UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Lucas Borges Aquino

Assessment of Agricultural Production Capacity Through Remote Sensing and Machine Learning

Belo Horizonte
2023

Lucas Borges Aquino

Assessment of Agricultural Production Capacity Through Remote Sensing and Machine Learning

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Nivio Ziviani
Co-Advisor: Adriano Veloso

Belo Horizonte
2023

[Ficha Catalográfica em formato PDF]

A ficha catalográfica será fornecida pela biblioteca. Ela deve estar em formato PDF e deve ser passada como argumento do comando `ppgccufmg` no arquivo principal `.tex`, conforme o exemplo abaixo:

```
\ppgccufmg{
      ...
      fichacatalografica={ficha.pdf}
}
```

[Folha de Aprovação em formato PDF]

A folha de aprovação deve estar em formato PDF e deve ser passada como argumento do comando `ppgccufmg` no arquivo principal `.tex`, conforme o exemplo abaixo:

```
\ppgccufmg{
     ...
     folhadeaprovacao={folha.pdf}
}
```

Dedico este trabalho à minha esposa, Isabella, meus pais, Marcelo e Soraya, e a meu irmão Marcelo.

Acknowledgments

A ESCREVER

"*Porquanto é o SENHOR quem concede sabedoria, e da sua boca procedem a inteligência e o discernimento.*"
(Provérbios 2:6)

Resumo

Neste estudo, apresentamos um sistema automático de análise da capacidade produtiva em áreas agrícolas utilizando de técnicas de aprendizado de máquina e sensoriamento remoto. Para isso, foram utilizadas imagens de satélite, classificando as culturas plantadas e os períodos de plantio. O uso de sensoriamento remoto apresenta diversos desafios, dentre eles, a frequência de captura de imagens por satélite, interferência de nuvens e baixa resolução espacial. A fim de mitigar esses desafios, o sistema utiliza técnicas de expansão geométrica para melhorar a resolução espacial e métodos de preenchimento de dados para superar a frequência limitada e a cobertura de nuvens.

O estudo tem foco no território brasileiro, em particular no cultivo de soja e milho, uma vez que estas são as principais culturas do setor agrícola do país. Este estudo contribui com a construção de um conjunto de dados de área produtivas, distribuídas por todo o território brasileiro e apresenta um fluxo de análise a fim de demarcar regiões produtivas, identificar áreas de plantio, determinar os períodos de plantio e colheita e, por fim, classificar as culturas. Análises comparativas entre técnicas de rotulagem manual e de rotulagem heurística, em aprendizado supervisionado, demonstram a vantagem de métodos de rotulagem manual, devido à alta diversidade de culturas e tipos de plantio.

Os resultados da demarcação de regiões produtivas e classificação de culturas apresentam alto desempenho, com rotulagens manuais alcançando uma área sob a curva (AUC) de aproximadamente 0.94 e rotulagens heurísticas alcançando uma AUC de aproximadamente 0.88. Com os resultados obtidos foi possível compreender a capacidade do sistema proposto em otimizar o processo de avaliação de capacidade produtiva. Além disso, o estudo apresenta uma proposta de abordagem futura na classificação de culturas baseadas em amostras, tendo em vista a alta performance dos modelos supervisionados de rotulagem manual.

Palavras-chave: Aprendizado de Máquina, Sensoriamento Remoto, Capacidade Produtiva, Classificação de Cultura, Rotulagem Manual, Rotulagem Heurística

Abstract

In this study, we present an automatic system for analyzing the productive capacity in agricultural areas using machine learning techniques and remote sensing. For this purpose, satellite images were used to classify the planted crops and planting periods. Remote sensing poses several challenges, including satellite image capture frequency, cloud interference, and low spatial resolution. To address these challenges, the system employs geometric expansion techniques to improve spatial resolution and data filling methods to overcome limited frequency and cloud coverage.

The study focuses on the Brazilian territory, particularly on the cultivation of soybeans and corn, as they are the main crops in the country's agricultural sector. This study contributes to the construction of a dataset of productive areas distributed throughout the Brazilian territory and presents an analysis flow to delineate productive regions, identify planting areas, determine planting and harvesting periods, and ultimately classify the crops. Comparative analyses between manual and heuristic labeling techniques in supervised learning demonstrate the advantage of manual labeling methods due to the high diversity of crops and types of planting.

The results of the demarcation of productive regions and crop classification show high performance, with manual labeling achieving an area under the curve (AUC) of approximately 0.94 and heuristic labeling achieving an AUC of approximately 0.88. The obtained results enable a better understanding of the proposed system's ability to optimize the evaluation of productive capacity. Additionally, the study presents a proposal for future approaches in crop classification based on samples, taking into consideration the high performance of supervised models with manual labeling.

Keywords: Machine Learning, Remote Sensing, Productive Capacity, Crop Classification, Manual Labeling, Heuristic Labeling

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Several remote sensing technologies have been developed for estimating the quality of the land and evaluating the productive capacity of specific regions. These technologies presents maps of crop distribution in the soil as well as the evaluation of socio-environmental issues in the land, such as deforestation and fires [2, 36].

The use of remote sensing for evaluating the productive capacity of the land faces several technical and conceptual challenges. The main difficulties are: frequency of images captured by satellites which have a frequency of three days; quality might be limited by the presence of clouds; and low spatial resolution which have a precision of ten meters. Also important is the understanding of a productive set (called stand) for its classification and delimitation and the high variety in planting patterns according to geolocation, so that in certain regions the planting of soybeans and corn is reversed in relation to the calendar of others, bringing one more difficulty in the culture detection process.

Brazilian agribusiness is among the ten largest exporters in the world [16]. The Brazilian market is mainly composed of a select group of crops, with 90% of the national agricultural territory consisting of soy, corn and sugar cane [12]. In the case of sugarcane, it has a concentrated geospatial behavior because of its direct relationship with ethanol and sugar plants [4].

When we observe the agricultural productivity of soy, corn and sugar cane, we understand that the available financial resource reduces the possibility of competition from small producers. Only producers with the ability to generate agricultural credit can effectively maintain the production cycle [31]. However, the concession of agricultural credit is associated with the productive capacity of the land [11]. To be evaluated it requires a face-to-face visit for estimating the quality of the land, a compilation of historical records to carry out a survey of the revenue capacity of the plantations, among other information. This results in a lengthy and costly process, both for the creditor and for the applicant [34].

