Spatiotemporal analysis of gap-filled high spatial resolution time series for crop monitoring

University of Natural Resources and Life Sciences (BOKU)

Departament of Landscape, Spatial and Infraestructure Sciences
Institute of Geomatics

Master's Degree in Agricultural Engineering/
Máster en ingeniería agronómica
Academic year: 2019 - 2020

Student:
Clara Rajadel Lambistos
from: Universitat Politècnica de València (UPV)
Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural (ETSIAMN)

Supervisors from BOKU:
Dr. Emma Izquierdo-Verdiguier and Professor Dr. rer. nat. Clement Atzberger

Tutores de la UPV:
Miguel Sánchez Marco y Aurea Cecilia Gallego Salguero

Vienna, 18th of November 2020

Abstract

Spatiotemporal analysis of gap-filled high spatial resolution time series for crop monitoring.

Reliable crop classification maps are important for many agricultural applications, such as field monitoring and food security. Nowadays there are already several crop cover databases at different scales and temporal resolutions for different parts of the world (e. g. Corine Land cover in Europe (CORINE) or Cropland Data Layer (CDL) in the United States (US)). However, these databases are historical crop cover maps and hence do not reflect the actual crops on the ground. Usually these maps require a specific time (annually) to be generated based on the diversity of the different crop phenologies. The aims of this work are two: 1- analyzing the multi-scale spatial crop distribution to identify the most representative areas. 2- analyzing the temporal range used to generate crop cover maps to build maps promptly. The analysis is done over the contiguous US (CONUS) in 2019. To address these objectives, different types of data are used. The CDL, a robust and complete cropland mapping in the CONUS, which provides annual land cover data raster geo-referenced. And, multispectral high-resolution gap-filled data at 30-meter spatial resolution used to avoid the presence of clouds and aerosols in the data. This dataset has been generated by the fusion of Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS). To process this large amount of data is used Google Earth Engine (GEE) which is a cloud-based application specialized in geospatial processing. GEE can be used to map crops globally, but it requires efficient algorithms. In this study, different machine learning algorithms: Random Forest and Support Vector Machine are analyzed to generate the promptest classification crop maps. This study presents the first results and the potential to generate crop classification maps using as less possible temporal range information at 30 meters spatial resolution.

Keywords: Crop classification, Google Earth Engine, Fusion data, Image processing, Classification algorithms, Vegetation indices

Student: Clara Rajadel Lambistos
Supervisors (BOKU): Dr. Emma Izquierdo-Verdiguier and Professor Dr. rer. nat. Clement Atzberger
Supervisors (UPV): Miguel Sánchez Marco y Aurea Cecilia Gallego Salguero

Vienna, 18th of November 2020

Resumen

Análisis espacio temporal de series temporales de imágenes de alta resolución espacial libres de huecos para el monitoreo de cultivos.

La obtención de mapas fiables de clasificación de cultivos es importante para muchas aplicaciones agrícolas, como el monitoreo de los campos y la seguridad alimentaria. Hoy en día existen distintas bases de datos de cobertura terrestre con diferentes escalas espaciales y temporales cubriendo diferentes regiones terrestres (por ejemplo, Corine Land cover (CORINE) en Europa o Cropland Data Layer (CDL) en Estados Unidos (EE. UU.)). Sin embargo, estas bases de datos son mapas históricos y por lo tanto no reflejan los estados fenológicos actuales de los cultivos. Normalmente estos mapas requieren un tiempo específico (anual) para generarse basándose en las diferentes fenologías de cada cultivo. Los objetivos de este trabajo son dos: 1- analizar la distribución espacial de los cultivos a diferentes regiones espaciales para identificar las áreas más representativas. 2- analizar el rango temporal utilizado para acelerar la generación de mapas de clasificación. El análisis se realiza sobre el contiguo de Estados Unidos (CONUS, de sus siglas en inglés) en 2019. Para abordar estos objetivos, se utilizan diferentes fuentes de datos. La capa CDL, una base de datos robusta y completa de mapas de cultivo en el CONUS, que proporciona datos anuales de cobertura terrestre rasterizados y georeferenciados. Así como, datos multiespectrales a 30 metros de resolución espacial, preprocesados para rellenar los posibles huecos debido a la presencia de nubes y aerosoles en los datos. Este conjunto de datos ha sido generado por la fusión de sensores Landsat y Moderate Resolution Imaging Spectroradiometer (MODIS). Para procesar tal elevada cantidad de datos se utiliza Google Earth Engine (GEE), que es una aplicación que procesa la información en la nube y está especializada en el procesamiento geoespacial. GEE se puede utilizar para obtener mapas de cultivos a nivel mundial, pero requiere algoritmos eficientes. En este estudio se analizan diferentes algoritmos de aprendizaje de máquina (machine learning): bosques aleatorios (RF) y máquinas de vectores de soporte (SVM) para analizar la posible aceleración de la obtención de los mapas de clasificación de cultivo. Este estudio presenta los primeros resultados para la generación de mapas de clasificación de cultivos utilizando la menor cantidad posible de información, a nivel temporal, con una resolución espacial de 30 metros.

Palabras clave: Clasificación de cultivos, Google Earth Engine, Fusion data, Procesamiento de imágenes, Algoritmos de clasificación, Índices de vegetación

Autor: Clara Rajadel Lambistos
Tutores (BOKU): Dr. Emma Izquierdo-Verdiguier y Professor Dr. rer. nat. Clement Atzberger
Tutores (UPV): Miguel Sánchez Marco y Aurea Cecilia Gallego Salguero

Viena, 18th de noviembre de 2020

Content

Figures

Tables

1. Introduction

Classification crop maps are used by the scientific community, agro-markets, governments, farmers, etc. to facilitate making decisions related to food and agriculture. However, crops are a dynamic landcover that needs permanent monitoring. This is especially important because their patterns are affected by external factors as climate change. All variables affecting crops, either inherent or external, make prompt classifications a big urgency. New technologies of remote sensing imagery are used to capture specific crop growth stages [28]. This section develops challenging topics in agriculture related to food security and agro-economy. The objective is to highlight the importance of monitoring crops and building remote sensing techniques. Moreover, up-to-date cloud computing technologies are presented to get valuable information from current satellite imagery. One of them is GEE centered on parallel geoprocessing.

1.1. Challenges in agriculture

Crop yielding primarily depends on weather conditions [12]. This makes plant behavior very unpredictable in time and it is reinforced by climate change that affects plants in biotic (pests and diseases) and abiotic manners (temperatures, droughts, etc) [27]. Therefore, climate change impacts the stability of food security and agro-market prices, conditioning at the same time future crop decisions. Climate change has not the same effect over every location on Earth. The template regions can be favoured by new conditions as warmer climates or longer periods of growth, while the tropical regions are supposed to lose production. One example of that is the case of cereals in Finland which production will increase in the future [8]. The variability and vulnerability depending on zones to climate change makes important monitoring crops. Climate change does not only vary depending on the areas, but also its effects can be contradictory. For example, incoming elevated CO_2 concentrations will boost the productivity of crops, but also change temperatures (Figure 1) and water availability [32]. High temperatures make crops grow quicker with less time to weight their grains and resulting in a lower final production. As well, they favour weeds growth and competence in resources with cropped species [27]. The result is a difficult understanding of how plants will behave. Weather disasters are also favoured with climate change, and these unpredictable events appear each year, for example, the August 10 windstorm derecho was a fast-unfolding disaster that affected millions of acres of corn, soybeans and other Midwestern farmland in the United States in 2020 [36]. Other consequences derived from climate change are droughts, floods, pests, and diseases. Moreover, farmer's decisions on which crops plant next season are conditioned by climate change, as well their crop practices, which leads to the emergence of new and different cultivation trends. This makes the variability between years in crop productions higher and more unpredictable. Crop instability affects the major actors in agriculture that are consumers and producers. The following sections detail its impact on food security and farmers.

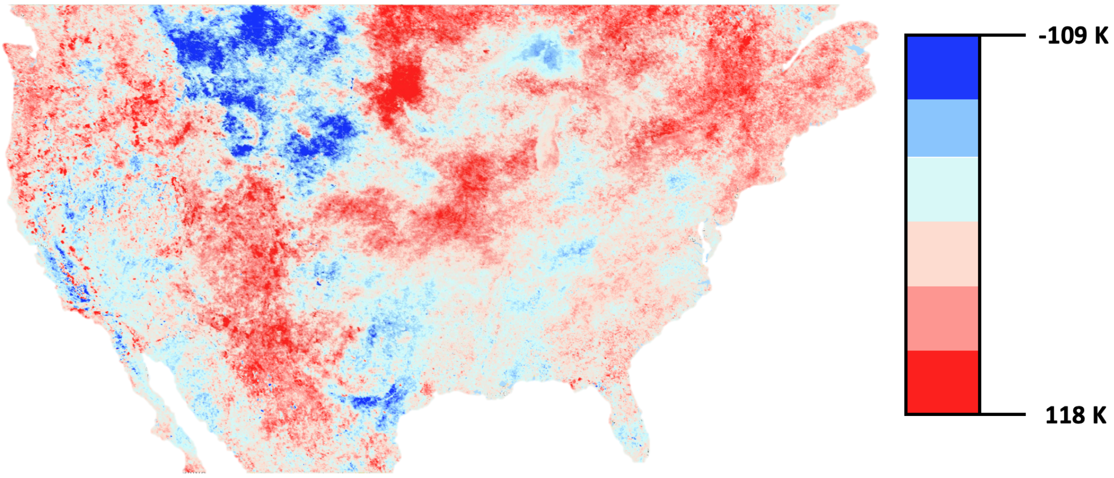


Figure 1 - Changes in mean surface temperatures (Kelvin) in the month of July in the United States between the periods of time: 2005-2009 and 2015-2019.

1.1.1. Food security

Climate change affects food security in the relocation of crop productions with impacts on prices, trade flows and food access [8]. The main concern is to achieve food security overall and that requires planning and logistics. Regarding food accessibility, not all countries are able to feed their population with enough quantity and diversity of food. Moreover, even if a country produces enough groceries, it does not ensure its availability at local level. Within countries, some areas may have suitable soils and may be more resilient to weather changes. The seasonal variability in food availability is reinforced by damages in plants from climate change. For this reason, through international and local trades, food surpluses need to be transported and distributed in the deficient areas and their markets [48]. On the other hand, food accessibility depends on market prices, locations, and "social contracts''. Detecting where crops are produced and in which quantities, helps to plan better their distribution and anticipate social policies. Strategies derived from crop monitoring can improve food availability and accessibility.

1.1.1.1. Future food demand

Supplying enough nourishment to all of us in the future will be even more challenging as growth dynamics is experiencing a huge rise in terms of population. According to FAO statistics, there will be around 9.7 billion people over the world in 2050. Population growth is highly noticeable in countries from Asia and Africa compared to Europe, North America, Oceania and Latin America. These last developed countries need to increase food incomes and their exports are relatively constant [9]. Moreover, Figure 2 shows how climate change impact is more concerning in tropical zones, where population is exponentially increasing. This will lead to an increase in imports of agricultural products in these areas. The world food production will need to grow up to 60-70% by 2050 [14]. Countries in temperate regions need to produce high quantities of nutrients to feed the population. Techniques to predict yearly and fast the production of crops and its locations are necessary.

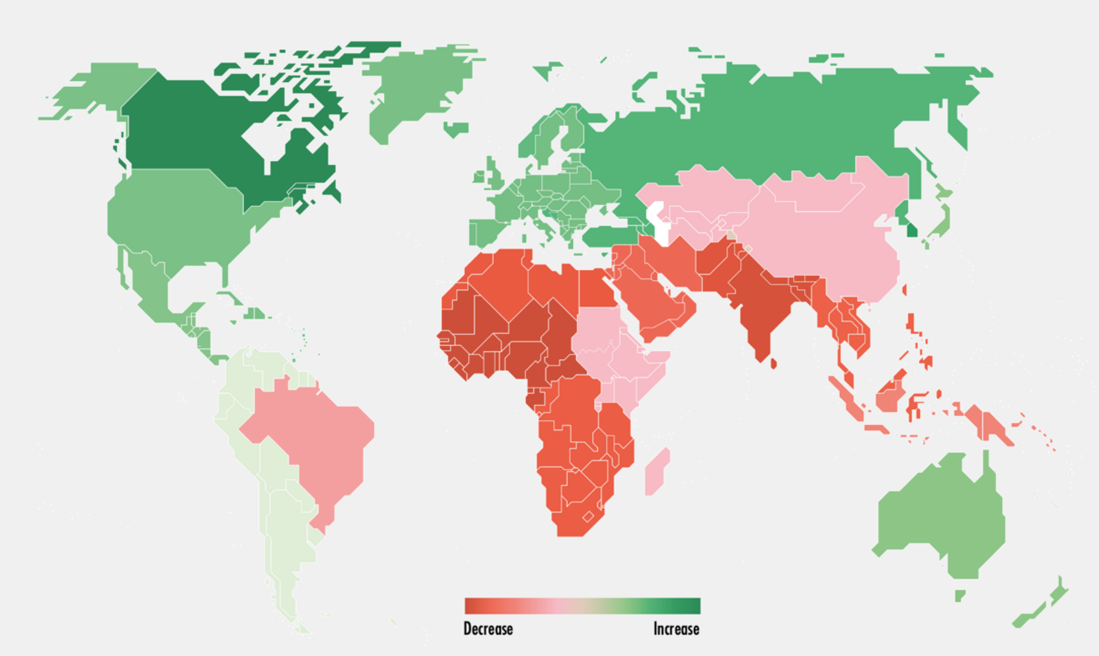


Figure 2 - Changes in agriculture production in 2050: Climate change relative to the baseline [9].

1.1.1.2. Importance of cereals

Cereals are major contributors into the intake of calories in those countries where the population is expanding [14]. They contain high energy values provided by the carbohydrates, mainly starches (65 - 75%), proteins (6-12%), and fat (1-5%) [39]. Therefore, cropping and monitoring cereals such as corn, rice, soybeans or wheat is of great relevance. Moreover, according to the National Agricultural Statistics Service (NASS) of the United States (US) Department of Agriculture (USDA) from 2008 to 2016, the production, domestic consumption, and exports of cereals in the US has increased between 16 - 18% while imports remain constant [45]. Indeed, the Corn Belt in the US provides more than one third of all the corn globally [32]. But feeding is not the only contribution of cereals, they also are inputs in many processes of other industries [29]. In this work, soybeans, corn, wheat, alfalfa, cotton and sorghum in the US are studied. Figure 3 shows the Normalized Difference Vegetation Index (NDVI) of these cereals in 2018. The NDVI characterizes the canopy vigor and serves as reference to evaluate the phenology in crops [52]. The peak of growth of these crops take place between May and August. Previous studies of crop classifications using remote sensing techniques claim that the classification maps can be acquired after the growing peak when crop phenology is best distinguished [11]. In this work, the best period/month of the year to crop classification by remote sensing images is analyzed. This is interesting because crop yielding data before harvest contributes to optimize grain distribution and planification.

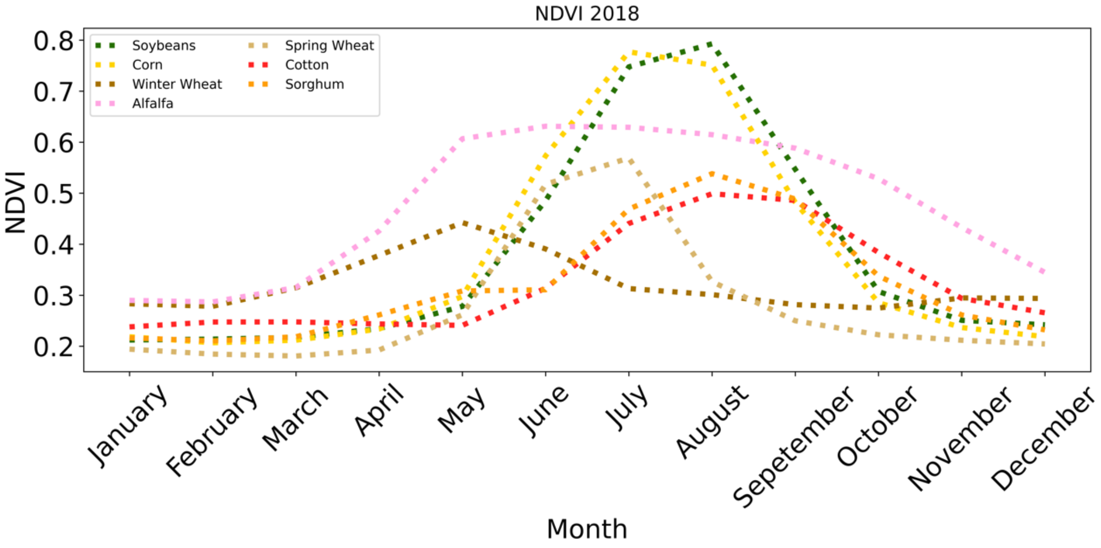
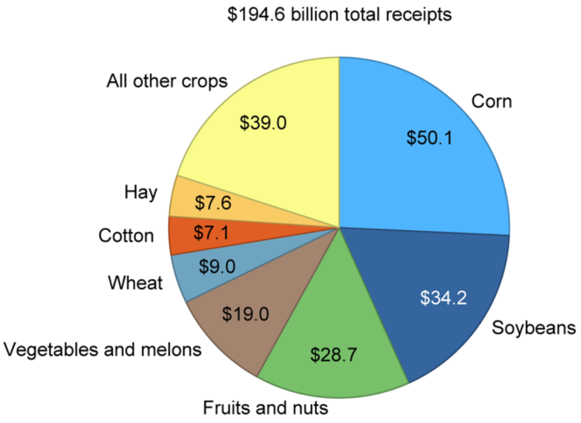


Figure 3 - NDVI profiles for 7different crop classes over the US in 2019.

1.1.2. Agro-economy

Individuals taking part in the agro-market suffer from crop variability and difficulty to predict final productions in time. In 2019 in the US, the total crop cash receipts in agriculture were $194.6 billion (Figure 4) being the 43.3% of this sum belonging to corn and soybeans [46]. In the figure wheat and cotton also appear as important cereals in terms of economy. This verifies the importance of these cereals on the economy and the need of providing timely information of their cropping. Examples of the need of crop near real time statistics are presented below.



Note: Components may not sum to total because of rounding. Data as of September 2, 2020.

Figure 4 - 2019 crop cash receipts ($ billion) [46].

Commodities and speculations

There are many factors which affect the price of crop commodities and its volatility, but the most important are seasonality (reinforced by climate change), energy prices, and excessive speculations [29]. The volatility on prices or "price risk" affects planted acreages in the way that high volatility tends to reduce acreages in some crops. Therefore, among others, global acreages depend on seasonal estimations [20]. Crop monitoring is fundamental to inform farmers, policy makers, and stakeholders. Objective estimations will lead to reduced risks in decisions and prevent excessive speculations, which also affect consumers with rise in prices. Being timely is crucial for making the best use of monitoring information because of the seasonality of crops. Nowadays, estimations are mostly based on surveys, but remote sensing can offer a more objective and timelier alternative to predict crop planted areas [2].

Crop insurances

Farmers rely on crop insurances to save their crop productions because averages in final cereal yieldings are extremely related with weather conditions. Since 2015 the main cause of crop loss in the US are weather-related causing even 85% of corn and soybeans losses [13]. Moreover, in low and middle-income countries farmers are most affected by climate change because their farming systems are more vulnerable [27]. Indemnity insurances depend on climatic variables, such as, rainfalls, temperatures, droughts, etc. Normally these variables are accepted within a range and out of this threshold farmers are compensated for their crop losses. It takes a long time to compensate farmers until damages are proved. Timely crop monitoring techniques through satellites can make the process of insurance much faster and economically affordable by assurance companies [9].

