GMOANALYSIS – USER MANUAL

VERSION 2.1.0 – 10 JULY 2014

INTRODUCTION

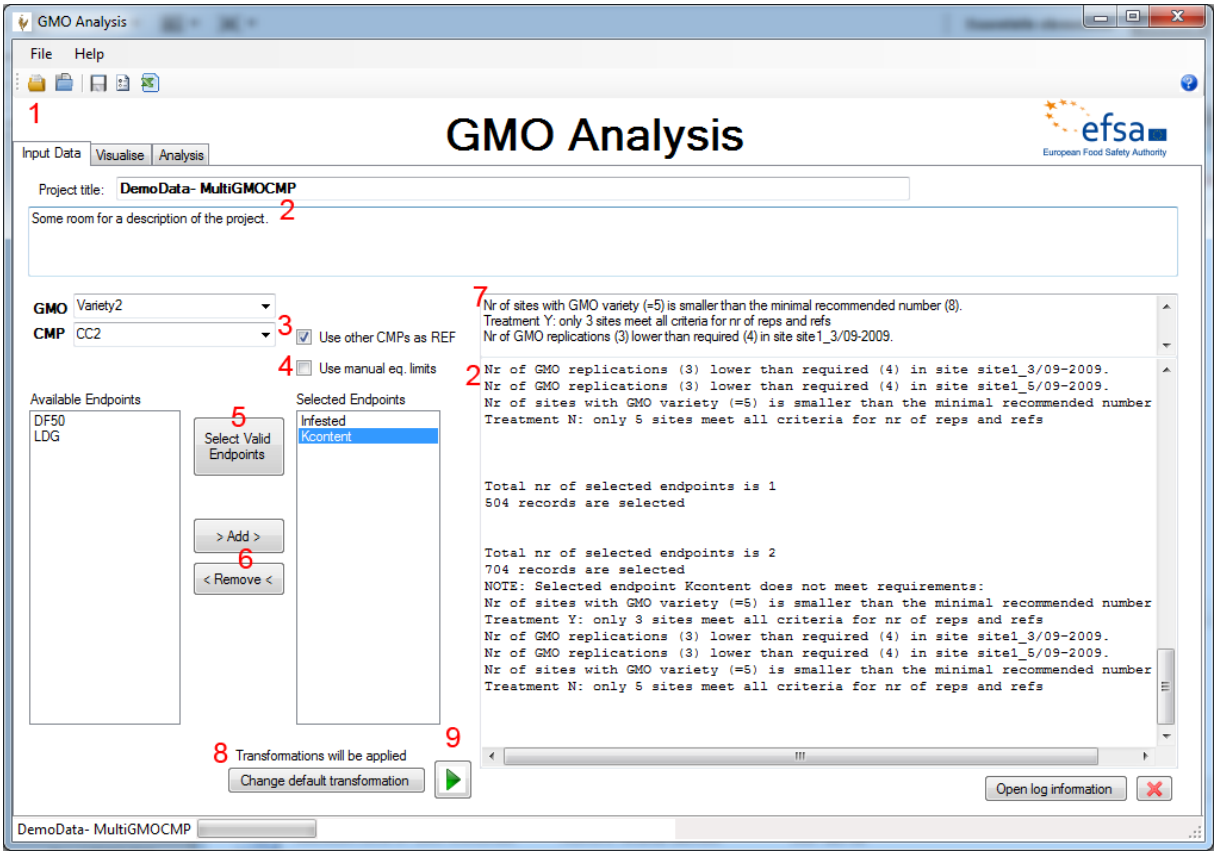
In the European Union (EU), the use of genetically modified organisms (GMOs) is regulated through a legal framework. According to EU legislation (Reg. (EC) No 1829/2003 and Dir. 2001/18/EC), GMOs can only be authorized for placing on the EU market following a scientific assessment of any risks that they may pose to human and animal health and the environment.

To guide and assist applicants for the preparation and presentation of genetically modified (GM) plant market authorisation applications, the GMO Panel has developed guidance for the risk assessment of GM plants and derived food and feed. This guidance describes principles, concepts, data requirements and issues to be considered when performing the risk assessment of GM plants and derived food and feed.

In 2009 the EFSA GMO Panel adopted an opinion on 'Statistical considerations for the safety evaluation of GMOs' in order to provide more detailed guidance on the performance of field trials and the statistical analysis of compositional data for the FF/MC risk assessment of GM plants. These statistical issues are described here. This software tool has been developed for EFSA by Wageningen UR to help applicants to perform the statistical analysis along these guidelines.

STARTING THE PROJECT

In the "Input Data" tab-sheet is the starting sheet of a project. In this sheet you can import and select the data of the analysis.



1. IMPORT EXCEL FILE 

Generally you will start an analysis by reading an Excel file with your data. The format of the file is described in the Annex. The Excel file can be imported using the menu (Import ExcelFile) or by clicking on the Import Excel File button in the toolstrip.

An example datafile is provided with the program in the Help directory (ExampleDataFile.xls).

The program will try to read data from worksheet Sheet1 in the Excel file. In the worksheet, the first row contains the name of the columns as defined in the format. The order of the columns is irrelevant, with the exception of the first column. This column should be ResultID. It indicates a unique number (integer) for each record in your data file. The reading stops after the first empty cell in ResultID.

A range of synonyms is available for the names of the required and optional columns. In addition, the program will read additional columns with quantitative or textual data, which will then be available as covariates to be corrected for in the modelling. You can always inspect the original Excel file by clicking the Excel button in the toolstrip.

