Pedro Marco Achanccaray Diaz

Crop Recognition in Tropical Regions based on spatio-temporal Conditional Random Fields from multi-temporal and multi-resolution sequences of remote sensing images

TESE DE DOUTORADO

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Programa de Pós-graduação em Engenharia Elétrica

Rio de Janeiro
January 2019

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Marco Achanccaray Diaz

Crop Recognition in Tropical Regions based on spatio-temporal Conditional Random Fields from multi-temporal and multi-resolution sequences of remote sensing images

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica.

Advisor : Prof. Raul Queiroz Feitosa
Co-advisor: Profa. Ieda Del'Arco Sanches

Rio de Janeiro
January 2019

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Pedro Marco Achanccaray Diaz

Crop Recognition in Tropical Regions based on spatio-temporal Conditional Random Fields from multi-temporal and multi-resolution sequences of remote sensing images

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica of PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica. Approved by the undersigned Examination Committee.

Prof. Raul Queiroz Feitosa
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Profa. Ieda Del'Arco Sanches
Co-advisor
Instituto Nacional de Pesquisas Espaciais – INPE

Prof. Franz Rottensteiner
Leibniz University Hannover – LUH

Profa. Leila Maria Garcia Fonseca
Instituto Nacional de Pesquisas Espaciais – INPE

Prof. Ruy Luiz Milidiú
Departamento de Informática – PUC-Rio

Prof. Rodrigo C. de Lamare
Centro de Estudos em Telecomunicações – PUC-Rio

Prof. Márcio da Silveira Carvalho
Vice Dean of Graduate Studies
Centro Técnico Científico – PUC-Rio

Rio de Janeiro, January 30th, 2019

Pedro Marco Achanccaray Diaz

The author received his bachelor's degree in Mechanical and Electrical Engineering at the Universidad Nacional de Ingenieria (UNI) in 2010. He obtained his master's degree in Electrical Engineering with emphasis on Signal Processing and Control at the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) in 2014. Since then, he has worked in the field of Digital Image Processing, Remote Sensing and Machine Learning.

Bibliographic data

Achanccaray D., P.

Crop Recognition in Tropical Regions based on spatio-temporal Conditional Random Fields from multi-temporal and multi-resolution sequences of remote sensing images / Pedro Marco Achanccaray Diaz; advisor: Raul Queiroz Feitosa; co-advisor: Ieda Del'Arco Sanches. – Rio de janeiro: PUC-Rio, Departamento de Engenharia Elétrica, 2019.

v., 87 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Reconhecimento de Culturas;. 3. Sensoriamento Remoto;. 4. Modelos Graficos Probabilisticos;. 5. Imagens óticas;. 6. Imagens de radar. I. Feitosa, R. Q.. II. Sanches, I. D.. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

To all those whose have a dream and ventured to make it true.

Acknowledgments

I am truly grateful to my advisor, Prof. Raul Queiroz Feitosa, for the encouragement, his advice, stimulating talks and generous support throughout the development of this thesis.

I would like to thank Prof. Christian Heipke and Prof. Franz Rottensteiner for their support during my stay in IPI, as well as all IPI members that help me a lot not only in the academic aspect, but also in the personal.

I thank my parents, Pedro and Mercedes, for their patience and unconditional love, my brothers, David, Saul and Ever, for their advice and supportive words and talks, and Karen, my love, for being so comprehensive and encourage me to continue and have faith.

I want to thank all my colleagues from the Computer Vision Lab for their companionship and valuable scientific discussions.

I also gratefully acknowledge the financial support of CNPq and its program Science without borders during my stay in Hanover, Germany.

Abstract

Achanccaray D., P.; Feitosa, R. Q. (Advisor); Sanches, I. D. (Co-Advisor). **Crop Recognition in Tropical Regions based on spatio-temporal Conditional Random Fields from multi-temporal and multi-resolution sequences of remote sensing images**. Rio de Janeiro, 2019. 87p. Tese de doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

The earth population growth has continuously increased the demand for agricultural production. Thus, acreage and crop yield information become increasingly important. Techniques based on satellite images are one of the most attractive options for agricultural monitoring over large areas. Most of the scientific works on this application were developed for temperate regions of the planet, which present a much simpler dynamics than those in tropical regions. In this context, the present thesis proposes a new automatic method based on Conditional Random Fields (CRF) for the crop recognition in tropical regions from multi-temporal and multi-resolution image sequences from orbital multi-sensors. Experiments were performed to validate several variants of the proposed method. We used public databases from two regions of Brazil that comprise sequences of optical and radar images with different spatial resolutions. The experiments demonstrated that the proposed method achieved a higher accuracy than methods based on a single image or sensor. Particularly, the reduction of the *salt-and-pepper* effect in the generated maps was noticed due, mainly, to the capacity of the method to capture contextual information.

Keywords

Crop Recognition; Remote Sensing; Probabilistic Graphical Models; Optical imagery; Radar imagery

Resumo

Achanccaray D., P.; Feitosa, R. Q.; Sanches, I. D.. **Reconhecimentos de culturas em regiões tropicais baseadas em campos aleatórios condicionais espaço-temporais a partir de sequências de imagens de sensoriamento remoto multitemporais e de múltiplas resoluções**. Rio de Janeiro, 2019. 87p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

O crescimento da população do planeta tem aumentado continuamente a demanda por produtos agrícolas. Assim, a informação quanto a áreas cultivadas e estimativas de produção se tornam cada vez mais importantes. Técnicas baseadas em imagens satelitais constituem uma das opções mais atrativas para o monitoramento agrícola sobre grandes áreas. A maior parte dos trabalhos científicos voltados a esta aplicação foram desenvolvidos para regiões temperadas do planeta, que apresentam um dinâmica muito mais simples da que se tem em regiões tropicais. Neste contexto, a presente tese propõe um novo método automático baseado em Campos Aleatórios Condicionais (CRF) para o reconhecimento de culturas agrícolas em regiões tropicais a partir de sequências de imagens multi-temporais e multi-resolução produzidas por diferentes sensores orbitais. Experimentos foram realizados para validar diversas variantes do método proposto. Utilizaram-se bases de dados públicas de duas regiões do Brasil que compreendem sequências de imagens óticas e de radar com diferentes resoluções espaciais. Os experimentos realizados demonstraram que o método proposto atingiu acurácias maiores do que métodos baseados em uma única imagem ou sensor. Particularmente, notou-se a redução do efeito sal-e-pimenta nos mapas gerados devido, mormente, à capacidade do método de capturar informação contextual.

Palavras-chave

Reconhecimento de Culturas; Sensoriamento Remoto; Modelos Graficos Probabilisticos; Imagens óticas; Imagens de radar

Table of contents

List of figures

List of tables

All men have stars, but they are not the same things for different people. For some, who are travelers, the stars are guides. For others they are no more than little lights in the sky. For others, who are scholars, they are problems... But all these stars are silent. You—you alone will have stars as no one else has them.

Antoine de Saint-Exupéry, *The Little Prince.*

1 INTRODUCTION

1.1 Motivation

Agricultural activities have to be monitored from local to global scales at high temporal frequency due to their dependency on physical landscapes and climatic conditions as well as seasonal patterns associated with crops' biological life cycle, growing conditions, pests and diseases, and excessive market speculation, leading to price spikes. Some of the challenges agriculture has to deal with are: to limit or reduce agriculture's environmental impacts, to confront increasing global food demand and to look for pathways to boost agricultural production.

The negative environmental impacts of agriculture are mainly related to threatening of biodiversity by land glades and habitat dissolution [4], Greenhouse gas (GHG) emissions from agricultural production, which represents 24% of global GHG emissions [5], and depletion of freshwater resources as 69% of freshwater [6] used by humans is for irrigation, and it is estimated to increase by about 19% in 2050 [7]. Furthermore, the increasing food demand is expected to last for three to four decades [1], leading to a per capita demand for crops to be doubled between 2005 and 2050 (see Figure 1), with strongest increases within economic groups C-E.

Looking for alternatives to increase agricultural production is another challenge to overcome, which could be achieved by intensification (adoption of better agronomic practices) or extensification (expansion of agricultural lands). According to [8], the preferred solution is the intensification to close the yield gap between realized productivity and the best that can be accomplished under the current conditions (see Figure 2 for an example of yield gaps from cereals at a global scale). Thus, to map crop type and acreage is crucial to achieve this goal by a better understanding of regional cropping practices and the influence of potential environmental threats, and in this way, supply national and multi-national agricultural agencies with an inventory of what was grown in certain areas and when. These maps serve the purpose of forecasting grain supplies (yield prediction), collecting crop production statistics, creation of crop rotation records, mapping soil productivity, identification of crop stress' factors, crop damage assessment due to storms and drought, and monitoring farming activity.

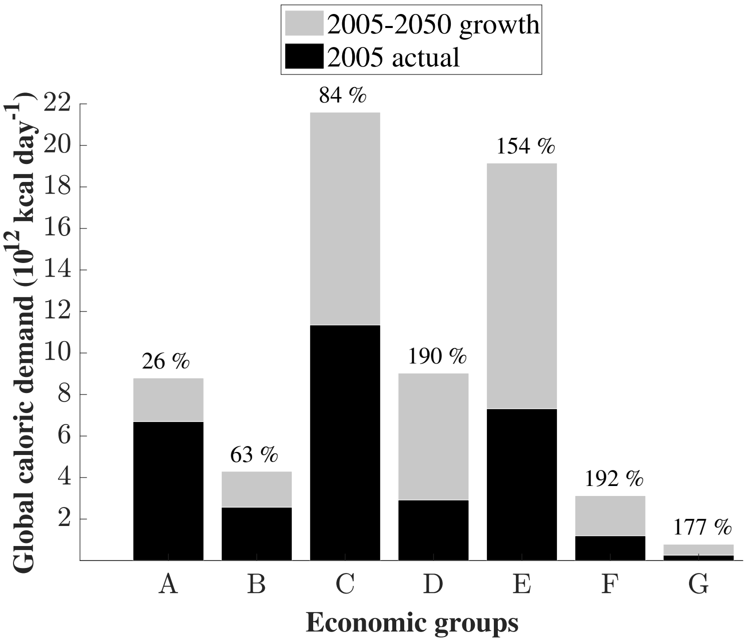


Figure 1: Global demand for crop calories per economic groups based on per capita Gross domestic product (GDP) ranking. Brazil is the economic group C (Modified from [1]).

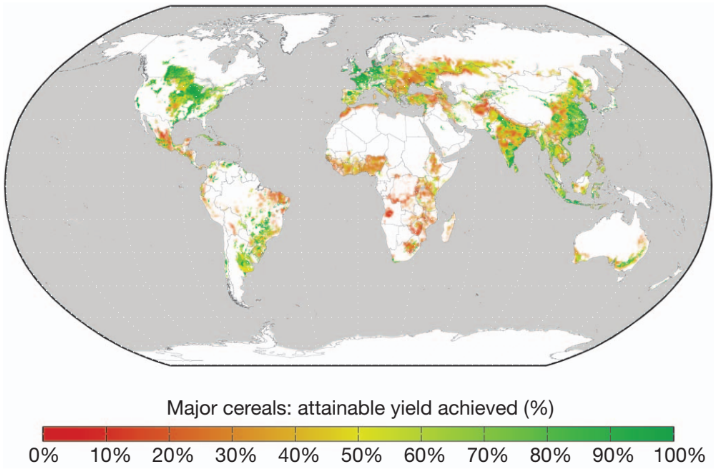


Figure 2: Average yield gaps for major cereal crops: *Corn*, *Wheat*, and *Rice*. The yield gap is the difference between the potential yield and the realized yield at a given location (Taken from [2]).

Agricultural monitoring systems should be able to provide timely information on crop production, status, and yield in a standardized and regular manner at the regional to the national level. Estimates should be provided as early as possible during the growing seasons and updated periodically through the season until harvest. Based on the information provided, stakeholders are enabled to take early decisions and identify geographically the areas with large variation in production and productivity. The system should provide homogeneous and interchangeable datasets with statistically valid precision and accuracy. Probably, only (satellite) remote sensing - combined with sophisticated modeling tools - can provide such information in a timely manner, over large areas, in sufficient spatial detail and with reasonable costs [9].

In remote sensing, crop maps are produced by supervised classification using Earth Observation (EO) images acquired at key phenological stages for optimizing class separability. The need for high amounts of cloud-free imagery impedes the employment of these approaches over large areas and in multiple years, making necessary the usage of complementary data from different sources such as LiDAR (Light Detection and Ranging), SAR (Synthetic Aperture Radar), among other active sensors, which are less affected by weather conditions.

In this context, the purpose of this work is to propose a method for crop mapping in tropical regions, more specifically in Brazil, which is the world's largest producer of sugarcane, coffee and orange juice and the second largest producer of soybeans [10]. This task is more challenging in tropical areas due to the high crop's dynamics generated by climatic, socio-economic, infrastructure and agricultural practices adopted (e.g. no-tillage, crop rotation, irrigation systems) [11].

In order to achieve this goal, Conditional Random Fields (CRF) are employed to build a model that is able to consider contextual information in different domains namely, spatial and temporal. Spatial context is important for crop mapping as pixels close to each other are likely to represent the same class. Also, temporal context is critical because the appearance, spatial distribution and orientation change over time due to crop development, which is strongly related to a crop's phenological stages.

1.2 Objectives

1.2.1 General Objective

The general objective of this work is to propose a model for crop recognition in tropical region based on Conditional Random Fields using sequences of remote sensing images from different sensors.

1.2.2 Specific Objectives

The specific objectives of this work are the following:

1. Introduce contextual information in spatial and temporal domains into a model based on Conditional Random Fields.

2. Exploit information from multiple sensors with the same spatial resolution.

3. Build a model able to deal with images with different spatial resolutions.

4. Generate datasets used for crop recognition in tropical regions and prepare them for public use in the scientific community.

5. Evaluate the influence of using different sensors and how they complement each other.

1.3 Contributions

The main contributions of this work are the following:

1. A novel method to recognize crops in tropical regions from sequences of remote sensing images based on Conditional Random Fields.

2. A Conditional Random Fields based model for crop recognition able to deal with image sequences generated by sensors of different spatial resolutions.

3. An evaluation of Conditional Random Fields and Convolutional Neural Networks to model spatial and temporal context for crop recognition tropical regions.

4. Two public datasets for crop recognition in tropical regions.

Regarding the last contribution, it was a joint effort between the National Institute for Space Research (INPE), Brazilian Agricultural Research Corporation (EMBRAPA), and the Pontifical Catholic University of Rio de Janeiro (PUC-Rio). Field works to record in-situ data from each region were performed by experts from INPE and EMBRAPA, while the acquisition and pre-processing of radar images have been carried out during the development of this thesis. A detailed description and more information about each dataset can be found in [11] and in [12].

1.4 Organization of the remaining parts of this thesis

Chapter 2 describes the related work available in the literature for crop recognition using different approaches such as Object-based Image Analysis, Probabilistic Graphical models and Deep Learning.

Chapter 3 provides the fundamental concepts and theory for a better understanding of the proposed method.

Chapter 4 introduces and explains the proposed method for crop recognition based on Conditional Random Fields (CRF).

Chapter 5 presents the datasets employed in this work, the experimental protocol followed in the experiments and the results obtained by the different variants evaluated in this thesis.

Chapter 6 summarizes the conclusions derived from the performed experiments and provides directions for further development of the proposed method.

2 RELATED WORK

This chapter presents an overview of different works on crop classification in different study areas, tropical or temperate regions, using a single image or a multi-temporal image sequence from the same or multiple domains.

Crop classification is a challenging task due to spatial and temporal changes crops experience within and between seasons. Agricultural prediction systems at big scales (entire cities or countries) have been proposed by the European Space Agency (ESA): Sen2-Agri[1] [13] and SENSAGRI[2] [14]. Sen2-Agri, a free and open source system, generates agricultural products (cloud-free composites, cropland masks, crop type maps, vegetation status maps) from Sentinel-2 and Landsat 8 times series along the growing season. SENSAGRI is still under development and aims to combine Sentinel-1 with Sentinel-2 and in-situ data to develop prototypes to estimate surface soil moisture, seasonal crop mapping, and crops statistics. These systems rely on a set of classification algorithms that are executed over the requested study area and their predictions are combined using a decision fusion algorithm (e.g. Dempster–Shafer [15, 16]).

Many approaches have been proposed so far for crop classification using either pixel-wise classification techniques (e.g. Random Forest, Support Vector Machines) [17–22], object-based methodologies [23–25], modelling context information by graphical models (e.g. Markov Random Fields, Conditional Random Fields) [26–35] or automatically learning representations by Neural Networks (e.g. Convolutional or Recurrent Neural Networks) [36–42].

Crop mapping using pixel-wise techniques is performed by classification algorithms trained upon pixel values, statistics or indices to map crops. Forkuor et al. [17] carried out a hierarchical classification using Random Forest (RF) trained with back-scatter intensities and spectral bands/indices from SAR and Optical imagery, respectively. Moreover, Tatsumi et al. [19] evaluated the employment of Tasseled-Cap bands' statistical variables using RF for crop recognition in Peru from a set of Landsat 7 images. Pixels can be assigned a crop type by comparison between pixel's candidate profile and

[1] Sentinel-2 for Agriculture (http://www.esa-sen2agri.org/)

[2] Sentinels Synergy for Agriculture (http://sensagri.eu/)

ideal crop curves based on spectral reflectance [18] or NDVI profiles [20]. Ancillary information, such as expert knowledge about crop's phenological stages, have been successfully employed to identify crops from Optical images [22] or phenological sequence patterns from a dense stack of SAR data [21]. These techniques usually present a *salt-and-pepper* effect in the produced maps, which can be overcame using post-processing methods such as majority voting inside each plot, among other strategies. Furthermore, they are agnostic to the type of entity that is classified.

Object-based Image Analysis (OBIA) has been adopted for many classification tasks. It comprises three main steps: generation of segments (objects), extraction of features from each segment, training and applying a classifier using the extracted features. In [23], Belgiu et al. assessed the usage of segments against pixels for crop mapping from Sentinel-2 time series using Time-Weighted Dynamic Time Warping (TWDTW) [43]. Clerici et al. [24] fused Sentinel-1A and Sentinel-2A data for land cover mapping in Colombia and performed a comparison between three classifiers: RF, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), being SVM the best one. Schultz et al. [25] developed an autonomous workflow for supervised object-based classification from multi-temporal Landsat 8 images in Brazil. In spite of the idea of spatial context obtained by the usage of segments, they diminish the spectral variability within them.

The employment of Probabilistic Graphical Models (PGMs) for crop recognition has increased during the last years due to their ability to capture context information in spatial and/or temporal domains. Methods based on Hidden Markov Models (HMMs) [44] have been proposed to learn spectral response variations along crops' cycles [26] or to represent vegetation dynamics [27] from multi-temporal image sequences. Notably, a HMM does not consider spatial context explicitly. In contrast, Markov Random Fields (MRF) [45] and Conditional Random Fields (CRF) [46] are able to capture spatial and temporal context. In [28], Liu et al. proposed a spatio-temporal MRF framework for multi-temporal classification and compared a global, a local and a pixel-wise model for temporal interactions to detect changes in forests. Likewise, Hagensieker et al. [29] introduced a spatio-temporal MRF for land use/land cover mapping, where the association potential is given by an Import Vector Machines (IVM) classifier and the spatial and temporal interaction potentials are represented by a Potts model and transition matrices from expert knowledge, respectively. In spite of the inclusion of spatial context, MRF based models are limited as the spatial interaction is a function of only labels, disregarding any dependence on the observed data.

Conditional Random Fields (CRF) are able to capture dependency on the observed data in both, the association and spatial interaction potentials. Hoberg et al. [30] proposed a multi-temporal CRF-based approach for crop type classification modeling the spatial and temporal interaction potentials by pixel-wise feature differences between two epochs. Later, Hoberg et al. [31] extended their work to consider images with different spatial resolutions for land cover classification and change detection. The spatial and temporal interaction potentials were modeled by label smoothing methods and a global transition matrix, respectively. Kenduiywo et al. [35] model a crop's phenology in conjunction with expert-based phenology knowledge in temperate regions by a higher order dynamic CRF. In our previous work, the estimation of the temporal interaction in a CRF-based approach for crop recognition is formulated as an optimization problem [32]. Then, spatio-temporal CRF models were developed for crop mapping in tropical areas where the association potential is given by a RF classifier, the spatial interaction potential is represented by a contrast-sensitive Potts model and the temporal interaction potential is estimated by a RF trained to learn transitions between adjacent epochs [33] or modeled by a transition matrix based on expert knowledge about possible and not possible transitions [34]. These approaches managed to achieve high accuracies, producing smooth classification maps due to the spatial and/or temporal context information embedded into the model. However, they still rely on handcrafted features, so that a study to select and extract suitable features for certain applications is still required.

The main advantage of Deep Learning (DL) models is their ability to learn representations automatically from data, allowing for better results than what can be achieved by using domain-specific handcrafted features. In addition, Convolutional Neural Networks (CNN) [47] are able to capture spatial context information by the application of different kernels during the convolution.

In [37], a comparison between supervised and unsupervised DL techniques for crop recognition is performed, where CNN-based approaches achieved the highest accuracies among the tested approaches. Kussul et al. [36] proposed a CNN-based architecture for land cover and crop type classification in Ukraine from multi-temporal and multi-source satellite imagery, where images from different sources are stacked to constitute a single image with multiple bands. Cue et al. [38] applied a Fully Convolutional Neural Network (FCN) for crop mapping in Brazil from a set of Sentinel-1A images, obtaining higher accuracy and less processing time than a CNN-based approach despite their computational cost. As CNN only considers spatial context, a common technique to include temporal context is to stack all images in the sequence to form a single image. However, stacking multi-temporal images restricts pattern recognition to a single feature space that may suffer from overlapping classes due to increased class variance [35].

In contrast, Recurrent Neural Networks (RNN) [48] are designed to capture the temporal context by introducing feedback to a neural network. In [40], an RNN-based classifier is employed for agricultural land cover mapping from a sequence of Sentinel-1A images in France. A comparison between two DL architectures (an RNN variant, long short-term memory (LSTM), and one-dimensional convolutional (Conv1D) based on an inception module) has been performed by Zhong et al. [42], the latter presenting the best results for crop classification using Landsat Enhanced Vegetation Index (EVI) time series. Hybrid methods have been proposed by [39] and [41], exploiting the main advantages of both, CNN and RNN. In [39], Bermudez et al. combined RNN and CNN for crop mapping in Brazil, achieving slightly better results than RF-based classification. An encoder structure with convolutional recurrent layers has been adapted by [41] to approximate phenological models for crop classification form a sequence of Sentinel-2 images.

3
FUNDAMENTALS

This chapter aims to provide the basics for a proper understanding of the proposed approach for crop mapping. First, a brief introduction about Remote Sensing is given comprising active and passive sensors, as well as levels in which data fusion can be performed. Then, Conditional Random Fields (CRF) are explained, as well as the proposed variants for multi-temporal and multi-resolution crop classification. Finally, some concepts related to Deep Learning (DL) are described centering on Convolutional Neural Networks (CNN).

3.1
Remote Sensing (RS)

According to Lillesand et al. [49], Remote Sensing (RS) is *"the science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation"*. This data might be acquired by different kinds of sensors. For instance, taking a photo using a smart-phone camera might be considered as remote sensing because it involves no direct contact of the sensor with the object of interest.

Electromagnetic RS makes use of electromagnetic energy sensors operated from airborne (aircraft, helicopters, and drones) and space platforms (satellites and space shuttle) to assist in inventorying, mapping and monitoring earth resources [49]. A summary of the advantages of each kind of platform is presented in Table 1. For agricultural applications like crop monitoring/mapping and yield estimation, it is necessary to cover vast extensions with a high temporal frequency to acquire images during the whole crop cycle (i.e. soil preparation, seeding, growing, harvesting and post harvest). Thus, space-borne sensors are a practical and cost-effective option for these applications. According to the illumination source sensors might be classified as passive or active.

