FINAL REPORT

MSc. Internship at the Spanish Council for Scientific Research

TOMÁS ROQUETTE DA CRUZ TENREIRO

MAY 2018

FINAL REPORT

MSc. Internship at the Spanish Council for Scientific Research

TOMÁS FERREIRA ROQUETTE DA CRUZ TENREIRO

Plant Sciences MSc. Internship Report

Crop Systems Analysis

CSA-70424

May 2018

920623239120

Supervisors: Dr. Ing. José António Jimenez-Berni (CSIC)

Examiner: Dr. Ing. Lammert Bastiaans (WUR)
Dr. Ing. João Vasco Silva (WUR)

Acknowledgements

First, I would like to mention my gratitude to Professor Elías Fereres Castiel (University of Córdoba & CSIC) for the given opportunity. It was a pleasure to cooperate during the last months as an intern with the department of agronomy of the Institute for Sustainable Agriculture (IAS) in Córdoba, a member institution of the Spanish National Council for Scientific Research (CSIC). The opportunity to conduct my MSc. Internship at CSIC-IAS was an extremely rich experience that added considerable value to my formation as an agronomist. The engineering-solution-driven perspective of the department, resulting from a serious agronomic approach that is followed in the context of Mediterranean scientific and consultancy projects, opened my perspective towards new ideas and ambitions.

It is equally important to thank Dr. José António Jimenez-Berni for receiving and integrating me in his projects. It was very inspiring and positively challenging to work side by side with a “high-tech” agronomy researcher, who introduced me several innovative methodologies and tools in the context of remote sensing, phenomics and data science. It was a pleasure to work with Dr. Berni and share ideas for future projects, always under an easy going and flexible atmosphere. A special thanks to Dr. José Alfonso Gómez Calero as well, for his collaboration and daily welcoming attitude in the laboratory, that resulted into a motivating working environment and constructive discussions about potential future professional steps.

Last but not least, I would like to mention the importance of Wageningen University supervision as well. A special thanks to Dr. Joao Vasco Silva for “keeping it agronomic”, by trying constantly to discuss the results of our work and the consequent implications under an agronomic perspective. Crop science deserves to be properly addressed in the context of big data and machine learning and Joao has always kept it clear! Thanks to Dr. Lammert Bastiaans for being involved as a supervisor/examiner, mostly during the last stage of this process, by adding his perspective on the development of the writing material and discussions on the project applicability to the agronomic study and management of cover crops.

To conclude, it is still important to thank the European Spatial Agency (ESA) for providing satellite Deimos-2 data to this project.

To all involved, muchas gracias, dank u wel, and obrigado!

Tomás Roquette Tenreiro

Wageningen, May 2018

Abstract

The current report aims to communicate the main purposes and outcomes of an academic internship conducted at CSIC-IAS in Córdoba, Spain, in the context of the MSc. degree program of Plant Sciences (specialization on Crop Science) at the University of Wageningen. The internship was supervised by both the University chair group of Crop Systems Analysis (CSA) and the department of Agronomy of CSIC-IAS, where the internship took place. The internship was focused on the development of scientific skills applied to agronomic research in the field of remote sensing, data analysis and machine learning. The project titled “*Mapping cover crop dynamics in Mediterranean perennial cropping systems through remote sensing and machine learning methods*”, aimed primarily to develop methodologies of big data analysis applied to the case of cover crop (CC) dynamics in perennial systems such as olive orchards. A model capable to map temporal variability of CC dynamics in local olive orchards was developed and validated through both biological imagery analysis software and field observations. The *cover crop mapper* model (CCMAP) has revealed satisfactory potential to be used not only at landscape scale (variability expressed at olive field level), but also at crop level (variability expressed with pixels of 10x10m). The temporal resolution of this model depends on the satellite imagery resolution used (5-10 days in cloud-free conditions for the case of Sentinel-2). The CCMAP was calibrated and tested with satellite data from two European spatial missions (Sentinel-2 and Deimos-2), accessed through the approval of a project proposal presented to the European Spatial Agency (ESA). Pixel classification has shown higher accuracy when distinguishing between high cover crop density (> 60% of soil covered by herbaceous species) and low-medium density (< 50%), than between low (<20%) and medium density ([20;50]%). A big data analysis conducted at field level has revealed a negative response of CC development optimal benchmark to field slope, which calls eventually for adjustments on political speech and conservation agriculture targets established for steeper plots. An exploratory exercise was conducted in order to estimate the relative effect of different fractions of vegetation cover on potential rates of soil erosion. The annual cover management factor of the Revised Universal Soil Loss equation (RUSLE) was calibrated for three different levels of cover crop fraction, measured by CCMAP, and relative changes of annual soil loss were predicted. We believe that CCMAP can be a powerful tool for both policy making and landscape planning as well as for cover crop or weeds integrated and precision management at crop level. However, further validation measurements are needed, and we believe that CCMAP accuracy can be substantially improved by atmospheric and topographic correction of the satellite imagery used for calibration.

Keywords: *Agronomy; Conservation agriculture; Mediterranean cropping systems; Olive orchards; Cover crops; Remote sensing; GIS; Satellite imagery; Photo sensing; European spatial agency; Sentinel-2; Deimos-2; Image-J; Google Earth Engine; Artificial intelligence; Machine learning; Big data; Geo-statistics; Digital ecology; Precision agriculture; Crop modelling; Landscape planning.*

List of Acronyms

BI	Bare soil index
BR	Bayesian ridge
CC	Cover crop
CCMAP	Cover crop mapper model
ESA	European Spatial Agency
EVI	Enhanced vegetation index
fCC	Fraction of cover crop
FGCC_{summer}	Fraction of green canopy cover (measured in summer assuming total absence of CC)
FGCC_t	Fraction of green canopy cover (measured at a certain date)
GEE	Google Earth Engine
MLR	Multi-linear regression
NDVI	Normalized difference vegetation index
NIR	Near infra-red
PLS	Partial least squares
PNOA	Plan Nacional de Ortofotografía Aérea
RF	Random forest
RUSLE	Revised Universal Soil Loss Equation
SWIR	Short-wavelength infrared

List of figures

Fig 1. Annual distribution of ET0 (mm) and annual cumulative precipitation (mm)	4
Fig 2. Annual evolution of NDVI time-series scheme.....	5
Fig 3. Map of study zones	6
Fig 4. Pan-sharpening process	7
Fig 5. PNOA binary classification	12
Fig 6. Deimos-2 binary classification	13
Fig 7. Histogram of FGCC by hand.....	14
Fig 8. The satellite effect (from Deimos-2 to Sentinel-2)	14
Fig 9. Radiation reflectance spectrum correspondent to each satellite band.....	14
Fig 10. R-G correlations to Y threshold.....	15
Fig 11. Modeled FGCC vs. predicted FGCC.....	16
Fig 12. Graphical scheme of potential ground cover fraction benchmarking	16
Fig 13. Correlations between FGCC and vegetation indexes	17
Fig 14. Modeled FGCC vs. predicted FGCC from vegetation indexes.....	18
Fig 15. Machine learning algorithms performance	19
Fig 16. Temporal mapping of FGCC	20
Fig 17. Mapping FGCC at different scales	21
Fig 18. Map FGCC at crop field level	21
Fig 19. Field validation stage pictures	22
Fig 20. Box-plots of validation stage.....	23
Fig 21. CC time-series – cover crop growth curves.....	24
Fig 22. Temporal analysis of soil protection level	24
Fig 23. The effect of olive plant density x CC on soil protection	25
Fig 24. Benchmarking fCC as a function of field slope.....	26
Fig 25. Scheme of the proposed project (CERSUL case-study)	34

Table of contents

Acknowledgements	iv
Abstract	vi
List of Acronyms	viii
List of figures	x
1. Introduction	1
1.1 Agronomic context of CC management in Andalusian olive orchards	1
1.2 Artificial intelligence, machine learning and precision agriculture	2
1.3 How was machine learning used in the current analysis?	3
2. Materials & Methods	4
2.1 Agro-ecological conditions	4
2.2 Data collection & monitoring of cover crop development	5
2.3 Biological image analysis	7
2.3.1 Estimation of FGCC	7
2.3.2 Validation of radiation reflectance spectrum & satellite effect	8
2.4 Model 1 vs. Model 2	8
2.5 Field validation	10
2.6 Time series of cover crop fraction	10
2.7 Calibration of RUSLE parameter-C	10
2.8 Other agro-ecological drivers of cover crop development	11
3. Results	12
3.1 Binary classification and satellite imagery	12
3.2 Radiation reflectance spectrum & satellite effect	14
3.3 Model 1: from RGB to Y threshold	15
3.4 Model 2: from satellite bands to FGCC	17
3.5 Field observations & validation	20
3.6 Time series of cover crop fraction	24
3.7 RUSLE calibration	25
3.8 The effect of field slope and area on CC development	26
4. Discussion	27
4.1 Model 1	27
4.2 Model 2	28

4.3 Field validation.....	29
4.4 Soil conservation through the use of cover crops.....	30
5. Potential application of CCMAP	32
6. Conclusion	36
References.....	37
Appendix.....	41

1. Introduction

1.1 Agronomic context of CC management in Andalusian olive orchards

About 1.5 Mha of olive orchards are found in the southern Spanish region of Andalusia, representing over 15% of the world olive surface (Guzmán, 2014). The socio-economic impact of such cropping systems is strongly visible, once these systems have been commercially managed since the 3rd century AD (Gomez et al., 2014). The local intensification of olive production systems that has been observed in the last century has resulted into increasing rates of soil and water erosion at crop level (Gomez et al., 2014; Gomez et al., 2011). Actually, some of the most critical rates of soil erosion in Mediterranean agriculture have been found in the local steep slopes of olive orchards and vineyards, where soil is frequently tilled to avoid crop-weeds competition (Laguna & Giráldez, 1990).

Regarding the sustainability of modern olive plantations, conservation agriculture has been proposing alternative management strategies such as the use of inter-row CC (sown or indigenous) during the period of lowest evaporative demand, with an effective control through herbicides or eventually tillage in spring, in order to avoid significant inter-specific competition during the lowest water availability months (Gomez et al., 2011; Gomez et al., 2009; Romero et al., 2007; Hernández et al., 2005). The main indigenous species used as CC in local olive orchards are the annuals gramineae *Avena sterilis* L., *Hordeum murinum* L., *Bromus madritensis* L., *Bromus rubens* L., *Phalaris paradoxa* L. and *Lolium rigidum* Gaudin (Cuenca & Aguilar, 1997; IFAPA, 2016). On the other hand, the use of sown species (such as the legumes *Vicia sativa* L. and *Trifolium* for biological N-fixation, the early-inflorescence gramineae *Brachypodium* for their self-seeding aptitude or even the common *Brassica napus* L. for its suppressive capacity of soil diseases (Vig et al., 2009; Garibaldi et al., 2009)) is also a potential strategy of ecological weed management applicable to the local systems (IFAPA, 2016).