1.1 Contributions

In this work, we present an automatic and efficient system for the analysis of productive capacity of large areas of plantation. We use satellite images for evaluating the productive capacity of specific regions. The system is able to (i) delimit the soil in planting regions, (ii) perform the demarcation analysis of the planting and harvesting periods and (iii) classify the crops present in each area. We focus on the Brazilian territory and the classification of plantings of corn and soybeans. The demarcation of regions for planting sugarcane is not necessary because they are close to ethanol and sugar industries. Our system is able to improve the agricultural productivity analysis process, ensuring accessibility and speed in the process of granting rural credit.

In summary, the main contributions of this work are:

- We built a dataset with 185,941 areas geographically spread throughout the Brazilian territory. The dataset created was constructed from three sources: public data containing 149,053 samples and a manually built dataset containing 35,158 samples from areas demarcated as productive, and 1,730 harvests classified manually with start and end dates and the identified crop.

- We developed an analysis pipeline that consisted of demarcating productive regions, selecting planting areas, treating historical series to demarcate planting and harvesting periods and classify the crops present in each area.

- In order to overcome the various challenges in remote sensing techniques, expansion techniques were applied in the process of demarcating productive regions, thus improve the spatial resolution of the images used, as well as methods of filling data in the historical series in order to circumvent the limited frequency and clouds in satellite images.

- We performed a comparative analysis between supervised and semi-supervised techniques for classifying crops. We show that there is a significant advantage of supervised classification techniques, mainly due to the high diversity of cultures and types of plantations in each region.

- Results of the techniques used for demarcation of productive regions and classification of harvest periods are highly performative and aligned with the reality of the data used. The predictive capacity of supervised models present $AUC \approx 0.94$ and unsupervised models present $AUC \approx 0.88$.

1.2 Thesis outline

The remainder of the thesis is organized as follows. Chapter 2 provides a discussion of relevant related work. Chapter 3 presents the dataset created for this study. In Chapter 4 we present our methods for image segmentation, productive area classification, season detection, crop classification. In Chapter 5 we report the results. In Chapter 6 we present concluding remarks.

Chapter 2

Related Work

In this chapter, we review the related work and existing literature on the topics of agricultural analysis, machine learning techniques, remote sensing applications, and crop classification. The focus of the review is on studies that have addressed challenges similar to those encountered in the presented study of analyzing productive capacity in agricultural areas using machine learning and remote sensing.

2.1 Agricultural Analysis using Machine Learning and Remote Sensing

Several researchers have explored the integration of machine learning algorithms with remote sensing data for agricultural analysis. Research by Vibhute [38], demonstrated the wide range of application that remote sensing and machine learning has in the agricultural sector, with applications such as Crop Inventory, Crop type classification, Crop identification, Crop condition monitoring, among others [21, 24]. The use of spectral indices such as NDVI has been explored as a way to demarcate planting regions and classify crops in different ways, results such as those presented by Musande [26] demonstrate the ability to correctly identify crops through remote sensing, reaching 93.12% accuracy in the cotton identification process.

Other studies show the advantage in the analytical process, in initially obtaining certain objects from an image through the segmentation process [32]. Despite the existence of several studies on the crop classification process, little has been explored in the combination of segmentation and classification techniques in order to demarcate planting regions as a whole, results such as those of Vieira [39] demonstrate the high predictive capacity of techniques directly associated with the raw information of the image, the addition of segmentation techniques can further increases the possibility of results.

2.2 Challenges in Remote Sensing for Agricultural Analysis

Remote sensing for agricultural analysis presents numerous challenges that need to be addressed to ensure accurate and reliable results. Researchers have investigated issues related to satellite image capture frequency, cloud interference, and low spatial resolution. Yang [42] shows how the limited spatial, spectral, radiometric and temporal resolutions methods can dificult the remote sensing process. More specifically in the application of remote sensing to agriculture, the need for images with high frequency and low presence of clouds is crucial [17].

In order to address problems such as the presence of clouds, studies have sought to apply techniques to remove obstructed information, and then reconstruct the resulting image based on temporal and spatial information. Lin [23] presents a spatial reconstruction approach based on the premise of low mutability of land cover, the premise is not applicable in the context of plant growth during a season, in view of the rapid plant growth of planted crops, however we can rely on the idea in order to replicate the temporal similarity proposed in the seasonality of regions of planting.

Another problem directly related to the use of remote sensing in agriculture is the low spatial resolution. sensors with long historical periods tend to have low spatial resolution, and high spatial resolution sensors tend to have a short historical interval, considering that the evolution towards high resolution sensors is something recent [33]. The increase in spatial resolution has allowed significant advances in the accuracy of remote sensing techniques [41], however the spatial limitation is still a significant factor when we aim to evaluate geographically smaller regions. Johnsson [20] demonstrates how even high resolution satellites have difficulty spatially delimiting regions when in the presence of high heterogeneity, he presents an approach that uses object-oriented and knowledge-based techniques in order to improve the delimitation process.

Finally, we have the low temporal resolution, the frequency of images from a satellite is related to the time taken for its rotation around the planet earth. When we consider the presence of clouds that prevent obtaining image information, the variability of the temporal frequency increases dramatically and emphasizes the risk of losing accuracy in the analysis. Zhang [44] presents an analysis of the accuracy of satellite-derived phenology, concluding that the absence of up to 16 days of images does not significantly compromise the analytical process, except in periods of phenological transition.

2.3 Crop Classification and Labeling Techniques

Labeling in supervised learning is a crucial part of the process, the quality of the selected data drastically defines the performance of models trained on them. Alonso shows how the quality of the labels of a training set significantly impacts the classification result [3], the approach of different labeling techniques can define the quality of a result as relevant or disposable. Automatic labeling techniques, or based on heuristics, are common in scenarios with large amounts of unlabeled and easily distinguishable data [1], the application of feedback loop systems has already proven capable of resulting in high-performance labeling methods, as presented by Boecking, where a small amount of feedback was enough to train models that achieve highly competitive test set performance[6].

