Phenological classification of crops in Northwest Argentina using 250-meter MODIS imagery

A thesis proposal by Jarrett A. Keifer

July 9, 2014

Abstract

Subtropical deforestation in Latin America is thought to be driven by demand for agricultural land, particularly to grow soybeans. However, existing remote sensing methods that can differentiate crop types to verify this hypothesis require high spatial or spectral resolution data, or extensive ground truth information to develop training sites, none of which are freely available for much of the world. Here, I propose a new method of crop classification using multi-temporal MODIS vegetation indices as a base image from which to extract crops using their phenologies. I test and refine this method in Kansas, USA using the USDA crop data layer as reference. I then test the applicability of the method to other regions of the world by applying it to data from Pellegrini, Santiago Del Estero, Argentina. The study is to examine if using phenological profiles in image classification is a viable method to verify the initial hypothesis that soybeans are driving deforestation in subtropical South America. The ability to map agricultural lands by crop type is crucial to understanding the geography and dynamics of land use and land cover change. Existing remote sensing methods that can differentiate crops by type require high spatial resolution data, high spectral resolution data, or extensive ground truth information to develop training sites, none of which are freely available for much of the world. As an alternative, I propose a new method of crop classification using multi-temporal MODIS vegetation indices as a base image from which to extract crops using their

phenologies. I test and refine this method in Kansas, USA using the USDA Cropland Data Layer as reference. I discuss the numerous factors that effect the application and accuracy of the method, the method's current limitations, and how the method might be further tested and refined.

Introduction

Deforestation has long been a concern throughout tropical South America. However, this process of land use/land cover (LULC) change from forest to other uses has been increasingly recognized in subtropical South America as a significant source of environmental degradation. Understanding the complex dynamics of subtropical deforestation is crucial given the prominent role of forests in debates about climate change, conservation, and the protection of endangered species (Geist and Lambin 2002; Zak, Cabido, and Hodgson 2004; Bonnie 2000; Houghton 1994; Sala et al. 2000).

Currently, many perceive growing demand for agricultural land—particularly land for soybeans—to be one of the greatest pressures on South American subtropical forests (Pengue 2005; Grau, Gasparri, and Aide 2005; Altieri and Pengue 2006). Remote sensing has given researchers a tool to classify land cover and measure deforestation, but the often used multi-spectral or multi-temporal image classification techniques require extensive ground truth information for the accurate classification of common crop types using widely-available data. Therefore, getting a complete picture of the dynamics of deforestation, including an understanding of agricultural pressures on forests, requires a significant expense for high spatial or high spectral resolution data, or for field time gathering training site data. The development of a tool that can efficiently and effectively extract crop types using widely-available imagery would be of value to the field.

The primary goal of my thesis is to develop and test a a phenological classification algorithm that can identify and extract crop types from a multi-date vegetation index sequence assembled using free and accessible data from the National Aeronautics and Space Administration's (NASA)

Moderate Resolution Imaging Spectroradiometer (MODIS) platform. I will first test the algorithm using five small test areas from across the state of Kansas using the U.S. Department of Agriculture's (USDA) crop data layer (CDL) as ground truth to derive reference crop phenologies and to test the accuracy of the classification. Then, once I have determined the best parameters for use, I will apply the method to the Department of Pellegrini in Santiago del Estero, Argentina (Fig. 1) during the 2013-2014 growing season to examine the method's applicability in subtropical South America.

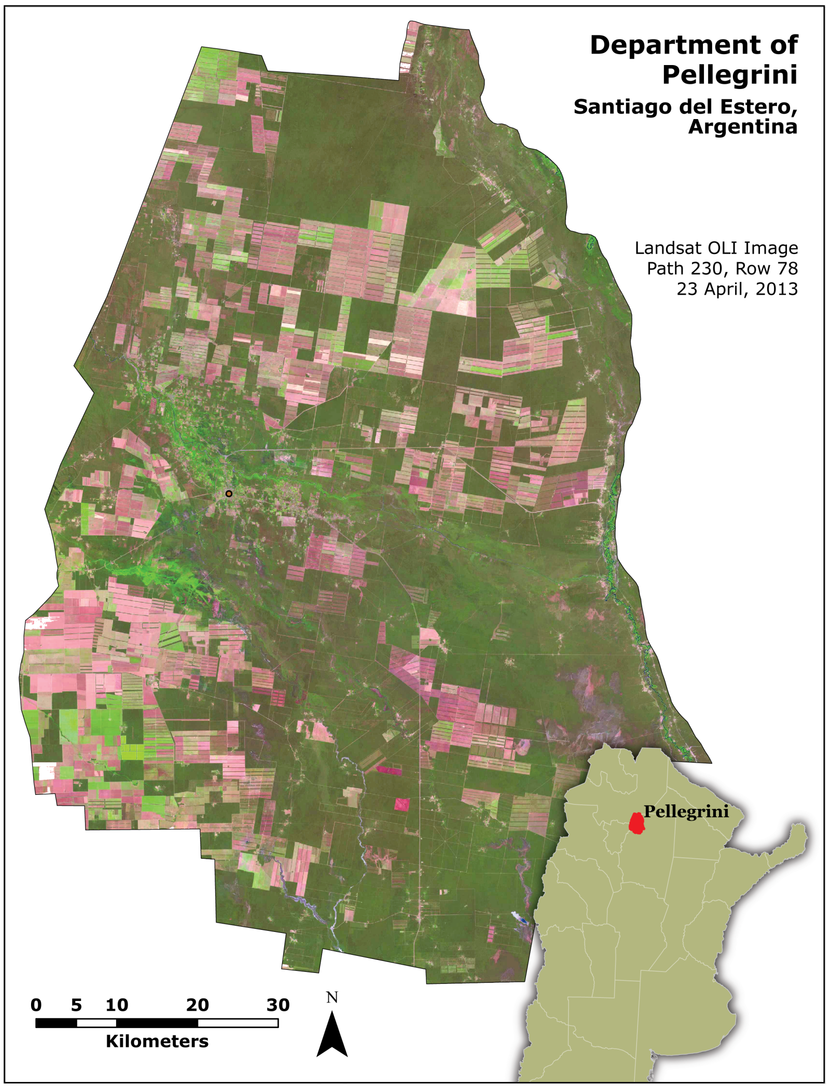


Figure 1: The department of Pellegrini, Santiago Del Estero, Argentina.

Background

Deforestation and the *Ley de Bosques* (Forest Act) in Argentina

The conversion of forestland to other uses has seriously impacted Argentina's forests. In 1915 it was remarked that 30 percent of the country had forest cover, but in 2001 only 10 percent remained forested (Secretaría de Desarrollo Sustentable y Política Ambiental [Argentina] 2001). Over the period 1998 to 2002, Argentina lost over 940,000 hectares of forest cover (Secretaría de Ambiente y Desarrollo Sustentable [Argentina] 2007). The high rate of deforestation concerned policymakers, and Law 26.331, or the *Ley de Bosques* (Forest Act), was voted into law in 2007 in an effort to preserve remaining native forests. Areas of native forest are defined to be those with forest cover of at least 20 percent native species, and that have trees of a minimum of 7 meters high. The law designates red, yellow, and green areas, each with different restrictions on clearing and use. Red is assigned to areas of "high conservation value," yellow is for areas that must be managed sustainably, and green allows "partial or total use" (Gulezian 2009: 25). Each provincial government was responsible for determining how to classify their native forest area, and each enacted the *Ley de Bosques* regulations under the *Ordenamiento Territorial de los Bosques Nativos* (Land Management Order for Native Forests, OTBN).

As a part of Law 26.331 ongoing land cover studies are done to examine the effectiveness of the legislation. Between 2006 and the passing of the law, 573,296 hectares of native forest cover were lost. From the passing of the law in 2007 and the classification of the OTBN areas in 2009, a further 473,001 hectares were deforested. From the enacting of the OTBN (in 2009) and 2011, some

459,108 hectares were found to have been lost (Secretería de Ambiente y Desarrollo Sustentable [Argentina] 2012). This suggests that, in the context of the native forest areas, the *Ley de Bosques* may have had a small effect in reducing deforestation, but overall levels still remain quite high. This has led some to question the effectiveness of the law at slowing cutting (Valpreda 2012; Greenpeace Argentina 2013), and calls for a better understanding of the driving forces of deforestation in Argentina.