However, despite the efforts of policy making (Knowler & Bradshaw, 2017; Piorr, 2003) and scientific research, the use of CC has not been fully adopted yet and a high variability regarding the fraction of ground cover is still observed in the region with actual mean soil erosion rates varying between 61 and 184 t ha⁻¹ year⁻¹ (Gomez et al., 2014). Therefore, a better understanding of the main drivers defining such variability seems to be required and the development of an up-scaled methodology of analysis can be a determinant step towards improved modelling and management of CC in olive orchards.

In perennial cropping systems, the analysis of CC reflectance signals through remote sensing must be disaggregated from the signals of other elements (as trees or bare soil surface), which vary largely according to the phenological dynamics of the crop species (relatively constant on evergreen species as *Olea europea* L., but fully dynamic on deciduous species as *Vitis vinifera* L. for instance). In this sense, considering the level of detail required, the use of remote sensing through satellite imagery, such as Landsat-8 or Sentinel-2, to link vegetation indexes to land cover fractions in homogeneous canopies (Sibanda et al., 2015), might be insufficient for a more detailed calibration of Phenotyping modelling due to lower spatial resolution (10-30 m). The large variability that is observed in terms of crop spatial arrangement and plant density along the region delivers additional challenges for a study of this kind, and therefore, higher spatial resolution images are required to complement the satellite imagery analysis (Hall et al., 2002).

By following this idea, we assume the need of different spatial resolution imagery for calibration purposes. The integration of several satellite databases at regional level results in an enormous amount of data that will require artificial intelligence

technologies to be properly managed. Regarding such an assumption, the main hypothesis is presented: *Modern artificial intelligence tools will substantially add predictive capacity to a model of this kind, in comparison to more fundamental statistics*. The development of such technology, in this context promoted by fundamental concepts of precision agriculture, requires the development of machine learning algorithms and other data science tools.

1.2 Artificial intelligence, machine learning and precision agriculture

Nowadays, artificial intelligence (AI) and machine learning are two popular terms widely mentioned and referred despite all the unclarity that still exists regarding what exactly they mean and how they function. A common misunderstanding that is frequently found is related to the existing differences between both. Although these two terms are often used as the same, AI is a much broader concept while machine learning is a specific application of it (Nasrabadi, 2007; Michalski et al., 2013). Artificial intelligence is mostly a fundamental concept that represents the capacity to turn technologies into automated systems capable to take decisions. In agriculture, the use of such systems has gained additional relevance with recent trends to step towards precision management and sustainable use of resources. Data collecting sensors and mining tools, decision supporting software, operational systems linked to GIS and others are some of the most common examples of how AI is serving agriculture to achieve higher efficiency and performance (Jones, 1993; Aitkenhead et al., 2003; Russ & Brenning, 2010; Wolfert et al., 2017).

On the other hand, to understand machine learning we must step into deeper levels of artificial intelligence by moving from the fundamental concept itself to an applied dimension of it. The entire concept is based on the assumption that as we learn from information, the “machines we build” may also do (Michalski et al., 2013). Information in the form of quantitative data is therefore the key for developing algorithms capable to learn from datasets. When the learning subsets of data are labeled with absolute conditions for decision making, the algorithms try to link clusters of analysed data to validation variables by following what is called supervised learning (Lantz, 2015). Supervised learning algorithms can operate in both a discrete or continuous level of prediction. Examples of supervised learning used for continuous prediction are known in the forms of regression analyses, when models are built based on the assumption that predictions will follow the same relations obtained from historical observations. But if a classification is what we aim for, supervised learning must follow other kind of reasoning by trying to cluster units of observation into multiple space dimensions according to relations between data and eventually by applying vector transformations in order to disaggregate what apparently cannot be distinguished (Noble, 2006). Examples of powerful supervised classifiers are support vector machines (SVM), Kernel functions and Naïve Bayes (Lantz, 2015). Several of the algorithms used for discrete classification, can also be used under a continuous context of prediction. Random forests are a good example, by constructing multiple decision trees at training level while outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees (Denisko & Hoffman, 2018).

The main driver of precision agriculture is data (Cox, 2002). Without data, there is absolutely no scope for precision management of farming systems. The main concept of “How much? When and where?” is directly related to the idea that spatial and temporal quantification of systems performance indicators enables more efficient solutions. Access to large cloud based historical data sets in combination with real time data supply are driving what is today known as the 4.0 industrial revolution where agriculture is being largely affected (Schwab, 2017; Wolfert et al., 2017). The digitalization of our agro-ecological systems has delivered enormous amounts of spatial and temporal specific data that reveals high potential for the development of automated systems of analysis. However, dealing with large amounts of data requires the

use AI to upscale the analysis efficiently (Wolfert et al., 2017). Precision agriculture is now stepping towards real time data supported systems where the use of computing technologies to collect, treat and analyse data is naturally essential. It is precisely here, where AI, machine learning and precision agriculture are co-existing.

1.3 How was machine learning used in the current analysis?

Looking at the specific case of CC management in Mediterranean olive orchards, we believe that the development of a smart measuring tool of spatial and temporal development of CC at larger scales will promote more integrated CC management (supported by real time data such as vegetation indexes, soil and weather data).

We get into a scale of “Big data” by accessing the database of the Spanish Ministry of Agriculture where about 750.000 olive fields (observation units) can be accessed through GIS. This amount multiplied by remotely collected data from satellite radiation bands and weather stations leads us into millions of quantifiable relations. Supervised learning might serve this analysis in two different ways: 1) classification algorithms for classifying vegetation pixels, the existence or absence of CC and levels of plant density; 2) regression analysis for determining relations between remotely measured variables and vegetation structure (examples are linear and multi-linear regressions, support vector machine regressions, random forest regressions and partial least square regressions). However, the use of large amounts of data might lead to what is commonly defined as over fitting, when correlations between input variables and output exist in the absence of any biological meaning or theoretical reasoning (Hawkins, 2004). In order to avoid over fitting it is important to 1) minimize the amount of predictive variables which increases the robustness of the model and its comprehension/use since solutions should not be provided in “black boxes”; and 2) integrate validation subsets from other methodologies of analysis (as manual imagery analysis and visual classification through field visiting).

2. Materials & Methods

2.1 Agro-ecological conditions

The study was conducted in the provinces of Córdoba (13771 km²) and Jaén (13496 km²), south of Spain. According to the Köppen Climate Classification system, the study area (252 km²) is characterized by the type Csa, commonly described as Mediterranean Climate. The period of study corresponds to the agricultural years of 2015-16 and 2016-17.

Regarding evapotranspiration (ET₀), both years have followed similar patterns with Córdoba revealing the highest values during summer (see figure 1A). However, differences between both provinces precipitation were larger. In comparison to Córdoba, Jaén annual rainfall was on average 45% less in 2015/16 and 35% less in 2016/17 (IFAPA, 2018). Average values of annual cumulative precipitation during the period of study were about 535 mm in Córdoba and 320 mm in Jaén (see figure 1B). No significant differences regarding annual cumulative precipitation were observed in Córdoba between both years but in Jaén the second year has significantly cumulated more rainfall than the first. A considerable concentration of rainfall was still observed at the end of April – beginning of May in both regions, mainly in Córdoba where an amount of 113 mm was observed in a three-week period, representing 20.5% of annual rainfall of the same year (IFAPA, 2018).

Fig 1 – A) Annual distribution of ET₀ (mm) and B) cumulative precipitation (mm) in the provinces of Córdoba and Jaén during the period of study (2015-2017). Data obtained from IFAPA (2018).

Local olive orchards are mostly found in cambisols, vertisols and regosols which typically reveal large differences in terms of texture, water holding capacity and exchange cationic capacity (ECC), resulting naturally into an expected large spatial variability regarding vegetation potential growth. The linkage between water availability (mainly driven by ET₀ and precipitation because the majority of plots are rainfed) and soil type will theoretically determine the development of CC. By mapping cover crop dynamics we aim also to understand if significant differences regarding CC development are observed between different combinations of soil type and precipitation zone (Córdoba vs. Jaén).

2.2 Data collection & monitoring of cover crop development

Data has been collected from three sources: 1) the SIGPAC platform (Spanish Ministry of Agriculture and Environment), 2) Google Earth Engine (satellite Sentinel-2) and 3) from the European Spatial Agency - ESA (higher resolution satellite imagery – satellite Deimos-2). SIGPAC is an open source platform, free of costs that can be consulted and downloaded at the Spanish Agricultural Ministry website (MAGRAMA). The access to Deimos-2 archived data was done through the approval of a project proposal that we submitted to the European Spatial Agency and Google Earth Engine (GEE) was accessed through Google Earth Engine Developers, a Google programming platform operated in JavaScript.

The quantification of spatial and temporal dynamics of cover crops was conducted using a key diagnostic variable that represents the fraction of soil covered by green vegetation (Patrignani & Ochsner, 2015). This variable was called fractional green canopy cover (FGCC) and represents the fraction of soil covered by photosynthetic active plants, when root systems are actually functioning, in order to exclude dry plant residues from this quantification. We started by obtaining the normalized difference vegetation index (NDVI) summer series from GEE and calculating mean values in order to represent FGCC without cover crops. We assumed that during summer the fraction of bare soil is maximum since ground vegetation is mechanically or chemically controlled before the highest evaporative demand period from June to September (see figure 2). Therefore, the FGCC of olive trees can be determined by measuring FGCC from summer imagery.

In this sense, the cover crop fraction (fCC) at any date can be estimated by equation 1. The idea behind this linear equation is based on the assumption that the evolution of fCC will be defined by a constant ($FGCC_{summer}$) and a time variable ($FGCC_t$), that follows typical curves of plant growth defined by log/exponential and stationary phases (Heuvelink & Dorais, 2005; Kruidhof et al., 2008), which can be smoothed by interpolation and adjusted to sigmoid signals (see figure 2). In this sense, the quantification of CC development might be obtained from the difference between what is real time remotely measured with satellite (or drone imagery) and the fraction of olive trees measured in summer and assumed as constant for mature plants, in the absence of pruning events or yield-reducing factors (Zarco-Tejada et al., 2004; Mezghani et al., 2016).

$$fCC = FGCC_t - FGCC_{summer} \quad (1)$$

Fig 2. Annual evolution of NDVI scheme. Blue dots represent the measured values of a random olive field from Sentinel-2 with GEE and red line-dots represent values adjusted to a sigmoid signal that represents the continuous evolution of NDVI time series.