You can use the File menu and toolstrip to save the data, settings and results in a GMO Analysis project file (xxx.gmo), after the Excel data have been read. The project file can be opened for later use and changes.

2. INFORMATION ON THE DATA

The name of the data file will become the name of the project. You can add an extra description of the project, if you like. This will be shown in the pdf documents containing the results.

General information about the data import will be displayed in the right large textbox, as well as some simple statistics of experiment. If needed, the log-text can be saved as a *.txt file on your temporary directory and opened in your associated default txt-editor (e.g. Notepad.exe) by clicking on the "Open log information" button.

The following checks are made when data are read:

- Checks for valid Excel file with Sheet1;
- Check for required and optional column names in the first row of the sheet; other column names are allowed but skipped.
- Check for types and remarks as described in the Annex;
- Check for non-duplicate ResultID;
- Check for match of Site and Location/Year combination. Optionally leave the SiteD empty and it will automatically use the Location/Year combination for this;
- Check for at least 5 records per endpoint;
- Later other checks are made per endpoint regarding nr of sites and nr of varieties in the experiment;

If you need to correct the data, please use Excel. You can open the Excel file within GMOAnalysis by pressing on the Excel button in the toolstrip. After making adjustments, save and close the Excel file, and import it again.

In the listbox Available Endpoints, an overview is given of all endpoints that have been found in your data-file.

3. MULTIPLE GMOS AND CMPS IN THE DATASET

If there are multiple GMO's and CMP's in the dataset, two listboxes will appear at the top, where you can select the combination of GMO and CMP that you want to analyse (red 4). To analyse another combination, you have to do a separate analysis (e.g. by saving the analysed project in a project file, open it and make the necessary changes, this allows you to keep all other settings the same).

If you have multiple CMP's in your data file, you can choose to use the other ones as REF variety using the checkbox. Note this is only allowed if the other CMPs are all non-GMO varieties.

4. SETTING EQUIVALENCE LIMITS MANUALLY

Sometimes, you want to set the equivalence limits manually, instead of determining it from the statistical analysis. If you want to use the manual equivalence limits, you can set the checkbox on this page. When selecting endpoints a new window will be opened, where you can enter the limits. You specify the limits relative to the mean of the comparator. This can be as a percentage (e.g. 80 and 120%) or as a multiplication factor, e.g. 1.2. In that case, the lower limit will be x/1.2 and the upper limit x*1.2 where x is the mean of the CMP in the analysis. When done, simply close the limits window by pressing on the right top cross of the window. Data are saved automatically when closing the window.

When manual limits is checked, an icon become available in the toolstrip to access the limits. By clicking on the limits icon in the toolstrip or by choosing it from the File menu, you can see and edit the limits. If you select manual limits, no reference varieties are required. However, if reference varieties are available they will be used in the analysis of the difference test.

5. SELECT ALL VALID ENDPOINTS

Click on the button “Select Valid Endpoints” to analyse all your endpoints at once that meet the requirements for the analysis. In the log-box at the right some information will be shown on any possible problems. If not all endpoints are selected by this button click, you either have errors in your data or the data of the non-selected endpoints do not meet the requirements as described in the EFSA guideline.

6. (DE)SELECTING ENDPOINTS

Even if an endpoint does not meet all requirements, it might be helpful to further visualize and/or analyse this endpoint. E.g. if a few sites have some missing varieties. You can still select the endpoints by selecting them in the Available Endpoints listbox and then pressing the Add button. You can select multiple endpoints at the same time using the Shift and Alt key with the mouse button.

Note that endpoints with errors cannot be analysed. Endpoints which do not meet the requirements can in general be analysed. Furthermore you can discard selected endpoints from the analysis by selecting them in the "Selected Endpoints" listbox and then pressing the button "Remove".

7. INFORMATION ON REQUIREMENTS FOR THE ENDPOINTS

You can see information on data errors and requirements for each individual endpoint by clicking on an endpoint in either the Available Endpoints and Selected Endpoints listbox. The information on errors and requirements will then be shown in the top right textbox. Note that the requirements on the number of reference varieties depends on whether you want to use manual equivalence limits or calculate them from the statistical model. If you use manual limits, no reference varieties are required. The requirements that are checked are:

- Minimum number of sites is 8.
- Minimum number of sites with GMO is 8.

- Minimum number of sites with GMO and CMP is 8.
- Minimum number of REFs is 6.
- For each site, the minimum number of replicates with REFs depends on the number of REFs. If there is only 1 REF, it should be available in at least 8 replications. It there are 2 REFs, they should both be available in at least 6 replications. It there are 3 or more REFs, they should be available in at least 4 replications.

Selecting an endpoint will display the results in the Data Checks box. Checks are made for both treated and untreated GMO separately. As stated before, you can still select an endpoint for visualization or analysis, by clicking on the ">Add>" button. In that case, a message that the requirements are not met will be printed in the output, and the analysis tables will show the endpoint name in a red font.

8. TRANSFORMING THE DATA

Before you look at the data and do the analysis, you can choose to select a data transformation which will be applied, e.g. a logarithmic transformation. You can specify the default data transformation depending on the ValueType of the EndpointID. E.g. a Fraction will by default be analysed using a arcsin sqrt transformation, and Counts using a sqrt transformation. In the Options window you can change the default data transformation for every ValueType. For each specific endpoint, you can later override the default transformation with your own choice of transformation in the Visualisation tab on the next page. Note, that the results of GMO Analysis will be on the ratio scale if all selected endpoints are logarithmically transformed, and on a difference scale of transformed values if one or more endpoints is transformed non-logarithmically.

