NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

MASTER THESIS

Statistical techniques for improving prediction in crop progress stages with meteorological and satellite data

---

Author:
Christina GIANNADAKI

Supervisor:
Dr. Samis TREVEZAS

Department of Mathematics

December 18, 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

Abstract

Department of Mathematics

Master In Statistical And Operational Research

Statistical techniques for improving prediction in crop progress stages with meteorological and satellite data

by Christina GIANNADAKI

Crop Progress Reports (CPRs) of the USDA are listing the weekly progress made in the different phenological stages of selected crops and in particular of corn. In this thesis, our goal was to predict the CPRs of a full year by taking into account available data from related features in a way that we can beat the predictions based on empirical means from historical data. For this reason, we used two features, the mean Normalized Difference Vegetation Index (NDVI) and the Accumulated Growing Degree Days (AGDDs). In order to achieve our target we implemented several modeling approaches, including Independent Mixture Models and Hidden Markov Models HMMs and we compared different type of estimators and predictors by taking into account both features or treating them separately, or making data transformations, such as differences. The results showed that the aforementioned models cannot predict better than the historical data. Finally, we managed to obtain better predictions by using Simple Linear Regression. This study can be extended in several directions for future work.

Περίληψη

Στατιστικές τεχνικές για τη βελτίωση της πρόβλεψης στα στάδια προόδου της καλλιέργειας με μετεορολογικά και δορυφορικά δεδομένα

Οι εκθέσεις προόδου της καλλιέργειας (CPR) του USDA παρουσιάζουν την εβδομαδιαία πρόοδο που σημειώθηκε στα διάφορα φαινολογικά στάδια των επιλεγμένων καλλιεργειών και ιδιαίτερα του καλαμποκιού. Σε αυτή την διπλωματική, ο στόχος μας ήταν να προβλέψουμε τα CPR ενός ολόκληρου έτους λαμβάνοντας υπόψη διαθέσιμα δεδομένα από συναφή χαρακτηριστικά με τρόπο που να μπορούμε να νικήσουμε τις προβλέψεις βάσει εμπειρικών μέσων από ιστορικά δεδομένα. Για το λόγο αυτό, χρησιμοποιήσαμε δύο χαρακτηριστικά, τον δείκτη κανονικοποιημένης βλάστησης (NDVI) και τις συγκεντρωτικές ημέρες καλλιέργειας (AGDDs). Προκειμένου να επιτευχθεί ο στόχος μας, εφαρμόσαμε αρκετές προσεγγίσεις μοντελοποίησης, συμπεριλαμβανομένων μοντέλων ανεξάρτητων μήξεων και κρυμμένα μοντέλα HMMs και συγκρίναμε διαφορετικούς τύπους εκτιμητών και προγνωστικών λαμβάνοντας υπόψη και τα δύο χαρακτηριστικά ή τη χωριστή επεξεργασία τους ή πραγματοποιώντας μετασχηματισμούς δεδομένων, όπως διαφορές. Τα αποτελέσματα έδειξαν ότι τα προαναφερθέντα μοντέλα δεν μπορούν να προβλέψουν καλύτερα από τα ιστορικά δεδομένα. Τέλος, κατορθώσαμε να λάβουμε καλύτερες προβλέψεις χρησιμοποιώντας απλή γραμμική παλινδρόμηση. Αυτή η μελέτη μπορεί να επεκταθεί σε διάφορες κατευθύνσεις για μελλοντικές εργασίες.

Acknowledgements

Undertaking this Master in Statistical and Operational Research, has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

I would like to first say a very big thank you to my supervisor Dr Samis Trevezas for all the support and encouragement he gave me, during all this months that we work together. For his guidance, understanding, patience and most importantly, his generosity. It has been a great pleasure and honor to have him as my supervisor. Without his guidance and constant feedback this Master would not have been achievable.

I am also grateful to Dr. Apostolos Mpournetas and to Dr. Loukia Meligotsidou for accepting to be members of committee for my thesis, it is a great honor for me. Especially, I would like to thank Dr. Loukia Meligotsidou for spending time read this thesis and providing useful suggestions about this thesis.

I would also like to thank all my professors during my studies in the Department of Mathematics of National and Kapodistrian University of Athens. Without their own will to make us fully trained so that we can move on with our lives, everything would be different. All this knowledge that I received in this Department will surely help me to evolve and thrive.

My deepest gratitude goes to all of my family members for the continuous support they have given me throughout my life. Also, I thank them for their patience and their understanding during the University years; I could not have done it without them.

Contents

Chapter 1

Features Description

1 Corn Progress Reports

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) issues Crop Progress Reports (CPRs) weekly during the growing season, listing the progress made in the different phenological stages of selected crops in major producing states and Agricultural Statistics Districts (ASDs) through field survey. Crop progress is based on survey data that are collected each week from early April to the end of November. The Crop progress percent values indicate the cumulative progress for each crop at key stages and is a non-probability survey which includes a sample of more than 5,000 reporters whose occupations provide them opportunities to make visual observations and frequently bring them in contact with farmers in their counties. Based on standard definitions, these reporters subjectively estimate progress of farmers' activities and progress of crops through their stages of development. They also provide subjective evaluations of crop conditions.

Because of this complex data collection process and the accuracy of the values we receive, we can collect information on a limited geographical area. Historical data are available via CPRs of NASS. Since on-site research is time-consuming and cost-effective, it is necessary to provide efficient and accurate estimation of crop progress stages. In this report, we focus on a particular type of crop. More specifically, we are interested in corn. This choice is motivated by recent papers on this subject ([56], [57]), where the authors used Hidden Markov models to predict CPR combining historical data, meteorological conditions and satellite data. Below we give a brief analysis of the corn phenological stages.

1.1 Phenological Stages Of Corn

Producers have several methods of staging corn. The leaf collar method uses a counting system of "collared" leaves during the vegetative growth stages and includes the first emerging round-tipped leaf in the count. Another method is similar but does not count the first emerging leaf, only the later, pointed-tipped leaves. A third method of staging is commonly used by the crop insurance industry and simply counts all visible leaves, whether rounded or pointed and collared or not. The knowledge of the later staging method, used to describe the stages of plant growth, is essential and with this method we will deal.

Detailed knowledge of the plant's growth process provides the means to improve the quantity and the quality of the yield. Additionally, plant symptoms occurring during certain growth stages help the grower determine the cause and effect of a deficiency, disease or other crop problem and take timely measures.

The duration of the organic cycle of corn ranges from 110-150 days depending on the type of hybrid and the environmental conditions. The development of hybrid seed corns is a very specialized procedure. Its production requires more time, expenses and expertise to produce than other commercial crops. Hybrid seed corn production involves the crossing of two inbred lines; that is hybridization. The two inbreds that are used in the process are referred to as male (the plant responsible for producing pollen) and female (the plant which produces the hybrid seed). Throughout the process extreme measures are taken to ensure the quality and purity of the seed being produced. Inbreds are crossed to create a variety that demonstrates certain characteristics, such as drought resistance or standability, or a variety that is produced for planting specifically in various conditions and climates. While it is true that hybrids have allowed for an increase in corn production, they have done much more. Hybrids allow for an efficient use of applied fertilizers. They also allow for resistance to a variety of insects and diseases, leading to higher quality corn.

Corn is a fixed growth plant with distinct stages of germination and reproductive development. The following basic stages of the biological cycle are identified :

Germination stage of development :

- S_1 : Pre-Season
- S_2 : Planted
- S_3 : Emerged

Reproductive stage of development :

- S_4 : Silking
- S_5 : Dough
- S_6 : Dent
- S_7 : Mature
- S_8 : Harvested

For a better understanding of the biological background we analyze each phenological stage of corn development.

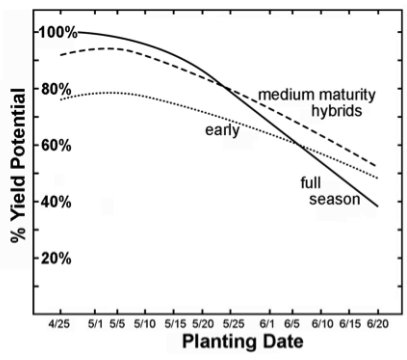
Pre-Season (S1)

The Pre-Season stage is added artificially as a phenological stage, but in reality it represents the time period from the beginning of CPR recordings until corn seed planting takes place. This approach is followed by [57] in order to facilitate the design of the model. In particular, this allows synchronization of CPR recordings from different years by initializing them at a common hypothetical stage.

Planted (S2)

Planting time can vary depending on the climate and the weather, but generally will begin in early Mid-April and will continue until mid to late May.

FIGURE 1.1: Percentage of Yield Potential relative to planting date [8].



Corn is planted for different uses such as grain, silage (grass or other green fodder compacted and stored in airtight conditions, typically in a silo, without first being dried, and used as animal feed in the winter), sweet corn etc. Early planting usually, but not always, results in maximum corn yields. Corn planted in late April or early May typically outyield either grain or silage corn planted after mid-May (see Figure 1.1). A general guideline for the best time to begin planting corn is about 10 days before the average date of the last $32^{\circ}F$ ($0^{\circ}C$) temperature in the spring.

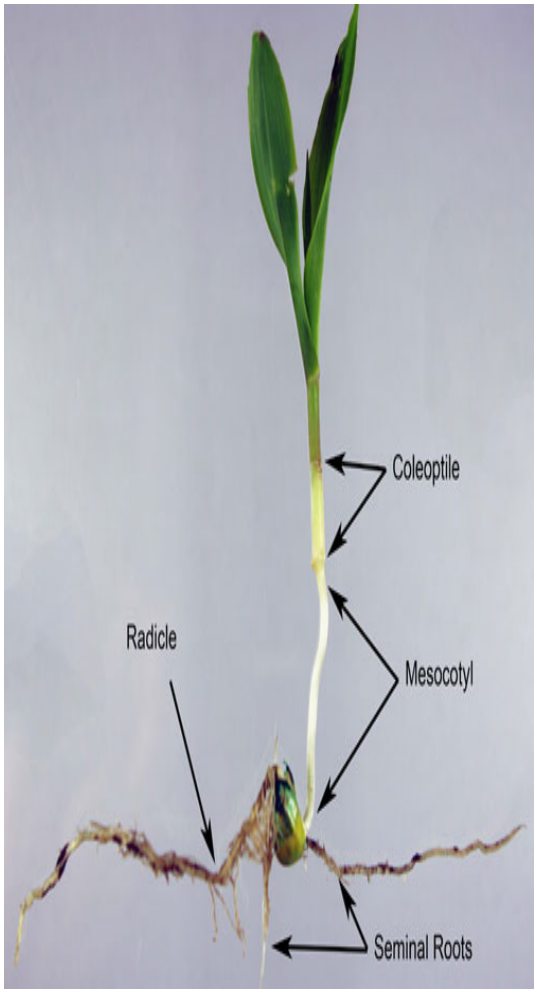
Corn planted in late May under dry soil conditions will consistently outyield corn planted in late April under wet soil conditions. Modern corn hybrids tolerate cold soil conditions and seed treatments protect corn from soil pest problems under extended emergence time due to cold soil temperatures. Planting depths of about 1.5 inches (3.81 cm) for silty clay or clay loam soils and 1.75 to 2.0 inches (4.445 to 5.08 cm) for silt loam and gravelly loam soils are recommended for April or early May-planted corn. Planting depths of about 1.75 to 2.0 inches for silty clay or clay loam soils and 2.0 to 2.5 inches (5.08 to 6.35 cm) for silt loam and gravelly loam soils are recommended for most planting dates in May. If soil conditions are dry in the top 2 inches in late May and early June, corn can be safely planted to a depth of 3 inches (7.62 cm) on silt loam and gravelly loam soils.

To achieve the full yield potential of an early planting date, full-season hybrids are necessary (see Figure 1.1). After the first or second week of May, however, the yield advantage of full-season vs. medium-season hybrids decreases when planted for grain. Furthermore, full-season hybrids may not mature, resulting in low test weight, and/or will have high grain moisture at harvest, if planted after the second week of May. Therefore, for grain production, full-season hybrids should be planted only in late April or during the first 2 weeks of May. For silage production, full-season hybrids can be planted until about May 20. Growers should not plant more than 30% of their crop to full-season hybrids (see Figure 1.1). The majority of corn acreage ($\sim$ 60%) should be planted to medium-season hybrids (100 and 200 growing degree days less than the growing degree days in a region for silage and grain, respectively). Finally, if planting must be delayed until early June, early-season hybrids (300-400 growing degree days less than the growing degree days in a region for silage and grain, respectively) are recommended.

Emerged (S3)

Once planted, the seed will eventually germinate and then emerge from the ground in approximately 7 days after germination (see Figure 1.3). Small variations exist and are mostly due to local environmental conditions. Below we analyze the two phases of plant growth, germination and emergence.

FIGURE 1.2: Corn Plant Structure [9]



Germination

Corn seed begins germination when the seed contains at least 30 percent moisture. The first seedling structure to emerge from the corn seed is the radicle (root), followed by the coleoptile (shoot) with the enclosed plumule (first leaves and growing point). (In Figure 1.2 we give an illustration of these characteristics.)

Emergence

Emergence of the radicle first allows the young seedling to anchor in the soil and obtain an adequate supply of water and for further development obtain extra water and nutrients. To emerge, the first internode on the corn plant (the mesocotyl (see Figure 1.2)) elongates toward the soil surface and continues until the coleoptile reaches light.

At this stage, the growing point is normally 1 to 1.5 inches below the soil surface.

The growing point remains below the soil surface for three to four weeks (during the period of Germination), protecting this growing point from physical injury, including frost, surface insects or grazing animals.

FIGURE 1.3: The size of corn at the Emerged stage [37]

Silking (S4)

The first stage of reproductive stages is the **Silking** stage. The plant is about 55 to 66 days after emergence. This stage begins when silks are visible and pollination occurs.
Pollination takes place when pollen grains contact the new, moist silks. A pollen grain grows down the silk and fertilizes the ovule in about 24 hours. Upon this fertilization, the ovule is a kernel. Silks grow about 1 to 1.5 inches per day. Normally, it takes two to three days for all silks on a single ear (see Figure 1.5) to emerge and be pollinated. Once this occurs, the kernels will develop and the grain will fill out (see Figure 1.4).

FIGURE 1.4: Corn development at Silking stage [41]

After the «Silking» stage and before the «Dough» stage, two other stages «Blister» and «Milk» are mediated. In summary, during these stages the development of corn is achieved as follows. The kernels are white and shaped like a blister. The cob is close to full size. Silks darken and dry. Kernels are in a steady and rapid period of seed-fill (this continues to Mature stage) (see Figure 1.5). Kernels are beginning to yellow on the outside but contain a milky white inner fluid (**starch accumulation**) (see Figure 1.6). Most of the kernels have grown out from the surrounding cob material. The endosperm cell division in each seed is complete and growth will be due to **cell expansion and starch accumulation**.

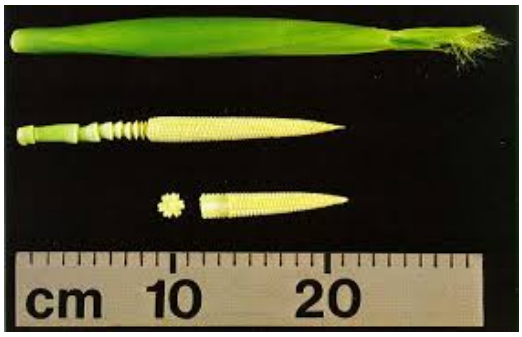


FIGURE 1.5: Blister: ear and shank [53]

FIGURE 1.6: Milk: ear and shank [53]

Dough (S5)

Bypassing the two intermediate stages, the next stage for us is the **Dough** stage. This stage appears approximately 26 days after Silking. The kernel has thickened to a pasty (doughy) consistency from the earlier milky state (starch has continued to accumulate and kernel moisture content has decreased). The embryo of the seed is growing while the kernels are just beginning to dry at the top (dent). Kernels have accumulated 50 percent of their dry weight and have about 70 percent moisture. During the Dough stage kernels become more yellow, but the appearance remain dull and matte (see Figure 1.7).