The access to archived data of the Deimos-2 satellite mission was approved by the European Spatial Agency (see map of zones requested to ESA in figure 3).

Fig 3. Schematic figure representing map of zones requested to the European spatial agency (Deimos-2). Black features represent all olive plots (>2 and <50ha) found in the provinces of Córdoba and Jaén.

In total, six different zones were requested (see more detailed information in table 1). The criteria behind imagery request were: 1) priority given to dates between November and May in order to complement PNOA imagery (obtained from summer observations), 2) the lowest percentage of clouds, 3) imagery containing the maximum amount of olive plots in order to maximize the amount of observation units and 4) assuming that temporal variability might be predicted from spatial variability, multiple zones were selected despite the overlapping of dates on some images.

Table 1. Technical description of obtained Deimos-2 imagery

Date and Acquisition Time (UTC)	Area(Sq.Km)	Clouds (%)	Polygon ID	required EPSG3042
2015-11-27 ; 10:43:22	43	0	4	
2015-11-27 ; 10:43:23	29	0	6	
2015-11-27 ; 10:43:25	48	0	5	
2016-03-28 ; 10:46:56	29	0	2	
2016-03-28 ; 10:46:58	61	1	3	
2017-05-30 ; 11:04:02	42	0	1	

We worked with pan sharpened satellite imagery (see figure 4). Pan sharpening is a common technique used to merge higher-resolution panchromatic data with lower-resolution multispectral data (Vivone et al, 2015). Maximizing features resolution is particularly important in the context of our image binary classification methodology since it adds precision to the identification of different kinds of features (olive canopies, herbaceous vegetation and vegetation observed in gully erosion zones) helping to classify as zero (non-vegetated pixel) or one (vegetated pixel).

Fig 4. Descriptive figure of multispectral false colour imagery, panchromatic imagery and pan sharpened imagery; obtained from Saeedi & Faez (2011).

2.3 Biological-image analysis

2.3.1 Estimation of FGCC

Biological-image analysis was conducted in order to estimate both $FGCC_{summer}$ and $FGCC_t$. Two different types of imagery were used for each parameter: 1) PNOA imagery for $FGCC_{summer}$ and 2) Deimos-2 imagery for $FGCC_t$.

An algorithm capable to analyse aerial pictures of vegetation was developed with Python. The algorithm estimates colour thresholds as a function of different colour scales (RGB and YUV). By defining colour ranges on which pixels are classified as vegetation, the determination of FGCC was possible through image binary classification: vegetation pixels were classified with the value 1 and no-vegetation pixels with the value 0. The automated binary classification algorithm developed with Python was compared with manual tests where colour thresholds were adjusted by hand with Fiji Image-J processing software (Schindelin et al., 2012).

The distribution of colour follows in the large majority of cases a histogram that tends to express different patterns depending on vegetation density, structure and other field conditions. Different colour patterns, apparently determined by several agro-ecological conditions, were observed. Spatial gradients of soil water content, different soil types, orientation and shaded zones are likely to be associated with this variability (Patrignani & Ochsner, 2015). The fact that not only the distribution of colour values change from plot to plot, but also the pattern of the histogram, did not allow us to predict accurately colour thresholds based on differential analysis of the colour histogram. In this sense, we have decided to explore linear correlations between RGB colour mean values and the colour threshold that optimizes binary classification by hand.

Several trials were conducted using different classification methods with Fiji ImageJ software (Schindelin et al., 2012), first in the province of Córdoba with PNOA images and later with Deimos-2 imagery for the province of Jaen as well. The classification method that has revealed the best results with Fiji ImageJ software was the *Max Entropy* (Gull & Skilling, 1984). The validation sets were built by hand, following visual adjustments of colour thresholds through this method.

In total, 300 olive fields were classified (50 for each zone) through hand adjustment of brightness (Y' threshold) with Fiji Image-J (with the *Max Entropy* method selected) and the mean values of both the Red and Green bands were measured for the same images. Colour band values were plotted against correspondent Y' threshold values, adjusted by hand, regression analyses and ANOVA were conducted in order to estimate correlations and correspondent statistical significance (p-value <0.05).

2.3.2 Validation of radiation reflectance spectrum & satellite effect

It is important to validate the biological meaning of radiation reflectance spectrum measured with satellite. In order to perform this, we collected from Sentinel-2 archived database, the summer time series of the 12 bands individually related to the same dates of PNOA imagery. Linear regressions and statistical significance tests were conducted for each of the 12 Sentinel-2 bands regarding their correlation with FGCC_{summer} (N=1600).

We have also quantified, through regression analysis and ANOVA, the correlation between NDVI measured by Sentinel-2 and Deimos-2 in order to determine conversion factors when determining FGCC from different satellites (the satellite effect).

2.4 Model 1 vs. model 2

In order to upscale the analysis to the level of province, we had to generate training and testing datasets from raw imagery. Model 1 was the initial model, developed to estimate FGCC based on the correlations found between Y' thresholds and R-G-B colour bands. The 300 olive fields labelled by hand were used to validate this model. Predicted FGCC values were plotted against measured FGCC and correlation accuracy considered according to the R², RMSE and p-value. Model 1 allowed us to increase substantially the amount of observation units, in order to explore the prediction capacity of machine learning methods in comparison to fundamental regression analyses obtained from standard vegetation indexes. Model 2 was then developed from datasets generated with model 1.

First, three different vegetation indexes were compared: 1) the normalized difference vegetation index (NDVI), 2) the enhanced vegetation index (EVI) and 3) the bare soil index (BI). We compared the prediction accuracy of these three indexes derived from Sentinel-2 data.

Both the normalized difference vegetation index (NDVI) and the enhanced vegetation index (EVI) follow the idea that live green plants absorb solar radiation in the photosynthetic active radiation (PAR) spectral region, which is used as a source of energy in the process of photosynthesis (Gates, 1980; Weier & Herring, 2000). However, at wavelengths longer than about 0.7 µm the photon energy is not large enough to synthesize organic molecules and, therefore, the absorption at these wavelengths would only lead to overheating of plant tissues (Gates, 1980). As a feedback mechanism, plants have the capacity to filter light. The chlorophyll pigment absorbs visible light (from 0.4 to 0.7 µm, where red is included) while reflecting near-infra light (from 0.7 to 1.1 µm). Following this conceptual idea NDVI can be determined through equation 2:

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (2)$$

However, the use of NDVI value turns out to be sensitive to several perturbing factors including atmospheric effects (in particular with respect to water vapour and aerosols), presence of clouds, soil effects (since soils tend to get darker when wet,

and therefore, depending on plant density, it might affect NDVI more than vegetation properties adding noise to the interpretation of results), anisotropic effects (different surfaces reflect light differently) and the sensor effect (or in this case, the satellite effect, since the correlations between vegetation indexes and FGCC are satellite specific). We have controlled the presence of clouds through the use of cloud masks functions in Google Earth Engine (GEE) and by getting satellite images with minimum percentage of clouds (all below 3%). However, atmospheric correction of imagery for the presence of smoke, and aerosols was not conducted for simplicity reasons). The anisotropic effect of human-made surfaces (e.g. roads, farm buildings, water dams, etc.) was taken into account by excluding all “human made features” from the analysed polygons.

In order to deal with constraints on modelling FGCC for higher plant density zones, we have also explored correlations with EVI. The enhanced vegetation index (EVI) is an 'optimized' vegetation index that has been widely used due to its ability to enhance the vegetation signal with improved sensitivity in higher biomass regions and through a de-coupling of the canopy background signal and a reduction in atmosphere noise (Kim et al., 2010). This index can be determined through equation 3 (where all constants are satellite specific, in this case for Sentinel-2):

$$EVI = \frac{2.5 (NIR-Red)}{(NIR+6*Red-7.5*Blue+1)} \quad (3)$$

The Bare Soil Index (BI) was also computed in order to explore differences between agricultural and non-agricultural vegetation. Bare soil, fallow lands, and non-agricultural vegetation tend to be enhanced when using the BI index (Jamalabad, 2004). This index integrates four different satellite bands through equation 4:

$$BI = \frac{(SWIR+Red)-(Blue+NIR)}{(SWIR+Red)+(Blue+NIR)} + 1 \quad (4)$$

Regression analyses and ANOVA were conducted to explore correlations between the each index and three different parameters: 1) FGCC measured by hand at a certain date (t), corresponding to the specific satellite image date; 2) FGCC at summer, specific PNOA orthophoto date; and 3) the fraction of cover crop (fCC) assumed to be the difference of FGCC registered at both previous dates ($FGCC_t - FGCC_{PNOA}$).

Additionally, four different supervised machine learning algorithms were developed with Python using the software platform *Anaconda* (Cano, 2015). Predicted FGCC was plotted against measured FGCC (N=1600, generated with model 1), and correlations were compared (considering the R^2 and RMSE values) for the four tested algorithms and with the correlations obtained from the previous indexes (NDVI, EVI and BI). The selected algorithms were: 1) a partial least squares regression (PLS), 2) a Bayesian ridge regression (BR), 3) a multi-linear regression (Linear) and 4) a random forest regression (RF). These algorithms took all satellite bands as “learning material” instead of only the ones used in the previous indexes calculation (see equation 2, 3 and 4), they operate at a higher dimensional space. After evaluation of statistical significance of each method (the four machine learning algorithms and the three indexes), the algorithm that best performed was used for calibration of model 2 (the cover crop mapper model - CCMAP).

2.5 Field validation

Field validation took place in several locations distributed in both the province of Córdoba and Jaén at the third week of April. A map of fCC was initially generated by model 2 (CCMAP), from Sentinel-2 imagery taken at the second week of April. We assumed no significant variations on cover crop fraction in a week-time period. The validation map was used to spot olive plots with different levels of fCC and olive plant density. Based on the validation map a route was designed. In total, the map was validated for 68 olive plots distributed around Córdoba, Castro del Río, Baena, Alcalá la Real, Almendinilla, Priego, Cabra, Lucena y Aguilar de la Frontera.

The method used for validation was based on visual observation and classification of fCC into four different levels: absence of cover crop, low (fCC < 20%), medium (fCC from 20 to 30%) and high density (fCC > 50%). Digital photographs were also taken, and voice descriptive comments recorded for further analysis at the office. For some plots we also tested the *Canopeo* application (Patrignani, 2015). However, this was done more under an exploratory way, not concisely, so we did not took a statistical analysis on the performance of the app or the differences between the app classification and our visual criteria. All measures, photos and comments were mapped according to the GPS coordinates measured by a smartphone.

Statistical analysis was conducted on the collected data. First, mean fCC values per unit of observation (clusters of pixels) were measured for the observed zones through zonal statistics in QGIS and box-plotted. Second, a probability distribution curve and a histogram were generated and the difference between each class (visually classified) mean values of fCC was statistically measured with the Tukey's honest significance test (HSD-test).