Table 1: Characteristics of main Remote Sensing platforms: Space-borne, Aircraft and Drones. Related costs to each platform are based on [3]

Platform	**Characteristics**
Space-borne	• Large area coverage. • Frequent and repetitive coverage of an area of interest. • Relatively low cost per unit area of coverage. • Low, medium and high spatial resolutions.
Aircraft	• Small to large coverage, depending on the altitude. • High cost per unit area of coverage for small projects. • Relatively low cost for bigger projects. • Medium and high spatial resolutions.
Drones	• Small area coverage. • Low cost per unit area of coverage only for small projects. • High and very high spatial resolutions.

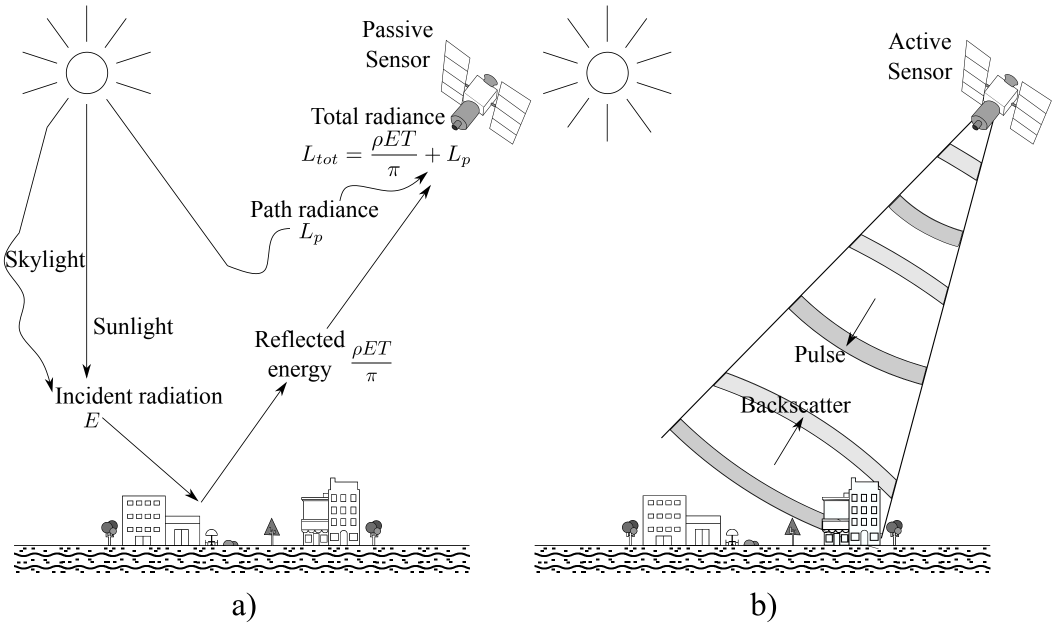


Figure 3: Passive and Active sensors. a) A passive sensor and the atmospheric effects that influence the radiance measured by the sensor L_{tot}, which is composed by the path radiance L_p and the reflected energy (a fraction of the incident radiation E by the sunlight and skylight). b) An active sensor emitting a pulse and receiving the back-scatter response after the pulse has interacted with a surface.

3.1.1 Passive sensors

A passive RS system is one that relies on energy that originates from sources other than the sensor itself, typically in the form of either reflected radiation from the sun or emitted radiation from earth surface features. Due to the atmospheric effects (see Figure 3a), the total radiance measured by the sensor L_{tot} is composed of the reflected energy (a fraction of the incident radiation E by the sunlight and skylight) and the path radiance L_p (reflected by the atmosphere). The reflected energy is computed as $\rho ET/\pi$, where ρ represents the reflectance of an object and T is the atmospheric transmittance.

The many forms of electromagnetic energy (e.g. visible light, radio waves, ultraviolet rays, among others) are commonly categorized by their wavelength within the electromagnetic spectrum (see Figure 4). Optical remote sensing operates within the *optical spectrum*, which extends from approximately 0.3 to 14 μm, including ultraviolet (UV), visible, near-, mid-, and thermal infrared (IR) wavelengths. Then, as different materials reflect and absorb differently at different wavelengths, objects can be differentiated by their spectral reflectance signatures in the remotely sensed images.

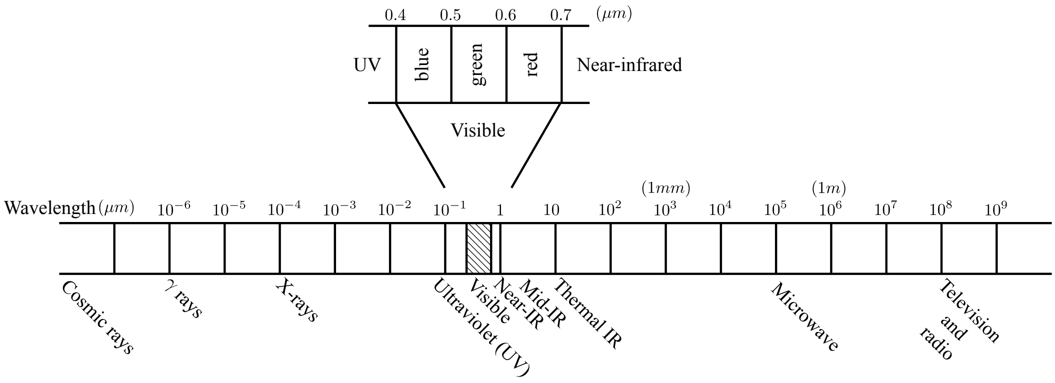


Figure 4: Electromagnetic spectrum (Modified from [1]).

3.1.2 Active sensors

An active remote sensing system is characterized by supplying its own illumination energy. Active sensors, such as Radio Detection And Ranging (Radar) and Light Detection and Ranging (LiDAR) systems, first emit energy and then measure the return of that energy after it has interacted with the object surface. Those who operates in the *microwave* portion of the electromagnetic spectrum (see Figure 4) have gained importance due to increasing amount of valuable environmental and resource information derived from them [49].

Microwave radiation has the ability of penetrating the atmosphere, depending on the wavelengths involved, under different conditions to "see through" haze, light rain and snow, clouds, and smoke. Also, it provides a different measure of earth materials that has not direct relationship with their counterparts in the visible or thermal portions of the spectrum.

Imaging Radar systems mainly use pulses where the energy from the antenna is confined to a very short interval of time (see Figure 3b). These pulses interact with the objects in a scene and some of them may be back-scattered to return toward the antenna, which records the intensity and phase shift of the returning pulses. Notice that as the electromagnetic energy has two components orthogonal to each other (electrical and magnetic), the sensor might be configured to transmit or receive different polarizations, which refer to the orientation of the electric field. In this way, the antenna can transmit and/or receive in either horizontal (H) or vertical (V) single polarizations (HH or VV, where the first letter indicates transmit and the second receive) or cross-polarization (HV or VH).

Synthetic Aperture Radar (SAR) systems operate on the principle of using the sensor motion along a track to simulate with a single physically short antenna an array of such antennas that can be linked together mathematically. It can be done by considering each position as an element of a single and long synthetic antenna (see Figure 5) as part of the data recording and processing procedures. SAR systems use a side looking geometry illuminating a surface at an oblique angle θ_l and an incidence angle θ_i, recording in this way in two directions, parallel to the sensor motion (azimuth) and orthogonal to its motion (range). Ground spatial resolutions, which is defined as the minimum possible distance between two objects to be distinguished, depend on pulse duration τ and antenna size L (see Figure 5). Range (ρ_r) and azimuth (ρ_a) resolutions are computed as follows:

$$\rho_r = \frac{c\tau}{2\sin\theta_i} \tag{3-1}$$

$$\rho_a = \frac{L}{2} \tag{3-2}$$

where c is the velocity of light (in vacuum) ($\approx 3 \times 10^8 m/s$). Thus, these resolutions define the minimum area, also known as cell, where back-scatter information might be recorded by the sensor. The *normalized back-scatter coefficient* σ^0 (sigma nought), defined as the back-scatter measured from a target area normalized per unit geometric cell area, depends on the received (P_R) and transmitted (P_T) power, the target's area to be recorded (A), the antenna gain (G) and its wavelength (λ), and range from antenna to target (R). This dependency is stated by the so-called *SAR equation*, which is given

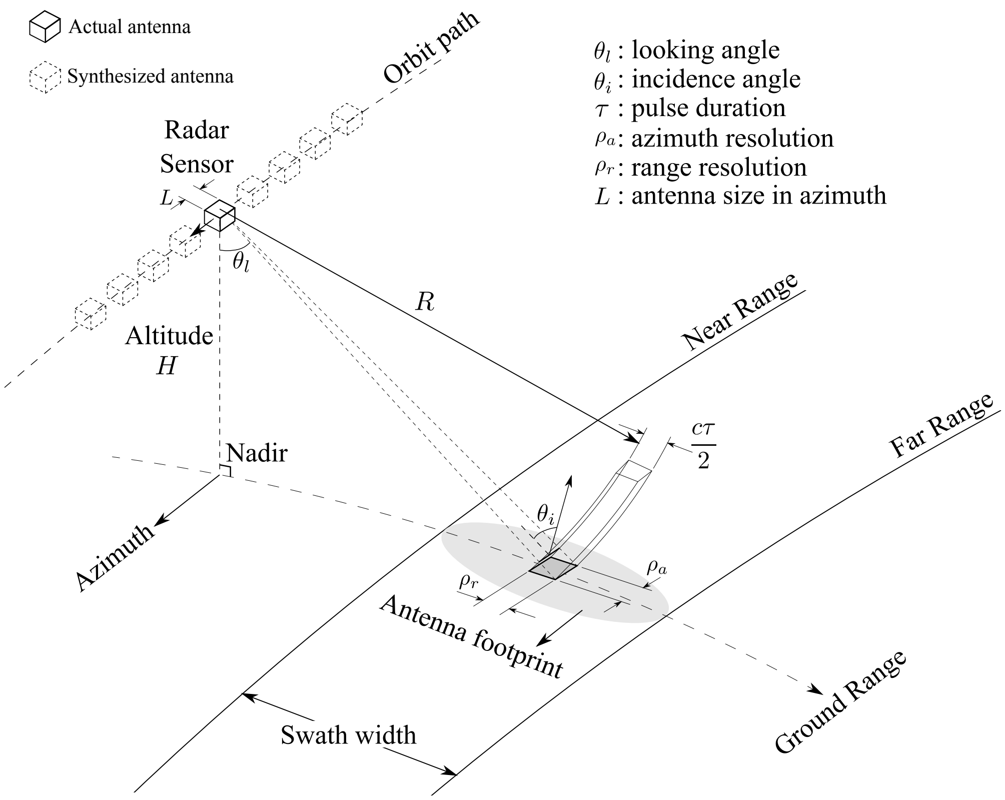


Figure 5: Components of a Synthetic Aperture Radar (SAR) system.

by:

$$\sigma^0 = P_R \frac{(4\pi)^3 R^4}{P_T A G^2 \lambda^2} \tag{3-3}$$

In crop mapping applications, the frequency or wavelength might affect crop back-scatter magnitude due to differences in dielectric properties (e.g. moisture) and relationship between wavelength and crop/leave size and/or canopy penetration. Polarization also alters crop discrimination as cross-polarization is able to retrieve geometry attributes (e.g. roughness and canopy structure) and it is very sensitive to height, shape and direction of land vegetation [50]. Furthermore, in multi-temporal analyses, back-scatter response is dominated by bare soil in early stages and by plant canopy in later stages. This correlation between crop growth stages and back-scatter magnitude suggests that knowledge about crop phenology is important for classification.

Figure 6 shows two remote sensing image taken on the same day, June 19th, 2018, by both a passive and an active sensor, corresponding to the municipality of Luis Eduardo de Magalhães, Bahia, Brazil. The image from a passive sensor was acquired by Sentinel-2A MSI (MultiSpectral Instrument) satellite with a spatial resolution of 10 m, where Figure 6a presents an RGB true color composition. Optical data are useful for crop mapping because the spectral response provides information about the changes in the moisture and chlorophyll content of crop leaves. The image from an active sensor was acquired by Sentinel-1A C-Band SAR, also with an spatial resolution of 10 m and VH polarization (see Figure 6b). From the back-scatter information under different polarizations it is possible to extract information to describe the structure, orientation distribution and dielectric constant characteristics of crops.

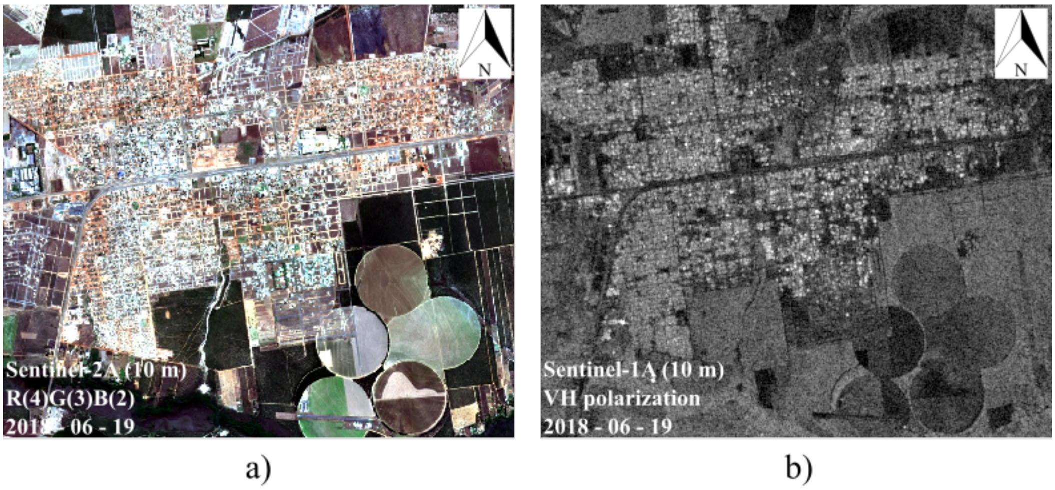


Figure 6: Images of Luis Eduardo de Magalhães municipality in Bahia state, Brazil acquired on the same date June 19th, 2018 from two different satellites: a) Passive sensor: Sentinel-2A MSI (MultiSpectral Instrument), 10 m spatial resolution, true color composition R(4)G(3)B(2), b) Active sensor: Sentinel-1A C-Band SAR (Synthetic Aperture Radar), 10 m spatial resolution, Vertical-Horizontal (VH) polarization.

3.2 Conditional Random Fields (CRF)

Conditional Random Fields (CRF) are discriminative undirected graphical models with the capability to consider contextual information. CRF were firstly introduced by Lafferty et al. [51] for one-dimensional text classification. Then, Kumar & Hebert [46] extended CRF for two-dimensional image classification using discriminative models for class associations at individual sites as well as interactions for neighboring sites.

Let $G = \{S, E\}$ be a graph with nodes S and edges E. Given a set of image sites $i \in S$, let $\mathbf{x} = \{\mathbf{x}_i\}_{i \in S}$ be the observed data and $\mathbf{y} = \{y_i\}_{i \in S}$ its corresponding labels, where $y_i \in \{l_1, ..., l_m\}$ and m the number of available classes. A CRF models the posterior probability $P(\mathbf{y}|\mathbf{x})$ of the set of labels $\mathbf{y}$ given the data $\mathbf{x}$ as follows:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_{i\in S} A(y_i,\mathbf{x}) + \theta\sum_{i\in S}\sum_{j\in N_i} I(y_i,y_j,\mathbf{x})\right) \tag{3-4}$$

where Z is a normalizing constant also called partition function, A and I are the unary term or association potential and the pair-wise term or interaction potential, respectively, and θ expresses the interaction potential weight relative to the association potential.

The association potential A relates to how likely an image site i takes a label y_i given the data $\mathbf{x}$. The interaction potential I expresses how labels at spatially neighboring sites i and $j \in N_i$ interact given the observed data $\mathbf{x}$, where N_i is the neighborhood of site i.

This CRF model, defined by Equation 3-4, is extended to model interactions between adjacent sites belonging to different epochs from a multi-temporal image sequence.

3.3 Multi-temporal CRF

Let's consider a set of T co-registered images from different epochs, with $t = 1, ..., T$ (see Figure 7) and a set of image sites $i \in S$, where i corresponds to the same geographical region in all epochs. Let $\mathbf{x} = \{\mathbf{x}_{i,t}\}_{i\in S,t\in T}$ be the data corresponding to the site i in epoch t and $\mathbf{y} = \{y_{i,t}\}_{i\in S,t\in T}$ its corresponding labels. Then, a multi-temporal CRF models the posterior probability $P(\mathbf{y}|\mathbf{x})$ of the set of labels $\mathbf{y}$ given the data $\mathbf{x}$ as given by:

$$\begin{aligned} P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\Bigg[\exp\Bigg(&\sum_{t\in T}\sum_{i\in S} AP^t(y_{i,t},\mathbf{x}) + \theta\sum_{t\in T}\sum_{i\in S}\sum_{j\in N_i} SIP^t(y_{i,t},y_{j,t},\mathbf{x}) \\ &+ \sum_{t\in T}\phi^t\sum_{i\in S}\sum_{k\in C_i} TIP^{tk}(y_{i,t},y_{i,k},\mathbf{x})\Bigg)\Bigg] \end{aligned} \tag{3-5}$$

where AP, SIP and TIP represent the Association, Spatial Interaction and Temporal Interaction Potentials, respectively. θ and $\phi^t \in \mathbf{\Phi} = \left\{\phi^1, \phi^2, ..., \phi^{T-1}\right\}$ are the spatial and temporal interaction potentials' weights.

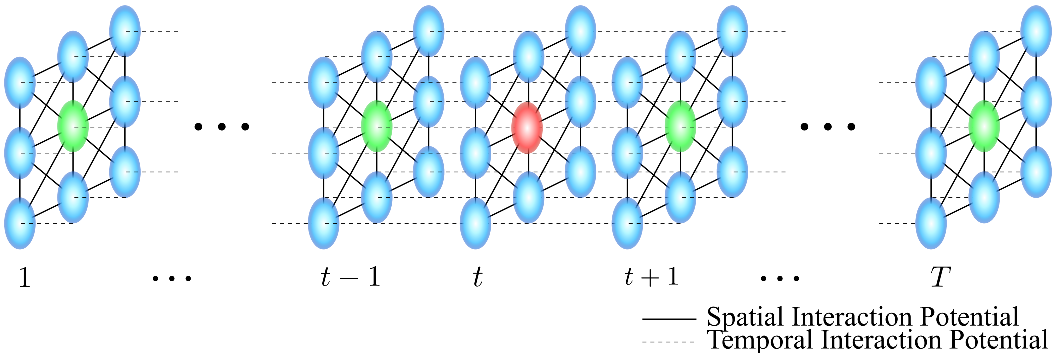


Figure 7: Multi-temporal sequence of T co-registered images represented as a graph where each node corresponds to an image site (e.g. a pixel or a segment) and each edge illustrates an interaction between neighboring pixels in spatial and temporal domains. Red and green nodes represent the same geographical region in different epochs. Solid and dashed lines symbolize the spatial and temporal interaction potentials, respectively.

3.3.1 Association Potential (AP)

The association potential $AP^t(y_{i,t}, \mathbf{x})$ measures how likely an image site i in epoch t will take a label $y_{i,t}$ given its feature vector $\mathbf{f}_{i,t}(\mathbf{x})$ that may depend on the entire image at epoch t or even on the whole multi-temporal image sequence. In this thesis the association potential is defined as:

$$AP^t(y_{i,t}, \mathbf{x}) = \log P(y_{i,t}|\mathbf{f}_{i,t}(\mathbf{x})) \tag{3-6}$$

where $P(y_{i,t}|\mathbf{f}_{i,t}(\mathbf{x}))$ is a local class conditional probability at image site i given $\mathbf{f}_{i,t}(\mathbf{x})$ and can be given by some discriminative classifier with a probabilistic output.

3.3.2 Spatial Interaction Potential (SIP)

The spatial interaction potential $SIP^t(y_{i,t}, y_{j,t}, \mathbf{x})$ measures how labels $y_{i,t}$ and $y_{j,t}$ at spatially neighboring sites i and j interact in epoch t, given the data $\mathbf{x}$ (see Figure 8). Contrast-sensitive smoothing methods, which penalize label changes unless a significant data variation occurs in neighboring sites, have been successfully applied for this purpose [52]. Then, according to these methods, the SIP is given by:

$$SIP^t(y_{i,t}, y_{j,t}, \mathbf{x}) = \begin{cases} p_{no-change}(i, j, t, \mathbf{x}) \text{ , if } y_{i,t} = y_{j,t} \\ p_{change}(i, j, t, \mathbf{x}) \quad \text{, if } y_{i,t} \neq y_{j,t} \end{cases} \tag{3-7}$$

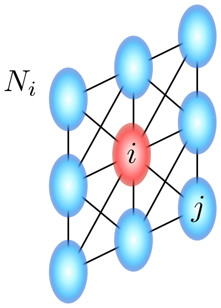


Figure 8: Spatial Interaction Potential (SIP), which measures how labels at spatially neighboring sites i and j ($j \in N_i$, N_i is the neighborhood of site i) interact given the data observed $\mathbf{x}$. Solid lines represent the SIP and red and blue nodes are the pixel of interest and its neighbors, respectively.

where p_{change} and $p_{no-change}$ represents the probabilities of label change and no-change between sites i and j, respectively. As it is desired to penalize label changes, $p_{no-change}$ usually must be higher than p_{change}.

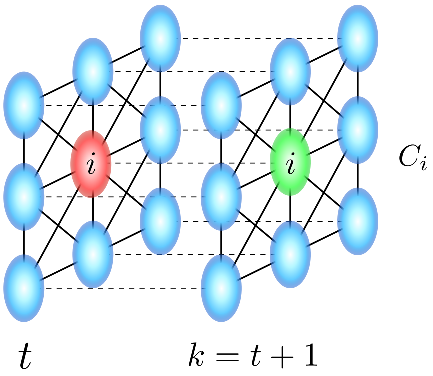


Figure 9: Temporal Interaction Potential (TIP), which measures how labels at sites representing the same geographical region in neighboring epochs t (red node) and k (green node) interact given the observed data $\mathbf{x}$. Solid and dashed lines symbolize the spatial and temporal interaction potentials, respectively.

3.3.3 Temporal Interaction Potential (TIP)

The temporal interaction potential $TIP^{tk}(y_{i,t}, y_{i,k}, \mathbf{x})$ measures how labels $y_{i,t}$ and $y_{i,k}$ at site i interact in epochs t and k , where epoch k is in the temporal neighborhood C_t of epoch t. (see Figure 9). Similar to the spatial interaction potential, TIP^{tk} can be represented by a transition matrix, whose element in row i and column k expresses how likely there is a transition between $y_{i,t}$ and $y_{i,k}$, given the observed data $\mathbf{x}$. If m_t and m_k are the number of classes in epochs t and k, respectively, $TIP^{tk}(y_{i,t}, y_{i,k}, \mathbf{x})$ will be a $m_t \times m_k$

matrix, whose $m_t * m_k$ elements represent transition probabilities to be estimated from training data. Alternatively, one can rely on expert knowledge, neglecting the dependency on the data (see next chapter).

