ARTMO'S NEW MACHINE LEARNING REGRESSION ALGORITHMS (MLRA) MODULE FOR SEMIAUTOMATIC MAPPING OF BIOPHYSICAL PARAMETERS

Jochem Verrelst[1], Juan Pablo Rivera[1], Jordi Muñoz-Mari[1], Jose Moreno, Gustavo Camps-Valls[1]

1. University of Valencia, Image Processing Laboratory, Valencia, Spain; jochem.verrelst@uv.es

ABSTRACT

As part of the scientific Automated Radiative Transfer Models Operator (ARTMO) software package a new module, being 'Machine Learning Regression Algorithms' (MLRA) Module, has been developed. ARTMO provides a seamless link between inputs and outputs required for running a suite of reflectance models both at the leaf level and at the canopy level. The toolbox facilitates consistent and intuitive user interaction, thereby streamlining model setup, running, storing and output plotting for any kind of optical sensor operating in the VNIR range. In this work ARTMO version 3 (V.3) is presented. It differs from earlier versions that it is completely redesigned in a modular architecture. As such, new models and modules can be easier implemented into the toolbox. Specifically, the MLRA Module enables to analyze the predictive power of various MLRA in an automated manner. Data can either come from field campaign or from simulations and options have been implemented such as controlling training/validation data partitioning, adding noise. As a showcase, the performance of all implemented MLRAs have been evaluated using the SPARC dataset (Barrax, Spain) and hyperspectral CHRIS spectral observations. In general, PLSR outperformed LR but best results were obtained with the nonlinear MLRAs. Most stable results were obtained by KRRR, closely followed by GPR. The advantage of the latter regressor is that insight in relevant bands and associated uncertainty estimates are delivered. Finally, we applied a GPR model to several images to map leaf chlorophyll content (LCC) and associated uncertainties to gain insight in the robustness of the model.

INTRODUCTION

With the forthcoming superspectral Sentinel-2 and Sentinel-3 missions and the planned EnMAP and PRISMA imaging spectrometers, operational Earth observation (EO) is reaching a state of maturity. This unprecedented data availability requires processing techniques that are easy and fast to apply to obtain information about the plants' growth or health status. For the last decade the family machine learning regression algorithms (MLRAs) emerged as powerful method for delivering biophysical parameters in an operational context. MLRAs have the potential to generate adaptive, robust relationships and, once trained, they are very fast to apply (1). Typically, machine learning methods are able to cope with the strong nonlinearity of the functional dependence between the biophysical parameter and the observed reflected radiance. They may therefore be more suitable candidates for operational applications. Effectively, algorithms such as neural networks (NNs) are already implemented in operational retrieval chains (e.g. CYCLOPES products). It remains nevertheless to be questioned whether NNs offer the most flexible tools for parameter estimation, gaining insights in the retrievals and evaluating retrieval performances. Besides, training NNs involve tuning several parameters that may greatly impact the final robustness of the model. In part, this is why in the recent years, NNs are being replaced by more advanced, simpler to train regression methods. Specifically, during the last two decades, the family of kernel methods (2) has emerged as an alternative to NNs in many scenarios. Kernel methods typically involve few and intuitive hyperparameters to be tuned, and can perform flexible input–output nonlinear mappings.

While MLRAs are recognized as powerful methods, by the broader remote sensing community they are also perceived as complicated. Moreover, some MLRAs need to be tuned, which requires expertise. To facilitate and automate the use of MLRAs, in this work we present a recently developed 'MLRA' Module that allows systematically analyzing and applying MLRA-developed models. This module is being implemented within the innovative toolbox called ARTMO: "Automated Radiative Transfer Models Operator" (3).

The following sections will first briefly describe the latest status of the ARTMO toolbox, followed by introduction of the most important components of the new MLRA Module. Subsequently the used data is described and first results that function as showcases are presented. A conclusion closes this paper.

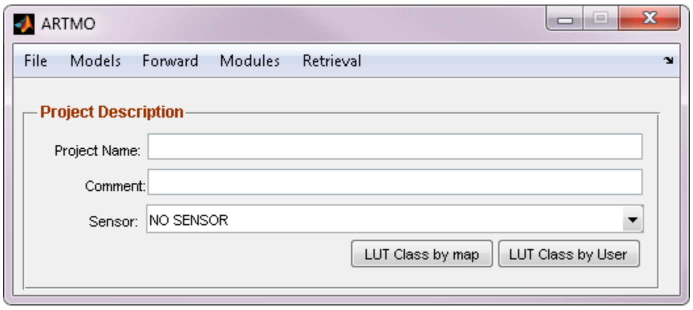
ARTMO V.3

A first version of ARTMO has been presented at the 7th EARSeL Imaging Spectrometry Workshop 2011 (Edinburgh, UK). In short, ARTMO brings multiple leaf- and canopy-RT models together along with essential tools required for semiautomatic retrieval of biophysical parameters in one graphical user interface. The toolbox, developed in Matlab, permits the user: i) to choose between various invertible leaf and canopy RTMs with varying complexity (e.g., PROSPECT-4, PROSPECT-5, 4SAIL, SLC, FLIGHT), ii) to choose between spectral band settings of various air- and space-borne sensors or defining new sensor settings, iii) to simulate a massive amount of top-of-canopy (TOC) reflectance spectra of any sensor in the range of 400 to 2500 nm based on look-up tables (LUT) which are then stored in a database, and finally, iv) to run various retrieval strategies.