2.6 Time series of cover crop fraction

Annual time-series of discrete fCC temporal variability were estimated with model 2 for the agricultural year of 2016/17 (N=16500). Theoretical sigmoid curves of plant growth (here representing the fraction of ground cover by CC) were estimated through polynomial regression of fCC plotted against time. The mean evolution of fCC as a function of time (expressed as day of the year, DOY) was estimated for the three different classes of CC fraction considered in the field validation stage. The low CC density mean curve was defined as the polynomial regression of all fCC measurements below the 30th percentile. The medium CC density mean curve was estimated from all points between the 30th and the 60th percentile and the high CC density mean curve from all points above the 60th percentile.

The representation of different scenarios of CC development is important to model soil management parameters that are essential to define the relative risk of soil loss according to the revised universal soil loss equation (RUSLE).

2.7 Calibration of RUSLE parameter C

RUSLE is an erosion model designed to predict the long-time average annual soil loss, A (ton ha⁻¹ year⁻¹), carried by runoff from specific field slopes under specific crop and management systems, computed according to equation 5:

$$A = R K L S C P \text{ (ton/ha/year)} \quad (5)$$

where R ($\text{MJ mm ha}^{-1} \text{ h}^{-1} \text{ year}^{-1}$) is a rainfall erosivity factor; K ($\text{ton ha}^{-1} \text{ MJ}^{-1} \text{ mm}^{-1}$) is soil erodibility; L is the slope length factor; S is a slope gradient factor; C is a cover management factor; and P is a support practice factor. The last four factors are dimensionless.

The model was calibrated by Gomez et al. (2003) for local conditions but the estimation of C was based on simulated scenarios dependent on linear relations between annual maximum fCC and olive plant density. In order to understand how representative those scenarios are, we have re-calibrated the RUSLE C -parameter for the three classes of CC fraction observed, assuming the other five parameters as constants.

Statistical analysis was developed in order to quantify the relative impact of different classes of CC fraction on the modelled potential annual soil loss. The controlling factors considered were olive plot density and fCC. ANOVA was conducted to explore how the potential annual soil erosion rate relates to each controlling factor.

2.8 Other agro-ecological drivers of cover crop development

We know from the RUSLE equation that effective soil erosion mitigation in steeper plots is strongly dependent on soil management. Additionally, lower olive plant density is also commonly observed in steeper plantations, which empowers the argument that soil conservation strategies on steeper fields must be largely focused on cover crop management. However, it is important to understand if the proposed targets for certain topographic conditions are realistic. Should policy making propose fCC targets independently on the system structure? Here we highlighted slope and area.

Therefore, a big data analysis was conducted in order to benchmark potential fCC as a function of crop field slope and area ($N=16500$). Following the concepts addressed by Cade & Noon. (2003) and Tenreiro (2017), a frontier analysis through quantile regressions of fCC plotted against crop field area and slope was conducted.

3. Results

3.1 Binary classification of satellite imagery

In total 600 images were analysed by hand with Fiji Image-J, 300 from PNOA and 300 from Deimos-2. The binary classification of PNOA imagery was quite straight forward since summer images have shown one single peak on the brightness histograms measured by Image-J for each olive field, in comparison to multiple peaks for the case of Deimos imagery.

For PNOA imagery, the single Y' histogram peak corresponded to the level of brightness intensity (in the colour scale of YUV) that, when selected through the *Max Entropy* method, was capable to classify quite accurately all vegetation pixels (see figure 5 A and B). In this case, since we were analysing summer imagery (absence of herbaceous species in inter-rows), the vegetation pixels classified through the selection of this Y' histogram peak corresponded directly to the olive trees canopy (see figure 5 A and B). However, the existing colour gradients, eventually due to a large variability of olive health conditions, did not allow us to work with a constant pattern of Y' threshold selection, which means that for some plots, the classification of olive canopies was better under slightly lower Y' values than the peak, but sometimes under higher (see the range spotted with a red circle in figure 5 C).

Fig 5 – A) A summer satellite image of a high density olive plot (>400 trees/ha). B) A summer satellite image of a low density olive plot (< 100 trees/ha). Pan sharpened images on the left, binary images on the right. Data obtained from PNOA. C) Example of typical Y' histogram measured from PNOA imagery. Red circle spots the range of best performing thresholds.

The variability of this relation, did not allow us to base or method on an automated differential analysis of the histogram, since the gradient of the best performing threshold was negative for some cases, positive for others, and eventually zero (for those where the maximum peak point was exactly the level of brightness that best performed). Considering this variability, the binary classification of PNOA has followed adjustments by hand of this threshold image by image. All values were recorded and a database was built.

For Deimos imagery, the existence of multiple peaks on the Y' histogram has led to additional struggles (see figure 6). Deimos images showed considerable gradients of colour related to three main factors: 1) soil colour variability, 2) presence/absence of cover crops and 3) olive trees canopy. Therefore, different histogram patterns were observed for different images (see figure 6 C). Some showing one single peak (mostly in cases of extremely high plant density, or for lower densities in the absence of cover crops simultaneously with a homogeneous soil colour), some with two peaks (for medium-high plant density, with non-uniform cover crops) and some with three peaks (when the three main factor were clearly visible and varying significantly).

Fig 6 – A) A winter satellite image of a high density olive plot (>400 trees/ha); pan sharpened image on the left, binary image on the right. B) A summer satellite image of a medium density olive plot (200 trees/ha) from PNOA above, and the correspondent Deimos-2 image below, taken at November (colour gradients leading to multiple peaks on the Y' histogram). Data obtained from Deimos-2. C) Example of two typical Y' histogram measured from Deimos imagery. Red circles spot the multiple peaks corresponding to different factors: soil colour gradients, cover crops and olive trees canopy.

For Deimos-2 imagery, the binary classification has followed adjustments by hand of the Y' threshold, image by image. All values were also recorded.

From these 600 images classified, corresponding to two images per olive plot, one from summer PNOA imagery and another from “winter” Deimos-2 imagery, the $FGCC_t$ values measured by hand followed a shape that adjusts to a log normal distribution (positive skew, see figure 7). It is still important to note that both FGCC below 10% or above 90% was observed on just a few plots (in total less than 2%).

Fig 7 – Histogram of FGCC by hand

3.2 Radiation reflectance spectrum & satellite effect

The satellite effect has shown to be represented by a linear regression, significant for both terms ($p\text{-value} < 0.05$) expressed as $NDVI_{sentinel} = (1.13NDVI_{Deimos}) - 0.116$ (see figure 8). By plotting band reflectance values measured by Sentinel-2 (N=1600) against FGCC, the following negative linear correlations were observed (see figure 9 and table 2). In general, all correlations were quite weak mainly for wavelengths close to the near infra-red (NIR) range, where theoretically, positive correlations would be even expected.

Fig 8 – The satellite effect

Fig 9 – Linear regression of Sentinel-2 bands (1-12) plotted against modelled FGCC (here called f_{cover}) [N=1600].

Table 2. Linear regression parameters, R^2 values of fitted lines and p-values with significant codes. Blue cells indicate the strongest correlations between satellite band and FGCC, and red cells indicate the weakest correlations

Sentinel-2 Bands		Linear regression ($y = ax + b$)			
		a	b	R-sq	p-value
B1	Aerosols	-4.65E-04	1.17	0.3	<2e-16 ***
B2	Blue	-3.92E-04	1.02	0.3	<2e-16 ***
B3	Green	-3.67E-04	1.02	0.4	<2e-16 ***
B4	Red	-2.97E-04	0.98	0.4	<2e-16 ***
B5	Red Edge 1	-2.78E-04	0.98	0.3	<2e-16 ***
B6	Red Edge 2	-1.86E-04	0.88	0.2	<2e-16 ***
B7	Red Edge 3	-1.41E-04	0.81	0.1	<2e-16 ***
B8	NIR	-1.36E-04	0.78	0.1	<2e-16 ***
B9	Water vapour	-3.40E-04	0.68	0.1	<2e-16 ***
B10	Cirrus	-5.83E-02	1.23	0.2	<2e-16 ***
B11	SWIR 1	-1.89E-04	1.03	0.3	<2e-16 ***
B12	SWIR 2	-8.95E-04	0.56	0.1	<2e-16 ***

Signif codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

3.3 Model 1: from RGB to Y' threshold

Considering the results of the ANOVA's conducted on the regressions between R-G-B and the Y' threshold manually adjusted with Fiji Image-J (see figure 10), we decided to work with regression A for PNOA imagery and D for Deimos imagery (see figure 10).

Fig 10 – Linear regressions between red (A and C) or green band (B and D) values and Y' threshold manually defined [N=300].

Model 1 was then developed based on the previous regressions in order to estimate the Y' threshold for more olive plots ($N=1600$) from both PNOA and Deimos imagery. First, the model estimated $FGCC_{summer}$ and $FGCC_t$ automatically for the same plots measured by hand. Measured $FGCC$ values were plotted against predicted values (see figure 11). Despite getting a linear relation of slope close to 1, statistical analysis revealed quite a low level of accuracy achieved by model 1 ($R^2 = 0.4$ and a non-significant intercept, $p\text{-value} > 0.05$, see figure 11).

Fig 11 – Modeled plotted against measured by hand FGCC from the Deimos validation subset [N=300].

The use of linear regressions based on color values to estimate the parameter Y' has revealed some limitations, this was visible when plotting $FGCC_{summer}$ against $FGCC_t$ from both the manually measured and the modeled validation subsets (see figure 12). We observed that model 1 concentrated the observation cloud in the case of the modeled set due to a “smoothing effect” of Y' on $FGCC$.

Fig 12 – Graphical scheme of potential ground cover fraction benchmarking. The filled black line 1:1 represents the minimum agronomic threshold below which observation units are assumed as outliers (or eventually juvenile olive plots due to considerable variations of trees canopy over time), assuming that $FGCC_t$ (Deimos) cannot be lower than $FGCC_{summer}$ (PNOA). Green and red filled lines represent the potential benchmark of ground cover fraction defined

by a linear regression of the 95th FGCC_t percentiles. The difference between FGCC_t and FGCC_{summer} represents fraction of cover crop (fCC).

The “smoothing effect” of predicted FGCC from Deimos imagery, in comparison to measured values by hand, is visible in the upper right zone of the observation cloud (see figure 11). The higher dispersion of points that is observed in that zone results from the fact that FGCC sensitivity to Y’ threshold adjustments increases with plant density.

3.4 Model 2: from satellite bands to FGCC

Despite the low R-sq values, positive logarithmic correlations between FGCC_t and both NDVI and EVI were found (see figure 13, A and D), revealing some limitations to predict FGCC on higher plant densities (as observed in model 1 stage). However, no significant correlations were found with BI. For simplification purposes, results are only shown for the first two indexes.

