

rTLC manual

Dimitri Fichou

07.10.2016

Introduction

rTLC is a web application for image processing and multivariate analysis of HPTLC chromatograms.

Different features are available:

- Chromatograms extraction from pictures
- Chromatograms preprocessing
- Variables selection
- Exploratory statistics
- PCA
- Cluster
- Heatmap
- Predictive statistics
- Model training
- Parameters tuning
- Cross validation
- Regression/classification
- New data prediction *Report output

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

Analytical pipeline

Data input

Chromatogram extraction

Demonstration data In the tab *Data Input*, select one of the demo files in the *Data to use* menu on the left (Figure 3). A picture should appear on the page, as well as a *Plate choice* menu and a table named *Horizontal dimension*.

Horizontal dimensions A chromatogram will be extracted between each pair of red and green vertical lines on the central image by taking the horizontal mean of pixels on each of the red, green and blue channels of the chromatogram. The gray scale is then calculated from the 3 channels at each Rf and for each track.

The number in the *Horizontal dimension* table must be modified in order to match each band of the chromatogram between a pair of red and green lines.

If the dimensions are available from the manipulation AND there wasn't **unnecessary cropping** of the image, this step should be straightforward. About the cropping, it is good practice to upload the totality of the images without a cropping that could be difficult to reproduce on other data in the future. It is possible to choose two conventions for those dimensions, *i.e.* calculation from the center to the band or from the exterior. The *Edge cut* parameter is here to control the zone of the band to extract, a value of 0 will extract all the band, whereas a bigger value will help to take only the center of the band. This operation must be

