SEMIAUTOMATIC RETRIEVAL OF BIOMASS BASED ON VEGETATION INDEX OPTIMIZATION AND MACHINE LEARNING METHODS FOR WINTER RAPESEED CROPS

Dessislava Ganeva

Space Research and Technology Institute – Bulgarian Academy of Sciences
e-mail: dganeva@space.bas.bg

***Keywords**: biophysical parameters retrieval, biomass, rapeseed, empirical regression models, machine learning methods*

***Abstract**: In order to evaluate crop condition, remote sensing technics together with regression models or machine learning regression algorithms (MLRA) are usually used. Historically Vegetation Indexes (VI) are employed coupled with regression models, and more recently MLRA are used, to estimate a given biophysical parameter. For the evaluation of fresh and dry biomass from winter rapeseed crop, the Automated Radiative Transfer Models Operator (ARTMO) package with Sentinel-2 images and ground data are used. The sampled pixels from Sentinel-2 images are evaluated as a single pixel and as averaged with the 8 closest. In order to better detect bare soil, samples from bare soil were included.*

The preliminary results show that the bare soil samples add to the determination power of the models and single pixel models give better results than the averaged pixels.

Although the MLRA and the regression models with VI have similar goodness-of-fit measures (i.e. MAE, RMSE, NRMSE, R^2), the resulting image of estimated fresh and dry biomass are better fitted for MLRA and almost not fitted with the regression models with VI. Because of the difficulties to interpret the results of those methods, of particular interest could become the MLRA that include uncertainty estimation, as the Gaussian Progress Regression Algorithm.

This approach allows a quick and broad view of the relation between remote sensed and ground data. As well as identify locally related correlations between the remote sensing and biophysical parameters.

ПОЛУАВТОМАТИЧНО ОПРЕДЕЛЯНЕ НА БИОМАСА НА БАЗА НА ОПТИМИЗИРАНЕ НА ВЕГЕТАЦИОННИ ИНДЕКСИ И МЕТОДИ ЗА МАШИННО ОБУЧЕНИЕ НА ПОСЕВИ ОТ ЗИМНИ РАПИЦА

Десислава Ганева

Институт за космически изследвания и технологии – Българска академия на науките
e-mail: dganeva@space.bas.bg

***Ключови думи**: извличане на биофизични параметри, биомаса, рапица, емпирични регресионни модели, методи за машинно обучение*

***Резюме**: Дистанционните методи на наблюдение използват регресионни модели или алгоритми на машинно обучение (ММО) за оценка на състоянието на посеви. Исторически се използват вегетационни индекси (ВИ) заедно с регресионни модели, а все по-често се използват ММО, за да се оцени даден биофизичен параметър. За оценката на свежа и суха биомаса на зимна рапица се използва софтуерното приложение ARTMO заедно с изображения от Sentinel-2 и наземни данни. Измерените пиксели от изображения на Sentinel-2 се оценяват като единичен пиксел и като осреднени с 8-те най-близки. За да се моделир по-добре участъци от почва без растителност, бяха включени проби от гола почва.*

Предварителните резултати показват, че моделите с почвени проби подобряват резулта на моделите, атези с единичните пикселни дават по-добри резултати от осреднените пиксели.

Макар че ММО и регресионните модели с ВИ имат сходни параметри за грешка и сходство (MAE, RMSE, NRMSE, R^2), получените изображение на свежата и суха биомаса след прилагане на модела е по-добре изразено с ММО и почти не е с регресионни модели с ВИ. Поради трудностите при интерпретирането на резултатите от тези модели, от особен интерес може да се превърнат моделите, които включват оценка на сигурност, като Gaussian Regression Algorithm Progress.

Предложеният метод дава бърз и широк поглед върху връзката между наземните данните и тези от дистанционно наблюдение и земята. Освен това позволява да се идентифицират на локално ниво корелации между дистанционни наблюдения и биофизичните параметри.