Fig 13 – Correlations found between NDVI and FGCC (A and B) and fCC (C); or EVI and FGCC (D and E) and fCC (F); at winter (t) and summer (PNOA) [N=143]. Results are only shown for plots with areas larger than 2 ha and smaller than 50 ha (N=300).

EVI has shown to be more effective to predict FGCC, due to higher R-sq values with linear regressions that “do not saturate” at higher levels of plant density (see figure 13, B and E). However, vertical dispersion of FGCC values turns these correlations to be only significant for one of the equation terms since p-values were larger than 0.05 for the regression intercept (see figure 13, B and E). Regarding fCC, the linear correlations found with NDVI or EVI were not quite well adjusted (R-sq < 0.3, see figure 13, C and F), despite being significant (p-value < 0.05).

We know from these results that there are clearly correlations between vegetation indexes and herbaceous species signals; however, such correlations seem to be quite limited for modeling accurately spatial and temporal development of cover crops in local olive orchards. It is still important to note that the lower R-sq values obtained for “winter period - t” might be due to the fact that multiple images were combined in the analysis of winter signals, adding variability, while one single image (from PNOA) was used for the summer period.

Predicted FGCC through the use of both NDVI and EVI were plotted against measured by hand (N=300) and R^2 , RSE and regression p-values were registered (see figure 14).

Fig 14 – measured FGCC plotted against predicted from NDVI and EVI [N=300].

In order to explore the added value of higher dimensional statistics, we analysed “new kinds of vegetation indexes” estimated through machine learning algorithms (see figure 15).

Fig 15 – measured FGCC plotted against predicted [N=1600]. Measured values were generated with model 1 and predicted values were generated through machine learning algorithms (model 2): A) Partial least squares regression (PLS); B) Bayesian ridge regression (BR); C) Multiple linear regression model (MLR); D) Random forest (RF); E) model coefficients for PLS, BR and MLR.

This implied a considerable increase in the number of observations to build both a training and a testing dataset large enough to apply the four selected algorithms. We up-scaled through model 1 our data to a higher level (N=1600). Per observation unit, the dimension of analysis increased from two satellite bands (in the case of NDVI) or three (in the case of EVI) to 12 bands, adding more "learning material to the machine".

3.5 Field observations & validation

Field validation stage started with some back office planning by mapping and interpreting the dynamics of estimated FGCC based on the MLR coefficients (see figure 15 E). We spotted extreme cases in order to discuss the effectiveness of our results and to approach the model under an agronomist perspective (an example can be seen in figure 16). Several maps of fCC at different dates were generated for the provinces of Córdoba and Jaén (see figure 16).

Fig 16 - Map of FGCC measured by model 2 at higher scale (1:150.000). Here we can observe the evolution of FGCC over time for a particular zone of Baena, province of Córdoba. Senescence of CC is observed after March 2017 for the majority of olive fields. Color transitions from green to orange represent cover crop senescence and consequent decrease of fCC, transitions to white represent full extinction of CC and consequent situation of bare soil. Zone A is located on Regosols and zone B on Cambisols.

The interactive maps generated by CCMAP, delivered different options under a user perspective:

- Selection of different periods of analysis.
- Selection of satellite images according to dates, zones and definition.
- Estimation of FGCC, fCC, and colour representation of different levels of plant density (see figure 17).
- Visualization of maps under false or natural colour.
- Visualization of Deimos satellite images used in the calibration stage.
- Visualization of CC development over time (see figure 16).
- Exporting maps according to zones of interest (see figure 18).

Fig 17 – Visualization of fCC map estimated by model 2 at different scales.

As an example of a map exported for a specific zone of interest, the following figure shows variability in one single olive plot in zone 5 (province of Jaén), where fCC was estimated at the end of November 2015 (see figure 18). Here we observe a spatial representation of fCC and we see that for the same olive density, variation within the same field is also observed.

Fig 18 – Map of FGCC measured by model 2 from satellite imagery of November 2015.

The selected validation plots have been visited (see map in the appendix A1) and the visual classification of fCC measurements were crossed with predicted values from CCMAP. Four classes of fCC were validated and grouped according to HSD-test results (see figure 19).

The first class, absence of CC, has shown significant correlation to the interval 0-20% of fCC. The second class, low CC density, has not shown to be significantly different from the third, medium CC density. Both were therefore grouped as Low-Medium class (20-50%). The fourth class has shown to be significantly different from all others with fCC values ranging from 60 to 100% (see some photos of plots with fCC > 60% in figure 19).

The model has still shown satisfactory performance on mapping differences between plots treated with herbicide (where not only a selection of species was observed but also progressive senescence) and plots managed under a conservation approach with mechanical control of weeds later in the season (larger and more diverse seedbank, good natural re-seeding capacity, higher species diversity, biological N fixation through higher concentration of Leguminosae, pictures of an example are shown in figure 19).

Fig 19 – Scheme with pictures taken at field validation stage. The three different classes of fCC after statistical analysis of our observations and an example of model prediction capacity of senescence under different CC treatments.

The initial classes, defined in *materials & methods*, were taken as a reference for visual classification. The visual classification (N=68) was then linked to the mean values measured by CCMAP for the same observation zones and statistical analysis conducted on the correlations found (see figure 20 and table 3).

Fig 20 – A) Box-plots of the initial four fCC (also called dFGCC) classes that we validated on fields; B) Histogram of fCC (also called dFGCC) levels, observations ranged from 0 to 80%.

Table 3 – Mean, median, minimum, maximum and standard deviation values of each initial class.

CC_level	Mean	Median	Min	Max	St.Dev
None	0.027763	0	0	0.15457	0.049037
Low	0.216327	0.21881	0	0.35311	0.103724
Medium	0.302084	0.31812	0.12387	0.51237	0.101959
High	0.632266	0.66898	0.24651	0.78294	0.16218

3.6 Time-series of cover crop fraction

Temporal dynamics of cover crop fraction were modelled through the analysis of fCC time-series (N=1600). The tree levels of fCC considered have fitted into polynomial regressions of degree five: High density (see figure 21 A), medium density (see figure 21 B) and low density (see figure 21 C).

Fig 21 – CC annual growth curves representing three different classes of CC density.

The previous figure represents the annual variation of fCC as a function of time (expressed as day of the year – DOY) for the three levels of ground cover fraction considered. We observed relatively early CC senescence once fCC starts decreasing in January, revealing that the majority of plots (even those with higher density of CC) have the soil unprotected against surface runoff and physical impact of raindrops during the critical periods of spring rainfall events (see figure 22).

Fig 22 – Qualitative representation of early fCC senescence derived from the high density polynomial function and the correspondent period of soil erosion risk due to late-spring rainfall events.

The integration of these functions in water balance or soil erosion modelling might be a promising step towards the development of real-time support decision software in the context of water and soil conservation in local olive groves (as shown by the schematic example in figure 22). An example applied to quantification of annual soil erosion risk was quantitatively tested through the *RUSLE* equation.

3.7 *RUSLE* calibration

The soil management parameter-*C* of *RUSLE* was calibrated based on the three representative annual CC growth curves. The box-plot analysis showed significant positive correlation between parameter-*C* and CC class (see figure 23 A, B and C). ANOVA has shown higher variability of parameter-*C* due to fCC than to olive density, revealing that annual soil erosion risk might be reduced on about 56% only by optimizing ground cover through better development of cover crops (see figure 23 D). Regardless the state of cover crops, relative reductions of annual soil erosion risk of about 23% are also predicted for scenarios of increased olive plant density (see figure 23 D).

Fig 23 – Box plots of C as a function of the CC class under A) low olive density (<30%), B) medium olive density (30-45%) and C) high olive density (>45%). ANOVA C mean plot for two controlling factors: olive and CC density.

The stronger correlation between parameter-C and CC class was also indicated by the differences of magnitude found in p-value test and by a lower Std. Error (see table 4). No significant correlation was found for the interaction between both controlling factors (see table 4), revealing that the effect of one variable is independent from the other (please note that parameter-C relative differences between different CC classes are constant for different olive densities (see figure 23 D).

Table 4 – ANOVA results of estimated parameter-C under two controlling factors: olive and CC density.

	Estimate	Std. Error	t-value	Pr (> t)
Olive_density (A)	0.0113521	0.0021294	5.331	1.02E-07
CC_class (B)	0.0464902	0.0019016	24.448	2.10E-16
factor_A : factor_B	0.0007924	0.0008803	0.900	0.368

3.8 The effect of field slope and area on CC development

The frontier analysis conducted on crop field area plotted against fCC did not show any significant pattern. No boundary function was therefore derived as a benchmark of this relation. However, for the case of field slope (N=16500), two possible benchmarks were graphically represented (see figure 24).

Fig 24 – Benchmarks of fCC as a function of field slope (%): A) a linear regression (where no cap was assumed); B) a negative logarithmic regression (where a cap was set at the 98th percentile in order to adjust the line to the higher density zone of the observation cloud).

4. Discussion

4.1 Model 1

Model 1 has shown some limitations that have consequently affected the quality of our validation set. The model accuracy on estimating FGCC from PNOA imagery was higher due to the existence of less peaks in the Y' histograms (see figure 5 and 6), which indicates that elements quantity and diversity enhances the sensitivity of FGCC to Y' threshold variations. This means that higher degrees of colour variation, as a result of larger plant species variability and soil gradients, added complexity to the calibration set and affected naturally the correlations found.

An example, of the high variability degree that we have worked with, is observed when analysing the correlations between FGCC and each satellite band (see figure 9). Despite the low R^2 values, the correlations found ($p\text{-value} < 0.05$) for wavelengths close to near-infrared (NIR) were not in line with what would be theoretical expected (see table 2). According to the theoretical reflectance spectrum of olive trees described by Zarco-Tejada et al. (2004), the reflectance at this wavelengths (close to 700 nm) should increase with FGCC, because higher amounts of chlorophyll per area imply higher reflectance of NIR radiation. This might indicate a considerable amount of noisy data due to stress such as lack of water or pressure of yield-reducing factors as described by Knippling (1970) and defined as the “red shift” phenomenon.

The “red-shift” phenomena is a plausible cause for the moderate accuracy of model 1. However, other factors might have also constrained the model's performance: lack of atmospheric correction for aerosols and smoke, insufficient spatial resolution of the satellite imagery and the integration of multiple dates (different sun angles) on the same regression analysis. It is definitely a valid argument that correlations between the Y' threshold and colour bands may not be the same for different dates, however, in order to minimize the amount of images requested to the European Spatial Agency (ESA), we assumed that CC temporal variability could be effectively represented by sufficient high spatial variability, and therefore, the CC temporal dynamics were estimated from large observation areas but only at three single dates.

