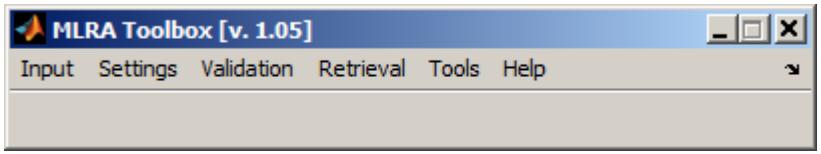
Machine Learning Regression Algorithms (MLRA) toolbox - v.1.05

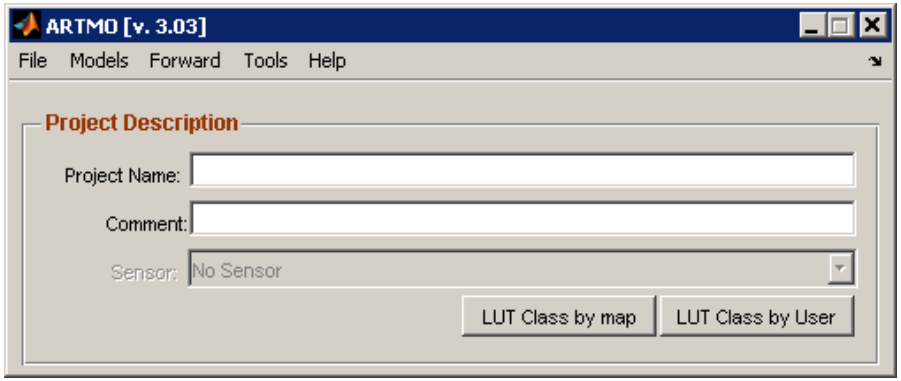
Tutorial: Training/Validating & Mapping with User data

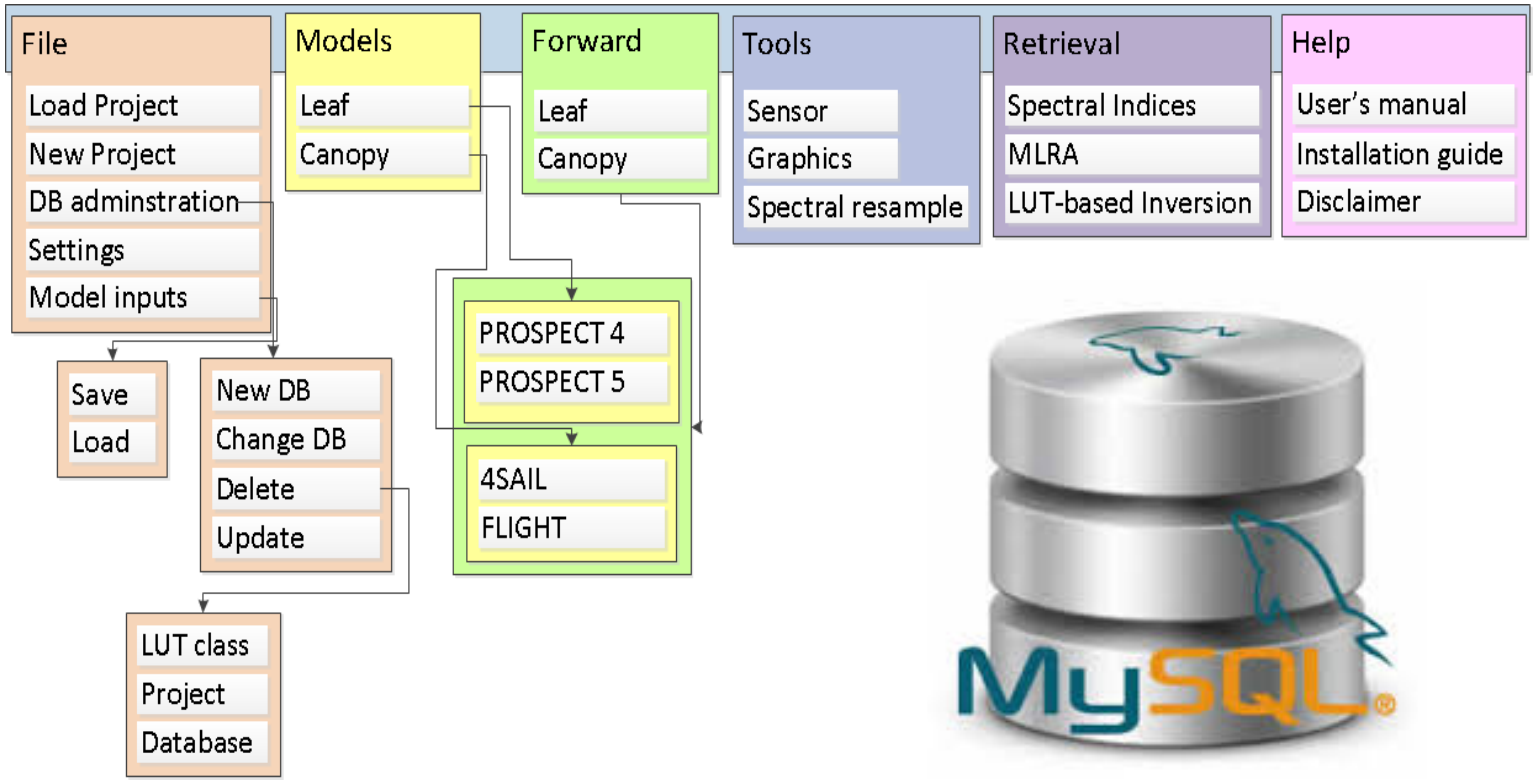


Jochem Verrelst
Jochem.verrelst@uv.es - http://ipl.uv.es/artmo/

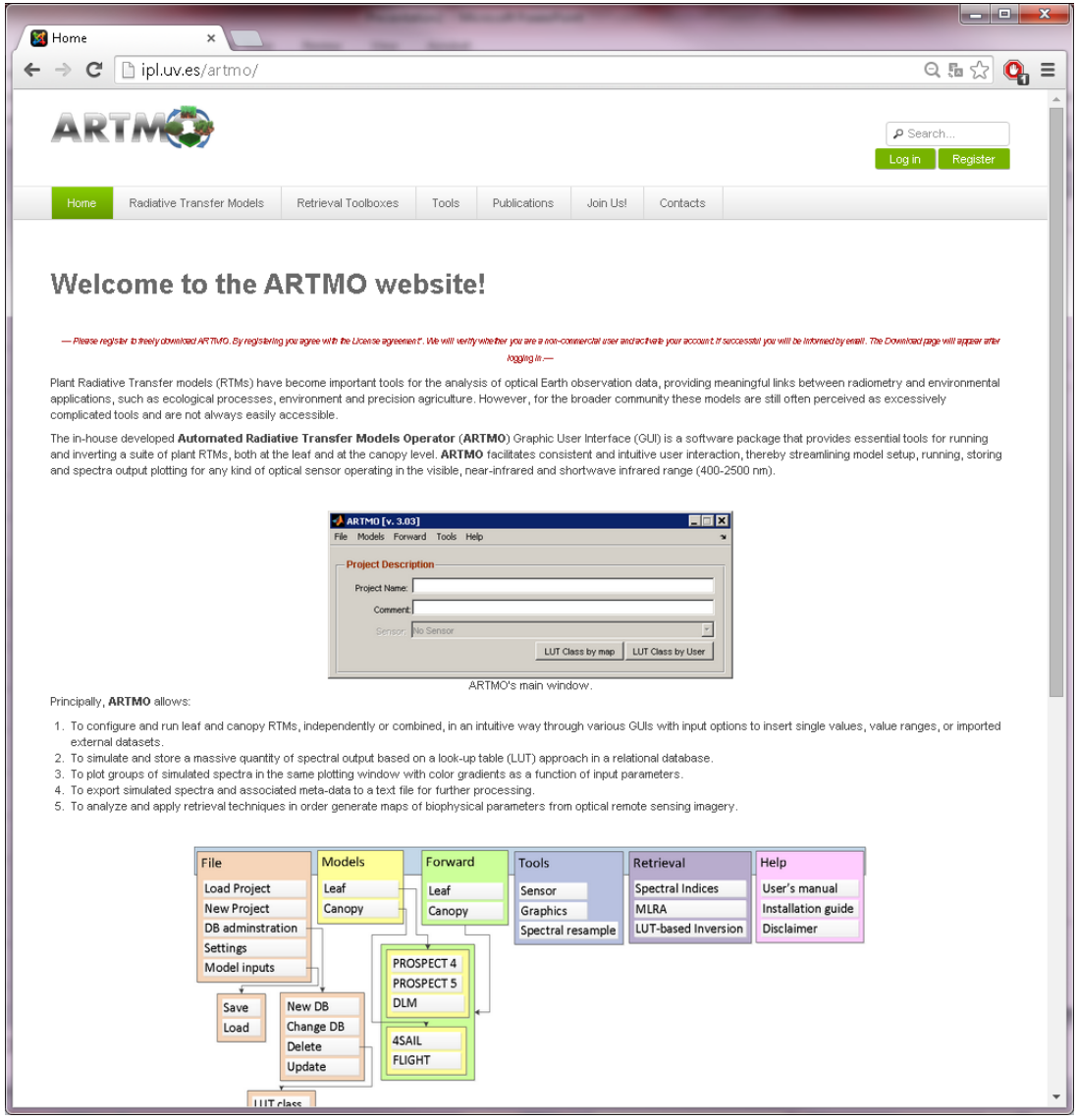
8-12 September 2014 – University of Valencia – Valencia, Spain