1.2. Crop forecasting

Agriculture land covers are dynamic, and agro-marketplace needs objective and timeliness crop forecasts to define basis in commodity markets. This information is valuable for farmers, agribusinesses, economic firms, university researches, government policy makers, and media. In the US, the USDA provides monthly acreage estimates depending on the crop for the US and also at global scale (for example winter wheat monthly estimates begin in May). USDA uses several information sources such as the NASS that estimates yearly the acreage of crops cultivated in the US [47]. For performing these estimates, NASS selects sample farms using satellite imagery, among others, and collects the data using survey methods (mail, phone, personal interview, or internet). However, surveys and field trips for obtaining this information usually are expensive and cost-timing [49]. Therefore, this process is time-demanding and requires excessive administrative intervention. As well, at global scales, objective farm surveys are not available. In this case, results are obtained from weather analysis, country reports, and satellite imagery. As well, not all countries have the robust data historial as the US and thus they can not obtain their own forecasts with good accuracy using these methods. For all these reasons, speed the crop forecasting process and make it more objective and simpler is highly needed. Moreover, even though the US generates around 20% of grain in the world, it still does not provide timely spatial estimates of production, that means no crop maps are generated on real time [8]. All of this is possible through improving remote sensing methods to obtain valuable satellite data. Remote sensing can not only help in identifying crops over large scales, but also in optimizing the selection of farm samples as those used by the NASS.

1.2.1. Remote sensing

In the previous sections, it is shown that crops are conditioned by several parameters, making them highly variable in space and time. Even there exist other monitoring practices (surveys, historical data, etc.) monitoring crop productions through large areas using remote sensing provides objective, valuable, and reliable data. The resultant information allows us to make evidence-based decisions being this useful for governments, stakeholders, farmers etc. [2]. Therefore, crop classification maps from remote sensing are important for many agricultural applications, such as field monitoring and food security. Principal requisites asked for obtaining maps are the temporal, spatial and spectral resolutions of the satellite sensors. Using heavy volumes of time-series imagery is important because agriculture is a dynamic landcover. As well is important, even the maps cover large areas having fine spatial resolutions, especially if they contain small farms. Once satellite images are obtained, they need to be processed by users to obtain the desired data. Nowadays, a high number of algorithms have been developed to perform classification and forecasting tasks. In the literature are used from basic algorithms as decision trees [51] to more complex models developed by deep learning processes. Some examples are artificial neural networks [38], support vector machines [15], convolutional neural networks [28] or random forests [37]. Nowadays, the most available maps over the world are provided by the European Space Agency (ESA) global land cover map (Figure 5) on the internet. However, its resolution is not enough when analyzing land covers at country level. At continental scale in the US is provided the Cropland Data Layer (CDL)[1] by the USDA. In this case, it provides a good resolution of 30 m, but this type of map is not currently available for almost all of the world. A methodology to obtain crop classification maps with good spatial resolutions and applicable over all the world is necessary to monitor agriculture.

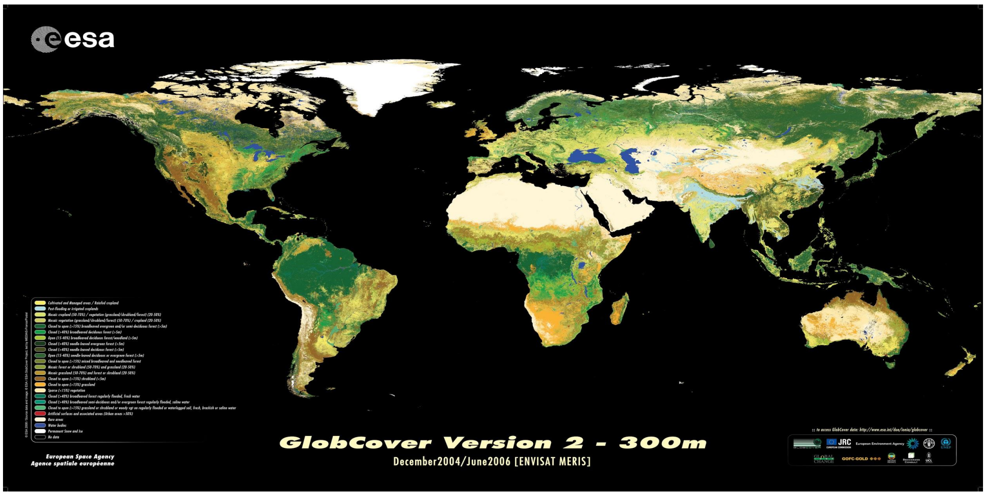

esa
GlobCover Version 2 - 300m
December2004/June2006 [ENVISAT MERIS]
European Space Agency
Agence spatiale européenne

Figure 5 - ESA (European Space Agency) global land cover map (ESA website global land cover map[2])

1.2.2. Prompt crop classification maps

Multi-temporal remote sensing provides the best opportunity to accurately and repeatedly obtain crop classification maps. However, even landcover maps over years provide valuable information, they are more useful when used in real time. To get the best value of them, crop classification maps have to be timely, provided before the harvest and uploaded regularly until the end of the season. This allows individuals involved in agriculture to forecast and make decisions on time. The prediction of crops before their harvest can prevent famine and help with food security strategies [33]. In the literature other works have obtained early classification maps, for example, Dahal, Wylie and Howard (2018) obtained an accuracy around 70% with 500 m of spatial resolution over the US by the beginning of September to classify major crops. Another example is the case of Konduri et al. (2020) mapping corn and soybeans with 231 m of spatial resolution over the US with 90% of accuracy in August. There are also studies mapping winter wheat by the end of April as the work of Skakun et al. (2017). Science should keep

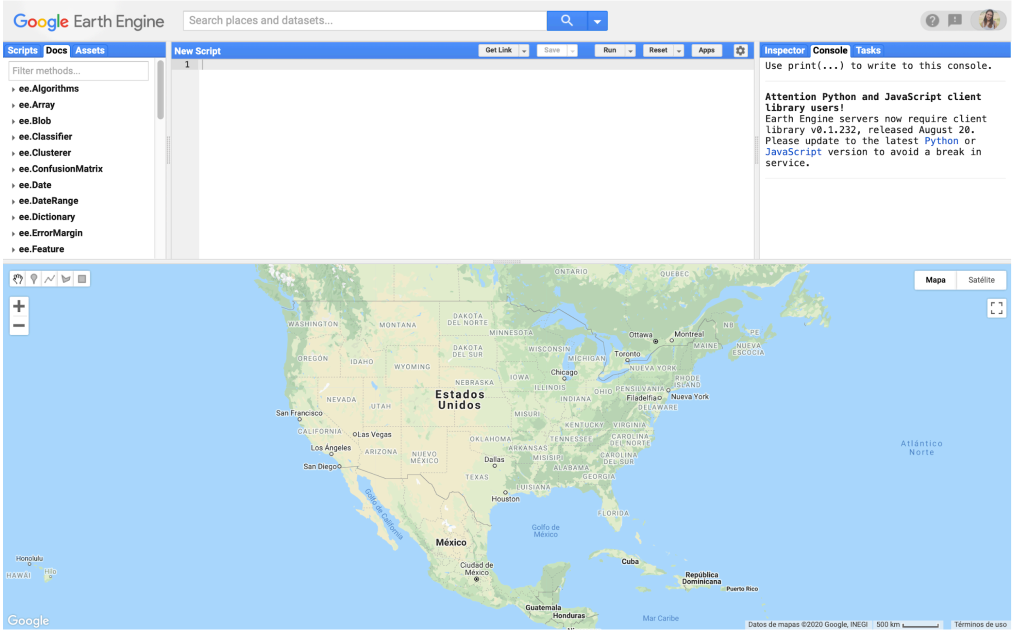
[1] CropScape viewer: https://nassgeodata.gmu.edu/CropScape/

[2] ESA website global land cover map: http://www.esa.int/Applications/Observing_the_Earth/Space_for_our_climate/ESA_global_land_cover_map_available_online

investigating how to provide timely maps obtaining good classification accuracies and with the best spatial resolutions.

1.2.3. Current potential tools

Currently, there are many efficient and potential tools to monitor and forecast crops promptly. Moreover, free satellite imagery datasets are easily available on the internet. For example, Landsat products are characterized for being free, with -high 30 m spatial resolution, and the longest running continuous program [50]. Among others, its imagery data can be processed in Google Earth Engine (GEE) [19] which is a cloud computing geospatial platform (Figure 6). GEE contains a wide catalog of satellite imagery (Landsat, Sentinel, and Moderate Resolution Imaging Spectroradiometer) allowing detection of changes and map trends to scientists. It also contains datasets with climate and weather data and digital elevation models (DEMS) which helps in crop monitoring tasks (GEE website[3]). This web application is developed in Javascript and contains specific functions to analyze huge quantities of images in short periods. As well, GEE has its python version (GEE Python API). The python API can be used downloading the package "ee". This is helpful because Python contains useful libraries for machine learning and supervised classifications, such as Scikit-learn, which can complement classification tools in GEE. Moreover, Python can be used in the open-source web application Jupyter Notebook[4] which easily allows to share live code and obtain instant visualizations. There exists some Geographic Information System (GIS) software which can also be used to work with satellite data. One example is QGIS[5] that is used to create, edit, and visualize geospatial information. There is no need to use only one of these tools, all of them contribute improving remote sensing tasks in different manners. A blend of them provides a robust set of tools for analyzing different crop mapping techniques.



```
Use print(...) to write to this console.

Attention Python and JavaScript client
library users!
Earth Engine servers now require client
library v0.1.232, released August 20.
Please update to the latest Python or
JavaScript version to avoid a break in
service.
```

Figure 6 - Google Earth Engine interface.

[3] Google Earth Engine website: https://earthengine.google.com

[4] Jupyter Notebook: https://jupyter.org

[5] QGIS: https://qgis.org

1.3. Objectives

Monitoring crops over large areas is vital for many reasons mentioned in the previous sections. The aim of this work is to obtain prompt classification maps over a continental magnitude, in this case over the US, and as soon as possible. In this way, maps are available when they are most needed and not later when their contribution is minor. To achieve this, first is proposed analyzing the multi-scale spatial crop distribution in the US to identify the most representative crop areas. The reduction is done over climatic regions and counties in the US. By selecting these reduced areas data collection is simplified diminishing costs and time of sampling. Secondly, it is pursued to analyze the temporal range used to generate crop cover maps to build promptly maps. This allows obtaining crop classification maps before the end of the year and probably before harvests. Different parameters concerning the classification will be analyzed and compared: study areas (climatic regions and counties), input datasets (satellite imagery and others), and classification algorithms.

2. Area of study and data

This chapter presents the study area used in this work and a detailed explanation about all the databases used in this thesis. It covers from the dependent variables (i. e. classes, targets) obtained from the Crop Data Layer to the independent variables (i. e. features or bands) calculated from different remote sensing satellites or using weather data. The input datasets used are all free available on the internet. As well, two spatial datasets are presented to divide the area of the CONUS into climatic regions and counties.

2.1. Area of study

The study area of this work is the contiguous (CONUS) United States (US) formed by the 48 adjoining states in the US and excluding Alaska, Hawaii, and all other offshore insular areas. The CONUS represents a continental magnitude that can serve as an example to apply in other large areas on the Earth as it contains 7,663,941 km^2 of land cover. The CONUS is chosen due to its presence of extensive areas dedicated to important cereals such as corn or soybeans. Figure 7 shows the main four agriculture zones in which the CONUS is divided: midwest, centre, south and west. The Midwest of the CONUS contains one of the most grain productive areas in the world. In this region, known as the Corn Belt, primarily are produced soybeans and corn which are key points in exportations, as well as oat, sorghum and wheat [31]. In the center of the CONUS is located the Great Plains, this is a flat area where are concentrated most of the country crop products. Climate in both areas is characterized for its variability with cold winters and hot summers. Recently, these zones are getting warmer due to gas emissions and also precipitation is rising [26]. The south of the CONUS is mainly destined for cotton and tobacco production. This area suffers from extreme weather events (floods, droughts, heat waves, etc.) and its agriculture has not experienced a big increase in recent years [22]. Finally, west of the CONUS is a mountainous area where agricultural practice is more restricted by geographical conditions. In general, over all the CONUS ranges of temperatures decrease from the south to the north.

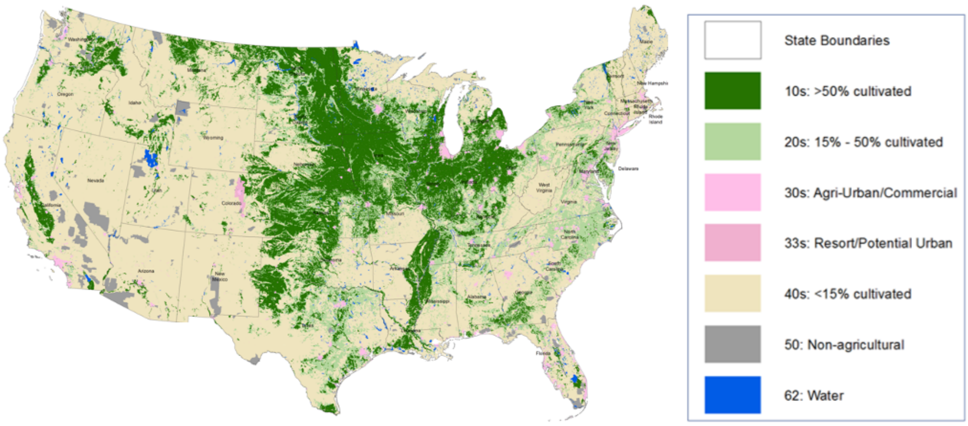


Figure 7 - Percent land use cultivation in the CONUS. (USDA Land Use Strata website[6]).

2.2. Data

[6] USDA Land Use Strata website: https://www.nass.usda.gov/Research_and_Science/stratafront2b.php

2.2.1. Cropland Data Layer (CDL)

The Cropland Data Layer (CDL) is a cropland cover product geo-referenced since 2008 throughout the CONUS and created using moderate resolution satellite imagery and extensive agricultural ground truth. This georeferenced shapefile has been developed by the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) to provide timely, accurate, and useful statistics in service to U.S. agriculture [6]. It is available as an annual database containing land cover information at a spatial resolution of 30 m. Its format is presented as raster GeoTIFF files, but a vector shapefile is also included for each state. CDL image is an accessible product in GEE platform under the name "USDA NASS Cropland Data Layers" and also it is possible to download the data on the NASS CropScape web application (USDA CropScape and CDL - metadata website[7]). The main aim of CDL generation is the identification of 100 different crops as well as non-cropped categories, classifying each pixel of the CONUS with a different type of land cover. The CDL product assessment provides high accuracy values compared with the ground truth. Since 2008 the methodology to generate the CDL has been standardized making it a robust and useful dataset [24]. The annual information provided by CDL is available the February of the following year once the classification has taken place. The CDL is appropriated and helpful data in several processes such as crop statistics measurements and as a training dataset in supervised classification methods. In this work, CDL 2018 and 2019 are used because they are the most current years available in the temporal series.

On the one hand, the CDL 2018 was used to analyze statistically multi-scale crop distribution over the CONUS and then, to identify the most representative areas based on its crop percentages. In addition, CDL 2018 was used in the benchmarking of the crop promptly classification maps over the temporal analysis range. On the other hand, the CDL 2019 was used in the benchmarking of the forecast processing in order to study the potential of the prompt classification maps in real-time. Figure 8 represents the CDL in 2019 as an example of this land cover layer over the CONUS.

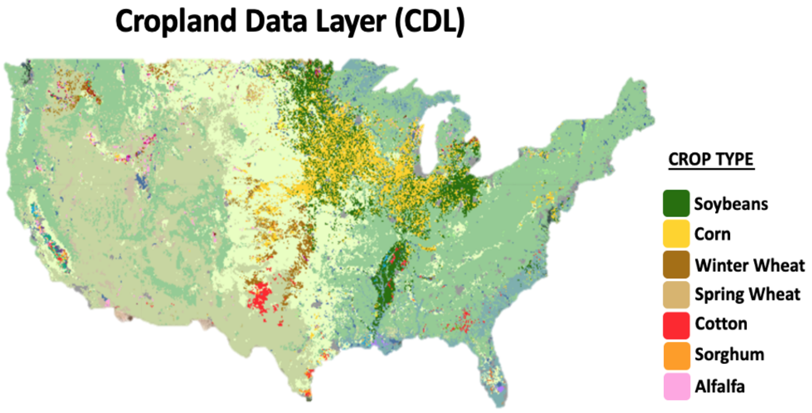


Figure 8 - Annually derived Cropland Data Layer (CDL) in 2019, CONUS. Legend: the 7 crops of interest in this thesis.

[7] USDA CropScape and CDL - metadata website: https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/meta.php

2.2.2. Satellite data: Landsat and MODIS

Landsat program is a group of satellites that provide global time series images. The program started in 1972 with the launch of Landsat 1 satellite and it is still producing images with Landsat 8 satellite. This group of satellites (i. e. from Landsat 1 to Landsat 8) follows a Landsat Data Continuity Mission (LDCM), offering the assurance that the currently studied applications can be used in the future. Landsat mission series are mainly managed by the United States Geological Survey (USGS) and the National Aeronautics and Space Administration (NASA). They are a valuable data source to scientists through cloud computing by using algorithmic approaches. Within their multiple applications, they are used to monitor and detect changes on the Earth's surface [50].

Landsat 7 and Landsat 8 were used to generate monthly gap filled and smoothed reflectance images at 30 m spatial resolution over the CONUS in this thesis. The Enhanced Thematic Mapper Plus (ETM+) on Landsat-7 at 30 m spatial resolution includes an additional 15-m spatial resolution panchromatic channel and a 60 m TIR band. The Landsat-8's Operational Land Display (OLI) has reinforced the previous trend of refined spectral bandpass and the addition of new channels with 12-bit quantification. It has an improved cirrus channel that serves to detect better clouds. They are all compatible with each other and technological advances are incorporated as new models are launched. The applications offered by Landsat satellites have increased considerably with improved data quality since the launch of Landsat 8. However, the spectral reflectances of the OLI sensor show differences compared with those in the ETM+ sensor (Figure 9), so if they are blended these differences should be adjusted. Moreover, Landsat satellites are sensitive to atmospheric effects, which makes gap-free surface reflectance images difficult to obtain. Correction of the gaps over the land is needed to perform satellite applications on the Earth and for that, multiple atmospheric correction algorithms have been developed. One of them is the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) implemented for Landsat-7 and the Landsat Surface Reflectance Code (LaSRC) for Landsat-8. Moreover, recently, data quality has been improved by the fusion of sensors and using temporal series [50].

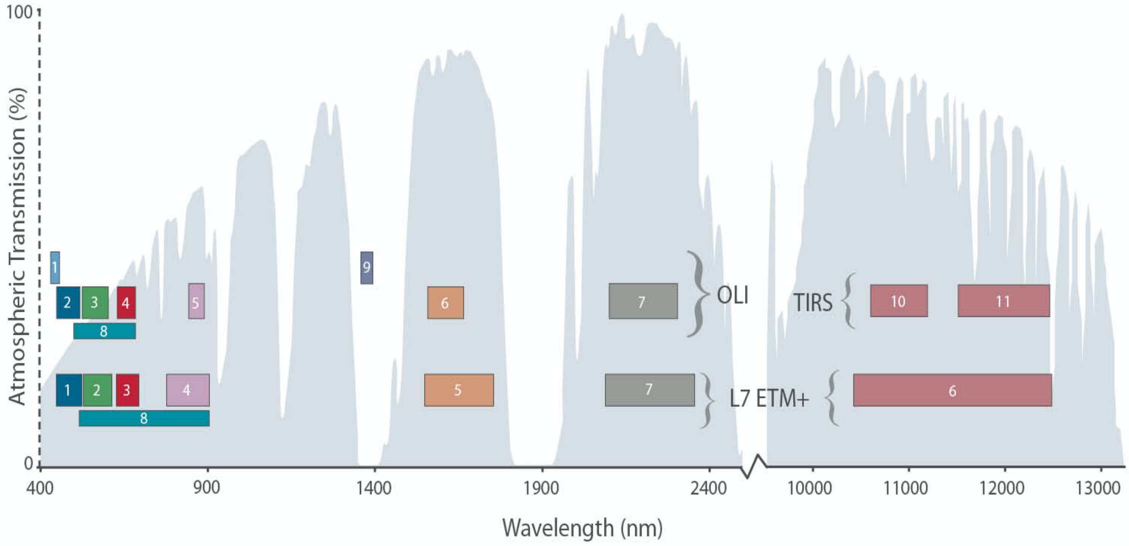


Figure 9 - Wavelength of Landsat 7 and 8 bands in the spectrum (USDA Landsat 8 - Wavelenghts image[8]).