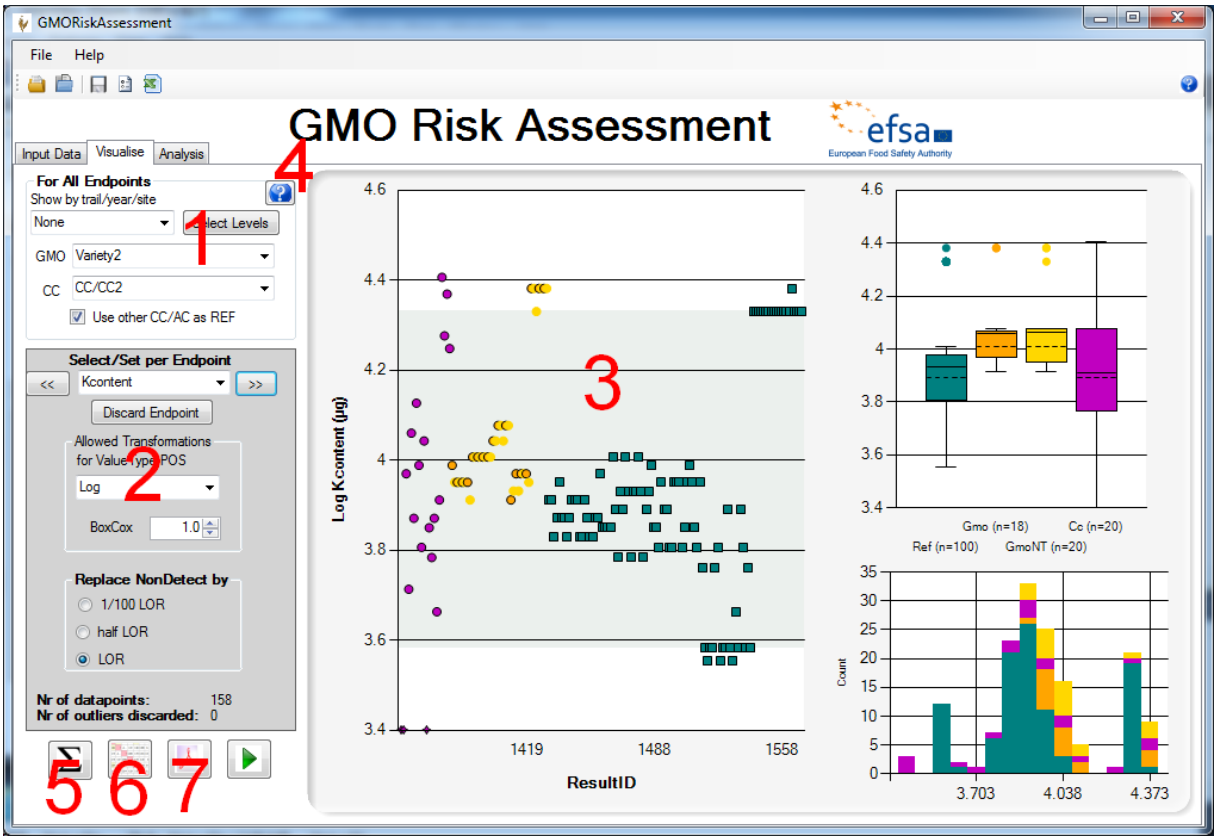
9. GO TO NEXT TAB

By clicking on the next tab-sheet [Visualise] or the green arrow , you can now examine your data in more detail and specify settings.

Note: this arrow is greyed out if no endpoints are selected.

VISUALISATION

In the “visualise” tab you can set defaults for ***all endpoints***, as well as specific settings ***per endpoint***, as well as excluding individual datapoints (outliers).



1. SELECTIONS FOR ALL ENDPOINTS

In the top left part of the window (see red 1), the general settings for ***all*** endpoints can be set. Here you can order the data set in different groups, e.g. years, locations or sites. The data in the point chart will be ordered and the different groups/levels are indicated by alternating striped areas. This is helpful to find abnormal sites. You can then decide to exclude such sites from the analysis by clicking on Select Levels and click which levels you want or do not want to include in the analysis. If you exclude levels from a grouping factor, these data are not shown anymore and will be excluded from the analysis.

Note 1: exclusion of certain levels will be used for all endpoints simultaneously**.**

Note 2: in case the GMO has different treatments, these groups are shown with different symbols in the graphs.

2. SELECTIONS AND SETTINGS PER ENDPOINT

The grey panel on the left (Select/Set per Endpoint) allows you to examine and change settings for each individual endpoint. You can select a specific endpoint from the dropdown list. Alternatively, you can scroll through the list by clicking on the left and right arrow at each side of the listbox. By clicking on the Discard Endpoint button below the list, this endpoint will be deselected for the analysis and added to the Available Endpoints (on the previous tab).

Depending of the ValueType of your endpoint, as specified in the Excel file, a number of choices for a transformation are available. For POS-values, the default transformation is LOG, for COUNT it is SQRT. You can change these default transformation settings in the Options menu (see later). For every ValueType at least the option Identity (=no) transformation is available. For CONT values, only Identity is available, as most transformations require that data are non-negative. If you choose a Box-Cox transformation, you have to provide the lambda parameter. The effect is shown in the graphs, especially the histogram at the bottom right.