Regarding spatial resolution variations, the comparison between vegetation indexes measured with different satellites (the satellite effect) has also shown some interesting relations (see figure 8). While Sentinel-2 has a spatial resolution of 10m, Deimos-2 has a spatial resolution of 4m. In this sense, for the sample tested [$N=1600$], higher-spatial resolution imagery seems to enhance NDVI values. In this case, we observed that NDVI estimated at 27th of November 2015 from Deimos-2 imagery plotted against the same index, but estimated from Sentinel-2 (at the same date), fitted into a linear regression with $R^2=0.86$ and $p\text{-value} < 0.05$ for both terms (see figure 8). Assuming that the expected relation between FGCC and NDVI will follow the similar patterns, it seems important to take into account the satellite that is used and the correspondent correction equations: this is what we called the satellite effect.

The correlations found with both the red and green colours should be understood at plant physiology level. The fact that both colours have shown correlations to levels of brightness capable to select vegetation results from the fact that light reflectance in these wavelengths is strongly correlated to the concentration of plant pigments (Arnon, 1971). While the two major pigments, chlorophyll a and b, are responsible for a large absorption of red light and reflectance of green light, the two minor ones, Xanthophyll and Carotenoids, absorb more green light while reflecting yellow (Arnon, 1971). Therefore, considering that the expected correlation between red and FGCC should be inversely proportional to the correlation between green and FGCC, the fact that both correlations with Y' threshold were positive (see figure 10), indicates that might not

exist a significant correlation between Y' threshold and FGCC. In this sense, higher levels of Y' threshold do not necessarily correspond to higher levels of plant density.

On the other hand, the binary classification method that was developed, has resulted into a “smoothing effect” on higher levels of plant density derived from DEIMOS imagery (see figure 12), which may contradicts the last point. Is it FGCC significantly correlated with the Y' threshold? Theoretical reasoning indicates that if there is a correlation, the linear regressions between Y' and red or green (see figure 10) should be inversely proportional instead. However, for higher levels of plant density this correlation seems to exist: we observed higher sensitivity of FGCC_Deimos to Y' threshold for higher levels of plant density (please note the larger dispersion of points in the top-right zone of figure 11).

Despite the lack of a clear explanation for the moderate accuracy ($R^2 = 0.4$) of model 1 (mostly constrained at higher levels of plant density as seen in both figure 11 and 12), but considering the duration of the internship, we have decided to move forward in order to explore the potential of machine learning to agronomic analysis of big data (model 2), which was also one of the main goals of this project.

4.2 Model 2

The results are mostly in line with one of the major hypothesis behind this project: *higher dimensional statistics through vector transformations and machine learning algorithms must add substantial value to modelling based on standard vegetation indexes*. Considering the complexity of factors that result into high variability of phenotypes at field level, finding strong correlations between NDVI or EVI and FGCC gets highly difficult. However, despite the moderate R^2 values, positive logarithmic correlations between FGCC_t and both NDVI and EVI were found (see figure 13, A and D), revealing limitations to predict FGCC on higher plant densities as previously demonstrated by model 1. On the other hand, FGCC estimation from summer imagery has shown to be more precise (higher R^2 values). The fact that correlations adjusted to linear regressions that do not “saturate” at higher levels of plant density (see figure 13, B and E) indicates that both NDVI and EVI seem to be quite accurate options for measuring olive trees fraction from summer imagery. However, despite higher R^2 values, the vertical dispersion of FGCC turned these correlations to be only significant for one of the equation terms (p-values were larger than 0.05 for the regression intercept) and therefore we cannot propose a significant function defining this correlation. By plotting the difference between the same vegetation index measured in winter (t) and summer against fCC, we observed relatively weak correlations for developing model 2.

Data analysis was then taken to a higher dimensional space through machine learning and the dataset generated with model 1 (N=1600) was used for both learning and testing purposes. In general, machine learning algorithms have increased the R^2 values of measured vs. predicted FGCC: from 0.36 in the case of NDVI, or 0.46 in the case of EVI to 0.5 in the case of MLR for instance (compare figure 14 to 15), and slightly decreased de RMSE values, revealing the higher potential of “bigger data” for modelling cover crops in olive orchards.

Two different types of results were obtained (see figure 15). In one hand, higher model robustness was observed with PLS, BR and MLR (see figure 15 A, B and C) algorithms while RF (see figure 15 D) has shown clearly an over fitting situation, by predicting FGCC much more precisely from the learning subset ($R^2 > 0.9$) than from the test set ($R^2 = 0.47$). This implies that the existing variability is not completely represented in the learning subset of RF and therefore, larger datasets would be required to fully interpret the algorithm performance.

For the generated dataset (N=1600), we know that multi-dimensionality has clearly improved model prediction capacity. Considering the MLR as an example (see figure 15 E), we know that a multi-linear function can be developed by multiplying the reflectance radiation measured at each satellite band by the correspondent estimated parameter for the same wavelength range (shown in figure 15 E by the red line). We decided to work with a MLR, because the R^2 values of both the learning (train) and testing (test) subsets were quite similar, indicating that the model performs well and homogeneously on both the data subsets.

4.3 Field validation

The annual evolution of cover crops during the year 2017 was successfully monitored by CCMAP as we clearly observe how fCC varies from zone to zone along the year (see figure 16). It is particularly interesting to observe different responses to different agroecological factors as well (please consider the example shown in figure 16, zones A and B). While zone A is predominately characterized by Regosols, zone B is located on Cambisols. From February to March 2017, we observed a clear decrease of fCC in zone A while in zone B it seems to substantially increase (see figure 16). This might be due to the higher water retention capacity of Cambisols, especially important during a dry year as 2017, which reveals the utility of CCMAP to represent the effect of soil spatial variability on cover crop temporal dynamics (IWG, 2007).

CCMAP generates maps that might be used at different scales (see figure 17 and 18). The development of a technical tool capable to represent fCC variability at different scales, allows its use not only for landscape planning and ecological studies at regional scale but also at crop field level, in order to integrate spatial data with precision agriculture management strategies. The model represents spatial variability of fCC within plots with homogeneous olive plant density (see figure 18), which gets additional importance under a farmer's user perspective.

The map generated from imagery corresponding to the second week of April has shown interesting results during the field validation stage. The validation has only deviated considerably on cases where CCMAP identified the presence of a high density cover crop and where soil had been ploughed in the previous days. In fact, by assuming no significant variations on cover crop fraction in a week-time period, we rejected, wrongly, the possibility of CC control during the same period. On those cases, the freshly mown biomass was still visible, mixed with soil, which confirmed, that despite not being capable to verify the level of fCC before control, there was a considerable amount of mown biomass which was taken as a validation indicator for the predictions of the model. However, those cases ($n < 5$) were not included on the statistical analysis of CC classes. For the few photos taken with *Canopeo*, we did not observe large differences ($> 10\%$) between fCC measured by the app and by the model.

CCMAP has shown higher accuracy when distinguishing between high cover crop density ($> 60\%$ of soil covered by herbaceous species) and low-medium density ($< 50\%$), than between low ($< 20\%$) and medium density ($[20; 50]\%$). This results from the fact that visual classification gets also more doubtful when distinguishing between low and medium levels of fCC. Which is relevant since the minimum levels of fCC proposed by EU policy (30% for slopes $> 10\%$) fall in this range and therefore, the use of tools capable to more precisely measure it seems to be necessary. In order to better validate CCMAP accuracy at these levels, we propose to use quantifiable field measurement methods instead of merely visual classification. The use of Fiji Image-J or Canopeo are two potential alternatives, however they are more time demanding and therefore, additional time would have been needed in order to integrate them in this particular stage.

The model estimates fCC based on colour light reflectance, which is proportional to chlorophyll concentration. The fact that some plots, fully covered by already senescing CC (reflecting more on the yellow wavelength range than on the green) were classified as lower density ones, does not necessarily mean that the level of ground cover is considerable lower for soil erosion control purposes. Attention must be paid to these cases. CC where herbicide has been applied, take from two to three weeks to fully die, and despite losing chlorophyll (and decreasing competition for water and nutrients with olive plants), they might still play a role in terms of soil protection, functioning as a mulch, by decreasing the physical impact of raindrops and by contributing positively to soil structure resistance through the still existing rooting systems.

It is still important to have a look at the histogram of fCC observed on field (see figure 20 B) and note that the large majority of fields during a rainy spring season (2018) were unprotected (fCC < 30%). For a year as 2018, this indicates unsustainable management and CCMAP might be used to understand where and why it occurs.

4.4 Soil conservation through the use of cover crops

The mean time-series of fCC, fitted into representative CC growth curves of each density class, showed potential to be integrated in soil erosion models such as the *RUSLE* calibrated by Gomez et al. (2003). In theory, herbaceous species growth curves would follow a mild density-dependent population growth up to an equilibrium at maximum soil cover (May & McLean, 2007). However, we observe a drop that results from the combination of two factors: a disruptive chemical or mechanical control of herbaceous vegetation and a decrease of temperatures during the winter period. The temporal analysis of this drop is essential for integrated management under a soil conservation perspective, without significant competition for water and nutrients with olive plants.

In general, early senescence was observed according to the annual precipitation distribution patterns. Despite not having linked CC dynamics to a water balance for the period of study, we assume that at least two critical rainfall events have occurred later in spring, when the modelled levels of fCC were below 20% for the highest density plots and below 15% for low-medium density plots (see figure 20). In fact, none of this levels of fCC respect the legislation concerning soil protection against surface run-off and consequent erosion on plots with slopes above 10% (which is the case of the large majority of local olive fields). Such evidence, points out to the need of understanding why CC are senescing too early, and which economic, cultural, operational and agro-ecological drivers are behind this trend.

Risk of soil erosion seems to be potentially controlled by better CC management. From the observed time-series of cover crop fraction (see figure 21 and 22), we may assume that a considerable delay on CC control and consequent later senescence might be a plausible and effective solution to decrease risk of soil erosion. However, despite the promising results regarding relative decrease of soil erosion risk (about 56%) due to better soil management (higher fCC), predicted by *RUSLE*, it is important to note that the calibration of parameter-C proposed by Gomez et al. (2003) is based on annual estimations that depend on the maximum annual fCC instead of dynamic estimations at a daily, weekly or monthly base. As we also see from the representative scheme (shown in figure 22), that high values of annual maximum fCC do not necessarily guarantee effective soil protection if the CC senescence starts too early. Therefore, methodological adjustments seem also to be required for a better understanding of the impact of CC management on the parameter-C.

