rTLC manual

Version : 11/11/2015

Table of Contents

Introduction

rTLC is an application made to facilitate the explotation of HPTLC pictures and specially statistics exploitation.

Differents features are available :
- Chromatograms extraction from pictures
- Chromatograms preprocessing
- Variables selection
- Exploratory statistics
    - PCA
    - Cluster
    - Heatmap
- Predictive statistics
    - Model training
        - Parameters tunning
        - Cross validation
        - Regression/classification
    - New data prediction
- Report Output

The application could be found at this url :
http://shinyapps.ernaehrung.uni-giessen.de/rtlc/

Data Input

Chromatograms extraction

Demonstration Data

In the tab Data Input, select one of the demo file in the *data to use* menu on the left (Figure 1-1).

A picture should appear on the page (Figure 1-2), as well as a *Plate choice* menu (Figure 1-3) and a table named *Horizontale Dimension* (Figure 1-4).

Horizontale Dimensions

A chromatogram will be extracted between each pair of red and green vertical line on the central picture by taking the horizontal mean of pixel on each channel red, green and blue of the picture (Figure 1-2).

You must modify the number in the Horizontale Dimension table in order to match each band of your picture between a pair of red and green line.

If the dimensions are available from the manipulation AND there wasn't **unecessary cropping** of the pictures, this step should be straighforward (Be carreful with the frame option in the default setting of winCats, the default option is to remove 2 mm of the plate so you must adapt the dimension or put this option to 0 mm).

It is possible to choose 2 conventions for those dimensions : *LINOMAT* or *ATS*, depending if the dimensions are considered from the outside of the band (LINOMAT) or from the center (ATS) (Figure 1-3).

The *tolerance* dimension is here to control the zone of the band you want to extract, a value of 0 will extract all the band where a bigger value will help to take only the center of the band.

This operation must be done for each plate of the study, the picture could be choose in the *Plate choice* menu on top of the picture (Figure 1-4). If your study contains 3 plate, there will be 3 choices in the dropbox menu and therefore, 3 rows in the *Dimension Table.*

Verticale Dimensions

The verticale dimension table is here to calculate the Retention Front value and redimension the number of pixels of the picture and therefore the number of point in the chromatogram (Figure 1-7).

Batch Table

Visit the batch table to visualise the batch (Figure 1-8).

You can edit this table and choose to remove some samples from the study (standard or outlier for example).

Chromatograms/Band Comparaison/Chromatograms Comparaison

In this three tab, you can visualize the chromatograms extracted from the plates (Figure 1-9).

Figure 1: Data Input. Demo file

Your Own Data

Now in the tab Data Input, choose to use *Your Own Data*.

There is two part :
- the independent variables : plate pictures with the band
- the dependent variables : batch file (in excel) with information on each band

You can upload your(s) plate(s) in the *browse* that appears on the left (Figure 2-1).

Proceed to the extraction like for the demonstration data

For the batch, there is two choice, it's possible to upload an excel file on the left side of the page (Figure 2-2) or it's possible to edit directly the batch file in the *batch* tab, the number of rows will correspond to the number of extracted chromatograms (Figure 2-3).

In case a excel file is uploaded, it must contain the information in the first sheet and have this kind of format :

Id	Colnames 1	Colnames 2	Colnames 3	...
1	...	...	...	...
2	...	...	...	...
3	...	...	...	...
...	...	...	...	...

The first row must be the name of the columns with at least 4 columns, the first column will be the id column used for the application to track the samples. There must be the same number of rows (without the first one) than chromatograms extracted.

In case one of the constraint is not respected, a green message will appear showing the user what is the problem.

This kind of format is called tidy data in the field of Data sciences and it's a good practice to adopt this kind of format for data collection : 1 observation is the association of independent variables and dependent variables, therefore, taking tracks of the data in a consistent way from the beginning of a study (or even multiple study) can save a lot of time later in the phase of data analysis.

Save the data extracted

In order to avoid the step of chromatogram extraction for a futur session, it's possible to save a file containing the chromatograms and the batch table with the *Save Chromatograms* button on the left of the page (Figure 2-9).

In an other session, choose to use *Saved Data* in the tab *Data Input*. And upload the file saved precedently in the browse button (Figure 3).