In the context of crop classification, the use of heuristics for demarcation is relatively common, studies such as Patel's show the ability to infer the stages of a crop using the planting and harvesting calendar together with remote sensing information [28]. We also observed the relative ease that exists in mapping crops such as soybeans and corn based on their phenological planting cycle, Zhong presents an accuracy of 87.2% in the process of agricultural mapping of soybeans and corn through regional classification techniques using concepts of phenological cycle [45].

2.4 Differences Between Aforementioned References and This Study

The references mentioned earlier primarily center around agricultural analysis employing remote sensing and machine learning techniques, encompassing tasks like crop classification and land evaluation. However, these references do not explicitly delve into the specific challenges associated with evaluating the productive capacity of large agricultural areas and its implications for rural credit in Brazil. To address this gap, our study takes a more focused approach, dedicating efforts to create an automatic and efficient system for analyzing the productive capacity of extensive plantations in Brazil, with a particular emphasis on corn and soybean crops.

In the realm of related work, various challenges in remote sensing for agricultural analysis are discussed, encompassing issues such as cloud interference, low spatial resolution, and limited temporal resolution. Our study acknowledges these challenges and implements suitable solutions. We apply expansion techniques to enhance spatial resolu-

tion, adopt methods for filling data in historical series to mitigate the effects of limited frequency and cloud-related problems in satellite images, and explore the advantages of utilizing supervised classification techniques for accurate crop classification.

Chapter 3

Dataset Construction

The dataset created for this study was constructed from three sources: public satellite images, public data and a manually built dataset, as detailed next.

3.1 Public Satellite Image Acquisition Datasets

The image data was captured by the Sentinel-2 satellite [29]. The satellite has multispectral information with a resolution of 10 to 60 meters, varying according to the band, with an approximate granularity of three days. All data were obtained using the Google Earth Engine tool [27], which provides already treated bands, specific indices and cloud filters.

We selected approximately 9 million good quality images for the Brazilian territory obtained after 2018, as we use an API from Earth Engine to provide the data, the image estimation was made considering the average frequency of the satellites, 3 days, the number of months and the number of areas. To identify the type of crops, frequency of planting and harvesting we extracted indices from the raw satellite bands. The main index was the Normalized Difference Vegetation Index (NDVI), which is used to quantify the greenness of plants and the density of vegetation on the soil [40].

3.2 Public Datasets

We used three public datasets for classifying productive lands: (i) Mapbiomas [35], with $51,957$ samples, (ii) TerraClass [9] with 46,183 samples, and (iii) CONAB [10] with 50,913 samples, totaling 149,053 samples. A conjuncture of soil mapping bases presenting

a relationship between the culture and the soil were used to demarcate the productivity of different cultures in the Brazilian territory.

3.3 Manually Built Datasets

The manually built datasets were constructed in a partnership with Tarken[1], an agricultural credit analysis company. Tarken provides a platform to optimize the entire credit cycle for clients, integrating the entire workflow from the registration form through credit approval. One important step in this workflow is to decide whether an area of land is considered productive, which harvest is carried on with start and end dates and the identified crop. The platform carry out a feedback loop system to obtain manual input data directly from producers and agricultural resellers.

The data was collected using a manual analysis system on satellite images, temporal graphs of the reported bands and assertive response from producers and resellers who made use of the platform. We build a database containing 35,158 samples from areas demarcated as productive, as well as 1,730 harvests classified manually, with start and end dates and the identified crop throughout the years of 2018 to 2022.

[1] www.tarken.com.br

Chapter 4

Methods

In this chapter we present our methods for image segmentation, productive area classification, season detection, crop classification.

4.1 Image Segmentation

The vast majority of studies related to the mapping of crops in the soil perform a spatial analysis, demarcating planting regions. However, the granularity of the marking is related to the spatial resolution of the satellite, which in most cases does not correspond to reality, considering that planting follows a pattern of behavior in pre-demarcated regions.

In order to classify planting with greater precision, we propose a stage of demarcation of planting regions based on the NDVI correspondence over time. For this, an extraction of the TIFF file of a given region was carried out with the NDVI information pixel by pixel to form a temporal series of the level of biomass in the soil for each pixel. Each image was grouped monthly with the mean value of the NDVI on each pixel, resulting on a series of 60 point per pixel.

Next, we perform a temporal clustering by grouping those that have similar behavior. For clustering, the Felzenszwalb Segmentation [14] algorithm was applied, followed by the construction of polygons based on the demarcation of discovered clusters, the algorithm uses a local threshold to determine the limit of each cluster, in our experiments it was possible to observe that this value behaves similarly throughout the distribution of the national territory, thus allowing the selection of a single value in all images.

Despite the segmentation being capable of delimiting the spatial separation of terrains with high precision, there is still a spatial limitation of the satellite itself. In the case of this study the spatial limitation is 10 meters, which results in certain polygons with high eccentricity, often related to roads or physical borders of terrains. We then perform a post-processing for removing polygons with high eccentricity and high proportion of area and perimeter. Finally, we make a spatial expansion of the resulting polygons using the

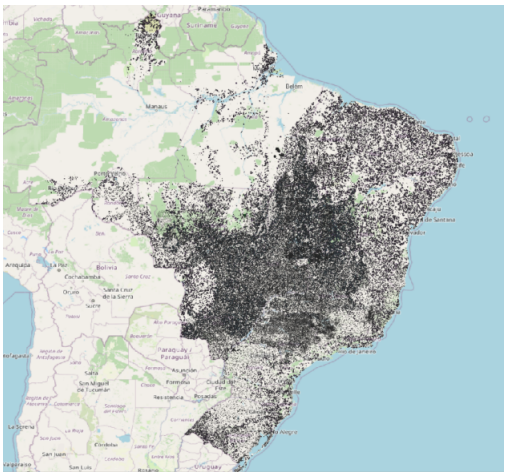

Figure 4.1: The distribution of areas covers almost the entire national territory, with the exception of the Amazon region, in view of the high density of the forest.

concept of Voronoi diagrams [5] and obtain 100% of ground cover, filling in the spaces generated by the aforementioned post processing.