Soy and its effects

The increase of soybean in Argentina has occurred at a rapid pace throughout the last two decades, making it the third largest producer of soy in the world (US Foreign Agricultural Service 2013). Necessarily, as soy production rises, so does its spatial extent and the intensity of cultivation methods. Currently, almost all of Argentina's soy production is using genetically modified (GM) varieties, specifically Monsanto's "Roundup Ready" beans (Greenpeace International 2005). The highly mechanized and input intensive nature of this crop type calls into question other environmental consequences of soybean cultivation, such as pesticide runoff, glyphosate-resistant weeds, and soil depletion (Pengue 2005).

A number of studies have addressed soy and deforestation in Northwest Argentina, but only one has used methods capable of mapping crop types in deforested areas (Volante et al. 2005). However, this study by the Argentine *Instituto Nacional de Tecnología Agropecuraia* (National Institute of Agricultural Technology, INTA) does not have well-documented methodology and has not been updated since 2005. Of the remainder, all used remote sensing techniques to only classify LULC and not specific crop types, leaving the effect of soy on LULC as an underlying assumption (Grau, Gasparri, and Aide 2005; Grau, Aide, and Gasparri 2005; Grau, Gasparri, and Aide 2008; Boletta et al. 2006; Gasparri and Grau 2009). While the extreme deforestation in Argentina is undeniable—and certainly soy plays a part—its role has not been examined in full, leaving unsubstantiated the perception of soy as the driving force in this process.

The goal of this research is to develop an image classification capable of mapping agricultural crops by type, allowing soy to be explicitly identified on remotely sensed imagery. The accurate and efficient mapping of soy distributions and their changes over time could allow further investigation of the roles of soy in deforestation. The direct and indirect effects soy crops have had on deforestation can thus be understood conceptually and systemically at both regional and local scales, which could lead to the development of more effective policies for land management.

Study Area

This study will use agricultural areas in Kansas, USA for testing and verification of the phenological classification method and will apply the classification method to Pellegrini, Santiago del Estero, Argentina to test its effectiveness in subtropical South America.

Kansas, USA

The state of Kansas is one of the big agricultural producers of the US. As one of the plains states, it is relatively flat across much of its extent, making it well suited to large highly-mechanized agro-industrial operations. In 2012, the three most extensive crops in the state were wheat, corn, and soybeans (Table 1), which are also the most abundant crops in Pellegrini, Argentina. Additionally, Kansas has been the focus of a number of previous studies into the use of MODIS time-series for

Table 1: Most extensive crops in Kansas, 2012
(adapted from US Department of Agriculture 2013).

	Acreage (1,000 acres)	Production (1,000 units)
Wheat	9,100	382,200
Corn	3,950	379,200
Soy	3,810	83,820
All Hay	2,750	4,340
All Forage	2,750	4,545
Sorghum	2,100	81,900

crop classification (Wardlow and Egbert 2002, 2005; Wardlow, Egbert, and Kastens 2007; Wardlow and Egbert 2008), and has a very detailed and easily-accessible crop cover dataset in the form of the USDA CDL, making it a natural choice for a preliminary study area to test my method.

Pellegrini, Santiago del Estero, Argentina

Santiago del Estero, a province in Northwest Argentina, has an area of 136,351 square kilometers, about the same as Arkansas, but a population of about 874,000 (INDEC 2010b). The entire province is classified within the *Parque Chaqueño* (Chaco forest), but the forested area has declined rapidly in the past fifteen years. Over the period 1998 to 2002, 306,055 hectares were deforested (Secretaría de Ambiente y Desarrollo Sustentable [Argentina] 2007). From 2006 through 2011, a further 701,030 hectares of forest were lost, 283,669 of which were after the enacting of the OTBN (Secretería de Ambiente y Desarrollo Sustentable [Argentina] 2012). Over both of these time periods Santiago del Estero experienced the highest levels of deforestation in all of Argentina.

The Department of Pellegrini is an administrative area in the Northwest corner of Santiago del Estero (Fig. 1). The department has an area of 6,944 square kilometers, or slightly larger than the state of Delaware, and a 2010 population of only 20,514 (INDEC 2010a). The primary municipality of the department is Nueva Esperanza, with a population of about 4,500. The frontier nature of Pellegrini seems to have limited deforestation in the department for some time, but the push for land has increased the rate of deforestation. Over the years 2001 to 2005, only 5,968 hectares were found to be deforested (Volante 2005). From 2006 to 2011 the area deforested increased to 75,349 hectares, some 39,480 hectares cut after the enacting of the OTBN, a rate much higher than previously witnessed (Secretería de Ambiente y Desarrollo Sustentable [Argentina] 2012). Of the area cleared post-OTBN, 2,181 hectares were in red areas, the highest clearing of that designation in the nation. The vast majority of clearing, however, was 29,796 hectares in yellow areas. While Pellegrini's total deforestation during the period 2006 to 2011 was not the highest in Santiago del Estero, as both Moreno Department and Alberdi Department had higher total deforestation, as a percent of total land area Pellegrini's deforestation occurred at a greater rate: 10.85 percent of Pellegrini's land area was cleared versus 10.45 percent and 7.91 percent of Moreno and Alberdi, respectively.

Volante et al. (2005) found Pellegrini's primary summer crop over the years 2000 to 2005 to be soy, averaging about 40,000 hectares cultivated per year. Corn was the second most frequent crop, occupying about 7,500 hectares per year. Kidney beans were the third most popular, averaging a total cultivation of about 2,500 hectares per year. The primary winter crop was wheat, though cultivation varied wildly from less than 10,000 hectares in 2002 to over 31,000 hectares in 2004.

Data and Methods

The differentiation of crop types in remotely-sensed imagery is not a straightforward process. The use of a vegetation index (VI), such as the normalized difference vegetation index (NDVI) or the enhanced vegetation index (EVI), can help identify crops by their specific VI values in an image.

NDVI is a normalized ratio of the red and near-infrared bands, and can be expressed mathematically as:

$$NDVI = \frac{\rho_{\ NIR} - \rho_{\ red}}{\rho_{\ NIR} + \rho_{\ red}} \tag{1}$$

where $\rho_{\ NIR}$ and $\rho_{\ red}$ are the measured surface reflectance in their respective bands. As a ratio, the index minimizes multiplicative noise, but has issues with non-linearity and additive noise (Huete et al. 2002).

With advances in calibration, atmospheric correction, and other noise removal techniques which are integrated into the MODIS data processing workflow, a ratioing index is less necessary. The EVI was specifically developed for the MODIS platform to help correct some of the deficiencies of the NDVI. It has better sensitivity to high biomass, canopy structure, and leaf area, and less susceptibility to atmospheric degradation. EVI is calculated as:

$$EVI = G\frac{\rho_{\ NIR} - \rho_{\ red}}{\rho_{\ NIR} + C_1 \times \rho_{\ red} - C_2 \times \rho_{\ blue} + L} \tag{2}$$

Again, each ρ is the measured surface reflectance in the respective band, after complete or partial atmospheric correction. The blue band is used to ”subtract” aerosol effects from the red band. Additionally, four coefficients are introduced: G is the gain factor, C_1 and C_2 are used in the aerosol calculation, while L "is the canopy background adjustment that addresses nonlinear, differential NIR and red radiant transfer through a canopy" (Huete et al. 2002: 196). The values of these coefficients as used in the MODIS EVI calculation are 2.5, 6.0, 7.5, and 1.0, respectively.

Some crops, such as soy and sugarcane, have very different spectral reflectance throughout their development and maturation, however others, such as soy and corn, can have very similar reflective curves, leading to overlapping VI ranges. Such overlap can make it impossible to determine a crop type with specificity using traditional approaches. To combat this, a time series of images can be used to find VI values throughout a year to develop a classification based on annual phenology rather than a single-date image (Gu et al. 2010; Wardlow and Egbert 2002, 2005; Wardlow, Egbert, and Kastens 2007; Wardlow and Egbert 2008; Zhang et al. 2003).

In this study, I will use 16-day MODIS NDVI and EVI composite images to perform the phenological VI classification. One specific advantage of MODIS data is its temporal resolution; as the satellite passes any given location on a daily basis, the likelihood of getting enough cloud-free data to develop a phenologic model is significantly increased over other common platforms like Landsat Thematic Mapper (TM) and Landsat Operational Land Imager (OLI), which only have repeat coverage every sixteen days. MODIS data, however, comes at the price of a reduced spatial resolution of 250 meters compared to Landsat's 30-meter pixels.