When all endpoints use a logarithmic transformation (ln or log10), the results of statistical analysis will be expressed as ratios, and variation as coefficients of variation (CVs). If at least one endpoint is transformed in another way (or not transformed) all results will be expressed as differences on the transformed scales, and variation as standard deviations on these scales.

If your data contain NonDetects an option is shown where you can specify a value for replacing the non-detect. Typically this is either the Level of Reporting (LOR) itself, half of its value, or zero. Non-Detects are shown as diamonds in the graph. In the case of a ln or log10 transformation, a zero is not allowed and a value of 0.01 is used before transformation.

3. INTERACTIVE GRAPHICS

Outliers can be spotted in the point chart. Generally it is very helpful to show the endpoints ordered by site (red 1). By clicking on a point in the point chart, it will be selected as an outlier (or vice versa). The symbol will then change into a cross instead of a square (REF) a circle (GMO or CMP) or diamond (NonDetects). The number of discarded points for this endpoint is shown in the bottom of the grey panel at the left. You specify outliers per endpoint, i.e. these are not treated as outliers for the other endpoints.

You can zoom in on the point chart by selecting a rectangular area, keeping your left mouse button down. Use the scrollbar and indicators on the axes to pan through the chart or to zoom out.

The grey area in the chart (red 3) is calculated as the range between the largest and smallest value in your dataset, still fitting within a distance of 1.5 times the size of the Inter-Quartile Range (IQR) from the IQR. The IQR is the difference between the 25 percentile and 75-percentile. You can adjust the size of the grey area by changing the value 1.5 in the Options window from the File-menu.

The IQR is also used for the boxplot on the top-right, but here it is calculated per group (over the sites). The coloured box indicates the IQR (25-75%). The whiskers have size 1.5*IQR and extend from the IQR above and below. Any points outside the whiskers are marked as dots and are potential outliers. The boxplots are shown for each of the different genotype groups with possible outliers (circles), mean (solid line) and median (dotted line). Note that you can adjust the size of the whiskers by changing the value 1.5 in the Options window from the File-menu.

In the Visualization tab of the Options menu you can also select the size of the markers and the colours to be used for the different groups.

The histogram at the bottom right can be used to see if your transformation is adequate. By selecting a different transformation (red 2) the shape of the histogram will change to reflect the distribution of the data after transformation. Ideally the histogram should look like a symmetric normal distribution.

Note that in the example here, one site clearly had larger values. This makes it more difficult to interpret the normality of the data. You could try to exclude the site from the data using "Select levels" button. Do not forget to add the site again, as it will otherwise be excluded from the analysis for all endpoints.

4. LEGEND/HEADER OF THE GRAPHS

An explanation of the meaning of the different colours and symbols is provided in a pop-up window by clicking on the Legend button in a corner of each of the graphs.

5. STATISTICS

After you have made your selections for the GMO variety and the corresponding CMP variety, and if necessary, discarded some levels from the Site/Location/Year factors, and you have inspected all the endpoints and set the correct transformation, chosen the replacement of Non-detects by LOR or another value and removed any outliers, you can view the descriptive statistics per endpoint by pressing the sigma button Σ.

6. CORRELATION

You can study the correlation between all the endpoints by clicking on the correlation button . The correlation is calculated by a pairwise comparison of the endpoints at the plot-level. It will pop up a new window where you can select thresholds for colouring the correlation. By default values above 0.7 are shown in pink and between 0.5 and 0.7 in yellow. You can adjust these values using the trackbar (sliders). Furthermore, you can adjust the default thresholds to be used for colouring the correlation matrices in the Options window.

By hovering over the matrix, you can see to which combination of endpoints this correlation belongs. Note that the correlation is calculated in the statistical package R, which should be installed on your computer. You can store the matrix as a .csv file which can later be opened in Excel.