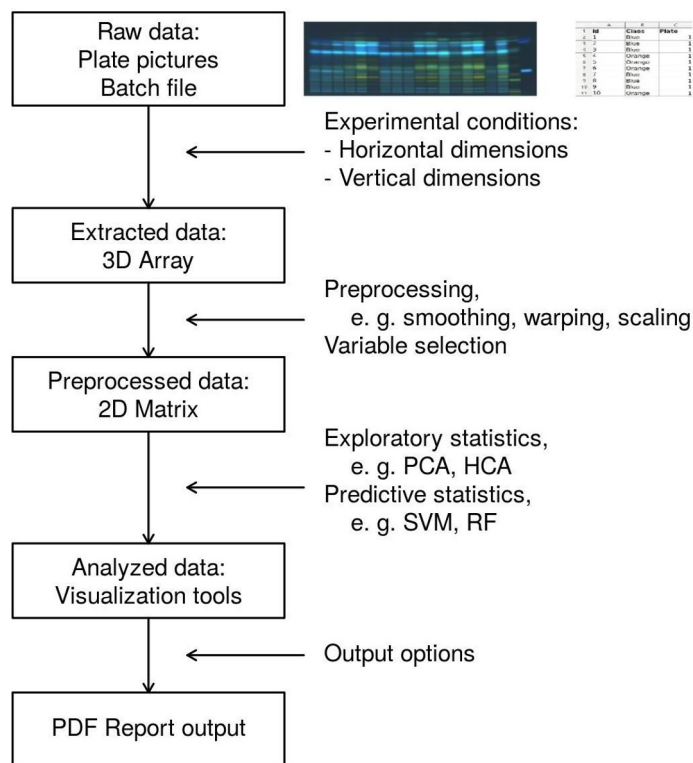


Figure 1: Analytical pipeline

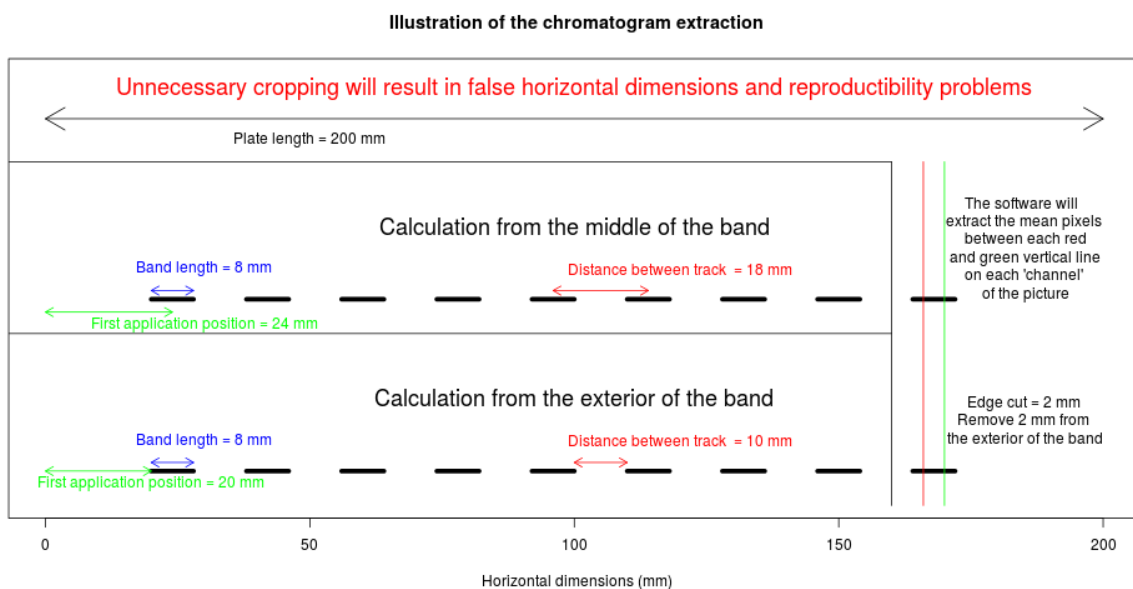


Figure 2: Illustration of the chromatograms extraction

rTLC V.1.0 Data input Data preprocessing Variables selection Exploratory statistics Predictive statistics Report output About/help

Data to use

☐ Your own data

☒ demo 1: Medicinal plants, 20 samples

☐ demo 2: Propolis dataset

☐ Saved data

☐ Predict data - QC

Filename

rTLC_checkpoint_1

Save Chromatograms

Filename

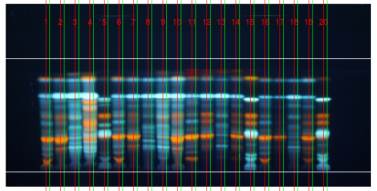
rTLC_zip_export

Save zip file with csv

Chromatogram extraction **Batch** Track plot Chromatogram comparison Densitogram comparison Image reconstruction Prediction (QC)

Plate choice

1 - rTLC_demopicture.JPG



Vertical dimensions (mm)

Pixel width: 128

Plate width: 100

Migration front: 70

Distance to lower edge: 8

Horizontal dimensions (mm)

	Plate length	First application position	Band length	Distance between track	Edge cut
1	200	23	6	8	2

Convention how to use the horizontal table

☐ Calculation from the exterior of the band

☒ Calculation from the middle of the band

Filename

TableDimensionSave

Save the Dimension table

Upload the saved table

Browse... No file selected

Figure 3: Data input

done for each plate of the study, the picture could be chosen in the *Plate choice* menu on top of the central image. If the study contains three plates, there will be 3 choices in the drop-box menu and therefore, 3 rows in the *Dimension table*. It's possible to save a dimension table as an excel file to use later, for example with the same study but with pictures under a different light.

Vertical dimensions The *Vertical dimension* table is here to indicate the *Migration front* value, the *Plate width*, the *Distance to lower edge* as well as the *Pixel width*. Those values allow the software to redimension the chromatograms and attribute a R_F for each pixel.

Chromatogram extraction **Batch** Track plot Chromatogram comparison Densitogram comparison Image reconstruction Prediction (QC)

Information to include in the track plot

☒ Reference ☒ Drug ☒ Information

Column filter: Keep only selected, if none, keep all.

Reference

Drug

Information

Exclude	ID	Reference	Drug	Information
<input type="checkbox"/>	1.00	T411	Mentha	Dry Extract EtOH60
<input type="checkbox"/>	2.00	B247	Sage	Dry Extract EtOH60
<input type="checkbox"/>	3.00	H329	Melissa	Dry Extract EtOH60
<input type="checkbox"/>	4.00	C244	Rosmarin	Dry Extract EtOH60
<input type="checkbox"/>	5.00	R411	Artichoke	Dry Extract EtOH60
<input type="checkbox"/>	6.00	N712	Mentha	Dry Extract EtOH60
<input type="checkbox"/>	7.00	X147	Sage	Dry Extract EtOH60

Figure 4: Batch tab

Batch table Visit the *Batch table* to visualize the batch. The table is editable and the *Exclude* option allow to exclude samples, outliers or standard for example. The checkbox on the left concern the informations

of the batch that should be passed to the *Track plot* title. Finally, the *Column filter* allow to exclude bigger part of the data set.