Save csv file for each channels

To export the chromatograms to an other software for further exploitation, it's possible to save each channel as a CSV file with observation as row and R_F as column. The files use ' ;' as separator. The download buttons are in the left part of the page (Figure 2-10).

Figure 2. Data Input. Your own data

Figure 3: Data Input. Saved Data

Data preprocessing

This tab allow different preprocessing in order to prepare the data for further analysis.

Preprocess Order

In the left side of the page, choose the order the preprocessing should appear (Figure 4-1). Available preprocessing are :

       -Smoothing : Savitzky-Golay transformation

       -Warping : Peak alignment (experimental)

       -Baseline correction

       -Scalling

       -Standart Normal Variate

       -Mean centering

Preprocess Details

For each preprocessing, a set of options are available, in each case, a link lead to an exhaustive explanation of the features (Figure 4-2).

Chromatograms/ Chromatograms Comparaison

In this two tab, you can visualize the results of the preprocessing (Figure 4-3).

Figure 4. Data Preprocessing

Variable Selection

This tab allow for variable selection in order to choose a channel or part of a channel. There is 12 possibility to choose a channel, a range and to include or not this range in the study (Figure 5-1). After this step, all selected data are combine into one data set that will be used for statistical study. The two plots on the left should help the user to understand the feature (Figure 5-2).
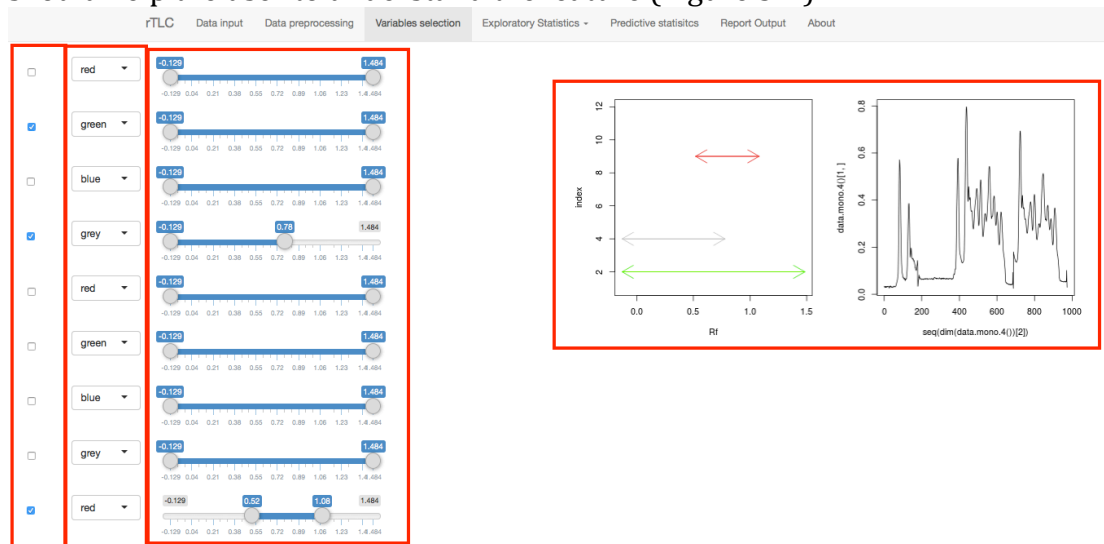

Figure 5. Variable Selection

Exploratory Statistics

PCA

This feature allow to perform Principal Componant Analysis on the dataset.

PCA tab

The principale plot is the score plot, a few options are available :
-choose the color/shape/label of the point according to one of the variable of the batch (Figure 6-1)
-choose the componant to plot (Figure 6-2)
-choose to calculate the ellipse and to plot it (Figure 6-3)
-Some aestetics parameters (Figure 6-5) :
>-hjust and vjust move the label on the plot
>-point size scale the point size on the plot

-Title of the plot (Figure 6-6)

On this page, there is also a table of the batch and the first 4 componant of the analysis and a Summary of the model which show the the cummulative variables of the first 5 and the 10th componant.

Loading

This tab show the loading plot of the PCA (Figure 6-7).
It's possible to choose the componant to study and to plot or not the minimum and maximum point on the graph according to a *neighbourhood* value. The resulting maximum and minimum value are shown in the field bellow.