Introduction

Part of the winter rapeseed crop monitoring consists in evaluation of the crop condition before and after winter. In this period the rapeseed crop develops its leaves and biomass. Therefore, quick and accurate retrieval of fresh or dry biomass is of importance for remote sensing monitoring of the winter rapeseed crops. This study presents biomass retrieval from Sentinel-2 images by parametric and non-parametric models

Materials and Method

Study area, description of the winter rapeseed fields and ground data

This study was carried out in East Danube plain in Bulgaria, over one growing season, from September 2017 to July 2018, on three mass fields sown with different hybrids of winter rapeseed, Fig. 1. The area is mostly flat, the soil has mainly sandy loam texture, the climate in this region is Moderate Continental with cold winters and hot summers (mean daily temperature 10.2 ºC), and an annual cumulative rainfall of 540 mm.

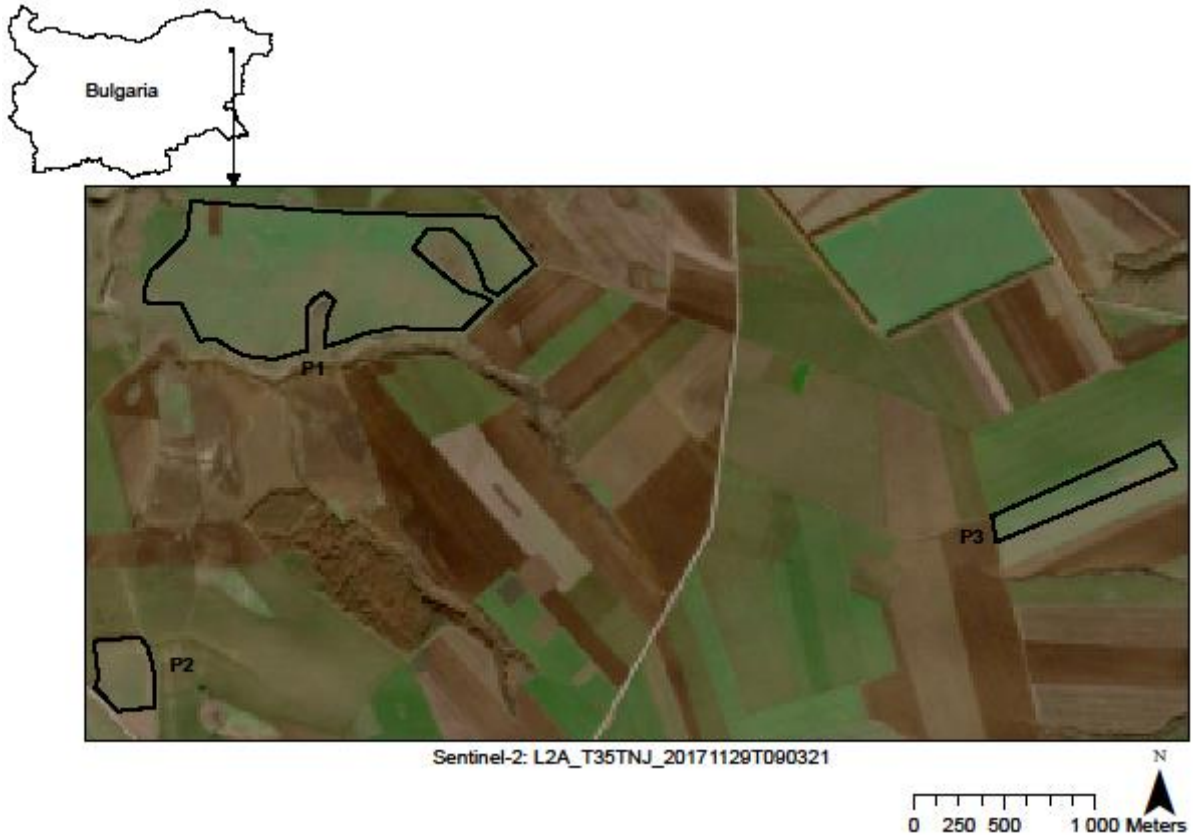

Fig. 1. Study area and winter rapeseed fields

Before the field campaign, the sample locations were identified. A literature review has shown a set of Vegetation Indices (VI) well correlated with important biophysical parameters for winter rapeseed, Table 1. Those indices were calculated for the studied rapeseed fields on the Sentinel-2 image from 12.11.2017, downloaded from Copernicus Data Open Hub in 2A product (https://scihub.copernicus.eu/dhus/#/home). The sample locations were positioned in order to capture as much as possible to heterogeneity of the fields in terms of biophysical parameters identified in Table 1 on the VI maps.