3.3.4 Inference

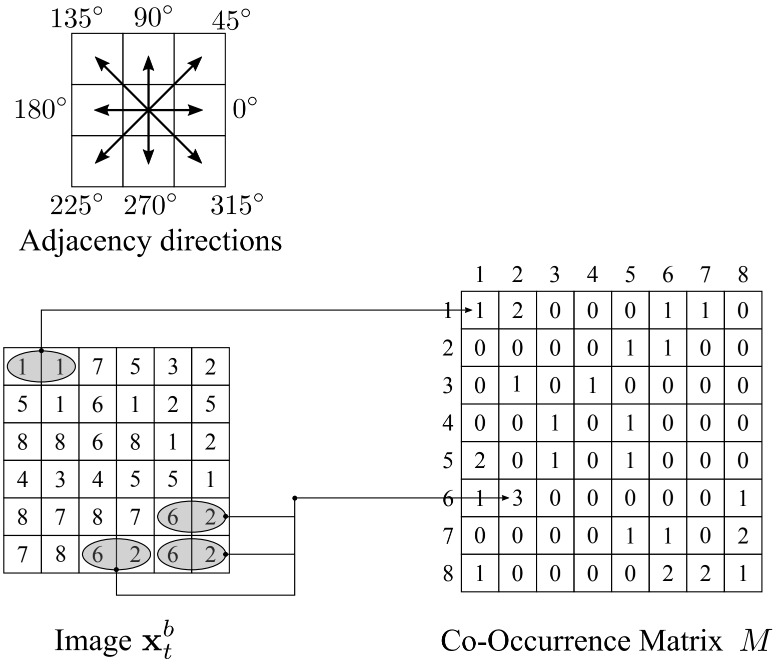
Exact inference, which is the task of finding the optimal label configuration $\mathbf{y}$ based on the proposed model in Equation 3-5, is computationally intractable for CRF, except for special cases in binary classification [46]. Approximations are usually employed to infer a solution such as pseudo-likelihood, mean field or Loopy Belief Propagation (LBP). In this thesis, sum-product LBP[53] was adopted, which is a standard approximate inference algorithm for undirected graphs with cycles and works by passing messages from each node to its neighbor nodes via edges to calculate its beliefs. Then, each node is assigned to the class with maximum belief. Further details about LBP description can be found in [54].

3.4 Gray-Level Co-Occurrence Matrix (GLCM)

Given an image $\mathbf{x}_t$, corresponding to epoch t, constituted by N_{bands_t} bands, and let $\mathbf{x}_t^b$ be the band b of image $\mathbf{x}_t$, where $b = 1, ..., N_{bands_t}$, which has a total of N_{levels} gray levels. Then, the Gray-Level Co-Occurrence Matrix (GLCM) - M, is a square matrix of order N_{levels}, where each element p_{ij} of M represents the number of occasions a pixels with intensity i is adjacent to a pixel with intensity j (see Figure 10). The adjacency might be defined to take place in each of the eight directions (0°, 45°, 90°, 135°, 180°, 225°, 270° and 315°) at certain distance of d pixels. For instance, in Figure 10, the computation of the matrix M is done from a 6×6 image with $N_{levels} = 8$, adjacency direction of 0° and distance $d = 1$ pixel.

Normalizing the matrix M by the total number of pixels pairs that satisfies the adjacency criterion, p_{ij} now represents the probability that a pixels pair (i, j) satisfies that adjacency criterion.

GLCM matrices capture texture properties but they are not directly employed for further analysis. Instead, numeric features are computed from them, which represent texture in a more compact way. Some of the most common features extracted from the GLCM matrix are the following [55].



1	1	7	5	3	2
5	1	6	1	2	5
8	8	6	8	1	2
4	3	4	5	5	1
8	7	8	7	6	2
7	8	6	2	6	2

	1	2	3	4	5	6	7	8
1	1	2	0	0	0	1	1	0
2	0	0	0	0	1	1	0	0
3	0	1	0	1	0	0	0	0
4	0	0	1	0	1	0	0	0
5	2	0	1	0	1	0	0	0
6	1	3	0	0	0	0	0	1
7	0	0	0	0	1	1	0	2
8	1	0	0	0	0	2	2	1

Figure 10: Example of how the Gray-Level Co-Occurrence Matrix M is generated from a 6 × 6 Image $\mathbf{x}_t^b$ with 8 gray levels.

Correlation

A measure of how correlated a pixel is to its neighbor over the entire image. Correlation ranges from 1 to -1 and is defined as follows:

$$\text{Correlation} = \sum_{i=0}^{N_{levels}-1} \sum_{j=0}^{N_{levels}-1} p_{ij} \frac{ijp_{ij} - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{3-8}$$

where μ_x, μ_y, σ_x and σ_y are the means and standard deviations of the matrices obtained by summing the rows or columns of M. They are defined by:

$$\mu_x = \sum_{i=0}^{N_{levels}-1} i \sum_{j=0}^{N_{levels}-1} p_{ij} \tag{3-9}$$

$$\mu_y = \sum_{i=0}^{N_{levels}-1} \sum_{j=0}^{N_{levels}-1} jp_{ij} \tag{3-10}$$

$$\sigma_x^2 = \sum_{i=0}^{N_{levels}-1} (i - \mu_x)^2 \sum_{j=0}^{N_{levels}-1} p_{ij} \tag{3-11}$$

$$\sigma_y^2 = \sum_{j=0}^{N_{levels}-1} (j - \mu_y)^2 \sum_{i=0}^{N_{levels}-1} p_{ij} \tag{3-12}$$

Homogeneity

Measures the spatial closeness of the distribution of elements in M to the diagonal. Homogeneity ranges from 0 to 1, achieving its maximum value when M is a diagonal matrix.

$$\text{Homogeneity} = \sum_{i=0}^{N_{levels}-1} \sum_{j=0}^{N_{levels}-1} \frac{1}{1+(i-j)^2} p_{ij} \tag{3-13}$$

3.5 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a kind of neural network specialized for processing data having grid-like topology such as time series (1D grid with regular time intervals) or image data (2D grid of pixels).

Introduced by LeCun et al. [47], LeNet-5 was the first CNN, outperforming other approaches and becoming the state-of-the art for hand-written digit classification.

A regular Neural Network (NN) transforms an input by putting it through a series of hidden layers. Every layer comprises a set of neurons, where each layer is fully connected to all neurons in the layer before. Finally, there is a last fully-connected layer —the output layer —that represents the class scores.

CNNs are quite different from regular NNs. First, the layers are organised in three dimensions: width (w), height (h) and depth ($N_{features}$). Additionally, the neurons in one layer do not connect to all the neurons in the next layer but only to a small region of it. Lastly, the final output will be reduced to a single vector of probability scores, organized along the depth dimension.

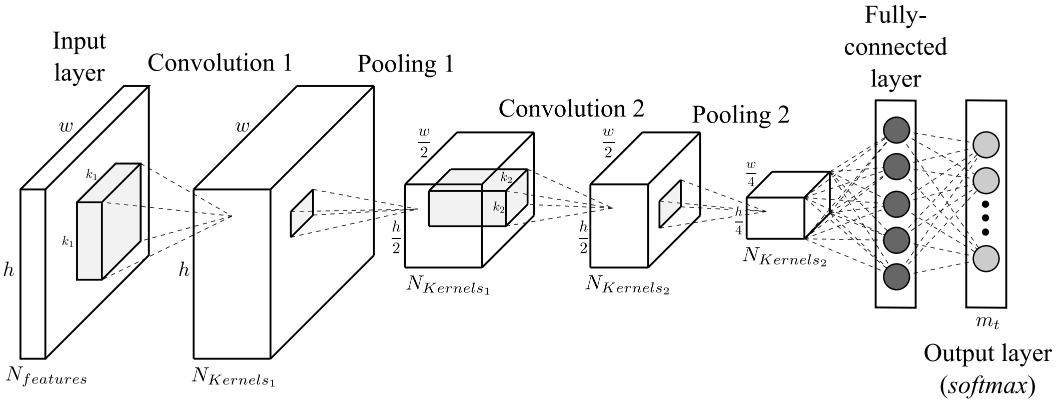


Figure 11: A CNN basic architecture with two convolutional layers.

Figure 11 illustrates a basic CNN architecture with two convolutional layers (with kernels sizes k_1 and k_2) + 2 × 2 pooling, a fully-connected layer, and finally, the output layer with m_t neurons, where m_t is the number of classes available in epoch t.

In the following, each block is detailed as well as other layers commonly used in CNN architectures.

Convolutional layer

The term convolutional layer refers to the mathematical operation *convolution*, which is a combination of two functions to produce a third function, merging two sets of information.

The convolution is executed on the input data $w \times h \times N_{features}$ with the use of a *filter* or *kernel* with size $k \times k \times N_{features}$, by sliding the kernel over the input to produce a feature map. At every location, an element-wise matrix multiplication is performed and the products are summed up onto the feature map. The feature map's dimensions depend on the dimensions w and h of the input data, and the number of kernels employed, $N_{kernels}$, as well as the kernel size and the stride.

Similar to regular NNs, activation functions are employed to create a non-linear output. For CNNs, the output of the convolutional layer is passed through the activation function. The most common activation functions are: *sigmoid*, *tanh*, *ReLU* and *Leaky ReLU*.

Pooling layer

Such a layer is usually inserted in-between successive convolutional layers in a CNN architecture. It reduces the spatial size of the feature map to reduce the amount of parameters and computation in the network. The most common is a 2×2 *max-pooling*, which applies a 2×2 filter with a stride of 2 to every depth slice of the feature map replacing all values inside the filter by the maximum value among them. Only the spatial dimensions are reduced by a factor of 2, while the depth dimension remains unchanged.

Batch Normalization

Batch Normalization (B.N.) [56] reduces the dependency on the initialization and improves convergence. It forces the set of activations throughout a network to have zero mean and unit variance for each training mini-batch.

Fully-connected layer

Neurons in this layer have full connections to all activations in the previous layer, as in regular NNs. Thus, their activations are computed by a matrix multiplication plus a bias offset.

Dropout

Dropout aims to reduce over-fitting [57] during the feature-learning procedure by randomly setting some activations to zero in each forward pass. It is like training a large ensemble of models that share parameters.

Weight decay

Weight decay is a standard trick to improve the generalization performance of neural networks by encouraging the weights to be small in magnitude [58]. It is performed by scaling the weights by a factor less than 1 in each iteration.

3.6 Accuracy assessment

Given the set of predicted labels $\hat{\mathbf{y}} = \{\hat{y}_{i,t}\}$ and their corresponding references $\mathbf{y} = \{y_{i,t}\}$, where $\hat{y}_{i,t}, y_{i,t} \in \{l_1, ..., l_m\}$ and m are the number of classes, the *confusion matrix* CM reports the classifier accuracy on a class-by-class basis. The element $CM(i,j)$ of the $confusion\ matrix$ contains the number of pixels of true class l_i assigned to class l_j by the classifier.

Table 2: Confusion Matrix—CM.

		Classification				
		l_1	l_2	l_3	$\cdots$	l_m
Reference	l_1	CM_{11}	CM_{12}	CM_{13}	$\cdots$	CM_{1m}
	l_2	CM_{21}	CM_{22}	CM_{23}	$\cdots$	CM_{2m}
	l_3	CM_{31}	CM_{32}	CM_{33}	$\cdots$	CM_{3m}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	l_m	CM_{m1}	CM_{m2}	CM_{m3}	$\cdots$	CM_{mm}

True positives (tp) are defined as the correct classifications for each class and, the number of tp is calculated by the diagonal of CM. *False positives* (fp) are those pixels that were erroneously considered as a part of a class and *false negatives* (fn) are those pixels that truly belongs to a class but were considered as part of an other class. The number fp and fn for all classes can be computed from CM by the sum of their corresponding column and row, respectively, of its elements excluding the main diagonal (tp).

Then, the *Overall Accuracy* (OA) is defined by:

$$OA = 100 \times \frac{\sum\limits_{i=1}^{m} CM_{ii}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m} CM_{ij}} \tag{3-14}$$

OA is a global metric that measures how many pixels were correctly classified with respect to the total number of pixels. It varies from 0 to 100 representing complete disagreement and perfect agreement between reference and predicted classification, respectively.

Precision, also known as *Correctness*, and *Recall*, also known as *Completeness*, are defined as follows:

$$Precision = \frac{tp}{tp + fp} \tag{3-15}$$

$$Recall = \frac{tp}{tp + fn} \tag{3-16}$$

Then, the $F1$-*score* is given by the harmonic mean of *Precision* and *Recall*, varying from 0 to 100 similar to OA.

$$F1 - score = 100 \times \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3-17}$$

As the $F1$-*score* is defined per class, the average $F1$-*score* is employed as a global metric which gives us a better idea about how accurate each class has been classified, disregarding the fact that some classes are more abundant than others.

$$\text{Average } F1 - score = \frac{\sum_{i=1}^{m} F1 - score_i}{m_t} \tag{3-18}$$

where $F1$-$score_i$ is the $F1$-$score$ of class l_i, and m_t is the number of available classes in epoch t.

4
CRF BASED CROP RECOGNITION MODEL

This chapter describes the proposed method based on CRF for crop recognition. In the following, we also describe the variants conceived to accommodate single or multiple sensors with potentially different spatial resolutions. Such variants are characterized by different definitions of AP, SIP, and TIP potentials as shown in the following.

Equation 3-5, which is below to facilitate reading, is a general model for multi-temporal images, which considers the spatial and temporal contexts.

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\left[\exp\left(\sum_{t\in T}\sum_{i\in S} AP^t(y_{i,t},\mathbf{x}) + \theta \sum_{t\in T}\sum_{i\in S}\sum_{j\in N_i} SIP^t(y_{i,t},y_{j,t},\mathbf{x}) + \sum_{t\in T}\phi^t \sum_{i\in S}\sum_{k\in C_i} TIP^{tk}(y_{i,t},y_{i,k},\mathbf{x})\right)\right] \quad (4\text{-}1)$$

It may be instantiated in different ways, called *variants* hereafter, depending on the setup of parameters θ and ϕ^t, and especially, on the design of association, spatial and temporal interaction potentials. In this work we propose some CRF model variants for crop recognition from multitemporal SAR and optical images. We also consider some variants to serve as baselines and to assess how different design choices impact the accuracy. In particular, we investigate the contribution of the spatial and/or the temporal interaction potentials by turning them selectively off, i.e., by setting the parameters θ and/or ϕ to zero. In the following we present the alternative designs considered in this work for the potentials that make up the model of equation 3-5.

4.1
Association Potential (AP)

The association potential $AP^t(y_{i,t},\mathbf{x})$ is given in this thesis by the logarithm of the posterior probability of class $y_{i,t}$ given a feature $\mathbf{f}_{i,t}(\mathbf{x})$. The posterior is computed by a discriminative classifier, formally:

$$AP^t(y_{i,t},\mathbf{x}) = \log P_C(y_{i,t}|\mathbf{f}_{i,t}(\mathbf{x})) \quad (4\text{-}2)$$

where the subscript C stands for the classifier used to estimate posterior probabilities.

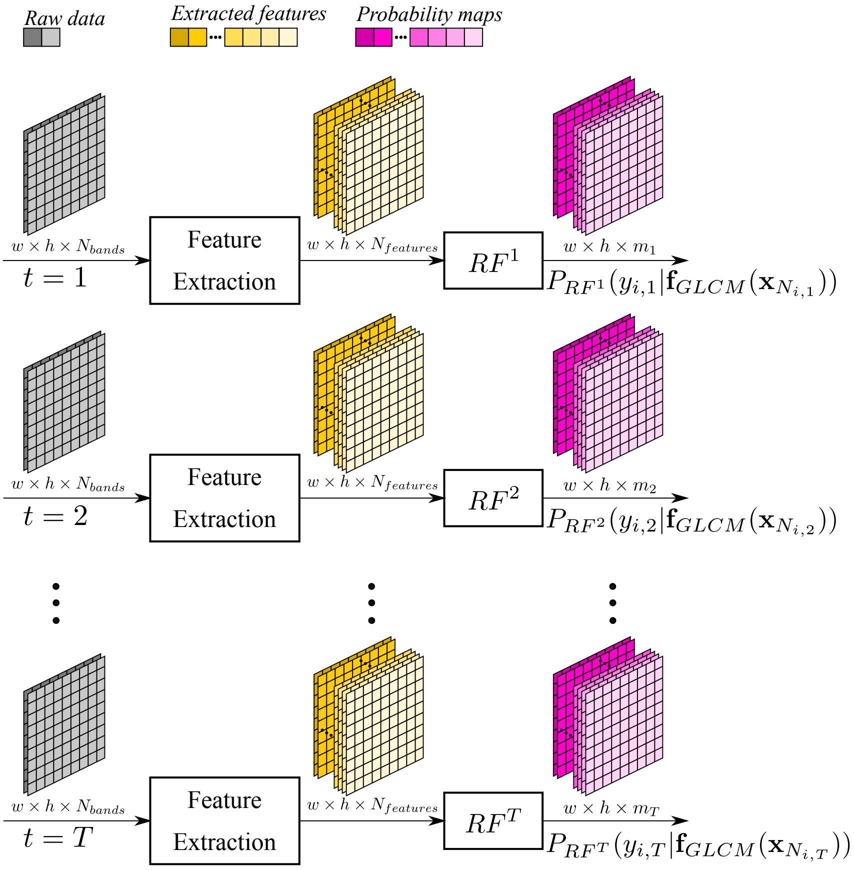
Three groups of model variants for the association potential are considered in the following.

Association Potential from RF classifier and GLCM features

Texture features extracted from GLCM, denoted henceforth $\mathbf{f}_{GLCM}(\mathbf{x}_{N_{i,t}})$, are frequently used for SAR data. Instead of considering the whole dataset ($\mathbf{x}$), such texture features are computed only upon a neighborhood $N_{i,t}$ centered at the i-th pixel location in epoch t. Yet, GLCM based features capture partially the spatial context. Some model variants that handle SAR data in our subsequent analysis rely on a Random Forest (RF) [59] classifier. In such variants, the association potential takes the form:

$$AP^t(y_{i,t}, \mathbf{x}) = \log P_{RF^t}(y_{i,t}|\mathbf{f}_{GLCM}(\mathbf{x}_{N_{i,t}})) \tag{4-3}$$

Figure 12 illustrates how the association potential is computed in the aforesaid variants.



Raw data
Extracted features
Probability maps
$w \times h \times N_{bands}$
$t = 1$
Feature Extraction
$w \times h \times N_{features}$
RF^1
$w \times h \times m_1$
$P_{RF^1}(y_{i,1}|\mathbf{f}_{GLCM}(\mathbf{x}_{N_{i,1}}))$
$w \times h \times N_{bands}$
$t = 2$
Feature Extraction
$w \times h \times N_{features}$
RF^2
$w \times h \times m_2$
$P_{RF^2}(y_{i,2}|\mathbf{f}_{GLCM}(\mathbf{x}_{N_{i,2}}))$
$w \times h \times N_{bands}$
$t = T$
Feature Extraction
$w \times h \times N_{features}$
RF^T
$w \times h \times m_T$
$P_{RF^T}(y_{i,T}|\mathbf{f}_{GLCM}(\mathbf{x}_{N_{i,T}}))$

Figure 12: Association Potential computed from GLCM features and a Random Forest (RF) classifier.

Association Potential from raw data and CNN

Some variants explored in our model rely on a CNN to learn features from single epoch data and to provide posterior probabilities. The posterior probabilities to a pixel at location i in epoch t is computed by a CNN taking as input the image patch $N_{i,t}$ centered at i in t. Patches are made of all available polarizations or spectral bands for SAR and optical data, respectively. The association potential takes the following form:

$$AP^t(y_{i,t}, \mathbf{x}) = \log P_{CNN^t}(y_{i,t}|\mathbf{x}_{N_{i,t}}) \tag{4-4}$$

Note that in this case the association potential also takes spatial context into account. Figure 13 illustrates how the association potential is computed in the CNN variants based on single epoch data.

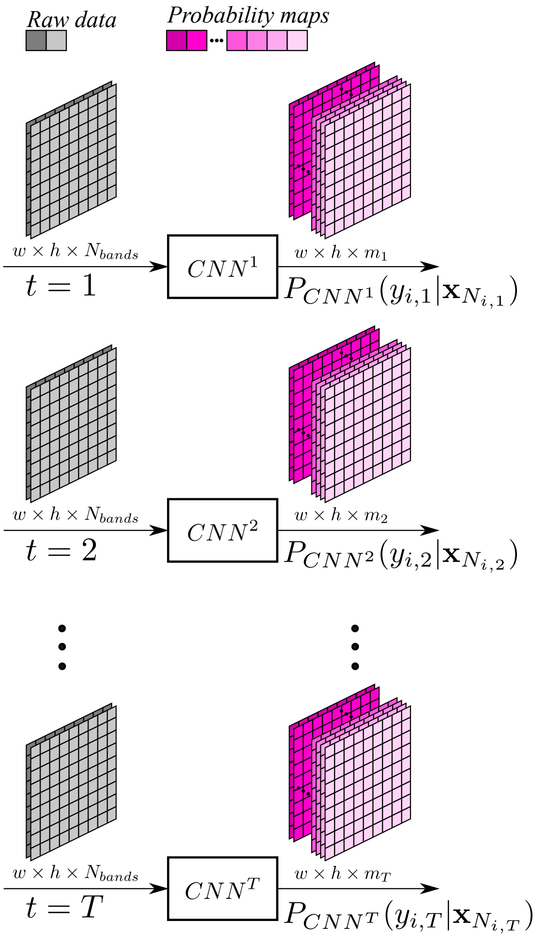

Raw data
Probability maps
$w \times h \times N_{bands}$
CNN^1
$w \times h \times m_1$
$t = 1$
$P_{CNN^1}(y_{i,1}|\mathbf{x}_{N_{i,1}})$
$w \times h \times N_{bands}$
CNN^2
$w \times h \times m_2$
$t = 2$
$P_{CNN^2}(y_{i,2}|\mathbf{x}_{N_{i,2}})$
$w \times h \times N_{bands}$
CNN^T
$w \times h \times m_T$
$t = T$
$P_{CNN^T}(y_{i,T}|\mathbf{x}_{N_{i,T}})$

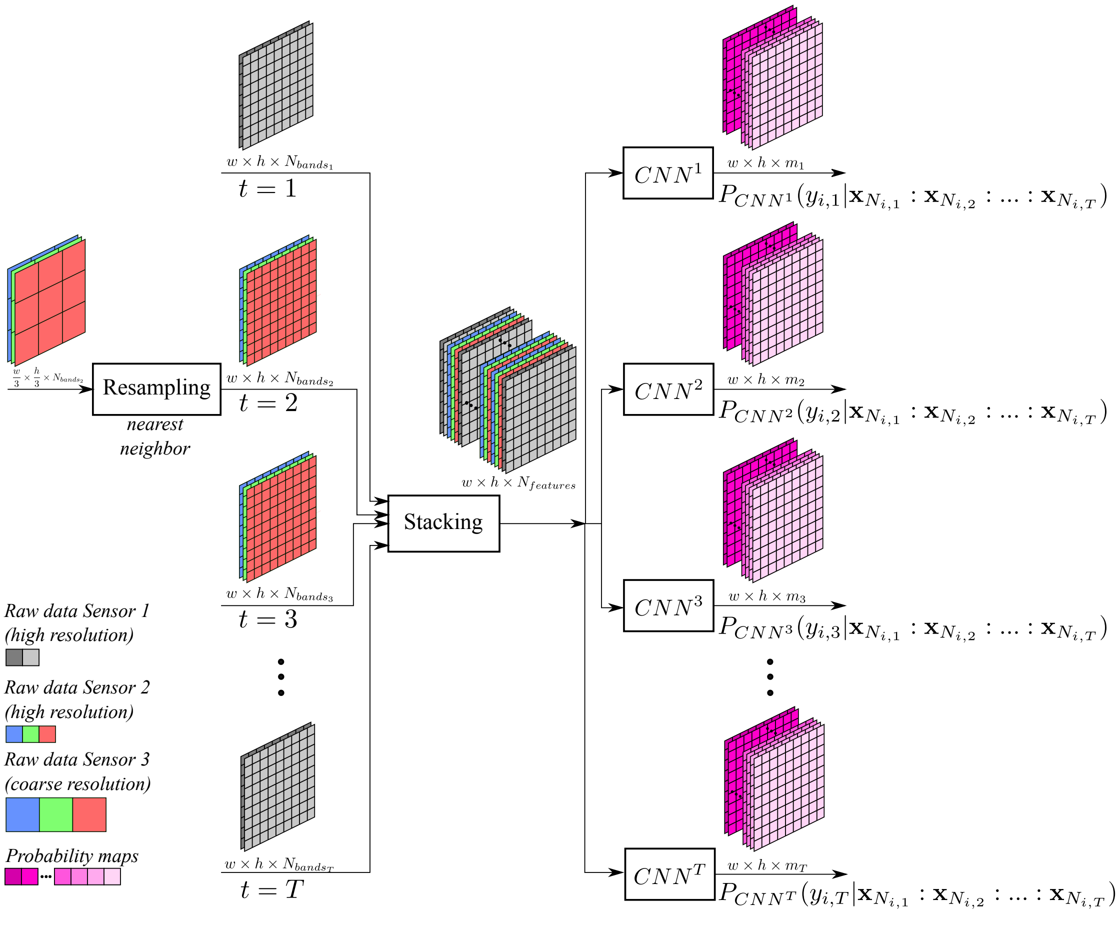
Figure 13: Association Potential computed from raw SAR or optical data by a Convolutional Neural Network.