FIGURE 1.7: Corn Kernels Appearance [41]

Dent (S6)

This stage is about 36 days after Silking. Nearly all kernels are dented or denting. Drying kernels show a small, hard, white layer on top. A white line (known as the milk line or starch line) can be seen across the kernel shortly after denting (starch line indicates maturity; it will advance toward the kernel tip with maturity). Kernels at this stage have about 55 percent moisture. At around 48 days after silking, all the kernels should be fully dented. The seed embryo is morphologically mature. Dry-matter accumulation in the kernels will cease soon. Dent stage occurs when the kernel crown turns the bright, shiny, dark yellow color of mature kernels and obtain a firm consistency (see Figure 1.8).

FIGURE 1.8: The kernel crown at Dent stage [41]

Mature (S7)

This stage is about 55 days after midsilk. All kernels have attained maximum dry weight. The starch line has advanced completely to the kernel tip and a brown or black layer is present (black layer progresses on the ear from the tip kernels to the basal kernels in about 10 days (see Figure 1.9)). At black layer, the average kernel moisture is 30 to 35 percent (varying with hybrids and environmental conditions). Maturity is when the kernel development is finished and "the crop is made".

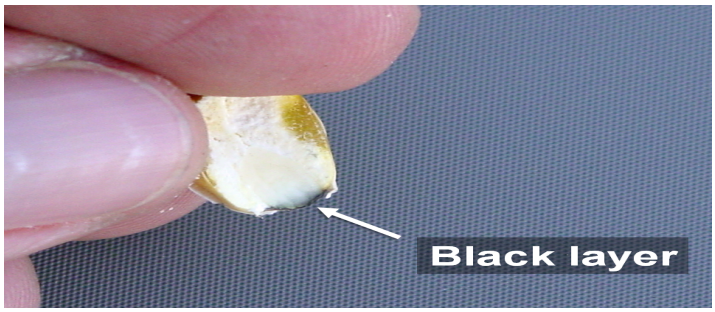


FIGURE 1.9: The characteristic black layer in the basal kernel [41]

Harvested (S8)

Harvest time will vary based on the variety of the corn and its intended use. Some corn used for silage may be harvested towards the end of August, while corn planted for animal feed needs to "dry down" before being harvested. This corn can be harvested and placed in a dryer, or it can be left in the field until it reaches approximately 15% moisture.

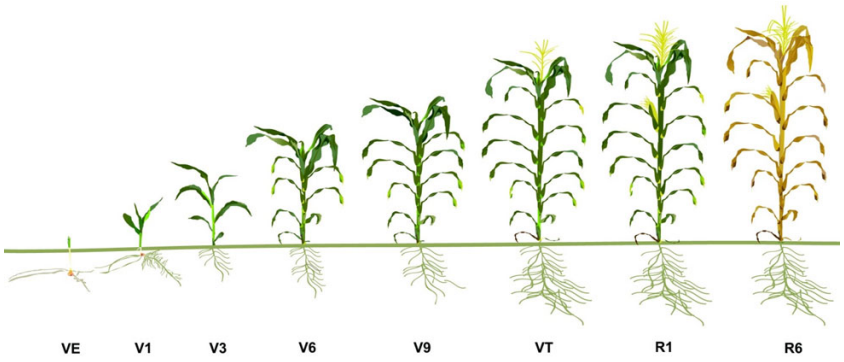


FIGURE 1.10: A representation for the corn growth and size on each phenological stage. [10]

As we mentioned above, our goal is to predict next years' CPRs. At this point we need to import two features that will be used for the prediction. The multisource features include the mean Normalized Difference Vegetation Index (NDVI) and Accumulated Growing Degree Days (AGDDs). The NDVI data are obtained from satellite data and for this reason we will describe in the next section the use of MODIS, specialized products which are used for this purpose.

2 MODIS

MODIS or Moderate Resolution Imaging Spectroradiometer is a key instrument aboard the Terra (originally known as "EOS AM-1") and Aqua (originally known as "EOS PM-1") satellites. Terra's orbit around the Earth is timed so that it passes from north to south across the equator in the morning, while Aqua passes south to north over the equator in the afternoon. With its sweeping 2,330-km-wide viewing swath, MODIS sees every point on our world every 1-2 days, acquiring data in 36 discrete spectral bands, or groups of wavelengths. These data will improve our understanding of global dynamics and processes occurring on the land, in the oceans, and in the lower atmosphere [1].

The first MODIS Flight Instrument, ProtoFlight Model or PFM, is integrated on the Terra (EOS AM-1) spacecraft. Terra successfully launched on December 18, 1999. The second MODIS flight instrument, Flight Model 1 or FM1, is integrated on the Aqua (EOS PM-1) spacecraft; it was successfully launched on May 4, 2002. These MODIS instruments offer an unprecedented look at terrestrial, atmospheric, and ocean phenomenology for a wide and diverse community of users throughout the world. The next image, shows the "EOS AM-1" scanning the Earth :

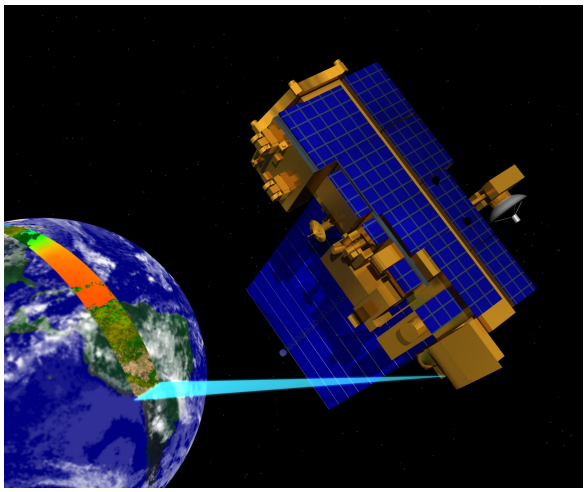


FIGURE 1.11: [6], NASA: Scientific Visualization Studio

Almost every day over the entire globe, the sensor monitors changes on the land surface, thereby building upon and extending the heritage begun by Landsat. MODIS sees changes in the Pacific phytoplankton populations that may signal the onset of the famous El Niño/La Niña climatic siblings well ahead of their arrival. MODIS also has a unique channel for measuring chlorophyll fluorescence. All plants bombarded with light begin to glow, or fluoresce, but in wavelengths that our eyes cannot see. The more plants fluoresce, the less energy they are using for photosynthesis. Thus, MODIS not only maps the distribution of phytoplankton, it also helps us gauge its health [2].

MODIS detectors measure 36 spectral bands between 0.405 and 14.385 μm, and it acquires data at three spatial resolutions – 250m, 500m, and 1,000m. Along with all the data from other instruments on board the Terra spacecraft and Aqua Spacecraft, MODIS data are transferred to ground stations in White Sands, New Mexico, via the Tracking and Data Relay Satellite System (TDRSS). The data are then sent to the EOS Data and Operations System (EDOS) at the Goddard Space Flight Center. The Level 1A, Level 1B, geolocation and cloud mask products and the Higher-level MODIS land and atmosphere products are produced by the MODIS Adaptive Processing System (MODAPS), and then are parceled out among three DAACs for distribution. Ocean color products are produced by the Ocean Color Data Processing System (OCDPS) and distributed to the science and applications community.

As just noted, MODIS products are available from several sources. MODIS Level 1 and atmosphere products are available through the LAADS web. Land Products are available through the Land Processes DAAC at the U. S. Geological Survey EROS Data Center (EDC). Cryosphere data products (snow and sea ice cover) are available from the National Snow and Ice Data Center (NSIDC) in Boulder, Colorado. Ocean color products and sea surface temperature products along with information about these products are obtainable at the OCDPS at GSFC. Users with an appropriate x-band receiving system may capture regional data directly from the spacecraft using the MODIS Direct Broadcast signal [1].

The following MODAPS–LAADS diagram provides a synoptic view of MODAPS' data flow dynamics stemming from its production activities that generate various data products that are archived and distributed by LAADS and other data centers. The primary objective of this diagram is to portray MODAPS as a central data provider, whose evolving SIPS components support and sustain both Level-1 and higher-level atmosphere and land data processing requirements for LAADS and its various NASA stakeholders [3].

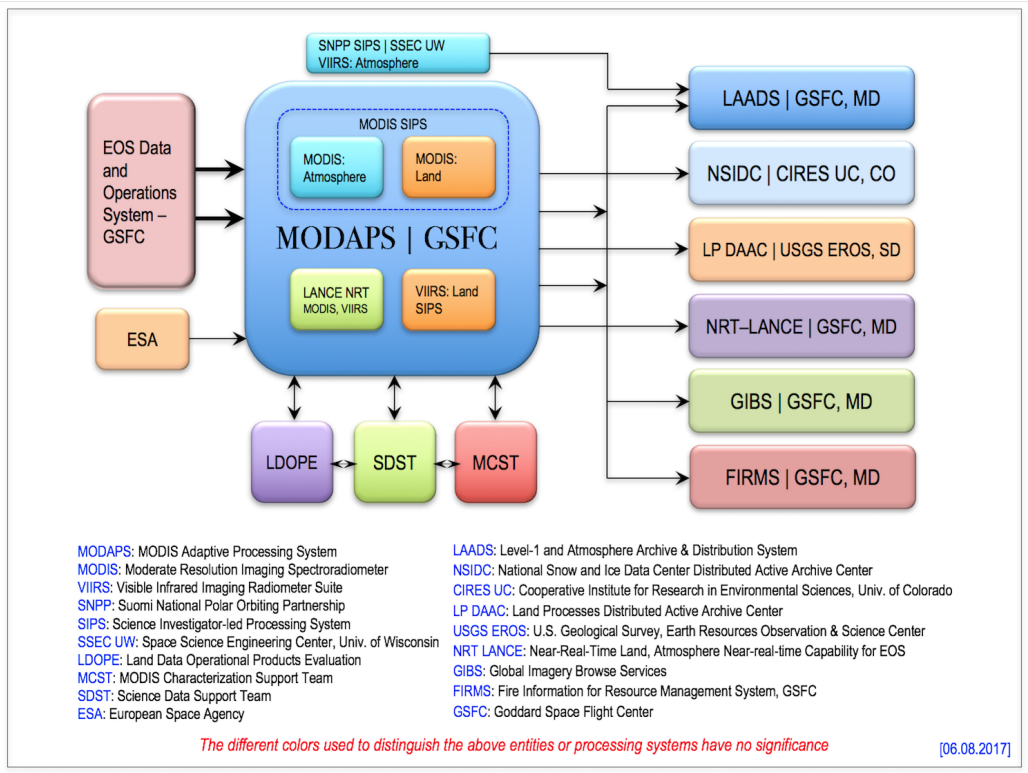


FIGURE 1.12: MODAPS-LAADS diagram [11]

The many data products derived from MODIS observations describe features of the land, oceans and the atmosphere that can be used for studies of processes and trends on local to global scales. MODIS is playing a vital role in the development of validated, global, interactive Earth system models able to predict global change accurately enough to assist policy makers in making sound decisions concerning the protection of our environment.

The MODIS Surface Reflectance products provide an estimate of the surface spectral reflectance as it would be measured at ground level in the absence of atmospheric scattering or absorption. Low-level data are corrected for atmospheric gases and aerosols, yielding a level-2 basis for several higher-order gridded level-2 (L2G) and level-3 products. MOD09GQ provides Bands 1 and 2 at a 250-meter resolution in a daily gridded L2G product in the Sinusoidal projection. Science Data Sets provided for this product include reflectance for Bands 1 and 2, a quality assurance rating and observation coverage [3].

Next is an example of MOD09GQ surface reflectance product. The corresponding MODIS data were collected on December 3, 2000 over Alabama, Mississippi and Florida. In the following image Band 2 (near-infrared) surface reflectance shown on a gray scale :

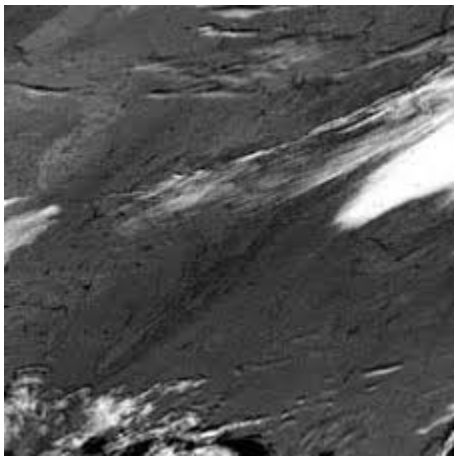


FIGURE 1.13: Image from MOD09GQ [63]

The MOD09GQ Version 6 product provides an estimate of the surface spectral reflectance of Terra MODIS 250 m bands 1 and 2 corrected for atmospheric conditions such as gasses, aerosols, and Rayleigh scattering. Along with the 250 m bands are the QC 250 m layer and five observation layers. This product is meant to be used in conjunction with the 500 m product (MOD09GA) where important quality and viewing geometry information is stored.

Some improvements and changes from previous versions are :

- Improvements to the aerosol retrieval and correction algorithm, and use of new aerosol retrieval look-up tables.
- Refinements to the internal snow, cloud, and cloud shadow detection algorithms. Uses BRDF database to better constraint the different threshold used.
- Processes ocean bands to create a new Surface Reflectance Ocean product and provide QA data sets for these bands.
- Improved discrimination of salt pans from cloud and snow, and flag salt pans in QA band.

Although crop progress metrics derived from satellite data may not necessarily correspond directly to conventional terrestrial phenological events [51], they implicitly link to the specific crop growth status. Vegetation Indices (VIs), especially the Normalized Difference Vegetation Index (NDVI), which reflects terrestrial crop cover and growth condition [52], are frequently utilized in crop progress studies. The methods of crop progress stage detection using VIs time series can be broadly grouped into four categories [26]: thresholds, derivatives, smoothing functions, and fitted models [15]. To be able to continue at the crop progress study we must analyze these indexes.

3 Vegetation Index

The exploration of outer space started in earnest with the launch of **Sputnik 1** by the Soviet Union on 4 October 1957, which was the first man-made satellite orbiting the Earth. Subsequent successful launches, both in the Soviet Union, and in the U.S., quickly led to the design and operation of dedicated meteorological satellites. Starting in 1960, the **TIROS** series of satellites embarked television cameras and radiometers. This was later (1964) followed by the Nimbus satellites and the family of Advanced Very High Resolution Radiometer instruments on board the National Oceanic and Atmospheric Administration (NOAA) platforms. The latter measures the reflectance of the planet in red and near-infrared bands, as well as in the thermal infrared.

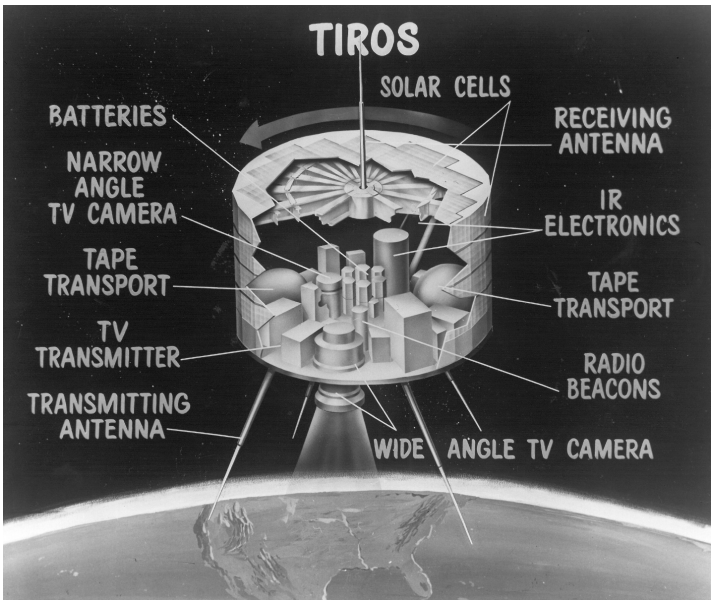


FIGURE 1.14: Television and Infrared Observations Satellites (TIROS) [12]

In parallel, NASA developed the Earth Resources Technology Satellite (ERTS), which became the precursor to the Landsat program. Despite the fact that these first sensors had minimal spectral analysis, they could distinguish germination and clouds among other targets, due to the fact that they included bands in the red and near the superficial radiation.