Table 1. Vegetation Indices used to position the location of the samples for the field campaign

Biophysical parameter	Vegetation Index	Formulation	Reference
AGB, biomass, number of plants per square meter after emergence	RVI (Ratio Vegetation Index)	NIR/Red	(Piekarczyk, Wójtowicz and Wójtowicz, 2006)
AGB, biomass	OSAVI (Optimized soil adjusted)	(1+L)(NIR-Red)/(NIR+Red+L) (L = 0.16)	(Han *et al.*, 2017)
LAI	SAVI (Soil adjusted vegetation index)	(1+L)(NIR-Red) / (NIR+Red+L) (L = 0.5)	(Hatfield and Prueger, 2010)
Canopy chlorophyll and nitrogen	CIredEdge	R783/R705−1	(Clevers and Gitelson, 2013)
Plan height	EVI (The enhanced vegetation index)	EVI=G x ((RNIR-Rred)/(RNIR+C1 x Rred-C2 x Rbleu +L)); G=2.5; C1=6; C2=7,5; L=1	(Bartoszek, 2014)
Vegetation Fraction	VARIgreen (Visible Atmospherically Resistant Indices)	(R550 – R670)/(R550 + R670)	(Fang *et al.*, 2016)
Number of plants per square meter after emergence	NDVI (Normalized Difference Vegetation Index)	(NIR-Red)/(NIR+Red)	(Piekarczyk, Wójtowicz and Wójtowicz, 2006)

During each field campaign, one before winter and one after winter, a sample was identified by its position measured by consumer hand held GPS device. The Aboveground Fresh Biomass (FBM) was harvested as described by (Cihlar *et al.*, 1987) from an area of 1m² and all plants cut, stored in paper bags and transported to a laboratory. In the laboratory the same day, each sample of FBM is weighted. The dry biomass (DBM) is obtained from a sample of the FBM, within 24 hours, by oven-drying at 105 °C until constant weight. Each field campaign produced 15 measurements. In total 30 samples of FBM and DBM were registered for the study.

Because of the meteorological conditions, some of the plants started growing immediately after sowing but many had more than a month delay. Particularly the plots P2 and P3 were with plants in very different phonological phases, from BBCH13 to BBCH19 (Weber, Bleiholder and Lancashire, 1991), during the before winter field campaign. This difference in the phonological phase was completely reduced after winter, where all plots were at BBCH50/BBCH51. One sample during before winter campaign was sampled from an area with plants in BBCH19 and it was clearly an outlier compare to the other samples, but not regarding the field condition.

Table 2. Characteristics of the fields, dates of the selected Sentinel-2 images, field campaigns, and the number of sampling locations per plot

					Before Winter		**After Winter**	
Field Code	Area (ha)	Planting Date	Sowing rate plant/ m²	Sampling locations	Sampling date	Sentinel-2 image (product)	Sampling date	Sentinel-2 image (product)
P1	137	3.09.2017 4.09.3017 5.09.2017	80	9	23.11.2017	29.11.2017 (2A)	1.04.2018	03.04.2018 (2A)
P2	10	20.08.2017	56	3	23.11.2017	29.11.2017 (2A)	1.04.2018	03.04.2018 (2A)
P3	15	4.09.2017	76	3	24.11.2017	29.11.2017 (2A)	1.04.2018	03.04.2018 (2A)

Remote sensing images and data

For this study all spectral bands of 10 m and 20 m spatial resolution from Sentinel-2 are used. It provided 10 spectral bands from 490 nm to 2190 nm, resampled at 10 m special resolution. The selected Sentinel-2 images for the study were the closest available cloud-free on the studied area and closest to the dates of field campaign (Table 2).