Track plot, Chromatogram comparison, Densitogram comparison Those three tabs allow to visualize the extracted chromatograms.

Your own data Now in the tab *Data input*, choose to use *Your own data*. There are two parts:

- 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. Proceed to the extraction like for the demonstration data. For the batch, there are two choices, it's possible to upload an excel file on the left side of the page 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. In case a excel file is uploaded, a few rules must be observed:

The first row must be the name of the columns There must be the same number of rows (without the first one) as chromatograms extracted.

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

Save the data extracted In order to avoid the step of chromatogram extraction for a future 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. In another session, choose to use *Saved data* in the tab *Data input*. And upload the file saved precedently in the browse button.

Save csv file for each channel To export the chromatograms to another software for further exploitation, it's possible to save each channel as a CSV file with observation as row and RF as column. The files use “;” as separator. The download buttons are on the left part of the page.

Data preprocessing

This tab allows 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. Available preprocessing are:

- Smoothing: Savitzky-Golay transformation
- Warping: Peak alignment (experimental)
- Baseline correction
- Scaling
- Standard Normal Variate
- Mean centering

Preprocess details For each preprocessing, a set of options are available, in each case, a link leads to an exhaustive explanation of the features.

Chromatogram comparison, Densitogram comparison In these two tabs, you can visualize the results of the preprocessing.