[8] USDA Landsat 8 - Wavelenghts image: https://www.usgs.gov/media/images/landsat-8-wavelengths

On the other hand, data acquired by NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) are also used in the fusion process (more details in *Section 3.2.1. Data Fusion: HISTARFM*) to obtain the gap-filled monthly reflectances at 30 meters. MODIS instruments acquire data in 36 spectral bands. Their imagery is also transformed to obtain atmospheric correction, among others, by the proposed Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm [30]. MODIS data are extensively used along with the Landsat missions due to their consistency and their complementary specifications. While Landsat data provides 30 m of spatial resolution and a revisit rate of 16 days, MODIS is daily provided and depending on the spectral characteristics of interest, it has spatial resolutions of 250 m, 500 m, and 1000 m. Their fusion, when retaining the spatial patterns of Landsat images and the temporal regularity of MODIS, allows them to produce monthly gap-free high-resolution images. The Landsat satellites used in this work are Landsat-7, and -8, as they are the ones that coincide in time with the MODIS Terra and Aqua platforms that were launched in 1999 and 2002 respectively (Table 1).

Table 1 - Characteristics of Landsat and MODIS sensors used in this work.

Mission/Platform	Instrument	Time span	Revisit time (days)	Spatial resolution (m)
Landsat 7	Enhanced Thematic Mapper plus (ETM+)	1999 - present	16	30
Landsat 8	Operational Land Imager (OLI)	2013 - present	16	30
Terra	MODIS	2000 - present	1	500
Aqua	MODIS	2002 - present	1	500

For their fusion, the reflectances in the visible, the infrared (NIR), and short-wave infrared (SWIR) wavelengths will be combined. These are the most multispectral bands used in the scientific community, and they are needed for calculating vegetation indices. The range of reflectance covered by the two respective Landsat and two MODIS sensors is shown in Table 2 along with their spatial resolution. Both sets of images, Landsat and MODIS, can be obtained from the GEE platform.

Table 2 - Landat and MODIS spectral channels and band-passes used in the fusion.

Band Name	Landsat 7	Landsat 8	Terra	Aqua
Blue	0.45-0.52	0.45-0.51	0.46-0.48	0.46-0.48
Green	0.52-0.60	0.53-0.59	0.55-0.57	0.55-0.57
Red	0.63-0.69	0.64-0.67	0.62-0.67	0.62-0.67
Near Infrared	0.77-0.90	0.85-0.88	0.84-0.88	0.84-0.88
SWIR I	1.55-1.75	1.56-1.65	1.63-1.65	1.63-1.65
SWIR II	2.09-2.35	2.10-2.30	2.11-2.16	2.11-2.16
Resolution (m)	30	30	500	500

No se encuentran elementos de tabla de ilustraciones.

2.2.3. Weather data: Daymet

Crop stages are highly correlated with soil and plant evapotranspiration, external temperatures, radiation and other agrometeorological parameters [4]. For this reason, the classification accuracy can be improved by measuring these parameters. However, obtaining data from weather stations has two problems: i) they provide rainfall and temperature data, but they do not always include other important climate parameters for crop analysis. ii) their spatial coverage is not sufficient in large areas. Therefore, gridded weather data, as Daymet, are usually commonly used since they provide continuous and completed spatial weather parameters georeferenced in the CONUS [35].

Daymet data are used as an input variable to analyze the improvement of the accuracy and look for a robust classification model. Daymet dataset is supported by NASA through the Earth Science Data and Information System (ESDIS) and the Terrestial Ecology Program and it is available in GEE. The continued development of the Daymet algorithm and processing is also supported by the Office of Biological and Environmental Research within the U.S. Department of Energy's Office of Science. It contains daily weather estimates from January 1, 1980, to the most recent full calendar year and is provided with a spatial resolution of 1 km for the CONUS among other areas such as Canada and New Mexico. Daymet datasets are a daily weather data (1 - 365 days of the year (DOYs)), so they are provided at the beginning of the next year. Daymet includes 7 variables all of them used in this work: duration of the daylight period, daily total precipitation, incident shortwave radiation flux density, snow water equivalent, daily maximum and minimum temperature, and daily partial pressure of vapor. These variables are represented in Table 3 together with their specifications [43]. To avoid the curse of dimensionality problems and to be consistent with the reflectance data, the monthly variable values were calculated.

Table 3 - Description of the weather data parameters in Daymet [43].

Name	Description	Min*	Max*	Units
dayl	Duration of the daylight period. Based on the period of the day during which the sun is above a hypothetical flat horizon.	0	86400	seconds
prcp	Daily total precipitation, sum of all forms converted to water equivalent.	0	200	mm
srad	Incident shortwave radiation flux density, taken as an average over the daylight period of the day.	0	800	W/m^2
swe	Snow water equivalent, the amount of water contained within the snowpack.	0	1000	kg/m^2
tmax	Daily maximum 2-meter air temperature.	-50	50	°C
tmin	Daily minimum 2-meter air temperature.	-50	50	°C
vp	Daily average partial pressure of water vapor.	0	10000	Pa

2.2.4. Ancillary data: Climatic regions and counties

Mapping croplands over large areas is challenging because of the large volumes of data needed and the discontinuous availability of cloud-free data. Obtaining samples involves elevated high computational costs and time for training the algorithm and obtaining the subsequent classification map. For avoiding these issues, from a spatial point of view, the area of study can be reduced [53]. In this study, to simplify the calculations, two databases are used as ancillary data to split spatially the huge area of the CONUS in smaller regions. It is suggested to divide the CONUS into its 9 climatic regions and its 3007 counties. The aim is to select a few regions and study them separately (Figure 10). Both splits represent homogeneous divisions but with two different sizes and have been obtained with the computing platform GEE. The states which belong to the same climatic region are grouped according to table 4. In the case of the counties, they are directly downloaded from the shapefile "TIGER: US Census Counties 2018" offered by the U.S. Census Bureau and available in GEE. This shapefile identifies each entity by linking it with a geographic identifier in the data and is primarily used for censuses and surveys.

Table 4 - States corresponding to the same climatic region.

🟨	🟦	🟪	🟥	🟧	🟩	🟦	🟪	🟫
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Washington Oregon Idaho	Montana Wyoming North Dakota South Dakota Nebraska	Minnesota Iowa Wisconsin Michigan	Missouri Illinois Indiana Ohio Kentucky West Virginia Tennessee	Pennsylvania New York Vermont New Hampshire Maine Massachusetts Rhode Island Connecticut New Jersey Delaware Maryland	Virginia North Carolina South Carolina Georgia Alabama Florida	Kansas Oklahoma Arkansas Mississippi Louisiana Texas	Utah Colorado Arizona New Mexico	California Nevada

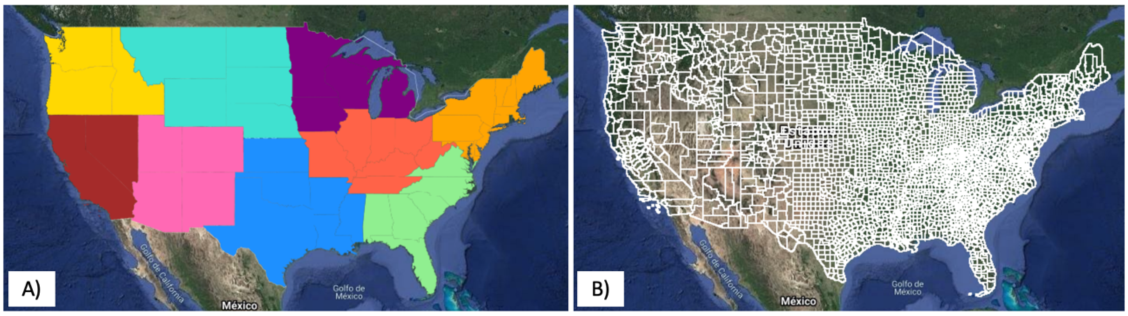


Figure 10 - Split of the CONUS taking into account the Climatic regions (A) and the counties (B).

3. Methodology

3.1. Introduction

The methodology of this work is based on a common processing chain in remote sensing fields. It covers the basic steps to generate classifications maps: fusion data, feature extraction/selection, classification and validation (Figure 11 in green). The process was developed in order to accelerate the classification looking for a spatial representativeness and the lower temporal acquisitions as far as possible. The reduction of spatial and temporal range was done losing the less information in the available data providing robust classification models.

The original data used in the training and test steps in the classification are the surface reflectance from the fusion of Landsat and MODIS sensors (described in the *Section 3.2.1. Data fusion: HISTARFM*), and weather data (for more information, see *Section 2.2.3. Weather data: Daymet*). The feature extraction step is focused on spectral indices (i.e. vegetation indices) that are highly recommended for distinguishing different agricultural crops. After that, a selection of features was done by means of a feature selection method focused on the similarity (or unsimilarity) of the features taking into account the crop classes. A benchmarking classification was done using the traditional machine learning supervised classification algorithms, and at last, they were validated by statistical measures.

The spatial reduction consists of reducing the study area from continental to regional magnitudes by selecting smaller areas within the CONUS whose crop distribution is similar to the crop distribution in the CONUS. Then the selected areas are classified (following the classification process). These results are considered to be similar as if the entire CONUS was classified (the spatial reduction methodology is represented in orange in Figure 11).

The temporal analysis consists of looking for the minimum period of time required to classify crops without loss accuracy in the results. This analysis was done in the CONUS as well as the best small areas obtained in the spatial analysis.

To assess the capacity of the model, a forecasting experiment over the CONUS is presented. It consists in obtaining crop type label data from 2018 for classifying in 2019. This experiment simulates a real case when CDL is not available in the year of classification.

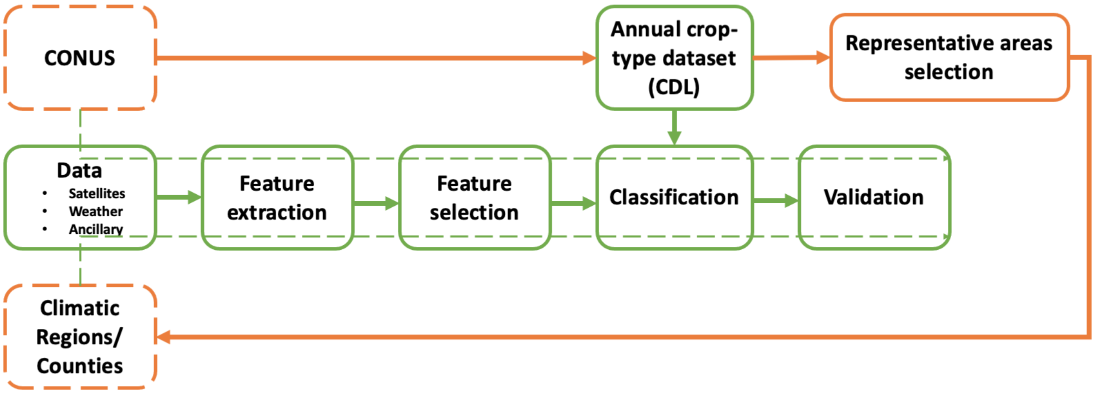


Figure 11 - Methodology flowchart. Orange: spatial reduction, Green: Steps of classification process.

3.2. Classification process

3.2.1. Data fusion: HISTARFM

Continuous and smoothed data is fundamental to the classification step since the gaps and missing data are a hindrance to obtain the classification maps. New models of data fusion are available currently and they allow us to obtain continuous regions with free cloud data. Fast and simple gap-filled methods such as Harmonic Analysis of Time Series (HATS, [17]) or Spatial and Temporal Adaptive Reflectance Fusion Model (STARFM, [16]) are restricted to small areas whereas new methods cover continental scales at high spatial resolution (e. g. Landsat temporal series).

The common way to obtain monthly smoothed spectral bands at 30 m is fusing Landsat and MODIS temporal series (see *Section 2.2.2 Satellite data: Landsat and MODIS* for more information of the data), however that is a great challenge. Focusing on crop mapping, especifically when mapping cropping cycles over large areas, acquiring fine resolution images has a difficulty because of unfavourable atmospheric conditions (clouds, aerosols, shadows, and strong angular effects). To acquire these images and avoid these problems, two solutions are proposed: i) take advantage of the existing high variety of sensors to combine them and thus mitigate individual limitations; ii) pre-process the images and improve the areas with noise and gaps [34].

In this work, we focus on combining Landsat and MODIS to obtain the spectral bands using the HIghly scalable temporal adaptive reflectance fusion model (HISTARFM) proposed by Moreno-Martínez, A et al. (2020) [34]. This model is based on the Kalman Filter (KF) proposed by Sedano et al. (2014) which is able to classify over large scales as it does not need a special parameter tuning [40]. The fusion images obtained by HISTARFM are free of spatial gaps due to clouds and aerosols and therefore, they are highly valuable for image processing applications such as crop image classification. The fusion steps are summarized as follows:

- Landsat and MODIS images are aggregated in monthly temporal resolution by calculating their mean values.
- A Bayesian estimator is used to predict Landsat observations for a given month. It blends the climatology of Landsat reflectances and MODIS reflectances. For that, first MODIS reflectances are downscaled applying a Landsat-MODIS fusion using a pixel-wise linear regression model. The outputs of the combination are used to estimate reflectance means and covariances.
- The reflectance predictions of the Bayesian estimator are then corrected by a Kalman Filter. It avoids errors from the Bayesian model using a bias correction. Two estimators are used because sometimes forecastings can lead to errors if the prediction models are biased. Finally, an unbiased mean reflectance value and its covariance error are estimated.

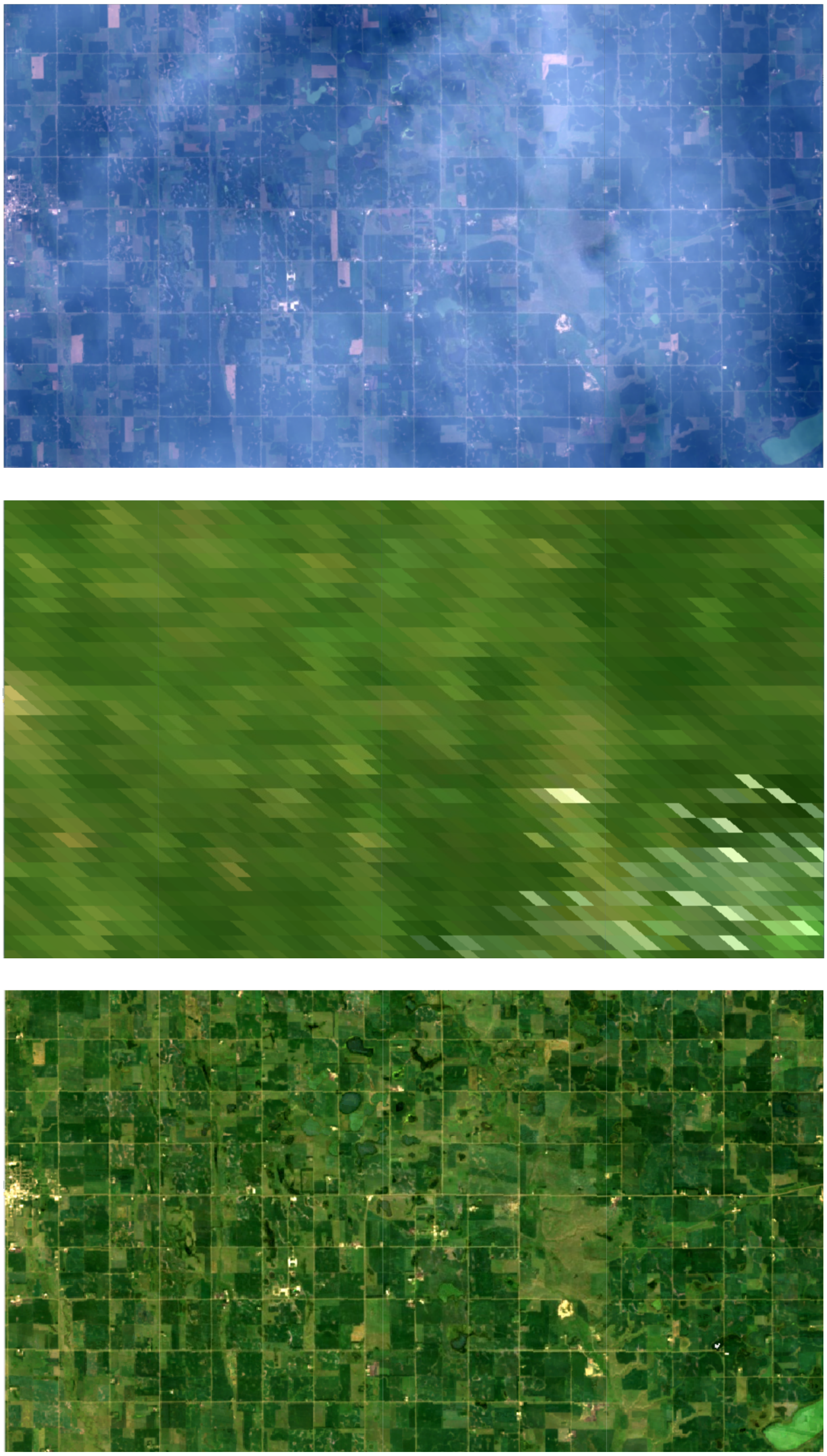


Figure 12 - Images of cropland cover at the same location from Landsat 8 (top), MODIS Terra (middle), and the fusion Landsat-MODIS with HISTARFM (botton). All images were obtained from GEE.

Six spectral bands (B1, B2, B3, B4, B5, and B7) gap-free monthly reflectance Landsat are obtained at 30 m spatial resolution. All these images are available in GEE[9] (with a user account) and they

[9] https://code.earthengine.google.com/?asset=projects/KalmanGFwork/GFLandsat_V1

were directly used from the platform to download the surface reflectance values for the training and testing datasets. Figure 12 shows an example of Landsat and MODIS images in August 2018, and the resultant image from their fusion with HISTARFM.

3.2.2. Feature extraction

The state-of-art of crop classification shows different approaches to increase the accuracy [1] [55] [56] [57]. These works have improved the results including informative features into the training data to train the supervised classifiers. Vegetation indices are ones of the most informative features in crop classification [23] so the combination gap-filled bands obtained by the HISTARFM method is used to obtain vegetation indices as new features for the classification. The near-infrared (NIR) is widely used to study the plant canopy along with important vegetation parameters such as water, pigments, carbohydrates, proteins, and other content [3]. Applications are also derived from the relation between NIR and visible, especifically between NIR and red. From this relation is created the Normalized Difference Vegetation Index (NDVI) which is the vegetation index VI most used [44]. However, using only single combinations of bands results in a lack of sensitivity. This limitation is even more noticeable when studying heterogeneous canopies because of their diversity. In this work several vegetation indices from different band combinations are used: basic VI, VI considering atmospheric effects, adjusted soil VI, etc. Finally, a total of 35 VI have been extracted from the spectral bands (see Table 5, [52]).

Table 5 - Vegetation Indices [52].