4.2 Productive Area Classification

After the demarcation of regions with temporal correspondence of biomass, we select regions that are productive, that is, those that are within the cycle of planting and harvesting. For this, we carry out a supervised classification using public soil mapping datasets demarcating productive or not productive polygons according to the percentage of coverage of the area marked as productive. In the end, we build a dataset with 185,941 areas geographically spread throughout the Brazilian territory. Figure 4.1 presents the diversity of the distribution of areas for for the national territory, with the exception of the Amazon region, in view of the high density of the forest.

This dataset were divided in five sets for using the 5-folds cross-validation technique, which were used in the training of a XGBoost [8] model, using as features the following six indices extracted from the collected satellite images:

i. Normalized Difference Vegetation Index (NDVI) [40]

ii. Enhanced Vegetation Index (EVI) [40]

iii. Canopy chlorophyll content index (CCCI) [15]

iv. Leaf Area Index (LAI) [13]

v. Normalized Difference Moisture Index (NDMI) [19]

vi. Normalized Difference Red Edge Index (NDREX) [37]

For each index, we aggregate the average values of the demarcated polygons, thus forming a single value for each index, polygon and date. In this way we arrive at a multivariate time series, from which we extract time characteristics including mean, minimum, maximum, variance and standard deviation

4.3 Season Detection

In order to demarcate the cultures of each harvest it is necessary to demarcate the planting and harvesting periods in time, given that certain planting regions are used for planting multiple crops, such as soybeans in summer and corn in winter. For this we use the NVDI index, which visually shows the behavior of the crop over time, being possible to observe "waves" in the periods of crop growth. This occurs due to the process of harvesting and plowing the soil, which reduces the level of biomass to close to 0, generating the effect of a fall in the curve, followed by growth associated with the next planting carried out.

However, satellite images are subject to noise, mainly due to the presence of clouds. Therefore, we appliy a filter to remove pixels identified with clouds using the information provided by the GEE on the probability of cloud presence in each pixel. Then it is necessary to apply a smoothing to the curve in order to remove visual noise using the filter Savitzky-Golay (SavGol) [7, 30] and obtaining the demarcation of the periods. Figure 4.2 shows that the filter is able to remove noise without compromising the behavior analysis of the series.

Considering that there is an extremely strong correlation between planting and rainy seasons, it is very common that most of the photos during the harvest period have high cloud coverage, reaching often 90% or more cloud coverage in the region, thus making analysis impossible. In order to circumvent the problem in question, a data imputation technique was proposed, for periods of long absence of information. Using the historical series itself, we trained a KNN [43] with the neighborhood information of each point, last observation, next observation, moving average of the last three observations and day of the year of the current observation, in order to predict the value in days of absence.

Then we apply a technique for demarcation of the valleys in the time series by using the second derivative of the series for picking the inflection points, thus demarcating all the valleys. However, due to the noise there are still markings of valleys that do not

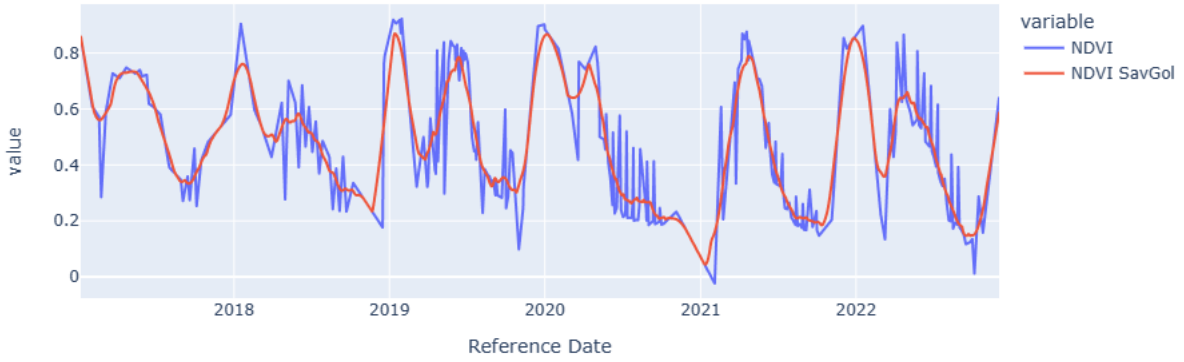

Figure 4.2: The NDVI graph is extremely noisy and we apply a Savitzky-Golay filter (SavGol) to remove noise without compromising the behavior analysis of the series.

correspond to periods of beginning or end of crops. To remove invalid markings we apply the following heuristic:

i. In the case of intervals of less than 60 days between a start and an end tag, mark both points as invalid.

ii. If the lowest NDVI value during the interval is greater than 0.8, mark both points as invalid.

iii. If the highest NDVI value during the interval is less than 0.1, mark both points as invalid.

iv. If the distance from the marked point to the peak in NDVI value is less than 0.1, that is, there is no real growth or decay of values, mark both points as invalid.

Finally, we select all the intervals marked as invalid and check the distance from it to the next valid interval, if it is less than 15 days, we group both intervals generating a single marking of harvests.

4.4 Crop Classification

After the demarcation of the harvest periods, we proceed to the final analysis for the classification of the cultures of each harvest. For that we carry out the following four approaches:

1. The application of a heuristic using the soybean and corn agricultural calendar [22, 25], that is, classifying both crops based on planting and harvesting dates in each geographic region.

2. Training a model using part of the dataset of 1,730 harvests classified manually (see Section 3.3).

3. Training a model using a dataset of 35,158 harvests classified with the heuristic (see step1).

4. Training a model using the two datasets built in steps 2 and 3.

The four approaches were evaluated on the dataset of manually classified crops (see Section 3.3) using a XGBoost model. The dataset of 1,730 harvests was divided using the 70/30 technique, thus resulting in 502 crops for evaluation and 1,228 for training. In all cases the multi class classification was evaluated for corn, soy or unknown considering the possibility of the existence of other crops in the middle of the dataset.

To train the model we perform an extraction of temporal features referring to the crop interval, namely, the NDVI peak of the period, the day of the year of planting, the day of the year of harvest, the Kurtosis of the curve and the Skewness of the curve, characteristics that were evaluated through the manual analysis of the curves already classified as potential differentiators of the main crops.