Each of the 16-day composite VI images will be used as a band in a multi-date time-series image representing an entire agricultural year. The images are numbered by the day of the year (DOY) of the last date in the image, so an image from DOY 17 is the composite of the images from January 2 through January 17. In Kansas, an entire agricultural year can be captured from January 2 through the following January 1. Therefore, a Kansas time-series image with 16-day composite imagery would require 23 bands, band one being the composite image from January 17, with each succeeding band progressing every sixteen days through the end of the year (thus band 2 is DOY 33, band 3 is DOY 49, band 4 is DOY 65, etc.). Technically, following this pattern will make the last band of the image be from January 4 the following year, but the MODIS composite numbering ”resets” at the end of each year, and band 23 ends up being from January 1.

In Argentina, as it is in the Southern Hemisphere and the seasons are inverted to those of the Northern hemisphere, the growing season shifts, as must the date range for the VI time-series in order to capture crop phenologies unbroken. That is, the time-series image for Pellegrini must begin mid-calendar-year to adequately capture the annual phenologies. To accomplish this, the time-series images begin with the 16-day composite image from DOY 193 (July 12 in common years) and end with the image from the following DOY 177 (June 26 in common years).

Gu et al. outlined that phenological statistics regarding vegetation development can be derived from a MODIS VI time-series, including “start-of-season time (SOST), start-of-season NDVI (SOSN), end-of-season time (EOST), end-of-season NDVI (EOSN), maximum NDVI (MAXN), maximum NDVI time (MAXT), duration of season (DUR), amplitude of NDVI (AMP), and seasonal time integrated NDVI (TIN)” (2010: 529). A principal component analysis (PCA) can then be used to extract the meaningful variation in the data. Similarly, Wardlow, Egbert, et al. (Wardlow and Egbert 2002, 2005; Wardlow, Egbert, and Kastens 2007; Wardlow and Egbert 2008) showed that a decision tree classifier can be used to classify vegetation time-series data into increasingly refined categories until specific crop types are isolated and classified. By beginning with a basic land cover classification (e.g. forest, urban, agriculture), crops in the agriculture class can be broken down into winter and summer varieties using peaks in the vegetation index (winter wheat will peak earlier in the year than summer crops like corn and soy). Then, using training sites of known crop types defined by ground truth data, a final crop classification can be assigned by finding pixel values for key dates where like crops can be differentiated. That is, using the growing season in the Northern Hemisphere as an example, if from the training sites we know crop A has VI values between 0.7 and 0.8 on June 26 and between 0.5 and 0.6 on August 29, while crop B is between 0.55 and 0.65 and 0.75 and 0.85 on the same dates, pixels in the summer crop class can be assigned one of these types by testing their pixel values on these dates. While the authors found this method to have about an 85 percent overall accuracy (Wardlow and Egbert 2005), the downside of this method is that it requires training sites with previously-determined crop types to produce a classification, which can be time consuming and expensive to acquire.

Masialeti, Egbert, and Wardlow (2010) found that VI values from one year have a significant correlation with values from other years. Comparing the phenological curves of crops formed by the NDVI values from 2001 MODIS data (from Wardlow and Egbert 2005) with those from 2005 MODIS data, the authors found the overall shape of each crop's curve is maintained year-to-year, with subtle shifts in the beginning of the curve (earlier or later planting), the maximum of the curve, and the spread of the curve (a longer or shorter growing season), depending on weather and other external variables (Fig. 2). They surmised, with a means to account for these shifts of the curve, one could use VI values from one year to classify those from another.

Sakamoto et al. (2005, 2010) has shown that MODIS time-series data can be used to find key dates in a crop's phenology, enabling better crop management strategies. Specifically, the authors' two-step filter (TSF) method uses a wavelet transformation and a constrained minimization function to find reference curves for a specific crop's phenological development, and then fits that curve

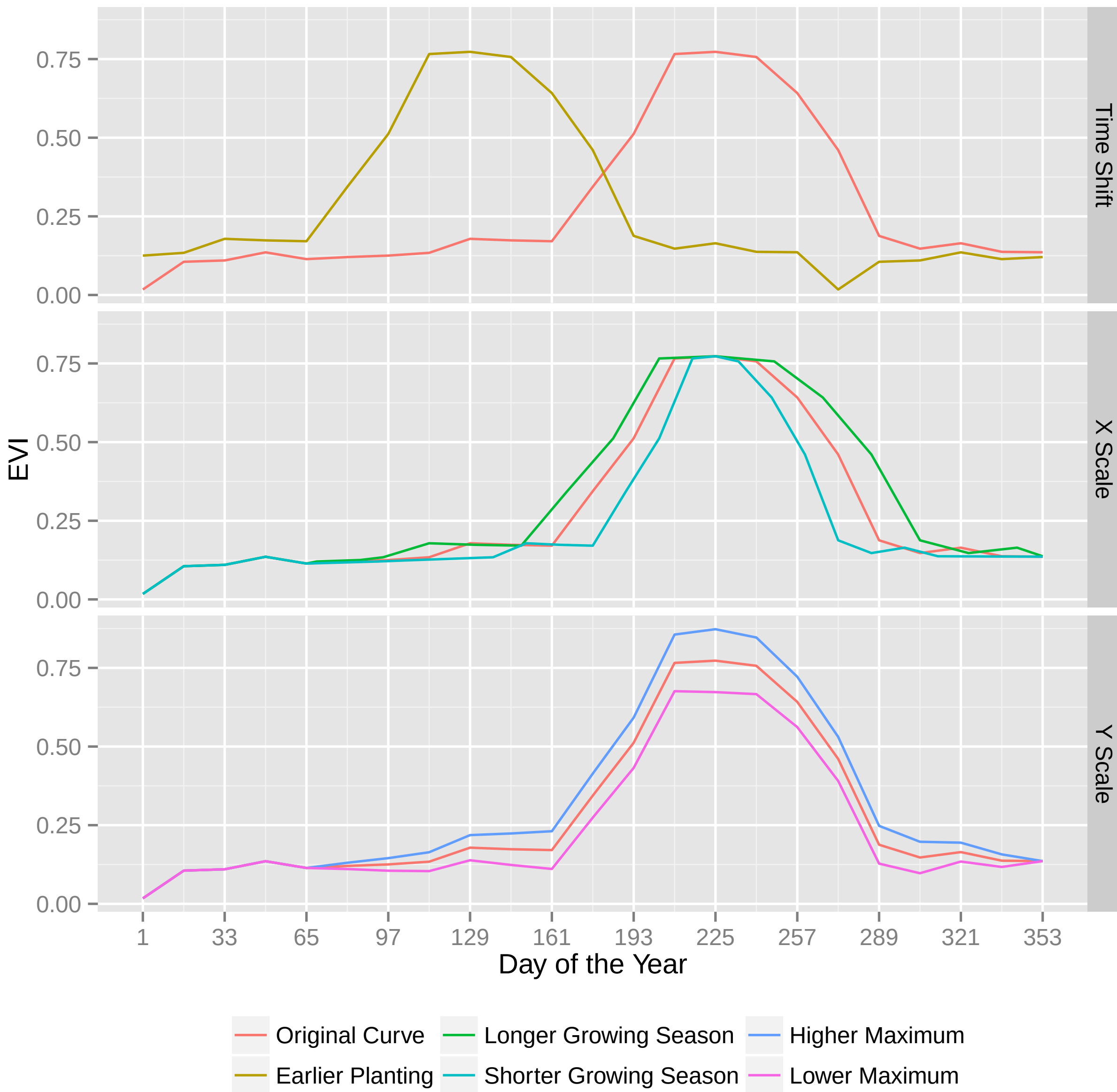


Figure 2: Examples of transformations of a crop's VI curve due to interannual variations in growing conditions. The original curve is that found for soy in the initial test area (see the Preliminary Results on pg. 18). The other curves were arbitrarily adjusted to illustrate each of the possible transformations.

to known pixels of that crop type to find the transition between developmental stages in the plants' growth. This TSF method demonstrates that reference curves can be fit to a pixel's values using a minimization function, accounting for the variations from the reference curve and the pixel curve. Therefore, unlike the previous multi-date VI classification approaches, which require trainings sites, this minimization method could be used without training sites to classify imagery by fitting previously-known references curves (i.e. not derived from ground truth data) to pixel values in a VI time-series; such is the foundation of the method I wish to test.