With the launch of the first **ERTS** satellite – which was soon to be renamed **Landsat 1**, on July 23, 1972 with its MultiSpectral Scanner (MSS)- NASA funded a number of investigations to determine its capabilities to remote sensing Earth. In a research conducted from the southern trip of Texas to the

US-Canada border, researchers Donald Deering and Robert Hass with the help of a mathematician Dr. John Schell subsequently developed the ratio of the difference of the red and infrared radiances over their sum as a means to adjust for or "normalize" the effects of the solar zenith angle. Originally, they called this ratio the "Vegetation Index" (and another variant, the square-root transformation of the difference-sum ratio, the "Transformed Vegetation Index"); but as several other remote sensing researchers were identifying the simple red/infrared ratio and other spectral ratios as the "vegetation index" they eventually began to identify the difference/sum ratio formulation as the normalized difference vegetation index.

4 NDVI

The normalized difference vegetation index (NDVI) is a simple graphical indicator that can be used to analyze remote sensing measurements, typically, but not necessarily, from a space platform, and assess whether the target being observed contains live green vegetation or not. NDVI was one of the most successful of many attempts to simply and quickly identify vegetated areas and their "condition".

Once the feasibility to detect vegetation had been demonstrated, users tended to also use the NDVI to quantify the photosynthetic capacity of plant canopies.

Since early instruments of Earth Observation, such as NASA's ERTS and NOAA's AVHRR, acquired data in visible and near-infrared, it was natural to exploit the strong differences in plant reflectance to determine their spatial distribution in these satellite images. The NOAA AVHRR instrument has five detectors, two of which are sensitive to the wavelengths of light ranging from 0.55–0.70 and 0.73–1.0 micrometers. With AVHRR's detectors, researchers can measure the intensity of light coming off the Earth in visible and near-infrared wavelengths and quantify the photosynthetic capacity of the vegetation in a given pixel (an AVHRR pixel is 1 square km) of land surface. In general, if there is much more reflected radiation in near-infrared wavelengths than in visible wavelengths, then the vegetation in that pixel is likely to be dense and may contain some type of forest. If there is very little difference in the intensity of visible and near-infrared wavelengths reflected, then the vegetation is probably sparse and may consist of grassland, tundra, or desert [5].

Definition 1.1. Let Red and NIR stand for the spectral reflectance measurements acquired in the red (visible) and near-infrared regions, respectively. Then, the Normalized Difference Vegetation Index (NDVI) is given by :

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1.1}$$

Remark. *Since Red and NIR take values in* $[0, 1]$*, the NDVI index takes values in* $[-1, 1]$*. Negative values of NDVI (values approaching* -1*) correspond to water. Values close to zero (*-0.1 *to* 0.1*) generally correspond to barren areas of rock, sand, or snow. Low, positive values represent shrub and grassland (approximately* 0.2 *to* 0.4*), while high values indicate temperate and tropical rainforests (values approaching* 1*) [7].*

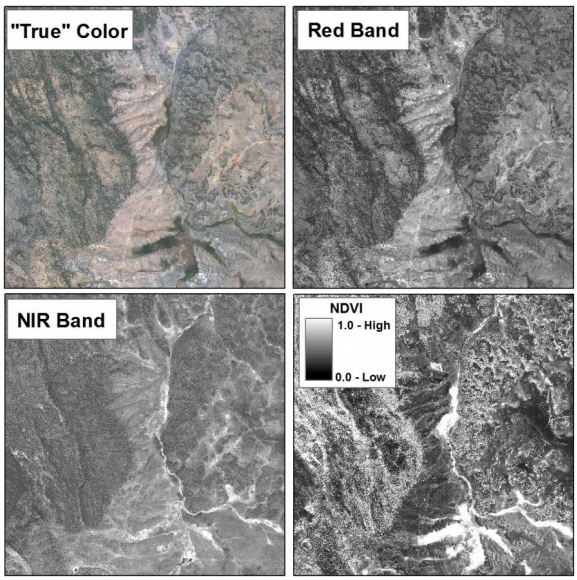


FIGURE 1.15: Example of an NDVI image calculated from an Ikonos image on an approximately 1 mi^2 area in Owyhee County, Idaho. The "true" color image (upper-left) shows encroaching juniper woodlands grading into a mosaic of montane sagebrush and semi-wet meadows. The Red (upper-right) and Near Infrared (lower-left) bands for this area each highlight different aspects of the area. From the NDVI image (lower-right), however, the junipers and semi-wet meadows are easily distinguishable.[4]

It should be noted that NDVI is functionally, but not linearly, equivalent to the simple infrared/red ratio (NIR/VIS). The advantage of NDVI over a simple infrared/red ratio is therefore generally limited to any possible linearity of its functional relationship with vegetation properties (e.g. biomass). It can be seen from its mathematical definition that the NDVI of an area containing a dense vegetation canopy will tend to positive values (say 0.3 to 0.8) while clouds and snow fields will be characterized by negative values of this index. The most important concept in the understanding of the NDVI algebraic formula is that, despite its name, it is a transformation of a spectral ratio (NIR/VIS), and it has no functional relationship to a spectral difference (NIR-VIS).

Also, the calculation of the NDVI value turns out to be sensitive to a number of perturbing factors including : atmospheric effects, clouds, soil effects, anisotropic effects and spectral effects. For these reasons, the NDVI should be used with great caution. In any quantitative application that necessitates a given level of accuracy, all the perturbing factors that could result in errors or uncertainties of that order of magnitude should be explicitly taken into account; this may require extensive processing based on ancillary data and other sources of information. More recent versions of NDVI datasets have attempted to account for these complicating factors through processing. In spite of many possible perturbing factors upon the NDVI, it remains a valuable quantitative vegetation monitoring tool when the photosynthetic capacity of the land surface needs to be studied at the appropriate spatial scale for various phenomena.

In addition, NDVI is often used around the world to monitor drought, forecast agricultural production, assist in forecasting fire zones and desert offensive maps. NDVI is preferable for global vegetation monitoring since it helps to compensate for changes in lighting conditions, surface slope, exposure, and other external factors. NDVI is a measure of the state of plant health based on how the plant reflects light at certain frequencies. For example, chlorophyll (a health indicator) strongly absorbs visible light, and the cellular structure of the leaves strongly reflect near-infrared light. When the plant becomes dehydrated, sick, afflicted with disease, etc., the spongy layer deteriorates, and the plant absorbs more of the near-infrared light, rather than reflecting it [59].

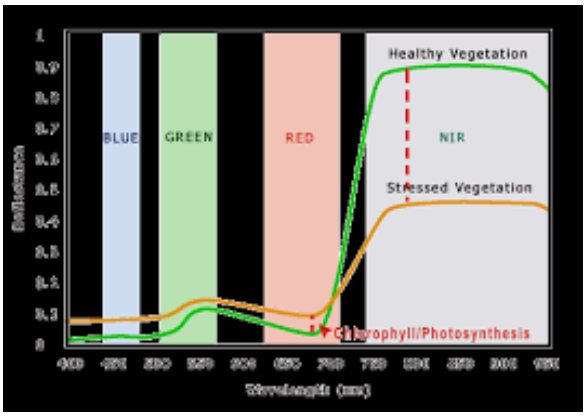


FIGURE 1.16: Plant health through wavelengths [59]

Thus, observing how NIR changes compared to red light provides an accurate indication of the presence of chlorophyll, which correlates with plant health. So, healthy vegetation (chlorophyll) reflects more near-infrared (NIR) and green light compared to other wavelengths. When you have high NDVI values, you have healthier vegetation. When you have low NDVI, you have less or no vegetation. Generally, if you want to see vegetation change over time, then you will have to perform atmospheric correction.

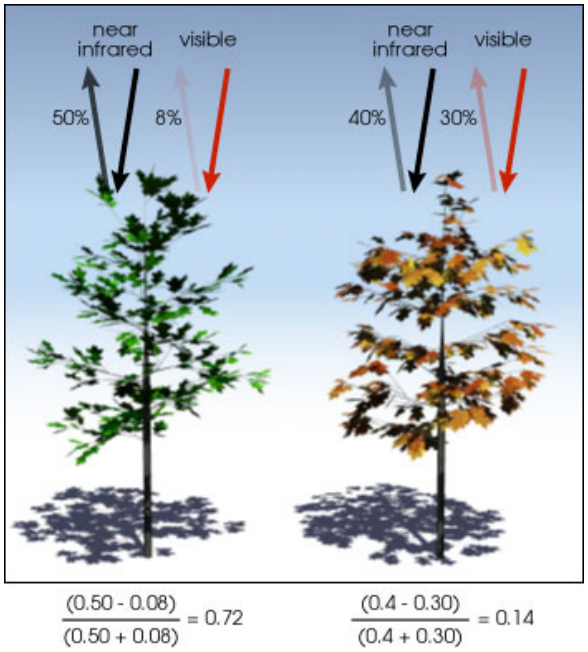


FIGURE 1.17: NDVI is calculated from the visible and near-infrared light reflected by vegetation. Healthy vegetation (left) absorbs most of the visible light that hits it, and reflects a large portion of the near-infrared light. Unhealthy or sparse vegetation (right) reflects more visible light and less near-infrared light. The numbers on the figure above are representative of actual values, but real vegetation is much more varied. (Illustration by Robert Simmon)

By transforming raw satellite data into NDVI values, researchers can create images and other products that give a rough measure of vegetation type, amount, and condition on land surfaces around the world. NDVI values can be averaged over time to establish «normal» growing conditions in a region for a given time of year. The uses of NDVI include assessing or monitoring: vegetation dynamics, biomass production, grazing impacts or attributes related to grazing management, changes in rangeland condition, vegetation or land cover classification, soil moisture, carbon sequestration or CO^2 flux etc. Further analysis can then characterize the health of vegetation in that place relative to the norm. When analyzed through time, NDVI can reveal where vegetation is thriving and where it is under stress, as well as changes in vegetation due to human activities such as deforestation, natural disturbances such as wild fires, or changes in plants' phenological stages.

5 Data Preprocessing

As we mentioned above, the NDVI is a measure of greenness of crop on spectral response of remote sensing image. Its measurements are dependent on the crop progress stage. We need to study the evolution of corn phenological stages and we have to be able to use NDVI as data. So, the raw daily pixel-based data need to be preprocessed. For this reason, a crop based mask should be applied. The data pre-processing mainly includes : **image compositing**, which composites daily NDVI images into weekly composite products by Maximum Value Composite (MVC) [39], and **image masking**, which eliminates non-corn pixels from weekly NDVI image with the mask of NASS's CDL.

5.1 Image Compositing

The current data are images from each day of the year. Our study is limited to specific weeks of the year, from the 13th to the 47th week. For each day, therefore, we have an NDVI image consisting of pixels. Instead of reasoning on a daily basis, it is common to work with weekly data. For this reason, weekly NDVI images are formed by computing for each pixel, the maximum value of the NDVI which was attained during the week. This practice allows for noise reduction and time synchronization with the CPR data which are reported on a weekly basis. In order to obtain the final aggregate data, the mean NDVI (per pixel) is computed only upon the pixels which are characterized as corn pixels. This is feasible by removing from the weekly NDVI image all the non-corn pixels with the help of image masking. [57].

5.2 Image Masking

Since 1997, the US Department Agriculture (USDA) National Agricultural Statistics Service (NASS) in conjunction with the Spatial Analysis Research Section (SARS), have worked together to provide timely, accurate and useful statistics for US agriculture. The National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA) produces the Cropland Data Layer (CDL) product, which is a raster-formatted, geo-referenced, crop-specific, land cover classification with a spatial resolution of 56 m . The CDL product utilizes rectified imagery to accurately and geospatially identify field crop types. CDL program inputs include medium resolution satellite imagery. USDA collected ground truth and other ancillary data, such as the National Land Cover Data set. It is a standardized GIS data layer of the nation's farms and fields and it was established to provide information for crop forecasting, estimation and data presentation for many agencies.

In 2009, the NASS Cropland Data Layer (CDL) program played an important role toward fulfilling this mission by providing operational in-season acreage estimates to the NASS Agricultural Statistics Board (ASB) and Field Offices (FOs) for 15 crops in 27 states. The 2009 CDL program covered many different crops, such as corn, soybeans, wheat, rice, cotton, etc. It provided updated acreage estimates throughout the growing season as increased quantities of farmers reported and satellite data became available [23].

CDL products have been used in a variety of research applications including assessing the utility of 500 m Moderate Resolution Imaging Spectroradiometer (MODIS) Time-Series Data for mapping corn and soybeans in the US [25], validating plant functional type maps developed from MODIS data using multisource evidential reasoning [60], examining the relationship between agricultural chemical exposure and cancer [44] to flood mapping assessment with satellite images [55].

It is very important that Crop Data Layers are available. In our case, in order to apply the CDL mask to weekly NDVI images, which would remove non-corn pixels, we used two packages of R : "cdlTools" and "Modistsp".

6 NDVI dataset

After the procedure of image compositing and masking (see previous section), masked weekly NDVI images are available. So, we can use 35 images, one for each week, where each one is formed by combining the pixels with the maximum value in each week and then submitted to the CDL mask. Therefore, in order to obtain our final NDVI dataset, we compute the average value of the pixels from each masked weekly NDVI image. Note that, finally, for each week we have a single observation which corresponds to the average (over all pixels) of the maximum weekly NDVI values. Consequently, the final NDVI dataset consists of 35 observations per year, one for each week, and this for every year of study. In order to emphasize this procedure, we also refer to this dataset as mean-max NDVI.

The preceding analysis concerning the NDVI is followed here by some illustrations. In Figure 1.18, we present the weekly maximum NDVI values at the pixel-level for some weeks of interest. As the colour becomes more green, NDVI values increase. In Figure 1.19 the mean maximum values of the NDVI are recorded. We can easily see again that in the weeks where the values have green colour, we take large mean max values. In Figure 1.20a the daily mean NDVI values are depicted. It is obvious that this curve is more variable and unstable, as compared to the weekly curve (Figure 1.20b). Finally, in Figure 1.21 we compare the smoothness of the curves between those formed by the weekly maximum NDVI values at the pixel-level and that of the mean max NDVI at the state-level. It is clear that the latter curve is smoother, with smaller deviations between observations.