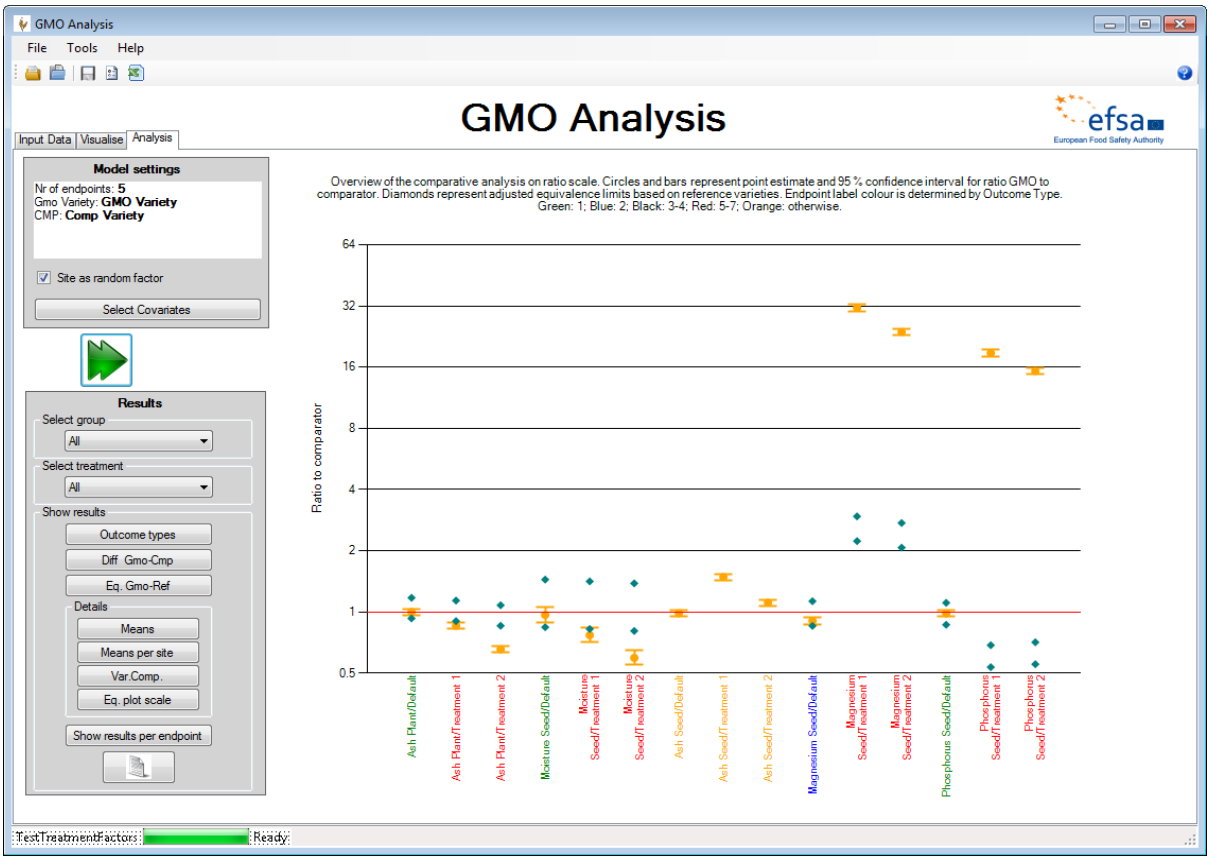
7. SAVE YOUR DATA

You can store the descriptive statistics as a PDF-file for future printing and reference. After the pdf is saved you are asked if you want to open it. The PDF will contain a picture of the three graphs for each endpoint. This will be at the original 100% zoom factor, i.e. any zooming in the graph is discarded.

If you are ready inspecting and adjusting the data, click the green arrow to go the analysis tabsheet, or click on the tabsheet Analysis. After you have chosen all the settings and outliers, it is strongly advised to save your data as a new GMO project (xxx.gmo file). To do so, just click on the Save Project button in the toolbar or select it from the File-menu. You can see the name of the current file/project in the bottom left of the window. If needed, you can open this project file again using the Open Project button from the toolbar, or using the File-Open Project menu.

ANALYSIS

In the “Analysis” tab-sheet you can start and evaluate the analysis.



1. DATA SUMMARY

You can see a short summary of your choices in the upper left corner: how many endpoints will be analysed, the names of the selected GMO and CMP.

2. SITES AS RANDOM OR FIXED FACTOR

Further there is a checkbox to choose whether you want to do the analysis for Sites (Year/Location combinations) as a random factor or as fixed factor. If you consider the sites to be a random selection from a large population of sites, and you are not so much interested in the specific effect of each site, but more in the variance introduced by these sites, you can specify the sites to be a treated as a random factor. If you have chosen these sites specifically and are interested in the means of the sites, you can specify this factor to be treated as fixed by unchecking the checkbox. The default choice is to treat sites as a random factor. If you have chosen to treat the sites as fixed effect, the resulting variance component for this factor will of course be zero.

3. COVARIATES

If your data file contains additional columns, these might specify covariates for the statistical analysis, and a ”Select covariates” button will be shown. Clicking this, a pop-up window will appear, allowing you to select covariates for inclusion in the statistical models. There are three possible forms of covariate: 1) as a numerical covariable (available for quantitative data only); 2) as a fixed cofactor, where all equal values in the column will define a group, and all groups are estimated by a fixed effect; and 3) as a random cofactor, where all equal values in the column will define a group, and the group levels are considered as random realizations of levels from a normal distribution on the transformed scale.

The Covariate pop-up window also allows to specify interactions between the covariates, with the exception of interactions involving more than one covariable (Note: such interactions can be included by adding an additional column in the input data file, containing the multiplication of mean-corrected columns).

The selected covariates and interactions will build up a partial model formula that will be used in all statistical models calculated by the R lmer function. The partial formula is shown, and can be edited by experienced R users to allow even more specific models, for example random coefficients models, or nested models. However, the choice of covariates should be restricted to the non-excluded covariates in the upper part of the pop-up window. Note, that editing the model formula will make the upper part of the window irrelevant, e.g. a covariate can be removed by editing although it is still shown as selected. Unclicking the Editing function will return the model formula to its default form.

Click *OK* in the Covariate selection window to fix the covariate model to be used in the analyses. Please note, that it cannot be guaranteed that the selection of covariates, covariate interactions and covariate models in the Covariate selection window leads to statistically sensible models. The responsibility for the appropriateness of the statistical model therefore is with the user of the program.

4. RUN ANALYSIS 

You can click on the green double arrow. In the bottom of the window the progress will be indicated by a progress bar and a status of the current endpoint. The data are analysed per endpoint. If you have set a grouping factor on in the visualisation tab, the analysis will be done for each group separately and shown combined.

The analysis is done with a REML mixed model in the statistical package R (using RDotNet for the communication). You can choose to log all R-commands and even all data in the Output tab of the

Options window. This allows you to run the analysis stand-alone and check for any errors that were encounterd. Of course it also allows you to extend the script with your own commands.