Association Potential from stacked data and CNN

A third group of model variants also implements the association potential by means of CNNs. The difference to the previous approach lies in the network input. Instead of considering only the data related to the epoch t the association potential is to be computed for all patches centered at location i through all epochs in the dataset up to T are stacked one upon the other forming an input tensor for a CNN. This technique, called *image stacking*, captures both the spatial and the temporal context. With some abuse of notation, we denote these association potential variants as:

$$AP^t(y_{i,t}, \mathbf{x}) = \log P_{CNN^t}(y_{i,t}|\mathbf{x}_{N_{i,1}} : \mathbf{x}_{N_{i,2}} : ... : \mathbf{x}_{N_{i,T}}) \tag{4-5}$$

where $\mathbf{x}_{N_{i,1}} : \mathbf{x}_{N_{i,2}} : ... : \mathbf{x}_{N_{i,T}}$ denotes the concatenation along the third dimension of patches starting with the initial epoch until epoch T. Figure 14 illustrates this group of variants.



$w \times h \times N_{bands_1}$
$t = 1$
$\frac{w}{3} \times \frac{h}{3} \times N_{bands_2}$
Resampling
nearest neighbor
$w \times h \times N_{bands_2}$
$t = 2$
$w \times h \times N_{bands_3}$
$t = 3$
$w \times h \times N_{bands_T}$
$t = T$
Stacking
$w \times h \times N_{features}$
CNN^1
$w \times h \times m_1$
$P_{CNN^1}(y_{i,1}|\mathbf{x}_{N_{i,1}} : \mathbf{x}_{N_{i,2}} : ... : \mathbf{x}_{N_{i,T}})$
CNN^2
$w \times h \times m_2$
$P_{CNN^2}(y_{i,2}|\mathbf{x}_{N_{i,1}} : \mathbf{x}_{N_{i,2}} : ... : \mathbf{x}_{N_{i,T}})$
CNN^3
$w \times h \times m_3$
$P_{CNN^3}(y_{i,3}|\mathbf{x}_{N_{i,1}} : \mathbf{x}_{N_{i,2}} : ... : \mathbf{x}_{N_{i,T}})$
CNN^T
$w \times h \times m_T$
$P_{CNN^T}(y_{i,T}|\mathbf{x}_{N_{i,1}} : \mathbf{x}_{N_{i,2}} : ... : \mathbf{x}_{N_{i,T}})$
Raw data Sensor 1 (high resolution)
Raw data Sensor 2 (high resolution)
Raw data Sensor 3 (coarse resolution)
Probability maps

Figure 14: Association Potential computed from stacked data by a Convolutional Neural Network.

When dealing with a multi-sensor dataset, we up-sample the lower resolution data using the *nearest neighbor* approach in order to obtain a common resolution throughout the data set. Nearest neighbor is employed as it does not alter the distribution of pixel values. The different grid sizes of some elements of Figure 14 points to different spatial resolutions and the up-sampling strategy.

4.2 Spatial Interaction Potential (SIP)

The spatial interaction potential $SIP^t(y_{i,t}, y_{j,t}, \mathbf{x})$ in the CRF model given in equation 3-5 has the whole dataset $\mathbf{x}$ as one of its input parameters. In this thesis we take a somewhat simpler formulation for the spatial interaction potential $SIP^t(y_{i,t}, y_{j,t}, \mathbf{x}_{i,t}, \mathbf{x}_{j,t})$, which requires only the data in spatially neighboring sites i and j ($j \in N_i$).

The spatial interaction potential is given by a contrast-sensitive Potts model [60] defined by:

$$SIP^t(y_{i,t}, y_{j,t}, \mathbf{x}_{i,t}, \mathbf{x}_{j,t}) = \begin{cases} p + (1-p)e^{-d^2_{ij,t}/2\sigma_t^2} & , y_{i,t} = y_{j,t} \\ 0 & , y_{i,t} \neq y_{j,t} \end{cases} \tag{4-6}$$

where $\mathbf{g}_{i,t}(\mathbf{x}_{i,t})$ is the feature vector of site i in epoch t.

It takes into account the similarity between adjacent site features vectors as given by its Euclidean distance, $d_{ij,t} = \|\mathbf{g}_{i,t}(\mathbf{x}_{i,t}) - \mathbf{g}_{j,t}(\mathbf{x}_{j,t})\|$. Notice that $\mathbf{g}_{i,t}(\mathbf{x}_{i,t})$ might be different from $\mathbf{f}_{i,t}(\mathbf{x}_{i,t})$ used for the association potential. σ_t^2 refers to the mean value of squared feature distances $d_{ij,t}$ computed during training. The parameter $p \in [0, 1]$ in Equation 4-6 controls the relative influence of the data-dependent and data-independent terms.

4.3 Temporal Interaction Potential (TIP)

The design of association and spatial interaction potentials presented in the previous sections is pretty standard in CRF modeling. The contributions of this thesis lie primarily in how the temporal interaction potential is designed to accommodate the high diversity in crop dynamics in tropical regions. Variants for single and multiple sensor sequences are proposed in the following.

4.3.1 Model for single sensor sequences

The general CRF model given by equation 3-5 admits that the temporal interaction potential might depend on the data ($\mathbf{x}$). Tuning such a model accurately would require a number of training samples that can't be afforded in most applications. Thus, we drop data dependency of the temporal interaction potential, which takes the form $TIP^{tk}(y_{i,t}, y_{i,k})$. In this case, the model must provide a potential estimate for each pair of classes, which amounts to estimating $m_t * m_k$ values, for an application involving m_t and m_k classes at epochs t and k, respectively. This would still require an amount of labeled samples that can hardly be obtained in most real applications.

The solution we propose relies on prior knowledge about transitions that may or may not occur between any pair of adjacent epochs present in the dataset. It is represented by a $m_t \times m_k$ transition matrix, the elements of which take either the value 0 or -∞, depending on whether the corresponding class transition is possible or not. In this way, the probability of any class configuration $\mathbf{y}$ containing at least one impossible class transition will be set to zero, and will consequently be discarded as a solution. On the other hand, the temporal interaction potential will take the same non-zero value determined by ϕ, for all class configurations containing no impossible class transitions.

4.3.2
Model for Multi Sensor Sequences

Generally, it is difficult to obtain a sequence of multi-temporal images from the same sensor at a regular time interval, due to different reasons, e.g. cloud coverage in case of optical images. A possible way to circumvent this hindrance is the usage of multi source data to fill the gaps in the sequence. However, these images may have different spatial resolutions, besides being of different domains (e.g. optical and radar).

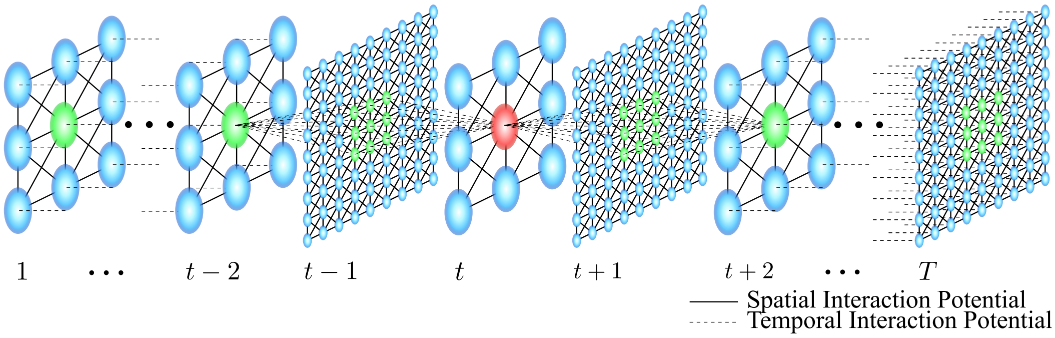


Figure 15: Multi-temporal sequence of T co-registered images with different spatial resolutions represented as a graph.

When dealing with images with different spatial resolutions, the edges of CRF graph defining the temporal neighborhood are modeled as illustrated in Figure 15 [31], where it is assumed that the spatial resolutions between two sensors differ by a factor of three in each spatial dimension. In this case, each pixel of the coarser image will be connected by temporal edges to nine pixels of the higher resolution image.

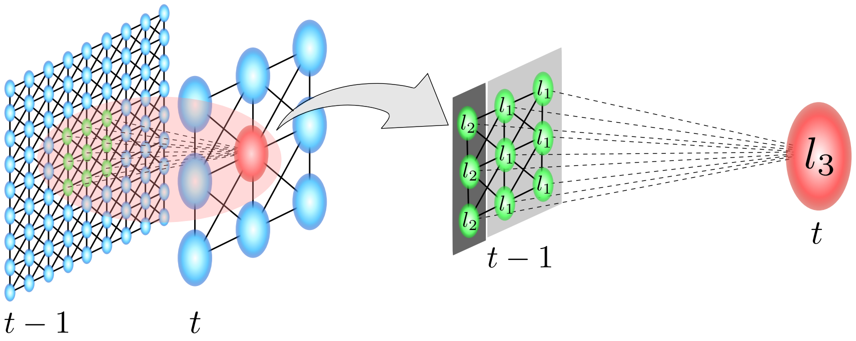


Figure 16: Images corresponding to two adjacent epochs with images of different spatial resolutions: fine in $(t-1)$ and coarse in (t).

This graph may lead to violation of the prior knowledge the temporal interaction model proposed in the previous section relies on, as will be demonstrated in the following simple example.

Let's suppose that a plot border goes through a 3×3 pixel array of a fine resolution image captured at $t-1$, as shown in Figure 16. A small part of such pixels belong to a plot covered by class l_2, while most of them belong to class l_1. In the coarser image at t this array corresponds to a single pixel. Let's further assume that an agriculture expert tells that between $t-1$ and t only two class transitions are possible in that area, namely $l_1 \rightarrow l_3$ and $l_2 \rightarrow l_4$. And let the ground truth of the pixel at t be l_3.

In this ground truth scenario, the class transition $l_2 \rightarrow l_3$ will occur, conflicting with prior knowledge constraints. Certainly, a temporal interaction potential designed as proposed in the previous section, would prevent assignment either to class l_2 at $t-1$ or to class l_3 at t or both, leading to classification errors in at least one of these epochs.

At first glance, this problem can be easily overcome by relaxing the constraints imposed by prior knowledge by modifying the transition matrix to change the temporal association potential so as to allow transition $l_2 \rightarrow l_3$ between $t-1$ and t. This simple solution has a deleterious effect on the model's accuracy since it renders eligible a number of class sequences that contradict the prior knowledge about crop dynamics in the target region.

In order to mitigate this shortcoming of the aforementioned solution, we propose the incorporation of higher order connections in the temporal domain between any pair of high spatial resolution images separated by a sequence of low resolution images, as depicted in Figure 17.

Again, such high order connections carry a m by m transition matrix, whose elements are either 0 or -∞, depending on whether the classes are compliant with the prior knowledge or not.

In summary, our approach relaxes the constraints regarding possible and non-possible class transitions along the temporal dimension when involving adjacent images of different resolutions. Additionally, the high order temporal edges reestablish such constraints in the next image of the sequence having the same fine resolution.

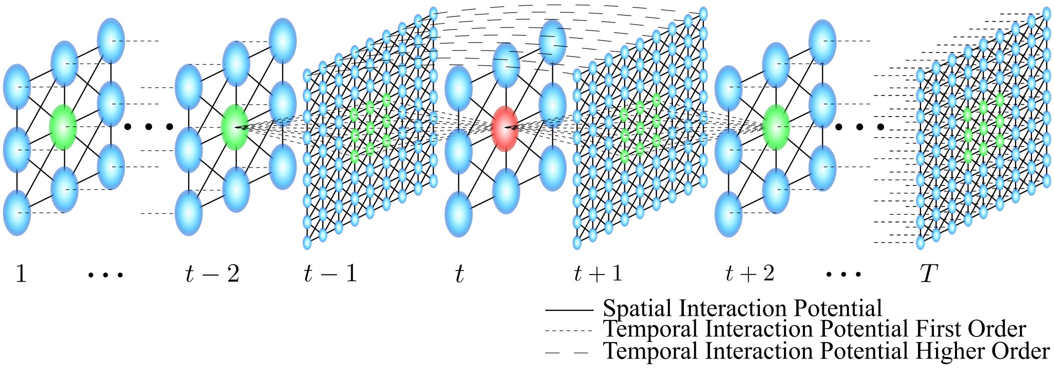

1
⋯
$t-2$
$t-1$
t
$t+1$
$t+2$
⋯
T
Spatial Interaction Potential
Temporal Interaction Potential First Order
Temporal Interaction Potential Higher Order

Figure 17: Multi-temporal sequence of T co-registered images with different spatial resolutions represented as a graph with higher order connections in the temporal domain between high spatial resolution images.

5 EXPERIMENTAL ANALYSIS

This chapter reports the experiments carried out in order to validate the method proposed in the previous chapter. First, the datasets used in the experiments are presented. Then, the experimental protocol followed for multi-temporal classification is described, and the parameter setup is detailed. Finally, the results obtained in the experiments are reported, in terms of *Overall Accuracy* (OA) and average $F1$-*score*.

5.1 Datasets

Two municipalities from Brazil, Campo Verde in Mato Grosso and Luis Eduardo Magalhães in Bahia, were selected to evaluate the proposed method in tropical regions. These datasets are publicly available for the scientific community.

The datasets corresponding to those municipalities were elaborated in cooperation with the National Institute for Space Research (INPE) and Brazilian Agricultural Research Corporation (EMBRAPA). They were in charge of the field works to collect data from the study areas to assign a label to each plot, while the acquisition and pre-processing of Sentinel-1A scenes was performed at PUC-Rio. These datasets received financial support from the Brazilian agencies CAPES[1] and CNPq[2] for Campo Verde [11], and from the ISPRS[3], for Luis Eduardo Magalhães [12].

5.1.1 Campo Verde, Mato Grosso

Campo Verde is a municipality of the Mato Grosso (MT) state in the central west region of Brazil (see Figure 18) at a latitude of $15°32'48''$ south and a longitude of $55°10'08''$ west, with an area of 4,800 km^2 and approximately with an altitude of 736 m [11]. It presents a *Tropical Aw* climate according to the Köppen—Geiger classification [61], with an average temperature of 22.3°C, average annual rainfall of 1,726 mm and latosoils as predominant soils.

[1] Coordination of Superior Level Staff Improvement

[2] National Counsel of Technological and Scientific Development

[3] International Society for Photogrammetry and Remote Sensing

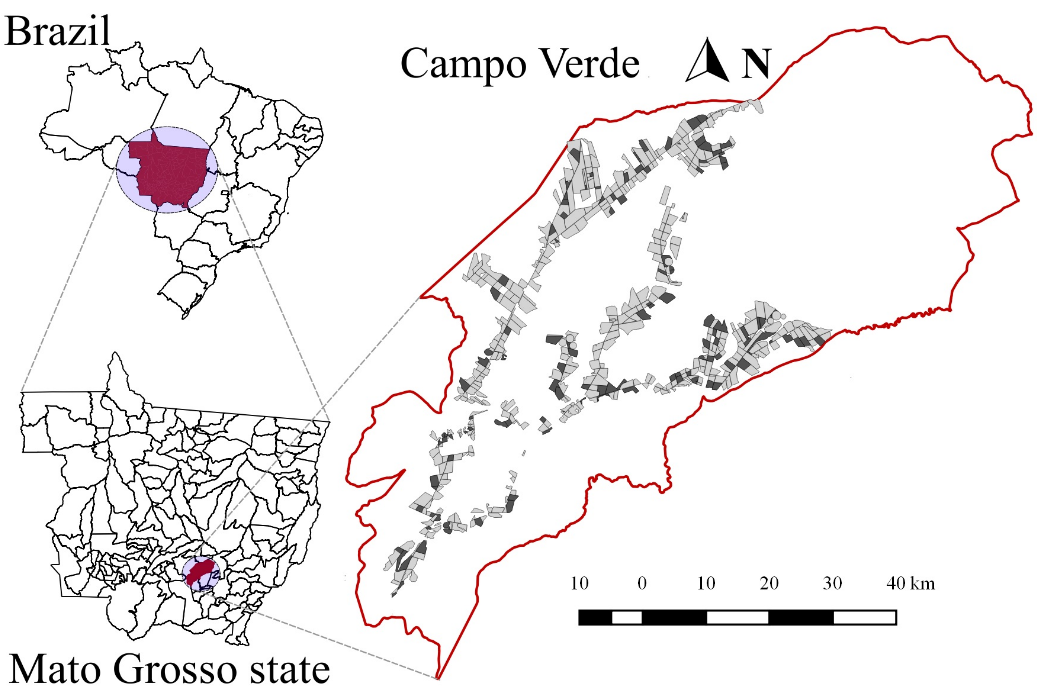


Figure 18: Location of the Campo Verde dataset in the state of Mato Grosso, Brazil. It comprises 513 fields divided into training (black fields) and validation (gray fields) sets with approximately 20% and 80%, respectively.

Table 3: Campo Verde dataset: Acquisition dates either from Sentinel-1A or Landsat 8.

Year	**Month**	**Date**	
		Sentinel-1A (10m)	**Landsat 8 (30m)**
2015	October	29	-
	November	10, 22	11
	December	04, 16	-
2016	January	21	-
	February	14	-
	March	09, 21	-
	April	-	19
	May	08, 20	05
	June	13	
	July	07, 31	08, 24

The dataset comprises a set of 513 fields[4] with their respective land-use classes and a set of Sentinel-1A and Landsat 8 images acquired between

[4] Available at: http://www.obt.inpe.br/agricultural-database/campoverde/

October 2015 and July 2016 (see acquisition dates in Table 3). Level-1 Interferometric Wide Swath (IWS) mode Ground Range Detected (GRD) Sentinel-1A products (C-band) in VV and VH polarizations were acquired the Copernicus Open Access Hub[5], geometrically corrected using SRTM's DEM and radiometrically calibrated to sigma nought (σ^0) back-scatter coefficient, converted to *db*, co-registered and geo-referenced to UTM 21S/WGS84, mosaicked and clipped. Level-1 Landsat 8 data products were acquired from the United States Geological Survey (USGS) Earth Resources Observation and Science Center, atmospherically corrected, mosaicked and clipped. Further details about this dataset can be found in [11].

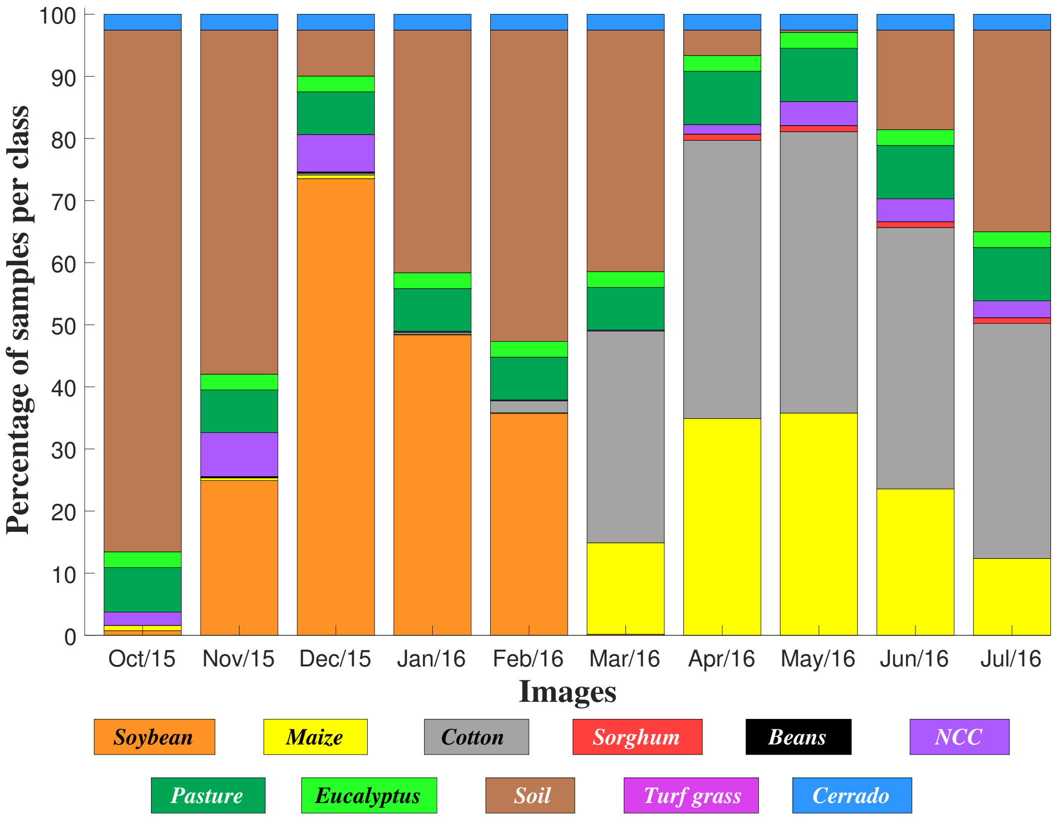


Figure 19: Percentage of samples per class in each epoch of the Campo Verde dataset.

The reference data was built by two field campaigns conducted between December 14th—18th, 2015 (first harvest, summer crops and rainy season) and May 9th—13th, 2016 (second harvest, dry season) to record the localization and land-use classes of each field. Land-use classes for the remaining months were assigned by visual classification. A total of 11 land-use classes are present: *Soybean*, *Maize*, *Cotton*, *Beans*, *Sorghum*, *Non-Commercial Crops—NCC* (*Millet*, *Crotalaria*, *Brachiaria*, *Grasses*), *Pasture*, *Turf grass*, *Eucalyptus*, *Cerrado*

[5] https://scihub.copernicus.eu/

(brazilian savanna), *Soil* (i.e. bare soil, soil with crop residues/weeds). Figure 19 summarizes the percentage of each class per month. Notice the presence of two main periods: from October 2015 to February 2016, where *Soybean* is the predominant crop, and from March 2016 to July 2016, where *Maize* and *Cotton* are the most abundant crops in this region.