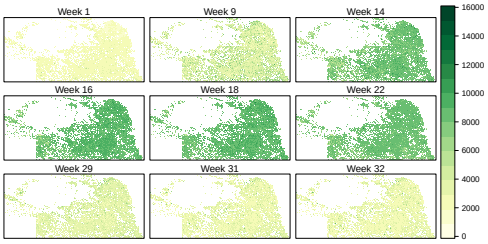


FIGURE 1.18: Presentation of the maximum pixel values from some weeks (2010, Nebraska)

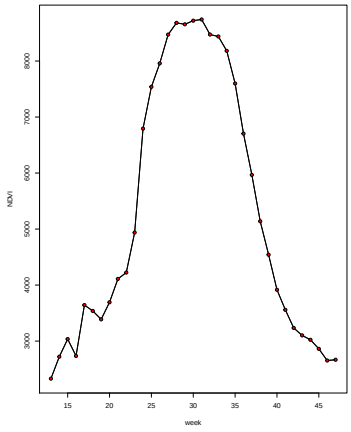
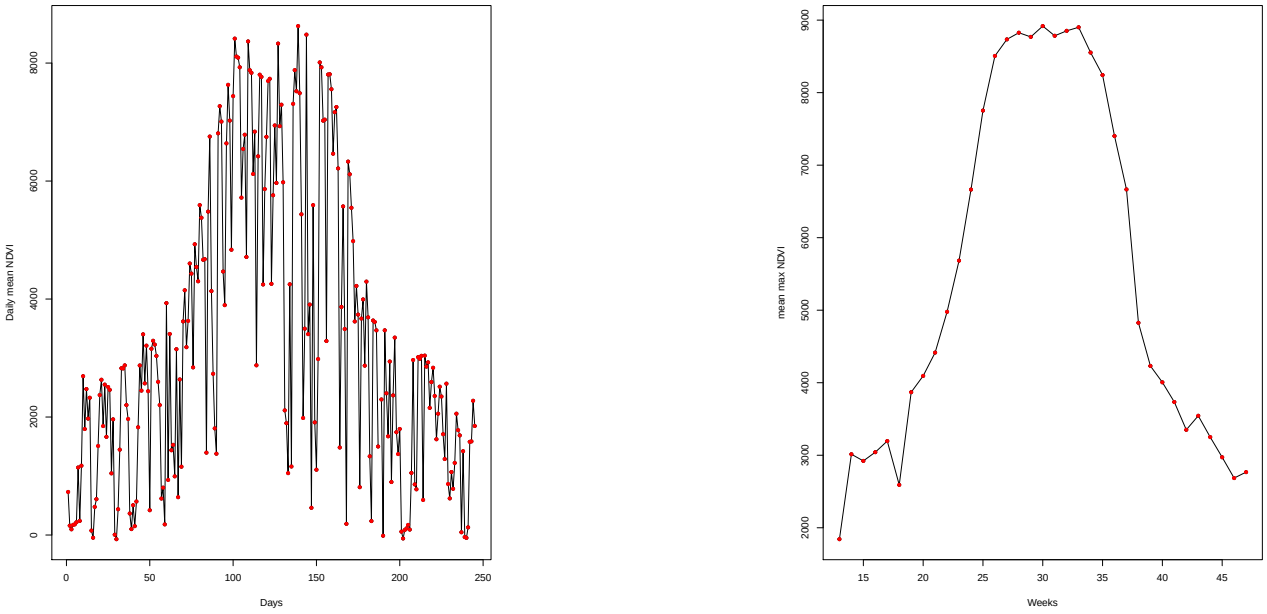


FIGURE 1.19: Mean Max NDVI (2010, Nebraska)



(A) Daily Mean NDVI (2007, Iowa)

(B) Mean Max NDVI (2007, Iowa)

FIGURE 1.20: Comparison of daily and weekly data

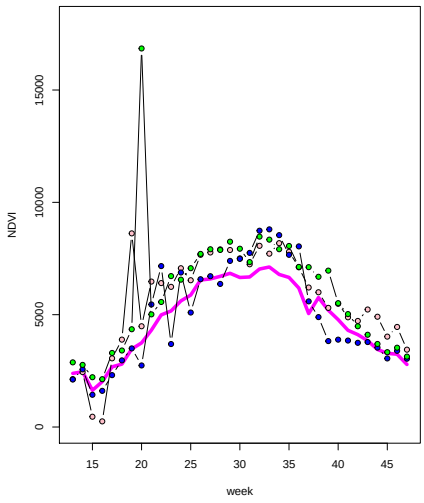


FIGURE 1.21: In this figure we have three dotted lines that stem from the maximum weekly values in three specific pixels and one line which is derived from the mean maximum values in all the state (2013, Nebraska).

7 Accumulated Growing Degree Days (AGDDs)

7.1 Individual AGDD curves

For many plants and animals, there is a specific number of growing degree days that must be accumulated to trigger a change in phenological status such as budburst in plants or egg hatching in insects. These are referred to as growing degree thresholds. If a growing degree threshold for a phenological transition in a particular organism is known, it is possible to assess how soon that transition is likely to be reached, by computing the accumulated growing degree days (AGDDs) over the course of the season.

In the studies of crop growth, temperature is often presented as growing degree days. The Growing Degree Day (GDD) corresponds to a measure of the daily accumulated heat. Accumulated growing degree days (AGDD) result from the summation of GDD within a specific period (it usually starts from the beginning of heating accumulation). The AGDD is a very important index and it is mainly used as an objective and stable way to estimate plant development rate and growing stages. This role will also be highlighted by the results of our study.

Now, we describe the basic elements for computing the AGDD, with the help of data obtained by meteorological stations. The stations that we used in this study for the states Nebraska, Illinois and Iowa are listed in Appendix B. The only temperatures that we need in order to extract the global AGDD are the ambient minimum and maximum temperatures and these are given by the meteorological stations. However, some other biological restrictions have to be taken into account. This, in particular, concerns the base temperature T_b. The lower developmental threshold temperature or base temperature for an organism is the temperature below which development stops. This threshold is determined by the organism's physiology and is independent of the method used to compute the degree days. Base thresholds vary with different organisms, but for cool crops grown in Nebraska, Illinois and Iowa 10^oC is often the best base temperature for predicting plant development [46]. In fact, there is also a restriction in the other direction. The development also ceases when temperatures exceed an upper threshold T_u. We will refer to this as an empirical maximum temperature threshold. In fact, evidence in this direction comes from previous studies, where results show that corn growth slows considerably at temperatures above 30^oC [57], [58], [62]. By taking into account all these factors we are now able to define our quantities of interest.

Definition 1.2. If T_b is the base temperature of the plant, T_u the empirical maximum temperature threshold, $T_{min}(t)$ and $T_{max}(t)$ the minimum and the maximum temperature at day t, then

(i) the *Growing Degree Days* at day t, denoted by GDD_t are defined by

$$GDD_t := \frac{T^*_{max}(t) + T^*_{min}(t)}{2} - T_b, \tag{1.2}$$

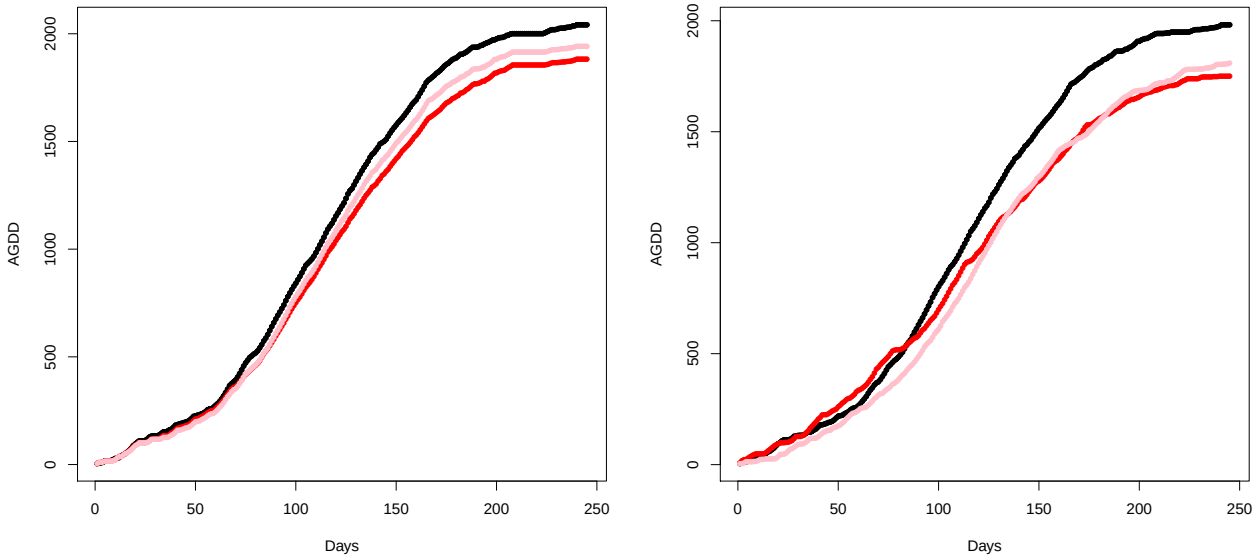
where $T^*_{max}(t) = \min\{T_{max}, T_u\}$ and $T^*_{min}(t) = \max\{T_{min}, T_b\}$,

(ii) the *Accumulated Growing Degree Days* until day t, denoted by $AGDD_t$ are defined by

$$AGDD_t := \sum_{u=1}^{t} GDD_u. \tag{1.3}$$

Note that in this definition of GDD, the daily maximum and minimum temperatures are truncated by the empirical maximum temperature threshold T_u (in this study, $T_u = 30^oC$) and the base temperature T_b (in this study, $T_b = 10^oC$) respectively. This happens because crop growth is highly sensitive to temperature as explained in the previous paragraph.

Finally, the $AGDD_t$ is defined above to be the accumulated GDD until day t, since it corresponds to the sum of the GDD values until day t. In fact, another way to interpret the AGDD is as the feature which measures the thermal age of the plant, which is more adapted to capture plant's evolution than the usual calendar time. In Figure 1.22 we give two examples from the state Nebraska. The first one depicts three different daily AGDD curves corresponding to three different meteorological stations (12^{th}, 25^{th}, 33^{rd} see Appendix B) for the same year (2002). The second one depicts three different daily AGDD curves corresponding to three different years (2002, 2004, 2008) for the same meteorological station (33^{rd} see Appendix B). Notice that there is a two-fold variability in the AGDD, inter-region and inter-year variability.



(A) Three different meteorological stations for one year

(B) Three different years for one meteorological station

FIGURE 1.22: Daily AGDDs

The AGDD curve is region specific, since it corresponds to the thermal time evolution associated to a specific region of interest. In order to be able to obtain a representative mean AGDD curve characteristic for all the state level (one for each state), we first have to introduce the concept of Thiessen Polygon.

7.2 Thiessen Polygon

As explained in the previous sections temperatures are not known in the pixel level, but only via recordings from the meteorological stations. Every recording can be represented as a point corresponding to the pixel which is associated to the specific coordinates of the meteorological station from which it was derived. With this procedure, we can form for each state and for each year of study so many AGDD curves as the number of selected meteorological stations. A natural question that arises is how we combine these curves in order to form a single mean curve representative for all the state. The sample mean may not be appropriate, since in this way each curve is considered to be equally important, which could be far from reality for several reasons. In fact, corn-pixels could be underrepresented or overrepresented in the vicinity of a specific meteorological station and this

could affect the overall mean. Another reason concerns the specific spatial distribution of the selected meteorological stations, which could be far from uniform.

A formal way to account for this problem is to associate a weight to each curve, which should be characteristic to its influence to the mean curve. This brings the problem of partitioning the state in such a way that each pixel is assigned to a specific meteorological station. In order to perform the partition, we could use the framework of **Thiessen polygons**, also known as **Voronoi diagram**.

Let us denote by $N = \{C_1, C_2, ..., C_n\}$ a set of n points, which are called centroids and correspond here to the coordinates of the meteorological stations. Each centroid C_s is associated with a polygon P_s which corresponds to the set of points of the region of interest (here the pixels of the state) which are closer to C_s, than any other centroid. The set of n centroids determines a set of n Thiessen polygons. The set of all polygons is called a Thiessen diagram [24]. In other words, Thiessen polygons are constructed around a set of climatological stations in such a way that all locations within a given polygon boundary are closer to the station enclosed by the polygon than to any other station.

With this approach, we are able to classify the pixels within polygons of supposedly constant temperature, the one of the nearest meteorological station, and this will serve as the basis for weighting the individual AGDD curves (one for each station). This will be discussed in more detail in the next subsection. Here is an illustration of the Thiessen Tessellation of Nebraska, Illinois and Iowa, with respect to the selected meteorological stations :

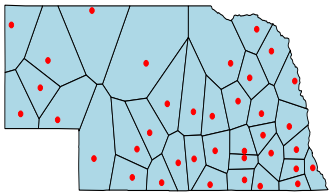


FIGURE 1.23: Nebraska Meteo Stations Tesselation

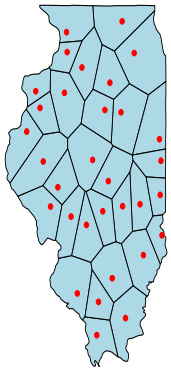


FIGURE 1.24: Illinois Meteo Stations Tesselation

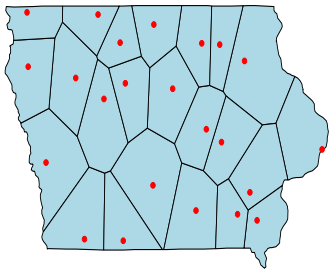


FIGURE 1.25: Iowa Meteo Stations Tesselation

7.3 State-level AGDD

In the previous subsection, we described the use of Thiessen polygons as a means to partition the state in substates of spatially constant temperature and derive a weighted mean AGDD curve. Now, we describe two ways to obtain a mean curve.

For a given year, let us denote by $AGDD(t,s)$ the value of AGDD, as defined in (1.3), for DOY (day of year) t and substate s, for a specific enumeration of the meteorological stations. We will refer to $AGDD(t,s)$ as the substate-level AGDD. A weighted mean AGDD curve at the state level is formed by combining the substate-level AGDD curves as follows:

$$AGDD(t) = \sum_s w_s AGDD(t,s), \tag{1.4}$$

where w_s corresponds to the weight associated to substate s. The weights are considered to be normalised, in the sense that $\sum_s w_s = 1$. Different ways of selecting the weights result in different state-level AGDD curves. We can opt for two natural choices of selecting the weights, and then compare their performance in the prediction problem. If we denote by A_s the area of the polygon P_s in the Thiessen Tesselation and N_s the number of corn pixels which are inside P_s, then the weights can be selected as follows:

(A) $w_s \propto A_s$, that is, each weight is proportional to the area occupied by Polygon P_s,

(B) $w_s \propto N_s$, that is, each weight is proportional to the number of corn pixels that can be found inside P_s.

The disadvantage of method A is that it does not take into account the number of corn pixels being inside each polygon. This could be a serious drawback, since an averaged AGDD curve should be representative for the crop for which it is destined for. On one hand, method B seems to surpass this problem, but on the other hand, it has the disadvantage that it does not take into account the area of the polygon. This could be important for example in the case that the corn pixels are very far from the centroid. The difference between these methods in the determination of the mean AGDD curve is illustrated in Figures 1.26 and 1.27 for two arbitrarily selected years.

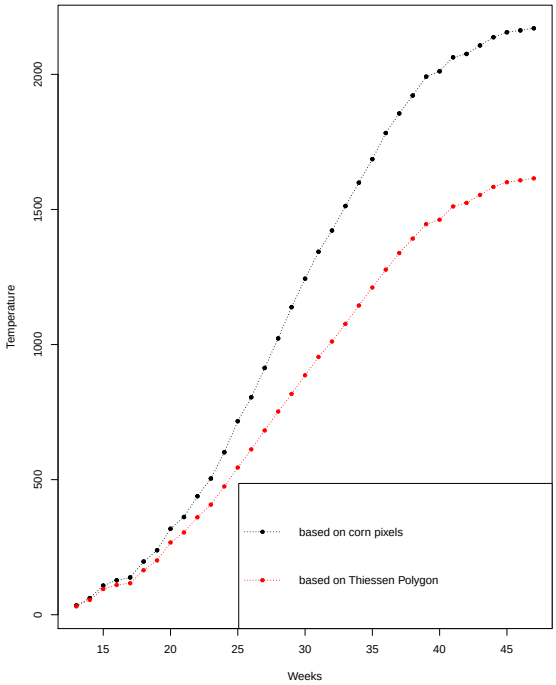


FIGURE 1.26: AGDDs values (2004, Nebraska)

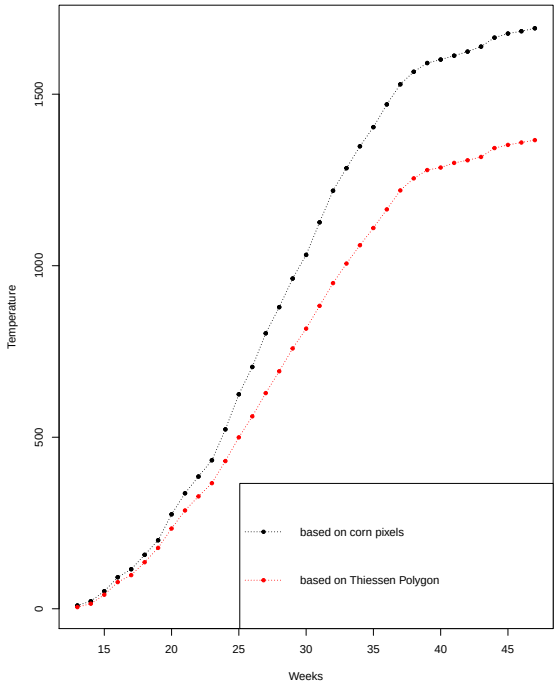


FIGURE 1.27: AGDDs values (2008, Nebraska)

Note that the AGDD values are systematically higher when we use the weights of method B (proportional to the number of corn pixels). This could be explained by the fact that corn planting is more dense in regions of higher temperatures to ensure better yield and consequently the mean thermal age should be higher when method B is selected. These methods of computing the AGDD will be compared in terms of their ability to predict the CPR at the state level.