In comparison to the first version, an updated version is presented here, being ARTMO 3 (V.3). Various major changes and new modules have been introduced in this version. The most important ones are briefly listed below:

- ARTMO has been completely redesigned and is now organized in a modular way. The modular architecture makes possible that easily RTM models can be added or removed. The idea behind this modular design is that the existing models can be seamless coupled with new or other types of models. For instance, it is foreseen to couple canopy models with atmospheric models so that to top-of-atmosphere radiance can be simulated and inverted.
- Internally, the MySQL database has been reorganized in a more efficient manner to support the modular architecture, to avoid redundancy and to speed up processing.
- Various new retrieval modules have been designed, based on parametric regression, non-parametric regression and physically based inversion. This led to the development of 1) 'Spectral Indices module', 2) 'Machine Learning Regression Algorithm module' and, 3) 'LUT-based inversion', respectively.

Figure 1 presents ARTMO V.3's main window. Compared to earlier versions, the main window has been considerably simplified. Now in the main window a new project can be initiated, a sensor chosen and a comment added, but all modules are accessible through drop-down menus at the top bar. These drop-down menus depend on the modules and tools found within the ARTMO folder and can thus easily be expanded. A systematic overview of the drop-down menu is provided in Figure 2. The focus of this paper is on 'Machine Learning Regression Algorithm module'.



Project Description

Project Name:

Comment:

Sensor: NO SENSOR

Figure 1: ARTMO V.3's main window

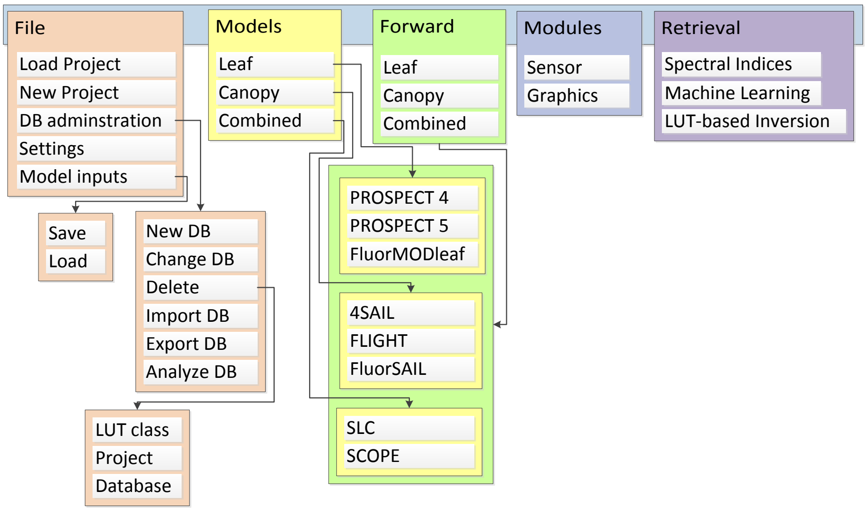


Figure 2: ARTMO's main modular architecture.

1. MLRA Module

The MLRA Module is again designed in a modular manner. Figure 3 displays the module as displayed by Matlab, while a schematic overview of the dropdown menu is shown below. In short, the MLRA module enables to apply and evaluate multiple MLRAs. The MLRAs have been categorized according to single-output, which means that one per parameter is developed; or multi-output, which means that the developed model can deliver outputs of multiple parameters. Two sorts of input data can be loaded into the module for training or validation of the relationships. On one hand, data can come from the RTMs. In this way an earlier generated LUT with simulated spectra can be imported. On the other hand, user-defined data can be imported, e.g. as collected during field campaigns. Options are available to merge and partition both datasets, e.g. training on the basis of RTM data and validation on the basis of user-defined field data. Furthermore, if a land cover map is loaded then for each land cover class different MLRA optimization strategies can be defined. Most importantly, when having validation data available, multiple MLRA strategies (e.g. with different noise and train/validation partitioning) can be analyzed against the validation and statistical results (e.g. R^2, RMSE, NRMSE) are stored in a relational database. The best performing strategy can then be loaded and applied to an image. In the following sections the most important modules will be explained in a bit more detail. These are: 1) 'MLRA settings', 2) 'Test results', and 3) 'Retrieval'.