For the analysis four specific non-standard installed libraries in R are needed: lme4, pbkrtest, lsmeans and arm. These libraries need to be installed. You can install them within R once manually (see the installation requirements). Note: this program uses the 32-bits / i386 version. Alternatively, this program will try to download and install the library during first start-up, but this may fail due to security restrictions of your machine or if no internet connection is available. If needed, you can cancel the calculations by pressing the red stop sign which becomes visible during calculations.

The analysis will always be carried out, even if the number of sites and reference varieties do not meet the requirements as described in the EFSA opinion. In the results, the number of sites and number of reference varieties is printed and a warning will be included in the output of the results if the prescribed requirements are not met. If requirements are not met, or if something could not be analysed properly, the endpoint name in the main tables (Outcome type, Difference Test and Equivalence Test), will have a red font. When hovering over it, the warning will be displayed as a tooltip.

If you have chosen manual equivalence limits (on the first page, when reading the data and selecting the endpoints), the equivalence limits shown in the results will be your manually chosen limits.

5. RESULTS

After the analysis is finished, you will see a chart of all the endpoints. It shows the difference test (point estimate and confidence interval of the GMO-CMP comparison) in orange. The (adjusted) equivalence limits are shown as green diamonds.

If all the endpoints are transformed to log-scale (as is the default for ValueType POS) the graph is shown on a multiplicative scale, as was suggested in the statistical consideration paper of EFSA. Thus confidence intervals are shown for the ratio of the GMO to its conventional counterpart. A ratio close to 1 indicates that the means are similar. Means are shown as (geometric) means on the original scale.

However, if some endpoints have a different transformation, e.g. Identity or SQRT, the multiplicative scale is not the right choice. In that case, ALL the results will be shown on the transformed (linear) scale. On this scale, we are interested in the difference (not ratio) between GMO and CMP, and this should be close to zero if the two varieties are similar. The previous mentioned multiplicative scale is a dimensionless scale, in which treatment effects are expressed in terms of proportional or percentage change, which can be easily compared over multiple endpoints. However, a difference on the transformed linear scale can simply be made smaller or larger by changing the unit of measurement (e.g. expressing it in milligram instead of microgram). Therefore such a plot should be studied with more care.

The graph corresponds with the graph in the EFSA GMO panel opinion

The labels of the endpoints shown at the X-axis are colour-coded , based on the outcome type as follows:

- Outcome type 1 (Equivalent): label is green;
- Outcome type 2 (Equivalent more likely than not): label is blue;
- Outcome type 3-4 (Non-equivalent more likely than not): label is black;
- Outcome type 5-7 (Non-equivalent): label is red;
- Otherwise: label is orange.

Clicking on a label of an endpoint on the X-axis of the graph will pop up a window with detailed information for the specific endpoint.

Buttons will appear in the left panel after the analysis to show the results in tabular form. If your data file contains multiple groups or multiple treatments for the GMO, you will see drop-down buttons to restrict the display to specific groups and/or treatments.

Clicking on the Show Results buttons will show tables over all the endpoints, as specified in the EFSA opinion on 'Statistical considerations for the safety evaluation of GMOs'. The three main tables are:

- the overall outcome type (combined equivalence and difference test)
- the results of the difference test
- the results of the equivalence test

More detailed information is provided in a table containing the overall means for GMO, CMP and REF, a table with the means per site, a table with the variance components and a table with plot scale assessment of equivalences. If the table heading does not provide sufficient detail, see the EFSA opinion document for a further explanation.

The rows in the tables can be sorted, by double clicking on a header. By clicking again, you can revert the ordering (up/down). Double clicking in a cell will pop-up the window with detailed information for the specific endpoint.

If the endpoint label in a table is red, then it did not meet all requirements, or something went wrong in the analysis. A tooltip text will provide the warning message. Double clicking on the cell will show a pop-up window with detailed information for the specific endpoint, also including the warning.

Do not forget to save your project (from the toolstrip or menu). You can specify your own name for the project. It may be helpful to indicate the selected Gmo treatment in the name of the project. Finally the results can be saved as Word docx document or a PDF file. Since the project name and description (sheet 1) is printed in the output, it is advised to save the project under the right name before saving the PDF/word-document.

6. ENDPOINT REPORT

The button 'Show results per endpoint' gives a pop-up window with all detailed results for one endpoint. This report will first show the conclusions of the difference and equivalence tests and then provide detailed information of the analysis. Note that this same window can also be reached by double-clicking endpoint labels in the over-alll summary graph, or by double-clicking cells in a table.

In the toolbar of this window the endpoint (and, if relevant, GMO treatment) can be chosen. The report per endpoint can be saved as a Word or pdf document. In order to help identification of points with large residuals, the toolbar also has buttons to show interactive versions of the residual plots for the various fitted statistical models: by hovering over points a tooltip will display the ResultID and Variety group.

By default the Results per endpoint windows and reports also include the statistical output of the REML analysis done by the lmer function in R, including standard errors for fixed effects. You can suppress the REML output by unchecking the corresponding checkboxes in the Output tab of the Options menu. The Options menu can be opened from the menu (File | Options) or toolbar in the main window.

7. TOTAL ANALYSIS REPORT

The lowermost button gives a pop-up window with all tables with combined results and all detailed results for all endpoints. This report will first show the conclusions of the difference and equivalence tests and then provide detailed information of the analysis. Note that this same window can also be

reached by double-clicking endpoint labels in the over-alll summary graph, or by double-clicking cells in a table.

The total analysis report can be saved as a Word or pdf document.