ARTMO v3.03


ARTMO [v. 3.03]
File
Models
Forward
Tools
Help
Project Description
Project Name:
Comment:
Sensor:
No Sensor
LUT Class by map
LUT Class by User


File
Load Project
New Project
DB adminstration
Settings
Model inputs
Save
Load
New DB
Change DB
Delete
Update
LUT class
Project
Database
Models
Leaf
Canopy
Forward
Leaf
Canopy
PROSPECT 4
PROSPECT 5
4SAIL
FLIGHT
Tools
Sensor
Graphics
Spectral resample
Retrieval
Spectral Indices
MLRA
LUT-based Inversion
Help
User's manual
Installation guide
Disclaimer
MySQL

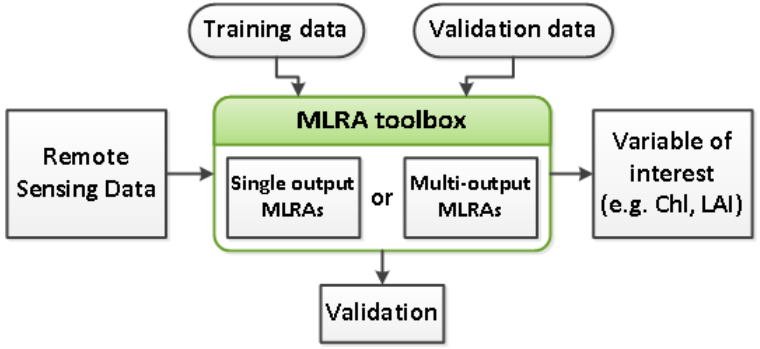
Website: http://ipl.uv.es/artmo/

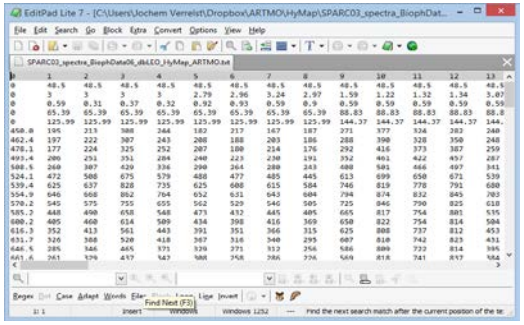


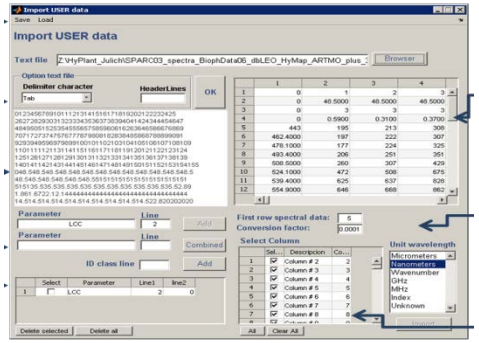
Outlook:

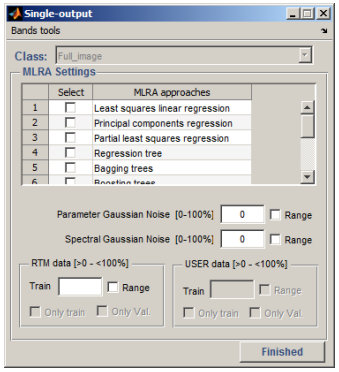
1. MLRA mapping based on User data

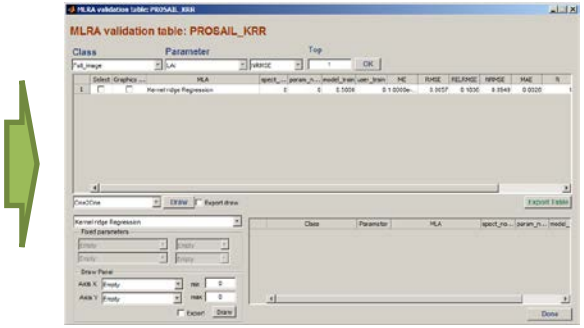
- User Input data
- MLRA setting
- Validation
- Mapping