Drawing extensively from Sakamoto et al. (2010), I am in the process of developing said method. From page 2151:

$$RMSE = \left[\frac{1}{365/s} \sum_{x=d,d+s,d+2s\ldots}^{365} \left(f\left(x\right) - g\left(x\right)\right)^2\right]^{\frac{1}{2}} \tag{3}$$

where s is the interval of the imagery, d is the starting date of the imagery, $f(x)$ is the phenological curve for a given pixel in a dataset, and x is the DOY. $g(x)$ is given by:

$$g(x) = yscale \times \ h\left(xscale \times \left(x_0 + tshift\right)\right) \tag{4}$$

Here, $yscale$ and $xscale$ are coefficients controlling the vertical and horizontal scaling of a reference curve $h(x_0)$, and $tshift$ is a constant representing the horizontal shift, in days, of $h(x_0)$ (Fig. 2). x_0 is the day of year in the shape model. Thus, if we minimize eq. 3 bounding $yscale$, $xscale$, and $tshift$ in $g(x)$ with reasonable values for each, we can calculate how well a given reference curve $h(x_0)$ can be made to fit the pixel values $f(x)$. Comparing the minimum RMSE of each of the reference curves used allows us to assign a confidence score that a pixel is one of the crops we are seeking. Using Bayesian probability theory or Dempster-Shafer theory of evidence, a pixel's confidence score for each of the crop types will determine the final crop classification and uncertainty of the classification (Jiang and Eastman 2000).

As this is a new approach to crop classification, a variety of variables need to be tested to see how they impact the resulting classification, including:

- The spatial distribution of pixels chosen to create the reference curves.
- The temporal distribution of pixels chosen to create the reference curves.
- The VI used for the classification.

To test these factors, I will iteratively run the analysis using five small sample areas dispersed across Kansas. With the USDA CDL as ground truth reference, I will test the classification in each of the sample areas using reference curves derived from one, two, three, and four of the sample areas, to see if multiple sites being averaged increases the accuracy of the reference curves, or introduces noise due to geographical discrepancies in season start, maximum intensity, or season length. Similarly, I will add multiple years of data to each sample location, to see if averaging curves over those years has a positive or negative effect on the classification results. Lastly, I will perform all of theses tests twice, once with MODIS NDVI data, and again with MODIS EVI data. From this testing, I can determine the best method for deriving crop reference curves, and use the reference curves from that method and apply them to classifying the data from my study area in Argentina.

The accuracy of the Argentina study area classification will need to be assessed, which will require ground truth data to verify whether the classification identified the pixels correctly. Unfortunately, data like the USDA CDL does not exist for Argentina. In order to acquire the necessary ground truth, I will create a set of control points using a stratified random approach. This will provided me with a random sample of points throughout each of the different classes identified (i.e.

corn, soy, wheat, not crops). Then, I will visit each of these points to determine the crop cover or land cover. I will use this ground truth data to construct a confusion matrix in order to check the producer, user, and overall accuracies of the classification.

Research Timeline

I hope to have finished the classification algorithm by the beginning of winter term 2014. At that point, I will run the testing on the Kansas sample areas to find the best reference curves. Once that testing is completed, I will use those curves to classify MODIS data of Pellegrini from the 2012 – 2013 growing season. This will provide me with a rough picture of the crop cover my method will find for 2013 – 2014, and from this I can use a stratified random sampling technique to better ensure my ground truthing will contain sufficient sample points in each land cover class. In March 2014 I plan to do fieldwork in Pellegrini to collect the ground truth data for these points. I have chosen the month of March specifically because it is in the middle of the summer growing season, and I can get ground truth data for summer crop identification as the crops are growing. Winter crops, such as winter wheat, will need to be verified by inquiring with field owners, or confirmed by visual interpretation of Landsat imagery. In spring term 2014, I will classify 2013 – 2014 data for Pellegrini, calculate the accuracy using the ground truth acquired in March, and begin writing. I plan to finish my work and writing over summer 2014 and present fall 2014.

Preliminary Results

To this point, I have been able to create scripts to import and assemble my multi-date images; to sample images and generate mean reference curves of the VI values for soy, corn, and wheat (Fig.

Table 2: Accuracy assessment of the initial results.

	Corn	Soy	Wheat	Other	Total	User Accuracy
Corn	260	22	6	103	391	67%
Soy	10	59	1	29	99	60%
Wheat	33	0	354	127	514	69%
Other	174	27	241	670	1112	60%
Total	477	108	602	929	2116	
Producer Accuracy	55%	55%	59%	72%		
						Overall: 63%
						Kappa: 0.443

3; and to take said reference curves and generate an image for each crop, of which the pixel values are the RMSE after constrained minimization, and an image with the best fit for each pixel (Fig. 4). For script development I am using MODIS 16-day EVI data from 2012. I have not completed the computation of the confidence scores of any given classification (a la fuzzy classification), but, for the sake of exploring this initial output, I used a threshold of .08 RMSE and classified the image according to the best fit. That is, for any pixels which had an RMSE of .08 or less for one or more of the crops tested, I used the best fit image values for those pixels as the values for classification (Fig. 5a). I used the USDA CDL for 2012, resampled to 250 meter pixels with the majority value (Fig. 5b), to check my classification. Building a confusion matrix for all of the pixels in the image resulted in Table 2. As shown, the accuracies for each of the crops ranged between 60 percent and 70 percent, with an overall accuracy of 63 percent, though the kappa value is somewhat low at 0.44. Nonetheless, considering this is the unoptimized algorithm with an arbitrary threshold value, I believe my results suggest this classification method to have potential.