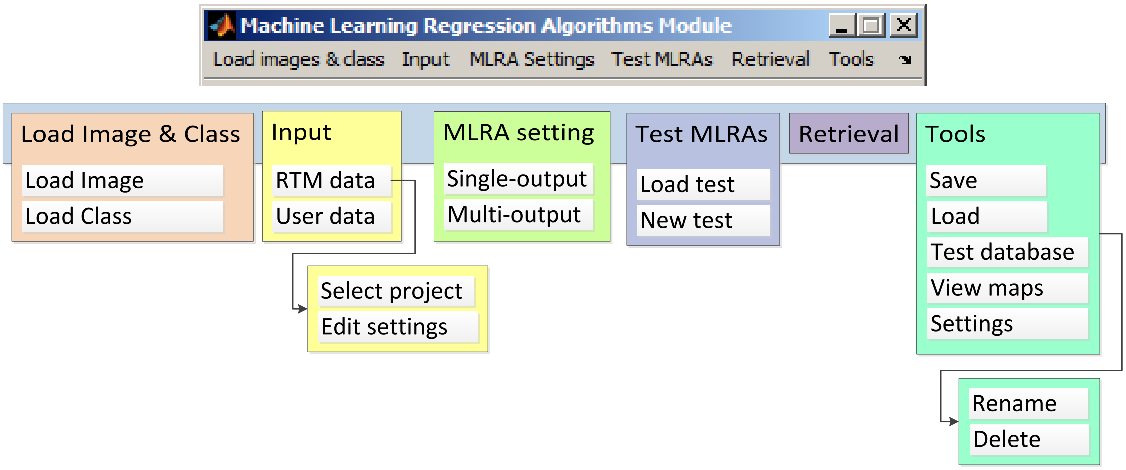
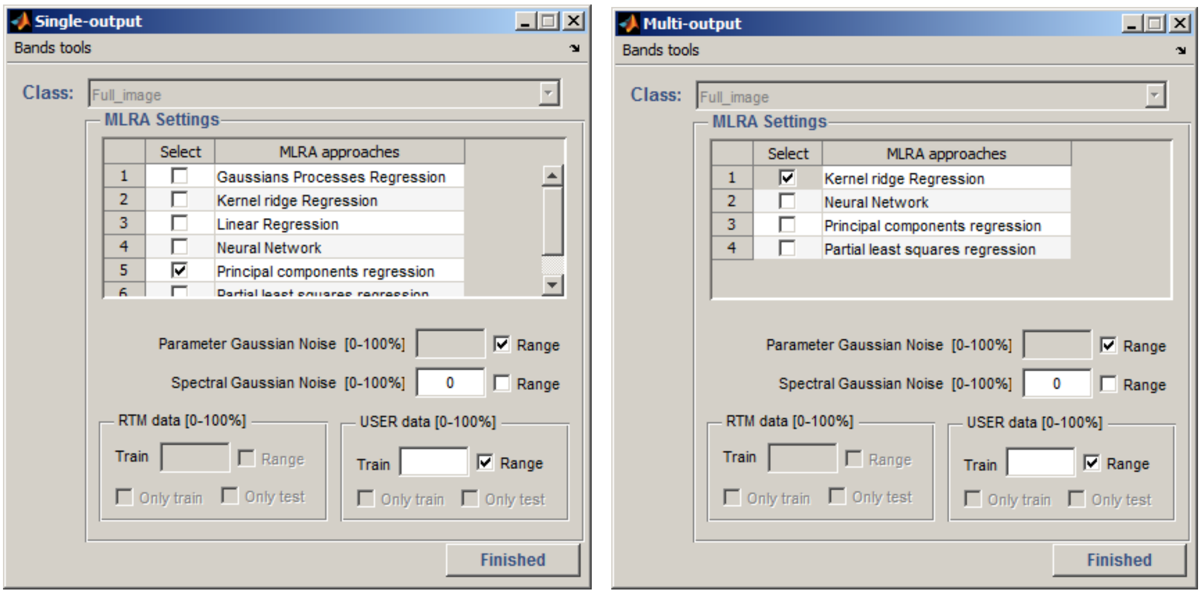


Figure 3. ARTMO's MLRA architecture.

2. MLRA settings

The following step addresses the analysis of multiple MLRA-based retrieval strategies. A first step to do is inserting RTM LUT-data or ground truth data through the Input module. Required input data refers to retrievable biophysical parameters and associated spectra. Once this is done the 'MLRA settings' module is activated (Figure 4). It can be opted to select either the single-output regressors or the multi-output regressors. Multi-output regressors include partial least square regression (PLSR), neural networks (NN) and kernel ridge regression (KRR). Note that these models can also be used into single-output. Other single-output regressors include principal component regressor (PCR), support vector regression (GPR), Gaussian processes regression (GPR) and the conventional linear regression (LR). The 'MLRA settings' GUI configures the nonparametric regression strategies through five successive steps. First, if multiple land cover classes have been defined (within the 'Load Image and Class' window) then retrieval strategies can be configured per land cover class. Second, multiple nonparametric regressors can be selected. Third, options to add Gaussian noise are provided. Noise can be added both on the parameters, as on the spectra. Here, a range of noise can be configured, so that multiple noise scenarios can be evaluated. The injection of noise can be of importance to account for environmental and instrumental uncertainties when synthetic spectra from RTMs are used for training. Fourth, the train/val data partition can be controlled by setting the percentage how much data from a RTM or user-defined is assigned to training or to validation. Both datasets can also be merged by selecting a portion of both datasets for training and validation. Also here multiple train/val partitions can be evaluated. Fifth, since each added band puts a burden on the computational load, an option to compute relevant bands has been added to overcome the Hughes phenomenon. For now the computation of relevant bands occurs through mutual information theory, but it is also foreseen to implement feature extraction techniques such as PCA. Also efforts are foreseen to enable identifying redundant samples, e.g. through active learning. Sample reduction may be particularly valid when MLRAs are trained with data coming from RTMs. Such dataset easily consists of several ten thousands simulations, but many of them can be considered as redundant, e.g. in cases where the spectral impact of RTM parameters is beyond the selected wavelengths.



Single-output

Bands tools

Class: Full_image

MLRA Settings

	Select	MLRA approaches
1	☐	Gaussians Processes Regression
2	☐	Kernel ridge Regression
3	☐	Linear Regression
4	☐	Neural Network
5	☑	Principal components regression
6	☐	Partial least squares regression

Parameter Gaussian Noise [0-100%] ☑ Range

Spectral Gaussian Noise [0-100%] 0 ☐ Range

RTM data [0-100%]: Train ☐ Range ☐ Only train ☐ Only test

USER data [0-100%]: Train ☑ Range ☐ Only train ☐ Only test

Multi-output

Bands tools

Class: Full_image

MLRA Settings

	Select	MLRA approaches
1	☑	Kernel ridge Regression
2	☐	Neural Network
3	☐	Principal components regression
4	☐	Partial least squares regression