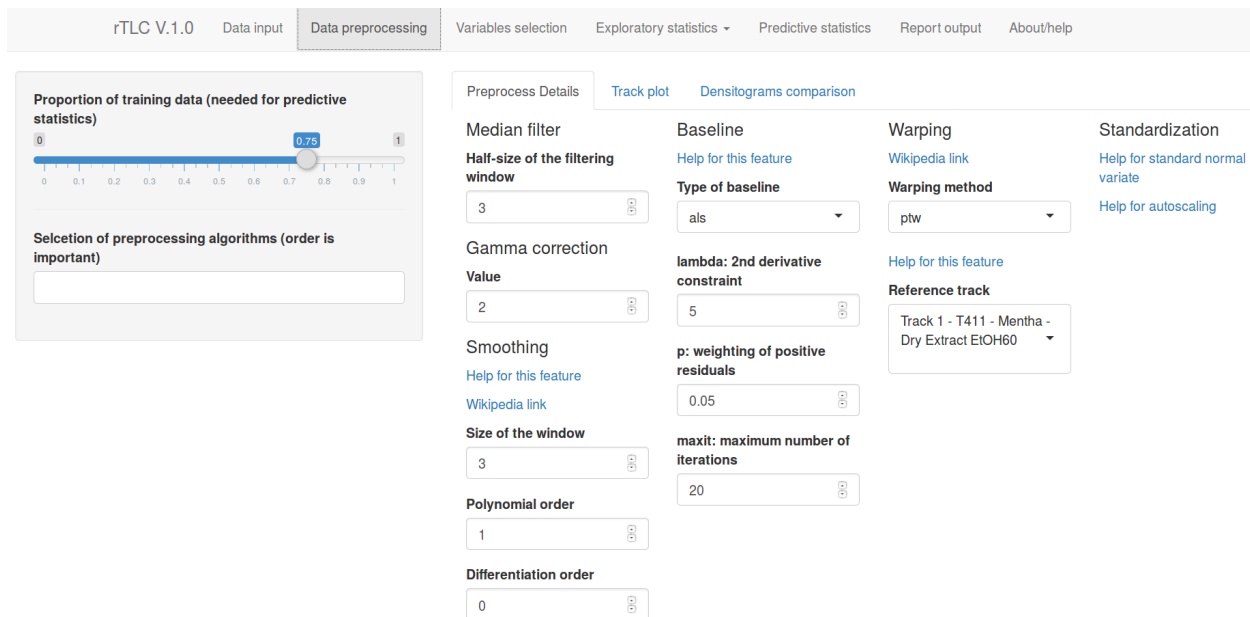


Figure 5: Data preprocessing

Variables selection

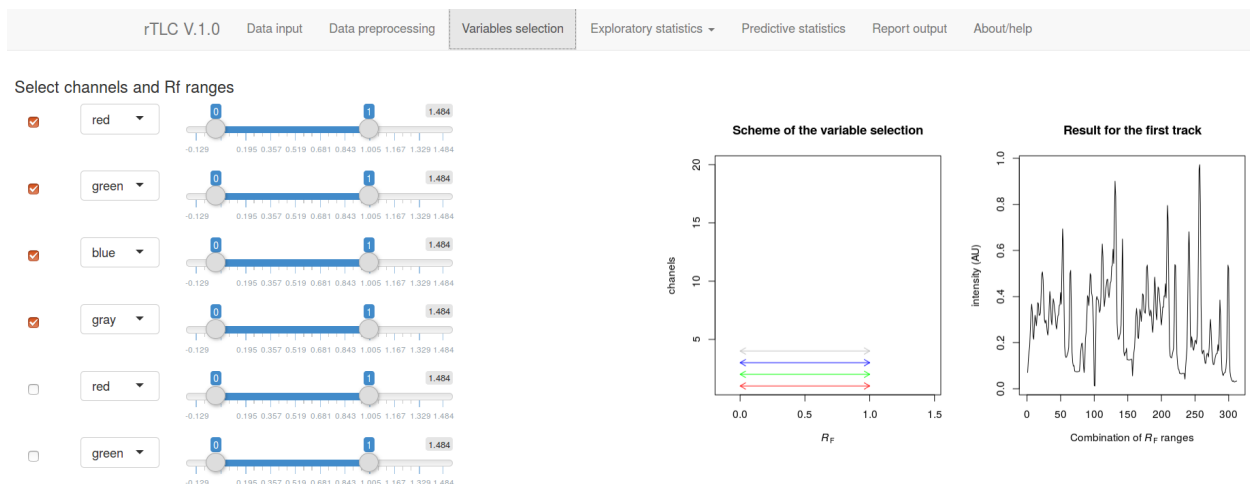


Figure 6: Variables selection

This tab allows for variable selection in order to choose a channel or part of a channel. There are 20 possibilities to choose a channel, a range and to include or not this range in the study. After this step, all selected data are combined 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.

Exploratory Statistics

PCA

This feature allows to perform PCA on the dataset.

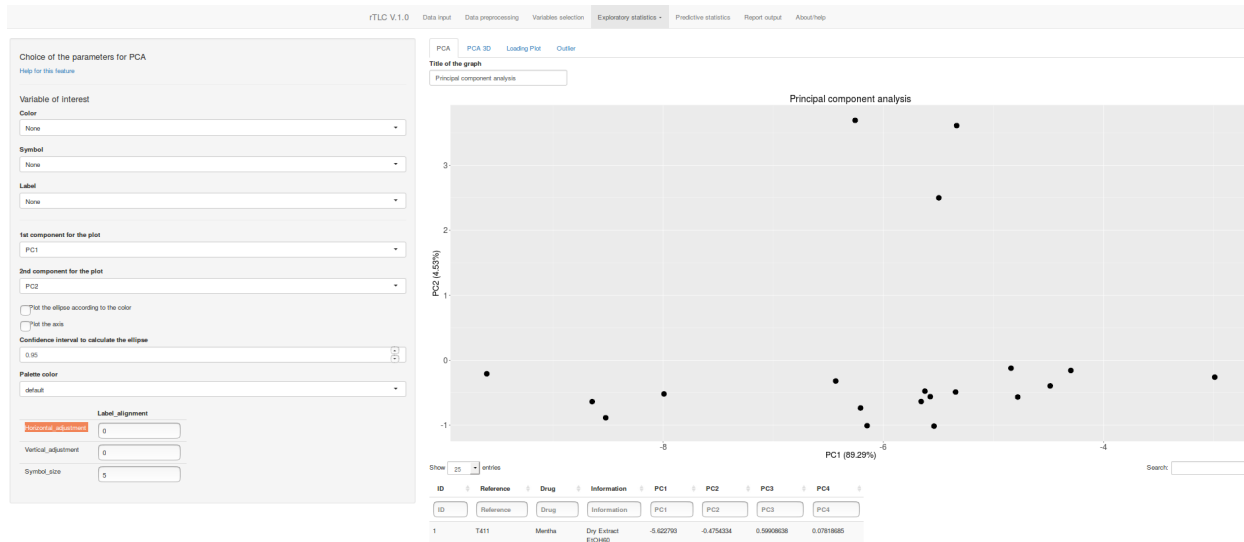


Figure 7: PCA

PCA tab The principal 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
- Choose the component to plot
- Choose to calculate the ellipse and to plot it
- Choose the color palette to use in the plot
- A few aesthetics parameters :
 - *Horizontal adjustment* and *Verticale adjustment* move the label on the plot
 - *Point size* scales the point size on the plot
- Title of the plot

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

Loading Plot This tab shows the loading plot of the PCA. It's possible to choose the component to study and to plot or not the minimum and maximum point on the graph according to the *Span of local minima and maxima*. The resulting maximum and minimum values are shown in the field below.