Chapter 5

Results

In this chapter we discuss the results obtained in each of the worked categories, image segmentation, productive area classification, season detection, crop classification.

5.1 Image Segmentation

In order to evaluate the quality of land segmentation a qualitative observation of the result was necessary, in view of the absence of datasets for comparisons in direct metrics. Despite this, the qualitative assessment for the case in question might be measured by the visualization contrasted with the satellite image. Figure 5.1 shows that the original segmentation clearly delimits the separations of the different productive areas, as well as the areas of native vegetation and constructions such as the headquarters of a farm or roads.

Figure 5.1: The image presents an example of segmentation carried out on top of spatial clippings which were later grouped.

Figure 5.2: We can observe the reduction in the number of polygons, as well as the total land cover being maintained, thus improving the quality of the real polygons.

We can also observe from Figure 5.1 that despite the good delimitation there are still noisy polygons among the demarcated ones, many with high eccentricity or are very small. For noisy polygons we apply the filter criteria followed by the Voronoi technique to expand the area and obtain 100% ground cover, as shown by the image of Figure 5.2.

Although we observe some unnecessary separations inside productive regions, they do not have a negative impact considering that, after classifying productive regions, we arrive at planting polygons. Even if they have more divisions than necessary, the biomass behavior in the subdivisions does not differ to the point of affecting season detections and crop classifications. Together with this, we can observe the high capacity of the model to distinguish non-productive regions as a single cluster, thus facilitating the classification analysis of productive areas described in the following section.

With the result presented, we carried out the segmentation process throughout the national territory, generating a database containing 18,223,305 demarcated areas. The delimitation process showed a high ability to differentiate regions based on the behavior of historical biomass. In this study, the focus was solely on the classification of the planting area, but based on the observed results, the technique in question has the possibility of being used in other processes for evaluating the behavior of biomass, such as detection of deforestation, loss of biomass in native vegetation, among others.

Figure 5.3: As we only plot the polygons marked as productive, we observe the false positive rate approaches zero and our biggest problem becomes false negatives.

5.2 Productive Area Classification

The classification of productive regions was evaluated with an extensive dataset of demarcated regions, so it was possible to observe both the numerical results of the binary classifier, as well as a qualitative visualization similar to the original segmentation. Figure 5.3 shows only productive stands plotted on the map. In spite of the fact that the classification is easy to demarcate non-productive regions, the existence of false negatives is an obvious problem, in order to mitigate the occurrence of false negatives, we sought to better understand the distribution of classes, which led us to consider a more tolerant classification threshold in order to increase recall. False negatives are often linked to anomalous behavior in the planting regions, as the planting pattern is often not uniform and we have polygons with historical series very different from the standard behavior of productive regions.

Despite the qualitative result demonstrating the lack of precision in the classification, when we expand the scope of analysis to the total set of data that we use, containing 185,941 areas, we have more direct metrics, the real result of the classifier, arriving at the results presented in Table 5.1, we can observe that corroborating the hypothesis observed in the qualitative analysis we have a low precision of 0.765. However, AUC [18] reaches the value of 0.94, which is a more robust metric for situations of imbalance of classes, which corresponds to our real scenario, considering the high density of productive regions in the analyzed areas, but their low spatial distribution, we can observe this scenario

in the base used, which contains 100 thousand non-productive samples and 80 thousand productive samples.

Table 5.1: Productive Classifier result metrics

Recall	F1 Score	Precision	Balanced Accuracy	AUC
0.904	0.829	0.765	0.878	0.940

Thus, we can observe the high reliability in the regions marked as productive, allowing subsequent analyzes to be carried out only on polygons that are correctly marked as productive, thus facilitating the marking of crops and classification of crops.

5.3 Season Detection

In order to exemplify the impact that the presence of clouds during the planting period can have on the demarcation, we selected an NDVI sample from a region with crop (soybean) and off-season (corn) production over the last five years. In this region there is a high presence of clouds in year 2022, thus compromising the demarcation of harvest periods and identification of periods of planting as observed in the image from Figure 5.4. However, we can observe in the image from Figure 5.5 that in the actual observation there was a harvest carried out in January, planting in February and in March a new culture was in place. In this way, when applying the KNN-Inputer, it was possible to reconstruct the curve based on the historical behavior, thus correctly demarcating the period of harvest and new planting, as observed in the image from Figure 5.6.

For the aforementioned example, we have a case of recurrent behavior, that is, planting in the evaluated region maintained the same pattern over the years. As we use a clustering technique based on the historical series itself, we run a risk in regions that have a high variance in planting behavior, cases in which the region changes the planted crop or the planting windows may result in incorrect reconstructions in the curve.

In order to more clearly assess this risk, we carried out an experiment with a region that underwent a drastic change in behavior, moving from planting Sugar Cane (annual crop) to alternating between Corn and Soybean, in which it can be seen that the impact did not compromise the analysis of the behavior of the curve, considering that we only carried out the correction in regions that spent more than 15 days without observable information, thus reducing the probability of correcting the curve as a whole, bearing in mind that the shorter period of harvests is 90 days. We can see in Figure 5.7 that the KNN method does not change the curve to the point of losing information about it.

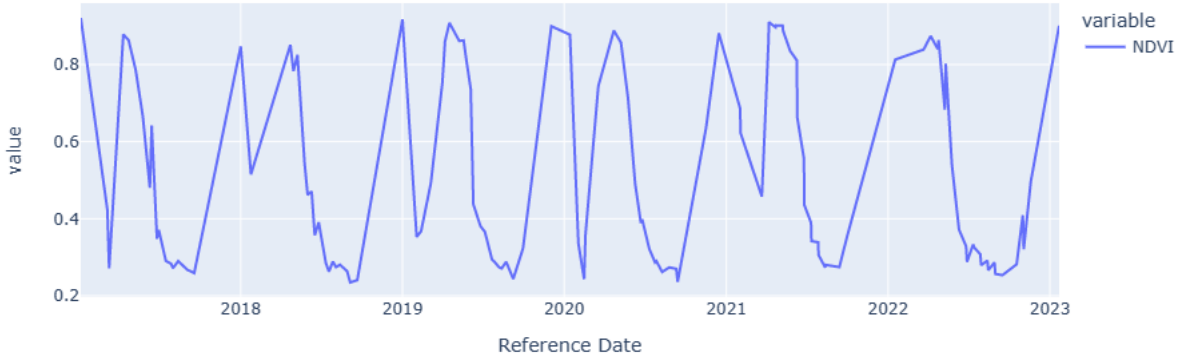

Figure 5.4: In the year 2022, the presence of clouds completely compromises the demarcation of the soybean harvest and corn planting.