8 Long-Term Daily and Monthly Climate Records from Stations Across the Contiguous United States

The United States Historical Climatology Network (USHCN) is a high-quality data set of daily and monthly records of basic meteorological variables from 1218 observing stations across the 48 contiguous United States. Daily data include observations of maximum and minimum temperature, precipitation amount, snowfall amount, and snow depth; monthly data consist of monthly-averaged maximum, minimum, and mean temperature and total monthly precipitation. Most of these stations are U.S. Cooperative Observing Network stations located generally in rural locations, while some are

National Weather Service First-Order stations that are often located in more urbanized environments. The USHCN has been developed over the years at the National Oceanic and Atmospheric Administration's (NOAA) National Climatic Data Center (NCDC) to assist in the detection of regional climate change. Furthermore, it has been widely used in analyzing U.S. climate. The period of record varies for each station. USHCN stations were chosen using a number of criteria including length of record, percent of missing data, number of station moves and other station changes that may affect data homogeneity, and resulting network spatial coverage.

Collaboration between NCDC and CDIAC on the USHCN project dates to the 1980s ([21]). At that time, in response to the need for an accurate, unbiased, modern historical climate record for the United States, the Global Change Research Program of the U.S. Department of Energy and NCDC chose a network of 1219 stations in the contiguous United States that would become a key baseline data set for monitoring U.S. climate. This initial USHCN data set contained monthly data and was made available free of charge from CDIAC. Since then it has been comprehensively updated several times [e.g., [22] and [28]]. The initial USHCN daily data set was made available through CDIAC via [40] and contained a 138-station subset of the USHCN. This product was updated by [29] and expanded to include 1062 stations. In 2009 the daily USHCN dataset was expanded to include all 1218 stations in the USHCN.

For our research, we find the daily records from three state of US : Nebraska, Illinois and Iowa. The number of meteorological stations of Iowa , Illinois , and Nebraska is 23, 33, and 37, respectively. Their meteorological stations that we use and we extract our data are represented in **Table 1.** (see Appendix B). For this thesis, we finally use data only from state Nebraska.

An illustration of these three states and the selected meteorological stations is following (see Figure 1.28). Stations are marked as circle dots, and colors are labeled for different states [57] (meteorological stations of Iowa (blue dots), Illinois (green dots), and Nebraska (red dots)).

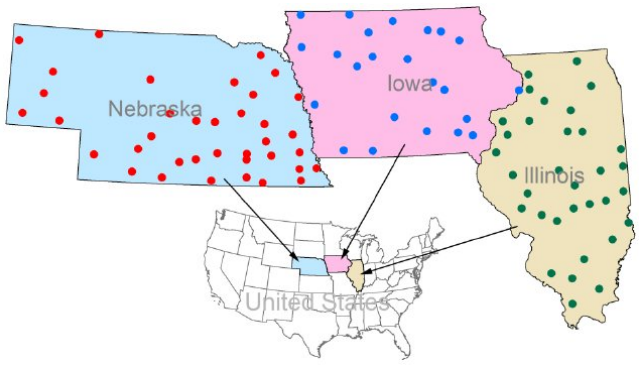


FIGURE 1.28: Illustration of three states of US and the meteorological stations

Chapter 2

Hidden Markov Models

Although initially introduced in the late 1960s, Hidden Markov Models have become very popular in the last several years. This has occurred, since this type of modeling works very well in practice for important applications, when applied properly. Also, the models are very rich in mathematical structure and hence can provide a theoretical base to a wide range of applications. The first application of Hidden Markov Models was in speech recognition (see L.Rabiner [48]). These models are also referred to as Markov sources or probabilistic functions of Markov chains in the literature.

However, neither the theory of Hidden Markov Models (HMMs), nor its applications is new. The basic theory was first introduced in a series of statistical papers by Leonard E. Baum and his colleagues in the late 1960s ([17],[18],[19], [20], [16]). Therefore, the theoretical properties of the simplest cases are proved in a series of papers by Baum et al. The term Hidden Markov Process mainly corresponds to a couple of stochastic processes $(X_t, Y_t)_{t\geq 0}$, where (X_t) is assumed to be an unobservable Markov chain that can be observed only indirectly through the process (Y_t). In this work we restrict ourselves to the case where the hidden processes have finite state space. The term Hidden Markov Model, instead of process, is used in the case where some parameters are unknown and have to be estimated from available data.

In the simplest case, where the observable process has also a finite spate space, the hidden Markov model can be characterized by the following five elements.

- The number of states N in the model. This corresponds to the number of states of the underlying hidden Markov process. The states often have some relation to the phenomena being modeled.
- The sample size M corresponding to the number of observations from the output variable of the system being modeled.
- The state transition probability matrix $P = (p_{ij})$, $1 \leq i, j \leq N$.
- The emission probabilities from state i : $R = (R_i(k))$, $1 \leq i \leq N$, $1 \leq k \leq M$.
- The initial state probability vector $\pi = (\pi_i), 1 \leq i \leq N$.

So, we can obtain an HMM as a two-layered process consisting of a hidden layer and an observable layer. The hidden layer produces a state sequence that is discrete and not observable ("hidden"), but generates the observation sequence on the basis of the state-dependent probability/density functions, where the latter possibility corresponds to the case where the observable sequence has a continuous state space. The first layer of an HMM is a Markov chain as introduced above. To define the second layer of an HMM, we need to specify a space of possible output values and a probability/density function for each state. In general, the output space can be any set including the real numbers, a vector space, or any kind of feature space. Also, there is a possibility for different state types. The observable variables can either be continuous or discrete, or mixed in some rare cases. Now, we give a formal definition of the simplest case of a Hidden Markov Model.

Definition 0.1. A hidden Markov Model is a bivariate discrete time stochastic process $(X_t, Y_t)_{t\geq 0}$, where

- (X_t) is an unobservable Markov chain and,
- (Y_t) is an observable sequence of conditionally independent random variables such that the conditional distribution of Y_t given (X_t), depends only on X_t.

In the sequel we refer to $f(x_{0:n}, y_{0:n})$ as the joint density of $(X_{0:n}, Y_{0:n})$ w.r.t. to the product measure $\nu_{n+1} \otimes \mu_{n+1}$, where ν_{n+1} refers to the $(n+1)$-dimensional counting measure and μ_{n+1} to the $(n+1)$-dimensional Lebesgue measure or counting measure, depending on the choice of the conditional distributions of $[Y_t \mid X_t = i]$. The joint density of a sequence of states and observations for the first-order HMM can be written as :

$$f(x_{0:n}, y_{0:n}) = f(x_0) f(y_0|x_0) \prod_{t=1}^{n} f(x_t|x_{t-1}) f(y_t|x_t), \tag{2.1}$$

where the notation $y_{0:n}$ is used as a shorthand for $y_0, \ldots, y_n$ and probabilities or conditional probabilities/densities are simplified by the notation $f(x_0)$, $f(x_t \mid x_{t-1})$ or $f(y_t \mid x_t)$, as a shorthand for $\mathbb{P}(X_0 = x_0)$, $\mathbb{P}(X_t = x_t \mid X_{t-1} = x_{t-1})$ or $\mathbb{P}(Y_t = y_t \mid X_t = x_t)$, when the observable variables are discrete or in the latter case this could also refer to a density $f_{Y_t|X_t}(y_t \mid x_t)$, when the observable variables are continuous. The above equation can be rewritten as :

$$f(x_{0:n}, y_{0:n}) = f(x_0) \prod_{t=1}^{n} f(x_t|x_{t-1}) \prod_{t=0}^{n} f(y_t|x_t) \tag{2.2}$$

1 Non-Homogeneous Normal Hidden Markov Model

Hidden Markov models (HMMs) are frequently used to analyse longitudinal data, where the same set of subjects is repeatedly observed over time. In this context, several sources of heterogeneity may arise at individual and/or time level, which affect the hidden process, that is, the transition probabilities between the hidden states. In this thesis, we use non-homogeneous HMMs (NH-HMMs) to face the heterogeneity problem. The non-homogeneity of the model allows us to take into account heterogeneity in time of the transition probabilities. So, the NH-HMM is categorized by the assumed existence of a finite, discrete-valued, hidden state process which follows a nonhomogeneous Markov chain. The basic Markov chain model (MC) consists of a time-invariant transition matrix which records the probabilities of a state change. Here, these transitions are also dependent on time t. Additionally, we also assume that the observations are continuous and the conditional densities correspond to that of a normal distribution. The resulting model is a non-homogeneous Normal Hidden Markov Model. In particular :

$$[Y_t|X_t = j] \sim \mathcal{N}_d(\mu_j, \Sigma_j),$$

where

$$f_{Y_t|X_t}(y_t \mid x_t = j) = (2\pi)^{-(d/2)} \det(\Sigma_j)^{-1/2} \exp\{(-0.5)(y_t - \mu_j)^\top (\Sigma_j)^{-1}(y_t - \mu_j)\}, \tag{2.3}$$

and the state-dependent mean vector μ_j and covariance matrix Σ_j are generally unknown and have to be estimated from the data.

Hidden Markov models consist of two kinds of variables, observable and hidden. Because of the existence of hidden variables, we can not directly maximize the resulting likelihood function. However, there are indirect maximization methods that are appropriate in such kinds of situations. The most popular method is maximizing with the EM algorithm [61].

2 Expectation Maximization Algorithm

The Expectation-Maximization (EM) algorithm is a general method of finding the maximum-likelihood estimate of the parameters of an underlying distribution from a given data set when the data are incomplete or have missing values. There are two main applications of the EM algorithm. The first occurs when the data indeed have missing values, due to problems or limitations of the observation process. The second occurs when optimizing the likelihood function is analytically intractable but the likelihood function can be simplified. This likelihood is greatly simplified by using data augmentation. Data augmentation and related Markov chain Monte Carlo algorithms enable us to perform either parameter simulation, multiple imputation or both. Parameter simulation and multiple imputation can be viewed merely as two different ways of extracting information from the same Markov chain. The latter application is more common in the computational pattern recognition community.

Let Z consist of observed variables $Y = (Y_1, ..., Y_k)$ and unobserved (missing or latent variables) $X = (X_1, ..., X_{N-k})$. We write $Z = (Y, X)$. With this notation the log-likelihood function for the observed data Y is $\ell_Y(\theta)$. The problem here is that the direct maximization of the likelihood may be very difficult. To maximize $\ell_Y(\theta)$ with respect to θ the idea is to follow an iterative procedure. The EM algorithm is an iterative way to approximate the maximum likelihood function. While maximum likelihood estimation can find the "best fit" model for a set of data, it does not work particularly well for incomplete data sets. Instead, the more complex EM algorithm can find model parameters even if you have missing data. It works by choosing random values for the missing data points, and using those guesses to estimate a second set of data. The new values are used to create a better guess for the first set, and the process continues until the algorithm converges to a fixed point. The EM estimate is only guaranteed to never get worse. Usually, it will find a peak in the likelihood but if the likelihood function has multiple peaks, the EM will not necessarily find the global maximum of the likelihood. In practice, it is common to start EM from multiple random initial guesses, and choose the one with the largest likelihood as the final guess for θ ([38], [13]).

We are particularly interested in the application of the EM algorithm for two types of models, the Independent Mixture Models (IMM) and hidden Markov models, where the unobserved data correspond directly to the unknown realisations of the underlying Markov chain.

Generally, we always start with an initial value $\theta^{(0)}$ and then we continue with the two basic steps of this algorithm, the **E-step** and the **M-step**.

- **E-step :** Compute $\mathbb{E}_{\theta^{(m)}}[\ell_{y,X}(\theta) \mid Y = y] =: \mathbb{Q}(\theta; \theta^{(m)})$.
- **M-step :** Update parameter $\theta^{(m)} \rightarrow \theta^{(m+1)}$
  $\theta^{(m+1)} \in \text{argmax}_\theta \mathbb{Q}_{\theta^{(m)}}(\theta)$.

The convergence of the EM algorithm is guaranteed under some regularity conditions that can be found in [27] and [64].

2.1 EM in IMMs

The Independent Mixture Models correspond to the special case of an HMM, where the hidden variables are independent and identically distributed. Their distribution determines the mixing coefficients in the mixture model. The two steps of the EM-algorithm in the special case of Independent Mixture Models (IMM) are analyzed here.
The data are in the form $Z = (X, Y)$, where $Y = Y_{0:n}$ and $X = X_{0:n}$ and the parameter $\theta = (\rho, \phi)$, where $\rho = (p_i)_{1 \leq i \leq s}$ with $p_i = \mathbb{P}(X_k = i)$, and $\phi = (\phi_i)_{1 \leq i \leq s}$, where ϕ_i corresponds to a parameterization of the probability/density function of $[Y_k \mid X_k = i]$. First we compute the complete data likelihood :

$$\begin{aligned}
\mathcal{L}_{y,x}(\theta) &= f(x_{0:n}; \rho) f(y_{0:n} \mid x_{0:n}; \phi) = \\
&= \left(\prod_{k=0}^{n} f(x_k; \rho) \right) \left(\prod_{k=0}^{n} f(y_k | x_k; \phi) \right) = \\
&= \left(\prod_{k=0}^{n} p_{x_k} \right) \left(\prod_{k=0}^{n} f(y_k \mid x_k; \phi) \right) = \\
&= \left(\prod_{k=0}^{n} \prod_{i} p_i^{1_{\{x_k = i\}}} \right) \left(\prod_{k=0}^{n} \prod_{i} (f_i(y_k; \phi))^{1_{\{x_k = i\}}} \right).
\end{aligned} \tag{2.4}$$

Then, we have :

$$\begin{aligned}
\ell_{y,x}(\theta) = \log \mathcal{L}_{y,x}(\theta) &= \sum_{k=0}^{n} [\log f(x_k; \rho)] + \sum_{k=0}^{n} [\log f(y_k | x_k; \phi)] = \\
&= \sum_{i} \left(\sum_{k=0}^{n} 1_{\{x_k = i\}} \right) \log p_i + \sum_{i} \left(\sum_{k=0}^{n} 1_{\{x_k = i\}} \right) \log f_i(y_k; \phi).
\end{aligned} \tag{2.5}$$