Index	Name	Definition	Other
BGI 2	Blue Green Pigment Index 2	$BGI2 = \frac{B}{G}$	
CRI500	Carotenoid Reflectance Index	$CRI500 = \frac{(1/B)}{(1/G)}$	
GDVI	Green Difference Vegetation Index	$GDVI = NIR - G$	
DVI	Difference Vegetation Index	$DVI = NIR - R$	

EVI	Enhanced Vegetation Index	$EVI = 2.5 * \frac{NIR - R}{NIR + C1R - C2B + L}$	C1=6
			C2=7.5
			L=0.5

ExG	Excess Green Index	$EXG = 2G - R - B$	
GEMI	Global Environmental Vegetation Index	$GEMI = (1 - 0.25) - \frac{(R - 0.125)}{(1 - R)} = \frac{2(NIR^2 - R^2) + 1.5NIR + 0.5R}{NIR + R + 0.5}$	
GLI	Green Leaf Index	$GLI = \frac{2G - R - B}{2G + R + B}$	
GNDVI	Green Normalized Difference Vegetation Index	$\frac{NIR - G}{NIR + G}$	
GRVI	Green Ratio Vegetation Index	$GRVI = \frac{NIR}{G}$	
Greenness index (G)	Greenness index	$G = \frac{G}{R}$	
IPVI	Infrared Percentage Vegetation Index	$IPVI = \frac{NIR}{NIR + R}$	

MCARI	Modified Chlorophyll Absorption Ratio Index	$MCARI = \frac{1.5 * [2.5(NIR - R) - 1.3(NIR - G)]}{\sqrt{(2NIR + 1)^2 - (6NIR - 5R) - 0.5}}$	
MNLI	Modified Nonlinear Index	$MNLI = \frac{(NIR^2 - R)(1 + L)}{NIR^2 + R + L}$	L=0.5
MSAVI2	Modified Secondary Soil-Adjusted Vegetation Index	$MSAVI2 = 0.5 * \left[(2NIR + 1) - \sqrt{(2NIR + 1)^2 - 8(NIR - R)}\right]$	
MSI	Moisture Stress Index	$MSI = \frac{SWIR1}{NIR}$	
MTVI	Modified Triangular Vegetation Index	$MTVI = 1.2 * [1.2 * (NIR - G) - 2.5(R - G)]$	
MTVI2	Modified Triangular Vegetation Index	$MTVI2 = \frac{1.5 * [1.2(NIR - G) - 2.5(R - G)]}{\sqrt{(2NIR + 1)^2 - (6NIR - 5R) - 0.5}}$	
NDGI	Normalized Differential Greenness Index	$NDGI = \frac{G - R}{G + R}$	
NDVI	Normalized Vegetation Index	$NDVI = \frac{NIR - R}{NIR + R}$	
NDWI	Normalized Difference Water Index	$NDWI = \frac{NIR - SWIR1}{NIR + SWIR1}$	
NGBDI	Normalized Green-Blue Difference Index	$NGBDI = \frac{G - R}{G + B}$	
NGRDI	Normalized Green-Red Difference Index	$NGRDI = \frac{G - R}{G + R}$	
NMDI	Normalized Multi-Band Drought Index	$NMDI = \frac{NIR - (SWIR1 - SWIR2)}{NIR + (SWIR1 - SWIR2)}$	
NLI	Non-linear Vegetation Index	$NLI = \frac{NIR^2 - R}{NIR^2 + R}$	
OSAVI	Optimized Soil-Adjusted Vegetation Index	$OSAVI = \frac{(1 + X)\ (NIR - R)}{NIR + R + X}$	X=0.16
$PSND_C$	Pigment Specific Normalized Difference c	$PSND_C = \frac{NIR - B}{NIR + B}$	
$PSSR_C$	Pigment Specific Simple Ratio for Carotenoids	$PSSR_C = \frac{NIR}{B}$	
RVI	Ratio Vegetation Index	$RVI = \frac{R}{NIR}$	
SAVI	Soil-Adjusted Vegetation Index	$SAVI = \frac{(NIR - R)(1 + L)}{(NIR + R + L)}$	L = 0.5
SGI	Sum Green Index	$SGI = \frac{NIR}{R}$	
SR2	Simple Ratio 2	$SR2 = \frac{NIR}{G}$	
TDVI	Transformed Difference Vegetation Index	$TDVI = \sqrt{0.5 + \left(\frac{NIR - R}{NIR + R}\right)}$	
VARI	Visible Atmospherically Resistant Index	$VARI = (G - R)(G + R - B)$	
VDVI	Visible-Band Difference Vegetation Index	$VDVI = \frac{2G - R - B}{2G + R + B}$	

3.2.3. Feature selection

Feature selection is a previous step in a classification process to reduce the dimensionality excluding redundant information and selecting the most relevant features [5]. This reduces computational complexity in the classification by reducing the number of input features. In this case, the feature selection is performed for the three input datasets: spectral features, spectral indices, and weather data. There are many methods to select features grouped as: a) filter methods, b) wrapper methods, and c) embedded methods. In this study is used a filter method, that means that only the subset of relevant features is taken. It is based on the coefficient of correlation of Pearson (R) that is defined by the relation between the covariance of the features divided by the product of their standard deviations (expression 1). This parameter measures the linear correlations and ranges between -1 and 1, meaning 0 no correlation and 1 and -1 positive and negative correlations, respectively. The correlation between all the features, taking into account the crop classes, is calculated and the features with the largest average R (the most correlated) are discarded [21]. A threshold is applied to the correlation, in this case the value is 2.5 times the standard deviation of the data for the three datasets. Features whose correlations raise this value are discarded. If the threshold increases less features are selected and if the threshold decreases, the number of features is higher.

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} * C_{jj}}} \tag{1}$$

* Where *C* refers to the covariance matrix. The covariance indicates the level to which two variables vary together (Numpy documentation[10]).

3.2.4. Classification

The efficiency of the results depends on the classification performance. Different types of classifiers are in the literature and they are split depending on either if they use labels in the training step: *supervised*, *semi-supervised* and *unsupervised* or whether they estimate statistics parameters from the training data: *parametric* and *non-parametric*. Several works are focused on supervised and *non-parametric* models to classify crops in remote sensing fields. In this thesis, the benchmarking of *supervised* and *non-parametric* classifiers is analyzed. The two most well-known machine learning classifiers in remote sensing [37] [15] are chosen for the analysis: Random Forest and Support Vector Machine.

3.2.4.1. Random Forest

Random forest (RF) is a supervised algorithm used in classification or regression tasks. RF is an ensemble of decision trees (i. e. ensemble classifier) that are fully grown and not pruned. RF ensemble avoids the overfitting in new samples of the decision tree classifier and combines its results by the maximum vote rule (i. e. selecting the most popular class) [7]. RF is fast and easy to implement providing accurate predictions even with high dimensionalities. The bagging process avoids overfitting because the variance in the classification is diminished. Moreover, it

[10] Numpy documentation: https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html

provides insight on the importance of each feature, and there is no need of preprocessing the data nor pruning [55].

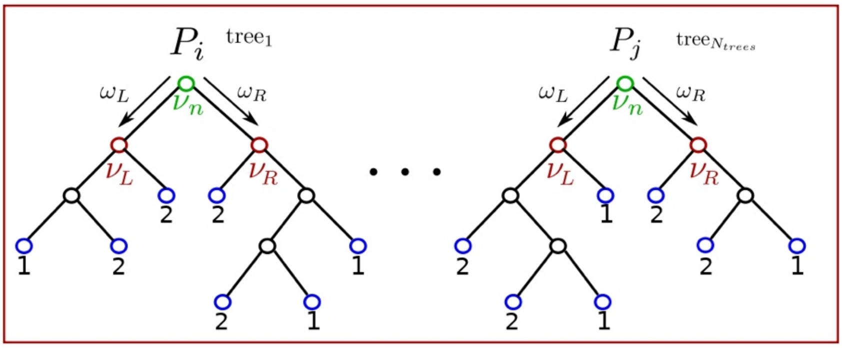


Figure 13 - Random Forest classifier with Pj partitions. Each tree has a root node vn and internal nodes (vL and vR). The partitions are recurrent split until reaching the terminal nodes (in blue). Each sample of the partition gets assigned a label (classes 1 or 2 in this case). ω_L and ω_R are the fraction of samples that fall in each node.

Two subsets of data are required in the RF classifier, train and test data. In each tree of the RF (see Figure 13) a different subset of training is chosen taking as many samples as in the original training set, but randomly and with replacement, which is known as bootstrapping. That means that samples can be used more than once in the same tree. The samples that are not included in the bootstrap form the out-of-bag (OOB) and are used to calculate the OOB error. All trees follow the same process as an unpruned decision tree. From a partition of the original data, the best feature and cutoff is selected by means of the Gini Index. Therefore, the number considered a split candidate in each node (i.e. the number of features of the data partition) has to be optimized as well as the number of trees in the RF. But additionally, the number of minimum samples required to be a leaf node and the minimum number of samples required to split an internal node are optimized.

Once the algorithm is trained, the testing data is used to validate the classification. For that, all the trees are aggregated, and the final classification is determined by the majority vote of all them. Table 6 shows the four RF parameters to be optimized in this work and their range of values.

Table 6 - Random Forest parameters

Parameters	**Abbreviation**	**Range**
Number of decision trees	N_{tree}	50 - 200
Number of features to split in each node	mtry	2 - X_{train}
Minimum samples leafs	$min_{samples\_lf}$	1 - 4
Minimum samples splits	$min_{samples\_sp}$	2 - 4

In this work, to generate a RF classifier, the RF is trained with a set of samples (training set) and is tested in an independent set of samples (test set). To avoid the randomness in the results of RF, ten repetitions per month are done, and the mean value is obtained.

3.2.4.2. Support Vector Machines

In classification, the classes can be linearly (Figure 14 A) or nonlinearly (Figure 14 B) separated. Remote sensing data usually are non-linear separated and for that reason, Support Vector

Machines (SVM) is a robust classifier method in the field [23]. The basic idea of the SVM is to map the data into a high dimensional space (called Hilbert space) and find the best hyperplane which splits the classes of the data linearly. The hyperplane is found by means of the Support Vectors (SV); these are the closer samples of the different classes to the hyperplane (Figure 14 C). The objective of SVM is to look for the hyperplane with maximum distance (i.e. large margin) to the SV without misclassifying the training samples. However, the SVM has the possibility to be permissive with the misclassifications by the parameter C. High values of C provide lower margin and then, the classification of the training samples correctly are the most important assuming that the training set contains the maximum variability. The main problem with it is the high possibilities to get overfitting. Otherwise, low values of C provide a larger margin of the hyperplane assuming a high number of misclassification samples in the training.

SVM always provides binary classifications either one-against-rest for multi-class cases or using only two classes. In this thesis, it is used the one-against-rest approach because seven classes are studied at once. Each class is discriminated from the others by a binary classifier.

Taking into account that the transformation to the Hilbert space is unknown, SVM uses a trick (i.e. kernel trick) to obtain the similarity between samples in the Hilbert space. The similarity is calculated by the dot products of the samples, that is the *kernel function*. Whether the kernel function is linear, the SVM provides a linear classifier. However, if the kernel function is non-linear, the SVM provides a non-linear classifier. Several kernel functions are able to be used in the SVM, but the most common is the Radial Basis Function (RBF). RBF is defined as:

$$k(x_i, x_j) = exp\left(-\frac{d(x_i, x_j)^2}{2\sigma^2}\right) \tag{2}$$

* Where here σ is the bandwidth of the kernel and *d* is the euclidean distance.

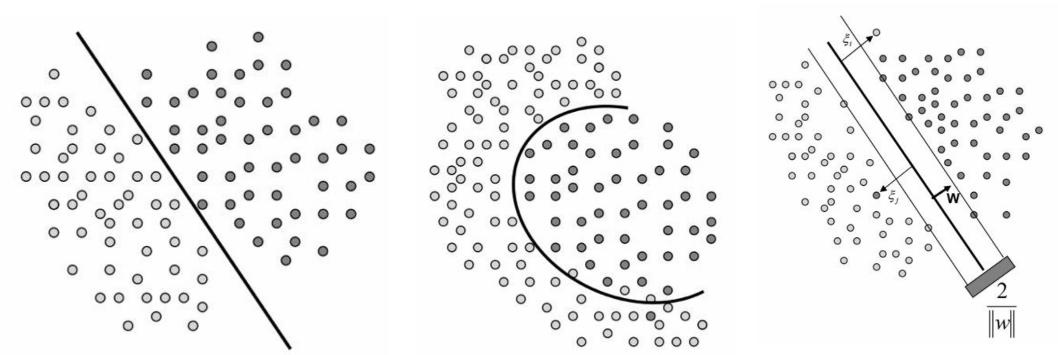


Figure 14 - From left to right: Lineal data (A), Non-linear data (B), Hyperplane (C).

In this work, the separation of the dataset into training and testing subsets is repeated 10 times per month and then the mean value is calculated. Moreover, two parameters of the SVM have to be optimized (see Table 7): the cost (C) and the bandwidth of the kernel (σ). Note that the bandwidth parameter is related with the γ value as follows:

$$\gamma = 1/(2 * ([0.5 - 30] * \sigma)^2 \tag{3}$$

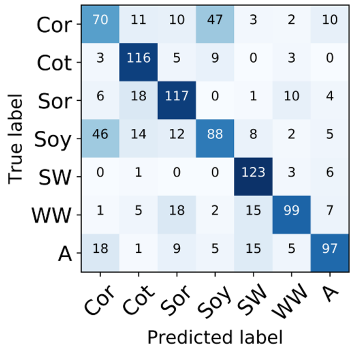
Table 7 - Support Vector Machine parameters.

Parameters	**Abbreviation**	**Range**
C	C	0.1 - 1000
Gamma	γ	$1/(2*([0.5-30]*\sigma)^2$

3.2.5. Validation

In remote sensing assessing the classification accuracies is an important task and there are many measures developed to calculate it. Moreover, the consistency of using a set of measures is better than using only one. To assess the accuracy during the training of both algorithms are calculated the Confusion Matrix, the Overall Accuracy (OA), and the Kappa Coefficient using the Sklearn Python library. The accuracies are used to analyze results from different experiments depending on the input data, area of study, and classification algorithms.

Confusion Matrix: in Figure 15 appears an example of a confusion matrix, the columns represent the reference data and are compared with the rows which are the predicted data obtained by the classifier. The main diagonal of the confusion matrix represents the data that is correctly assigned to its label [42].



	Cor	Cot	Sor	Soy	SW	WW	A
Cor	70	11	10	47	3	2	10
Cot	3	116	5	9	0	3	0
Sor	6	18	117	0	1	10	4
Soy	46	14	12	88	8	2	5
SW	0	1	0	0	123	3	6
WW	1	5	18	2	15	99	7
A	18	1	9	5	15	5	97

Figure 15 - Example of confusion matrix. (Notation: Cor: corn, Cot: cotton, Sor: sorghum, Soy: soybeans, SW: spring wheat, WW: winter wheat, and A: alfalfa).

Overall Accuracy: is the sum of the diagonal of the confusion matrix divided by the total of elements. In other words, it is the probability that a new individual will be correctly classified. Given the confusion matrix N = (n_{ij}) the overall accuracy is defined by the expression 4 [18]:

$$O^c = \frac{\sum_{i=1}^{k} n_{ii}}{|T|} \tag{4}$$

* Being |T| the number of testing samples.

Kappa Coefficient: follows the Cohen's kappa [10] shown in expression 5. It measures the observed correct classifications against if they were classified randomness. The value of Kappa is calculated by:

$$k = (p_0 - p_e)/(1 - p_e) \tag{5}$$

* Where p_0 is the empirical probability of agreement on the label assigned to any sample the overall accuracy) and represents the percentage of samples correctly classified by chance. Table 8 represents the quality of the classification algorithm for the kappa values.

Table 8 - Meaning of Kappa values.

Kappa	Interpretation
< 0	Poor
0 - 0.2	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1	Almost perfect

3.3. Case studies

The classification process explained in the previous section was used to perform different case studies. The first case study is the spatial reduction that focuses on selecting smaller areas within the CONUS with a crop distribution similar to the entire CONUS. Next is proposed a temporal analysis to identify the month at the earliest when a near-real time crop classification map can be obtained. The last case study is the forecast analysis to analyze how the best results of the previous cases would perform using training samples from one year before the testing takes place.

3.3.1. Spatial reduction: crop distribution in representative areas

The main objective of this analysis is to simplify the obtention of training samples over the CONUS from a spatial point of view. The main idea is to find climatic regions and/or counties (see *Section 2.2.4. Ancillary data: Climatic regions and counties*) whose crop areas represent the whole in the CONUS. The spatial reduction is interesting for two reasons: i) reducing the sampling area diminishes considerably time and economic costs of sampling, and ii) it allows to consider the possibility of studying the methods of this work in other continental areas in the Earth where an annual cropland dataset is not available.

In this study, the CDL is used to determine the crop distribution in the CONUS in 2018. Within the CDL, there is a long list of land covers, but to simplify the study only are considered the seven most extensive crops: corn, soybeans, winter wheat, alfalfa, spring wheat, cotton, and sorghum. The percentage of each crop is calculated over the area in the CONUS, climatic regions, and counties using the expression 6. Finally, areas with similar percentages to the CONUS are selected.

$$Crop\ over\ area * \ (\%) = \frac{number\ of\ pixels\ of\ the\ crop\ in\ the\ area *}{total\ number\ of\ pixels\ in\ the\ area *} x100 \tag{6}$$

* Where "Crop over area" is either referred to the CONUS, each climatic region or each county.

Once the percentage of the crops is calculated in the CONUS, and in the climatic regions and counties, it is possible to determine which of them are representative for each crop (see Figure 16). The reduction magnitude from the CONUS size to the representative areas is noticeable.

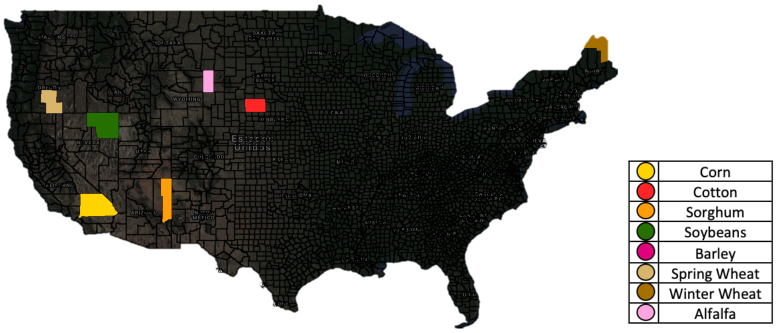


Figure 16 - Example of crop representative areas in the CONUS.

To define which areas are most representative for each crop a statistical method is applied: for each crop of interest is selected the region whose crop percentage is closer to the CONUS percentage and its difference is less than 5%. At the end, each crop has its most representative area.

In order to assess if these counties are still representative in the following years, it is calculated their temporal correlation with the CONUS for ten years (2009-2018). This correlation is based on crop area percentages and is calculated using the coefficient of determination R^2.

3.3.2. Temporal analysis: prompt crop classification mapping

Prompt classification mapping is an important issue in agricultural crop mapping. Government agencies or insurance companies, as few examples, are interested in crop classification as soon as possible. Therefore, the analysis of this case study is focused on knowing the earliest month to crop mapping without loss accuracy in the classification process. Crop classification maps are influenced by the type of the input dataset, machine learning algorithm, and reference data. In this work, the best combination of these different variables has been studied.

Regarding the input data set, the analysis is focused on the CONUS and the representative areas. In both cases, three types of categories are considered: 1- spectral features (SF), 2- SF and spectral indices (SI), and 3- SF, SI and weather data (WD). The SF are the gap-filled bands obtained by the HISTARFM method (*Section 3.2.1. Data Fusion: HISTARFM*), the SI are the vegetation indices extracted using the SF (*Section 3.2.2. Feature Extraction*) and the the WD are the Daymet weather data (*Section 2.2.3. Weather Data: Daymet*). As Figure 17 shows they are combined to be evaluated using RF and SVM classifiers.

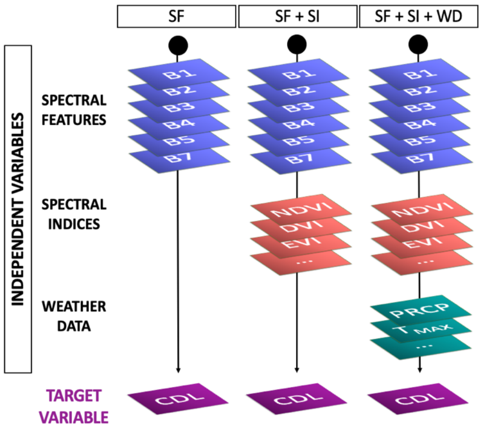


Figure 17 - Input datasets. SF means spectral features; SI means spectral indices; and WD means weather data.

In order to obtain the earliest month to the crop classification, accumulative months are added in the classification process in the three categories of the input data. The earliest month is going to be identified by the OA of the classifiers since adding new information (i.e. adding subsequent months) would not make a huge difference in the results. After that, a feature selection method is applied to select the most important features in the classification. Finally, the best category combination and the earliest month are used to generate the classification map.

3.3.3. Forecasting analysis: near-real time prediction

Previous case studies were focused on the spatial and temporal analysis of a specific year (2018) in order to know the best area, input data, earliest month, and classifier. Therefore, once the best conditions (data, area, algorithm and time) are selected, they are used to classify the crops doing a forecasting experiment (Figure 18). It consists in using training data from 2018 and then classifying in 2019, assuming that this is the current season year. The objective is to avoid the collection of training data by using target labels (CDL) from the year before. The assessment of the classifier is done using a test data set obtained randomly over the CONUS in 2019. Additionally, the classification map of 2019 is generated for the crops: soybeans, corn, cotton, spring wheat, winter wheat, sorghum and alfalfa, and the rest are masked.