Parameter Gaussian Noise [0-100%] ☑ Range

Spectral Gaussian Noise [0-100%] 0 ☐ Range

RTM data [0-100%]: Train ☐ Range ☐ Only train ☐ Only test

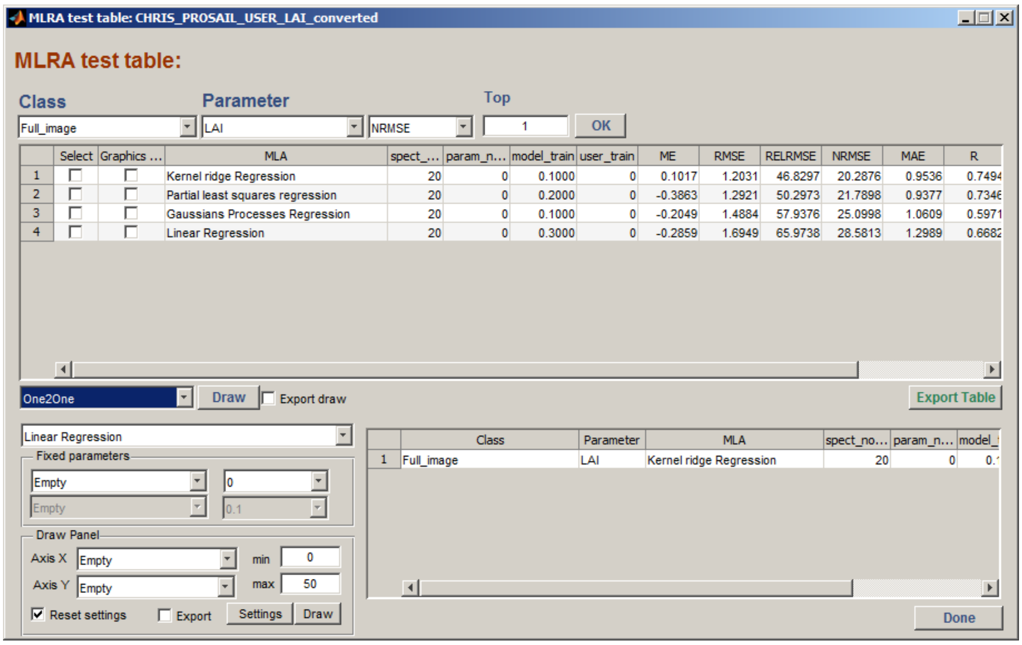
USER data [0-100%]: Train ☑ Range ☐ Only train ☐ Only test

Figure 4: GUI to configure the MLRA settings for single-output [left] and multi-output [right].

Finally, once having the train/val data partitioning defined and MLRA settings configured, then those scenarios can be run. This is done through naming a new test in the 'Test MLRA. Subsequently, one-by-one all retrieval strategies over the configuration ranges are analyzed.

3. Test results

All results are automatically saved in a MySQL database. This has the advantage that a large number of results can be stored in a systematic manner and that results can be easily queried. Only validation results are presented in the 'MLRA test table', as those are the results that matter (training results alone may face the problem of overfitting). The estimations are evaluated against the validation dataset through a wide range of evaluation statistics, being R^2, RMSE, NRMSE, MAE, ME. In the table the best performing results is shown according to selected Class (if configured), parameter and statistic (figure 5). Further, various options to display the results are provided. For instance, 1:1-line, plotting the sigmas (relevant bands) of GPR, and then matrices of performances along ranges such as noise and train/val partitioning (see further showcases in Results). Finally, by clicking on 'Retrieval' an analyzed regression function can be selected for each retrievable parameter (e.g. the best one). Such regression function can then be applied to an image.



MLRA test table: CHRIS_PROSAIL_USER_LAI_converted

MLRA test table:

Class: Full_image
Parameter: LAI
NRMSE
Top: 1

	Select	Graphics ...	MLA	spect_...	param_n...	model_train	user_train	ME	RMSE	RELRMSE	NRMSE	MAE	R
1	□	□	Kernel ridge Regression	20	0	0.1000	0	0.1017	1.2031	46.8297	20.2876	0.9536	0.7494
2	□	□	Partial least squares regression	20	0	0.2000	0	-0.3863	1.2921	50.2973	21.7898	0.9377	0.7346
3	□	□	Gaussians Processes Regression	20	0	0.1000	0	-0.2049	1.4884	57.9376	25.0998	1.0609	0.5971
4	□	□	Linear Regression	20	0	0.3000	0	-0.2859	1.6949	65.9738	28.5813	1.2989	0.6682

One2One
□ Export draw

Linear Regression

Fixed parameters: Empty, 0; Empty, 0.1

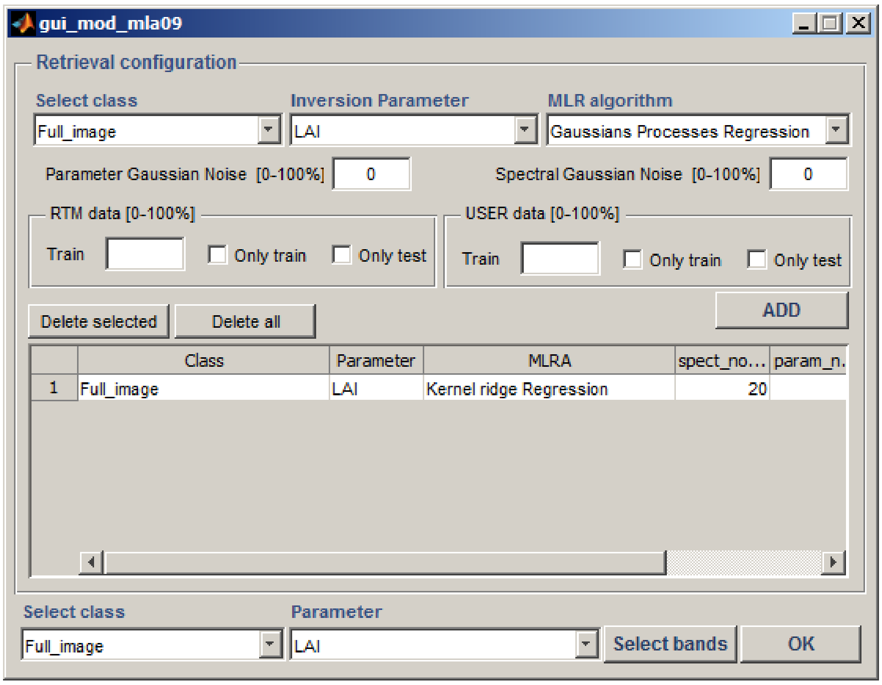
Draw Panel
Axis X: Empty; min: 0
Axis Y: Empty; max: 50
☑ Reset settings
□ Export

	Class	Parameter	MLA	spect_no...	param_n...	model_
1	Full_image	LAI	Kernel ridge Regression	20	0	0.1

Figure 5: GUI with results of tested configurations.