Figure 5.5: We can observe the evolution on 01/22, 02/22 and 03/22 of the land being harvested and replanted.

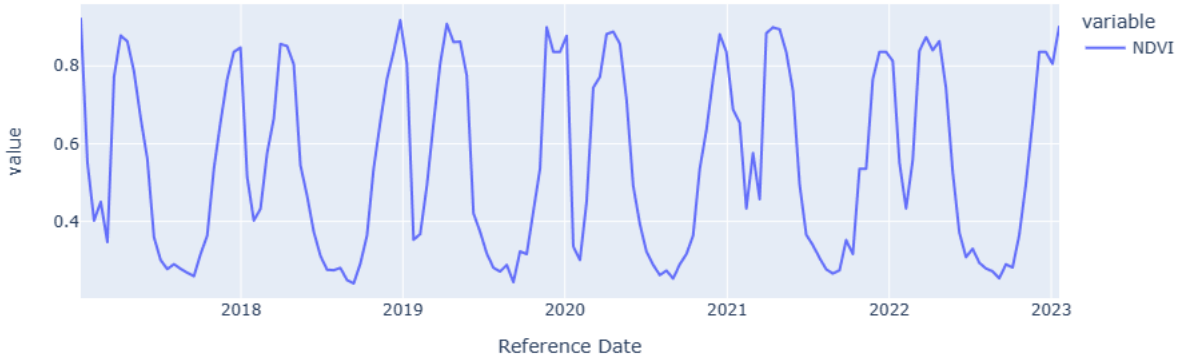

Figure 5.6: We can observe how the curve in 2022 is reconstructed while not losing the behavior of previous years.

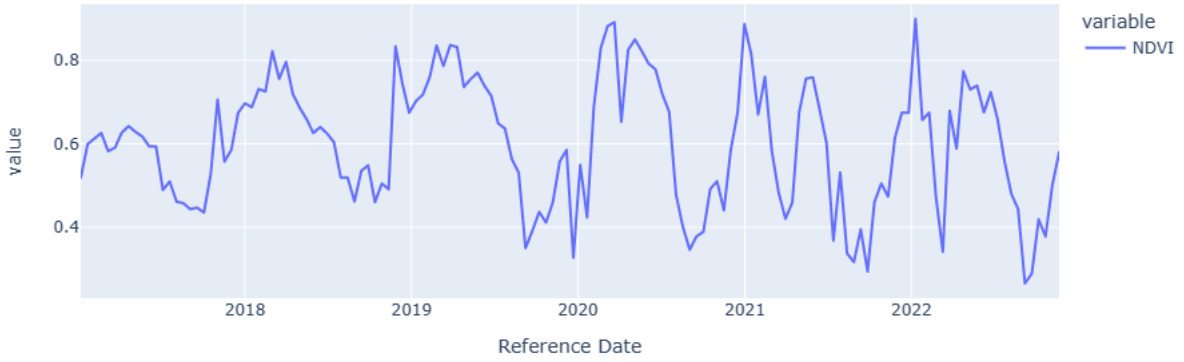

Figure 5.7: Until the year 2021 we have a sugarcane plantation, with annual harvest periods, in 2021 and 2022 we have the planting of soybeans and corn.

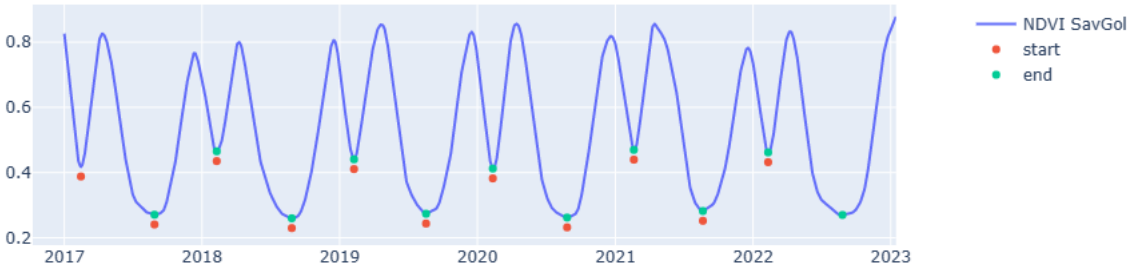

Figure 5.8: The markings for the beginning of the end overlap in the soybean and corn crops, it is common for there to be planting right at the time of the previous crop's harvest.

Finally, we mark the harvest intervals using the smoothed curve, despite the preprocessing complexity, the demarcation process is considerably simple and low risk, considering that we only seek the valley points on the curve and correct the intervals with the heuristics, the results are quite reliable in the demarcation of the intervals, being penalized only in the precision of exact planting and harvesting dates that are displaced due to the application of the filters. We can observe in Figure 5.8 that the seasons intervals are correctly marked, however the start dates suffer some noise with the smoothing of the original graph.

5.4 Crop Classification

With the crop ranges demarcated, we can now classify the crops in each range, using the information extracted between planting and harvesting dates. In order to evaluate the classification capacity, we used the 502 manually demarcated seasons. Initially, we evaluated the calendar heuristic as a crop classification technique, using the planting and harvesting dates for demarcation, we arrived at the results presented in the confusion matrix 5.2, with a balanced F1 of .827 being the approach with the worst result, very related to crops that have a calendar similar to corn and soybeans but differ in other aspects, such as speed of growth or harvest technique.