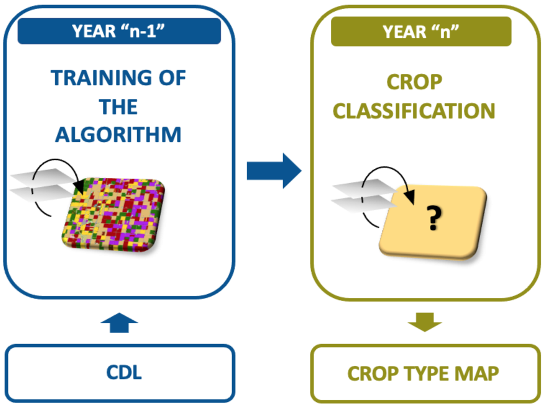


Figure 18 - Flowchart of the forecasting case.

4. Results

This chapter presents the results obtained for the three study cases described in the methodology chapter: the search of the representative areas (spatial reduction), the prompt crop classification mapping (temporal analysis) and the real time prediction (forecasting analysis). In the three cases the classification process, explained in the *Section 3.2. Classification Process* is used to compare the classification and/or to predict the classification maps. As the classification process is highly affected by the train and test selection, the analysis of the training and test selection is also shown.

4.1. Spatial reduction

Two approaches are developed to find representative small regions within the CONUS in 2018. The first is based on finding areas where the percentage of crops is similar to the overall CONUS; the second consists in studying if these small regions have a similar temporal evolution of crop areas as the CONUS. The first case is used to analyze how the representative areas work using the available information of the year to be processed. The second serves to check if the selected regions can be used in following years.

4.1.1. Crop percentages

In 2018 the CONUS area was covered at 16.8% by crops and the rest was filled by other land covers such as shrublands, grasslands/pastures or forests (called non-crops). In this work only are studied the most expansive crops in 2018: soybeans, corn, winter wheat, alfalfa, spring wheat, cotton, and sorghum. "Other-hay/non-alfalfa" and "fallow/idle-cropland" were also included in the nine more extensive cropland covers, but they do not represent specific crop types. According to Figure 19 even the crops of interest are seven over 108 in total, they occupied rather more area (14,21%) than the others (1,7%). It is also observed that the most extensive crops are soybeans and corn having an important weight in the final crop productions. The percentages of these seven crops are also in Figure 20.

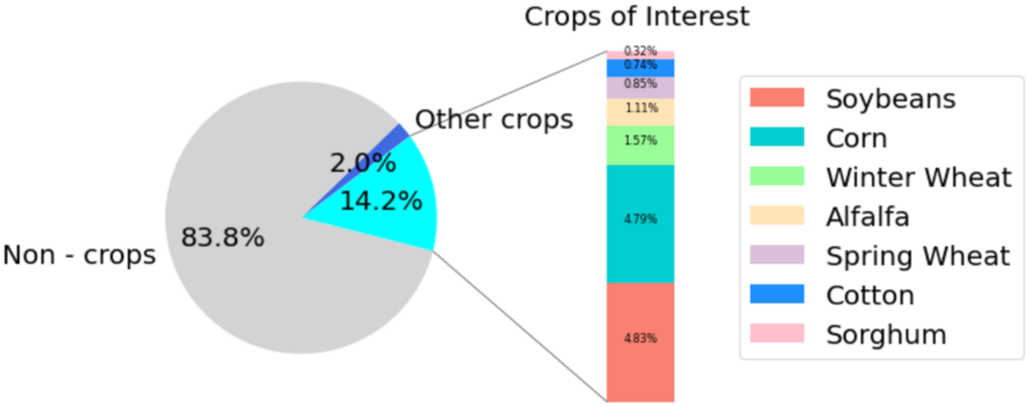


Figure 19 - Non-crops, interest crops and other crops in the CONUS for 2018.

For each crop of interest, it is going to select the climatic regions with the closest crop percentage to the CONUS in 2018. However, only alfalfa class has a difference of percentage lower than 5% compared to the CONUS for the climate region 5 (Northeastern region), whereas the rest of classes are out of the established threshold (5%). Therefore, the climatic regions are not representative of the crop area percentages in the CONUS. Figure 20 shows the climatic regions and counties with the closest crop percentages to the CONUS. It is noticeable that climatic regions are far from CONUS results. This happens because the climatic regions have larger areas than the counties and the crop classes are concentrated in determined areas in the CONUS. This leads to an excess or shortage of crop acreage percentages of the climatic regions compared to all the CONUS.

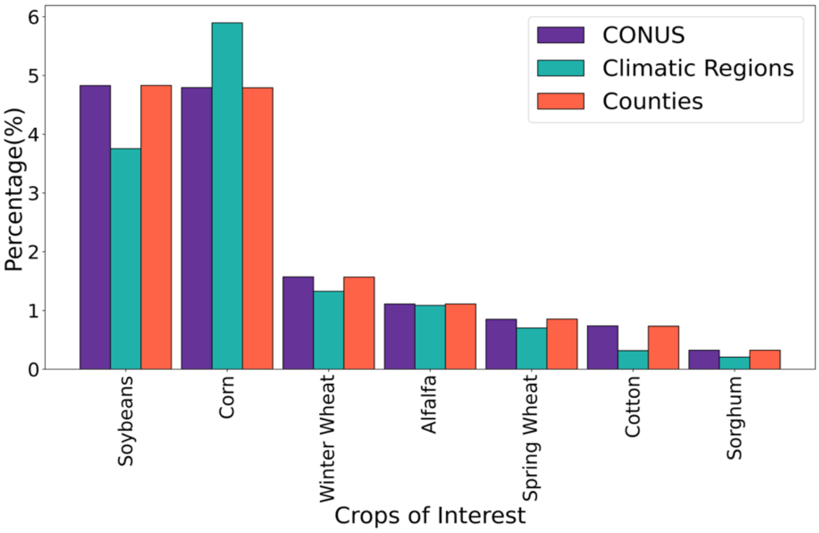


Figure 20 - Comparison of crop percentages in the CONUS and, in the closest climatic region/county for the crop of interest in 2018.

However, as Table 9 shows, there are seven counties (one per crop of interest) whose percentages only differ 0,7% from the respective crop percentage in the CONUS. These results are far better than the proposed threshold (5% of difference). So, these seven counties serve as representation of the crop percentages in the CONUS. Figure 21 shows the spatial soybeans example for the CONUS and for the Waukesha county. As we see the distribution of soybeans are centered mainly in the northeastern part of the country, close to the Big Lakes. Waukesha county has a random distribution of soybeans in the county (i.e. there is no predominant area in the county). In the following sections the counties Waukesha, Baker, Midland, Madera, Lake, Bowie, and McPherson are chosen as representative areas of the CONUS and then, they are used in the classification process to analyze the prompt classification and compare the results with the CONUS.

Table 9 - Comparison of CONUS and counties crop percentages (2018).

Crop	Percentage CONUS (%)	County Name	Percentage Counties (%)	Difference (%)
Soybeans	4,828538	Waukesha County	4,830390	0,038365
Corn	4,792732	Baker County	4,790296	0,050803
Winter Wheat	1,572511	Midland County	1,569186	0,211472
Alfalfa	1,110682	Madera County	1,111094	0,037088
Spring Wheat	0,849714	Lake County	0,853225	0,413214
Cotton	0,737493	Bowie County	0,732842	0,630627
Sorghum	0,322989	McPherson County	0,322623	0,112990

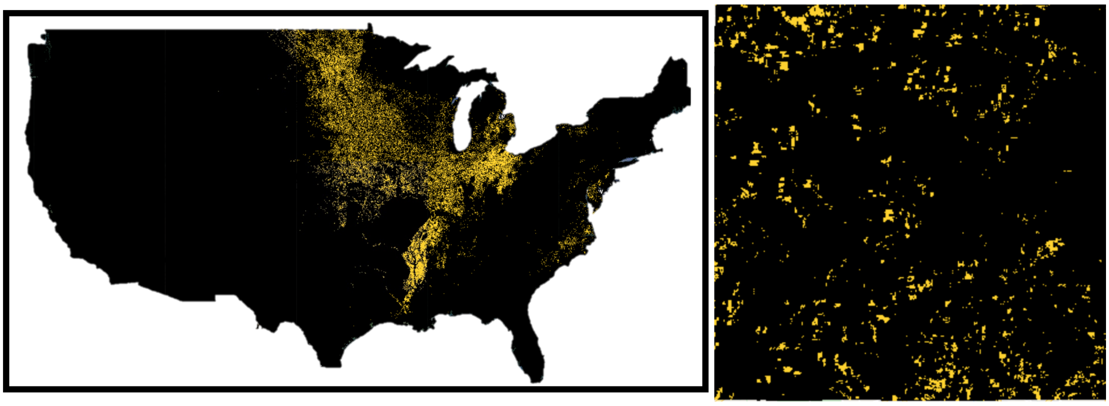
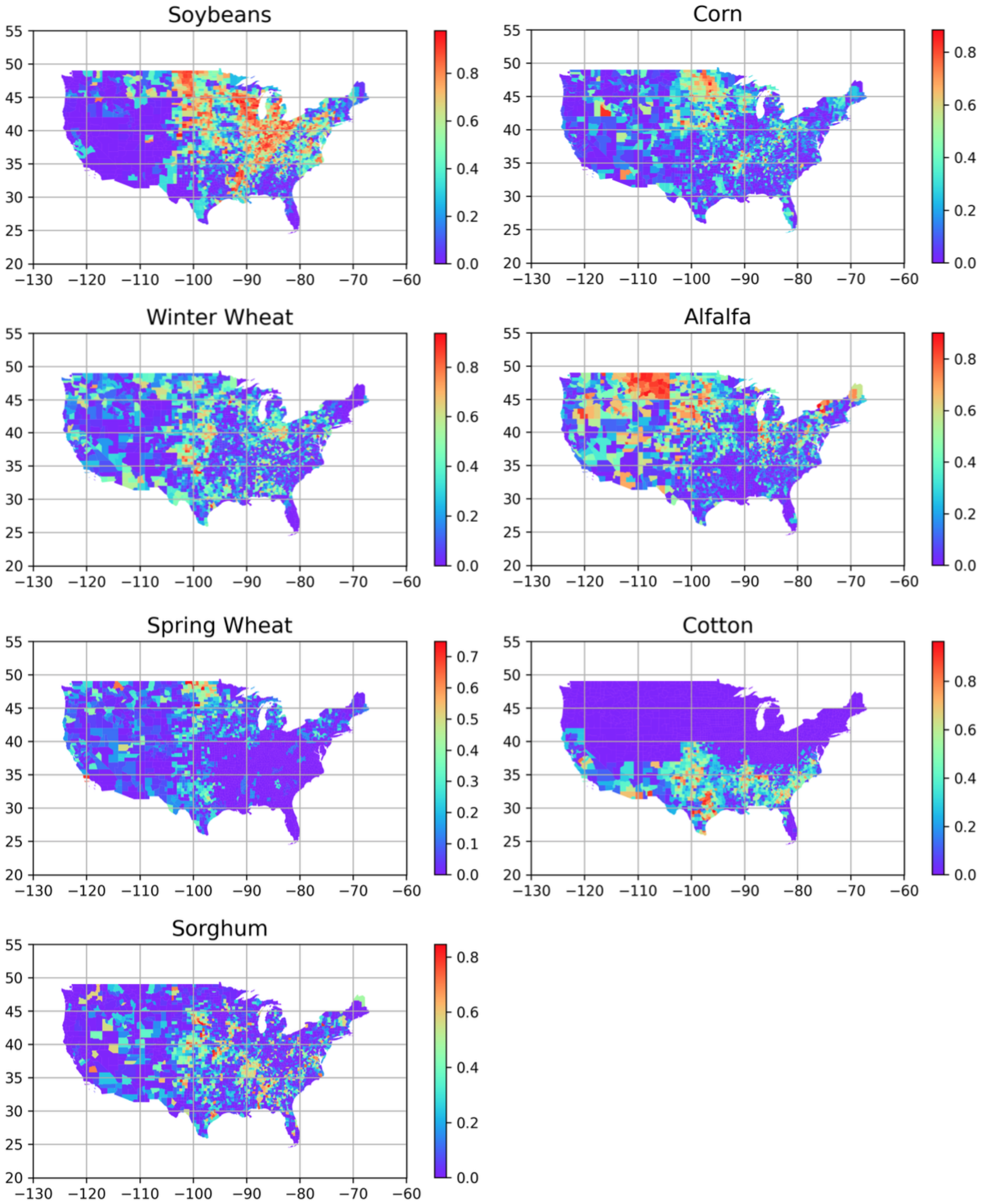


Figure 21 - Soybeans distribution in the entire CONUS (left) and Waukesha county (right) for 2018.

4.1.2. Temporal correlation

In this approach, the coefficient of determination R^2, is calculated to extract the correlation between the counties and the CONUS in the last 10 years in terms of crop area with the CONUS. Figure 22 shows the R^2 for all the counties and the CONUS for each crop of interest. The correlation in many counties follows similar behaviour as the CONUS does. It's worth mentioning that the R^2 of soybeans class is higher than 0.9 in quite a few counties, which are distributed in all the east of the CONUS. In the case of Alfalfa class, R^2 is also quite high, especially in Montana state. Cotton class is focussed on determined areas in the south of the US, but the class still has several counties with R^2 overcoming 0.9. As well as the winter wheat class that has few counties whose R^2 superior to 0.9 although most of them do not exceed 0.7. The corn and sorghum classes have the best correlations around 0.8 in counties scattered within the CONUS. And spring wheat class presents the worst correlations with a maximum equal to 0.75. This means all the interest crops have counties with high correlation area percentages regarding to the CONUS.



Soybeans

Corn

Winter Wheat

Alfalfa

Spring Wheat

Cotton

Sorghum

Figure 22 - Coefficient of determination (R2) between the area occupied for each crop in the CONUS and counties from 2009 to 2018. The colorbar represents the value of R2.

Focusing on the selected counties in the previous section, Table 10 shows the R^2. In these cases, only the area of soybeans class in Waukesha county has a correlation higher than 0.6 with the CONUS between 2009 and 2018. And taking into account that the soybeans area difference between the county and the CONUS is 0.04%, one may conclude that the soybeans area percentage would be similar in 2019 over Waukesha and all the CONUS.

Table 10 - R^2 of the area crop percentage between CONUS and counties between 2009 - 2018.

Crop	County name	R^2
Soybeans	Waukesha County	0.61
Corn	Baker County	0.12
Winter Wheat	Midland County	0.41
Alfalfa	Madera County	0.29
Spring Wheat	Lake County	0.00
Cotton	Bowie County	0.29
Sorghum	McPherson County	0.09

Note that, the following case studies are focused on one year (i.e. the processing year). Therefore, the classification benchmarking is presented in the next sections with all the CONUS and the selected counties (previous section), although the temporal correlation with the CONUS are not too high.

4.2. Training and testing data

The first classification step requires defining a representative training and test sets from the classification area. In this work, a set of 3500 samples from the CDL was randomly selected and after that, the 30% of this set were used for training the classifiers and the rest were used for testing (i.e. 70%). This process was done for both, the CONUS and the previously selected counties. Figures 23 and 24 show the random selection of the 3500 samples (500 points per class) represented in QGIS software. Figure 23 represents the training and testing subsets of all the CONUS. The figure shows that spring wheat samples are mostly placed in the north, alfalfa in the west, soybeans and corn in the northeast, cotton in the south, and sorghum and winter wheat in the centre of the CONUS. That is because the classes are influenced by the location in the CONUS. This is logic because these crops are mainly produced in those zones. Even that, each class also has point samples in other zones along the CONUS which justifies their representativeness. It is also shown that training and testing points are located in different plots leaving enough distance to have differences between training and testing datasets.

Figure 24 represents the training and testing data over the selected counties (Waukesha, Baker, Midland, Madera, Lake, Bowie, and McPherson). 500 samples are obtained from smaller regions as are the seven counties and as we see in the figure, they are sufficiently separated between training and testing points.

Relating to the counties, two ways were used to select the training and test samples. In the first one (the mixed case), all the classes were selected from different counties as shown in Figure 24. In the second one (one class per county case), training data was chosen selecting each class from the county where the class is the main class. The objective of that is to get samples for a specific class only in small regions and generate a classifier for predicting in the entire CONUS. For each class the training samples are selected on its most representative county following the results in *Section 4.1.1. Crop percentages* and table 9.

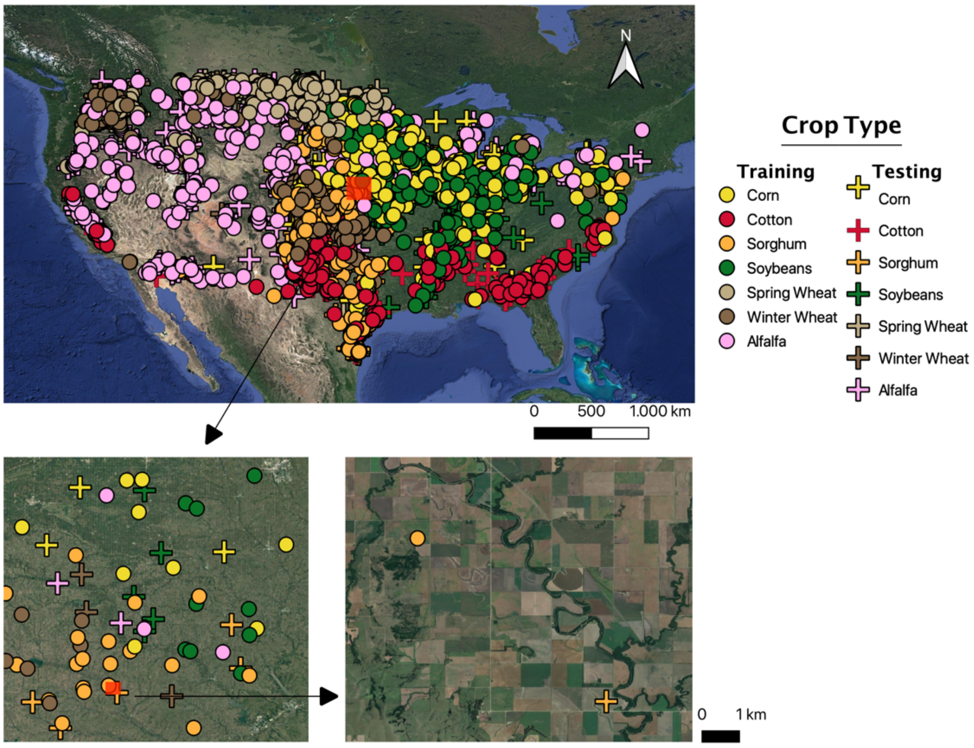


Figure 23 - Training and testing samples over all the CONUS.

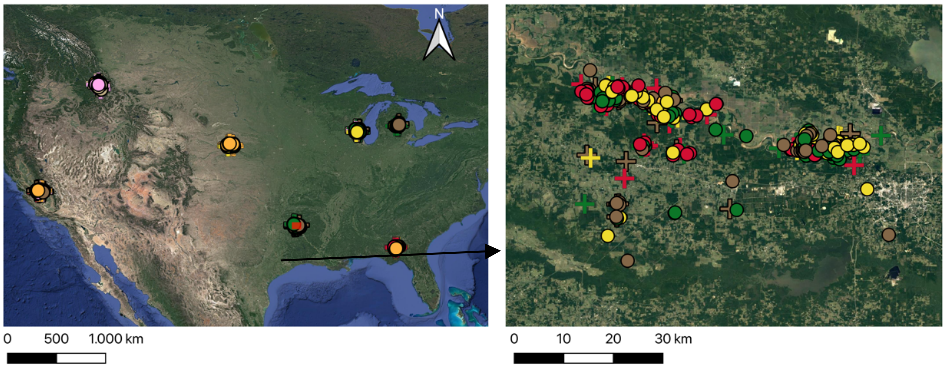


Figure 24 - Training and testing samples over the selected counties.