Models calibration and evaluation

The biomass retrieval was carried out with parametric and non-parametric regression methods (Verrelst *et al.*, 2015) with the Automated Radiative Transfer Models Operator (ARTMO) package (http://ipl.uv.es/artmo/). The aim was to identify the best method for retrieval of FBM and DBM for rapeseed and study the influence on the models of two main aspects:

- By adding 3 samples of bare soil to each ground data campaign. In total 6 more samples to the 30 of vegetation/soil from the ground data campaign. It was expected to have better fitted model when adding bare soil data.
- By averaging the values of the 9 closest pixels to the sample location of the remote sensed image. As the smallest pixel size of Sentinel-2 images is 10m² and the position of the sample during the campaign was measured with a consumer GPS device, the sample location could have been closest to the edge of the pixel or even on the edge of a neighbor pixel.

Another aspect that was studied is the influence of the outlier sample from the before winter field campaign. The tested scenarios (different input data for the models) are: 1) All biomass (30) samples with remote sensing data from 1 pixel. 2) All biomass and 6 bare soil (36) samples with remote sensing data from 1 pixel. 3) All biomass (30) samples with remote sensing data from 9 pixels. 4) All biomass and 6 bare soil (36) samples with remote sensing data from 9 pixels. 5) All biomass without the outlier (29) samples with remote sensing data from 1 pixel. 6) All biomass without the outlier and 6 bare soil (35) samples with remote sensing data from 1 pixel. 7) All biomass without the outlier (29) samples with remote sensing data from 9 pixels. 8) All biomass without the outlier and 6 bare soil (35) samples with remote sensing data from 9 pixels.

The models were validated with leave-one-out cross-validation, because of the small sample size. Even if FBM and DBM are highly correlated with Pearson's correlation (r) of 0.99, both biophysical variables were modeled.

The tested parametric regression methods consist in applying linear, exponential, logarithmic, power and polynomial fitting functions to VI of 2 or 3 bands, as described in Table 1, and one additional VI of three bands, 3BI=(B1-B2)/(B1+B3). All possible fitting functions are executed with all possible VI and all bands. For each test scenario the best performing model was recorded, as well as the best OSAVI and SR ones.

The tested non-parametric models are: Least squares linear regression (LSLR), Principal components regression (PCR), Partial least squares regression (PLSR), Kernel Ridge Regression (KRR), Gaussian Progress Regression (GPR) and the Variational Heteroscedastic variant of the Gaussian Progress Regression (VHGP). Some of the models, such as GPR, perform uncertainty evaluation. The GPR and VHGP calculate a Coefficient of Variation (CV = σ/μ), where σ is the Standard Deviation (SD) around the estimated biomass and μ the mean estimated biomass. CV provides relative uncertainty of the estimated parameters in %.

Results and discussion

All models were ranked on NRMSE (Normalized RMSE in %, NRMSE=100*RMSE/range of biomass measured). The NRMSE was selected because it is not influenced by the data unit (Richter *et al.*, 2012) and therefore can compare accuracy across different parameters. The approach that is adopted in this study is to first for each scenario select the best ranked model in term of NRMSE. Then each model is applied to both selected remote sensing images and linear regressions function performed between the simulated and measured values (all biomass measures with the outlier and without the bare soil additional samples). The results from the correlations that have $R^2 > 0.514$ are considered significant at $\alpha = 0.05$, because of the small sample size (Rogerson, 2001).

The best performing models, ranked by NRMSE, Table 3, are for the parametric and non-parametric models the ones with the scenario 8. However, when those models were applied to the remote sensing images and linear regressions function performed between the simulated and measured values, they had a poor fit (Fig. 2).