- **E-step**: Compute the Expectation

$$\begin{aligned}\mathbb{Q}(\theta;\theta^{(m)}) = \mathbb{Q}(\rho,\phi;\theta^{(m)}) &=: \mathbb{E}_{\theta^{(m)}}(\ell_{y,X}(\theta) \mid Y = y)\\ &= \sum_i \left(\sum_{k=0}^{n} \mathbb{E}_{\theta^{(m)}}[1_{\{X_k=i\}} \mid Y_k = y_k]\right) \log p_i + \sum_i \left(\sum_{k=0}^{n} \mathbb{E}_{\theta^{(m)}}[1_{\{X_k=i\}} \mid Y_k = y_k]\right) \log f_i(y_k;\phi)\\ &= \sum_i \left(\sum_{k=0}^{n} w_{k,i}^{(m)}\right) \log p_i + \sum_i \left(\sum_{k=0}^{n} w_{k,i}^{(m)}\right) \log f_i(y_k;\phi),\end{aligned} \tag{2.6}$$

where $\mathbb{E}_{\theta^{(m)}}[1_{\{X_k=i\}} \mid Y_k = y_k] = \mathbb{P}_{\theta^{(m)}}(X_k = i \mid Y_k = y_k) = w_{k,i}^{(m)}$. From Bayes' Theorem we can compute the quantity $w_{k,i}^{(m)}$, so :

$$\begin{aligned} w_{k,i}^{(m)} &= \mathbb{P}_{\theta^{(m)}}(X_k = i \mid Y_k = y_k)\\ &= \frac{\mathbb{P}_{\theta^{(m)}}(X_k = i) f_i(y_k;\phi_i^{(m)})}{\sum_i \mathbb{P}_{\theta^{(m)}}(X_k = i) f_i(y_k;\phi_i^{(m)})}. \end{aligned} \tag{2.7}$$

In the case of a Gaussian IMM and since $\phi_i = (\mu_i, \Sigma_i)$ we have :

$$f_i(y_k;\phi_i) \quad = \quad (2\pi)^{-(d/2)} \det(\Sigma_i)^{-1/2} \exp\{(-0.5)(y_k - \mu_i)^\top (\Sigma_i)^{-1}(y_k - \mu_i)\}.$$

M-step : The update equations for the parameters of the model are :

$$\mu_i^{(m+1)} = \frac{\sum_{k=0}^{n} w_{k,i}^{(m)}\, y_k}{\sum_{k=0}^{n} w_{k,i}^{(m)}}, \tag{2.8}$$

$$\Sigma_i{}^{(m+1)} = \frac{\sum_{k=0}^{n} w_{k,i}^{(m)}\, (y_k - \mu_i^{(m+1)})^\top (y_k - \mu_i^{(m+1)})}{\sum_{k=0}^{n} w_{k,i}^{(m)}}. \tag{2.9}$$

2.2 EM in HMMs

In this case $\theta = (\alpha, \rho, \phi)$, where an additional parameter vector α is needed for the initial probabilities. The two steps of the EM-algorithm in the case of an HMM are analyzed here. Again, the algorithm starts with an initial value $\theta^{(0)}$ and then continues with the **E-step** and the **M-step** [43].

E-step : Compute the Expectation :

$$\mathbb{E}_{\theta^{(m)}}\left[\log \mathcal{L}_{y_{0:n},X_{0:n}}(\theta) \mid \mathrm{y}\right] = \mathbb{Q}_{\theta^{(m)}}(\theta).$$

The evaluation of the $\mathbb{Q}$-function given above necessitates the decomposition of the joint density $f(x_{0:n}, y_{0:n})$. So,

$$f(x_{0:n}, y_{0:n}; \theta) = f(x_0; \alpha) \prod_{k=1}^{n} f(x_k \mid x_{k-1}; \rho) \prod_{k=0}^{n} f(y_k \mid x_k; \phi) =$$

$$= \left(\prod_i \alpha_i^{1_{\{x_0=i\}}}\right) \left(\prod_{k=1}^{n} \prod_{i,j} p_{ij}^{1_{\{x_{k-1}=i, x_k=j\}}}\right) \left(\prod_{k=0}^{n} \prod_i (f_i(y_k; \phi_i))^{1_{\{x_k=i\}}}\right), \tag{2.10}$$

where α is the initial vector of probabilities. Then,

$$\log f(x_{0:n}, y_{0:n}; \theta) = \left(\sum_{i=1}^{s} 1_{\{x_0=i\}} \log \alpha_i\right) + \left(\sum_{k=1}^{n} \sum_{i,j} 1_{\{x_{k-1}=i, x_k=j\}} \log p_{ij}\right) + \left(\sum_{k=0}^{n} \sum_i 1_{\{x_k=i\}} \log f_i(y_k; \phi_i)\right).$$

This leads to a $\mathbb{Q}$-function which is equivalent to :

$$\begin{aligned}
\mathbb{Q}_{\theta^{(m)}}(\theta) &= \mathbb{E}_{\theta^{(m)}}[\log f(X_{0:n}, y_{0:n})) \mid Y = y] = \sum_{i=1}^{s} \mathbb{E}_{\theta^{(m)}}[1_{\{X_0=i\}} \mid y] \log \alpha_i + \\
&+ \sum_{i,j=1}^{s} \left(\sum_{k=1}^{n} \mathbb{E}_{\theta^{(m)}}[1_{\{X_{k-1}=i, X_k=j\}} \mid y]\right) \log p_{ij} + \sum_{i=1}^{s} \left(\sum_{k=0}^{n} \mathbb{E}_{\theta^{(m)}}[1_{\{X_k=i\}} \mid y]\right) \log f_i(y_k; \phi_i) = \\
&= \sum_{i=1}^{s} w_{0,i}^{(m)} \log \alpha_i + \sum_{i,j=1}^{s} \left(\sum_{k=1}^{n} w_{k,i,j}^{(m)}\right) \log p_{ij} + \sum_{i=1}^{s} \sum_{k=0}^{n} w_{k,i}^{(m)} \log f_i(y_k; \phi_i),
\end{aligned} \tag{2.11}$$

where

$$w_{k,i}^{(m)} = \mathbb{P}_{\theta}^{(m)}(X_k = i \mid \mathrm{y}), \tag{2.12}$$

$$w_{k,i,j}^{(m)} = \mathbb{P}_{\theta}^{(m)}(X_{k-1} = i, X_k = j \mid \mathrm{y}). \tag{2.13}$$

In the case of a Gaussian HMM and since $\phi_i = (\mu_i, \Sigma_i)$:

$$f_i(y_k; \phi_i) \quad = \quad (2\pi)^{-(d/2)} \det(\Sigma_i)^{-1/2} \exp\{(-0.5)(y_k - \mu_i)^\top (\Sigma_i)^{-1}(y_k - \mu_i)\}.$$

M-step : The update equations for the parameters of the model are :

$$\mu_i^{(m+1)} = \frac{\sum_{k=0}^{n} w_{k,i}^{(m)}\, y_k}{\sum_{k=0}^{n} w_{k,i}^{(m)}}, \tag{2.14}$$

$$\Sigma_i{}^{(m+1)} = \frac{\sum_{k=0}^{n} w_{k,i}^{(m)}\, (y_k - \mu_i^{(m+1)})^\top (y_k - \mu_i^{(m+1)})}{\sum_{k=0}^{n} w_{k,i}^{(m)}}. \tag{2.15}$$

2.3 The Baum-Welch Algorithm

The Baum–Welch algorithm uses the well known EM algorithm to find the maximum likelihood estimate of the parameters of a hidden Markov model given a set of observed feature vectors. Its application can be found principally in electrical engineering, computer science, statistical computing and bioinformatics. It makes use of the forward-backward algorithm [45].

All we need to compute in order to have explicit solutions in the maximization problem are the two following quantities which emerged from the E-step of the EM algorithm (see Equation (2.7) in the case of IMMs and Equations (3.19) and (3.20) in the case of HMMs) :

$$\begin{aligned} w_{k,i}^{(m)} &= \mathbb{P}_{\theta^{(m)}}(X_k = i \mid \mathrm{y}), \\ w_{k,i,j}^{(m)} &= \mathbb{P}_{\theta^{(m)}}(X_{k-1} = i, X_k = j \mid \mathrm{y}). \end{aligned}$$

The basic idea is to compute efficiently the above quantities (and consequently their sums) by a **forward-backward decomposition**.

2.4 Forward-Backward Analysis

The Forward and Backward decomposition needed to compute the weights derived from the E-step of the EM algorithm in HMMs, is analyzed here. The basic decompositions are given by the following equations :

$$w_{k,i} = \frac{\alpha_k(i) b_k(i)}{\mathcal{L}_n}, \tag{2.16}$$

$$w_{k,i,j} = \frac{\alpha_{k-1}(i) p_{ij} f_j(y_k) b_k(j)}{\mathcal{L}_n}, \tag{2.17}$$

where

$$\alpha_k(i) = \mathbb{P}(X_k = i, y_{0:k}), \tag{2.18}$$

$$b_k(i) = f(y_{(k+1):n} \mid x_k = i), \tag{2.19}$$

$$\mathcal{L}_n = f(y_{0:n}), \tag{2.20}$$

$$f_j(y_k) = f(y_k \mid x_k = j). \tag{2.21}$$

For simplicity of interpretation, we keep the symbol $\mathbb{P}$, whenever a hidden variable appears in an event, as in Equation (2.18).

Proof.

$$w_{k,i} = \mathbb{P}(X_k = i \mid y_{0:n}) = \frac{\mathbb{P}(X_k = i, y_{0:n})}{f(y_{0:n})} = \frac{\mathbb{P}(X_k = i, y_{0:k}) f(y_{(k+1):n} \mid x_k = i, \cancel{y_{0:k}})}{f(y_{0:n})} = \\ = \frac{\alpha_k(i) f(y_{(k+1):n} \mid x_k = i)}{f(y_{0:n})}.$$

So,

$$\boxed{w_{k,i} = \frac{\alpha_k(i) b_k(i)}{\mathcal{L}_n}}.$$

Now, we prove Equation (2.17).

$$w_{k,i,j} = \mathbb{P}(X_{k-1} = i, X_k = j \mid y_{0:(k-1)}, y_k, y_{(k+1):n}) = \frac{\mathbb{P}(X_{k-1} = i, X_k = j, y_{0:(k-1)}, y_k, y_{(k+1):n})}{f(y_{0:n})},$$

and consequently,

$$w_{k,i,j} = \frac{A \times B}{\mathcal{L}_n},$$

where

$$A = \mathbb{P}(X_{k-1} = i, y_{0:(k-1)})\mathbb{P}(X_k = j \mid X_{k-1} = i, \cancel{y_{0:(k-1)}}),$$
$$B = f(y_k \mid \cancel{x_{k-1} = i}, \cancel{y_{0:(k-1)}}, x_k = j) f(y_{(k+1):n} \mid \cancel{x_{k-1} = i}, x_k = j, \cancel{y_{0:(k-1)}}, \cancel{y_k}).$$

So,

$$\boxed{w_{k,i,j} = \frac{\alpha_{k-1}(i) p_{ij} f_j(y_k) b_k(j)}{\mathcal{L}_n}}$$

□

2.5 Forward-Backward Equations

In the previous section we decomposed the weights $w_{k,i}$ and $w_{k,i,j}$ with the help of $\alpha_k(i)$ and $b_k(j)$. These quantities are computed recursively via the forward-backward equations. In particular, a forward step is needed to compute the $\alpha_k(i)$, for all $k = 0, 1, \ldots, n$, and then a backward step, to compute the $b_k(i)$, for $k = n, n-1, \ldots, 0$. This algorithm is known as the Forward-Backward algorithm in HMMs. First we analyse the Forward step.

Forward Equations

$$\begin{aligned} \alpha_0(j) &= \mathbb{P}(X_0 = j) f_j(y_0) \\ \alpha_k(j) &= (\sum_{i=1}^{s} \alpha_{k-1}(i) p_{ij}) f_j(y_k), \quad k = 1, 2, \ldots, n. \end{aligned}$$

Proof.

$$\alpha_0(j) = \mathbb{P}(X_0 = j, y_0) = \mathbb{P}(X_0 = j) f(y_0 | x_0 = j) = \mathbb{P}(X_0 = j) f_j(y_0),$$

$$\alpha_k(j) = \mathbb{P}(X_k = j, y_{0:k}) = \sum_{i=1}^{s} \mathbb{P}(X_{k-1} = i, X_k = j, y_{0:(k-1)}, y_k) =$$

$$= \sum_{i=1}^{s} (\mathbb{P}(X_{k-1} = i, y_{0:(k-1)}) \mathbb{P}(X_k = j \mid X_{k-1} = i, \cancel{y_{0:(k-1)}}) f(y_k \mid x_k = j, \cancel{x_{k-1} = i}, \cancel{y_{0:(k-1)}})) =$$

$$= \left(\sum_{i=1}^{s} \alpha_{k-1}(i) p_{ij} \right) f_j(y_k).$$

□

Then, we analyze the Backward step.

Backward Equations

$$\begin{aligned} b_n(1) &= 1, \quad \forall 1 \leq i \leq s, \\ b_k(i) &= \sum_{j=1}^{s} p_{ij} f_j(y_{k+1}) b_{k+1}(j), \quad k = n-1, n-2, \ldots, 0. \end{aligned}$$

Proof.

$$b_n(i) = f(y_{(n+1):n} \mid x_n = i) = 1, \quad k = n, \quad since \quad n+1 > n,$$

$$b_{n-1}(i) = f(y_{n:n} \mid x_{n-1} = i) = f(y_n \mid x_{n-1} = i), \quad k = n-1,$$

$$b_k(i) = f(y_{(k+1):n} \mid x_k = i) = \sum_{j=1}^{s} \mathbb{P}(X_{k+1} = j, y_{(k+1):n} \mid X_k = i) =$$

$$= \sum_{j=1}^{s} \mathbb{P}(X_{k+1} = j \mid X_k = i) f(y_{k+1}, y_{(k+2):n} \mid x_{k+1} = j, x_k = i) =$$

$$= \sum_{j=1}^{s} p_{ij} f(y_{k+1} \mid x_{k+1} = j) f(y_{(k+2):n} \mid x_{k+1} = j, \cancel{y_{k+1}, x_k = i}) =$$

$$= \sum_{i=1}^{s} p_{ij} f_j(y_{k+1}) b_{k+1}(j), \quad k < n-1.$$

□

In the Non-Homogeneous case the only difference is the dependence of p_{ij} on time. In this case, the equations are :

$$w_{k,i} = \frac{\alpha_k(i)b_k(i)}{\mathcal{L}_n},$$
$$w_{k,i,j} = \frac{\alpha_{k-1}(i)p_{ij}(k)f_j(y_k)b_k(j)}{\mathcal{L}_n},$$

for the weights and

$$\begin{aligned}
\alpha_0(j) &= \mathbb{P}(X_0 = j)f_j(y_0), \\
\alpha_k(j) &= \left(\sum_{i=1}^{s} \alpha_{k-1}(i)p_{ij}(k)\right) f_j(y_k), \quad k = 1, 2, \ldots, n, \\
b_n(1) &= 1, \quad \forall 1 \leq i \leq s, \\
b_k(i) &= \sum_{j=1}^{s} p_{ij}(k)f_j(y_{k+1})b_{k+1}(j), \quad k = n-1, n-2, \ldots, 0,
\end{aligned}$$

for the Forward and Backward equations respectively.

Chapter 3

Predictions with HMM

1 Introduction

With the aim of real-time estimation and prediction of crop progress stages, in this thesis, a Hidden Markov Model (HMM) based method, including also an Independent Mixture Model (IMM), combining multisource features is being presented. The multisource features include the mean Normalized Difference Vegetation Index (NDVI) and the Accumulated Growing Degree Days (AGDDs) already analyzed in Chapter 1. As we shall see later, these methods will not give us satisfying results and for this reason we will also employ Linear Regression in order to achieve better results.

The study area covers Nebraska, which is a state of the United States. Also, similar research can be carried out in other states of the United States, like Illinois and Iowa, as can be found in [57]. The results of the experiments which were conducted in [57] were assessed and validated by the Crop Progress Reports (CPRs) of the National Agricultural Statistics Service (NASS) as described in Chapter 1. In this chapter we will illustrate the proposed methodology, but we will also consider some improvements on the initial modeling approach in order to increase its predictive performance. We will also discuss some problems related to this methodology.

