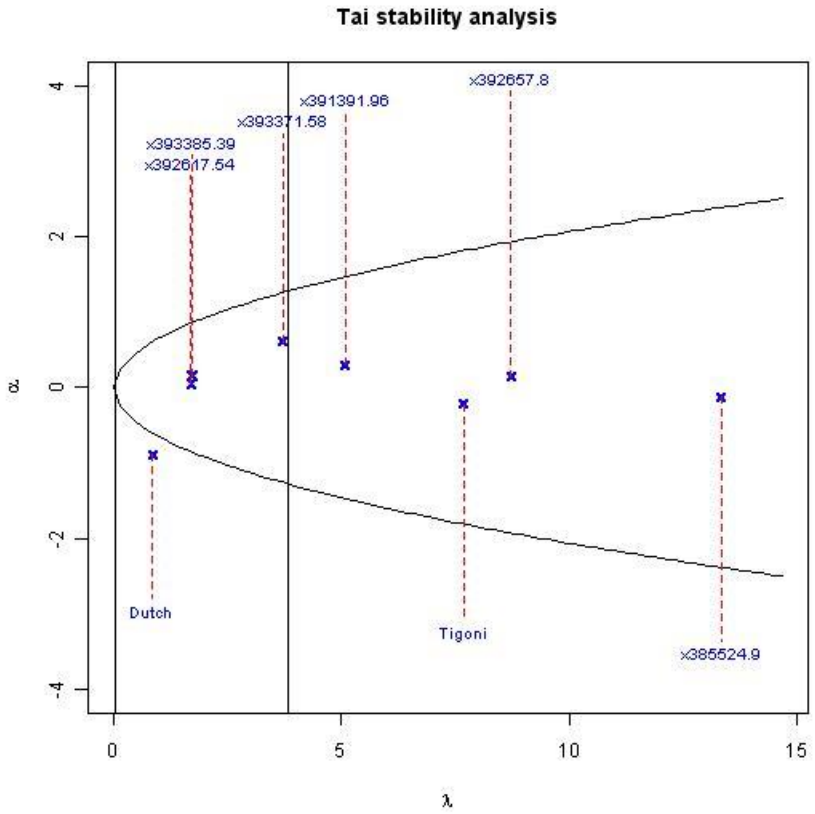


Genotypes near the origin (e.g. 392617.54 and 393385.39) are non sensitive to environmental interactive forces and those distant from the origin (e.g. Dutch and 385524.9) are sensitive and have large interactions. Points of either genotypes or environments which are near each other (e.g. those for environments Kisima and Kibirichia) have similar interaction patterns while points distant from each other (e.g. 393371.58 and Dutch) have different. Genotypes which are close to an environment should have good performance in that environment (e.g. 385524.9 in Timau or 393371.58 in Limuru) while genotypes that are far from an environment should have a bad performance in that environment (e.g. Dutch in Limuru or 392657.8 in Timau). Finally, remember that this graph doesn't explain all the

interaction structure. In this example, the first two principal components explain 96.4% so there is a 3.6% which is missing in the graph. If you want, you can think in something like a third dimension where this missing percentage is included, and hence, there is a distance between the genotypes and environments in this third dimension that we are not seeing. Therefore, it is advisable to check any conclusion you can get from the graphs with the numbers in the interaction means and interaction effects table before going on.

The mathematical trick behind AMMI is just a singular value decomposition of the interaction effects matrix. The principal components for AMMI are then just the singular vectors associated to the genotypes and the environments in this decomposition. There is a similar approach to AMMI, called GGE (Genotype main effects and Genotype by Environment interaction effects) model that applies a singular value decomposition to the matrix of residuals after removal of the environment main effects (while AMMI works with the interaction effects matrix, that is, the matrix of residuals after removal of both, environmental and genotype effects). You can have a funny tour trough the differences between these two approaches with Gauch (2006), Yan et al. (2007), and Gauch et. al (2008).

The Tai's stability analysis is a variant of the regression analysis. In the graph below the estimated values of α and λ are shown for each genotype together with some confidence limits. The α stability parameter of Tai is similar to the slope in the regression analysis. The upper and lower confidence limits around $\alpha = 0$ (which corresponds to a slope equal to one in the regression analysis) are shown in the graph (parabola shape). In the same way, a confidence limit around $\lambda = 1$ is shown with two vertical lines. (Maybe Robert could go further in the interpretation of the graph)