Outlier

This tab is for outlier detection, i.e. point that should be removed because they are too different from the dataset.
It's possible to choose the number of componant of the PCA to include in the test and the quantile to use for the cutoff The Mahalanobis distance is used and the classical and robust test are calculated (Figure 6-3).
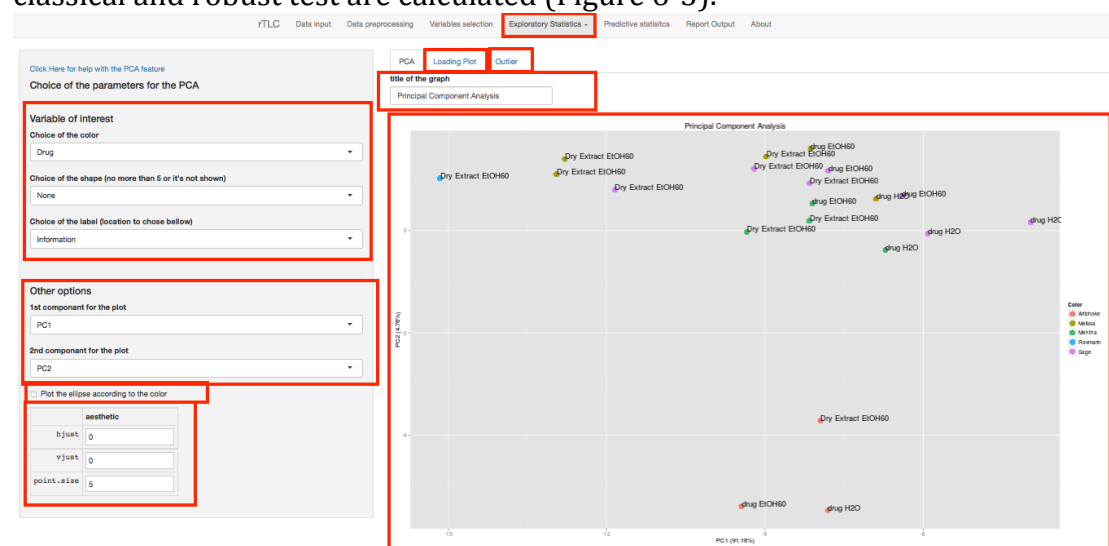
Figure 6. Exploratory statistics. PCA

Cluster

This feature allow to perform cluster analysis on the dataset.

The options available are :

-Choice of the variable of interest in the batch (Figure 7-1).

-Choice of the distance method (Euclidean is the more used) (Figure 7-2).

-Choice of the cluster method (Figure 7-3).

-Number of cluster to cut in the tree (Figure 7-4).
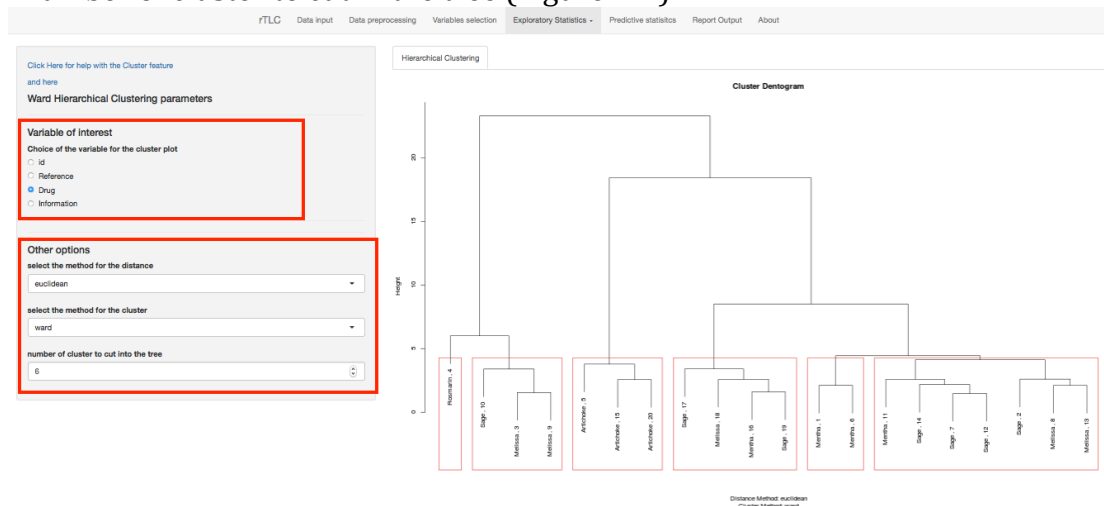