4. Retrieval

In the Retrieval GUI it is also possible to directly configure a relationship and apply to an image to map a parameter. Hence, the user can select the required land cover class (if available), the retrievable parameter, the regressors and train/val partitioning. Similarly, noise can be added to the spectra or parameters and the size of the training data can be selected. Multiple retrieval strategies can be added, e.g. for each retrievable parameter another one. Finally, by clicking on OK, the input images can be loaded and the output maps will be written away in an ENVI format.



gui_mod_mla09

Retrieval configuration

Select class: Full_image
Inversion Parameter: LAI
MLR algorithm: Gaussians Processes Regression

Parameter Gaussian Noise [0-100%]: 0
Spectral Gaussian Noise [0-100%]: 0

RTM data [0-100%]
Train: □ Only train □ Only test

USER data [0-100%]
Train: □ Only train □ Only test

	Class	Parameter	MLRA	spect_no...	param_n.
1	Full_image	LAI	Kernel ridge Regression	20	

Select class: Full_image
Parameter: LAI

Figure 6: MLRA retrieval GUI

USED DATA FOR SHOWCASES

A diverse field dataset, covering various crop types, growing phases, canopy geometries and soil conditions was collected during SPARC (Spectra bARrax Campaign). The SPARC-2003 and SPARC- 2004 campaigns took place in Barrax, La Mancha, Spain (coordinates 30º3´N, 28º6´W, 700 m altitude). In the 2003 campaign, carried out on 12-14 July, biophysical parameters were measured within a total of 113 Elementary Sampling Units (ESU) among different crops. ESU refers to a plot size compatible with a pixel size of about 202 m. In the 2004 campaign, carried out on 15-16 July, the same field data were collected within a total of 18 ESUs among different crops. Leaf chlorophyll content (LCC) was derived by measuring within each ESU about 50 samples with a calibrated CCM-200 Chlorophyll Content Meter. LCC values obtained in the SPARC 2003 campaign show good agreement with these obtained in the SPARC 2004 campaign. Green LAI was derived from canopy measurements made with a LiCor LAI-2000 digital analyser. Each ESU was assigned to a LAI value, which was obtained as a statistical mean of 24 measures (8 data readings x 3 replications) with standard errors between 5 and 10%. For both years, we have a total of 9 crops (garlic, alfalfa, onion, sunflower, corn, potato, sugar beet, vineyard and wheat), with field-measured values of LAI that vary between 0.4 and 6.3 and LCC between 2 and 55 µg/cm^2. Further details on the measurements can be found in (4). Additionally, 60 random spectra over bare soils, man-made surfaces and water bodies were added to broaden the dataset to non-vegetated samples (i.e., with a biophysical LCC and LAI value of zero).

During the campaign hyperspectral CHRIS images were acquired. CHRIS provides high spatial resolution hyperspectral data over the VNIR spectra from 400 to 1050 nm. It can operate in different modes, balancing the number of spectral bands, site of the covered area and spatial resolution because of on-board memory storage reasons. We made use of nominal nadir CHRIS observations in Mode 1 (62 bands, maximal spectral information) for the four SPARC campaign days, where field measurements of surface properties were measured in conjunction with satellite overpasses. CHRIS Mode 1 has a spatial resolution of 34~m at nadir. The spectral resolution provides a bandwidth from 5.6 to 33 nm depending on the wavelength. The images were geometrically corrected followed by atmospheric correction (see (4) for details).

SHOW CASES

The performance of the different regressors were evaluated along gradients of changing training/validation distributions (from 5 to 95% training, with steps of 5%) and increasing Gaussian noise levels (from 0 to 20% with steps of 2%). SPARC field data was used for training and validation and associated spectral data came from CHRIS. Models were developed both for LCC and LAI. Validation results are presented in the form of NRMSE, which allows evaluating across different parameters. As a guideline, remote sensing users (e.g., GMES) require an error threshold below 10%

When comparing these matrices, the following observations can be made. To start with the conventional LR as a reference it can be observed that this regressor is performing poorly when having only relatively few training portion available. In fact only acceptable results (NRMSE <10%) occurred when having more than 90% of the data used for training. Adding noise appeared to be of a less important factor, though in general results worsened with increasing noise. PCR performs more stable, but results are rather poor for all scenarios. Second, the in RS widely PLSR regressor already greatly improved, especially when being fed by a relatively low training portion. It can also be observed that only after introducing more than 10% noise results started to degrade. Nevertheless, when inspecting Table 1 where best results are presented it can be observed that PLSR is performing considerably worse than the nonlinear MLRAs. This can be explained by the functioning of PLSR and PCR that conducts transformations in the linear space. Conversely, NN, KRR and GPR conduct transformations in the nonlinear space and can therefore perform more flexible. Regarding NN, the erratic pattern is most notable. Hence, it appears that while being to deliver very accurate results in some cases the regressor is rather unstable, with a high probability of delivering poor results. This erratic behavior can be explained by the complicated training phase that develops highly specialized models, but therefore easily faces the problem of overfitting. The non-robustness is a major disadvantage of NN. From all regressors KRR yielded most impressive results. It not only led to best performing results (see table 1), but also proved to be stable with increasing noise levels. Also this regressor benefits from increasing portion of training data, but already excellent results can be obtained with about 80% training data. Note also that KRR was fastest trained, which overall, makes this as a very powerful regressor. Finally, GPR is another promising MLRA regressor. Although performing more unstable than KRR, it also leads to excellent performances when having 80% or more training. The regressor, though, is somewhat more affected by added noise, and also some configurations gave no results (darkest blue). Nevertheless, GPR is particularly of interest because of unique additional features; 1) it provides insight in relevant bands when developing the model; and, 2) it provides uncertainty estimates associated with the mean predictions.