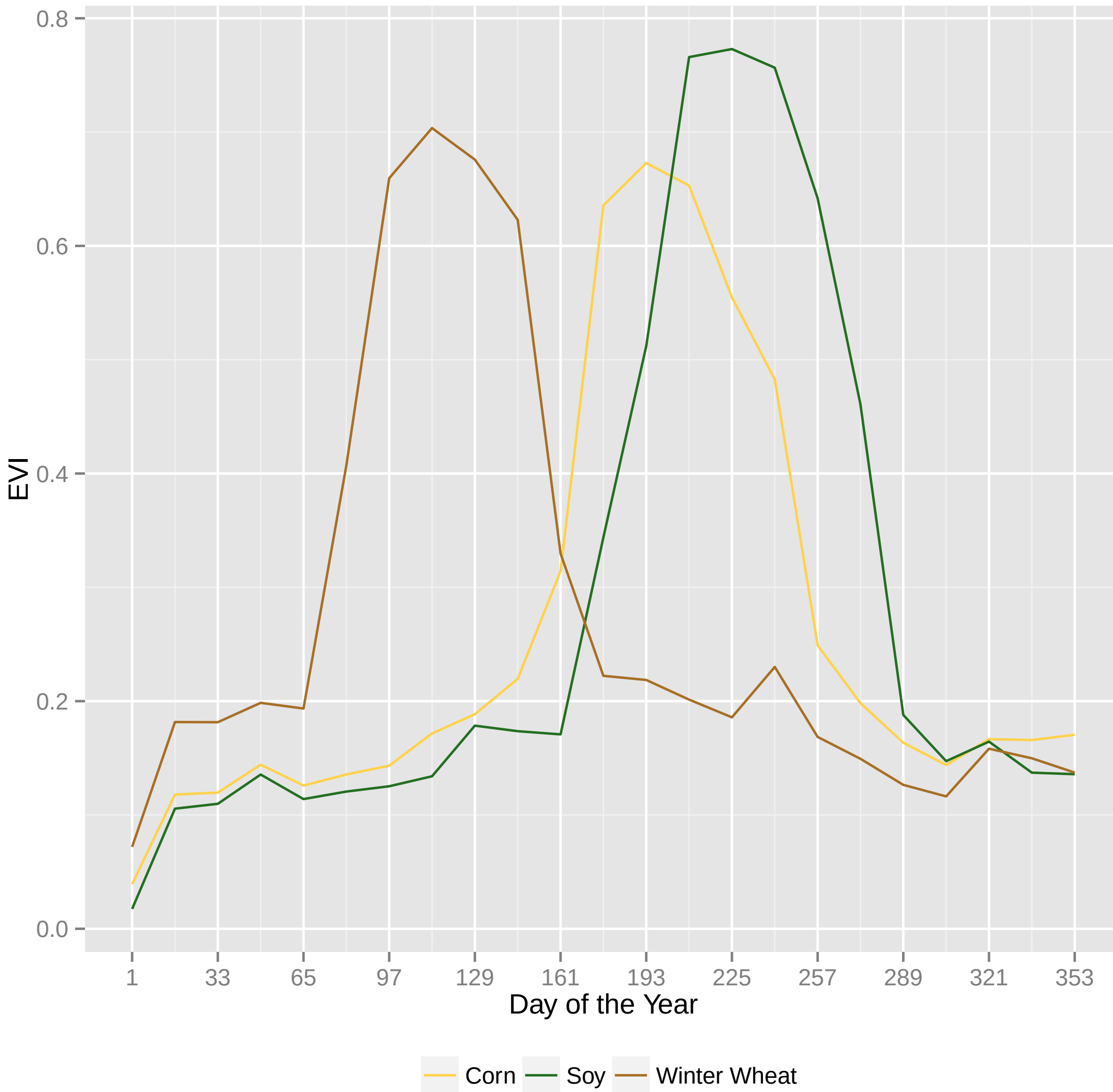
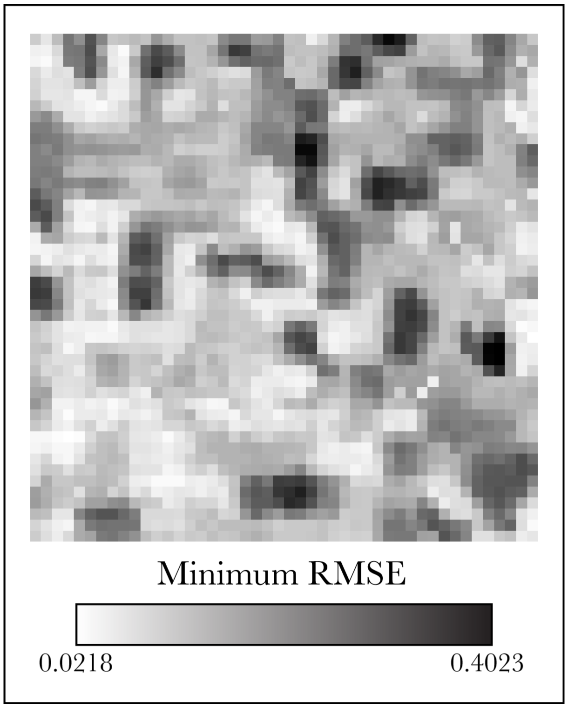
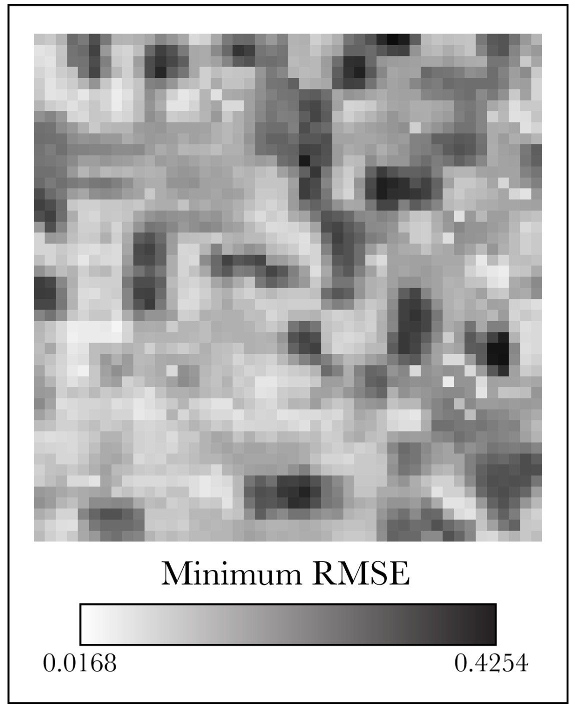


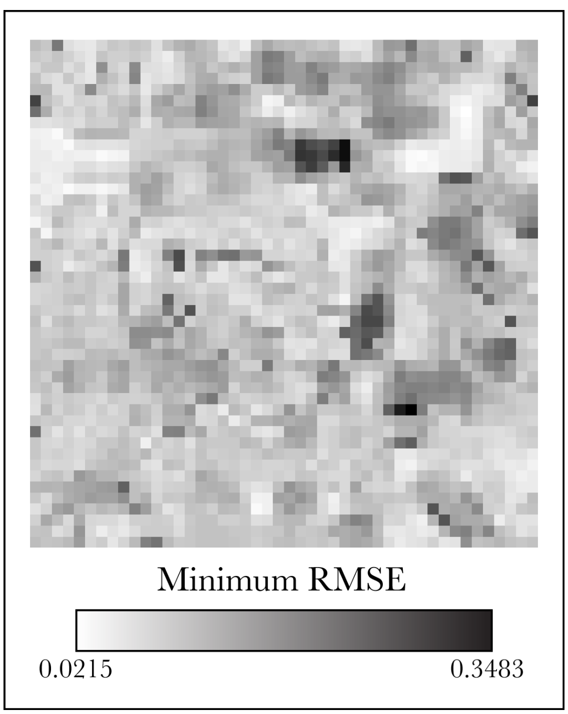
Figure 3: Mean phenological curves found for corn, soy, and winter wheat. Each curve is generated from the mean EVI values of four pixels of the respective crop from an initial test area in 2012 Kansas data.



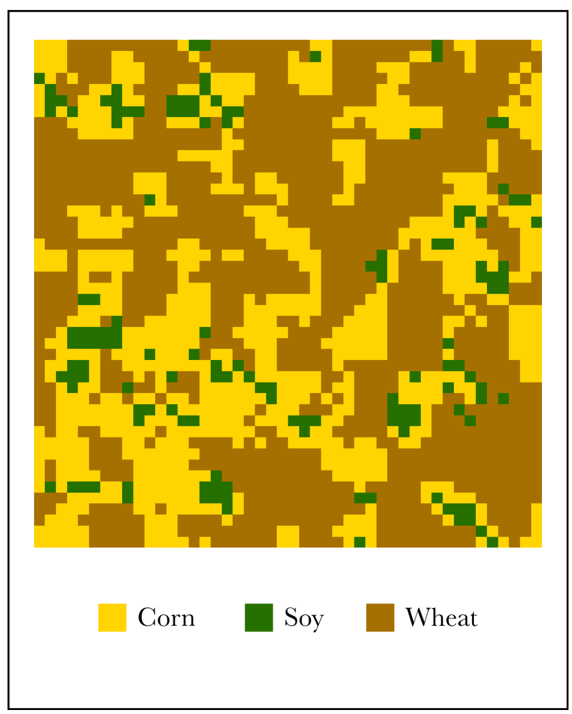
(**a**) Fit of corn reference curve.



(**b**) Fit of soy reference curve.