5.1.2 Luis Eduardo Magalhães, Bahia

Luis Eduardo Magalhães (LEM) is a municipality in the state of Bahia in the north-east region of Brazil (see Figure 20) at a latitude of 12°05′31″ south and longitude 45°48′18″ west, with an area of 4,000 km^2 approximately and with an altitude of 720 m [12]. It presents a Tropical Aw climate according to the Köppen—Geiger classification [61], with an average temperature of 24.2 °C, average annual rainfall of 1,511 mm and yellow latosol as predominant soil in this region.

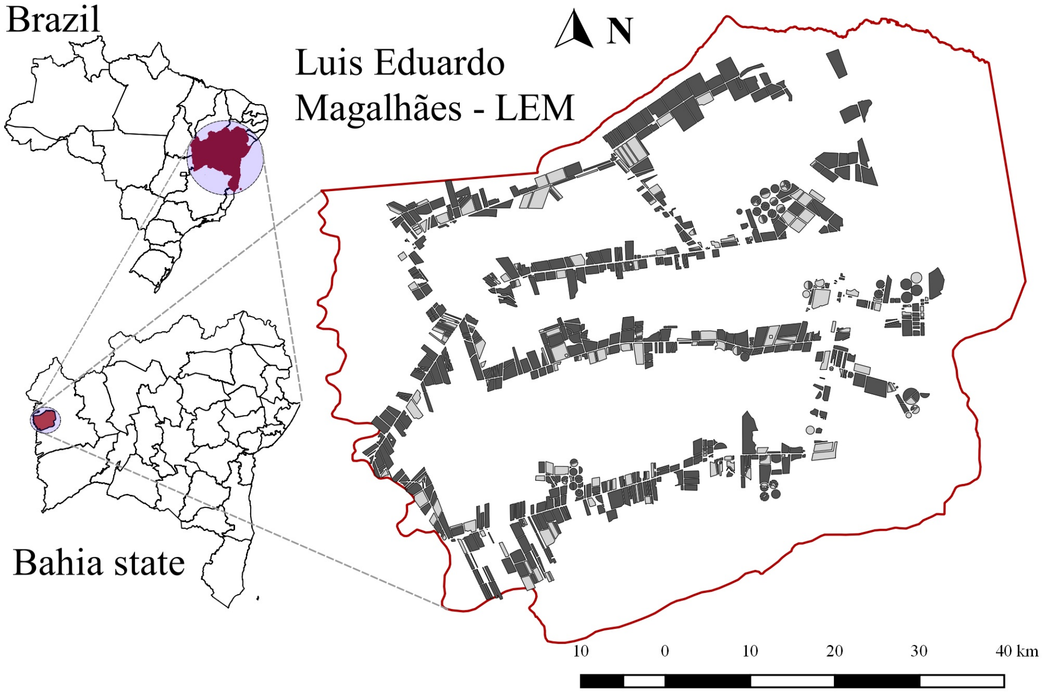


Figure 20: Location of the LEM dataset in the state of Bahia, Brazil. It comprises 794 fields divided into training (black fields) and validation (gray fields) sets with approximately 75% and 25% respectively.

The dataset comprises a set of 794 fields[6] with their respective land-use classes and a set of Sentinel-1A, Sentinel-2A/-2B and Landsat 8 images acquired between June 2017 and June 2018 (see acquisition dates in Table 4).

[6] Available at: http://www.obt.inpe.br/agricultural-database/lem/

Similar to Campo Verde, Level-1 IWS mode GRD Sentinel-1A products (C-band) in VV and VH polarizations were acquired, geometrically corrected and radiometrically calibrated, converted to *db*, co-registered and geo-referenced, mosaicked and clipped. Level-1C Sentinel-2A/-2B images were acquired in top-of-atmosphere reflectance from the Copernicus Open Access Hub, atmospherically corrected using the Sentinel-2 Atmospheric Correction (Sen2Cor) algorithm, mosaicked and clipped. Level-2 Landsat 8 data products were acquired from the USGS Earth Resources Observation and Science Center in surface reflectance, mosaicked and clipped. Further details about this dataset can be found in [12].

Table 4: LEM dataset: Acquisition dates either from Sentinel-1A, Sentinel-2A/2B and Landsat 8.

		Date		
Year	**Month**	**Sentinel-1A (10m)**	**Sentinel-2A/2B (10m)**	**Landsat 8 (30m)**
2017	June	12, 24	-	15
	July	06, 30	29	01, 17
	August	11, 23	03	02, 18
	September	04, 16, 28	07	19
	October	10	17, 22	21
	November	03, 15, 27	-	-
	December	09, 21	-	-
2018	January	02, 14, 26	-	-
	February	07, 19	-	-
	March	03, 15, 27	-	-
	April	08, 20	20, 30	-
	May	02, 14, 26	10	01
	June	07, 19	14, 19, 24	02, 18

The reference data was built by two field campaigns conducted between June 26th—30th, 2017 (second harvest, dry season) and March 14th—19th, 2018 (first harvest, wet season) to record the geographic coordinates, land-use classes of each field and its phenology phase. Land-use classes for the other months to cover one crop year (from June 2017 to June 2018) were assigned by an experienced image interpreter. A total of 14 land-use classes are present: *Soybean*, *Maize*, *Cotton*, *Coffee*, *Beans*, *Sorghum*, *Millet* (commercial and non-commercial millet), *Eucalyptus*, *Pasture/Grass*, *Hay*, *Cerrado*, *Conversion Area* (a cerrado field that has been recently deforested), *Uncultivated Soil* (bare soil, soil with crop residues/weeds) and *Not Identified* (areas irrigated by central pivot). Figure 21 summarizes the percentage of each class per month. Notice the presence of two main periods: from June 2017 to November 2017, where there is mainly *Uncultivated Soil* as well as other minor classes like *Millet*, *Hay* and *Maize*, and from December 2017 to June 2018, where *Soybean* is the most abundant crop in this region with the presence of *Maize* and *Coffee*.

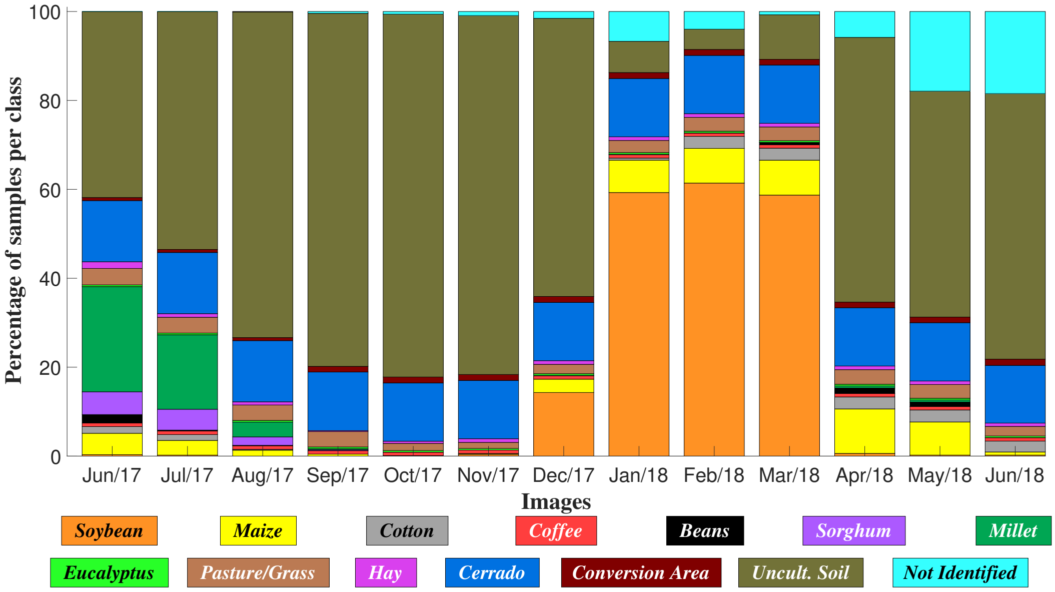


Figure 21: Percentage of samples per class in each epoch of the Luis Eduardo Magalhães dataset.

5.2 Experimental Setup

The method presented in Chapter 4 and its variants were tested for Crop Recognition using the datasets in Section 5.1.

The following nomenclature was employed to refer to each variant:

$$\{C\}_{Stack} - ASTH$$

where $\{C\}$ stands for the classification algorithm used for the association potential, which might be either a random forest (RF) or a convolutional neural network (CNN). The term *Stack* might be present or not depending on how the classification algorithm $\{C\}$ was trained, if it was upon a single image or an image stack, correspondingly. The subsequent four letters, A, S, T and H are related to which potentials are considered in the CRF model: the association potential (AP), the spatial interaction potential (SIP), the temporal interaction potential (TIP), and finally, higher order connections in the temporal domain, respectively.

For instance, RF-AS will denote a CRF model comprising the association and spatial interaction potentials (AP+SIP), where the AP is given by an RF classifier trained upon a single epoch. Moreover, CNN_{Stack}-$ASTH$ will represent a CRF model considering all three potentials (AP+SIP+TIP) with higher order connections in the temporal domain, where the AP was provided by a CNN trained upon an image stack.

For the variants that rely on handcrafted features, texture and spectral features were extracted from SAR and Optical images, respectively.

Correlation, *Homogeneity*, *Mean* and *Variance*, as in [35], were extracted from the Gray-Level Co-occurrence Matrix (GLCM) using 3 × 3 windows per polarization (VV and VH) in four directions (0°, 45°, 90° and 135°), producing a 32 dimensional feature vector. Spectral features corresponding to bands $R(4)$, $G(3)$, $B(2)$, $NIR(8)$ and $NDVI$ for Sentinel-2A/-2B, and bands from 1 to 7 and $NDVI$ for Landsat 8, were extracted, generating 5 and 8 dimensional feature vectors, respectively. The $NDVI$ was calculated as in Equation 5-1, using bands 8 and 4 for Sentinel-2A/-2B and bands 5 and 4 for Landsat 8.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{5-1}$$

The Random Forest (RF) classifier was employed with 250 random trees [62] and depth up to 25 levels. RF generates an ensemble of randomized decision trees during training. Then, the probabilistic output provided for the association potential (AP) is obtained by the ratio of the sum of all votes for a class $y_{i,t}$ of image site i and epoch t ($V_{y_{i,t}}$), and the total number of trees (N_{Trees}), as given by:

$$P_{RF^t}(y_{i,t}|\mathbf{f}_{GLCM}(\mathbf{x}_{N_{i,t}})) = \frac{V_{y_{i,t}}}{N_{Trees}} \tag{5-2}$$

As both datasets are highly unbalanced (see Figures 19 and 21), which could affect the classification results, under-/over-sampling was applied to the more/less abundant classes. For Campo Verde, 50,000 and 5,000 samples per class were employed from high (Sentinel-1A) and coarse (Landsat 8) spatial resolution images, respectively. For LEM, 10,000 and 2,000 samples per class were considered from high (Sentinel-1A and Sentinel-2A/-2B) and coarse (Landsat 8) spatial resolution images, correspondingly. These approaches are referred hereafter as RF-A, RF-AS and RF-AST, according to Chapter 4.

Table 5: CNN architecture configuration.

Layer	Output Size
- Input	$9 \times 9 \times N_{features}$
- Convolutional Block (3×3 Conv., B.N.[7], *Leaky ReLU*)	$9 \times 9 \times 80$
- Max Pooling (2×2)	$5 \times 5 \times 80$
- Convolutional Block (1×1 Conv., B.N., *Leaky ReLU*)	$5 \times 5 \times 80$
- Convolutional Block (3×3 Conv., B.N., *Leaky ReLU*)	$5 \times 5 \times 96$
- Max Pooling (2×2)	$3 \times 3 \times 96$
- Fully Connected	256
- B.N.	256
- *Leaky ReLU*	256
- Dropout	256
- *Softmax*	m_t

The variants that rely on a CNN for the AP, the network architecture employed is summarized in Table 5, where the input of the network was directly the pixel values (back-scatter responses in VV and VH polarizations for Sentinel-1, spectral bands R(4), G(3), B(2), *NIR*(8) (the bands with 10m spatial resolution) for Sentinel-2A/-2B, and bands from 1 to 7 for Landsat 8) resulting in $N_{features}$ dimensional feature vectors. The parameter setup of the CNN was: Adam optimizer, learning rate of 10^{-3}, weight decay of 10^{-4}, dropout of 0.35, batch size equals to 128 with 500 epochs and early stop to break after 10 epochs without improvement. Data augmentation was performed with flips and rotations only for minor classes. The output is a m_t dimensional vector with the local posterior probabilities, where m_t is the number of classes available in epoch t, Hereafter, these approaches are referred as *CNN-A*, *CNN-AS*, *CNN-AST* and *CNN-ASTH* according to Chapter 4.

For the stacked variants, whether using a single sensor or multiple sensors, the pixel values of each image were concatenated and employed as the CNN's input. For multiple sensors variants, when the images from different sensors have different spatial resolutions, the stack was done by up-sampling the coarse resolution images, using *nearest neighbor* method. For instance, in Campo

[7] Batch Normalization

Verde dataset, Landsat 8 images (spatial resolution 30 m) were re-sampled to 10 m as Sentinel-1A images. Hereafter, these approaches are referred as CNN_{Stack}-*A*, CNN_{Stack}-*AS*, CNN_{Stack}-*AST* and CNN_{Stack}-*ASTH*.

The SIP's weight θ (see Equation 3-5) presented in Chapter 4 might be selected empirically or by using an optimization algorithm. We carried out experiments varying the value of θ from 0.5 to 10 with steps of 0.5 for fixed sequences in each dataset. Then, the θ that produced the best result was selected for our experiments. As described in Section 4.3, in our model the value of parameter ϕ is irrelevant. For that reason, in our experiments we set it to one.

As for the spatial interaction potential SIP, the parameter σ_t^2, which is the mean value of squared feature distances, was computed during training and the parameter p, which controls the relative influence of the data-dependent and data-independent terms (see Equation 4-6) was set to 0.5.

Regarding the temporal interaction potential, the transition matrices TIP^{tk} between epochs t and k were given by an expert who provided the information about the possible and non-possible transitions between adjacent months during an agricultural year.

Table 6: Example of a transition matrix given by an expert with information about the possible and non-possible transitions between adjacent months during and agricultural year for Campo Verde dataset.

		t+1										
	Class	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
	1	0	-∞	-∞	-∞	-∞	-∞	-∞	-∞	0	0	-∞
	2	-∞	0	-∞	-∞	-∞	-∞	-∞	-∞	-∞	0	-∞
	3	-∞	-∞	0	0	-∞	-∞	-∞	-∞	-∞	0	-∞
	4	-∞	-∞	-∞	0	-∞	-∞	-∞	-∞	-∞	0	-∞
	5	-∞	-∞	-∞	-∞	0	-∞	-∞	-∞	-∞	-∞	-∞
t	*6*	-∞	-∞	-∞	-∞	-∞	0	-∞	-∞	-∞	-∞	-∞
	7	-∞	-∞	-∞	-∞	-∞	-∞	0	-∞	-∞	-∞	-∞
	8	-∞	-∞	-∞	-∞	-∞	-∞	-∞	0	-∞	-∞	-∞
	9	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞	0	-∞	-∞
	10	0	0	0	0	-∞	-∞	-∞	-∞	-∞	0	-∞
	11	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞	0

Table 6 shows an example of a transition matrix given by an expert with information about the possible and non-possible transitions between two adjacent epochs t and $t+1$, where there are 11 classes in each epoch. 0 and

-∞ represent the possible and non-possible transitions between two classes, respectively. For instance, it is possible to have plots belonging to classes *1*, *2*, *3* and *4* in epoch t, and to class *10* in epoch $t+1$.

Experimental Protocol

Our CRF based model delivers structured predictions, i.e. the class labels of all pixels in each epoch comprised in the sequence are predicted. Our datasets, Campo Verde and LEM, comprise 19 and 51 multi-temporal images, either from SAR or Optical sensors, respectively. Thus, it is possible to draw more than 10^4 and 10^{11} sub-sequences of different lengths from Campo Verde and LEM, respectively. Instead of running experiments on all these possible sub-sequences, we decided to assess the classification performance of our method on just one image per month, assuming that all earlier images in the dataset are available as input. We also tested sequences comprising only SAR and SAR+Optical data.

The distribution of train and test sets were approximately 20%/80% for Campo Verde (see black and gray fields in Figure 18) and 75%/25% for LEM (see black and gray fields in Figure 20). All pixels in a plot were assigned either for training or testing. For Campo Verde, training and test sets were selected using stratified random sampling over the plots taking as reference the images from May due to the presence of almost all classes. Plots were selected taking care of having samples for all classes in all epochs. Almost 20% of plots were selected for training, simulating a scenario with limited labeled samples. For LEM, as it is a public benchmark, the sets are already defined and provided with the data and the reference. It reproduces a scenario with plenty of labeled samples.

5.3 Results

Experiments were carried out using sequences comprising only SAR images (single sensor classification) and both sensors, optical and SAR (multiple sensor classification) using both datasets, Campo Verde and LEM.

5.3.1 Single sensor sequences

5.3.1.1 Campo Verde

Figures 22 and 23 summarize the results obtained for Campo Verde dataset in terms of OA and average $F1$-*score* for sequences comprising only Sentinel-1A images, a total of 14 epochs (see Table 3 for acquisition dates) from October 2015 to July 2016. The horizontal axis contains the image being classified according to the protocol explained in Section 5.2. In these figures we present just one result per month, always related to the most recent epoch.

For each epoch, there are three bar groups: light, dark and hatched. Each of them corresponds to one out of three distinct ways of capturing spatial and temporal context. In the first bar group (light), the spatial context was embedded in the texture features (GLCM). For the second bar group (dark), spatial context was learned by a CNN. In the third set of bars (hatched), a CNN learned attributes that incorporate both, the spatial and temporal contexts. Within each of these groups, the first bar relates to the accuracy of the association potential (AP) only, the second bar refers to a CRF combining both, the association and the spatial interaction potentials (AP+SIP), and the third bar represents the accuracy of a CRF model comprising all three potentials (AP+SIP+TIP).

Each epoch in the plots is indicated along the horizontal axis by a number, a date and a category, which denote the number of images comprised in the sequence, the acquisition date of the image being classified, and the type of image, SAR or Optical, respectively.

In case of image 1 (October) the sequence consists of a single image (1). So, results involving temporal context are not presented. Yet, these results were consistent with the results of other epochs, as explained in the following.

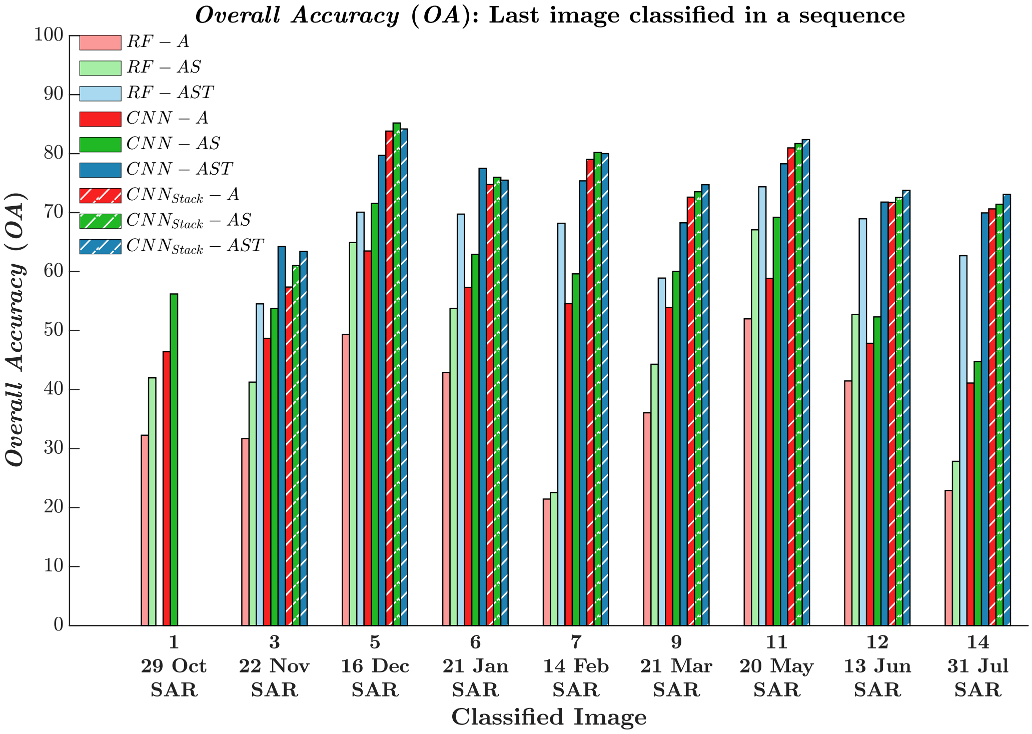


Figure 22: *Overall Accuracy* (OA) of different model variants for crop recognition in Campo Verde dataset comprising only Sentinel-1A images.

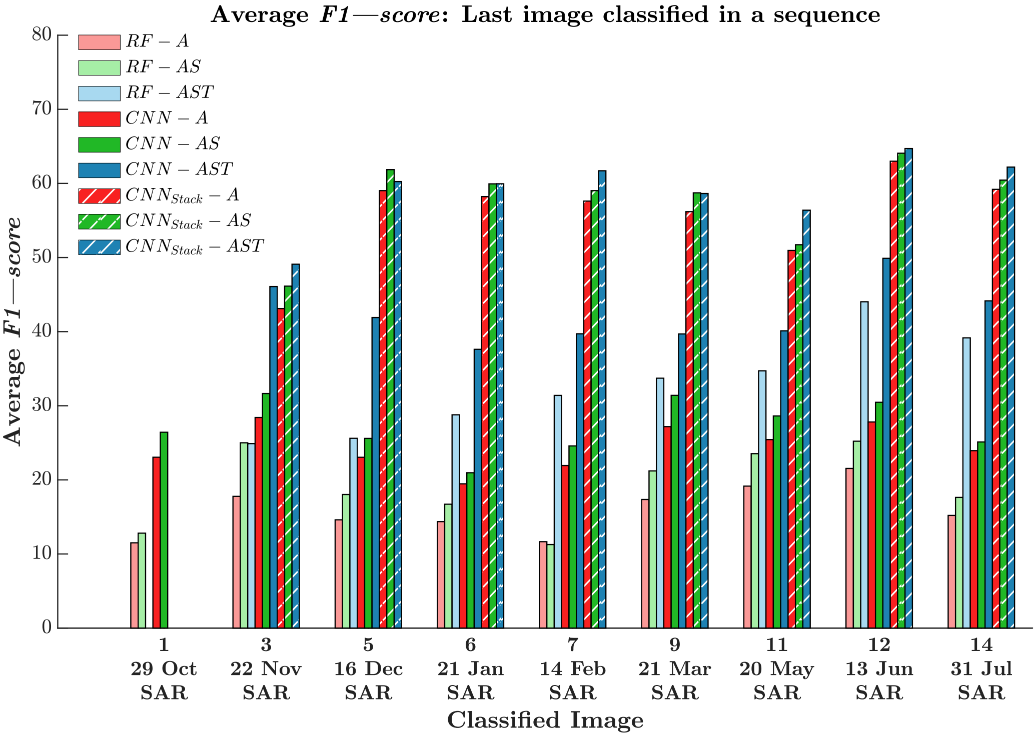


Figure 23: Average $F1$-*score* of different model variants for crop recognition in Campo Verde dataset comprising only Sentinel-1A images.