References

- Eberhart, S. A. and Russell, W. A. (1966). Stability Parameters for Comparing Varieties. *Crop Sci*. 6: 36-40.
- Gauch, H. G. (2006). Statistical analysis of yield trials by AMMI and GGE. *Crop Sci.* 46: 1488-1500.
- Gauch, H. G., Piepho, H. P., and Annicchiarico, P. (2008). Statistical analysis of yield trials by AMMI and GGE: Further Considerations. *Crop Sci.* 48: 866-889.
- Gollob, H. F. (1968). A Statistical Model which combines Features of Factor Analytic and Analysis of Variance Techniques. *Psychometrika*. 33(1): 73-114.
- Finlay, K. W., and Wilkinson, G. N. 1963. The Analysis of Adaption in a Plant-Breeding Programme. *Aust. J. Agric. Res*. 14: 742-754.
- Shukla, G. K. (1972). Some Statistical Aspects of Partitioning Genotype-Environmental Components of Variability. *Heredity*. 29: 237-245.
- Tai, G.C.C. (1971). Genotypic stability analysis and its application to potato regional trials. *Crop Sci*. 11:184-190.

- Yates, F., and Cochran, W. G. (1938). The Analysis of Group Experiments. J. Agric. Sci. 28: 556-580.
- Yan, W., Kang, M. S., Ma, B., Woods, S., and Cornelius, P. L. (2007). GGE biplot vs. AMMI analysis of genotype-by-environment data. *Crop Sci.* 47: 643-655.
- Yang R.-C., Crossa, J., Cornelius, P. L., and Burgueño, J. (2009). Biplot Analysis of Genotype x Environment Interaction: Proceed with Caution. *Crop Sci.* 49: 1564-1576.

Annex A: Small test of different selection indexes for sweetpotato

Jens Riis-Jacobsen and Shiphar Mulumba, Nairobi, November 16, 2010

One of the functionalities to be included in CloneSelector is Selection Indexes, and to define what methods to use Shiphar Mulumba and Jens Riis-Jacobsen have undertaken a small case study to compare some of the options for developing indexes.

The LSD index was taken from the Fieldbook Manual and is used in maize breeding. The K and Min indexes were taken from Elston (1963).

Selection indexes using equal weights for each trait

Genotype	LSD index	Rank LSD	K index	Rank K	Min Index	Rank Min
W119 06-39	1.09	1	10.97	1	2.79	1
MUSG 0703-37	0.98	2	6.73	2	1.43	2
Libertada	0.11	3	1.82	3	0.00	3
UW119 06-32	0.05	4	0.64	4	0.00	3
UW119 06-296	-1.46	5	0.27	5	0.00	3
UXIPHONE 06-1	-1.82	6	0.13	6	0.00	3

Selection indexes using weights for traits

Genotype	LSD index	Rank LSD	K index	Rank K	Min Index	Rank Min
W119 06-39	0.14	1	0.0000549	1	0.000014	1
MUSG 0703-37	0.10	2	0.0000336	2	0.000007	2
Libertada	-0.02	4	0.0000091	3	0.000000	3
UW119 06-32	0.03	3	0.0000032	4	0.000000	3
UW119 06-296	-0.10	5	0.0000013	5	0.000000	3
UXIPHONE 06-1	-0.43	6	0.0000007	6	0.000000	3

Conclusions on index comparison:

1. The K and Min indexes were created to have neutral weight, and effectively an introduction of weights did not change the ranking for any of the two indexes. For LSD the introduction of weights did change the ranking, though only moderately.
2. The Min index discards all clones that have the minimum value in a trait. In our case we analyzed 6 traits, and 4 different clones had the minimum value in at least one trait. Due to this 4 clones all had the index value cero. The K and LSD indexes rank all clones.
3. When no weight was applied LSD and K index resulted in the same ranking of the clones, and the Min index also identified the 2 top clones.

Other considerations

1. In sweetpotato we have a very large number of traits, and some such as root yield are included in various traits. Even if no weights are applied it would still appear necessary to permit the breeder to decide which traits to include in the index, and which to leave out.

Formulas used:

1. Using LSD

Selection index= (selection direction*selection weight*(X1-mean of X1))/LSD1 + ……. +selection weight*(Xn-mean of Xn))/LSDn

2. Using k

Selection index=((X1-K1)*(X2-K2*(Xn-Kn))* selection direction*selection weight

Where K= (n*Ximinimum-Ximaximum)/(n-1)

n=number of varieties/germplasm

Each trait is also multiplied with Selection direction (-1 or +1) and Selection weight, but no effect was observed of Selection weight

3.Using X minmum

Selection index = ((X1-minmum value of X1)*(X2 -minimum value of X2)*…….*(Xn - minimum value of Xn))

Where x=trait of interest

Each trait is also multiplied with Selection direction (-1 or +1) and Selection weight, but no effect was observed of Selection weight