Figure 7. Exploratory Statistics. Cluster

Heatmap

This feature allow to perform and visualise heatmap, choose the variable of interest and visualise the result either with the normal heatmap or with the interactive heatmap.

Predictive Statistics

This tab allow you to train a predictive model for classification or regression.

Options

Train/Test split

In a first time, the data set should be split in two, the test set and the train set. The train set will be use to train the data and the test set will be use to verify the result of the training on an independant part of the dataset (Figure 8-1).

Classification/Regression

Depending of the problem, one option should be choose in order to train the system on the good type of data. 3 choices are available :
Classification 2 classes, Classification multiclass and Regression.
(Figure 8-2).

Choice of the variable of interest

Choose the variable to be train with from the batch. What should be predicted. It must be in accordance with the Classification Regression choices otherwise an error will be return (Figure 8-3).

Algorythm

Choose which machine learning algorythm should be used, some of them are only available for classification or for regression (Figure 8-4).
Only a subset of available algorithm is available, other could be add, just contact us. The list of all model available could be found here :
http://topepo.github.io/caret/modelList.html

Tuning Options

The training will try every combinaison of every parameters of the grid in order to optimise the performance of the model and choose the better parameters.

Validation

-Validation method (Figure 8-5):
>   -bootstrap
>   -reapeated cross validation
>   -leave one out cross validation

-Summary metrics (Figure 8-6) :
>   Which summary metrics to use for the tunning

-Cross validation k-fold or resampling iterations (Figure 8-7):
>   Number of k-fold or resampling

-Number fo repeat (k-fold only) (Figure 8-8):
>   Number of time to repeat the process of validation

Grid

This area contains the tuning length, i.e. the maximum number of parameters to test a each parameters. It is also possible to choose the different parameters manually in the Grid table for fine tuning (Figure 8-9).

Launch

Once all the options are choosen, press the *Train* button to launch the analysis, note that you must visit an other tab to really launch the anaysis (Figure 8-10).

Results

Confusion matrix

This tab is used for classification problem, the confusion matrix of the true result vs the predicted results are shown (Figure 8-11).
It's possible to choose to show only the result for the training data, for the test data or for both.
More metrics are also produce to explore more closely the comparaison like accuracy, specificity and selectivity.
Note that for 2 class problem, one of the class is choose as positive and the other as negative to calcul specificity and selectivity.
For multiclass problem, this calcul is made for each class as one against all.

Prediction table

This tab show the prediction table for all data, it's possible to filter according to the use in the training set or not, to the prediction class etc… (Figure 8-12)

Algorithm information

This tab give more information about the algorithm used during the training, in particular, what are the tunning parameters (Figure 8-13).

Model Summary

This tab summarise important information of the tunning, it's possible to extract the information for each row of the tunning grid ad for each metrics.
Also important information describe how the tunning took place (Figure 8-14).

Tunning Curve

This tab show the evolution of the metric choose for the tunning dependending of each parameters of the algorithm (Figure 8-15).

Figure 8. Predictive Statistics

Model Download and New data prediction

Once the good model with the good preprocessing, the good variable selection, the good tunning parameters is made. It's possible to download a file that could be then upload at the beginning of the process (Figure 8-16).
In the first tab Data Input, choose to use *Predicted data – QC* (Figure 9-1).

Upload the batch and picture file as previously and also a model file created in an other session (Figure 9-2).

Proceed to the chromatograms extraction with the dimension table and visit the tab *Prediction (QC only).* The prediction for each chromatograms should appear (Figure 9-3).

Figure 9. Predict new data

Report output

In this tab, it's possible to download a pdf report, choose the content of this document and push the *Download the report* button (Figure 10).

Figure 10. Report Ouput