Additional advantages of delaying CC senescence should also be discussed here: one relates to the capacity of harvesting earlier in rainy years due to less susceptibility of soil to durable water saturation and another lies more under a theoretical

ecology perspective regarding the CC species re-seeding capacity. This last point is extremely important to improve management of soil seed bank as an ecological weed/CC management strategy (Kruidhof et al., 2008). The soil seed bank dynamics might be explained by the *Allee effect: a population that falls below a certain threshold tends to extinction* (May & McLean, 2007). Therefore, the selection of rapid seed formation species (as those belonging to the genus *Bromus* L.) gains additional relevance under a multiple-year perspective in order to avoid soil seed bank population to fall below that threshold too early.

When it comes to soil conservation we may therefore immediately think about keeping it covered. We know from the *RUSLE* equation that steeper plots are more susceptible to soil erosion due to the effect of parameter-S. Frequently, technical advice for steeper scenarios, is mostly focused on CC management. But how plausible is to talk about CC management on plots with slopes higher than 30, 40 or even 60%? We may realise from the boundary functions derived (see figure 24) that field slope is an agro-ecological factor that clearly limits the benchmark for CC development.

Independently on the mathematical equation chosen to represent this constraint, a negative response is visible, calling for adjustments on political speech and technical advice for extreme cases. Thresholds should not be defined regardless the agro-ecological conditions of each system, and here, we may even mention that more than blindly promoting CC in steeper traditional plantations, the intensification of olive groves (more plants per ha) deserves to be also seen as an alternative to compensate this limitation, giving to the system the opportunity to cover soil with deeper rooting system species, capable of better exploring soil layers under a steeper topography. Other interesting alternatives for steeper plots might be the use of inert mulches or other organic protective compounds, (capable to decrease the factor-R in *RUSLE* equation) or eventually mobile plastic sheets on inter-rows. Therefore, for steeper cases, the key might be in finding alternative practices to cover crops, capable to play a similar role without losing economic viability, or even a more deep re-think of the system, as we are currently observing through intensification with higher densities of olive plants.

5. Potential applications of CCMAP

Under a solution-driven perspective, it is interesting to put CCMAP into the context of consultancy services. How can this kind of technology (or methodologic approach) add value under an agro-business perspective? In the discussion of results, we have tried to keep some examples “on the table”: *precision management of ground cover, remote analysis of CC development, zonal quantification of inter-specific competition degrees, calibration of soil erosion models, landscape planning, control and monitorization of cross-compliance, technical advice for policy making*, etc. However, a more detailed proposal for a consultancy product was also developed in the context of this internship. The idea was explored under a farming cooperative organisation perspective, integrating remote data collection tools as CCMAP with benchmarking analysis of big data. The “case-study” took as an example the Portuguese cooperative organisation *CERSUL*.

Digital cooperative organizations

Taking agronomic consulting to the “Big level”

1. The particular case of CERSUL

CERSUL is a Portuguese agro-cooperative organization located in the city of Elvas, Alentejo. The organization was founded in 1990 with the main purpose of grouping local cereal growers at commercial level. The idea that “together we are stronger” has driven the institution to focus on joined storage and commercialization of cereal grains, with common acquisition of inputs and technical advice. Recently, the increasing pressure of global markets on cereal production has been forcing local farmers to diversify the traditional systems, mainly based on cereal crops, into higher value perennial crops to which Portugal has particularly differentiated conditions. In this region, we talk mainly about olive and almond orchards. Currently, the organization counts more than 175 members and it is responsible for more than 10% of the national cereal production (60 kton year⁻¹).

2. CERSUL towards modern fruticulture

In recent years, considerable investments have been done in Alentejo towards modern fruticulture. The three main cases are related to vineyards, olive and almond orchards. In the particular case of CERSUL, a group of members has recently (2016) founded a company based on similar values of the mother institution but specially applied to production, industrial transformation and commercialization of olive oil. This company integrates now more than 30 members with a market volume of about 8M€. The group aims to produce, transform and sell more than 20 kton of olives (4% of Portuguese national production) before 2020.

3. Precision optimization of agro-economic efficiency

Modern fruticulture concerns largely about efficiency, not only for environmental purposes (in order to comply with legislation and market requirements) but also under a cost-efficient agronomic perspective (energy prices volatilization, growing market demand, water availability, labour costs through taxes and costly associated bureaucracy). Local agro-economic efficiency is largely affected by access to low-cost inputs (mainly energy), higher marked prices (that have shown significant inter-annual variation in the case of olive oil), control of yield limiting/reducing factors (in local conditions, mainly water stress through drip irrigation) and by adding value factors (product certifications and subsidies). Agro-economic efficiency can be distinguished in two levels: the agronomic level and the financial level.

At the agronomic level, the optimization of resource use efficiency is the main driver of systems efficiency. This might be done through two different paths: 1) sustainable intensification (increasing yields without additional use of resources such as water or energy, or in other words, additional net income due to higher revenue at constant costs of production) or 2) extensification (by decreasing level of resource use without yield loss). Previous benchmarking studies reveal that both paths are potential options for increasing the agro-economic efficiency of European production systems. In Córdoba for instance, controlled deficit irrigation has been an effective practice towards extensification of olive production systems. Irrigation systems adjusted to levels of crop water allowable depletion instead of evapotranspiration demands are showing promising scope for decreasing resource use per unit of yield. However, dynamic analysis of water demand adjusted to temporal and spatial variability of ground cover is still quite an unexplored field and it might for instance determine the development of precision irrigation schedules better adapted to soil conservation farming systems.

At the financial level, support decision tools capable to help CERSUL members to effectively comply with legislation and certification requirements might also add substantial value in forms of time saving, access to lower input costs or higher market prices. We believe that spatial and temporal mapping of ground cover dynamics for instance, might be an important tool to effectively quantify the agronomic and environmental impact of ecosystem services and therefore to better determine the access to cross compliance of EU policy and consequent direct payments.

4. CERSUL in the digital era (“Internet of Things and Big Data”)

We believe in the positive impact that digital cooperative organizations might promote. However, the smart use of resources in the context of a living environment requires the human dimension of territory to be taken as the central element of development. In the context of Alentejo, there is no point for smart development if it leads to social desertification.

The Internet of Things linked to Big Data analysis (automated analysis of massive amounts of data) under a pure agro-cooperative perspective is showing worldwide potential to turn farming into a more transparent sector, not only in terms of marketing but also and mainly in terms of joined agronomic performance. Working networked is the key for preserving minorities, build prices and therefore to empower economic activities in our rural world. We believe that digital services will drive those networks supported by historic and real time data more efficiently from now on.

In a first stage, it could be interesting to explore the development of a historical data collection platform in order to build a big data base capable to provide material for agro-economic benchmarking at plot, field or grower level. As an example, we propose an integrated analysis of agronomic performance based on econometrics, machine learning and remote sensing.

The methodological plan could consist on two different types of analyses structured by several stages:

Proposed analysis one (for all members):

1. Performing a survey in order to get all data available from previous years (yields, labour demand, energy consumption, variable and fixed costs, dates of interventions as biocide spraying, pruning operations, irrigation events, etc.)
2. Geo-mapping data (linking data to physical location of fields, or eventually management zones based on soil properties, plant density or other factors).
3. Perform an analysis of benchmark based on historical data through statistical analysis, clustering methods and GIS tools.
4. Development of a visualization platform for CERSUL members capable to provide interactive information with maps of resource use efficiency variability, scope for improvement, predicted performance under different simulated scenarios (income variability, water availability, energy consumption, labour use, etc) based on machine learning methods.
5. Maps of crops maturation combined with real time and forecast weather data for harvest and storage planning.

Proposed analysis two (mainly for olive growers):

1. Analysis of satellite imagery with Google Earth Engine tools.
2. Development of an image classification model based on satellite data and drone imagery (plant density, weeds coverage, zones of stress, other suggestions).
3. Field validation phase of the model (at least for two growing seasons).
4. Development of an interactive platform where members can consult dynamic state of their crops regarding cover crops, weeds, water consumption, etc. Integration of real time and forecast data on energy prices, harvestable product prices, weather, other risks (such as susceptibility to forest fires in near areas).

Fig 25. Representative scheme of the proposed project and potential products/solutions of interest for digital consulting integrated in a digital cooperative organization.

5. “Food for thought”

- A step towards integrated support decision software is being taken by multiple agro-business companies. Decision supported by real time data, dynamic visualization of benchmarking and other possibilities from the combination of big data analyses with machine learning and fundamental statistical methods is nowadays a determinant key on the business plans of companies such as Monsanto, IBM, John Deere, Fertiberia, Agrovision, BovControl, CropX, and many others. We believe the power is even bigger for network organizations like CERSUL. Examples of digital cooperation are becoming very popular in United States (interesting article to read: *GILPIN, L. How Big Data Is Going to Help Feed Nine Billion People by 2050. TechRepublic. 2015*).
- The integration of scientific research in the private sector has been driven development through solution driven research by combining critical mass from both worlds. The Spanish Council for Scientific Research (CSIC) is an institution with more than 70 years of experience in the development of scientific research and consultancy. Particularly, the Institute of Sustainable Agriculture (CSIC institute at Córdoba) is well known for its agronomic research applied to Mediterranean conditions and mainly to the case of olive cropping systems. The opportunity to develop a joined project would be an interesting chance to get both worlds closer.
- It is an interesting exercise to put things into a perspective of area! The investments of a cooperative organization must be considered according to the total area managed by its members. It is important to understand how much this could represent per area in relation to the added value of the final product.
- The added value of digital agro-ecosystems must be considered under two perspectives: 1) a cost marginal perspective and 2) a risk management perspective. The optimization of resource use must drive the adoption of this kind of technology: productivity of inputs such as labour or energy instead of additional yield at higher marginal costs; and minimization of risk through improved planning of operations (harvest campaign) supported by real time data and predictions based on forecasted scenarios supported by econometrical analysis of historical data.

6. Conclusion

Our methodology has shown some limitations resulting from operational decisions based on the time and budget constraints of this project. However, we must note that it was still possible to access a satisfactory amount of quality data in the time given and under the absence of a concrete financial budget. This was mostly possible thanks to the support of European Spatial Agency data sharing programs for scientific purposes.