Table 3: Best performing models, ranked by NRMSE

	VI	FF	Bands	MAE	RMSE	RRMSE	NRMSE %	R	R2
FBM_SI_BareSoil_9Pixels_NoOutliner	3BI	polynomial	490;705;865	226	288	34	10	0,94	0,88
MLRA_BareSoil_9Pixels_NoOutliner_ML_KRR_FBM	KRR			216	298	35	11	0,93	0,87

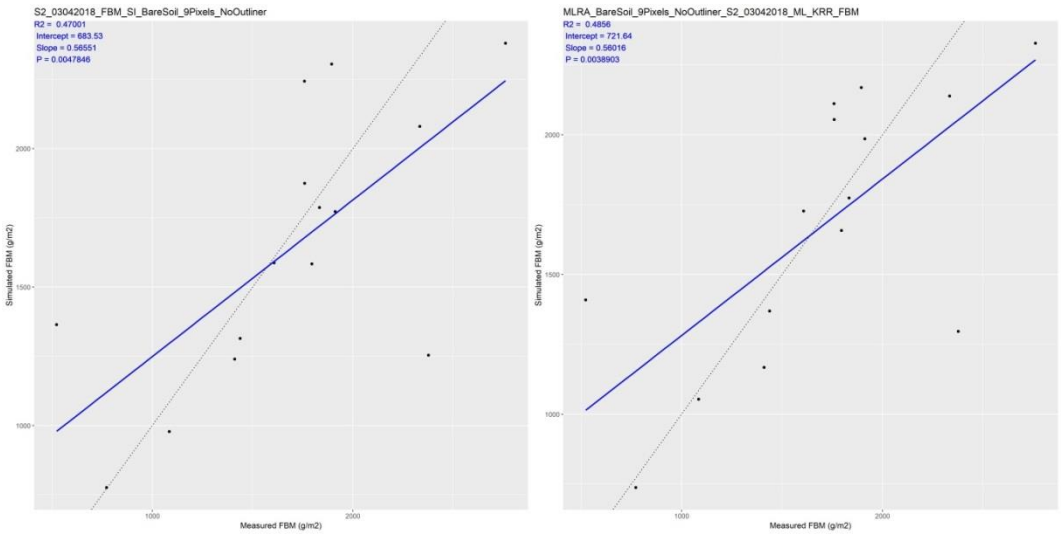

Fig. 2. Best performing models, ranked by NRMSE, applied to the remote sensing images and linear regressions function performed between the simulated and measured values

There was not a parametric model that provided satisfactory results for the before winter retrieval. Neither of the models achieve an error threshold under 10% that is the typical remote sensing end user requirements (Caicedo *et al.*, 2014).

Therefore, a different approach was followed by trying to find the best model that gives good result when applied to the remote sensing images. Following this approach the, best results were the one with scenario 2, Table 4. The models applied to the remote sensing images (Fig. 3), show that the lower the biomass the higher is the uncertainty. The comparison between both models and the orthophoto obtained by and Unmanned Aerial Vehicle (UAV) with RGB camera shows good overall estimation of the more and less vegetated area in the plots.

Table 4: Best Results by applying the model to the remote sensing image and evaluating the fit between the Measured and Estimated values

Model Name	Model	Bands	MAE	RMSE	RRMSE	NRMSE %	R	R2
MLRA_BareSoil_1Pixel_ML_GP_DBM	GPR	560;740;490;842	35	52	50	16	0,85	0,73
MLRA_BareSoil_1Pixel_ML_VHGP_FBM	VHGP	560;490;740;842	333	474	54	17	0,83	0,69

Test Name	R2 plot	Intersept	Slop	P
MLRA_BareSoil_1Pixel_S2_29112017_ML_GP_DBM	0,89	16,81	0,71	1,65E-07
MLRA_BareSoil_1Pixel_S2_03042018_ML_GP_DBM	0,73	57,08	0,70	5,23E-05
MLRA_BareSoil_1Pixel_S2_29112017_ML_VHGP_FBM	0,84	170,36	0,65	1,36E-06
MLRA_BareSoil_1Pixel_S2_03042018_ML_VHGP_FBM	0,71	529,47	0,67	7,95E-05