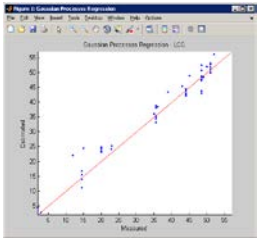


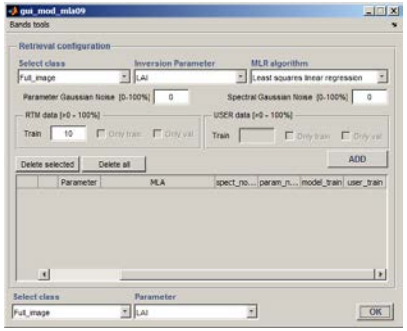


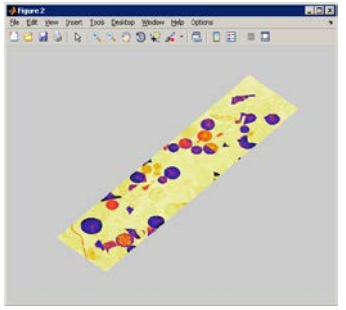




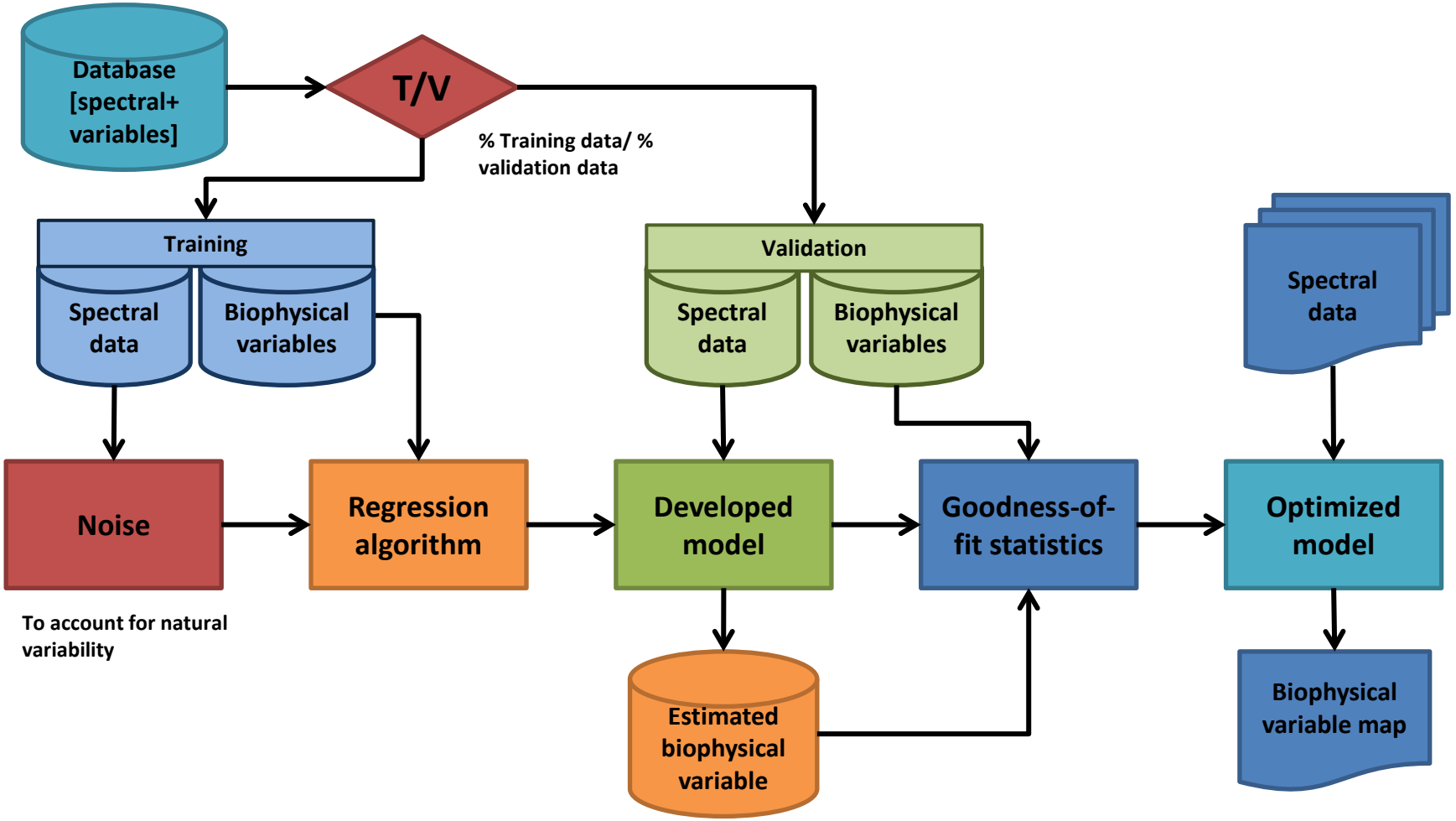








Schematic overview for systematic evaluation of nonparametric regression models to estimate biophysical variables

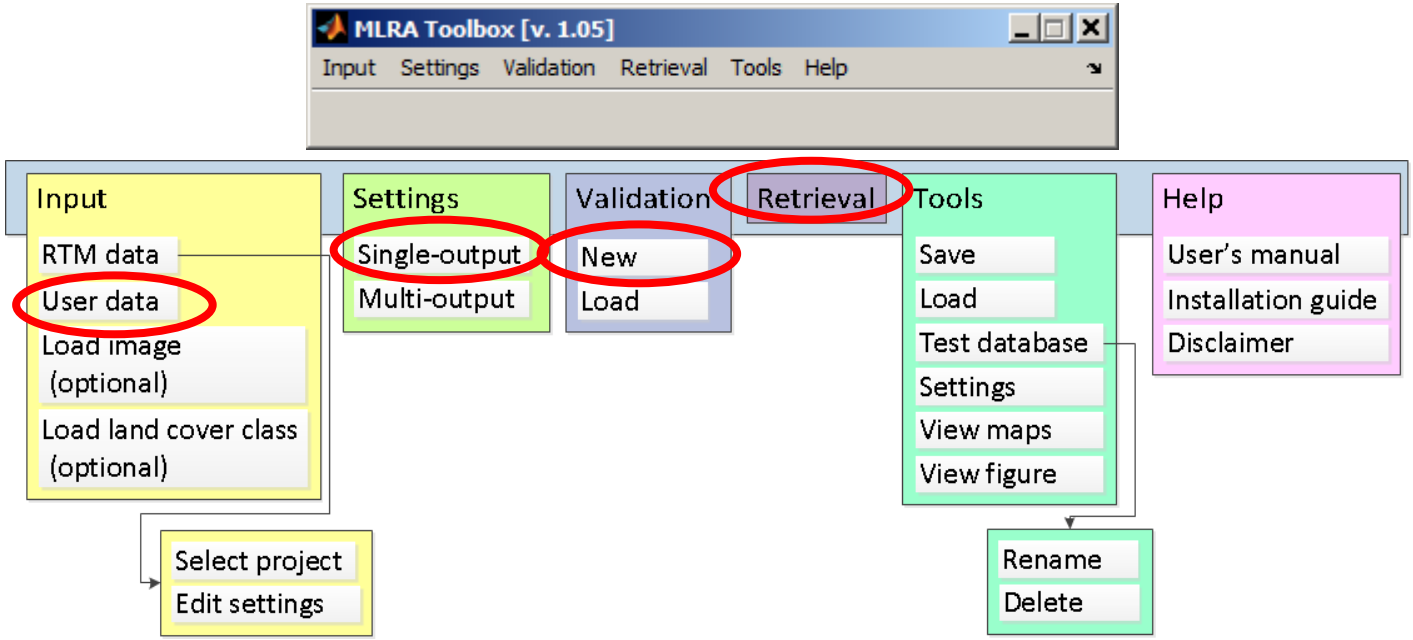


MLRA mapping based on User data

*The **example** will demonstrate how to train and validate single-output MLRA models and apply it to an image using a field dataset for training. Consult the **Manual** for more details.*

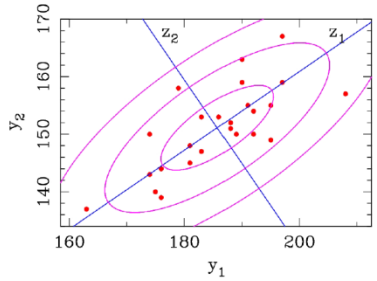
The procedure will be as follows:

1. ***User data:** Insert field data for training and validation*
2. ***Single-output:** Choose single-output MLRA models and define training/testing partitioning*
3. ***Validation:** Validate the defined MLRA strategies*
4. ***Retrieval:** Apply the best one to a remote sensing image.*

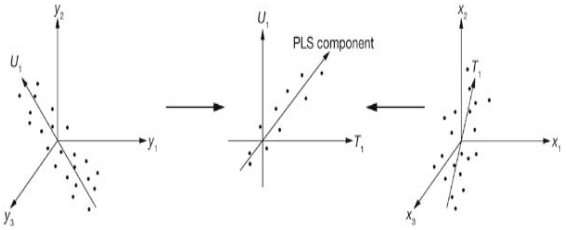


- # Non-parametric models
  - **SimpleR** [Camps-Valls et al., 2013]
  - http://www.uv.es/gcamps/code/simpleR.html

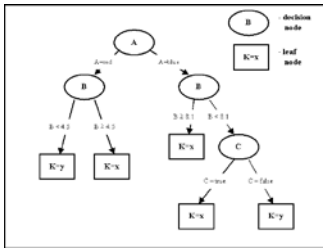
Principal component regression– PCR



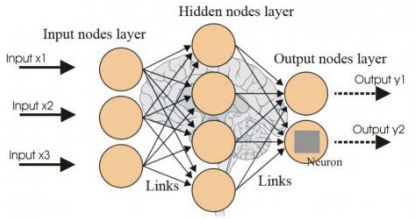
Partial least squares regression– PLSR



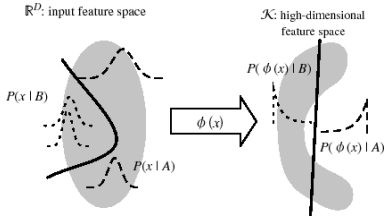
Decision Trees – DT



Neural networks – NN

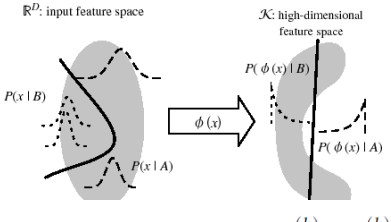


Kernel ridge regression – KRR



$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2)).$$

Gaussian processes regression - GPR



$$K(\mathbf{x}_i, \mathbf{x}_j) = \nu \exp\left(-\sum_{b=1}^{B} \frac{(\mathbf{x}_i^{(b)} - \mathbf{x}_j^{(b)})^2}{2\sigma_b^2}\right)$$

Also:

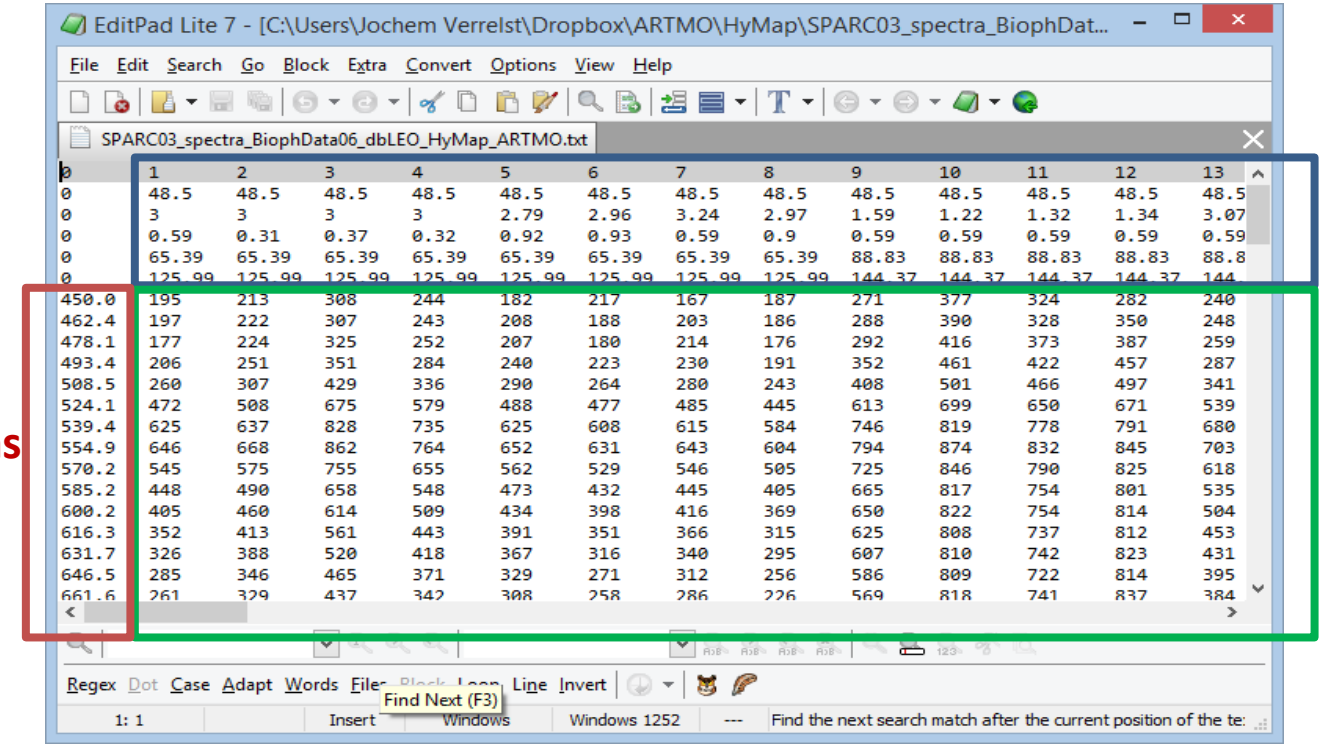
- Elastic Net (ELASTICNET)
- Bagging trees (BAGTREE)
- Boosting trees (BOOST)
- Neural networks (NN)
- Extreme Learning Machines (ELM)
- Support Vector Regression (SVR)
- Relevance Vector Machine (RVM)
- Variational Heteroscedastic Gaussian Process Regression (VHGPR)

Input: User data (e.g. field data)

User data for training and validation requires one input file, including:

1. Biophysical parameters (e.g., LAI, chlorophyll content,…)
2. Associated spectra (e.g., obtained from a remote sensing image)

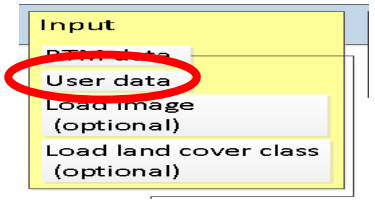
User data need to be organized in a matrix format in plain text file, according to example below:

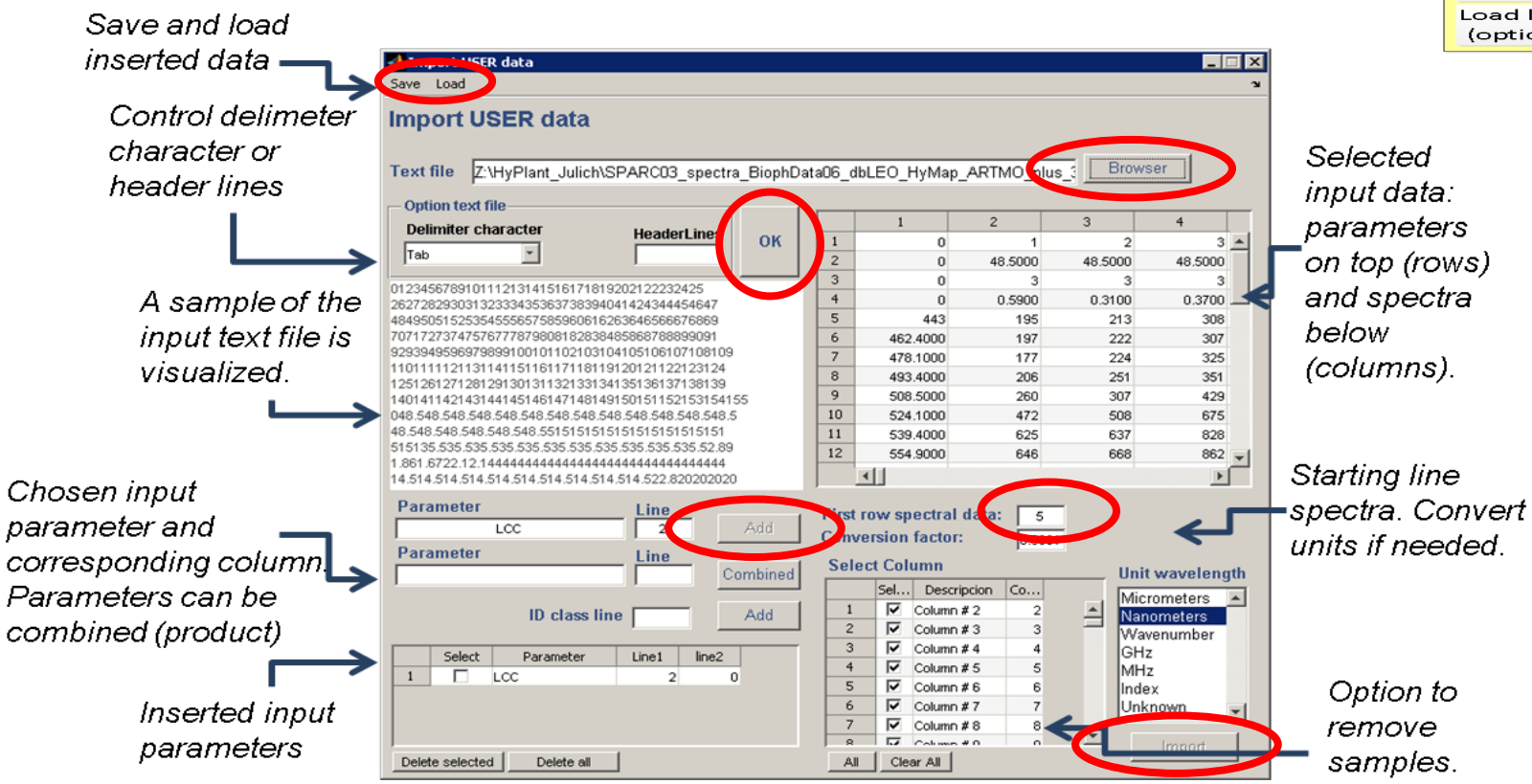


0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5
0	3	3	3	3	2.79	2.96	3.24	2.97	1.59	1.22	1.32	1.34	3.07
0	0.59	0.31	0.37	0.32	0.92	0.93	0.59	0.9	0.59	0.59	0.59	0.59	0.59
0	65.39	65.39	65.39	65.39	65.39	65.39	65.39	65.39	88.83	88.83	88.83	88.83	88.8
0	125.99	125.99	125.99	125.99	125.99	125.99	125.99	125.99	144.37	144.37	144.37	144.37	144
450.0	195	213	308	244	182	217	167	187	271	377	324	282	240
462.4	197	222	307	243	208	188	203	186	288	390	328	350	248
478.1	177	224	325	252	207	180	214	176	292	416	373	387	259
493.4	206	251	351	284	240	223	230	191	352	461	422	457	287
508.5	260	307	429	336	290	264	280	243	408	501	466	497	341
524.1	472	508	675	579	488	477	485	445	613	699	650	671	539
539.4	625	637	828	735	625	608	615	584	746	819	778	791	680
554.9	646	668	862	764	652	631	643	604	794	874	832	845	703
570.2	545	575	755	655	562	529	546	505	725	846	790	825	618
585.2	448	490	658	548	473	432	445	405	665	817	754	801	535
600.2	405	460	614	509	434	398	416	369	650	822	754	814	504
616.3	352	413	561	443	391	351	366	315	625	808	737	812	453
631.7	326	388	520	418	367	316	340	295	607	810	742	823	431
646.5	285	346	465	371	329	271	312	256	586	809	722	814	395
661.6	261	329	437	342	308	258	286	226	569	818	741	837	384

- Make sure to fill up the whole Matrix! In case of empty cells, use **NaN** and remove those samples in the following step.
- Make sure that wavelengths are the same as the remote sensing image! They need to match. A band selection or band transformation can be later done in **Settings**.