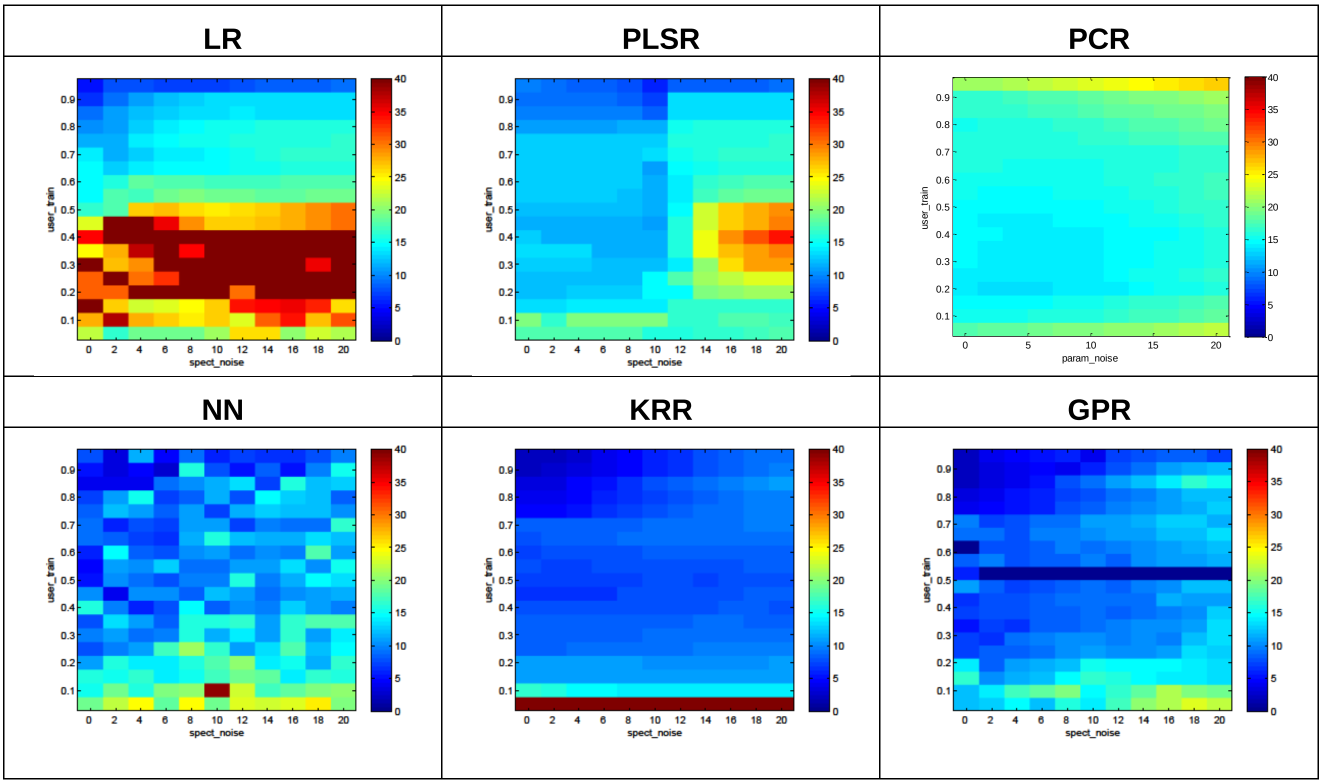
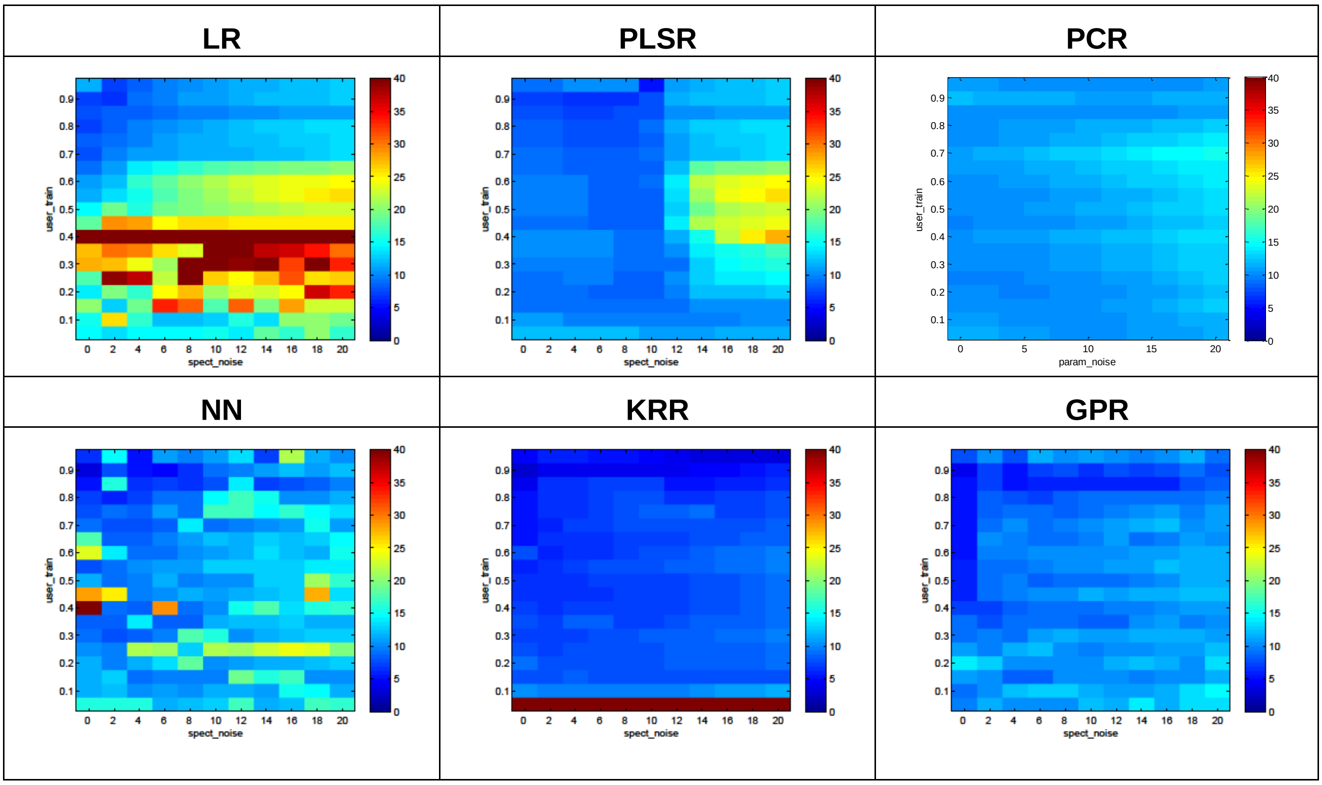