Table 5.2: Heuristic confusion matrix

	Soy	Corn	Others
Soy	161	2	54
Corn	0	131	25
Others	3	7	119

Then we have the machine learning approach with heuristic labeling techinique, in which we train a classification model using a dataset of 35,158 seasons built with the mentioned heuristic, the model in question was evaluated in the same 502 seasons and obtained the confusion matrix 5.3 that demonstrates a small improvement over the direct heuristic, but not very expressive with a balanced F1 of .829, facing similar problems in the classification of other cultures.

Table 5.3: Heuristic labeling confusion matrix

	Soy	Corn	Others
Soy	162	2	53
Corn	0	131	25
Others	3	7	119

Then we have the manual labeling techinique, using a separate base of 1,170 harvests for training, in this case we can observe a significant improvement in relation to the previous approaches, mainly due to the better ability to differentiate the cultures marked as "others", in relation to corn and soybean labels. In the confusion matrix 5.4 we can see how the reduction of false classifications of others was significant, thus reaching a balanced F1 of .942.

Finally, we carried out the joint approach, using both datasets, heuristic and manual, for model training. In this approach we observed a significant improvement in relation

Table 5.4: Manual labeling confusion matrix

	Soy	Corn	Others
Soy	201	2	14
Corn	0	151	5
Others	6	2	121

to the first two, however a worsening in relation to the manual approach, reaching a balanced F1 of .882, again having as a major problem the distinction of crops marked as "others" in relation to maize and the soybean.

Table 5.5: Joint approach confusion matrix

	Soy	Corn	Others
Soy	177	2	38
Corn	0	143	13
Others	3	5	121

We can therefore understand that the main limitation of heuristic techniques in the context of the problem in question are the unknown varieties of planting, as we deal with unusual planting scenarios, or even uncommon crops, we are confronted with several cases of crops in periods similar to the of soy and corn, or cases of soy and corn "out of season" that is, planted in periods outside the agricultural calendar. Even with the application of the heuristic added to the model, the classification based only on date results in an incorrect conduction for the model, causing it to lose its predictive capacity of separating different cultures and the two main ones.

Chapter 6

Conclusions

The agricultural productivity analysis process is important to ensure accessibility and speed in the process of granting rural credit. In order to evaluate the ability to carry out agricultural monitoring using remote sensing, we developed an analysis pipeline that consisted of demarcating productive regions, selecting planting areas, treating historical series to demarcate planting and harvesting periods, and finally a comparative analysis of manual and heuristic labeling approaches to classifying crops at season intervals. In order to overcome the various challenges in remote sensing techniques, expansion techniques were applied in the process of demarcating productive regions, thus helping the spatial resolution of the satellites used, as well as methods of filling data in the historical series in order to circumvent the limitations frequencies and clouds in satellite images.

It was possible to verify that the predictive capacity of the model trained on the manual dataset has a significant advantage over the model trained on the heuristic dataset, with values of 0.94 of F1 in relation to 0.88 respectively. This is directly related to the presence of several behavior outliers in planting patterns, which we can observe with crops outside the common period, as well as the presence of exception crops that, despite making up only 10% of the national territory in absolute terms , have a significant impact when looking at regions as a whole.

The study in question allows the evolution in three main fronts, initially it was possible to observe the high capacity of demarcation of regions using the temporality of the biomass added of clustering techniques, with the results presented we observe the possibility of expanding the technique in question to approach demarcators of deforestation, loss of biomass, invasion of the preservation area, among others.

Secondly, we have the productive areas classification approach, with the observed results it was possible to understand the lack of classification of non-productive regions, the information used in this approach is restricted to aspects of the local vegetation, however there are a series of spectral indexes that can be used to identify non-productive regions, such as information regarding the presence of buildings, water, rock formations, soil type, etc.

Finally, we have the evaluation of planting periods, and classification of cultures, in both cases it is possible to observe the direct relationship between the NDVI curve,

with the planting intervals and the planted culture. The high predictive capacity in the given scope demonstrates the effectiveness of the approach by observing characteristics of the culture, with this we can approach curve demarcation techniques based on similarity, techniques similar to the speech demarcation approach in sound waves, such as dynamic time warping.

With the aforementioned propositions, it is possible to expand the study in question to a different classification, including in the process the demarcation of more specific crops in addition to corn and soybeans, thus allowing a more diverse assessment of planting contexts and the application of the proposed flow in regions with greater diversity of crops.

References

[1] Sagheer Abbas, Syed Ali Raza, MA Khan, Muhammad Adnan Khan, Kiran Sultan, Amir Mosavi, et al. Automated file labeling for heterogeneous files organization using machine learning. *Computers, Materials & Continua*, 74(2), 2023.

[2] Cláudio Aparecido de Almeida, Alexandre Camargo Coutinho, Júlio César Dalla Mora Esquerdo, Marcos Adami, Adriano Venturieri, Cesar Guerreiro Diniz, Nadine Dessay, Laurent Durieux, and Alessandra Rodrigues Gomes. High spatial resolution land use and land cover mapping of the brazilian legal amazon in 2008 using landsat-5/tm and modis data. *Acta Amazonica*, 46:291–302, 2016.

[3] Omar Alonso. Challenges with label quality for supervised learning. *Journal of Data and Information Quality (JDIQ)*, 6(1):1–3, 2015.

[4] Paulo Arruda. Perspective of the sugarcane industry in brazil. *Tropical plant biology*, 4:3–8, 2011.

[5] Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[6] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision: Learning useful heuristics for data labeling, 2021.

[7] Jin Chen, Per Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita, and Lars Eklundh. A simple method for reconstructing a high-quality ndvi time-series data set based on the savitzky–golay filter. *Remote sensing of Environment*, 91(3-4):332–344, 2004.

[8] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[9] Terra Class. Projeto terra class cerrado. 2013.

[10] Companhia Nacional de Abastecimento. Mapeamento agrícolas conab, 2019.

[11] Eduardo Rodrigues de Castro and Erly Cardoso Teixeira. Rural credit and agricultural supply in brazil. *Agricultural Economics*, 43(3):293–302, 2012.

[12] Lívia CP Dias, Fernando M Pimenta, Ana B Santos, Marcos H Costa, and Richard J Ladle. Patterns of land use, extensification, and intensification of brazilian agriculture. *Global change biology*, 22(8):2887–2903, 2016.