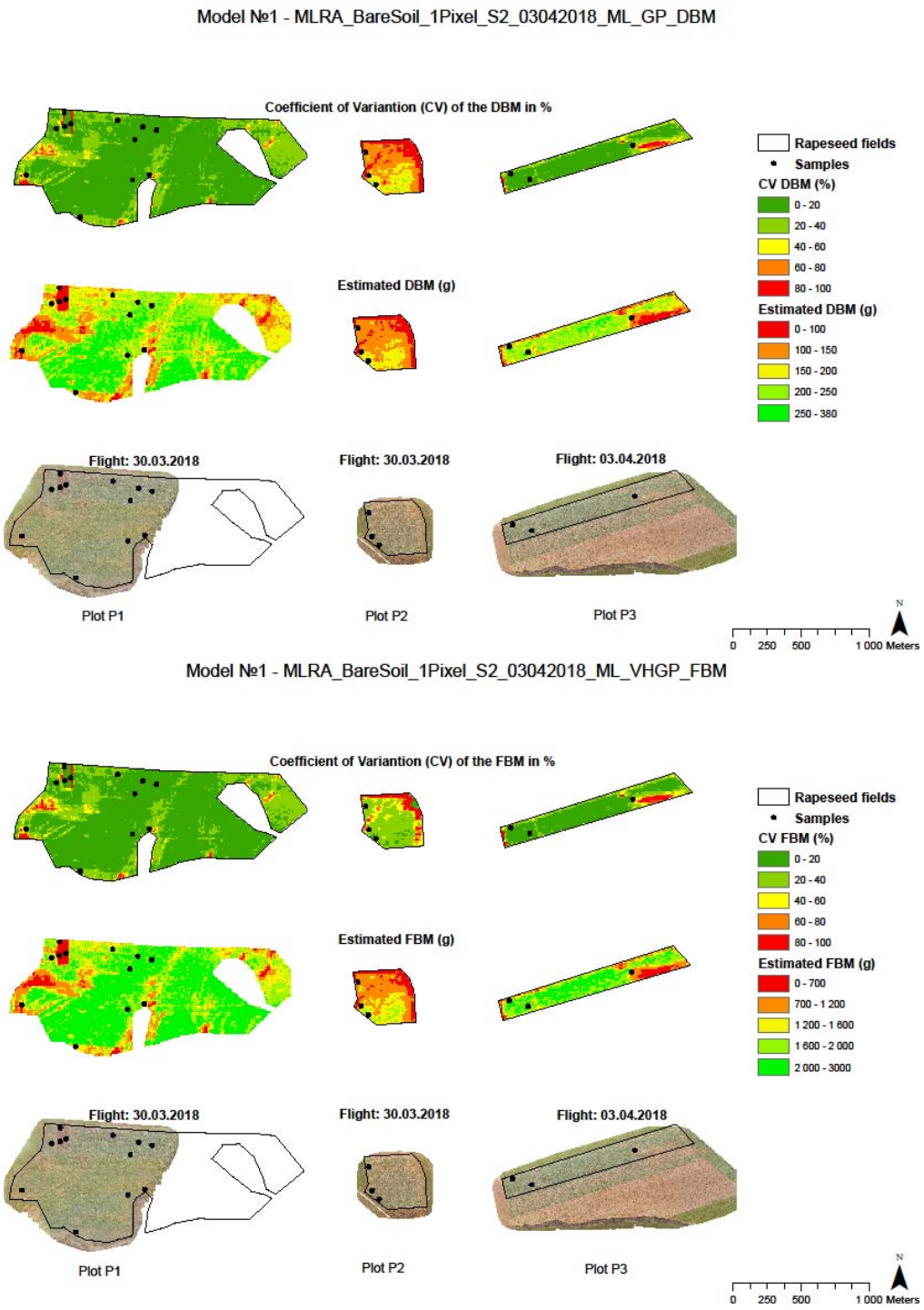

Fig. 2. Estimated Fresh and Dry biomass and Coefficient of Variance (CV) for the models with highest R² Plot

Conclusion

Both, parametric and non-parametric, models perform well for the period after winter when there is much more biomass and the bare soil is less visible than before winter. Nothing is gained by averaging the closest 9 pixels to the sample location from the remote sensing image and adding bare soil samples in the model increase its performance. Even if according to the NRMSE the models that exclude the outlier perform better than the ones with the outlier, when applied to the remote sensing maps they give poorer results. This is probably because even if the outlier stands out compare to the other samples, it was not an isolated event in two of the three studied plots. By including the outlier into the model calibration, it gives better representatively of the actual field data. The best results are with the GPR models and it is in accordance with study of Caicedo *et al.*, 2014.