Outlier This tab is for outlier detection, i.e. points that should be removed because they are too different from the dataset. It's possible to choose the number of components 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 tests are calculated.

Cluster

This feature allows to perform cluster analysis on the dataset. The options available are:

- Choice of the variable of interest in the batch.
- Choice of the distance method (Euclidean is the most common).
- Choice of the cluster method (Ward is the most common).
- Number of clusters to cut the tree in.

Heatmap

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

Predictive statistics

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

The screenshot shows the 'Predictive statistics' tab in the rTLC V.1.0 software. The interface is divided into several sections:

- Choice of the algorithm:** A dropdown menu set to 'Random Forest'.
- Choice of the variable:** Radio buttons for 'Reference' (selected), 'Drug', and 'Information'.
- Type:** Radio buttons for 'Classification' (selected) and 'Regression'.
- Train:** A button to initiate training.
- Filename:** A text input field containing 'rfStandard.Normal.Variate'.
- Download model:** A button with a download icon.
- Cross validation:** A section with a 'Wikipedia link', a 'Help for this feature' link, a 'Validation method for tuning' dropdown set to 'repeatedcv', a 'Performance metric' dropdown set to 'Accuracy', a 'Number of folds or resampling iterations' spinner set to 5, and a 'For repeated k-fold cross-validation: number of complete sets of folds' spinner set to 1.
- Grid:** A section with a 'Tuning length' spinner set to 10 and a list of 'mtry' values (2, 10, 18, 27, 35, 44, 52, 61, 69, 78) each with a corresponding spinner.

Figure 8: Predictive statistic

Training/Test split

In a first time, the data set should be split in two, the test set and the training set. The training set will be used to train the data and the test set will be used to verify the result of the training on an independent part of the dataset. This option is present in the *Preprocessing* tab as the split is applied before the preprocessing.

Classification/Regression

Depending on the problem, one option should be chosen in order to train the system on the good type of data.

Choice of the variable of interest

Choose the variable to be trained with from the batch. What should be predicted. It must be in accordance with the Classification/Regression choices, otherwise an error will be returned, for example if regression is asked on non-numeric data.

Algorithm

Choose which machine learning algorithm should be used, some of them are only available for classification or for regression. Only a subset of available algorithms is available, others could be added, just contact us. The list of all models available could be found here: <http://topepo.github.io/caret/modelList.html>

Tuning options

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

Cross validation

- Validation method:
 - Bootstrap
 - Repeated cross validation
 - Leave one out cross validation
- Summary metrics: Which summary metrics to use for the tuning
- Cross validation k-fold or resampling iterations: Number of k-fold or resampling
- Number to repeat (k-fold only): Number of times to repeat the validation process

Grid

This area contains the tuning length, *i.e.* the maximum number of parameters to test on each parameters. It is also possible to choose the different parameters manually in the Grid table for fine tuning.

Training itself

Once all the options are chosen, press the *Train* button to launch the analysis, note that you must visit another tab to really launch the analysis.

Validation Metrics

This tab is used to verify the performance of the model, a confusion matrix is shown for the classification problem and a plot of predicted values against the real value is shown for the regression problem. It's possible to choose to visualize the result for the Test data, the Training data and the cross-validation data, *i.e.* the data used during the optimization phase of the training.

Prediction table

This tab shows 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. . .

Algorithm information

This tab gives more information about the algorithm used during the training, in particular, what are the tuning parameters.

Model Summary

This tab summarizes important information of the tuning, it's possible to extract the information for each row of the tuning grid and for each of the metrics. Also important information describes how the tuning took place.

Tuning Curve

This tab shows the evolution of the metric chosen for the tuning depending on each parameter of the algorithm.

Model Download and New data prediction Once the good model with the good preprocessing, the good variable selection, the good tuning parameters is made. It's possible to download a file that could be then uploaded at the beginning of the process. In the first tab *Data Input*, choose to use *Predicted data – QC*. Upload the batch and picture file as previously and also a model file created in another session. Proceed to the chromatograms extraction with the dimension table and visit the tab *Prediction (QC)*. The prediction for each chromatogram should appear.

Report output

In this tab, it's possible to download a report, choose the content of this document as well as the format. It is also possible in the right side to download the PCA data to use them in other software.