The features that we include in this study concern the Accumulated Growth Degree Days (AGDDs) and the mean NDVI. It would be better if we could insert one more feature to this problem which is derived from the masked weekly NDVI images. This is called fractal dimension and it measures the roughness of the corn NDVI image. It can be used as an index of heterogeneity to reflect corn growth status. Unfortunately, we remarked a certain instability in computing the values of this index and, hence, we preferred to exclude this index from our analysis. According to the NASS's CPRs, crop progress stages in the state-level are represented as progress percentages. Below, we present some models designed to estimate corn progress percentages at the state-level. Among these models, a Hidden Markov Model based method is presented. This is highlighted because, in the field of agriculture, numerous researches are concerned with the use of HMMs and multi-temporal remote

sensing images for automatic land cover classification incorporating knowledge of phenology into the classification process.

2 Specifying an HMM

Motivated by the work of [57], we present the modeling of the evolution of the CPR statistics via an appropriate HMM. In this model, the observable variables (multisource features) include the mean NDVI and the AGDDs, while the unobservable (hidden) variables are the corn progress stages. Specifically, the corn progress stage can be assumed as the state of a non-homogeneous Markov process. The hidden stages consist of : pre-season, planted, emerged, silking, dough, dent, mature and harvested (see Chapter 1). The stage which represents the period when corn hasn't been planted, that is the pre-season stage, is added as an artificial stage to facilitate the design of the model.

In a typical Gaussian HMM, where the underlying Markov chain is homogeneous, the unknown parameter vector θ, contains the initial probability vector α, the transition matrix P and a vector ϕ, related to the means and the covariance matrices of the state dependent Gaussian distributions.

Now, if we assumed a time invariant transition probability matrix P, then all weeks along life cycle would share the same transition probability matrix. However, this is not appropriate to model corn growth. Indeed, in corn life cycle the probability of moving from the current stage to the next one generally increases with the week index depending on biophysical mechanisms and external factors driving corn plant growth. Thus, the transition probabilities are time dependent and the resulting Markov chain is non-homogeneous.

In a non-homogeneous HMM, the transition probabilities $p_{ij}(t)$ should also be added as unknown parameters. In this application's context though, there are some specificities which can simplify parameter estimation. In particular, the HMM is of the left-right type and, consequently, once a state is abandoned the system cannot return to this state again. Additionally, when in state i, the system can only move to state $i+1$ or remain in its current state. For this reason, the non-zero probabilities in the i_{th} row of the transition matrix P(t) are $p_{ii}(t)$ and $p_{i,i+1}(t)$ and, since they sum to one, there is a single unknown parameter in each row of P(t).

Another characteristic feature of this application, which simplifies the estimation procedure, is the possibility to estimate the aforementioned transition probabilities directly from the historical data of CPR statistics. This is possible, since the CPR statistics actually correspond to the estimated values of the marginal probabilities $p_i(t) = \mathbb{P}(X_t = i)$. From these probabilities we can estimate the transition probabilities $p_{ii}(t)$ and $p_{i,i+1}(t)$ as it will be described in the sequel.

In this model, the normalized CPR or stage prior $w_{t,i} \equiv p_i(t)$ can be interpreted as the area ratio of stage i occupancy at time t for a specific administrative unit. As we mentioned earlier, the NASS's CPRs record the progress percentages of each growth stage by the percent complete (area ratio) in the state-level. Here we denote the percent complete of stage i at time t by α_i^t. Corn phenological stages are unimodal in the life cycle. In Figure 3.1 we illustrate the percent complete of the eight stages at weeks 13 to 47, in Nebraska, in year 2011, as well as the corresponding $w_{t,i}$ (Figure 3.2) that can be computed from the following relation :

$$w_{t,i} = \begin{cases} \alpha_i^t & \text{if } i = N, \\ \alpha_i^t - \alpha_{i+1}^t & \text{if } 1 < i < N, \\ 1 - \alpha_{i+1}^t & \text{if } i = 1. \end{cases} \tag{3.1}$$

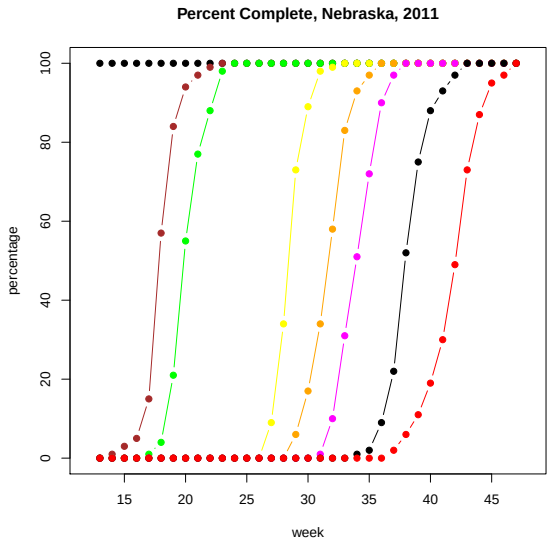


FIGURE 3.1: Percent Complete of stages, Nebraska (2011)

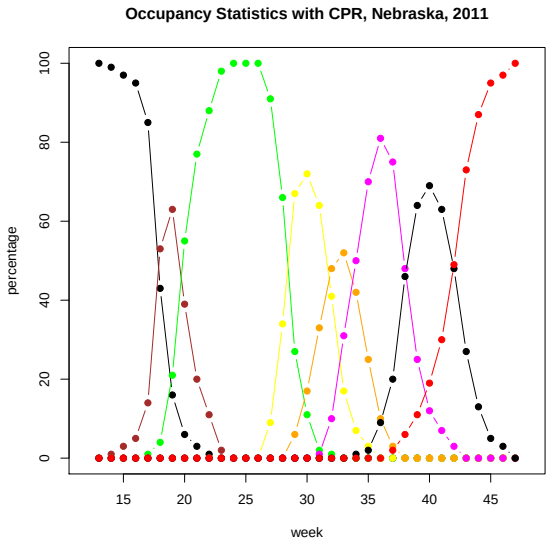


FIGURE 3.2: Occupancy Statistics with CPR, Nebraska (2011)

A representation of the features that we use in this study (Mean Max NDVI, AGDD) is also given for the years 2002-2008 in Figures 3.3 and 3.4.

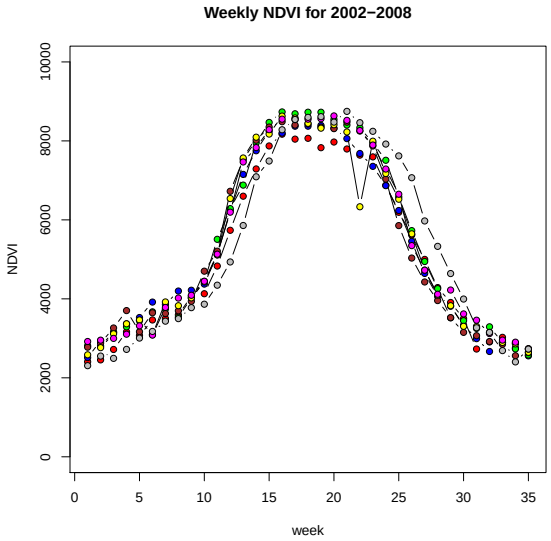


FIGURE 3.3: Weekly NDVI for 2002-2008

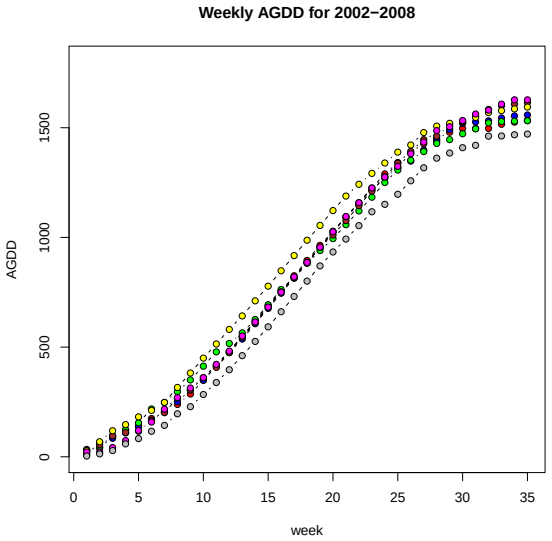


FIGURE 3.4: Weekly AGDD for 2002-2008

3 Estimation with the Independent Mixture Model

A degenerate HMM model of type M0-M0, which is the special case when (X_n) is an i.i.d. sequence of random variables, is called Independent Mixture Model (see Chapter 2). In this case, the serial dependence is irrelevant and the model is characterized by independent repetitions of

$$Y_t = \begin{cases} Y_{t,1} \sim \mathcal{N}(\mu_1, \Sigma_1), & \text{if } X_t = 1 \\ Y_{t,2} \sim \mathcal{N}(\mu_2, \Sigma_2), & \text{if } X_t = 2 \\ \vdots & \\ Y_{t,N} \sim \mathcal{N}(\mu_N, \Sigma_N), & \text{if } X_t = N, \end{cases} \tag{3.2}$$

where

$$\mathbb{P}(X_t = i) = w_{t,i}, \quad \sum_{i=1}^{N} w_{t,i} = 1. \tag{3.3}$$

So, the observation density is given by

$$f(y_t) = \sum_{i=1}^{N} w_{t,i} f_i(y_t; \phi_i), \tag{3.4}$$

where $w_{t,i}$ can be regarded as the weight (or the mixing coefficient) of the i_{th} component, $\phi_i = (\mu_i, \Sigma_i)$ and $f_i(\cdot)$ is the density of the Normal distribution $N(\mu_i, \Sigma_i)$.

Note that in this model, we do not assume that (Y_t) are identically distributed, since the non-homogeneity is reflected here to the possibility of having time-dependent mixing coefficients.

In a Gaussian IMM, the serial dependence is ignored and the unknown parameters for the degenerate Markov chain correspond only to the unknown mixing coefficients. The latter can be estimated directly from the historical data of the CPR statistics, by taking the mean over all possible available years which are considered for the estimation process.

In addition, in a Gaussian IMM, the parameters that have to be estimated are the means and the covariance matrices. As we saw in Chapter 2, in this type of model, and since $\phi_i = (\mu_i, \Sigma_i)$, we have :

$$f_i(y_t; \phi_i) = (2\pi)^{-(d/2)} \det(\Sigma_i)^{-1/2} \exp\{(-0.5)(y_t - \mu_i)^\top \Sigma_i^{-1}(y_t - \mu_i)\},$$

and the updates from the $(m+1)_{th}$ iteration of the EM algorithm are :

$$\mu_i^{(m+1)} = \frac{\sum_{t=0}^{n} w_{t,i}^{(m)} y_t}{\sum_{t=0}^{n} w_{t,i}^{(m)}}, \quad i = 1, \ldots, N, \tag{3.5}$$

$$\Sigma_i^{(m+1)} = \frac{\sum_{t=0}^{n} w_{t,i}^{(m)} (y_t - \mu_i^{(m+1)})^\top (y_t - \mu_i^{(m+1)})}{\sum_{t=0}^{n} w_{t,i}^{(m)}}, \quad i = 1, \ldots, N. \tag{3.6}$$

In this special case of IMM, the weights are computed as follows :

$$w_{t,i}^{(m)} = \frac{\hat{w}_{t,i} f_i(y_t; \phi_i^{(m)})}{\sum_{j=1}^{N} \hat{w}_{t,j} f_j(y_t; \phi_j^{(m)})}, \quad i = 1, \ldots, N, \tag{3.7}$$

where $\hat{w}_{t,i}$ correspond to the estimated values of $w_{t,i}$ which we obtain directly from the CPR statistics. The above algorithm can be easily extended when we have observations from multiple years specifically, if we assume that data for L training years are available, then the procedure is the same as with that described in Chapter 2. So, the new updates from the $(m+1)_{th}$ iteration of the EM algorithm are :

$$\mu_i^{(m+1)} = \frac{\sum_{\ell=1}^{L} \sum_{t=0}^{n} w_{\ell,t,i}^{(m)} y_{\ell,t}}{\sum_{\ell=1}^{L} \sum_{t=0}^{n} w_{\ell,t,i}^{(m)}}, \quad i = 1, \ldots, N, \tag{3.8}$$

$$\Sigma_i^{(m+1)} = \frac{\sum_{\ell=1}^{L} \sum_{t=0}^{n} w_{\ell,t,i}^{(m)} (y_{\ell,t} - \mu_i^{(m+1)})^\top (y_{\ell,t} - \mu_i^{(m+1)})}{\sum_{\ell=1}^{L} \sum_{t=0}^{n} w_{\ell,t,i}^{(m)}}, \quad i = 1, \ldots, N, \tag{3.9}$$

and the weights are computed as follows :

$$w_{\ell,t,i}^{(m)} = \frac{\hat{w}_{t,i} f_i(y_{\ell,t}; \phi_i^{(m)})}{\sum_{j=1}^{N} \hat{w}_{t,j} f_j(y_{\ell,t}; \phi_j^{(m)})}, \tag{3.10}$$

where $\hat{w}_{t,i} = \frac{1}{L} \sum_{\ell=1}^{L} \hat{w}_{\ell,t,i}$, where $\hat{w}_{\ell,t,i}$ is the value obtained from the CPR statistics for stage i occupancy, at year ℓ and week t.

4 Estimation with the HMM

In the following Proposition we indicate the way that the transition probabilities can be computed based on the marginal probabilities.

Proposition 1. Let (X_n) be an N-state Markov chain, with the property that $p_{ij}(t) = 0$, for $j \notin \{i, i+1\}$, $i = 1, 2, \ldots, N-1$ and $p_{NN}(t) = 1$. Then, the remaining transition probabilities can be computed from the marginal probabilities $w_{t,i}$ as follows :

$$p_{ii}(t) = 1 - \frac{\sum_{k=i+1}^{N} (w_{t,k} - w_{t-1,k})}{w_{t-1,i}}, \qquad \text{if } i \neq N \quad \text{and} \quad w_{t-1,i} \neq 0, \tag{3.11}$$

$$p_{i,i+1}(t) = \frac{\sum_{k=i+1}^{N} (w_{t,k} - w_{t-1,k})}{w_{t-1,i}}, \qquad \text{if } w_{t-1,i} \neq 0. \tag{3.12}$$

Proof. The last stage, N, is absorbent, since no other state follows. Consequently,

$$\boxed{p_{NN}(t) = 1}.$$

Now,

$$\mathbb{P}(X_t = N) = \mathbb{P}(X_{t-1} = N-1) p_{N-1,N}(t) + \mathbb{P}(X_{t-1} = N) p_{NN}(t) \Rightarrow$$

$$w_{t,N} = w_{t-1,N-1} p_{N-1,N}(t) + w_{t-1,N} \Rightarrow$$

$$\boxed{p_{N-1,N}(t) = \frac{w_{t,N} - w_{t-1,N}}{w_{t-1,N-1}}},$$

assuming $w_{t-1,N-1} \neq 0$. Consequently,

$$\boxed{p_{N-1,N-1}(t) = 1 - \frac{w_{t,N} - w_{t-1,N}}{w_{t-1,N-1}}}.$$

Let us now assume that the assertion holds for the i_{th} state, and we will prove that it holds for state $i-1$.

$$\mathbb{P}(X_t = i) = \mathbb{P}(X_{t-1} = i-1)p_{i-1,i}(t) + \mathbb{P}(X_{t-1} = i)p_{i,i}(t) \Rightarrow$$
$$w_{t,i} = w_{t-1,i-1}p_{i-1,i}(t) + w_{t-1,i}p_{i,i}(t).$$

Since (3.11) holds, by the inductive hypothesis, we get

$$w_{t,i} = w_{t-1,i-1}p_{i-1,i}(t) + w_{t-1,i} - \sum_{k=i+1}^{N}(w_{t,k} - w_{t-1,k}).$$

We conclude that :

$$\boxed{p_{i-1,i}(t) = \frac{w_{t,i} - w_{t-1,i} + \sum_{k=i+1}^{N}(w_{t,k} - w_{t-1,k})}{w_{t-1,i-1}}},$$

provided that $w_{t-1,i-1} \neq 0$. The above expression coincides with (3.12) for $i-1$ and thus the assertion holds for state $i-1$. This completes the proof, since (3.11) follows directly from (3.12).