In general, the CCMAP has shown robustness by performing well in different zones and dates. However, we believe that CCMAP accuracy can be substantially improved by re-calibration of the datasets used as learning material by model 2. In fact, it is important to mention that most of the deviations are explained by the uncertainty of the validation sets generated by model 1 and used by model 2. The enormous amount of phenotype traits (G x E) involved when this kind of analyses are up scaled to higher levels, e.g. multiple combinations of genotypes (weed/cover crop species/cultivars) with environmental factors (spatial variability of soil physical and chemical properties and crop management), leads to an increase of complexity related to the development and interpretation of radiative transfer models (RTM) in the context of plant science. In this sense, automated measurement of FGCC from relatively raw data (images that were not atmospherically corrected for instance) and collected at only three different dates, was one of the main struggles within this process. The regressions used in model 1 stage have underestimated FGCC for higher plant densities, which was called the “smoothing effect” and seen as one of the main limitations of our methodology. We believe that this can be partially solved by using higher-spatial resolution imagery, eventually through drone flights, and/or conducting an exhaustive atmospheric correction of images against the effect of aerosols and smoke. On the other hand, model 2 has shown better accuracy, with relatively better statistical correlations between predicted and measured data. Higher spatial dimensional analysis through machine learning has added accuracy to the correlations found through fundamental regressions with standard vegetation indexes, delivering additional possibilities for remote sensing of perennial cropping systems. The combination of both steps, model 1 and 2, resulted in what is considered to be the CCMAP model which has also confirmed its predictive capacity during the validation stage.

At landscape level, low levels of ground cover were observed for the majority of plots under a particularly rainy year, which calls for a better understanding of the main reasons behind such reality. Alternative paths must be adopted, not only on technical advice but also on policy making and political speech. Specific site agro-ecological conditions need to be taken into account when establishing realistic targets and additional attention should be given to other conservational strategies where even intensification (increasing olive plants per ha) might be seen as a potential way of preventing soil erosion when adverse conditions to the development of cover crops are observed.

To finalize, it is essential to highlight that machine learning and remote sensing within this context have shown high potential to be integrated into multiple tools/solutions capable to serve at both scientific and farming level.

References

- Aitkenhead, M. J., Dalgetty, I. A., Mullins, C. E., McDonald, A. J. S., & Strachan, N. J. C. (2003). Weed and crop discrimination using image analysis and artificial intelligence methods. *Computers and electronics in Agriculture*, 39(3), 157-171.
- Arnon, D. I. (1971). The light reactions of photosynthesis. *Proceedings of the National Academy of Sciences*, 68(11), 2883-2892.
- Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8), 412-420.
- Cano, J. L. (2015). Instalación de Python científico con Anaconda en Windows| 0.0-Curso Python científico.
- Cox, S. (2002). Information technology: the global key to precision agriculture and sustainability. *Computers and electronics in agriculture*, 36(2-3), 93-111.
- Cuenca, A. D., & Aguilar, M. J. C. (1997). La Cobertera vegetal en el olivar: un método de producción agroecológico. *Quaderns agraris*, 33-44.
- De La, M. G. M. V., & Morales-Fernández, E. J. (2015). Denominaciones de origen protegidas (DOP) y turismo gastronómico: una relación simbiótica en Andalucía. *GRAN TOUR, Revista de Investigaciones Turísticas*, (6).
- Denisko, D., & Hoffman, M. M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, 115(8), 1690-1692.
- Garibaldi, A., Gilardi, G., Clematis, F., Gullino, M. L., Lazzeri, L., & Malaguti, L. (2009). Effect of green Brassica manure and Brassica defatted seed meals in combination with grafting and soil solarization against *Verticillium* wilt of eggplant and *Fusarium* wilt of lettuce and basil. In VII International Symposium on Chemical and Non-Chemical Soil and Substrate Disinfestation 883 (pp. 295-302).
- Gates, David M. (1980) *Biophysical Ecology*, Springer-Verlag, New York, 611 p.
- Gómez, J. A., Battany, M., Renschler, C. S., & Fereres, E. (2003). Evaluating the impact of soil management on soil loss in olive orchards. *Soil use and management*, 19(2), 127-134.
- Gómez, J. A., Battany, M., Renschler, C. S., & Fereres, E. (2003). Evaluating the impact of soil management on soil loss in olive orchards. *Soil use and management*, 19(2), 127-134.
- Gómez, J. A., Guzmán, M. G., Giráldez, J. V., & Fereres, E. (2009). The influence of cover crops and tillage on water and sediment yield, and on nutrient, and organic matter losses in an olive orchard on a sandy loam soil. *Soil and Tillage Research*, 106(1), 137-144.

- Gómez, J. A., Infante-Amate, J., De Molina, M. G., Vanwalleghem, T., Taguas, E. V., & Lorite, I. (2014). Olive cultivation, its impact on soil erosion and its progression into yield impacts in Southern Spain in the past as a key to a future of increasing climate uncertainty. *Agriculture*, 4(2), 170-198.
- Gómez, J. A., Llewellyn, C., Basch, G., Sutton, P. B., Dyson, J. S., & Jones, C. A. (2011). The effects of cover crops and conventional tillage on soil and runoff loss in vineyards and olive groves in several Mediterranean countries. *Soil Use and Management*, 27(4), 502-514.
- Gull, S. F., & Skilling, J. (1984, October). Maximum entropy method in image processing. In *IEE Proceedings F (Communications, Radar and Signal Processing)* (Vol. 131, No. 6, pp. 646-659). IET Digital Library.
- Guzmán, J. R. A. (2004). Geografía de los paisajes del olivar andaluz. Consejería de Agricultura y Pesca, Junta de Andalucía, Sevilla.
- Hall, A., Lamb, D. W., Holzapfel, B., & Louis, J. (2002). Optical remote sensing applications in viticulture-a review. *Australian Journal of Grape and Wine Research*, 8(1), 36-47.
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1), 1-12.
- Hernández, A. J., Lacasta, C., & Pastor, J. (2005). Effects of different management practices on soil conservation and soil water in a rainfed olive orchard. *Agricultural Water Management*, 77(1), 232-248.
- Heuvelink, E., & Dorais, M. (2005). Crop growth and yield. *CROP PRODUCTION SCIENCE IN HORTICULTURE*, 13, 85.
- IFAPA. (2016). Available at: https://www.juntadeandalucia.es/agriculturaypesca/portal/export/sites/default/comun/galerias/galeriaDescargas/minisites/raif/Fichas_Fitopatologicas/OLIVAR_MALAS_HIERBAS.pdf (Accessed 8th of March 2018).
- IFAPA. (2018). Available at <http://www.juntadeandalucia.es/agriculturaypesca/ifapa/ria/servlet/FrontController> (Accessed 6th of March 2018).
- IWG, W. (2007). World reference base for soil resources 2006, first update 2007. Rome: FAO.
- Jamalabad, M. S. (2004). Forest canopy density monitoring using satellite images. In *Geo-Imagery Bridging Continents XXth ISPRS Congress, Istanbul, Turkey, 2004*.
- Jones, J. W. (1993). Decision support systems for agricultural development. In *Systems approaches for agricultural development* (pp. 459-471). Springer, Dordrecht.
- Kim, Y., Huete, A. R., Miura, T., & Jiang, Z. (2010). Spectral compatibility of vegetation indices across sensors: band decomposition analysis with Hyperion data. *Journal of Applied Remote Sensing*, 4(1), 043520.
- Knipling, E. B. (1970). Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. *Remote sensing of environment*, 1(3), 155-159.

- Knowler, D., & Bradshaw, B. (2007). Farmers' adoption of conservation agriculture: A review and synthesis of recent research. *Food policy*, 32(1), 25-48.
- Kruidhof, H. M., Bastiaans, L., & Kropff, M. J. (2008). Ecological weed management by cover cropping: effects on weed growth in autumn and weed establishment in spring. *Weed Research*, 48(6), 492-502.
- Laguna, A., & Giráldez, J. V. (1990, September). Soil erosion under conventional management systems of olive tree culture. In *Proceedings of the seminar on interaction between agricultural systems and soil conservation in the Mediterranean belt* (p. 8).
- Lantz, B. (2015). *Machine learning with R*. Packt Publishing Ltd.
- Marr, D. (1977). Artificial intelligence—a personal view. *Artificial Intelligence*, 9(1), 37-48.
- May, R., & McLean, A. R. (Eds.). (2007). *Theoretical ecology: principles and applications*. Oxford University Press on Demand.
- Mezghani, Mouna & Gouta, Hassouna & Laaribi, I & Labidi, Foued & J Agri, Int & Agri, R. (2016). Leaf area index and light distribution in olive tree canopies (*Olea europaea* L.) *International Journal of Agronomy and Agricultural Research (IJAAR)*.
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.
- Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4), 049901.
- Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, 24(12), 1565.
- Patrignani, A., & Ochsner, T. E. (2015). Canopeo: A powerful new tool for measuring fractional green canopy cover. *Agronomy Journal*, 107(6), 2312-2320.
- Piorr, H. P. (2003). Environmental policy, agri-environmental indicators and landscape indicators. *Agriculture, Ecosystems & Environment*, 98(1), 17-33.
- Romero, P., Castro, G., Gomez, J. A., & Fereres, E. (2007). Curve number values for olive orchards under different soil management. *Soil Science Society of America Journal*, 71(6), 1758-1769.
- Russ, G., & Brenning, A. (2010, June). Data mining in precision agriculture: management of spatial information. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* (pp. 350-359). Springer, Berlin, Heidelberg.
- Saeedi, J., & Faez, K. (2011). A new pan-sharpening method using multiobjective particle swarm optimization and the shiftable contourlet transform. *ISPRS Journal of photogrammetry and Remote Sensing*, 66(3), 365-381.
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., ... & Tinevez, J. Y. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7), 676.
- Schwab, K. (2017). *The fourth industrial revolution*. Crown Business.

- Sibanda, M., Mutanga, O., & Rouget, M. (2015). Examining the potential of Sentinel-2 MSI spectral resolution in quantifying above ground biomass across different fertilizer treatments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 110, 55-65.
- Tenreiro, T. R. (2017). Water and nutrient use efficiency of ware potato and winter wheat cropping systems in the Netherlands - A benchmarking analysis of Dutch arable farming. MSc. Thesis of Plant Sciences (Wageningen University), 18-46.
- Vig, A. P., Rampal, G., Thind, T. S., & Arora, S. (2009). Bio-protective effects of glucosinolates—A review. *LWT-Food Science and Technology*, 42(10), 1561-1572.
- Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., ... & Wald, L. (2015). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5), 2565-2586.
- Weier, J., & Herring, D. (2000). Measuring Vegetation (NDVI & EVI). NASA Earth Observatory.
- Wolfert, S., Ge, L., Verdouw, C., & Bogaardt, M. J. (2017). Big data in smart farming—a review. *Agricultural Systems*, 153, 69-80.
- Zarco-Tejada, P. J., Miller, J. R., Morales, A., Berjón, A., & Agüera, J. (2004). Hyperspectral indices and model simulation for chlorophyll estimation in open-canopy tree crops. *Remote sensing of environment*, 90(4), 463-476.

Appendix

A1. Field validation stage map

Fig A1 – Map of field validation route and selected olive plots for cross-validation of CCMAP.