Figure 8: LCC validation NRMSE matrices along gradients of increasing noise (X-axis) and increasing training partitioning (Y-axis).

Table 1: Best evaluated strategies per regressor from results shown in Figure 8.

MLRA	Spectral noise [%]	training [%]	RMSE	NRMSE [%]	R2
Kernel ridge Regression (KRR)	0	95	0.97	1.89	0.998
Gaussians Processes Regression (GPR)	0	90	1.03	2.02	0.997
Neural Network (NN)	6	90	1.50	2.95	0.995
Linear Regression (LR)	0	95	2.71	5.31	0.988
Partial least squares regression (PLSR)	10	95	2.90	5.69	0.991
Principal components regression (PCR)	4	20	7.18	13.6	0.89

The same exercise was repeated but then for the estimation of LAI. NRMSE matrix results are provided in Figure 9, and best performing results are listed in Table 2. Essentially the same results appeared, with PCR performing stable but poor and PLS and LR performing poorer than the nonlinear MLRAs; NN performing rather unstable and KRR performing most robust and GPR in between these two.


LR

PLSR

PCR

NN

KRR

GPR

Figure 9: LAI validation NRMSE matrices along gradients of increasing noise (X-axis) and increasing training partitioning (Y-axis).

Table 2: Best evaluated strategies per regressor from results shown in Figure 9.

MLRA	Spectral noise [%]	training [%]	RMSE	NRMSE [%]	R2
Kernel ridge Regression (KRR)	0	90	0.15	2.75	0.99
Neural Network (NN)	0	90	0.19	3.42	0.99
Gaussians Processes Regression (GPR)	0	90	0.22	3.98	0.99
Partial least squares regression (PLSR)	10	95	0.21	5.55	0.99
Linear Regression (LR)	2	90	0.36	6.65	0.96
Principal components regression (PCR)	4	20	0.60	9.95	0.90

Finally, the best performing strategy can be applied to images of interest. Because statistical methods are often criticized for their lack of portability, particularly GPR is a promising regressor to apply for mapping. The delivery of associated uncertainty estimates allows us to provide insight on a per pixel basis when applied to any image. Therefore the consecutive approach was to apply the best performing GPR strategy a multitude of CHRIS images acquired over various sites across the world and were atmospherically corrected with the BEAM toolbox (Figure 10). For brevity, only LCC results are shown.