By default the total analysis report also includes the statistical output of the REML analysis done by the lmer function in R, including standard errors for fixed effects. You can suppress the REML output by unchecking the corresponding checkboxes in the Output tab of the Options menu. The Options menu can be opened from the menu (File | Options) or toolbar in the main window.

OPTIONS

You can specify/alter the default settings in the program using the Options menu. The options menu can be opened from the main menu (File | Options) or by pressing the Options icon on the toolstrip. It consists of 4 tab pages.

1. IMPORT DATA OPTIONS

Choose a default transformation for different ValueTypes. In many cases it is advisble not to analyse the data on the original scale, but after a suitable transformation. The ln-transformation is most often used for this. This program allows you to set the default transformation per different ValueType. Which transformations are allowed depends on the ValueType of the endpoint.
E.g. for POS, allowed transformations are Ln (Log (x)), Sqrt (square root), Log10 (10Log(x)) , BoxCox ((y^a - 1)/a), where a is the lambda parameter to be specified.
For counts the default transformation is Sqrt, but also a Ln_1 (log (x+1)) transformation is available.
For fractions between 0 and 1, the ArcSin of the square root (ArcSinSqrt) is an often used transformation to obtain approximate normality.
In the visualisation tab of the program you can override the default transformation on an Endpoint-by-endpoint basis, e.g. based on the histogram chart.

There is also an option to start the program in full-screen mode or standard window size.

2. VISUALIZATION OPTIONS

Choose colour, marker size and size of potential outlier area (outside grey horizontal bar in the point chart. The outlier range factor (ORF) is used to construct the horizontal grey area in the pointchart. This is the area within which most points are likely to be. Points outside this area are potential outliers. The grey area is calculated as (1+ORF) * the inter-quartile range (IQR), where IQR is the range where 25-75% (mid 50%) of all points are located.
It is also used for calculating the size of the whiskers in the boxplots. The default ORF is 1.5.

3. CORRELATION OPTIONS

Set the default thresholds for the lower correlation and upper correlation in the correlation matrix window. Cells with an absolute correlation value above the upper threshold (default 0.7) are coloured pink, between lower and upper yellow. Below lower is white.

4. OUTPUT OPTIONS

Here you can choose to suppress REML summary output (both in the Endpoint Report window as in the Word/PDF-documents).
A standard summary output from R also prints the correlation matrix, which can be large. You can suppress storing this matrix in the summary using this setting. Note: this option only has an effect

during the analysis.
Furthermore you can choose to save the R-commands used to analyse the mixed models in a R-source file (you can specify the name, if you leave it open it will be stored in a temporary GMO directory). Besides only storing the commands, you can also save the (transformed) data in R-structures, which will be included in the source file. This allows you to run this source R-file stand-alone in R, e.g. for easier debugging, detailed analysis or to add your commands.

You can click on Reset to restore the Factory defaults of the program. By clicking Cancel, you will restore the defaults before you opened the Options window. By clicking OK, you will store the newly chosen options for future use.

TROUBLESHOOTING

1. THE PROGRAM CANNOT CONNECT TO R

The main known problem is that GMOAnalysis cannot properly connect to R. The program will try to connect to R automatically, using RDotNet.dll and the registry settings of R. Sometimes this does not work as planned, probably due to security reasons. Depending on the error message, you should do the following:

1. Make sure that R (version 2.15 or higher) is installed properly as described above.
2. The GMO Analysis program requires installation of four R-libraries (*lme4*, *pbkrtest*, *lsmeans, arm*). If they are not yet available, it will attempt to automatically download and install them the first time GMO Analysis is run. If this fails, for example due to security reasons, you can try to install the libraries manually. For manual installation, start R (32-bits version / i386). At the R-prompt, type the following eight lines:

```
install.packages("lme4")
library(lme4)
install.packages("pbkrtest")
library(pbkrtest)
install.packages("lsmeans")
library(lsmeans)
install.packages("arm")
library(arm)
```

3. If you still get the message that it is not working, look in the error-log-file which will open if an R-error is found during startup. Check if the PATH environment variable in the log file contains your R-directory. If it does not contain the R-directory, then there is a (security) problem to set this variable within the GMOAnalysis software. To solve this, you should add the R-directory to the environment variable PATH manually as described next.

2. MANUALLY CHANGING THE PATH ENVIRONMENT VARIABLE TO SEARCH FOR R

GMOAnalysis tries to automatically set the correct path for R and its libraries during startup. If this fails, you need to change the PATH variable by hand, which may require administration rights on your computer.

To add the R-directory to the PATH environment variable for Windows 7, do the following.

- Look where the *R.exe* file is located, e.g., *C:\Program Files\R\R-3.0.2\bin\i386*. Copy this directory string to the clipboard or write it down.
- Now you need to go to the environment variable section in system properties: Open the Control Panel and click on the System icon. This will open a System Control Panel. Click left on Advanced System Settings, which opens the System Properties window, with tab Advanced. Right down you

can find the Environment Variables button. Click this button and in the window Environment Variables, go to the System Variables section, where you select the variable PATH (Path) from the list.

- Choose Edit... Now set the PATH variable manually by adding the R-directory string to the SYSTEM environment variable PATH (make sure to insert a semi-colon between directories). You can add it at the end or in front. Click OK. This should solve the problem.