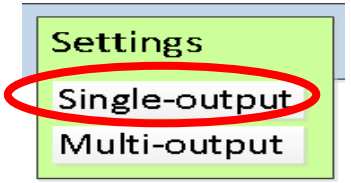
Import User data window





1. **Browser**: Import User data file.
2. Inspect if right data in left panel. By clicking on **OK** data will appear in right panel.
3. Define a row with a parameter to its line. Click on **Add**. Multiple parameters can be define by repeating this step. Parameters can be combined.
4. **Define the row where spectra starts.**
5. If needed, **convert** spectral data.
6. Option to **remove** samples.
7. Configured input data can be **save**d and **load**ed as .m file.
8. Finally, click on **Import**.

Single-output Settings

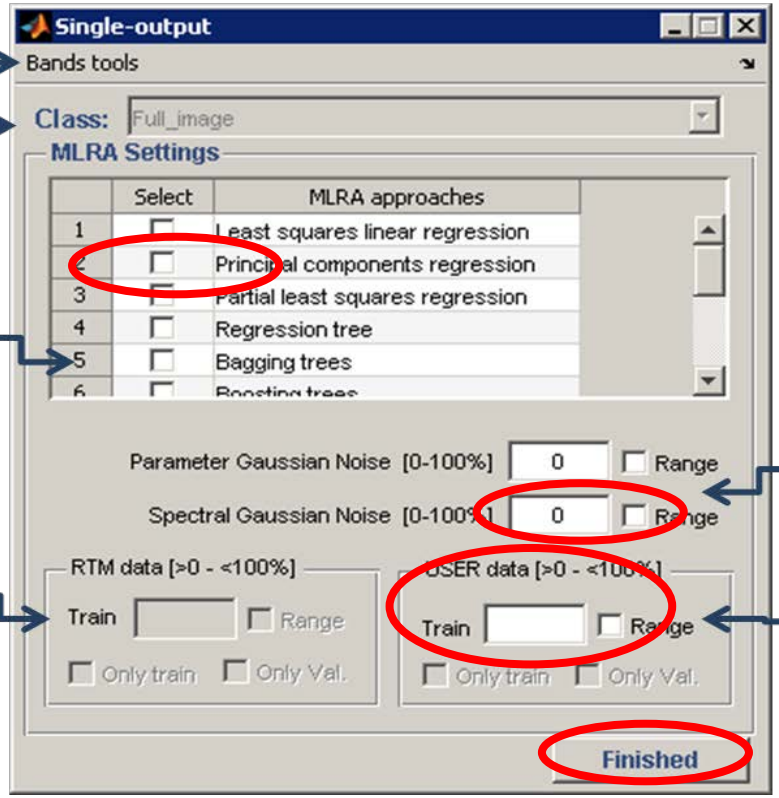


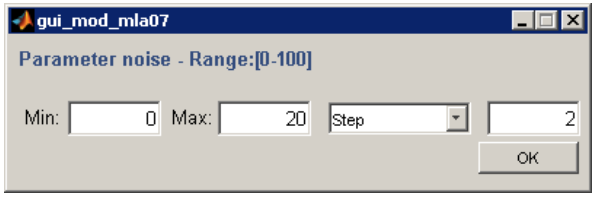
Option to apply a band subselection or transformation

If a land cover map has been provided, per clas scan be configured.

Multiple regressors can be selected

If RTM data is inserted it can serve for training or validation





Range options:

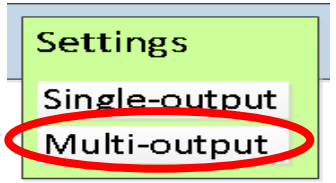
- Step
- Distribution

Options to add noise to parameters and spectral data

Options to control the training/validation partitioning for user data

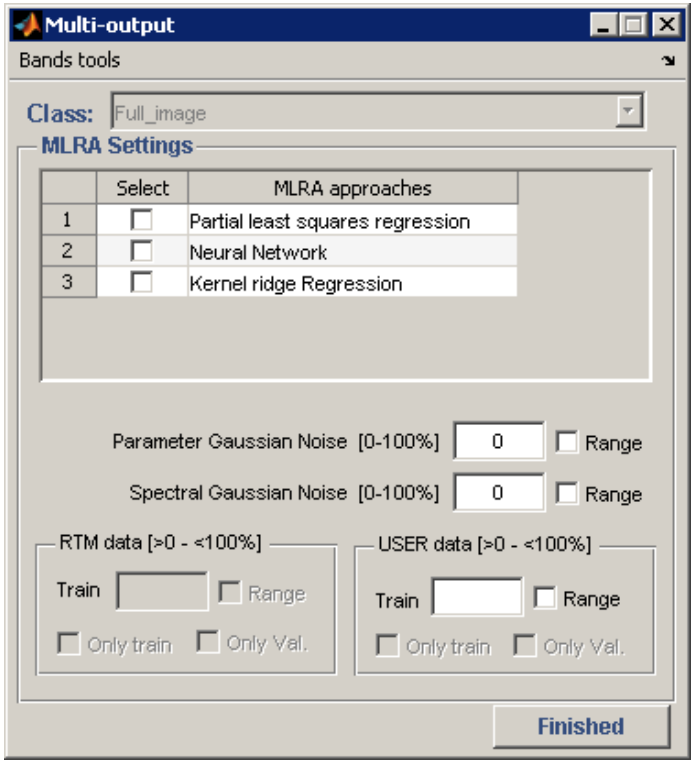
1. Select the **nonparametric regression models** to be trained and validated.
2. The option to add Gaussian **noise** is provided. A range of noise scenarios can be applied.
3. Select the User data **training/validation partitioning**. This will randomly partition the input data in a training and validation dataset. **Make sure to keep some data for validation (thus < 100% training).** Also a range of training/validation partitioning scenarios can be applied. **If no validation is required, go directly to Retrieval.**
4. Click on **Finished.**

Multi-output Settings



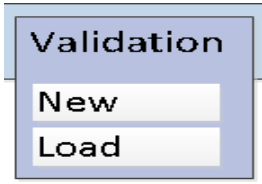
When multiple input variables have been selected, the following regression algorithms provide multiple-outputs with the same single model:

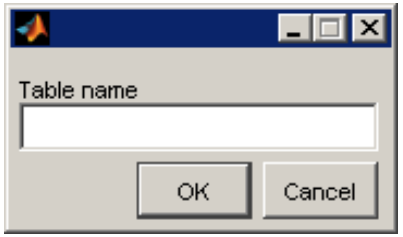
- partial least square regression (PLSR)
- neural networks (NN)
- kernel ridge regression (KRR).



The same options as Single-output are provided

Validation



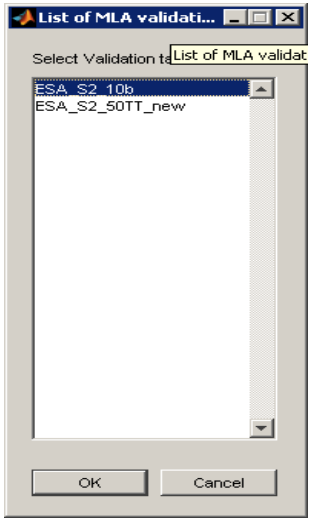


- Start a **New** validation: provide a name

- All MLRA scenarios will be trained using training data and validated against validation data according to goodness-of-fit indicators:
  - R, R^2, RMSE, RELRMSE, NRMSE, ME, MAE (see manual)

- Results will be automatically stored in a MySQL table.