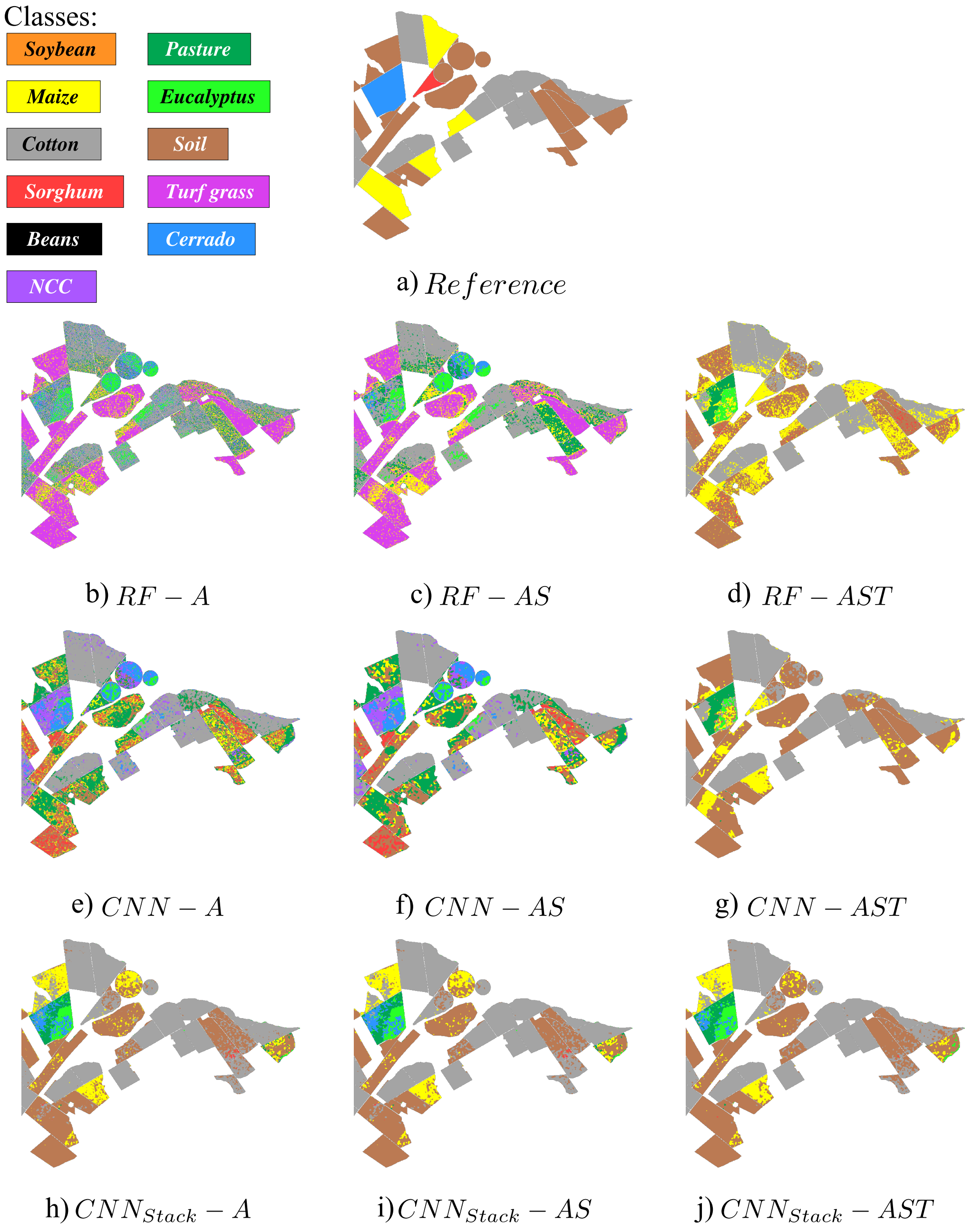
In most epochs, the three intermediate (dark) bars are higher than their correspondent in the left three (light) bars groups, revealing the importance of spatial context and the CNN ability to capture it. Similarly, the three right most (hatched) bars in each epoch are consistently higher than their counterparts in the same epoch. This shows that the consideration of temporal context, as captured by the CNN upon the stacked image, allowed for an accuracy improvement in nearly all epochs, having reached up to 29% in terms of OA and 36% in terms of $F1$-*score* in some epochs.

Comparing the greenish bars in each epoch with the corresponding reddish bars on their left, we found that the inclusion of the spatial interaction potential (SIP) in our CRF model consistently brought accuracy gains. Such improvements were more significant for the RF classifier runing on GLCM features (light bars), having declined for the CNN working on stacked features (hatched bars), being about 1% to 3.5%, in terms of OA and $F1$-*score*. It is worth noticing in the intermediate (dark) bars for all epochs, that the CRF spatial interaction potential (SIP) managed to improve the accuracy even for the CNN mono-temporal features, which already take spatial context into consideration.

Similarly, the inclusion of the temporal interaction potential (TIP) in the CRF model improved the performance for nearly all tested sequences, as can be seen by comparing the blueish with their correspondent greenish bars for all three bar groups in all epochs. Indeed, this benefit is significant for the light and dark bars and moderate to low for the hatched bars. It is particularly significant that the CRF temporal interaction potential was able to improve accuracy, both in terms of OA and $F1$-*score*, even for the CNN features learned from the image stack (hatched bars), which already capture temporal context.

Figure 24 presents snips of the predictions delivered by the different model variants for a sequence length equal to 14, classifying the 14th image in the sequence (July 31st). Each column in the figure from left to right represents CRF variants consisting of *AP*, *AP*+SIP, and *AP*+SIP+TIP, respectively. Each row from top to bottom shows the reference (Figure 24a) and variants based on handcrafted features (Figure 24b-d), CNN trained upon single images (Figure 24e-g), and CNN trained upon image stacks (Figure 24h-j), respectively.

The results in Figure 24 improve from left to right and from top to bottom. In particular, Figure 24b (*RF-A*) shows the *salt-and-pepper* effect typical of pixel-wise classification approaches, which is reduced in Figure 24c (*RF-AS*) thanks to the spatial interaction potential (SIP) of the CRF model. This effect is further reduced in Figure 24d (*RF-AST*), as the CRF model incorporates both, the spatial (SIP) and the temporal (TIP) interaction potentials.



a) $Reference$

b) $RF - A$ c) $RF - AS$ d) $RF - AST$

e) $CNN - A$ f) $CNN - AS$ g) $CNN - AST$

h) $CNN_{Stack} - A$ i) $CNN_{Stack} - AS$ j) $CNN_{Stack} - AST$

Figure 24: Snips of the predictions delivered by the different model variants for a sequence length equal to 14, classifying the 14th image (July 31st) in Campo Verde dataset with a single sensor.

Accuracy improves and the *salt-and-pepper* effect reduces for the single epoch CNN variant, as shown in Figure 24e to g. Figure 24h to j ($CNN_{Stack} - A$, $CNN_{Stack} - AS$ and $CNN_{Stack} - AST$) show further improvements due to the image stacking approach that captures the temporal information.

5.3.1.2 LEM

Figures 25 and 26 summarize the results obtained for LEM dataset in terms of OA and average $F1$-*score* for sequences comprising only Sentinel-1A images, a total of 16 epochs (see Table 4 for acquisition dates) from December 2017 to June 2018. Experiments were carried out only in this period. The data from June 2017 to June 2018 were not considered because the target area was mostly covered by bare soil, with crop residues from previous harvest or weeds, belonging to class *Uncultivated Soil* (see Figure 21)

Analogous to the previous section, each bar group (light, dark and hatched) relates to different ways of capturing spatial and/or temporal context. The variant based on GLCM texture features and the CNN monotemporal variant capture spatial context only, whereas the CNN with feature stacking exploits both spatial and temporal context.

As in the foregoing section, within each group of three bars, the first bar relates to the accuracy of the association potential (AP) only, the second bar refers to a CRF combining both, the association and the spatial interaction potentials (AP+SIP), and the third bar represents the accuracy of a CRF model comprising all three potentials (AP+SIP+TIP).

For sequences up to 11 epochs, the results for LEM database in Figures 25 and 26 presented the same behaviour observed for Campo Verde. The light bars, representing accuracies achieved from GLCM features, were lower than the dark bars, corresponding to CNN features learned from spatial context, which were in turn inferior to the accuracies attained with the use of CNN features learned from image stacking, as shown by the hatched bars.

The conclusions drawn from experiments on Campo Verde regarding the contribution of CRF spatial and temporal interaction potentials holds for the results recorded in the experiments on LEM. The CRF model comprising the spatial interaction potential combined with the association potential (AP+SIP) (greenish bars) managed to improve accuracy in relation to the a model based on the association potential alone (AP) (blueish bars). Such benefit was significant even in the variants that used CNN learned features that rely on spatial context, as shown by the dark and hatched bars, respectively.

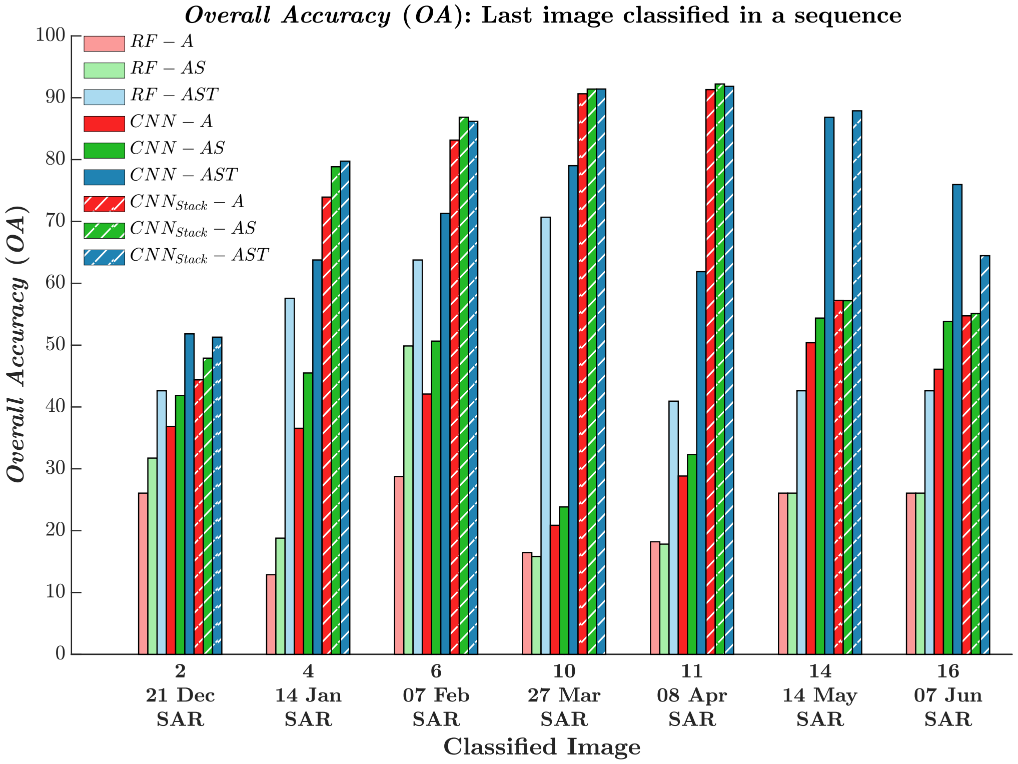


Figure 25: *Overall Accuracy* (OA) of different model variants for crop recognition in LEM dataset comprising only Sentinel-1A images.

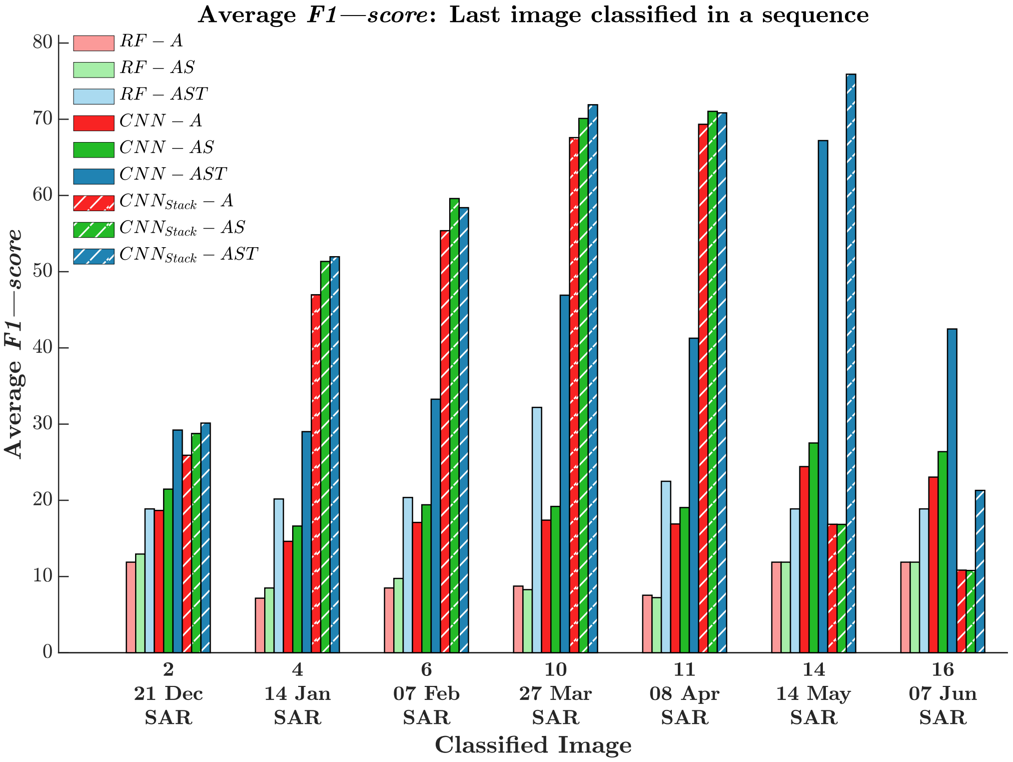


Figure 26: Average $F1$-*score* of different model variants for crop recognition in LEM dataset comprising only Sentinel-1A images.

The contribution of the temporal interaction potential TIP ranged from 9% to 55% in terms of OA and from 6% to 40% in terms of $F1$-*score* for the experiments on GLCM (light bars) and on spatial context aware features learned by CNN (dark bars).

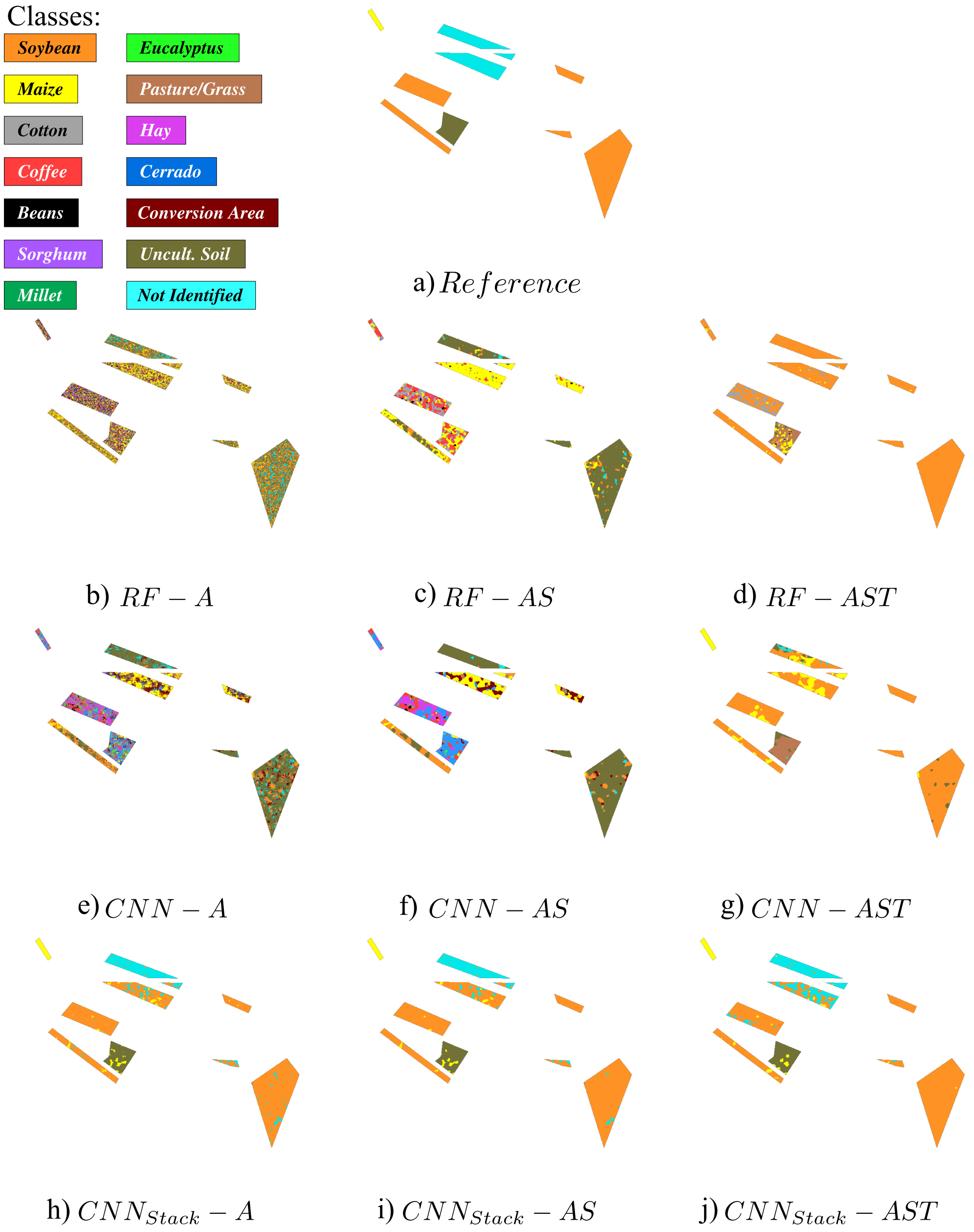
Such impact on accuracy for features learned by a CNN upon image stack (hatched bars) was mostly low, even slightly detrimental, as in epochs 6 and 11. A similar marginally negative effect was also observed for sequence 5 and 6 in Campo Verde.

The results for epoch 14 and, specially for epoch 16 diverged from the pattern observed thus far. Note that the $F1$-*score* for the CNNs on the feature stack (dashed bars) were mostly lower than the corresponding results for spatial only CNNs (dark bars). Besides, the full CRF model comprising AP, SIP and TIP performed for the stacked image variant much better than a similar model without TIP, actually, also in terms of OA. The reason for such unexpected results can be explained as follows.

CNNs are known to be highly demanding in terms of labeled training samples. As the ratio between the number of parameters to the number of labeled samples increases, the CNN tends to generalize poorly. Longer sequences imply in deeper image stacks, and consequently in deeper kernels at the first CNN layer. Therefore, more parameters have to be learned from the available labeled samples. Not surprisingly, we noticed in our experiments that the CNN with feature stacking variant (hatched bars) started declining for image sequences longer than 14. It is significant that the spatio-temporal CNN (hatched bars) performed poorer than the single epoch CNN (dark bars) in terms of $F1$-*score* for epoch 16. In this experiment the underlying CNN architecture was not able to generalize and did not capture the temporal context properly. Equally remarkable was the accuracy gain brought by the full CRF model (hatched blueish bar) in relation to the simpler CRF model comprising only the association and spatial interaction potentials (AP+SIP) for sequences 14 and 16.

Figure 27 shows snips of the predictions delivered by the model variants for a sequence length equal to 10, whereby the results refer to the last image in the sequence, the 10th one (March 27th).

From left to right, each column represents CRF variants considering only AP, AP+SIP, and AP+SIP+TIP, respectively. From top to bottom, each row shows the reference (Figure 27a) and the results of variants based on texture features (Figure 27b-d), CNN trained upon single images (Figure 27e-g), and CNN trained upon image stacks (Figure 27h-j).



a) $Reference$

b) $RF - A$ c) $RF - AS$ d) $RF - AST$

e) $CNN - A$ f) $CNN - AS$ g) $CNN - AST$

h) $CNN_{Stack} - A$ i) $CNN_{Stack} - AS$ j) $CNN_{Stack} - AST$

Figure 27: Snips of the predictions delivered by the different model variants for a sequence length equal to 10, classifying the 10th image (March 27th) in LEM dataset with a single sensor.

Moving from left to right and from top to bottom, the results improve, getting more similar to the reference and smoother as more context information is being exploited. The *salt-and-pepper* effect, common in pixel-wise approaches, is apparent in Figure 27b (RF-A), but diminished with the addition of the spatial interaction potential SIP in the CRF model, as shown in Figure 27c (RF-AS). It was further attenuated when the CRF model incorporated the temporal interaction potential (TIP) (see Figure 27d).

The results of the CNN monotemporal variants shown in the third row

(Figure 27e-g) are clearly better than the results shown in the second row produced by variants based on GLCM features. This demonstrates the benefits of using CNN to capture spatial context from single images.

Even further improvements were obtained by the image stack variants, as shown in Figure 27h-j (CNN_{Stack}-A, CNN_{Stack}-AS and CNN_{Stack}-AST), In these variants the spatial and temporal context were captured by a CNN trained upon a multi-temporal image stack.

5.3.2 Multi Sensor sequences

5.3.2.1 Campo Verde

Figures 28 and 29 summarize the results obtained for Campo Verde dataset in terms of OA and average $F1$-$score$ for sequences comprising both Sentinel-1A and Landsat 8 images, a total of 19 epochs (see Table 3 for acquisition dates) from October 2015 to July 2016. The horizontal axis indicates the image being classified according to the experimental protocol described in Section 5.2.

The Campo Verde dataset contains two main periods defined by the cycle of the most abundant crops: from October 2015 to February 2106, being *Soybean* the most abundant crop, and from March 2016 to July 2016, with *Maize* and *Cotton* as the major crops.

We report in the following just results relative to the period from March 2016 to July 2016. Only one result per month is presented, always related to the most recent epoch. This period was selected due to its higher dynamics (presence of more crops and crop transitions) and to the availability of images from sensors of different spatial resolutions (see Table 3 for acquisition dates from each sensor).

The results obtained on data from a single sensor (Section 5.3.1) clearly demonstrated the superiority of CNN based variants combined with feature stacking. For this reason, in our experiments on multi sensor data we did not consider variants based on handcrafted features.

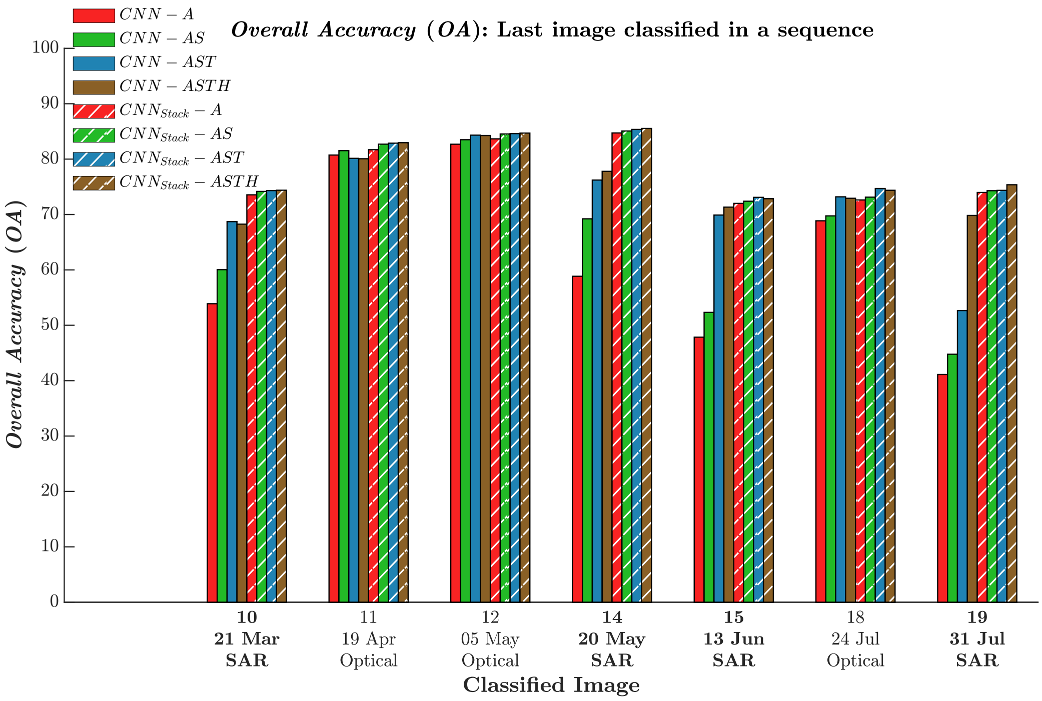


Figure 28: *Overall Accuracy* (OA) obtained by different approaches for crop recognition in Campo Verde dataset using sequences of images from multiple sensors.