Acknowledgment

This project is supported by Research Grant Award № ДФНП-17-43/26.07.2017 from the Bulgarian Academy of Sciences and it is done with the collaboration of Prof. E. Rumenina.

References:

1. Bartoszek, K. (2014) 'Usefulness of MODIS data for assessment of the growth and development of winter oilseed rape', *Zemdirbyste-Agriculture*, 101(4), pp. 445–452. doi: 10.13080/z-a.2014.101.057.
2. Caicedo, J. P. R., Verrelst, J., Muñoz-marí, J., Moreno, J. and Camps-Valls, G. (2014) 'Toward a Semiautomatic Machine Learning Retrieval of Biophysical Parameters', *EEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING*, 7(4), pp. 1249–1259.
3. Cihlar, J., Dobson, M. C., Schmugge, T., Hoogeboom, P., Janse, A. R., Baret, F., Guyot, G., Le Toan, T. and Pampaloni, P. (1987) 'Procedures for the description of agricultural crops and soils in optical and microwave remote sensing studies', *International Journal of Remote Sensing*, 8(3), pp. 427–439. doi: 10.1080/01431168708948651.
4. Clevers, J. G. P. W. and Gitelson, A. A. (2013) 'Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3', *International Journal of Applied Earth Observations and Geoinformation*. Elsevier B.V., 23, pp. 344–351. doi: 10.1016/j.jag.2012.10.008.
5. Fang, S., Tang, W., Peng, Y., Gong, Y., Dai, C., Chai, R. and Liu, K. (2016) 'Remote Estimation of Vegetation Fraction and Flower Fraction in Oilseed Rape with Unmanned Aerial Vehicle Data', *Remote Sensing*, 8(416), pp. 1–19. doi: 10.3390/rs8050416.
6. Han, J., Wei, C., Chen, Y., Liu, W., Song, P., Zhang, D., Wang, A., Song, X., Wang, X. and Huang, J. (2017) 'Mapping Above-Ground Biomass of Winter Oilseed Rape Using High Spatial Resolution Satellite Data at Parcel Scale under Waterlogging Conditions', *Remote Sensing*, 9(3), p. 238. doi: 10.3390/rs9030238.
7. Hatfield, J. L. and Prueger, J. H. (2010) 'Value of using different vegetative indices to quantify agricultural crop characteristics at different growth stages under varying management practices', *Remote Sensing*, 2(2), pp. 562–578. doi: 10.3390/rs2020562.
8. Piekarczyk, J., Wójtowicz, M. and Wójtowicz, A. (2006) 'ESTIMATION OF AGRONOMIC PARAMETERS OF WINTER OILSEED RAPE FROM FIELD REFLECTANCE DATA', *Acta Agrophysica*, 8(1), pp. 205–218.
9. Richter, K., Atzberger, C., Hank, T. B. and Mauser, W. (2012) 'Derivation of biophysical variables from Earth observation data : validation and statistical measures', *Journal of Applied Remote Sensing*, 6(1), pp. 1–24. doi: 10.1117/1.JRS.6.063557.
10. Rogerson, P. A. (2001) *Statistical method for Geography*. First edit. London: Sage Publications.
11. Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J. P., Veroustraete, F., Clevers, J. G. P. W. and Moreno, J. (2015) 'Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties - A review', *ISPRS Journal of Photogrammetry and Remote Sensing*. International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), 108, pp. 273–290. doi: 10.1016/j.isprsjprs.2015.05.005.
12. Weber, Bleiholder and Lancashire (1991) 'Échelle BBCH des stades phénologiques du colza (Brassica napus L. ssp napus)'.