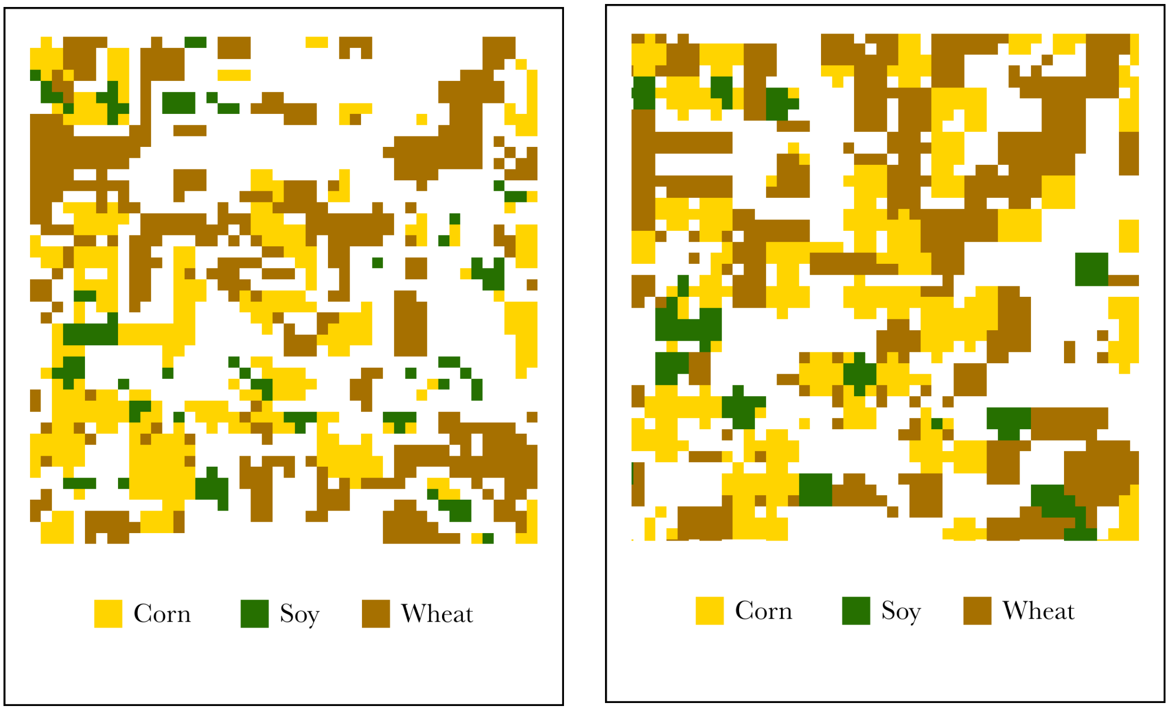


(**c**) Fit of wheat reference curve.



(**d**) Best fit image.

Figure 4: Initial test results.



(**a**) Initial classification of corn, soy, and wheat.

(**b**) 2012 USDA CDL, resampled to 250m pixels to match MODIS data.

Figure 5: Initial classification and ground truth

Anticipated Outcomes

Upon the completion of my research, I expect to have a working tool which can be used as an economical and effective means of crop classification. With the results of my testing, I will know how changes in the generation of the reference curves effects the results of the classifier. Using this tool and knowledge, I will generate crop maps of my study areas, and quantify their accuracies. The results of this study should be of value to those working in the field of remote sensing and those investigating LULC change, particularly in regard to deforestation and agriculture. I hope this work will be the basis for future investigations into soy's role in Argentina's deforestation.

Appendix 1 – Project Code

The code for this project is all written in python version 2.7.5, and has various dependencies, listed below:

- GDAL 1.8 with python bindings
- Scipy 0.12.0
- Numpy 1.6.2

The project is currently split into three scripts: one to import the VI images and build the composite multi-date VI image (Script 1); one to extract the mean reference curves from pixels of known crop type in an image (Script 2, page 27); and one to use minimization to fit said reference curves to pixel curves, writing the RMSE to output images and finding the best fit (Script 3, page 28).

Script 1: Building multi-date images

```
from osgeo import gdal
from osgeo.gdalconst import *
import os, sys

rootDIR = ”/Users/phoetrymaster/Documents/School/Geography/Thesis/Data/MODIS_KANSAS_2012/”
#rootDIR = ”/Users/phoetrymaster/Documents/School/Geography/Thesis/Data/MODIS 7_2012-2013/
outName = ”test”

newfoldername = ”kansas”

find = ”EVI”
ext = ”.hdf”
drivercode = ”ENVI”
ndvalue = -3000
projection = ”PROJCS[\”Sinusoidal\”,GEOGCS[\”GCS_Undefined\”,DATUM[\”D_Undefined\”,SPHEROI


########## METHODS ##########
```

```


def find_files(searchdir, ext):
    hdfs = []

    for root, dirs, files in os.walk(searchdir):
        for f in files:
            if f.upper().endswith(ext.upper()):
                foundfile = os.path.join(root, f)
                hdfs.append(foundfile)

    return hdfs


def create_output_dir(root, name):
    dirpath = os.path.join(root, name)

    if os.path.isdir(dirpath):
        count = 1
        dirpath_ = dirpath + ”_”
        while 1:
            dirpath = dirpath_ + str(count)
            count += 1
            if not os.path.isdir(dirpath):
                break

    os.makedirs(dirpath)
    return dirpath


def create_output_raster(outFile, cols, rows, bands, datatype, drivername=”GTiff”):
    driver = gdal.GetDriverByName(drivername)
    driver.Register()

    outds = driver.Create(outFile, cols, rows, bands, datatype)

    return outds


def get_output_params(filepath):
    image = gdal.Open(filepath, GA_ReadOnly)

    if image is None:
        raise Exception(”Could not open ” + filepath)
```

```

    rows = image.RasterYSize
    cols = image.RasterXSize
    band = image.GetRasterBand(1)
    bandtype = band.DataType
    geotransform = image.GetGeoTransform()
    projection2 = image.GetProjection()

    image = ""

    return rows, cols, bandtype, geotransform, projection


def get_hdf_subdatasets(hdfpath):
    hdf = gdal.Open(hdfpath, GA_ReadOnly)

    if hdf is None:
        raise Exception("Could not open " + hdfpath)

    sds = []
    hdfsds = hdf.GetSubDatasets()

    for data in hdfsds:
        sds.append((data[0], data[0].split(" ")[-1]))

    hdf = ""

    return sds


def main():
    outdir = create_output_dir(rootDIR, newfoldername)
    print "\nOutputting files to : {0}".format(outdir)

    print "\nFinding HDF files in directory/subfolders: {0}".format(rootDIR)
    hdfs = find_files(rootDIR, ext)
    print "\tFound {0} files.".format(len(hdfs))

    print "\nGetting images to process of type {0}...".format(find)
    toprocess = []

    for hdf in hdfs:
        sds = get_hdf_subdatasets(hdf)
        for ds in sds:
```

```
            if find.upper() in ds[1].upper():
                toprocess.append(ds[0])
                print "\t\t{0}".format(ds[0])

    bands = len(toprocess)
    print "\tFound {0} images of type {1}.".format(bands, find)

    print "\nGetting output parameters..."
    rows, cols, datatype, geotransform, projection = get_output_params(toprocess[0])
    print "\tParameters: rows: {0}, cols: {1}, datatype: {2}, projection: {3}.".format(row

    outfile = os.path.join(outdir, outName) + ".tif"
    print "\nOutput file is: {0}".format(outfile)

    outds = create_output_raster(outfile, cols, rows, bands, datatype, drivername=driverco
    print "\tCreated output file."

    print"\nAdding bands to output file..."
    for i in range(0, bands):
        print "\tProcessing band {0} of {1}...".format(i + 1, bands)
        image = gdal.Open(toprocess[i])
        band = image.GetRasterBand(1)

        outband = outds.GetRasterBand(i + 1)

        print "\t\tReading band data to array..."
        data = band.ReadAsArray(0, 0, cols, rows)

        print "\t\tWriting band data to output band..."
        outband.WriteArray(data, 0, 0)
        outband.SetNoDataValue(ndvalue)
        outband.FlushCache()

        del data, outband
        image = ""

    print "\tFinished adding bands to output file."

    print "\nSetting transform and projection..."
    outds.SetGeoTransform(geotransform)
    outds.SetProjection(projection)

    outDS = ""

```

```
    print ”\nProcess completed.”


########## PROCEDURE ##########


if __name__ == ’__main__’:
    sys.exit(main())
```

Script 2: Getting reference curves from known pixels

```
from osgeo import gdal
from osgeo.gdalconst import *
from math import floor

imagepath = ”/Volumes/J_KEIFER/Thesis/Data/ARC_Testing/test1.dat”

soylocs = [(6002, 2143), (5944, 2102), (5746, 2183), (5998, 2171)]
cornlocs = [(5997, 2139), (5940, 2096), (6051, 2230), (5691, 1998)]
wheatlocs = [(5993, 2136), (5937, 2080), (5935, 2076), (5921, 2217)]
refstoget = {”soy”: soylocs, ”corn”: cornlocs, ”wheat”: wheatlocs}


gdal.AllRegister()

img = gdal.Open(imagepath, GA_ReadOnly)

if img is None:
    raise Exception(”Could not open ” + imagepath)

bands = img.RasterCount

print ”Found {0} bands in input image.”.format(bands)

refs = {}

for key, val in refstoget.items():
    print ”Processing {0} coordinates:”.format(key)
    dict = {}
    for i in range(0, bands):
        band = img.GetRasterBand(i+1)
        print ”\tProcessing band {0}”.format(i+1)
        values = []
```

```
        for loc in val:
            print ”\t\tGetting position {0}”.format(loc)
            values.append(int(band.ReadAsArray(int(floor(loc[0])), int(floor(loc[1])), 1,
        dict[(i*16+1)] = sum(values) / float(len(values))
        band = ””
    refs[key] = dict

print refs

img = ””
```