After obtaining the training and testing data in the CONUS and counties (mixed and one crop per county), the precision of the RF and SVM classifiers for different training-testing was analyzed. Results are presented in four different cases depending on the type of training and testing: A) training and testing in the CONUS, B) training and testing in the counties (mixed case), C) training in the counties (mixed case) and testing in the CONUS, and D) training in counties (only one crop per county) and testing over the CONUS.

Figure 25 represents the OA mean for ten repetitions for the input data: spectral features, spectral indices, and weather data, using RF and SVM classifiers. The OA for the first case (A) is close to the 90% being the best OA obtained in this study. For case B, the OA is up almost 80%. However, when the classifier is training using small areas and testing in a continental scale, cases C and D, the OAs decrease as expected. These are challenging cases because of crop phenological development variation across climate zones [25]. Figure 25 shows that the SVM is still getting an OA close to 70%. However, the case D suffers a huge decrease in the OA below to 50%. According to these results, only cases A and B are considered to obtain the training and test samples in the next case studies.

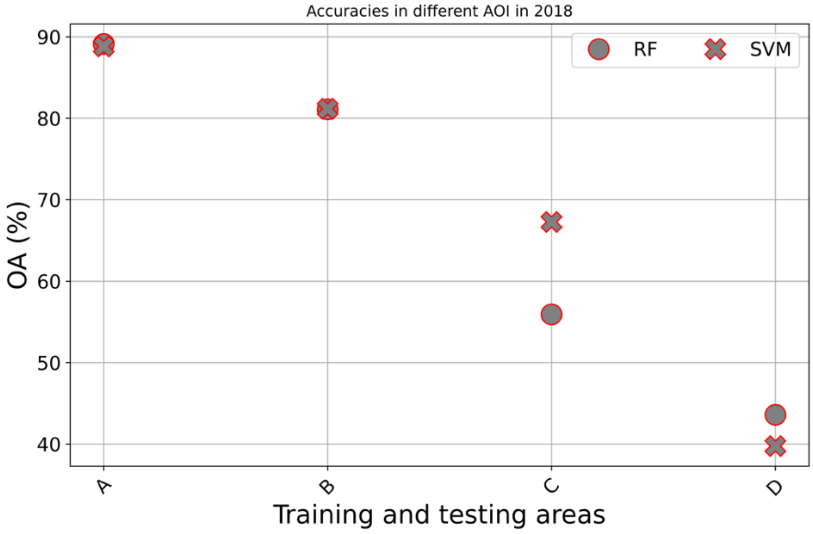


Figure 25 - OA for different train and test areas selection. [Notation: A: training CONUS - testing CONUS; B: training selected counties - testing selected counties; C: training selected counties - testing CONUS; D: training selected counties (one crop per county) - testing CONUS].

4.3. Temporal analysis

From the training and testing locations, three different datasets were generated depending on the type of spectral information and data were added:

- Dataset 1: spectral features (SF)
- Dataset 2: SF and spectral indices (SF + SI)
- Dataset 3: SF, SI and weather data (SF + SI + WD)

Figure 26 represents the OA using RF and SVM for each of these datasets. The results show a monthly OA mean of ten repetitions for each dataset during 2018. Regarding the type of dataset, the figure shows that the OA follows a similar tendency for the three datasets. Especifically, the SF+SI+WD clearly overcomes the two other datasets for all the months. That is, as input information increases so does the resulting accuracies. SI and WD provide new valuable information to the SF in the classification. Thus, SF+SI+WD provides the best results because the combination of the three dataset contains the most complete information in this study. That is, WD contributes to improve the accuracy. Therefore, satellite information can be complemented by the use of weather data to make more robust the classifiers and then obtain better accuracies.

Highlighting the turning point (August) in the curves, the dataset 1 and 2 obtained similar OA and dataset 3 overcame them although the difference is not very big. For example, using an SVM the difference between use or not weather data is around 3%. The SF+SI dataset obtained an OA of 85.2% meanwhile the SF+SI+WD to 88.8%.

As it has commented, August is the month when the curves become stable. Therefore, it is possible to get a classification at early September (when all the August data have been collected) with a similar OA than using the full information of the year.

Note that, the performance described for the three databases is equal when the classifiers were trained using the CONUS and the counties samples.

Regarding the classifiers, the performance of both of them are similar for the three databases. However, the use of the training samples from the CONUS makes the classifiers get higher OA for all the months.

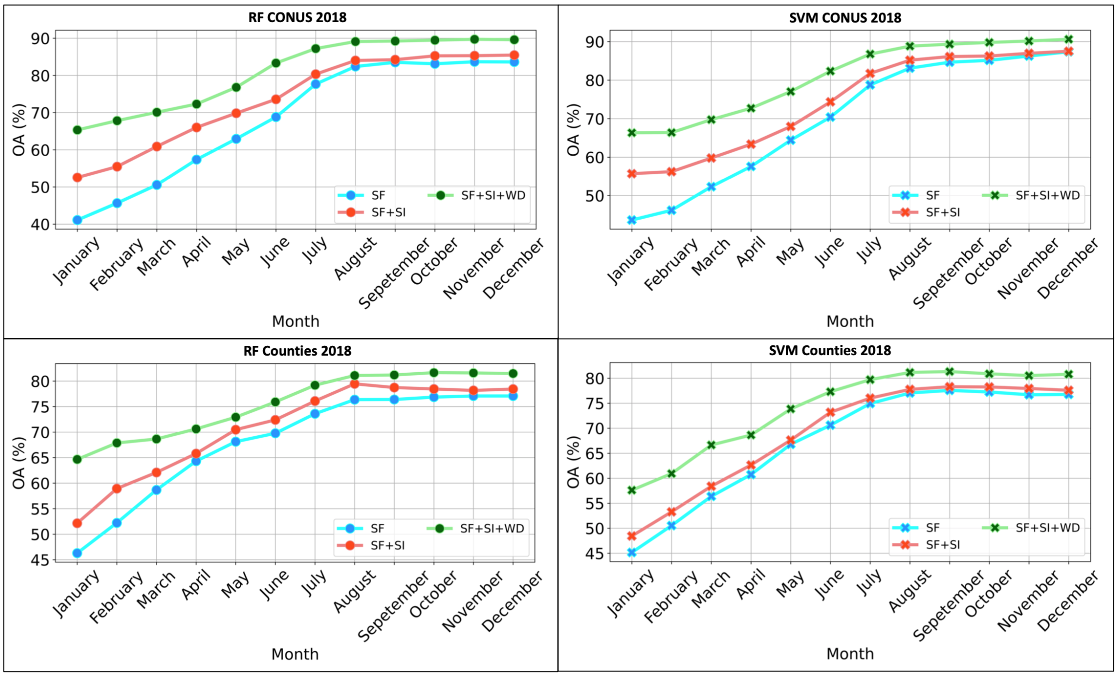


Figure 26 - OA average versus the months for the three datasets using RF and SVM trained by CONUS and counties samples. [Notation: SF: spectral features; SI: spectral indices (vegetation indices); WD: weather data].

Focusing on the classifiers training with the CONUS samples and the SF+SI+WD database, the boxplot of RF and SVM for the different runs were depicted. Figure 27 shows that RF contains less variance in its results than SVM. This is because assuming the randomness of the RF, the training set was not changed for the different runs whereas for the SVM it was necessary. Note that the SVM has bigger variability in its results from May to June and its results are slightly lower than RF. Even in August, SVM provides the best accuracy (90.57%) the average of RF performs better (89.11%) than the average SVM (88.82%). Therefore, the final crop classification was calculated using the RF classifier.

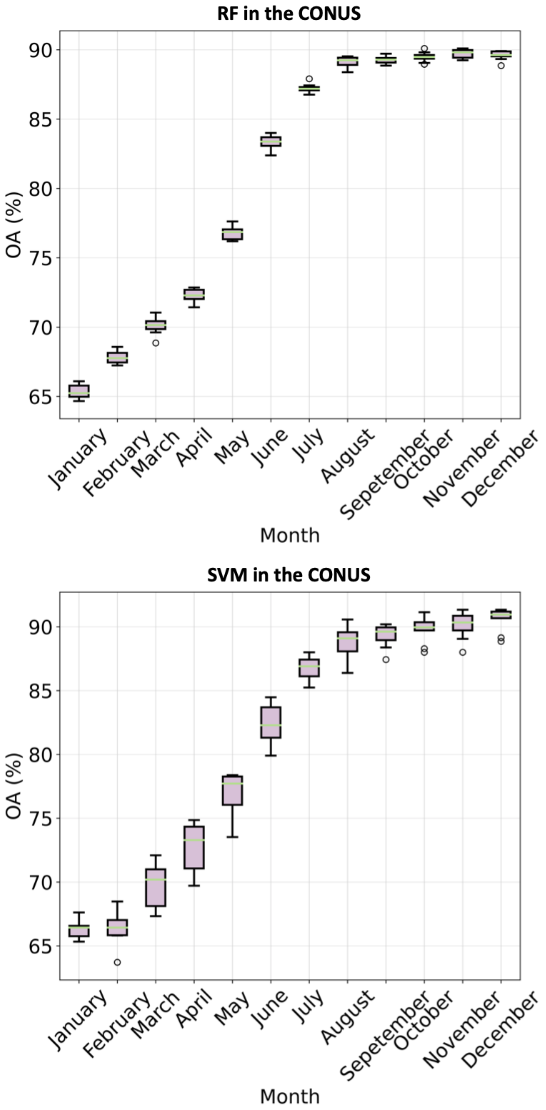


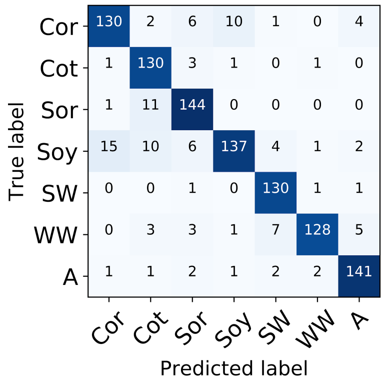
Figure 27 - Boxplot of OA using RF (top) and SVM (bottom) trained with the CONUS samples and the database 3 in 2018.

Table 11 summarizes the results for the turning point using different study areas, datasets and algorithms. As the table shows, best results are obtained with all the variables used in this study (SF + SI + WD), using training data from the CONUS and the RF classifier. Therefore, the final classifier was generated under these conditions. Table 11 shows the optimized parameters of this case. The results obtained with this classifier overcome 89% of OA with kappa coefficient equal to 0.88. The CM is represented in Figure 28 and shows good results in general for all the classes. However, there is a misclassification of soybeans, which is confused for corn and cotton classes. Additionally, the sorghum class is misclassified with cotton.

Table 11 - Optimized parameters of RF classifier for generating a crop classification map over the CONUS.

Month of classification	August
Input data	Spectral features + spectral indices + weather data
Algorithm	Random Forest
Number of trees	167
Number of features to split per node	119
Minimum samples leaf	1
Minimum samples splits	3

Figure 28 - CM of the RF classifier using SF+SI+WD from January to August and training with samples from the CONUS.



True label \ Predicted label	Cor	Cot	Sor	Soy	SW	WW	A
Cor	130	2	6	10	1	0	4
Cot	1	130	3	1	0	1	0
Sor	1	11	144	0	0	0	0
Soy	15	10	6	137	4	1	2
SW	0	0	1	0	130	1	1
WW	0	3	3	1	7	128	5
A	1	1	2	1	2	2	141

Notation: Cor: corn, cot: cotton, sor: sorghum, soy: soybeans, sw: spring wheat, ww: winter wheat, and a: alfalfa.

4.3.1. Crop type response

As commented in *Section 3.2.4. Classification*, the classifiers were generated month by month accumulating the information of the previous months. Therefore, the CMs of the month classifers are available for all the year and the temporal crop response to the classification is possible to analyze. In this work, the visualization of CM and the analysis of the monthly improvement per crop in the classification was done. Figure 29 depicts CMs for each month of 2018 using the best classifier obtained in the classification. Note that test subsets contained the same 150 samples per class for all the months. On the one hand, the best class in August is Sorghum with 144 correct test points followed by Alfalfa class. On the other hand, the crop with less samples correctly classified is winter wheat with almost 86 % of the samples well classified. Corn class improves more its classification than the rest of the classes from January to August. But, at the same time, it is the class with a high number of misclassification because it is difficult to differentiate from soybeans since both classes are cultivated in the same region of the CONUS (see Figure 23 in *Section 4.2. Training and testing data*). These misclassifications start to decrease from July being low and constant. Cotton only improves 13 samples from January to August being the class with less increment classification improvement. This is due to the cotton class having a high number of samples well classified from January. The figure also shows that there are misclassification samples in soybeans, which was assigned as sorghum during all the year, and it does not improve adding more monthly information. Focusing on the end of the year, alfalfa, sorghum and soybeans are almost perfectly classified.

Finally, it is noticeable that some crops can be classified before, for example, in May cotton, sorghum, wheats and alfalfa reach more than 70% of accuracy. To analyze the best month to classify each crop, their percentiles were calculated. From the 85% percentile the values stabilize, so this percentile can be used to look for the first month where this value is reached. Following this asumption, Table 12 shows that cotton, sorghum and alfalfa fulfill the idea of classifying in August, winter wheat could be better predicted the month before and corn one month after. However, the classification for spring wheat and soybeans improves in October. In this month, spring wheat shows a minimal improvement in the classification and it would change if another percentile was used. Apart from that, harvests could influence in soybeans classification.

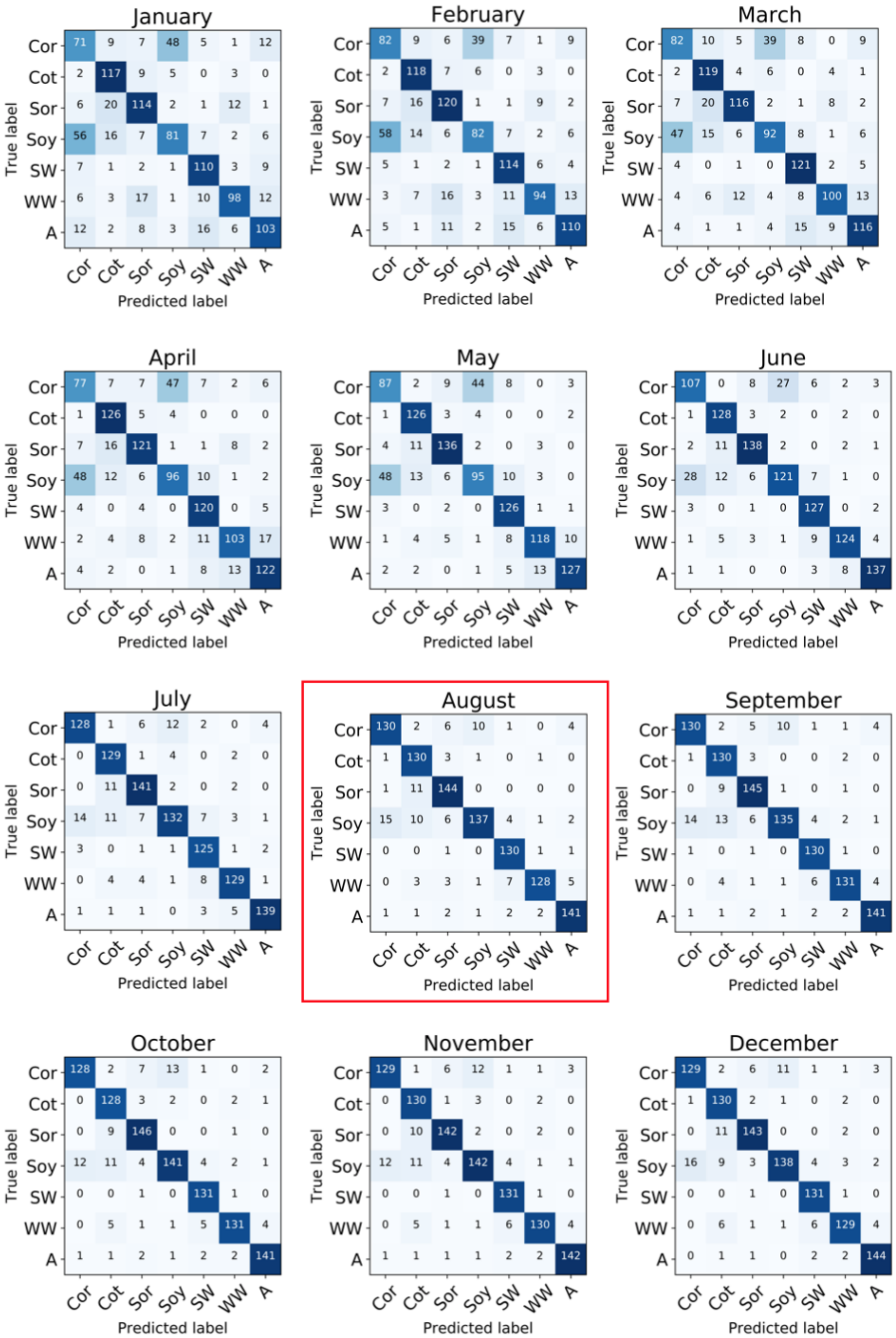


Figure 29 - CM per month of 2018 for the best classification model. The notation is equal to the Figure 28.

Table 12 - Months where 85% percentiles are reached.

class	***p85***	***month***
Cor	129	7
Cot	130	8
Sor	144	8
Soy	139	10
SW	131	10
WW	130	9
A	141	8

4.3.2. Feature selection

Spectral indices, spectral features and weather data provide valuable information for training a RF classifier. However, all this information is not required to obtain robust models. Selecting the most important features allows us, in general, to obtain similar results than using the original dataset. Thus, a feature selection of the three databases is performed using the 3500 samples from January to August 2018 over all the CONUS. Figure 30 shows the first features with more importance for each database. For the three databases, the correlation differences between the most important feature and the last selected are not very big. In the three cases the tendency decreases very smoothly. The feature selection method based on correlation (see *Section 3.2.3. Feature Selection*) chose only ten features over 42 spectral bands for the SF database. In this case, NIR and red bands in May and June are important bands being the NIR the two most important bands. The blue, red and SWIR bands of June are also included in the selected features, being the features from June the large number. When the feature selection method was applied to the SF+SI database, seven over 322 features were selected, highlighting that all of them were vegetation indices. That indicates the importance to calculate the vegetation indices to crop classification mapping. In this case, most of the features were selected from May. It is noticeable that NDVI is not included in this list, even other studies endorse its ability to differentiate between crops [52]. The last case (i.e. adding WD), ten features over 364 were selected. Only one vegetation index was included in the most important features list (TDVI for January) and the rest of the features were focused on precipitation (from January to August). The duration of the daylight in January is the most important feature, so it was a reason for analyzing it (Figure 31).

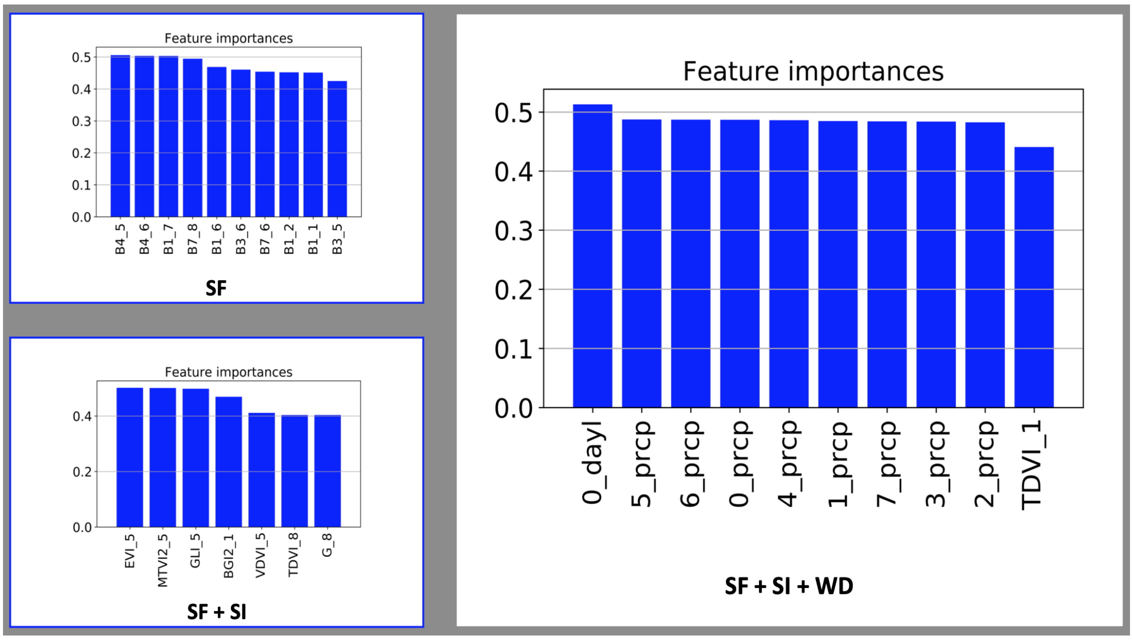


Figure 30 - Feature importances of spectral features (SF), spectral indices (SI), and weather data (WD) databases. The number represents the month (for weather data starting from 0: ex. January = 0)

Figure 31 represents the boxplots of the day length of January for each crop for training and test subsets. It is possible to see four groups of crops out of seven. Cotton presents higher values of daylight time followed by the sorghum. However, the variability of the Sorghum makes share values with winter wheat and soybeans. The crops: soybeans, corn, winter wheat and alfalfa have quite close similar durations of daylight. And the crop with less time of light during the day is the spring wheat. This division follows a location pattern in the CONUS. Generally spring wheat is in the north, where the daylight is lower, while cotton is in the south with longer periods of light.