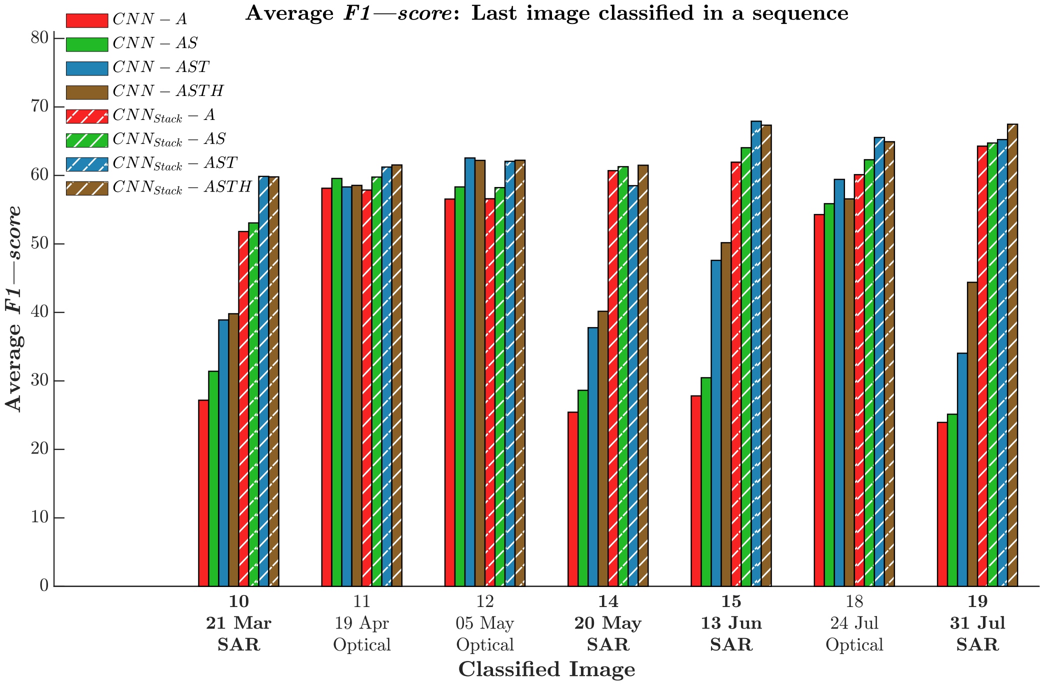


Figure 29: Average $F1$-*score* obtained by different approaches for crop recognition in Campo Verde dataset using sequences of images from multiple sensors.

Figure 28 and 29 present the results of our experiments on multi sensor data. For each epoch, there is a bar group which can be separated in two groups: dark and hatched. Each of these groups is related to a different way of capturing spatial and/or temporal context: by a CNN trained per epoch, or by a CNN trained upon an image stack, respectively.

Each group of bars comprises four bars. The first bar relates to the accuracy of the association potential (AP) only, the second bar refers to a CRF combining both, the association and the spatial interaction potentials (AP+SIP), the third bar represents the accuracy of a CRF model comprising all three potentials (AP+SIP+TIP), and the fourth bar portrays a CRF model using higher order interactions in the temporal domain, as presented in Section 4.3.2.

Recall that the dates the results refer to are given along the horizontal axis by a number, a date and a category, which stand for the sequence length, the acquisition date of the image being classified, and the image domain (SAR or Optical), respectively.

Notice that for almost all epochs, the four right most (hatched) bars are higher than their correspondent four left most (dark) bars, supporting the conclusion drawn from single sensor experiments (Section 5.3.1) about the improvements obtained by capturing spatial and temporal context using a CNN trained upon stacked images. These improvements reached up to 33% in terms of OA and 40% in terms of $F1$-*score* in some epochs.

As expected, the variants that do not consider temporal interaction (reddish and greenish bars) achieved higher accuracies for optical data (epochs 11, 12 and 18) than for SAR data (epochs 10, 14, 15 and 19), because spectral features are more discriminative than back-scatter intensities for mono-temporal crop classification.

Concerning the contributions of CRF spatial and temporal interaction potentials, the results here are consistent with the conclusions drawn the previous section. CRF models comprising the association and the spatial interaction potentials (AP+SIP) (greenish bars) achieved better accuracies than models based on the association potential alone (AP) (reddish bars). The accuracy gains amounted up to 10% in terms of OA and 4% in terms of $F1$-*score*. Moreover, with the addition of the temporal interaction potential (AP+SIP+TIP) to the CRF model (blueish bars), improvements of up to 17% in terms of OA and $F1$-*score* were obtained.

The inclusion of higher order interactions in the temporal domain (brownish bars) brought slight improvements in many epochs, ranging from 0.1% to 1% in terms of OA and from 0.2% to 10% in terms of $F1$-*score*, being more prominent for epochs 14 and 19. Notice that such epochs are characterised by SAR data after the use of optical data, exactly when the higher order interactions were expected to be more effective. Yet in these epochs the improvements in terms of OA were low. This is probably because the high-order temporal edges improve results mostly around plot borders, which correspond in our datasets to a small proportion of whole the test area.

Nevertheless, the high order temporal interactions caused an accuracy drop in epoch 18 of up to 3% in terms of $F1$-*score* and about 0.3% in terms of OA. After analysing the generated crop maps we concluded that the observed $F1$-*score* reduction was due to miss-classifications of *Turf grass* samples, which corresponds to less than 0.02% of the test area (see Figure 19). Therefore, we consider that this result is not representative of the performance associated to temporal higher order CRFs.

Figure 30 shows snips of the predictions delivered by the different model variants for a sequence length equal to 14, classifying the 14th image (May 20th).

The first column of pictures represents variants based on CNN, while the second column relates to those based on feature stack. From top to bottom, the row show first the reference, then the results of CRF variants that consider only the AP (Figure 30b-c), the AP+SIP (Figure 30d-e), the AP+SIP+TIP (Figure 30f-g), and AP+SIP+TIP+Higher order (Figure 30h-i), respectively.

As we move from left to right and from top to bottom in Figure 30, the predictions become closer to the reference getting a smoother appearance due to the consideration of more contextual information. Even capturing spatial context using a CNN per epoch in Figure 30b (CNN-A), the *salt-and-pepper* effect is apparent, which diminished by the successive inclusion of more spatial and/or temporal context by means of the SIP in Figure 30d (CNN-AS), the TIP in Figure 30f (CNN-AST), and the higher order interactions in Figure 30h (CNN-$ASTH$).

Feature stacking based variants, CNN_{Stack}-A in Figure 30c, CNN_{Stack}-AS in Figure 30e, CNN_{Stack}-AST in Figure 30g, and CNN_{Stack}-$ASTH$ in Figure 30i, improved their counterparts based on single epoch CNN, stressing the capacity of CNN to capture not only spatial context, but temporal context.

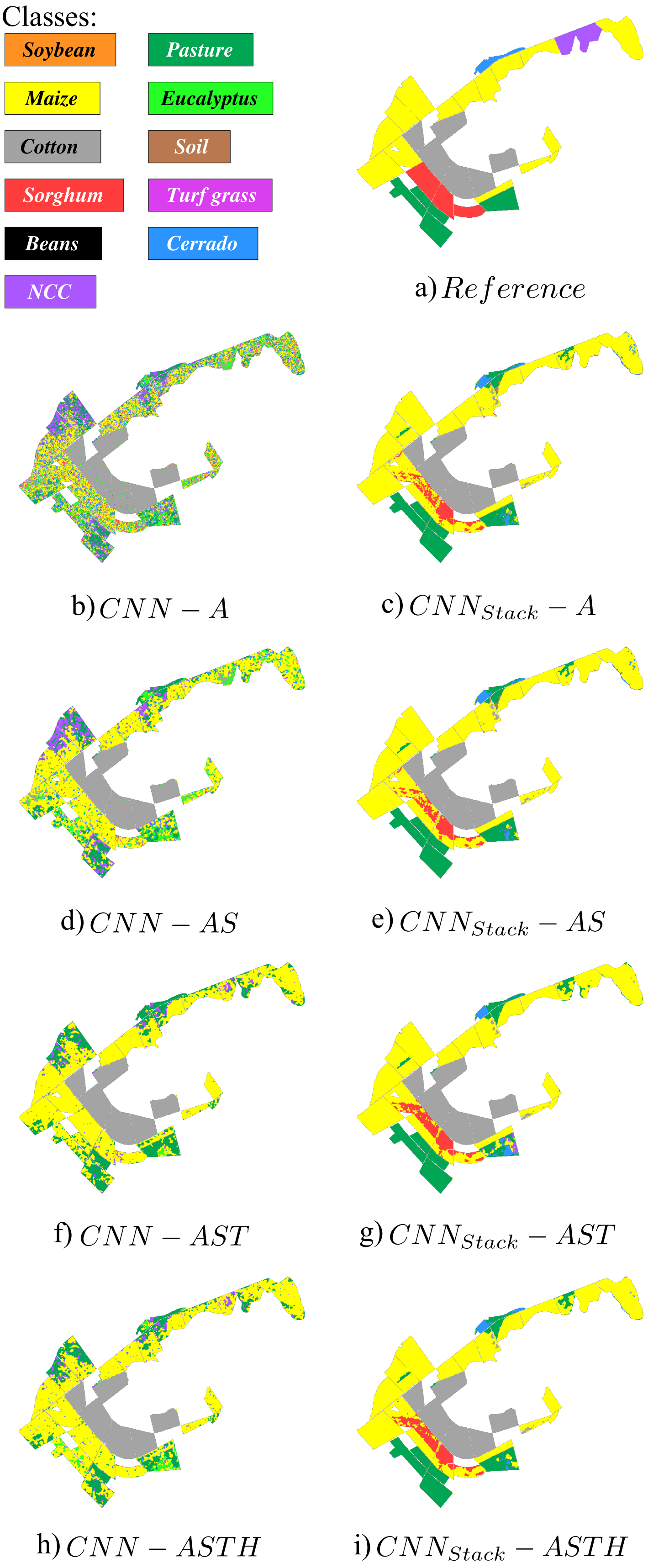
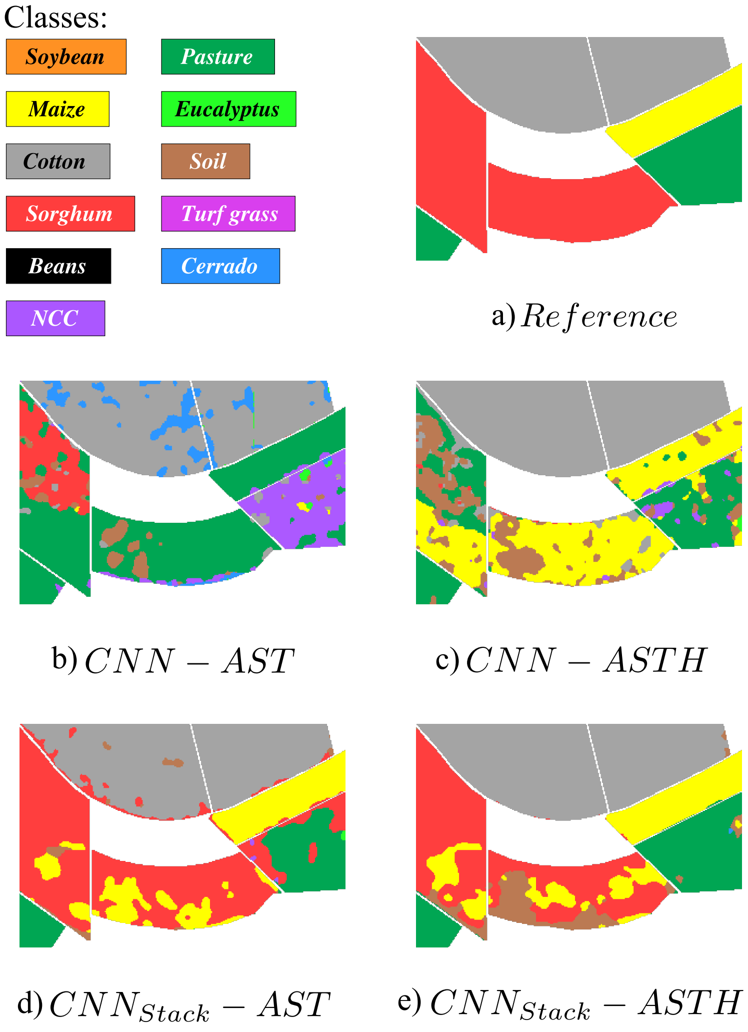


Figure 30: Snips of the predictions delivered by the different model variants for a sequence length equal to 14, classifying the 14th image (May 20th) in Campo Verde dataset with multiple sensors.

Figure 31 presents snips of the predictions obtained by four CRF variants for a sequence length equal to 19, classifying the 19th image (July 31st).

From left to right, each column represents CRF variants considering the three potentials $AP{+}SIP{+}TIP$, and those including higher order interactions. From top to bottom, each row relates to the reference (Figure 31a) and variants based on CNN (Figure 31b-c), and feature stacking (Figure 31d-e), respectively.

Higher order CRF variants, CNN-$ASTH$ (Figure 31c) and CNN_{Stack}-$ASTH$ (Figure 31e), brought improvements in relation to their counterparts that involve $AP{+}SIP{+}TIP$, CNN-AST (Figure 31b) and CNN_{Stack}-AST (Figure 31d), by reducing miss-classified pixels (e.g. as in *Cotton* plots) and correcting class transitions (e.g. as in *Maize* and *Pasture* plots).



a) $Reference$

b) $CNN - AST$

c) $CNN - ASTH$

d) $CNN_{Stack} - AST$

e) $CNN_{Stack} - ASTH$

Figure 31: Snips of the predictions delivered by four CRF variants for a sequence length equal to 19, classifying the 19th image (July 31st)

5.3.2.2 LEM

Figures 32 and 33 summarize the results obtained for LEM dataset in terms of OA and $F1$-*score* for sequences comprising Sentinel-1A, Sentinel-2A/-2B and Landsat 8 images, a total of 26 images (see Table 4 for acquisition dates) from December 2017 to June 2018. As in preceding sections, the image being classified is indicated in the horizontal axis according to the protocol explained in Section 5.2. In these figures we present just one result per month, always associated to the most recent epoch.

Similar to the previous section, the bar groups of each epoch consists of two sub-groups: dark and hatched. Each of them refers to different alternatives to introduce spatial and/or temporal context information into the model: by a CNN trained per epoch (exploitation of local spatial context), or by a CNN trained upon an image stack (taking advantage of local spatial and global temporal context), respectively.

Each group of bars comprises up to four bars. The first bar refers to the accuracy of the association potential (AP) alone, the second bar relates to a CRF considering the association and the spatial interaction potentials (AP+SIP), the third bar represents the accuracy of a CRF model composed of the three potentials (AP+SIP+TIP), and the fourth bar portrays a CRF model including higher order interactions in the temporal domain.

As the first 12 epochs involve a single sensor (see Table 4), the first five groups of bars are identical to those presented in Section 5.3.1.2. The reader is referred to that section for a detailed description and explanation of those results. Thus, only results related to epochs 14 to 21 will be analyzed in the following paragraphs.

Regarding the influence of CRF spatial and temporal potentials as well as the higher order interactions, the results are consistent with the conclusions drawn in the preceding sections. CRF variants comprising the association and spatial interaction potentials (AP+SIP) (greenish bars) managed to improved their counterparts based on the association potential alone (AP) (reddish bars) in up to 7% in terms of OA and 4% in terms of $F1$-*score*. Moreover, the introduction of temporal context into the CRF model through the temporal interaction potential (TIP) (blueish bars) brought accuracy gains of up to 28% in terms of OA and 49% in terms of $F1$-*score*. Finally, additional improvements were obtained by the addition of higher order interactions in the temporal domain with accuracy gains of up to 2% in terms of OA and $F1$-*score*.

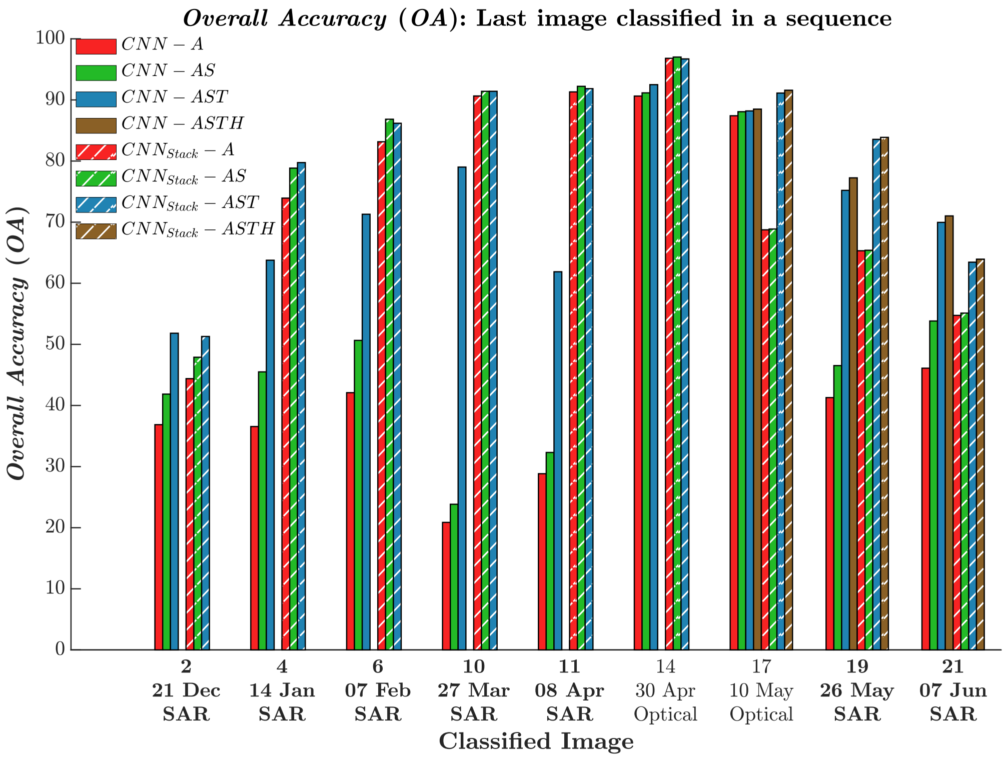


Figure 32: *Overall Accuracy* (*OA*) obtained by different approaches for crop recognition in LEM dataset using sequences of images from multiple sensors.

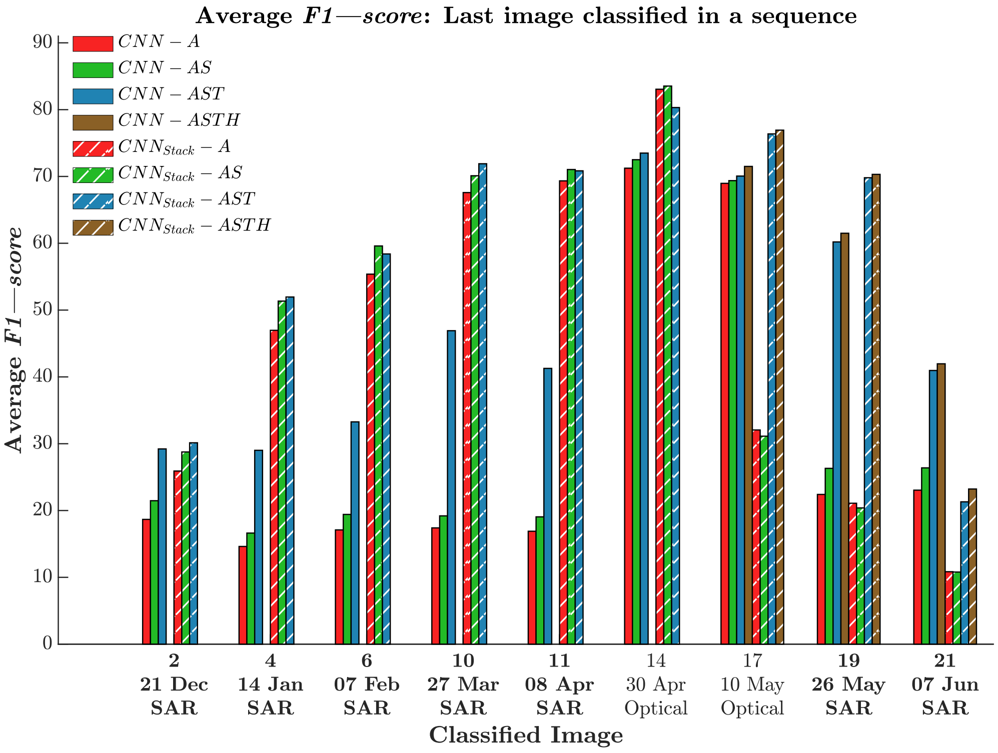


Figure 33: Average $F1$-*score* obtained by different approaches for crop recognition in LEM dataset using sequences of images from multiple sensors.

Variants based on the association potential alone (AP) (reddish) and those comprising both the association and the spatial interaction potentials (AP+SIP) (greenish) achieved higher accuracies classifying Optical epochs (14 and 17) than SAR epochs (19 and 21), strengthening what was concluded from Campo Verde experiments for multiple sensors (Section 5.3.2.1).