Script 3: Using reference curves to calculate fit for each pixel

```
from osgeo import gdal
from osgeo.gdalconst import *
import os
import numpy
from numpy import sum
from scipy import interpolate
from scipy import optimize

gdal.UseExceptions()

imagepath = ”/Users/phoetrymaster/Documents/School/Geography/Thesis/Data/ARC_Testing/ClipT
rootdir = ”/Users/phoetrymaster/Documents/School/Geography/Thesis/Data/OutImages/”

newfoldername = ”Testing”

drivercode = ’ENVI’
ndvalue = -3000

startDOY = 1
thresh = 500
bestguess = 0
fitmthd = ’SLSQP’


refs = {
    ’soy’: {1: 174.5, 97: 1252.25, 65: 1139.5, 209: 7659.0, 273: 4606.75, 337: 1371.75, 17
            49: 1355.25,
            129: 1784.75, 257: 6418.0, 321: 1644.5, 305: 1472.75, 193: 5119.75, 289: 1878.
            81: 1205.5, 225: 7729.75, 145: 1736.25, 161: 1708.25, 353: 1358.25, 113: 1340.
    ’corn’: {1: 392.25, 97: 1433.25, 65: 1258.5, 209: 6530.0, 273: 1982.5, 337: 1658.5, 17
```

```
             49: 1441.25, 129: 1885.25, 257: 2490.25, 321: 1665.75, 305: 1439.0, 193: 6728
             177: 6356.75,
             241: 4827.25, 81: 1355.75, 225: 5547.5, 145: 2196.5, 161: 3143.25, 353: 1704.
    'wheat': {1: 719.75, 97: 6594.75, 65: 1935.25, 209: 2013.5, 273: 1493.5, 337: 1498.25,
              49: 1985.25, 129: 6758.0, 257: 1685.75, 321: 1582.5, 305: 1163.25, 193: 2186
              241: 2301.0, 81: 4070.5, 225: 1858.0, 145: 6228.5, 161: 3296.5, 353: 1372.5,
}


########## METHODS ##########


def create_output_raster(outFile, cols, rows, bands, datatype, drivername="GTiff"):
    driver = gdal.GetDriverByName(drivername)
    driver.Register()

    outds = driver.Create(outFile, cols, rows, bands, datatype)

    return outds


def create_output_dir(root, name):
    dirpath = os.path.join(root, name)

    if os.path.isdir(dirpath):
        count = 1
        dirpath_ = dirpath + "_"
        while 1:
            dirpath = dirpath_ + str(count)
            count += 1
            if not os.path.isdir(dirpath):
                break

    os.makedirs(dirpath)
    return dirpath


def get_sort_dates_values(vals, threshhold=-3000):
    """Gets the DOY dates (the keys) in a list from dictonary vals and sorts those, placin
    (list x0). Then the function iterates over these values and gets the corresponding val
    they are lower than an optional threshhold value (-3000 default = NoData in MODIS imag
    the list y. x and y are then returned."""

    x = vals.keys()
```

```
    x.sort()
    y = []

    for i in x:
        if vals[i] < threshhold:
            y.append(threshhold)
        else:
            y.append(vals[i])

    return x, y

def find_fit(valsf, valsh, bestguess, threshhold, mthd="TNC"):

    x0, y0 = get_sort_dates_values(valsf, threshhold=threshhold)
    x1, y1 = get_sort_dates_values(valsh)

    tck = interpolate.splrep(x1, y1)

    fun = lambda x: ((1 / 22.8125 * sum(
        (valsf[i] - (x[0] * interpolate.splev((x[1] * (i + x[2])), tck))) ** 2 for i in x0
                         1. / 2))

    bnds = ((0.6, 1.4), (0.6, 1.4), (-10, 10))

    res = optimize.minimize(fun, (1, 1, bestguess), method=mthd, bounds=bnds)

    return res.fun, res.x, res.message


########## PROCEDURE ##########


try:
    outdir = create_output_dir(rootdir, newfoldername)
    print "\nOutputting files to : {0}".format(outdir)

    gdal.AllRegister()

    #Open multi-date image to analyze
    img = gdal.Open(imagepath, GA_ReadOnly)

    if img is None:
        raise Exception("Could not open: {0}".format(imagepath))

```

```
#Get image properties
cols = img.RasterYSize
rows = img.RasterXSize
bands = img.RasterCount
geotransform = img.GetGeoTransform()
projection = img.GetProjection()

print "Input image dimensions are {0} columns by {1} rows and contains {2} bands.".for


#Create output rasters for each crop type to hold residual values from fit and arrays
print "\nCreating output files..."
outfiles = {}
outdatasets = {}
outarrays = {}
for key in refs:
    outfile = os.path.join(outdir, key) + ".tif"
    outfiles[key] = create_output_raster(outfile, cols, rows, 1, GDT_Float32, driverna
    outfiles[key].SetGeoTransform(geotransform)
    outfiles[key].SetProjection(projection)
    outdatasets[key] = outfiles[key].GetRasterBand(1)
    outarrays[key] = numpy.zeros(shape=(rows, cols))
    print "\tCreated file: {0}".format(outfile)

#Create output raster for bestFit
outfile = os.path.join(outdir, "bestFit") + ".tif"
fitimgfile = create_output_raster(outfile, cols, rows, 1, GDT_Byte, drivername=driverc
fitimgfile.SetGeoTransform(geotransform)
fitimgfile.SetProjection(projection)
fitimg = fitimgfile.GetRasterBand(1)
fitarray = numpy.zeros(shape=(rows, cols))
print "\tCreated file: {0}".format(outfile)


#Iterate through each pixel and calculate the fit for each ref curve; write residuals
for row in range(0, rows):
    for col in range(0, cols):
        valsf = {}
        print "Pixel r:{0}, c:{1}:".format(row, col)
        for i in range(0, bands):
            band = img.GetRasterBand(i+1)
            measured = int(band.ReadAsArray(col, row, 1, 1))
            valsf[startDOY + i*16] = measured
        count = 1
```

```
            fit = {}
            for key, val in refs.items():
                res, transforms, message = find_fit(valsf, val, bestguess, threshhold=thre
                outarrays[key][row, col] = res
                print "\t{0}: {1}, {2}, {3}".format(key, res, transforms, message)
                fit[res] = count
                count += 1
            fitarray[row, col] = fit[min(fit.keys())]

    #Write output array values to files
    print "Writing output files..."
    for key, values in outdatasets.items():
        outdatasets[key].WriteArray(outarrays[key], 0, 0)

    fitimg.WriteArray(fitarray, 0, 0)
    print "\nProcess finished."

except Exception as e:
    print e

finally:
    print "\nClosing files..."
    try:
        fitimg = None
        fitimgfile = None
    except:
        pass
    try:
        for key, value in outdatasets.items():
            outdatasets[key] = None
    except:
        pass
    try:
        for key, value in outfiles.items():
            outfiles[key] = None
    except:
        pass
```

References

Altieri, M. A., and W. Pengue. 2006. GM soybean: Latin America's new colonizer. *Seedling* (Jan.): 13–17.

Boletta, P. E., A. C. Ravelo, A. M. Planchuelo, and M. Grilli. 2006. Assessing deforestation in the Argentine Chaco. *Forest ecology and management* 228:108–114.

Bonnie, R. 2000. Counting the cost of deforestation. *Science* 288 (5472): 1763–1764.

Gasparri, N. I., and H. R. Grau. 2009. Deforestation and fragmentation of Chaco dry forest in Northwest Argentina. *Forest ecology and management* 258 (6): 913–921.

Geist, H. J., and E. F. Lambin. 2002. Proximate causes and underlying driving forces of tropical deforestation. *BioScience* 52 (2): 143–150.