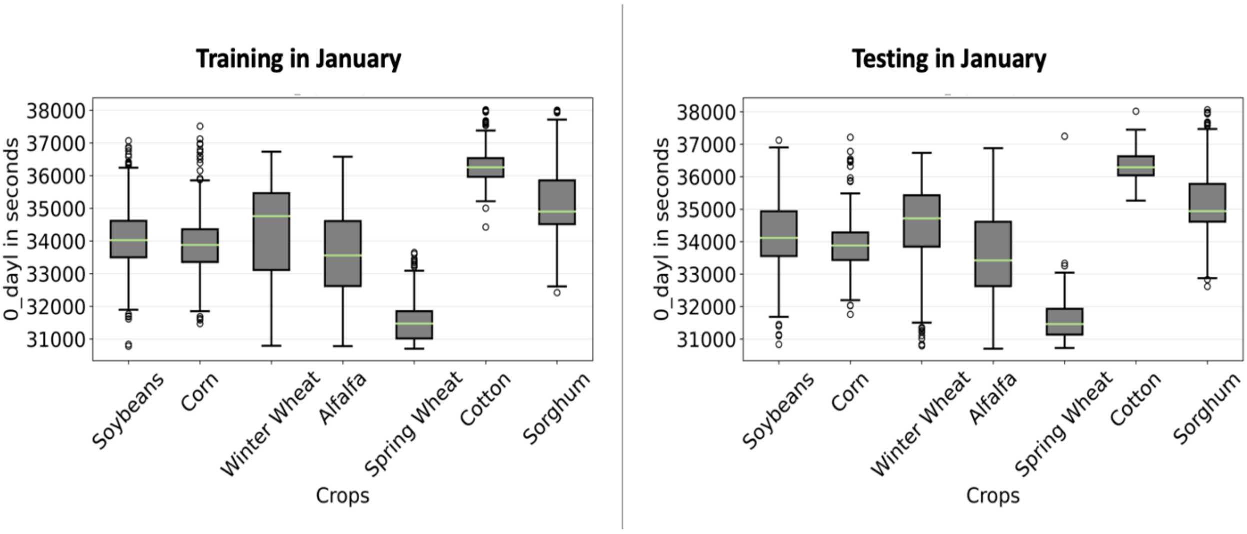


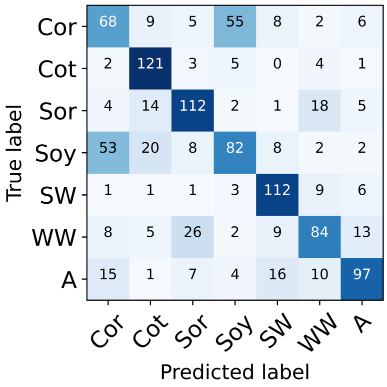
Figure 31 - Boxplots of the daylenght in January (in seconds) per crop of interest.

The features selected (with SF+SI+WD data) were used to classify the seven crops of interest with a RF. The optimal parameters of the RF obtained by 5-fold cross-validation are shown in the Table 13 together with the OA and kappa obtained. As the table presents the OA is not very high (lower than 64%). Additionally, the CM is plotted in the Figure 32. It is possible to see that the RF had problems with all the classes but especifically, to differentiate the corn and soybeans and winter wheat and sorghum. Therefore, this step was useful for analyzing which features are important and their importance but taking into account that RF does not suffer as much as other classifiers the dimensionality problem, the classification map was not generated with the features selected.

Table 13 - Optimized parameters of RF classifier for generating a crop classification map over the CONUS using only the selected features: daylenght in January, precipitations from January to August, and TDVI in January.

Trees	167
Features split	6
Min samples node	2
Min samples split	3
OA [%]	63.55
Kappa	0.58

Figure 32 - CM obtained using a RF classifier generated using the optimized parameters of table 13 and the features selected from January to August. The notation used here is equal to Figure 28.



4.3.3. Classification map

To finish the temporal analysis, the classification map of the best classifier is generated. In Figure 33.A, the CDL with all the land-cover classes over the CONUS are represented. The classes of interest mask are plotted in Figure 33.B. And the classification map obtained using the RF classifier and masked with the crops of interest is shown in Figure 33.C. This image shows that the classified crops are distributed accordingly with the CDL. However, a deeper insight into the classification in different zones over the CONUS provides a better visualization of the results. The result is shown in Figure 34.

In Figure 34 are represented classification maps in specific and small areas which contain the crops of interest. The area A) shows a clear misclassification of the winter wheat by spring wheat. In the case of D) corn is clearly misclassified with cotton. And in E) the clear misclassify class is soybeans by sorghum. Furthermore, soybeans and corn are not clearly distinguished (B). Therefore, the location of examples seems that alternates the performance of the classifier. For example, in area A the classifier misclassifies the winter wheat, whereas it is almost perfectly classified in area C. That is because area A is a non common area and area C is a common place for cultivating winter wheat (see Figure 23 in *Section 4.2. Training and testing data*). In any case, these are small numbers of misclassification compared to all the crop acreage in the CONUS and the high accuracy of the classifier.

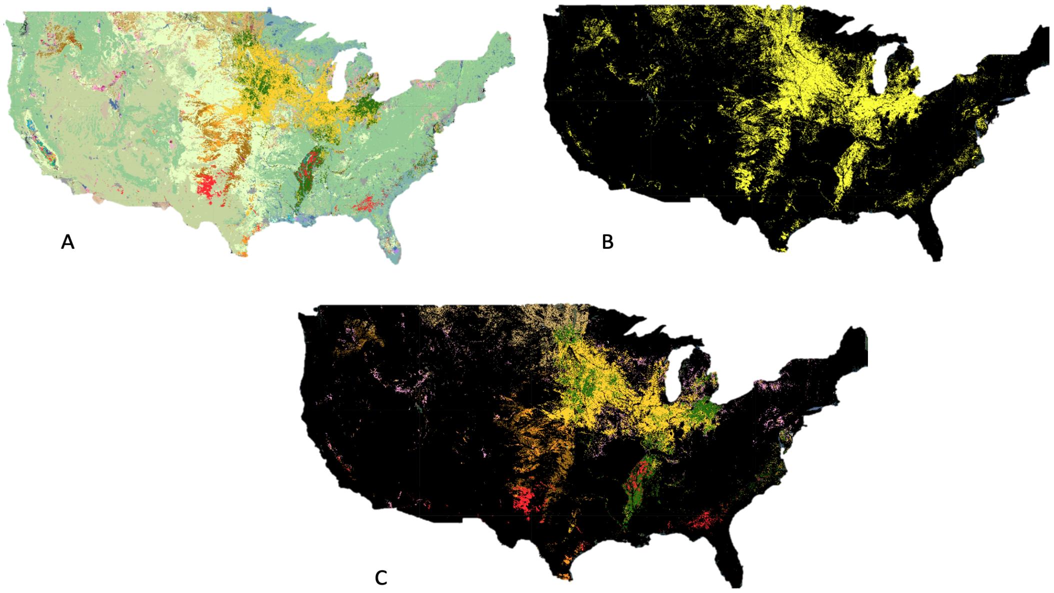
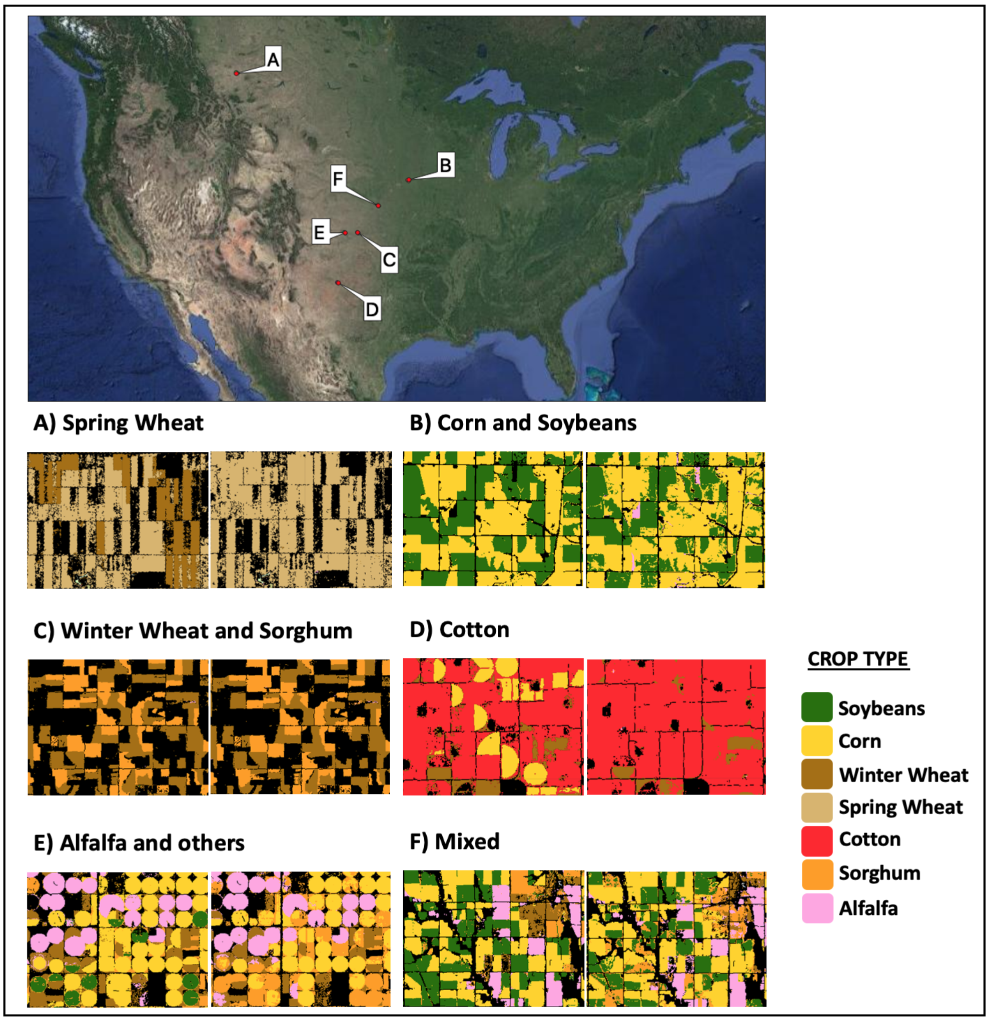


Figure 33 - A: CDL, B: mask of crops of interest, and C: crop classification map masked with the crop of interest in 2018.



A) Spring Wheat

B) Corn and Soybeans

C) Winter Wheat and Sorghum

D) Cotton

E) Alfalfa and others

F) Mixed

Figure 34 - Comparison between the CDL (left) and the crops classified (right) in small areas over the CONUS in 2018 with RF and the input datasets: spectral features, spectral indices, and weather data.

Taking into account the spatial pattern followed by the classifier to distinguish between crops, Figure 35 represents the duration of the daylight in January (i.e. the most important feature (*Section 4.3.2 Feature selection*)) along the CONUS. The relevance of this feature is being explained by the location of the crops because it changes along the latitude (north - south). Spring wheat grows mainly in the north, in zones where the length of the day is shorter compared with those in the south where cotton is planted. Then, other crops are in these areas the classifier misclassifies them. As it has commented before, winter wheat is misclassified with spring wheat in the north (Figure 34: A) and soybeans are classified as cotton in the south (Figure 34: D). Also, in the area B, there were misclassifications. However, looking at the misclassifications in the areas E and F, which have similar latitude, the soybeans are classified like sorghum. Therefore, the duration of the day light is not the unique variable classifying by location. Another variable that is dependent on the location and was in the list of the selected features is the precipitation.

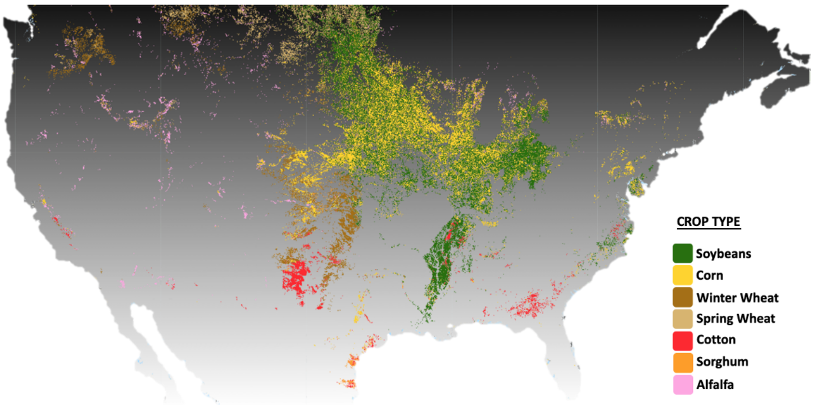


Figure 35 - Mean of day length in January 2018. Grey scales represent values from 29975 s (dark) to 38354 s (light).

Figure 36 shows that west areas of the CONUS have significantly less precipitations than in the east. This explains why soybeans are classified as sorghum even though they grow at the same latitude. When the samples are taken in central-west zones the classifier tends to overpredict sorghum instead of identifying such as soybeans.

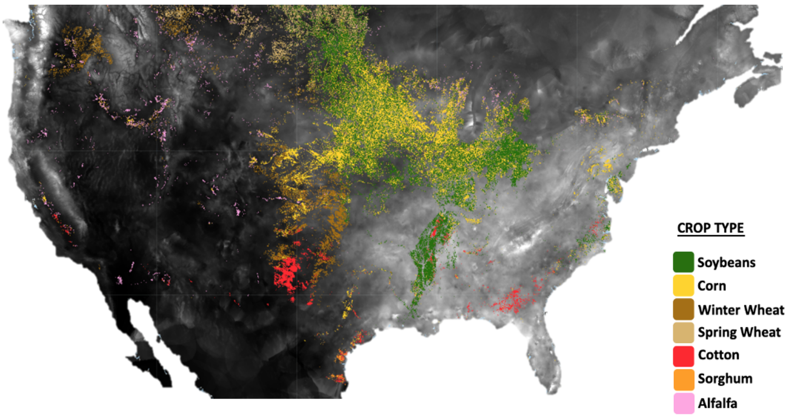


Figure 36 - Mean of the precipitations from January to August 2018. Grey scales represent values from 0.09 mm (dark) to 7.66 mm (light).

In order to analyze whether the classification follows the same spatial pattern without the weather dataset, an RF classifier was trained over the CONUS using only the SF+SI. Figure 37 presents the CM per month for the seven classes. It is observed that the misclassifications of soybeans as sorghum diminishes compared with the classification using weather data. As well, the cotton improves its accuracies while monthly data is added. That means, the weather data included location information and the classifier without these data probably was based on the phenological information. Therefore, it seems that without weather data the location pattern is not that much reinforced. However, the classifications are better using the three datasets in general.

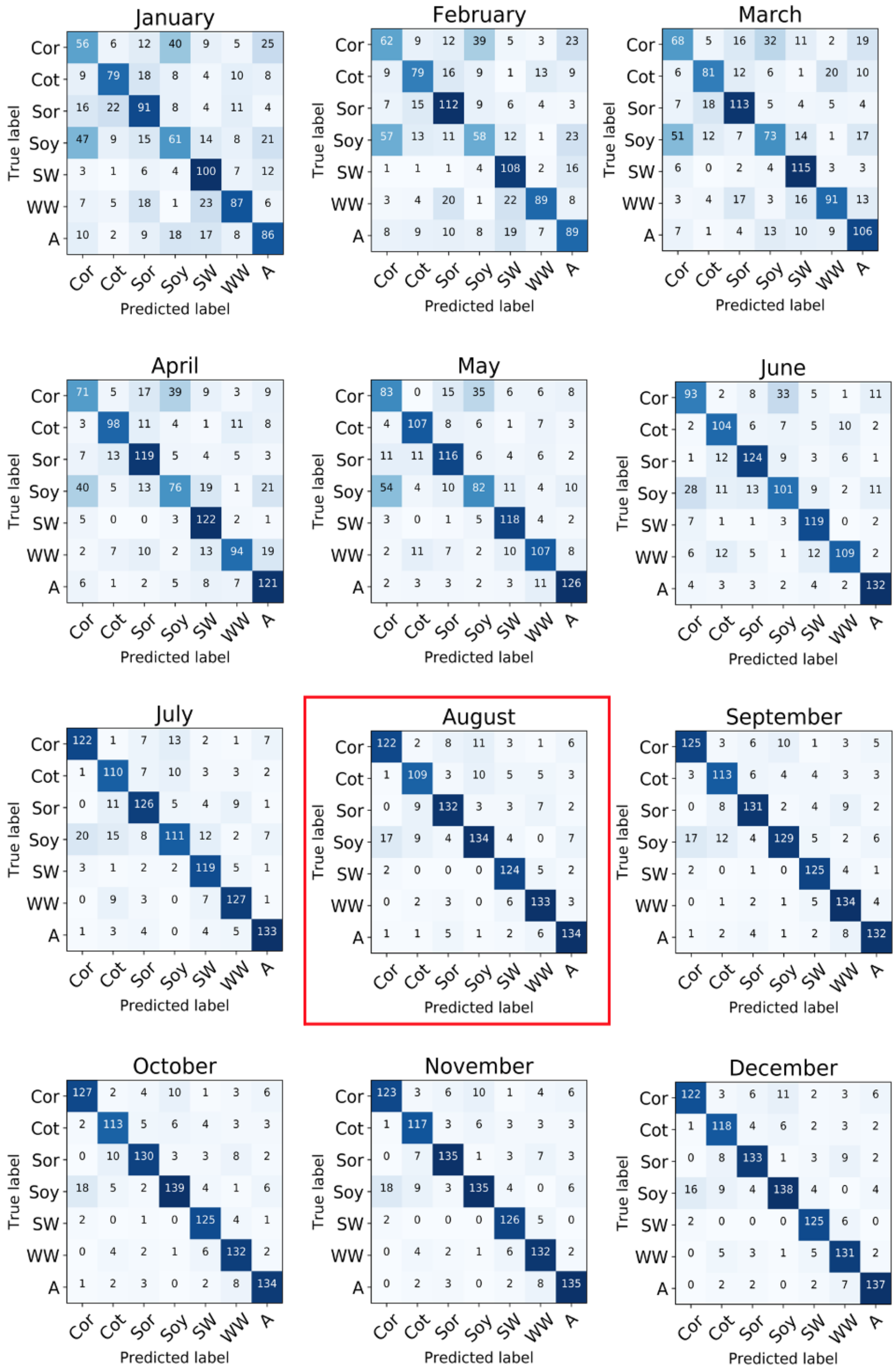


Figure 37 - CM obtained with a RF classifier using SF+SI data per month in 2018. Notation is equal to Figure 28.