- When finished, an overview table will appear (*see next slide*). Such overview table can also be consulted when selecting: **Load.** A window with generated validation results will appear: (see next slide)



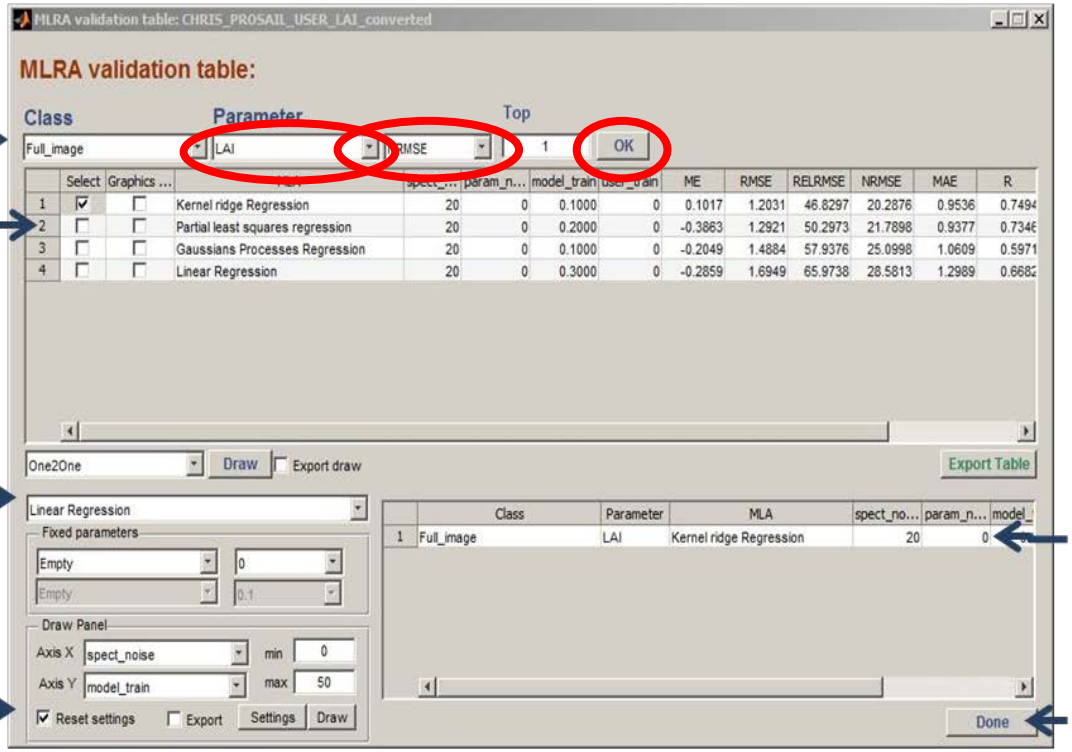
MLRA validation

Options to organize statistics per Class, Parameter, stat. and # of best results

Overview of best performing MLRAs. The check boxes allow to select a MLRA model for retrieval, the graphics for graphics outputs.

Graphic options: 1:1-line, sigmas (GPR), and 2D correlation matrices.

Options to control graphics properties.



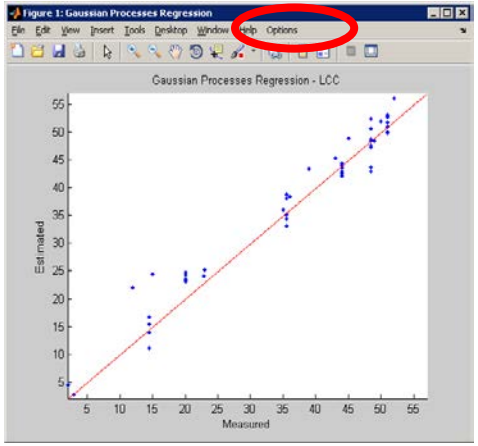
Option to export a table to a .txt file.

*Selected MLRA model. With **Done** it will be moved to **Retrieval.***

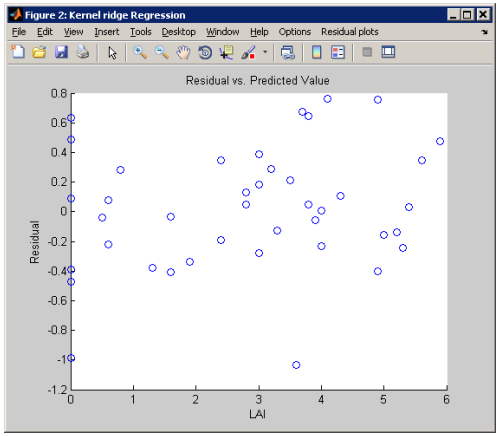
1. Choose how to **sort outputs,** according to parameter, statistic and number of top results per regressor. Click on **OK**.
2. **Select** a **MLRA scenario for retrieval** (e.g. the top performing one). It will move to lower panel. When clicking on **Done** it will move to the Retrieval window (slide 11).
3. Select a MLRA scenario for **Graphics plottings: 1:1-line measured vs. predicted.** For GPR additional band relevance information will be provided. **Make sure to have User data loaded, because the selected model will be regenerated.**
4. In case ranges were introduced (noise, training/validation partitioning), validation results can be plotted in a **2D-matrix**. Results are plotted according to selected parameter and statistic.

Examples of validation results

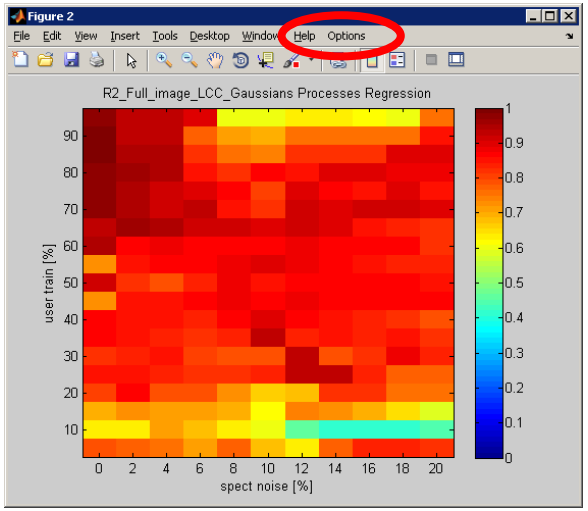
Measured data (validation) vs. estimated data along 1:1-line



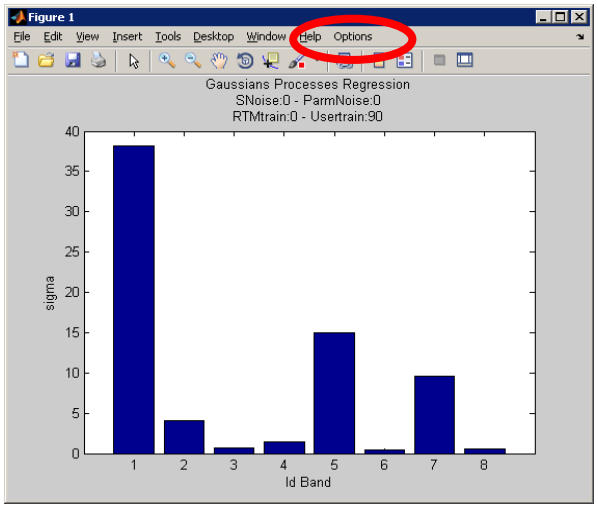
Residuals



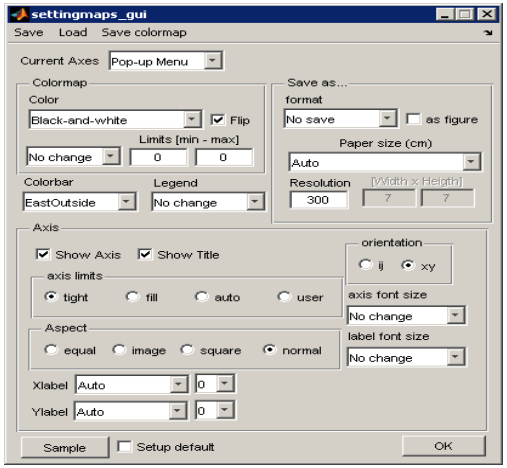
2D-matrix with validation results according to selected parameter and statistic.