Grau, H. R., T. M. Aide, and N. I. Gasparri. 2005. Globalization and soybean expansion into semiarid ecosystems of Argentina. *AMBIO: A journal of the human environment* 34 (3): 265–266.

Grau, H. R., N. I. Gasparri, and T. M. Aide. 2005. Agriculture expansion and deforestation in seasonally dry forests of Northwest Argentina. *Environmental conservation* 32 (2): 140.

———. 2008. Balancing food production and nature conservation in the neotropical dry forests of northern Argentina. *Global change biology* 14 (5): 985–997.

Greenpeace International. 2005. The expanding soybean frontier: Argentina's dangerous reliance on genetically engineered soybean. *Greenpeace briefing* (Jan.). Available at: http://www.greenpeace.org/international/Global/international/planet-2/report/2005/11/the-expanding-soybean-frontier.pdf (last accessed: 27 Oct. 2013).

Greenpeace Argentina. 2013. Ley de Bosques: 5 años con pocos avances (Jan.). Available at: http://www.greenpeace.org/argentina/Global/argentina/report/2013/bosques/Ley%20de%20Bosques.%205%20a%C3%83%C2%B1os%20con%20pocos%20avances%20FINAL.pdf.

Gu, Y., J. F. Brown, T. Miura, W. J. Van Leeuwen, and B. C. Reed. 2010. Phenological classification of the United States: A geographic framework for extending multi-sensor time-series data. *Remote sensing* 2 (2): 526–544.

Gulezian, S. E. 2009. Environmental politics in Argentina: The Ley de Bosques. Honors thesis, University of Vermont.

Houghton, R. A. . A. 1994. The Worldwide Extent of Land-Use Change. *BioScience* 44 (5): 305–313.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote sensing of environment* 83 (1): 195–213.

Instituto Nacional de Estadística y Censos. 2010a. Provincia de Santiago del Estero, departamento Pellegrini: Población total por sexo e índice de masculinidad, según edad en años simples y grupos quinquenales de edad, año 2010 [Department of Pellegrini, Santiago del Estero Province: Total population, sex, and index of masculinity, by age in single years and groups of five]. *Censo nacional de población, hogares y viviendas 2010.* Available at: http://www.censo2010.indec.gov.ar/CuadrosDefinitivos/P2-D_86_133.pdf (last accessed: 14 May 2013).

———. 2010b. Provincia de Santiago del Estero: Población total por sexo e índice de masculinidad, según edad en años simples y grupos quinquenales de edad, año 2010 [Santiago del Estero Province: Total population, sex, and index of masculinity, by age in single years and groups of five]. *Censo nacional de población, hogares y viviendas 2010.* Available at: http://www.censo2010.indec.gov.ar/CuadrosDefinitivos/P2-P_Santiago_del_estero.pdf (last accessed: 14 May 2013).

Jiang, H., and J. R. Eastman. 2000. Application of fuzzy measures in multi-criteria evaluation in GIS. *International journal of geographical information science* 14 (2): 173–184.

Masialeti, I., S. Egbert, and B. D. Wardlow. 2010. A comparative analysis of phenological curves for major crops in Kansas. *GIScience & remote sensing* 47 (2): 241–259.

Pengue, W. A. 2005. Transgenic crops in Argentina: The ecological and social debt. *Bulletin of science, technology & society* 25 (4): 314–322.

Sakamoto, T., B. D. Wardlow, A. A. Gitelson, S. B. Verma, A. E. Suyker, and T. J. Arkebauer. 2010. A two-step filtering approach for detecting maize and soybean phenology with time-series MODIS data. *Remote sensing of environment* 114 (10): 2146–2159.

Sakamoto, T., M. Yokozawa, H. Toritani, M. Shibayama, N. Ishitsuka, and H. Ohno. 2005. A crop phenology detection method using time-series MODIS data. *Remote sensing of environment* 96 (3-4): 366–374.

Sala, O. E., F. S. Chapin, J. J. Armesto, E. Berlow, J. Bloomfield, R. Dirzo, E. Huber-Sanwald, L. F. Huenneke, R. B. Jackson, and A. Kinzig. 2000. Global biodiversity scenarios for the year 2100. *Science* 287 (5459): 1770–1774.

Secretaría de Ambiente y Desarrollo Sustentable [Argentina]. 2007. Informe sobre deforestación en Argentina [Bulletin on deforestation in Argentina]. Available at: http://www.ambiente.gov.ar/archivos/web/UMSEF/File/deforestacin_argentina_v2.pdf (last accessed: 29 Oct. 2013).

Secretaría de Desarrollo Sustentable y Política Ambiental [Argentina]. 2001. *Primer inventario nacional de bosques nativos [first national inventory of native forests]*. Available at: http://aplicaciones.medioambiente.gov.ar/archivos/web/UMSEF/File/PINBN/nueva_version_manuales/pinbn_manual_cartografia_sig.pdf (last accessed: 13 May 2013).

Secretería de Ambiente y Desarrollo Sustentable [Argentina]. 2012. Monitoreo de la superficie de bosque nativo de la República Argentina, período 2006-2011: Regiones forestales Parque Chaqueño, Selva Misionera, y Selva Tucumano Boliviana [Monitoring of the land cover of the native forests of the Argentine Republic, Period 2006-2011: Forest regions of the Chaco, Misionera Forest, and Tucumano-Boliviana Forest]. Available at: http://www.ambiente.gov.ar/archivos/web/UMSEF/file/LeyBN/monitoreo_bn_2006_2011_ley26331.pdf (last accessed: 27 Oct. 2013).

US Department of Agriculture. 2013. Crops: Kansas agricultural statistics. *USDA national agriculture staticstics service monthly crop production reports* (Jan.). Available at: http://www.nass.usda.gov/Statistics_by_State/Kansas/Publications/Crops/Production/2013/cropjan.pdf (last accessed: 25 Nov. 2013).

US Foreign Agricultural Service. 2013. World agricultural production. *World agricultural production circular series* (Apr.). Available at: http://www.fas.usda.gov/psdonline/circulars/production.pdf (last accessed: 27 Oct. 2013).

Valpreda, J. R. 2012. The protection of natural forests in Argentina: Effective actions or words on paper? *INK-ideas, numbers, and knowledge* 1 (1). Available at: http://ink-journal.com/index.php/ink/article/view/9#.Unr-s5TF2fu (last accessed: 6 Nov. 2013).

Volante, J. N., H. P. Paoli, A. R. Bianchi, Y. E. Noe, and H. J. Elena. 2005. Análisis de la dinámica del uso del suelo agrícola del noroeste Argentino mediante teledetección y SIG. Período 2000-2005. [Analysis of the dynamics of the use of agricultural land in Northwest Argentina using remote sensing and GIS. Period 2000-2005.] Available at: http://inta.gob.ar/documentos/analisis-de-la-dinamica-del-uso-del-suelo-agricola-del-noroeste-argentino-mediante-teledeteccion-y-sig.-periodo-2000-2005/at_multi_download/file/Analisis_de_la_dinamica_del_uso_del_suelo.pdf (last accessed: 13 May 2013).

Wardlow, B. D., and S. L. Egbert. 2002. Discriminating cropping patterns for the US Central Great Plains region using time-series MODIS 250-meter NDVI data–Preliminary results. In *Pecora 15 and land satellite information IV conference,* 10–15.

———. 2005. State-level crop mapping in the US Central Great Plains agroecosystem using MODIS 250-meter NDVI data. In *Pecora 16 symposium,* 25–27.

———. 2008. Large-area crop mapping using time-series MODIS 250m NDVI data: An assessment for the U.S. Central Great Plains. *Remote sensing of environment* 112 (3): 1096–1116.

Wardlow, B. D., S. L. Egbert, and J. Kastens. 2007. Analysis of time-series MODIS 250m vegetation index data for crop classification in the U.S. Central Great Plains. *Remote sensing of environment* 108 (3): 290–310.

Zak, M. R., M. Cabido, and J. G. Hodgson. 2004. Do subtropical seasonal forests in the Gran Chaco, Argentina, have a future? *Biological conservation* 120 (4): 589–598.

Zhang, X., M. A. Friedl, C. B. Schaaf, A. H. Strahler, J. C. Hodges, F. Gao, B. C. Reed, and A. Huete. 2003. Monitoring vegetation phenology using MODIS. *Remote sensing of environment* 84 (3): 471–475.