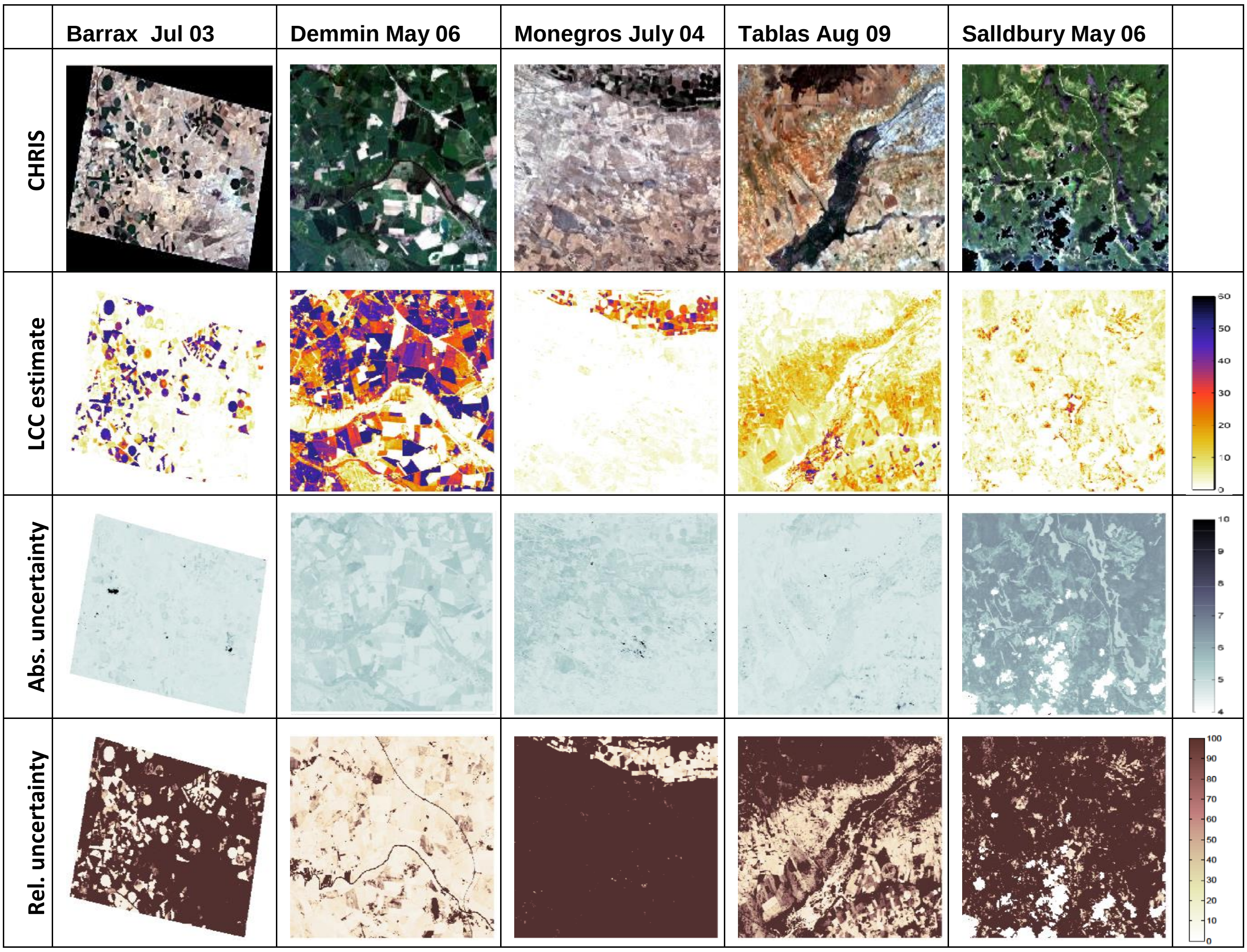


Figure 10: CHRIS images [top], mean LCC estimates [2nd row], absolute uncertainties [3th row] and relative uncertainties [bottom].

These maps are briefly discussed. Since the LCC model was trained with field data obtained in Barrax and spectral data during the CHRIS overflight, the first column can therefore be considered as the reference image. The absolute uncertainties provide insight in the model performances; the lower the uncertainty the more robust the obtained retrieval. When then transporting the GPR model to other images it can be observed that the same degree of uncertainties is to be found over the other images. Only the ‘Saldbury’ image (Canada) led to systematic poorer uncertainties, probably due to poorer atmospheric correction because of lower sunlight intensity. However, as an uncertainty of e.g. 5 µg/cm^2 against a mean estimate of 50 µg/cm^2 is more reliable than e.g. against a mean estimate of 10 µg/cm^2 it may be more valuable to provide relative uncertainties (σ/µ). Such maps are provided at the bottom of the figure. From those maps it can be clearly observed that retrievals were processed with more certainty over some areas than to others. This is also clearly visible over the reference image. In fact, it appears that retrievals with high certainty occur over the irrigated vegetated parcels. These were also the areas that were sampled during the SPARC campaign. Conversely, LCC was retrieved with low certainty over the dried-out lands.

Also hardly training data was collected on these areas. The same pattern is also visible over the other sites; over the agricultural ‘Demmin’ sites low uncertainties were encountered, while over the dried out land at ‘Monegros’ and ‘Las Tablas’ higher uncertainties occurred. Also ‘Saldbury’ was processed with on the whole poor uncertainties, suggesting that this image was less suitable for retrieving LCC with the GPR model. Overall, the relative uncertainties revealed the areas with robustly retrieved estimates and the areas that would benefit from a denser sampling scheme.

CONCLUSIONS

In this work ARTMO version 3 (V.3) is presented. It is designed in a modular architecture and consists of various new modules. Specifically, the ‘MLRA Module’ enables to analyze the predictive power of various nonparametric regressors. Multiple options have been implemented, e.g., controlling training/validation data partitioning and adding noise. Data can come either from field campaigns or from simulations as generated by radiative transfer models. When a land cover maps is loaded then per class a different retrieval strategy can be evaluated. The predictive powers of multiple MLRAs were evaluated. Kernel ridge regression (KRR) and Gaussian Processes regression (GPR) were evaluated as best performing and most robust. Moreover, GPR provides additional uncertainty estimates which enables evaluating the portability of the model.

ACKNOWLEDGEMENTS

This paper has been partially supported by the Spanish Ministry for Science and Innovation under projects: AYA2010-21432-C02-01 and CSD2007-00018.

REFERENCES

1. Hastie T, R Tibshirani & J H Friedman. 2009. The elements of statistical learning: data mining, inference, and prediction (2nd ed.). New York: Springer-Verlag.
2. Camps-Valls G, & L Bruzzone (Eds.). Dec 2009. Kernel methods for Remote Sensing Data Analysis. UK: Wiley & Sons.
3. Verrelst J, J P Rivera, L Alonso & J Moreno, 2011. ARTMO: an Automated Radiative Transfer Models Operator toolbox for automated retrieval of biophysical parameters through model inversion. Proceedings of EARSeL 7th SIG-Imaging Spectroscopy Workshop, Edinburgh, UK.
4. Verrelst J, J Muñoz, L Alonso, J Delegido, J P Rivera, G Camps-Valls & J Moreno. 2012. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel- 2 and -3. Remote Sensing of Environment 118, 127–139.