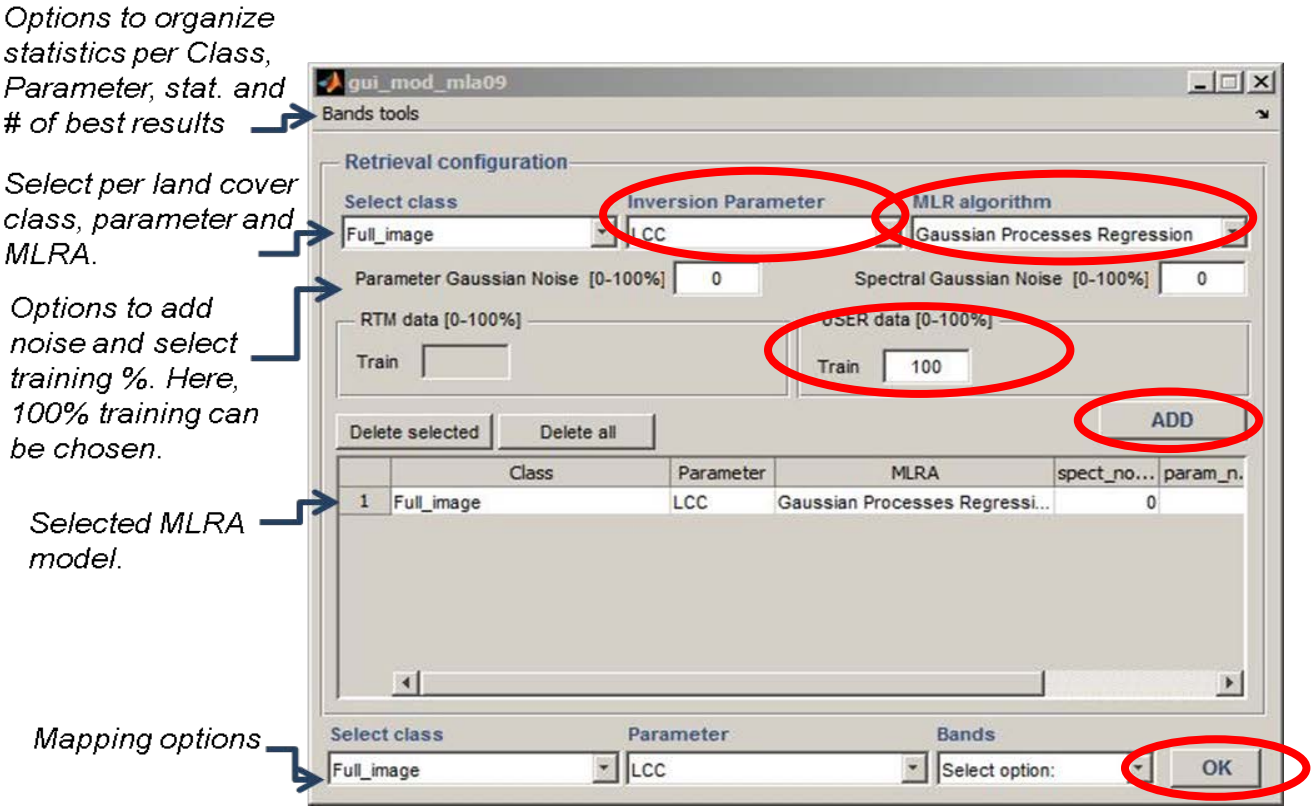
For GPR sigmas (band relevance) are provided. The lower the sigma, the more important the band.



When clicking on **Options**, options are provided to control the the figure properties and export it.



Retrieval



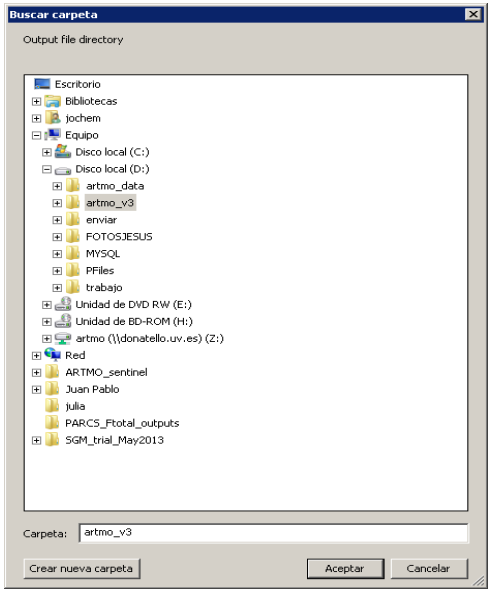
Instead of going through the validation procedure, **one can also choose to immediately train a regression model and apply it to a remote sensing image**. User data has for training to be first inserted (see slides 4 & 5).

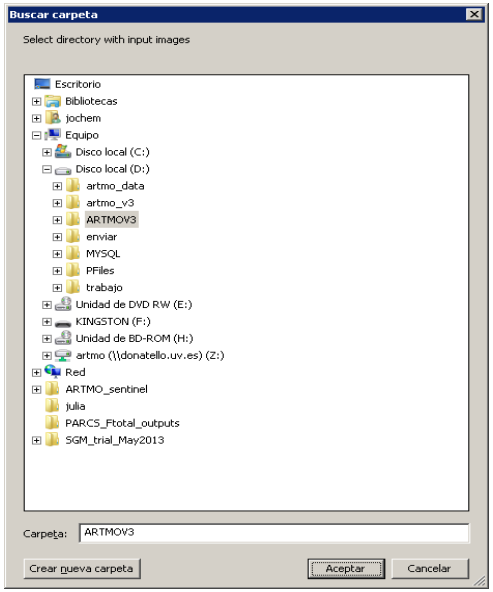
1. Select the **parameter** and **the regression model**.
2. Optionally **noise** can be added.
3. Select the **training partitioning**. Here 100% training data can be applied. The configuration need to be **ADD**ed and chosen model will appear in the down panel. **In case a model has been selected during the validation step, it will directly appear in that panel.**
4. Band tools options are provided (e.g. spectral subset, PCA).
5. When clicking on **OK**, the mapping procedure will start (see next slide).

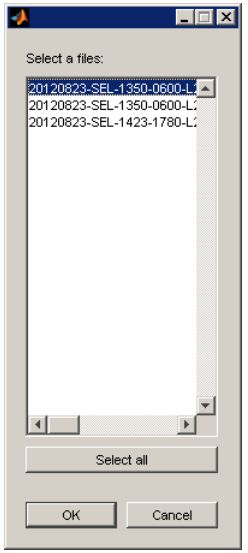
Retrieval

The mapping of selected biophysical parameter requires the following steps:

1. Select the **Output directory**
2. Select the directory with **Input images**
3. Images according to **ENVI file format** (including .hdr file) will be identified and listed. **Multiple images can be selected**. They will be processed one-after-another.
4. When the processing is done, the output maps can be viewed. Select one through **Open Map** and click on **PREVIEW**.

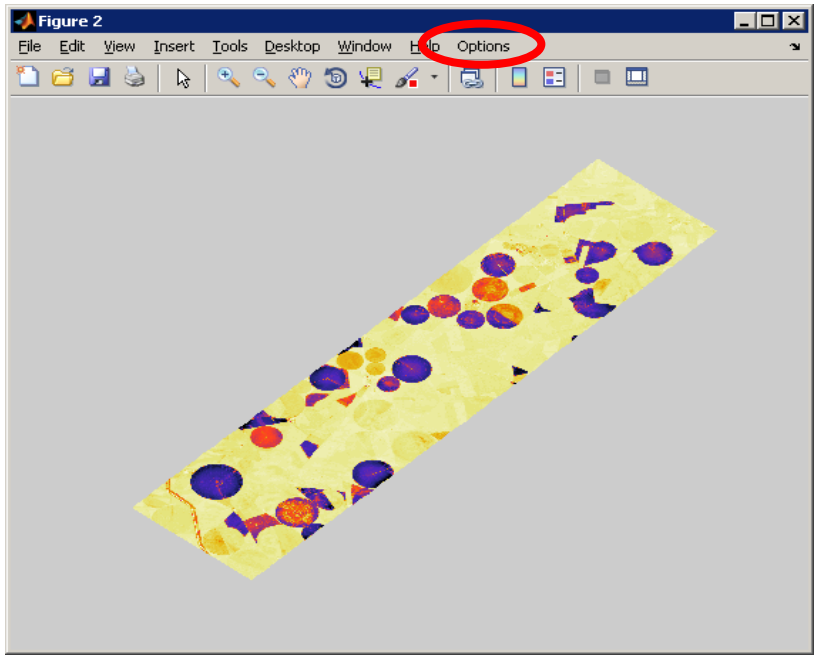


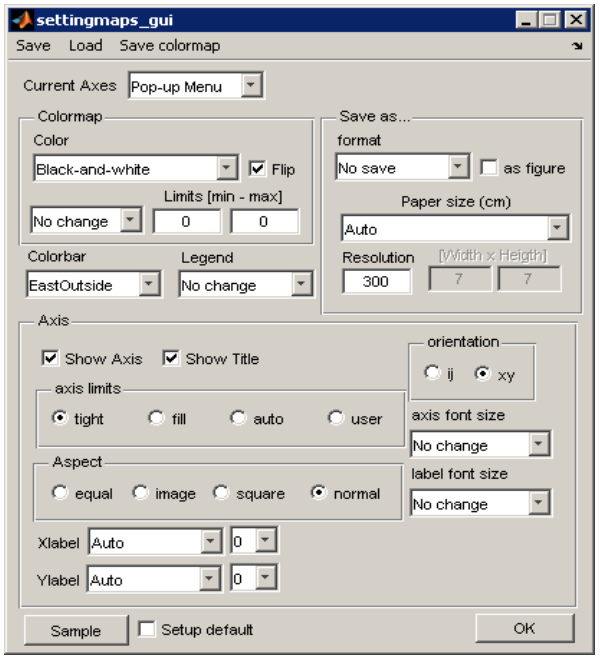




A drop-down list will show the provided output layers. One output map can then be peviewed.