[13] Hongliang Fang, Frederic Baret, Stephen Plummer, and Gabriela Schaepman-Strub. An overview of global leaf area index (lai): Methods, products, validation, and applications. *Reviews of Geophysics*, 57(3):739–799, 2019.

[14] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004.

[15] Glenn Fitzgerald, Daniel Rodriguez, and Garry O'Leary. Measuring and predicting canopy nitrogen nutrition in wheat using a spectral index—the canopy chlorophyll content index (ccci). *Field crops research*, 116(3):318–324, 2010.

[16] Food and Agriculture Organization of the United Nations. Faostat statistical database, 2015.

[17] Feng Gao and Xiaoyang Zhang. Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities. *Journal of Remote Sensing*, 2021, 2021.

[18] J.A. Hanley and Barbara Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 05 1982.

[19] Suming Jin and Steven A Sader. Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances. *Remote sensing of Environment*, 94(3):364–372, 2005.

[20] Katarina Johnsson. Segment-based land-use classification from spot satellite data. *Photogrammetric engineering and remote sensing*, 60(1):47–54, 1994.

[21] R. Kalpana, Sivaraj Natarajan, S. R. Mythili, D. Esther Shekinah, and J. Krishnarajan. Remote sensing for crop monitoring – a review. *Agricultural Reviews*, 24:31–39, 2003.

[22] M Kaster and JRB Farias. Regionalização dos testes de vcu-valor de cultivo e uso de cultivares de soja-terceira aproximação. 2011.

[23] Chao-Hung Lin, Po-Hung Tsai, Kang-Hua Lai, and Jyun-Yuan Chen. Cloud removal from multitemporal satellite images using information cloning. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):232–241, 2013.

[24] Ji-hua Meng and Bingfang Wu. Study on the crop condition monitoring methods with remote sensing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, 01 2008.

[25] Pecuária e Abastecimento Ministério da Agricultura. Calendário agrícola milho, 2022.

[26] Vijaya Musande, Anil Kumar, and Karbhari Kale. Cotton crop discrimination using fuzzy classification approach. *Journal of the Indian Society of Remote Sensing*, 40, 12 2012.

[27] Onisimo Mutanga and Lalit Kumar. Google earth engine applications, 2019.

[28] J. H. Patel and M. P. Oza. Deriving crop calendar using ndvi time-series. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-8:869–873, 2014.

[29] Darius Phiri, Matamyo Simwanda, Serajis Salekin, Vincent R Nyirenda, Yuji Murayama, and Manjula Ranagalage. Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12(14):2291, 2020.

[30] William H Press and Saul A Teukolsky. Savitzky-golay smoothing filters. *Computers in Physics*, 4(6):669–672, 1990.

[31] Nicholas Rada and Constanza Valdes. Policy, technology, and efficiency of brazilian agriculture. *USDA-ERS Economic Research Report*, (137), 2012.

[32] Md Rahman and S. Saha. Multi-resolution segmentation for object-based classification and accuracy assessment of land use/land cover classification using remotely sensed data. *Journal of the Indian Society of Remote Sensing*, 36:189–201, 06 2008.

[33] Rajendra P. Sishodia, Ram L. Ray, and Sudhir K. Singh. Applications of remote sensing in precision agriculture: A review. *Remote Sensing*, 12(19), 2020.

[34] Xin Song, Li Li, and Lei Xiao. Review of research on credit risk management for rural credit cooperatives. *Journal of Risk Analysis and Crisis Response*, 7(1), 2017.

[35] C Souza and Tasso Azevedo. Mapbiomas general handbook. *MapBiomas: São Paulo, Brazil*, pages 1–23, 2017.

[36] Carlos M. Souza, Julia Z. Shimbo, Marcos R. Rosa, Leandro L. Parente, Ane A. Alencar, Bernardo F. T. Rudorff, Heinrich Hasenack, Marcelo Matsumoto, Laerte G. Ferreira, Pedro W. M. Souza-Filho, Sergio W. de Oliveira, Washington F. Rocha, Antônio V. Fonseca, Camila B. Marques, Cesar G. Diniz, Diego Costa, Dyeden Monteiro, Eduardo R. Rosa, Eduardo Vélez-Martin, Eliseu J. Weber, Felipe E. B. Lenti, Fernando F. Paternost, Frans G. C. Pareyn, João V. Siqueira, José L. Viera, Luiz C. Ferreira Neto, Marciano M. Saraiva, Marcio H. Sales, Moises P. G. Salgado, Rodrigo Vasconcelos, Soltan Galano, Vinicius V. Mesquita, and Tasso Azevedo. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17), 2020.

[37] Corey N Thompson, Wenxuan Guo, Bablu Sharma, and Glen L Ritchie. Using normalized difference red edge index to assess maturity in cotton. *Crop Science*, 59(5):2167–2177, 2019.

[38] Amol D Vibhute and Bharti W Gawali. Analysis and modeling of agricultural land use using remote sensing and geographic information system: a review. *International Journal of Engineering Research and Applications*, 3(3):081–091, 2013.

[39] Carlos Vieira, Paul Mather, and Paul Aplin. Agricultural crop classification using the spectral-temporal response surface. 07 2002.

[40] John Weier and David Herring. Measuring vegetation (ndvi & evi). nasa earth observatory. *Washington, DC, USA*, 2000.

[41] Chenghai YANG. High resolution satellite imaging sensors for precision agriculture. *Frontiers of Agricultural Science and Engineering*, 5(4):393, 2018.

[42] Daiqin Yang, Zimeng Li, Yatong Xia, and Zhenzhong Chen. Remote sensing image super-resolution: Challenges and approaches. pages 196–200, 2015.

[43] Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85(11):2541–2552, 2012.

[44] Xiaoyang Zhang, Mark A. Friedl, and Crystal B. Schaaf. Sensitivity of vegetation phenology detection to the temporal resolution of satellite data. *International Journal of Remote Sensing*, 30(8):2061–2074, 2009.

[45] Liheng Zhong, Lina Hu, Le Yu, Peng Gong, and Gregory S Biging. Automated mapping of soybean and corn using phenology. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119:151–164, 2016.