4.4. Forecasting: training 2018 - testing 2019

This section presents the results obtained by a RF classifier trained with the data from the year before of the prediction. That is, training with data from 2018 and predicting and testing with data from 2019. The objective is to demonstrate if it is possible to classify when the training and testing data are not from the same year. This is done because in the year of the classification the CDL is not still provided. The forecasting is performed using the RF classifier and all the input data: SF+SI+WD. In this case, the overall accuracy achieves 82.2% and the kappa coefficient is 0.79. These results demonstrate that it is possible to classify crop types in one year using training data from the previous year. Figure 38 shows the map over all the CONUS for 2019 using training samples from 2018 and its comparison with the CDL map of 2019. The figure shows that the classification is based on location patterns even more noticeable in this case, especially for corn and soybeans. At the west are primarily classified corn while at the east soybeans at continental scale. Therefore, these forecasts are useful for estimating crop acreages because of their accuracy, but when looking into determined areas it is better to use more spatially distributed training data from the same year when the classification takes place.

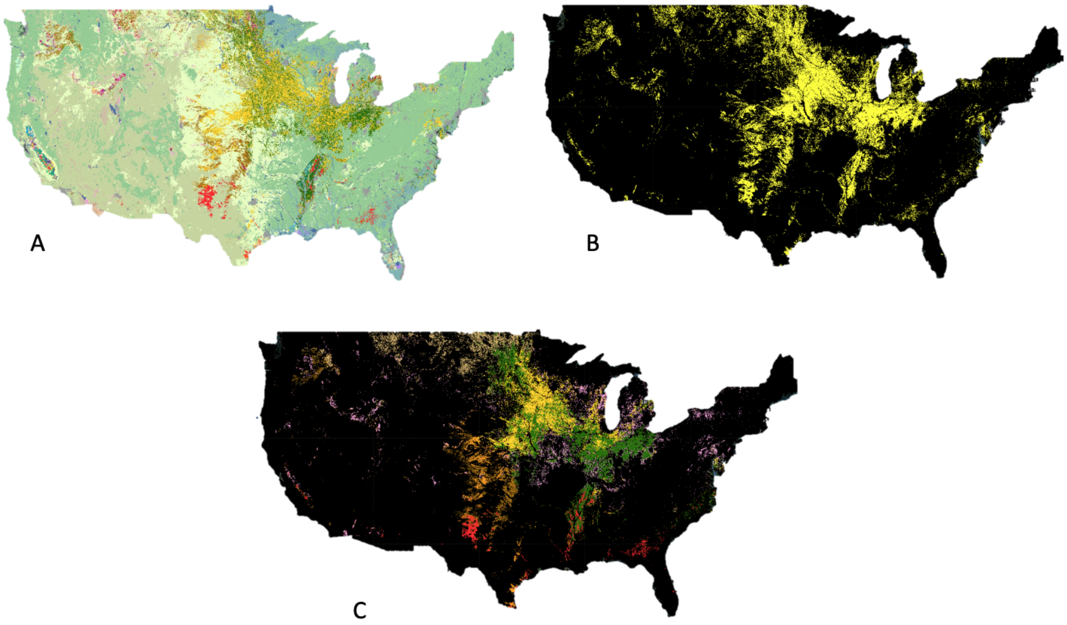


Figure 38 - Crop classification map over the CONUS in 2019. A: CDL (in 2019), B: crop of interest mask, and C: Crop classification map masked with the crops of interest in 2019.

5. Conclusions

In this study a method for obtaining as fast as possible crop classification maps over the CONUS has been tested in 2018. The method is based on a spatial reduction (from continental to county level) and a temporal analysis to detect the earliest month when accuracies are optimal. The study has been completed with a forecasting experiment to obtain a crop classification map in 2019 using available labelled data from 2018. It has been performed using a fusion of Landsat and MODIS satellite data (30 m of spatial resolution) and Daymet weather data. The process was developed with the cloud computing GEE application and python programming language.

Results have been assessed calculating the OA and Kappa coefficient for each different study area (CONUS, climatic regions or counties), input data (spectral features, spectral indices, and weather data), algorithm of classification (RF or SVM), and month of classification.

First was analyzed a spatial approach to find representative crop areas within the CONUS, it was shown that there are counties with similar crop percentages to the entire CONUS. Seven counties were selected representing each of them an area percentage of one different crop (soybeans, corn, winter wheat, spring wheat, cotton, sorghum, and alfalfa) over the CONUS in 2018. However, their temporal correlation over ten years (2009-2018) of their crop percentages compared to the CONUS was low. That means that percentages calculated in following years over the same counties are probable to not be comparable with those in all the CONUS. Therefore, the selected counties were only used to study how to perform classification maps over small areas distributed in the CONUS. A better selection of representativeness should be analyzed in further studies, see *Section 5.1. Further studies*.

Secondly, a temporal analysis was performed to obtain the promptest classification maps during the crop season. It was demonstrated that it is possible to obtain accurate crop classification maps of soybeans, corn, winter wheat, spring wheat, cotton, sorghum, and alfalfa over the CONUS with accuracies up to 90% at the end of August. From August accuracies are similar than the following month until the end of the natural year. The same study was performed over the seven selected counties in the previous section, and it was shown that classifications over the entire CONUS provided greater accuracies than when the area is reduced to the counties. However, the counties analysis facilitates the collection of samples and the accuracies obtained are still greater than 75%. For both cases, best accuracies were obtained using all the full input data, that is spectral features, spectral indices, and weather data. In particular, weather data gave the most valuable information to distinguish between crop types. Especifically, the duration of the daylight in January and the precipitations until August (month of classification) were the most important features in the classification over all the CONUS. When weather data was not included, the vegetation indices which best contributed to the classification were: SR2, GRVI, SGI, GEMI, GDVI, DVI, and MCARI. And even though the NDVI was inserted as a feature in the study it was not included in the list. Regarding the classification algorithms, both RF and SVM are robust classification algorithms that provided accuracies greater than 90% from August in 2018 in the CONUS. However, RF provided better accuracy than SVM. Finally, it was shown that the OA are stabilized in all cases from August, being its end the selected time of the prompt classification. Even that, it was shown by individual analysis of the crops, that depending on the crop it is possible to obtain good classification a few months early or a few months later. For example, in May cotton, sorghum, wheats, and alfalfa reached 70% of accuracy.

Hereafter, classification maps were obtained for the end of August 2018 with RF and using spectral features, spectral indices and weather data. It was shown that the classifier is influenced by location (latitude, longitude) patterns and affected to distinguish between crops. That means, over large areas specific crops are mainly planted in specified regions, and for that, it is a challenge to detect them when they are planted out from their most common area. On the one hand, in northern areas spring wheat is more probable to be planted; thus, the classifier tends to predict crops as spring wheat in this area, and the same occurs in the south with cotton. On the other hand, corn and soybeans tend to be misclassified with each other because of their similar locations and growth profile curves. As well, soybeans have been confused with sorghum since January and according to the confusion matrix this error remains permanent until December. This happens when soybeans are in areas where the sorghum is mainly produced. It was observed that weather data seems to favour the location-pattern classifications: duration of the daylight seems to differentiate between latitudes and precipitations between longitudes. But not only weather data is affecting this way to distinguish between crops, because the classification still depends on geographical position when weather data are not included. Other factors could be the phenology events that variate along the space depending on the environmental growing conditions [25]. Therefore, it is possible to conclude that crops always have differences inherents (environmental, phenological, etc.) depending on their locations when cropping over large areas. So, classifiers tend to distinguish crop types basing their decision in geospatial patterns, and in the cases where crops are out of their typical region, they can be misclassified. This means that the classification can be improved if more samples are taken for the training and they are more spatially distributed. It was demonstrated that selecting small areas within the CONUS is an alternative to avoid location troubles in classifications over large areas. For example, selecting representative counties which accuracies are up to 75%. In the following section (*Section 5.1. Further studies*) is presented a new alternative for selecting representative counties.

The forecasting experiment showed that using training samples from the year before of the classification (2018) allows to obtain accuracies around 80% in the end of August 2019. However, in this case, maps are even more based on the location of the crops, favouring their classification in the main region where they are cropped (location patterns). Therefore, this approach is valid to obtain statistics because its accuracies are high.

To sum up, optimal accuracies (up to 80%) can be achieved by the end of August, or even before depending on the crop type, using preprocessed HISTARFM satellite imagery and weather data over the US. Knowing the location and acreage planted of each crop during the growing season ultimately provides a better transparency in the food distribution. This can contribute to make decisions related with food security and agro-economy for governments, farmers, consumers and markets. Moreover, according to this thesis the information that crop classification maps provides is free for everyone with access to the internet.

5.1. Further studies

The results show great OA for classifying the main seven crops over the CONUS during the season (at the end of August). Even that, some proposals are presented below in order to complement or improve the results obtained in this thesis:

- *Crop type correlation between counties and the CONUS*
  Regarding the results in *Section 4.4.1. Crop percentages*, not all crop percentages in the selected counties are correlated temporarily with CONUS percentages. The future objective is to estimate crop percentages in the CONUS using temporarily correlated counties. Analyzing the crop area relation (R^2) between the CONUS and counties over 10 years provides a blend of counties with high temporal correlations that can predict

crop areas in all the CONUS. A selection of more than one county per crop could be made to calculate the average of their estimations. The resulting counties could be used to study total crop productions in the CONUS.

- *Binary classification in counties*

  The previous proposal can be complemented by conducting a more specific classification over each county. Classification over counties when each county represents a concrete crop can be performed separately and using a binary method: i.e. corn vs non-corn.

- *Train the classifier over several years*

  Due to the variation of the phenologies in crops along the time, one classifier that performs correctly in one year is not supposed to perform the same in the next year. Training the algorithm along several years helps to adapt to interannual variations in plant behaviour, and thus obtain a most robust method.

5.2. Contributions

This work has lead in several scientific publications and international conferences, as is described as follow:

1. Conference of “Geo for Good summit” (online conference, from 20th to 21st of October, 2020): a poster entitle “Looking for near-real time crop high resolution mapping classification using GEE” by **Rajadel-Lambistos, C.**; Izquierdo-Verdiguier, E.; and Moreno-Martínez, A. was presented.

2. American Geoscience Union (online conference from 1st to 17th December, 2020): The work entlite “Early crop mapping at continental scales derived from reconstructed high spatial resolution images” by **Rajadel-Lambistos, C.**; Izquierdo-Verdiguier, E.; Moreno-Martinez, A.; Atzberger, C.; Beguería, S.; Maneta, M.; Kimball, J.; Camps-Valls, G.; and Running, S. W. was accepted to an Oral presentation on the 8th of Decembre.

3. The paper entitled “Optimizing timing of crop classification by high spatial resolution images at continental scales" is in process status.

6. References

[1] Aguilar, R., Zurita-Milla, R., Izquierdo-Verdiguier, E., & A De By, R. (2018). A cloud- based multi-temporal ensemble classifier to map smallholder farming systems. *Remote sensing*, *10*(5), 729.

[2] Atzberger, C. (2013). Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote sensing*, *5*(2), 949-981.

[3] Badgley, G., Field, C. B., & Berry, J. A. (2017). Canopy near-infrared reflectance and terrestrial photosynthesis. *Science Advances*, *3*(3), e1602244.

[4] Bairagi, G. D., & Hassan, Z. U. (2002). Wheat crop production estimation using satellite data. *Journal of the Indian Society of Remote Sensing*, *30*(4), 213.

[5] Blessie, E. C., & Karthikeyan, E. (2012). Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, *6*(3), 385-394.

[6] Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, *26*(5), 341-358.

[7] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[8] Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z. (2018). A high- per- formance and in-season classification system of field-level crop types using time- series Landsat data and a machine learning approach. *Remote sensing of environment*, *210*, 35-47.

[9] Carroll, E., Chang, J., Lodi, L., Rapsomanikis, G., Zimmermann, A., & Blandford, D. (2018). The state of agricultural commodity markets 2018: agricultural trade, climate change and food security. Rome.

[10] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

[11] Dahal, D., Wylie, B., & Howard, D. (2018). Rapid Crop Cover Mapping for the Conterminous United States. *Scientific Reports, 8*(8631), 12. doi:10.1038/s41598-018-26284-w

[12] Dighe, S., Jagdale, A., & Chadchankar, A. (2018). Crop Production Prediction System. *International Journal of Research in Engineering, Science and Management,* *1*(12), 2581-5792.

[13] Enowski, H., S. Gerlt, A. Hungerford "How Do Adverse Planting Conditions Affect Crop Insurance Payment Timing?" *farmdoc daily* (10):159, Department of Agricultural and Consumer Economics, University of Illinois at Urbana- Champaign, September 2, 2020.

[14] Food, F. A. O. Agriculture Organization of the United Nations.(2017). The future of food and agriculture. *Trends and challenges. Rome*.

[15] Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on geoscience and remote sensing*, *42*(6), 1335-1343.

[16] Gao, F., Masek, J., Schwaller, M., & Hall, F. (2006). On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote sensing*, *44*(8), 2207-2218.

[17] Ghafarian Malamiri, H. R., Zare, H., Rousta, I., Olafsson, H., Izquierdo Verdiguier, E., Zhang, H., & Mushore, T. D. (2020). Comparison of Harmonic Analysis of Time Series (HANTS) and Multi-Singular Spectrum Analysis (M-SSA) in Reconstruction of Long-Gap Missing Data in NDVI Time Series. *Remote Sensing*, *12*(17), 2747.

[18] Gómez, D., & Montero, J. (2011). Determining the accuracy in image supervised classification problems. *In EUSFLAT-LFA 2011 European Society for Fuzzy Logic and Technology. Advances in Intelligent Systems Research*, 1 (1), 342- 349. Atlantis Press, Paris.

[19] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, *202*, 18-27.

[20] Haile, M. G., Kalkuhl, M., & von Braun, J. (2014). Inter-and intra-seasonal crop acreage response to international food prices and implications of volatility. *Agricultural Economics*, *45*(6), 693-710.

[21] Haindl, M., Somol, P., Ververidis, D., & Kotropoulos, C. (2006, November). Feature selection based on mutual correlation. In *Iberoamerican Congress on Pattern Recognition* (pp. 569-577). Springer, Berlin, Heidelberg.

[22] Ingram, K. T., Dow, K., Carter, L., Anderson, J., & Sommer, E. K. (Eds.). (2013). *Climate of the Southeast United States: variability, change, impacts, and vulnerability* (p. 341). Washington, DC: Island Press.

[23] Izquierdo-Verdiguier, E., & Zurita-Milla, R. (2020). An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *88*, 102051.

[24] Johnson, D. M. (2019). Using the Landsat archive to map crop cover history across the United States. *Remote Sensing of Environment*, *232*, 111286.

[25] Konduri, V. S., Kumar, J., Hargrove, W. W., Hoffman, F. M., & Ganguly, A. R. (2020). Mapping crops within the growing season across the United States. *Remote Sensing of Environment*, *251*. doi:https://doi.org/10.1016/j.rse.2020.112048.

[26] Kukal, M. S., & Irmak, S. (2018). Climate-driven crop yield and yield variability and climate change impacts on the US Great Plains agricultural production. *Scientific Reports*, *8*(1), 1-18.

[27] Kuo, C. G., Schreinemachers, P., Schafleitner, R., & Wopereis, M. (2020). Vegetables and Climate Change: Pathways to Resilience. *World Vegetable Centre*, 13.

[28] Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), 778-782.

[29] Kwas, M., Paccagnini, A., & Rubaszek, M. (2020). Common factors and the dynamics of cereal prices. A forecasting perspective.

[30] Lyapustin, A., Wang, Y., Korkin, S., & Huang, D. (2018). MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques*, *11*(10).

[31] Lofgren, B., & Gronewold, A. (2012) Water Resources Sector Midwest Technical Input Report National Climate Assessment. *Great Lakes Integrated Sciences and Assessments (GLISA) Center .* Retrieved November 1 from http://glisa.umich.edu/media/files/NCA/MTIT_WaterResources.pdf.

[32] Mase, A. S., Gramig, B. M., & Prokopy, L. S. (2017). Climate change beliefs, risk perceptions, and adaptation behavior among Midwestern US crop farmers. *Climate Risk Management*, *15*, 8-17.

[33] Maponya, M. G., Van Niekerk, A., & Mashimbye, Z. E. (2020). Pre-harvest classification of crop types using a Sentinel-2 time-series and machine learning. *Computers and Electronics in Agriculture*, *169*, 105164.

[34] Moreno-Martínez, Á., Izquierdo-Verdiguier, E., Maneta, M. P., Camps-Valls, G., Robinson, N., Muñoz-Marí, J., Sedano, F., Clinton, N., & Running, S. W. (2020). Multispectral high resolution sensor fusion for smoothing and gap-filling in the cloud. *Remote Sensing of Environment*, *247*, 111901.

[35] Mourtzinis, S., Edreira, J. I. R., Conley, S. P., & Grassini, P. (2017). From grid to field: Assessing quality of gridded weather data for agricultural applications. *European Journal of Agronomy*, *82*, 163-172.

[36] National Weather Service / NOAA. (2020). "Midwest Derecho - August 10, 2020, Updated: 10/8/20 12 pm". Retrieved September 8, 2020 from *www.weather.gov*.

[37] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*(1), 217-222.

[38] Pandey, A., & Mishra, A. (2017). Application of artificial neural networks in yield prediction of potato crop. *Russian Agricultural Sciences*, *43*(3), 266-272.

[39] Sarwar, M. H., Sarwar, M. F., Sarwar, M., Qadri, N. A., & Moghal, S. (2013). The importance of cereals (Poaceae: Gramineae) nutrition in human health: A review. *Journal of Cereals and Oilseeds, 4*(3), 32-35. doi:10.5897/JCO12.023

[40] Sedano, F., Kempeneers, P., & Hurtt, G. (2014). A Kalman filter-based method to generate continuous time series of medium-resolution NDVI images. *Remote Sensing*, *6*(12), 12381-12408.

[41] Skakun, S., Franch, B., Vermote, E., Roger, J. C., Becker-Reshef, I., Justice, C., & Kussul, N. (2017). Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment*, *195*, 244-258.

[42] Story, M., & Congalton, R. G. (1986). Accuracy assessment: a user's perspective. *Photogrammetric Engineering and remote sensing*, *52*(3), 397-399.

[43] Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. B. (2014). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2*. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).

[44] Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, *8*(2), 127-150.

[45] USDA United States Department of Agriculture. National Agricultural Statistics Service. (2018) *Department of Agriculture*. Retrieved November 1, 2020, from https://www.nass.usda.gov/Publications/Ag_Statistics/2008/Chap01.pdf

[46] USDA United States Department of Agriculture. (2020, September). *Economic Research Service, Farm Income and Wealth Statistics.* Retrieved November 1, 2020, from https://www.ers.usda.gov/data-products/farm-income-and-wealth-statistics/data-files-us-and-state-level-farm-income-and-wealth-statistics/.

[47] Vogel, F. A., & Bange, G. A. (2020). Understanding USDA crop forescasts. *National Agricultural Statistics Service and World Agricultural Outlook Board, Office of the Chief Economist, U.S. Department of Agriculture,* 1554.

[48] Wageningen University. (2020). Sustainable Food Security: Food Access Course (edx).

[49] Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018, June). Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 1-5).

[50] Wulder, M. A. et al. (2019). Current status of Landsat program, science, and applications. *Remote sensing of environment*, *225*, 127-147.

[51] Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, *97*(3), 322-336.

[52] Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, *2017*.

[53] Yang, N., Liu, D., Feng, Q., Xiong, Q., Zhang, L., Ren, T., Zhao, Y., Zhu, D., & Huang, J. (2019). Large-scale crop mapping based on machine learning and parallel computation with grids. *Remote Sensing*, *11*(12), 1500.

[54] Zafari, A., Zurita-Milla, R., & Izquierdo-Verdiguier, E. (2017, October). Integrating support vector machines and random forests to classify crops in time series of Worldview-2 images. In *Image and Signal Processing for Remote Sensing XXIII* (Vol. 10427, p. 104270W). International Society for Optics and Photonics.

[55] Zafari, A., Zurita-Milla, R., & Izquierdo-Verdiguier, E. (2019). Evaluating the performance of a random forest kernel for land cover classification. *Remote sensing*, *11*(5), 575.

[56] Zurita-Milla, R., Balasubramanian, K. I., Izquierdo-Verdiguier, E., & de By, R. A. (2017). The combination of multiple kernel learning and one-class classifier in classifying smallholder cotton fields. *Development*, *34*(4), 723-736.

[57] Zurita-Milla, R., Izquierdo-Verdiguier, E., & Rolf, A. (2017, June). Identifying crops in smallholder farms using time series of WorldView-2 images. In *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*(pp. 1-3). IEEE.