3. SPECIFIC R LIBRARY COMMANDS MAY HAVE CHANGED

The R-libraries each have different versions, which may lead to conflicts. We experienced this with *pbkrtest*. In case you suspect that this is the case, please try to install the newest versions of the packages for the appropriate R version. You can do this by running the following line in R:

```
update.packages()
```

4. READING OF AN EXCEL FILE FAILS

When an Excel file cannot be fully loaded, a possible reason is that the last column is ignored in the reading process. Making a clean copy of the data in excel to a new file will solve this problem.

ANNEX

INPUT DATA FORMAT

Element name	Mandatory	Type	Remark
ResultID	Y	Integer	For unique reference to the original row in the Excel file. May not contain empty cells. The first empty cell of ResultID is used as a stopping sign of the data, so all data (rows and columns) after an empty ResultID cell are ignored.
SiteID	N	Text	A site is a combination of Location and HarvestDate. It will be automatically constructed from LocationID and HarvestDate if not present, but it can also be manually provided. In the latter case, it must be set for every row and it must match exactly 1 Location per site. HarvestDate is then ignored. Eg. L1/2013a and L1/2013b for two harvests in 2013 at Location L1.
SiteText	N	Text	Optional, can be empty. It may contain a description of the SiteID.
Country	Y	Text	Can be an empty cell, otherwise it should preferably be a country code following the ISO 3166-1 alpha-2 code: http://en.wikipedia.org/wiki/ISO_3166-1
LocationID	Y	Text/Number	Can be either an integer (whole number) or a text. No empty cells allowed
SowingDate	N	Date	Excel-date format, if available.
HarvestDate	Y	Date	Excel-date format. It can be specified as day-month-year (e.g. 31-07-2013), month-year (Jul-2013) or only year (e.g. 2013). No empty cells are allowed. If SiteID is not available, levels of SiteID are constructed from the combination of Location and HarvestDate. Thus, if multiple harvest dates per year are available, this will result in different sites. Note: preferably use only year or month-year.
CropID	Y	Text	Short name/identifier of the crop. Can be at most 20 characters long and may contain any Unicode characters. Should be the same ID in the whole file.
CropText	N	Text	Further description of the crop if needed. Should be the same in the whole file, if supplied.
BlockID	Y	Text/Integer	Can be either an integer (whole number) or a text. No empty cells allowed. Every different value is treated as a different block.

PlotID	N	Text/Integer	Can be either an integer (whole number) or a text. Allows to specify multiple measurements per plot.
TestMaterial	Y	{CMP,GMO,REF}	One of these 3 text strings. No empty cells allowed. CMP=comparator (conventional counterpart or otherwise), GMO = the genetically modified organism, which is the genotype to be tested, REF = reference variety.
VarietyText	Y	Text	Name of the variety. It is used to distinguish between the different REF's, but also for multiple CMP's and GMO's. No empty cells allowed.
Treatment	Y	Text	Can be used for the GMO-variety to distinguish between different treatments.
TreatmentText	N	Text	A description of the type of treatment for the GMO.
EndpointID	Y	Text	Unique identifier string for the endpoint/analyte of no more than 10 characters for an Endpoint/Feature. Should begin with a character [a,z, A-Z] and followed by any of [a-z, 1-9, '-', '_'] or spaces.
EndpointText	N	Text	Full description of the EndpointID (unicode)
EndpointGroup	N	Text	Can be used to group endpoints in logical groups. This grouping is used in the analysis output and reports.
EndpointValue	Y	Number or empty cell	Meaning and allowed value depends on ValueType (see further). If ValueType is: POS: strictly positive number COUNT: whole non-negative number CON: any number (positive or negative) FRACTION: whole number between zero and EndpointTotal. The fraction is calculated as EndpointValue/EndpointTotal. ND: the EndpointValue represents the LOR (Level of reporting) for Non-Detects. Empty cells are used for missing observations.
EndpointTotal	Y/N	Number or empty cell	It is necessary, if ValueType is FRACTION. Then it should be a positive number, used as denominator for the fraction (e.g. total nr of plants). Otherwise it is an empty cell.
ValueType	Y	{POS, FRACTION, CON, COUNT, INDICATOR, ND}	CON = continuous measurement, can be any number, also below zero. No transformations are allowed for CON. COUNT=EndpointValue is a count, i.e. a whole non-negative number FRACTION=the Endpoint measurement is

			calculated as EndpointValue/EndpointTotal. ND = to indicate a Non-Detect, i.e. a measurement below Level of Reporting. POS= EndpointValue should be strictly positive number. INDICATOR: strictly 0/1 measurements. This one is not used. No empty cells are allowed.
Unit	Y	Text	Unit of the measurement of the EndpointID, e.g. mg/g, %, g, days etc. Empty cells are allowed.
Tissue	N	Text	Optional, is combined with EndpointID. To be used to distinguish between endpointIDs measured at different plant parts. E.g. for dry weight of seed and of root, you can have EndpointID = Dry Weight and Tissue = Seed and Tissue = Root. In that case, the EndpointIDs are treated separately.

NOTE: more restrictions apply, e.g.

- Only one option is allowed for Unit and EndpointText per EndpointID/Tissue combination.
- For ValueType, one option from (CON,POS,COUNT,FRACTION) is allowed per EndpointID and also ND for a non-detect measurement.
- Only one TreatmentText is allowed for a certain Treatment
- A site is a location/year combination. The unique combination of Location and HarvestDate is used as a site level.
- You can specifiy a randomized block design per location, with multiple measurements per plot by using LocationID, BlockID and PlotID, but also completely randomized designs with multiple measurements per plot.
- There is some flexibility in the name of the headings. E.g. you can specifiy Site or SiteID, and so on.