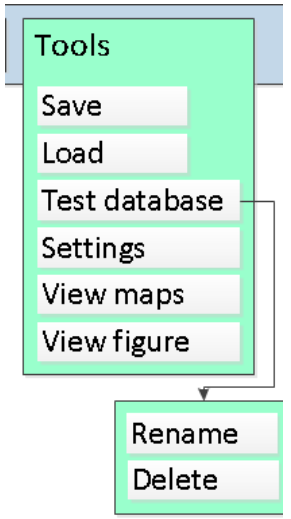
Final maps



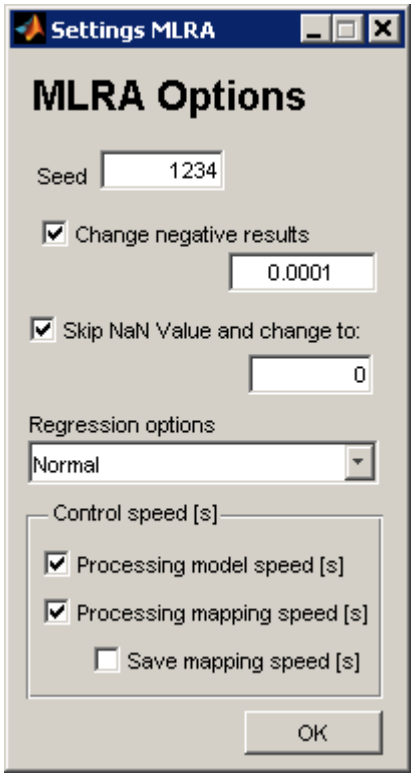


- Visualization of an output layer. In **Options**, map properties can be controlled (e.g. color scale, color table).
- Make sure to orient the map according to **ij** for **correct orientation**.
- The map can be **saved** according to various vector or bitmap formats. Redundant white space around the figure will be automatically removed.
- Settings can be set as **default** – will be automatically applied to subsequent maps.
- Click on **Sample** to visualize the map. Click on **OK** to save it away.

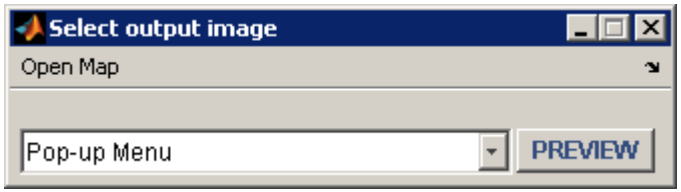
Tools

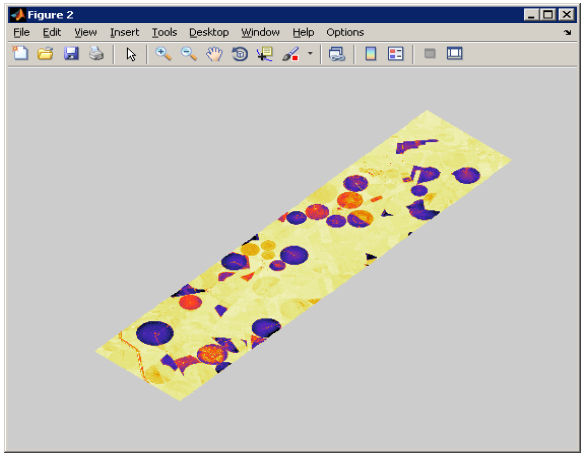


Options

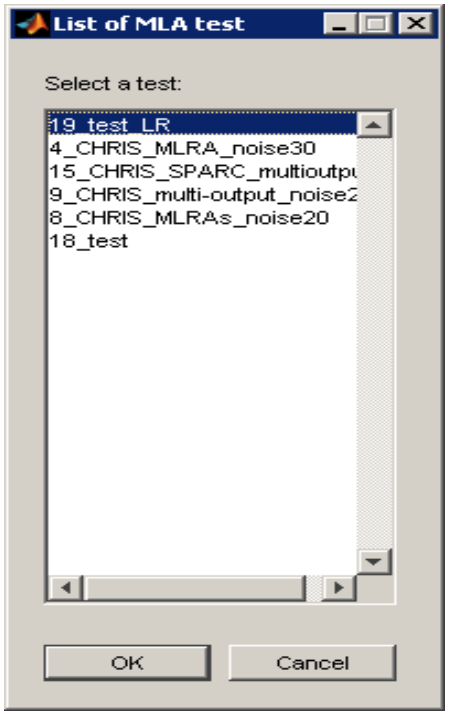


View a map – select a band





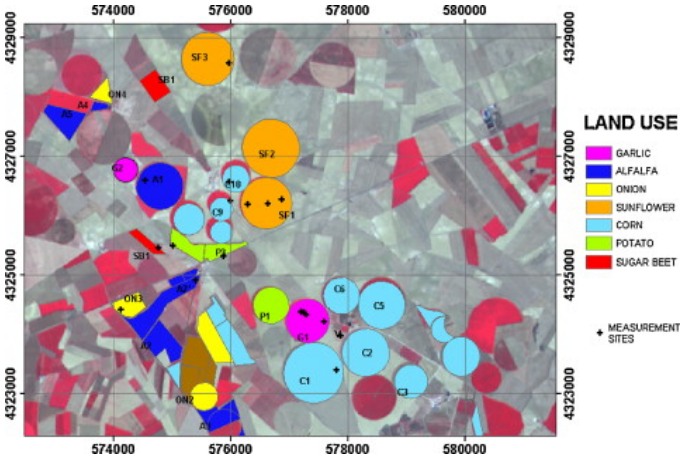
Delete validation tables



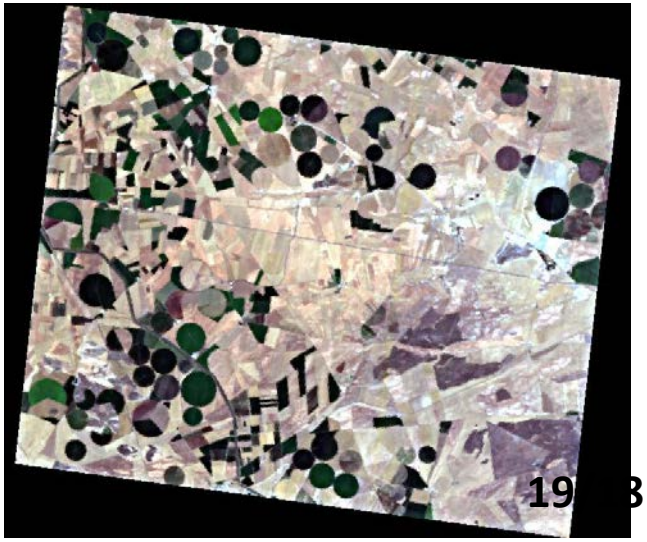
Field & RS data

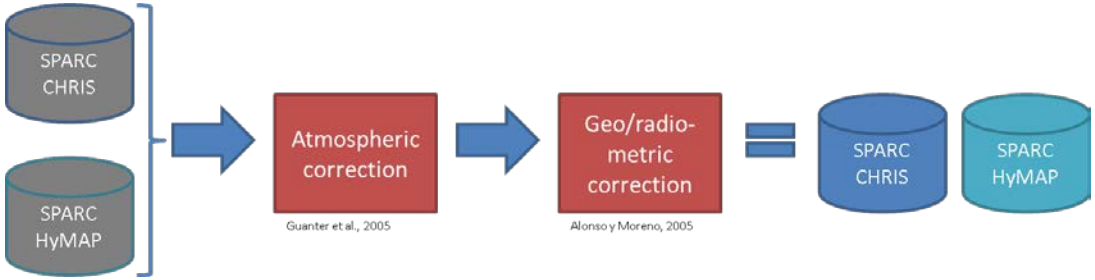
SPARC data set (July 2003; **Barrax**, Spain)

- **Field data** (135 points):
  - **Leaf Chl** measured with CCM-200
  - **LAI** measured with LiCor LAI-2000
  - **FVC** measured with hemispherical photographs
  - 30 additional **bare soil samples**



- **Spectral data**:
  - **CHRIS mode 1** (62 bands; 34m) nadir spectra (July 2003). Data has been resampled to Sentinel-2 (20 m: 8 bands)
  - **HyMAP** (125 bands; 5 m)





Excercise

- ✓ Evaluate the performance of MLRAs using a field dataset and remote sensing data.
- ✓ Apply the best performing regression algorithm to a RS image.