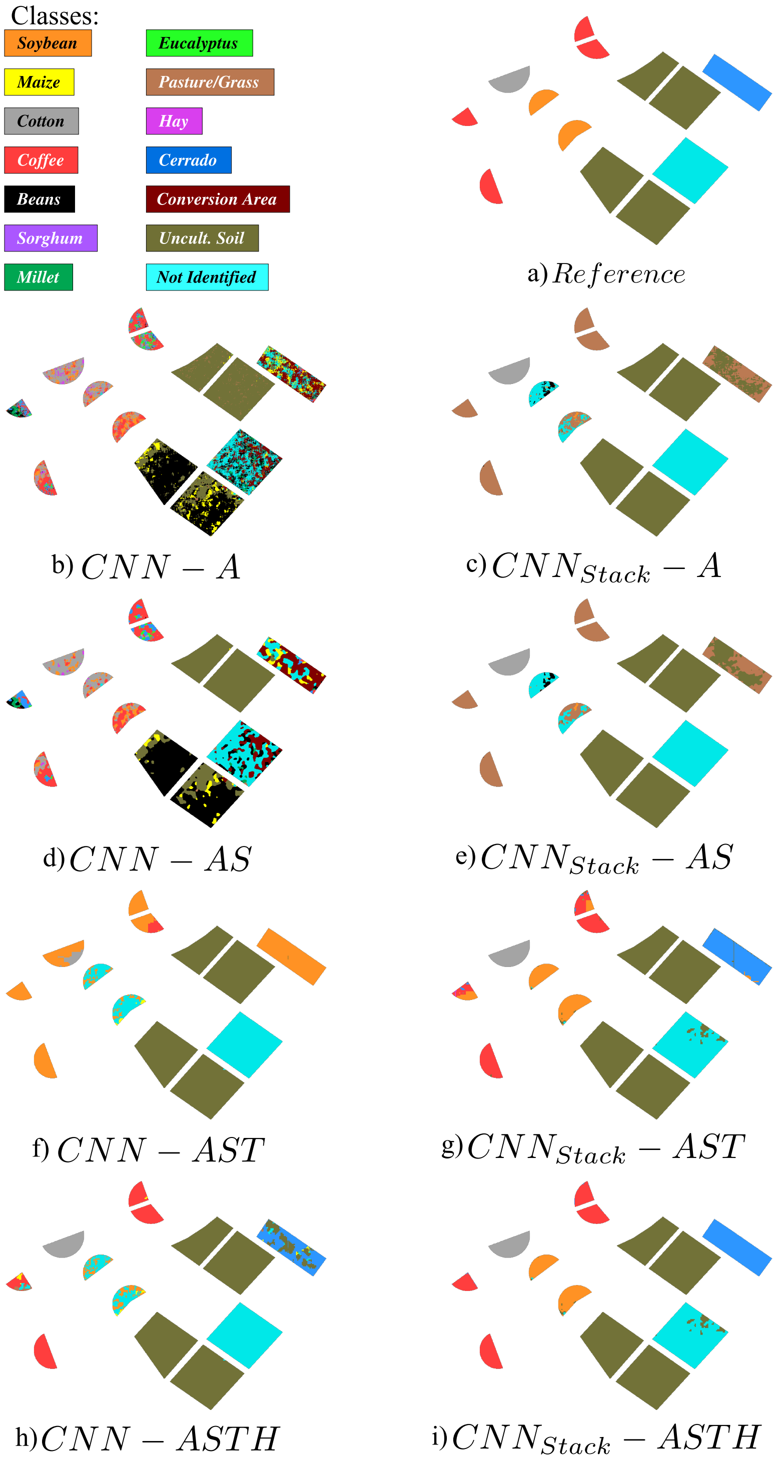
In epoch 14, the CNN stack variants (hatched bars) improved the results of their monotemporal counterparts (dark bars) in up to 6% in terms of OA and 12% in terms of $F1$-*score*. Among the stack variants, the one comprising all three potentials, CNN_{Stack}-AST (blueish hatched bar), obtained lower accuracy than a CRF model considering only the association and spatial interaction potentials, CNN_{Stack}-AS (greenish hatched bar). The accuracy dropped 0.3% in terms of OA and 3.2% in terms of $F1$-*score*. This indicates that the addition of TIP for epoch 14 affected predominately the classification of less abundant classes.

The pattern observed in previous epochs was not preserved in epochs 17 and 19, as observed in the results for LEM using a single sensor (Section 5.3.1.2). In terms of $F1$-*score*, comparing correspondent variants based on feature stack (hatched bars) and spatial CNN (dark bars), the former achieved greater accuracies only for CRF models comprising all three potentials (blueish) and higher order CRF (brownish). However, for those considering only the association potential (AP) (reddish bars) or both the association and the spatial interaction potentials (AP+SIP) (greenish bars), dark bars were higher. This effect is even more pronounced for epoch 21, when the CNN variants trained upon image stack (hatched bars) got worse results than those based on spatial only CNN (dark bars).

The explanation for this behaviour is similar what was stated in Section 5.3.1.2 for single sensor classification on the same months, May and June. For models comprising only the AP (reddish bars), or the AP+SIP (greenish bars), a CNN trained upon image stack (hatched bars) was not able to generalize properly due to the increasing number of parameters to be learned. Recalling that longer sequences imply deeper image stacks and equally deeper kernels in the first CNN layer.

Notice that it takes place during the months where sudden changes in class distribution arose, from April to June. Thus, as many crops cycles are finishing, not relevant information is introduced into the image stack to classify epochs on that period.

Figure 34 presents snips of predictions delivered by the different model variants for a sequence length equal to 19, classifying the 19th image (May 26th).



a) $Reference$

b) $CNN - A$

c) $CNN_{Stack} - A$

d) $CNN - AS$

e) $CNN_{Stack} - AS$

f) $CNN - AST$

g) $CNN_{Stack} - AST$

h) $CNN - ASTH$

i) $CNN_{Stack} - ASTH$

Figure 34: Snips of the predictions delivered by the different model variants for a sequence length equal to 19, classifying the 19th image (May 26th) in LEM dataset with multiple sensors.

The first column corresponds to variants based on spatial only CNN while the second one represents those based on CNN trained upon feature stacks. From top to bottom, each row relates to the reference (Figure 34a) and CRF variants comprising only the AP (Figure 34b-c), the AP+SIP (Figure

34d-e), the AP+SIP+TIP (Figure 34f-g), and higher order CRF (Figure 34h-i), respectively.

From left to right and from top to bottom, as more context information is being considered, the predictions get closer to the reference. For monotemporal CNN based variants, the inclusion of the SIP in the CRF model (Figure 34d) (CNN-*AS*) improved the results in relation to what had been obtained by the association potential alone (Figure 34b) (CNN-*A*). The further inclusion of the TIP (Figure 34f) (CNN-*AST*), reduced the *salt-and-pepper* effect, providing smoother predictions. In contrast, all image stack variants got smoother results even with CRF considering only the AP.

It is interesting to compare the results of single image CNN variants with and without the temporal interaction potential (CNN-*AS* vs. CNN-*AST*). Some plots miss-classified by the CNN-*AS* variant (Figure 34d) were fixed by the inclusion of the temporal interaction potential (see Figure 34f). The results got even better for the high order variant (CNN-*ASTH* in Figure 34h). A similar trend is observed by comparing the results in Figure 34e, g and i.

6 CONCLUSIONS

A novel method for crop recognition in tropical regions from sequences of multi sensor remote sensing images based on Conditional Random Fields (CRF) was proposed in this work. The method exploits contextual information in the spatial and temporal domains for a proper characterization of crops development. Furthermore, it is capable to work with sequences of remote sensing images from different domains/sensors with different spatial resolutions.

Experiments to validate the proposed method were carried out over public datasets of two municipalities in Brazil: Campo Verde, Mato Grosso state and Luis Eduardo Magalhães (LEM), Bahia. Those two datasets are a contribution of the present thesis. They were built in cooperation with partners from National Institute for Space Research (INPE) and from Brazilian Agricultural Research Corporation (EMBRAPA).

The proposed CRF model relies on three main terms called: the association potential, the spatial interaction potential and the temporal interaction potential. Variants of the proposed model were created, by considering only one, two or all three potentials, to assess the influence of each of them over model accuracy.

Three different designs for the association potential were tested: a Random Forest (RF) trained upon handcrafted features, a Convolutional Neural Network (CNN) trained upon a single image, and a CNN trained upon an image stack. For the spatial interaction potential, a contrast-sensitive Potts model was employed, which penalizes class changes unless a significant data variation between neighboring image sites occurs. Prior knowledge about possible and impossible temporal class transitions in adjacent epochs were used to model temporal interaction potential.

The accuracies obtained by the proposed CRF model in our experiments were up to 85% in terms of *Overall Accuracy* (OA) and 68% in terms of $F1$-*score* for Campo Verde, and up to 92% in terms of OA and 83% in terms of $F1$-*score* for LEM, demonstrating the capacity of the proposed method to recognize crops in tropical regions with complex dynamics and different agricultural practices like the selected study areas.

About using CNN to capture contextual information

For Campo Verde dataset, variants based on CNN trained upon single images outperformed their counterparts based on RF trained upon handcrafted features in up to 27% in terms of OA and 39% in terms of $F1$-*score*. Moreover, CNN trained upon image stacks managed to further improve those results in up to 29% in terms of OA and 31% in terms of $F1$-*score*. The same pattern was observed for LEM dataset with improvements of up to 70% in terms of OA and 54% in terms of $F1$-*score* obtained with a CNN trained upon image stacks rather than upon single images to provide the association potential.

About the influence of each CRF potential

The addition of the spatial interaction potential to the CRF model led to accuracy improvements of up to 16% in terms of OA and 4% in terms of $F1$-*score* for Campo Verde, and of up to 21% in terms of OA and 5% in terms of $F1$-*score* for LEM. Further accuracy gains were reached with the consideration of the temporal interaction potential, specifically, up to 25% in terms of OA and 19% in terms of $F1$-*score* for Campo Verde, and 56% in terms of OA and 40% in terms of $F1$-*score* for LEM. In this manner, we demonstrated the benefits of each potential, being the temporal interaction potential the one that brought most accuracy improvements. The importance of prior knowledge about class transitions between adjacent epochs was clearly demonstrated by the aforesaid results. Moreover, as the CRF model took more context information into account, the classification predictions got smoother, and the *salt-and-pepper* effect diminished.

About using higher order connections

We proposed the addition of higher order connections in the temporal domain for datasets with images from different domains/sensors and spatial resolutions, such as Sentinel-1A (SAR), Sentinel-2A/2B, and Landsat 8 (optical) images. The higher order CRF models obtained significantly better results than models using only the three potentials, with accuracy gains of up to 17% in terms of OA and 10% in terms of $F1$-*score* for Campo Verde, and 2% in terms of OA and $F1$-*score* for LEM. In addition, the high order edges were able to correct plots miss-classified by the other tested variants.

Future directions

The proposed method will be evaluated for crop recognition in temperate regions, which possesses different crops dynamic than tropical regions.

Additionally, different alternatives to exploit spatial and temporal context will be explored. For instance, other deep architectures to provide the association potential such as Fully Convolutional Networks (FCN) ans Recurrent Neural Networks (RNN), to capture spatial and temporal context, respectively..

References

1 THENKABAIL, P. S.. **Land resources monitoring, modeling, and mapping with remote sensing**. CRC Press, 2015.

2 MUELLER, N. D.; GERBER, J. S.; JOHNSTON, M.; RAY, D. K.; RAMANKUTTY, N. ; FOLEY, J. A.. **Closing yield gaps through nutrient and water management**. Nature, 490(7419):254, 2012.

3 MATESE, A.; TOSCANO, P.; DI GENNARO, S. F.; GENESIO, L.; VACCARI, F. P.; PRIMICERIO, J.; BELLI, C.; ZALDEI, A.; BIANCONI, R. ; GIOLI, B.. **Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture**. Remote Sensing, 7(3):2971–2990, 2015.

4 DIRZO, R.; RAVEN, P. H.. **Global state of biodiversity and loss**. Annual Review of Environment and Resources, 28(1):137–167, 2003.

5 EDENHOFER, O.; PICHS-MADRUGA, R.; SOKONA, Y.; AGRAWALA, S.; BASHMAKOV, I.; BLANCO, G.; BROOME, J.; BRUCKNER, T.; BRUNNER, S.; BUSTAMANTE, M. ; OTHERS. **Summary for policymakers**, 2014.

6 FOOD; OF THE UNITED NATIONS (FAO) AQUASAT, A. O.. **Water uses - water withdrawal ratios by continent**, 2018.

7 WORLD WATER ASSESSMENT PROGRAMME (WWAP), U.. **United Nations world water development report 4: managing water under uncertainty and risk**. Paris:UNESCO, France, 3 edition, 2012.

8 TILMAN, D.; BALZER, C.; HILL, J. ; BEFORT, B. L.. **Global food demand and the sustainable intensification of agriculture**. Proceedings of the National Academy of Sciences, 108(50):20260–20264, 2011.

9 MACDONALD, R. B.; HALL, F. G.. **Global crop forecasting**. Science, 208(4445):670–679, 1980.

10 PICOLI, M. C. A.; CAMARA, G.; SANCHES, I.; SIMÕES, R.; CARVALHO, A.; MACIEL, A.; COUTINHO, A.; ESQUERDO, J.; ANTUNES, J.; BEGOTTI, R. A. ; OTHERS. **Big earth observation time series analysis for**

monitoring brazilian agriculture. ISPRS Journal of Photogrammetry and Remote Sensing, 145:328–339, 2018.

11 SANCHES, I. D.; FEITOSA, R. Q.; DIAZ, P. M. A.; SOARES, M. D.; LUIZ, A. J. B.; SCHULTZ, B. ; MAURANO, L. E. P.. **Campo Verde database: Seeking to improve agricultural remote sensing of tropical areas**. IEEE Geoscience and Remote Sensing Letters, 15(3):369–373, March 2018.

12 SANCHES, I. D.; FEITOSA, R. Q.; ACHANCCARAY, P.; MONTIBELLER, B.; LUIZ, A. J. B.; SOARES, M. D.; PRUDENTE, V. H. R.; VIEIRA, D. C. ; MAURANO, L. E. P.. **Lem benchmark database for tropical agricultural remote sensing application**. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-1:387–392, 2018.

13 SPACE AGENCY ESA, E.. **The sen2agri system: Sentinel-2 for agriculture**, 2019.

14 SPACE AGENCY ESA, E.. **Sentinels synergy for agriculture - sensagri**, 2019.

15 DEMPSTER, A. P.. **Upper and lower probabilities induced by a multivalued mapping**. In: CLASSIC WORKS OF THE DEMPSTER-SHAFER THEORY OF BELIEF FUNCTIONS, p. 57–72. Springer, 2008.

16 SHAFER, G.. **A mathematical theory of evidence**, volumen 42. Princeton university press, 1976.

17 FORKUOR, G.; CONRAD, C.; THIEL, M.; ULLMANN, T. ; ZOUNGRANA, E.. **Integration of optical and synthetic aperture radar imagery for improving crop mapping in northwestern Benin, West Africa**. Remote Sensing, 6(7):6472–6499, 2014.

18 LIU, M. W.; OZDOGAN, M. ; ZHU, X.. **Crop type classification by simultaneous use of satellite images of different resolutions**. IEEE Transactions on Geoscience and Remote Sensing, 52(6):3637–3649, 2014.

19 TATSUMI, K.; YAMASHIKI, Y.; MORANTE, A. K. M.; FERNÁNDEZ, L. R. ; NALVARTE, R. A.. **Pixel-based crop classification in peru from Landsat 7 ETM+ images using a random forest model**. Journal of Agricultural Meteorology, 72(1):1–11, 2016.

20 ÖZÜM DURGUN, Y.; GOBIN, A.; KERCHOVE, R. V. D. ; TYCHON, B.. **Crop area mapping using 100-m Proba-V time series**. Remote Sensing, 8(7):1–23, 2016.

21 BARGIEL, D.. **A new method for crop classification combining time series of radar images and crop phenology information**. Remote Sensing of Environment, 198:369 – 383, 2017.

22 HEUPEL, K.; SPENGLER, D. ; ITZEROTT, S.. **A progressive crop-type classification using multitemporal remote sensing data and phenological information**. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 86(2):53–69, Apr 2018.

23 BELGIU, M.; CSILLIK, O.. **Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis**. Remote sensing of environment, 204:509–523, January 2018.

24 CLERICI, N.; CALDERÓN, C. A. V. ; POSADA, J. M.. **Fusion of Sentinel-1A and Sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, colombia**. Journal of Maps, 13(2):718–726, 2017.

25 SCHULTZ, B.; IMMITZER, M.; FORMAGGIO, A. R.; SANCHES, I. D. A.; LUIZ, A. J. B. ; ATZBERGER, C.. **Self-guided segmentation and classification of multi-temporal Landsat 8 images for crop type mapping in southeastern Brazil**. Remote Sensing, 7(11):14482–14508, 2015.

26 LEITE, P. B. C.; FEITOSA, R. Q.; FORMAGGIO, A. R.; DA COSTA, G. A. O. P.; PAKZAD, K. ; SANCHES, I. D.. **Hidden markov models for crop recognition in remote sensing image sequences**. Pattern Recognition Letters, 32(1):19–26, Jan. 2011.

27 SIACHALOU, S.; MALLINIS, G. ; TSAKIRI-STRATI, M.. **A hidden markov models approach for crop classification: Linking crop phenology to time series of multi-sensor remote sensing data**. Remote Sensing, 7(4):3633–3650, 2015.

28 LIU, D.; SONG, K.; TOWNSHEND, J. R. ; GONG, P.. **Using local transition probability models in markov random fields for forest change detection**. Remote Sensing of Environment, 112(5):2222 – 2231, 2008. Earth Observations for Terrestrial Biodiversity and Ecosystems Special Issue.

29 HAGENSIEKER, R.; ROSCHER, R.; ROSENTRETER, J.; JAKIMOW, B. ; WASKE, B.. **Tropical land use land cover mapping in Pará (Brazil) using Discriminative Markov Random Fields and multi-temporal**

TerraSAR-X data. International Journal of Applied Earth Observation and Geoinformation, 63:244–256, 2017.

30 HOBERG, T.; MÜLLER, S.. **Multitemporal crop type classification using conditional random fields and rapideye data**. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVIII-4/W19:115–121, 2011.

31 HOBERG, T.; ROTTENSTEINER, F.; FEITOSA, R. Q. ; HEIPKE, C.. **Conditional random fields for multitemporal and multiscale classification of optical satellite imagery**. IEEE Transactions on Geoscience and Remote Sensing, 53(2):659–673, 2015.

32 DIAZ, P. M. A.; FEITOSA, R. Q.; SANCHES, I. D. ; COSTA, G. A. O. P.. **A Method to Estimate Temporal Interaction in a Conditional Random Field Based Approach for Crop Recognition**. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLI-B7:205–211, 2016.

33 ACHANCCARAY, P.; FEITOSA, R. Q.; ROTTENSTEINER, F.; SANCHES, I. D. ; HEIPKE, C.. **Spatio-temporal conditional random fields for recognition of sub-tropical crop types from multi-temporal images**. In: ANAIS DO XVIII SIMPóSIO BRASILEIRO DE SENSORIAMENTO REMOTO. SãO JOSé DOS CAMPOS, p. 2539–2546. INPE, 2017.

34 ACHANCCARAY, P.; FEITOSA, R. Q.; ROTTENSTEINER, F.; SANCHES, I. D. ; HEIPKE, C.. **Spatial-temporal conditional random field based model for crop recognition in tropical regions**. In: 2017 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), p. 3007–3010, July 2017.

35 KENDUIYWO, B. K.; BARGIEL, D. ; SOERGEL, U.. **Higher Order Dynamic Conditional Random Fields ensemble for crop type classification in radar images**. IEEE Transactions on Geoscience and Remote Sensing, 2017.

36 KUSSUL, N.; LAVRENIUK, M.; SKAKUN, S. ; SHELESTOV, A.. **Deep learning classification of land cover and crop types using remote sensing data**. IEEE Geoscience and Remote Sensing Letters, 14(5):778–782, May 2017.

37 CASTRO, J. D. B.; FEITOZA, R. Q.; ROSA, L. C. L.; DIAZ, P. M. A. ; SANCHES, I. D. A.. **A comparative analysis of deep learning tech-**

niques for sub-tropical crop types recognition from multitemporal optical/sar image sequences. In: 2017 30TH SIBGRAPI CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES (SIBGRAPI), p. 382–389, Oct 2017.

38 ROSA, L. E. C. L.; HAPP, P. N. ; FEITOSA, R. Q.. **Dense fully convolutional networks for crop recognition from multitemporal sar image sequences.** In: IGARSS 2018 - 2018 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 7460–7463, July 2018.

39 CASTRO, J. B.; FEITOSA, R. Q. ; HAPP, P. N.. **An hybrid recurrent convolutional neural network for crop type recognition based on multitemporal sar image sequences.** In: IGARSS 2018 - 2018 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, p. 3824–3827, July 2018.

40 NDIKUMANA, E.; HO TONG MINH, D.; BAGHDADI, N.; COURAULT, D. ; HOSSARD, L.. **Deep recurrent neural network for agricultural classification using multitemporal sar sentinel-1 for camargue, france.** Remote Sensing, 10(8):1–16, 2018.

41 RUSSWURM, M.; KÖRNER, M.. **Multi-temporal land cover classification with sequential recurrent encoders.** ISPRS International Journal of Geo-Information, 7(4):1–18, 2018.

42 ZHONG, L.; HU, L. ; ZHOU, H.. **Deep learning based multi-temporal crop classification.** Remote Sensing of Environment, 221:430 – 443, 2019.

43 MAUS, V.; CÂMARA, G.; CARTAXO, R.; SANCHEZ, A.; RAMOS, F. M. ; DE QUEIROZ, G. R.. **A time-weighted dynamic time warping method for land-use and land-cover mapping.** IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9:3729–3739, 2016.

44 HORST, B.; MICHAEL, C. T.. **Hidden Markov Models: Applications in Computer Vision**, volumen 45. World Scientific, 2001.

45 GEMAN, S.; GEMAN, D.. **Stochastic relaxation, gibbs distributions, and the bayesian restoration of images.** In: READINGS IN COMPUTER VISION, p. 564–584. Elsevier, 1987.

46 KUMAR, S.; HEBERT, M.. **Discriminative Random Fields.** International Journal of Computer Vision, 68(2):179–201, 2006.

47 LECUN, Y.; BOTTOU, L.; BENGIO, Y. ; HAFFNER, P.. **Gradient-based learning applied to document recognition**. Proceedings of the IEEE, 86(11):2278–2324, Nov 1998.

48 GOODFELLOW, I.; BENGIO, Y. ; COURVILLE, A.. **Deep learning**. MIT press, 2016.

49 LILLESAND, T.; KIEFER, R. ; CHIPMAN, J.. **Remote Sensing and Image Interpretation**. John Wiley & Sons, 2014.

50 WANG, D.; SU, Y.; ZHOU, Q. ; CHEN, Z.. **Advances in research on crop identification using SAR**. In: AGRO-GEOINFORMATICS (AGRO-GEOINFORMATICS), 2015 FOURTH INTERNATIONAL CONFERENCE ON, p. 312–317. IEEE, 2015.

51 LAFFERTY, J. D.; MCCALLUM, A. ; PEREIRA, F. C. N.. **Conditional Random ields: Probabilistic models for segmenting and labeling sequence data**. In: PROCEEDINGS OF THE EIGHTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, ICML '01, p. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

52 SCHINDLER, K.. **An Overview and Comparison of Smooth Labeling Methods for Land-Cover Classification**. IEEE Transactions on Geoscience and Remote Sensing, 50:4534–4545, 2012.

53 FREY, B. J.; MACKAY, D. J. C.. **A revolution: Belief Propagation in Graphs with cycles**. In: PROCEEDINGS OF THE 1997 CONFERENCE ON ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 10, NIPS '97, p. 479–485, Cambridge, MA, USA, 1998. MIT Press.

54 BISHOP, C. M.; MITCHELL, T. M.. **Pattern Recognition and Machine Learning**. Springer, 2014.

55 GONZALEZ, R. C.; WOODS, R. E.. **Digital Image Processing (3rd Edition)**. Pearson, 2002.

56 IOFFE, S.; SZEGEDY, C.. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. arXiv preprint arXiv:1502.03167, abs/1502.03167, 2015.

57 SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I. ; SALAKHUTDINOV, R.. **Dropout: A simple way to prevent neural networks from overfitting**. Journal of Machine Learning Research, 15:1929–1958, 2014.

58 ZHANG, G.; WANG, C.; XU, B. ; GROSSE, R. B.. **Three mechanisms of weight decay regularization**. 1810.12281, 2018.

59 BREIMAN, L.. **Random Forests**. Mach. Learn., 45(1):5–32, 2001.

60 SHOTTON, J.; WINN, J.; ROTHER, C. ; CRIMINISI, A.. **Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context**. International Journal of Computer Vision, 81(1):2–23, 2009.

61 PEEL, M. C.; FINLAYSON, B. L. ; MCMAHON, T. A.. **Updated world map of the Köppen-geiger climate classification**. Hydrology and Earth System Sciences, 11(5):1633–1644, 2007.

62 HASTIE, T.; TIBSHIRANI, R. ; FRIEDMAN, J. H.. **The elements of statistical learning: data mining, inference, and prediction, 2nd Edition**. Springer series in statistics. Springer, 2009.