□

As we saw earlier, $f(y_t)$ is a mixture of Gaussian distributions, where $f_i(y_t)$ is given by :

$$f_i(y_t) = N_i(y_t \mid \mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^d \mid \Sigma_i \mid}} \exp\{-\frac{(y_t - \mu_i)^\top \Sigma_i^{-1}(y_t - \mu_i)}{2}\}, \tag{3.13}$$

where μ_i is the mean vector, Σ_i is the covariance matrix and d is the dimension of the observation space. In our work, $d = 2$ since two kinds of features were selected in this study, the mean NDVI and the AGDDs.

In this model, the only parameters that have to be estimated are the means and the covariance matrices. Given an observation sequence $Y_0, ..., Y_n$, we can estimate μ_i and Σ_i using the EM algorithm. The initial parameter vector $\theta = (\rho, \phi)$ where ρ includes the initial probability vector and all unknown transition probabilities together with $\phi = \{\mu, \Sigma\}$. Since ρ is estimated differently from the historical data, it can be excluded from the likelihood function corresponding to the HMM.

Estimation with the EM algorithm, will be done in the same way as in Chapter 2. So, the updated equations of the EM-algorithm take the following form:

$$\mu_i^{(m+1)} = \frac{\sum_{t=0}^{n} w_{t,i}^{(m)} y_t}{\sum_{t=0}^{n} w_{t,i}^{(m)}}, \tag{3.14}$$

$$\Sigma_i^{(m+1)} = \frac{\sum_{t=0}^{n} w_{t,i}^{(m)} (y_t - \mu_i^{(m+1)})^\top (y_t - \mu_i^{(m+1)})}{\sum_{t=0}^{n} w_{t,i}^{(m)}}. \tag{3.15}$$

Notice that all we need to compute in order to obtain explicit solutions in the maximization problem is the following quantity :

$$w_{t,i}^{(m)} = \mathbb{P}_{\theta^{(m)}}(X_t = i \mid y). \tag{3.16}$$

The above probability corresponds to the posterior probability of occupying stage i at time t, given all available data from a certain year. It also corresponds to the estimated area proportion occupied by stage i at time t. In contrast with a certain year, among different years observations can be considered independent. In the sequel, since the data concern observations from different years,we describe the corresponding estimation.
In order to present the two steps of the EM-algorithm we start with the joint density which is now modified and takes the form $f(x_{1:L,0:n}, y_{1:L,0:n})$. So,

$$f(x_{1:L,0:n}, y_{1:L,0:n}; \theta) = \left(\prod_{\ell} f(x_{\ell,0}; \alpha)\right)\left(\prod_{\ell}\prod_{k=1}^{n} f(x_{\ell,k} \mid x_{\ell,k-1}; \rho)\right)\left(\prod_{\ell}\prod_{k=0}^{n} f(y_{\ell,k} \mid x_{\ell,k}; \phi)\right) =$$

$$= \left(\prod_{\ell}\prod_{i} \alpha_i^{1_{\{x_{\ell,0}=i\}}}\right)\left(\prod_{\ell}\prod_{k=1}^{n}\prod_{i,j} p_{ij}^{1_{\{x_{\ell,k-1}=i, x_{\ell,k}=j\}}}\right)\left(\prod_{\ell}\prod_{k=0}^{n}\prod_{i}(f_i(y_{\ell,k}; \phi_i))^{1_{\{x_{\ell,k}=i\}}}\right), \tag{3.17}$$

where α is the initial vector of probabilities. So,

$$\log f(x_{1:L,0:n}, y_{1:L,0:n}; \theta) = \left(\sum_{\ell}\sum_{i=1}^{s} 1_{\{x_{\ell,0}=i\}} \log \alpha_i\right) + \left(\sum_{\ell}\sum_{k=1}^{n}\sum_{i,j} 1_{\{x_{\ell,k-1}=i, x_{\ell,k}=j\}} \log p_{ij}\right) +$$
$$+ \left(\sum_{\ell}\sum_{k=0}^{n}\sum_{i} 1_{\{x_{\ell,k}=i\}} \log f_i(y_{\ell,k}; \phi_i)\right).$$

This leads to a Q-function which is equivalent to :

$$\begin{aligned}\mathbb{Q}_{\theta^{(m)}}(\theta) &= \mathbb{E}_{\theta^{(m)}}[\log f(X_{1:L,0:n}, y_{1:L,0:n})) \mid Y = y] = \sum_{i=1}^{s}\sum_{\ell}\mathbb{E}_{\theta^{(m)}}[1_{\{X_{\ell,0}=i\}} \mid y]\log\alpha_i + \\ &+ \sum_{i,j=1}^{s}\left(\sum_{\ell}\sum_{k=1}^{n}\mathbb{E}_{\theta^{(m)}}[1_{\{X_{\ell,k-1}=i, X_{\ell,k}=j\}} \mid y]\right)\log p_{ij} + \sum_{i=1}^{s}\left(\sum_{\ell}\sum_{k=0}^{n}\mathbb{E}_{\theta^{(m)}}[1_{\{X_{\ell,k}=i\}} \mid y]\right)\log f_i(y_{\ell,k};\phi_i) = \\ &= \sum_{i=1}^{s}\sum_{\ell} w_{\ell,0,i}^{(m)}\log\alpha_i + \sum_{i,j=1}^{s}\left(\sum_{\ell}\sum_{k=1}^{n} w_{\ell,k,i,j}^{(m)}\right)\log p_{ij} + \sum_{i=1}^{s}\sum_{\ell}\sum_{k=0}^{n} w_{\ell,k,i}^{(m)}\log f_i(y_{\ell,k};\phi_i),\end{aligned} \tag{3.18}$$

where

$$w_{\ell,k,i}^{(m)} = \mathbb{P}_{\theta}^{(m)}(X_{\ell,k} = i \mid \mathrm{y}), \tag{3.19}$$

$$w_{\ell,k,i,j}^{(m)} = \mathbb{P}_{\theta}^{(m)}(X_{\ell,k-1} = i, X_{\ell,k} = j \mid \mathrm{y}). \tag{3.20}$$

Since the parameters which correspond to the Markov chain are estimated in a different way from the historical data and are considered here to be fixed, the only parameters that have to be estimated with the EM algorithm are ϕ_i, $1 \leq i \leq s$. In the M-step of the algorithm we take the following update equations :

$$\mu_i^{(m+1)} = \frac{\sum_{\ell=1}^{L}\sum_{k=0}^{n} w_{\ell,k,i}^{(m)} y_{\ell,k}}{\sum_{\ell=1}^{L}\sum_{k=0}^{n} w_{\ell,k,i}^{(m)}}, \tag{3.21}$$

$$\Sigma_i{}^{(m+1)} = \frac{\sum_{\ell=1}^{L}\sum_{k=0}^{n} w_{\ell,k,i}^{(m)} (y_{\ell,k} - \mu_i^{(m+1)})^{\top}(y_{\ell,k} - \mu_i^{(m+1)})}{\sum_{\ell=1}^{L}\sum_{k=0}^{n} w_{\ell,k,i}^{(m)}}. \tag{3.22}$$

5 Predictions with the IMM and the HMM

For real-time estimation of the corn progress stages, we implement a general HMM (including an IMM) framework with multi-source features. The results concern the state of Nebraska of the United States with selected years from 2002 to 2011 during the corn growing seasons, with the 13_{th} week as the starting time point and the 47_{th} week as the last time point.

The assessment of the predictive ability of the models is based on the RMSPE (Root Mean Square Prediction Error). The RMSPE is a measure of discrepancy between the predicted and the observed values, thus lower values indicate a better predictive performance. As we mentioned earlier (see Chapter 1), the pre-season stage is not included in the error evaluation since it corresponds to an artificial stage, and it is added only to facilitate the design of the model.

For prediction purposes we first computed the RMSPE for 120 different scenaria corresponding to $120 = \binom{10}{7}$ different combinations of 7 among 10 possible years (2002-2011). For each scenario we used the 7 selected years to train the model (training set) and the 3 remaining years (testing set) for prediction. Then, the predictive performance is assessed by this sample of RMSPE values, by computing the empirical mean and 95% empirical confidence intervals over all possible 120 evaluations for each week. Furthermore, note that the sample size is small (120), so we preferred not to make random selections. In the sequel, we present this procedure formally.

Let $MSPE(k,t)$, $k = 1,2,\ldots,120$, $t = 1,2,\ldots,35$, be the Mean Square Prediction Error for scenario k and week t. Then, if we denote by T_k and P_k the sets of the 7 training years and 3 testing years (used for prediction) for scenario k, respectively, we have

$$MSPE(k,t) = \frac{\sum_{i=2}^{8}\sum_{\ell\in P_k}(\hat{w}_{\ell,t,i}^{(k)} - w_{\ell,t,i})^2}{21}, \tag{3.23}$$

where $\hat{w}_{\ell,t,i}^{(k)}$ denotes the prediction that we take for $w_{\ell,t,i}$ (depending on the method) from scenario k and for year ℓ, week t and state i. We recall also that $w_{\ell,t,i}$ refers to the observed value of the corresponding CPR statistic.

There are different methods to obtain the prediction $\hat{w}_{\ell,t,i}^{(k)}$. First of all, this depends on the modeling approach, here, the IMM or the HMM approach. But, it could also depend on the type of the estimator or/and the predictor that we choose for a given modeling approach. This will be highlighted in the sequel. In any case, the final measure for each week t and scenario k is computed as follows:

$$RMSPE(k,t) = \sqrt{MSPE(k,t)}. \tag{3.24}$$

The sample $\{RMSPE(k,t)\}_{1\leq k\leq 120}$ can then be used for assessing the overall predictive performance by evaluating for each week t, the value of $RMSPE(t) = \frac{1}{120}\sum_{k=1}^{120} RMSPE(k,t)$, that is the average value of the measure over all available scenaria, and the associated empirical 95% confidence interval.

As previously mentioned, independently of the selected modeling approach, we can get different predictions by choosing different estimators in the training phase. Here, we opted for two types of estimators, a simple one that we call Moment Estimator (ME) and the Maximum Likelihood Estimator (MLE). In the sequel, we present both estimators and we compare their performance.

Moment Estimator

A very simple way to estimate μ_i and Σ_i results from their interpretation as expectations in the following sense :

$$\mu_i = \mathbb{E}[Y_{\ell,t} \mid X_{\ell,t} = i], \tag{3.25}$$

$$\Sigma_i = \mathbb{E}[(Y_{\ell,t} - \mu_i)^\top (Y_{\ell,t} - \mu_i) \mid X_{\ell,t} = i]. \tag{3.26}$$

Since the sample $\{y_{\ell,t}\}_{\ell,t}$ can provide information for each one of them, and for all $1 \leq i \leq 8$, we need an empirical estimator which takes into account this particular setting.

The above conditional expectations can be naturally estimated as weighted averages, with the weights being proportional to $w_{\ell,t,i}$, which corresponds to the probability of the event that we condition on. This results in a type of moment estimator of the form :

$$\tilde{\mu}_i = \frac{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i} y_{\ell,t}}{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i}}, \tag{3.27}$$

$$\tilde{\Sigma}_i = \frac{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i} (y_{\ell,t} - \tilde{\mu}_i)^\top (y_{\ell,t} - \tilde{\mu}_i)}{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i}}, \tag{3.28}$$

where T denotes the training set.

Note: The above equations can be rewritten in the form :

$$\tilde{\mu}_i = \sum_{\ell \in T} \sum_{t=1}^{35} \tilde{w}_{\ell,t,i} y_{\ell,t},$$
$$\tilde{\Sigma}_i = \sum_{\ell \in T} \sum_{t=1}^{35} \tilde{w}_{\ell,t,i} (y_{\ell,t} - \tilde{\mu}_i)^\top (y_{\ell,t} - \tilde{\mu}_i),$$

where $\tilde{w}_{\ell,t,i} = \frac{w_{\ell,t,i}}{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}}$ and expresses the percentage of information given by $y_{\ell,t}$ for the parameters μ_i and Σ_i.

Note also that the resulting estimators have exactly the same form with those obtained in the update equations of the EM algorithm, which are identical in the IMM and the HMM methods. The principal difference lies in the way that the $w_{\ell,t,i}$'s are estimated. In the moment estimator, $w_{\ell,t,i}$ is estimated directly by the empirical mean of the historical data, without appealing to a recursive evaluation as in the case of the EM algorithm.

Maximum Likelihood Estimator

As we mentioned in the previous paragraph the update equations of the EM algorithm are identical for both type of models, the IMM and the HMM. Nevertheless, the resulting MLEs differ since the weights $w_{\ell,t,i}^{(m)}$ are computed differently in these models. For a training set T, the update equations are given by :

$$\mu_i^{(m+1)} = \frac{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{(m)} y_{\ell,t}}{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{(m)}}, \tag{3.29}$$
$$\Sigma_i^{(m+1)} = \frac{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{(m)} (y_{\ell,t} - \mu_i^{(m+1)})^\top (y_{\ell,t} - \mu_i^{(m+1)})}{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{(m)}}. \tag{3.30}$$

In the case of an IMM (see Equation (3.10)),

$$w_{\ell,t,i}^{(m)} = \frac{\hat{w}_{t,i} f_i(y_{\ell,t}; \phi_i^{(m)})}{\sum_{j=1}^{8} \hat{w}_{t,j} f_j(y_{\ell,t}; \phi_j^{(m)})}, \tag{3.31}$$

while in the case of an HMM,

$$w_{\ell,t,i}^{(m)} = \mathbb{P}_{\theta^{(m)}}(X_{\ell,t} = i \mid y_{\ell,1:35}), \tag{3.32}$$

and they are computed by the forward-backward algorithm (see Section 2.4 from Chapter 2). Instead of (3.32), one could also opt for an alternative evaluation of the weights. In particular, the smoothing probability could be substituted with the corresponding filtering probability given by

$$w_{\ell,t,i}^{(m)} = \mathbb{P}_{\theta^{(m)}}(X_{\ell,t} = i \mid y_{\ell,1:t}). \tag{3.33}$$

Last but not least, since the initialization for the EM-algorithm could be important in order to obtain a solution very near to the MLE, a good choice for the initial value is a critical issue. In this application, the moment estimator can be obtained directly and a good solution for initialization.

Predictions

The final procedure needed to evaluate the predictive performance of each method, is described at the beginning of Section 5 (Chapter 3) and Equations (3.23) and (3.24). For each scenario k, the predictions are denoted by $\hat{w}_{\ell,t,i}^{(k)}$. If $\theta^{(k)} = (\rho^{(k)}, \phi^{(k)})$ refers to the parameter estimate obtained from the k-th training set (independently of the type of model or estimator), then we consider 3 types of predictors :

(A)

$$\hat{w}_{\ell,t,i}^{(k)} = \frac{\hat{w}_{t,i}^{(k)} f_i(y_{\ell,t}; \phi_i^{(k)})}{\sum_j \hat{w}_{t,j} f_j(y_{\ell,t}; \phi_j^{(k)})}, \quad \ell \in P_k, \quad t = 1, \ldots, 35, \tag{3.34}$$

where the prediction is based on an IMM approach,

(B)

$$\hat{w}_{\ell,t,i}^{(k)} = \mathbb{P}_{\theta^{(k)}}(X_{\ell,t} = i \mid y_{\ell,1:35}), \quad \ell \in P_k, \quad t = 1, \ldots, 35, \tag{3.35}$$

where the prediction corresponds to the smoothing probability by taking into account all the available data for year ℓ, and

(C)

$$\hat{w}_{\ell,t,i}^{(k)} = \mathbb{P}_{\theta^{(k)}}(X_{\ell,t} = i \mid y_{\ell,1:t}), \tag{3.36}$$

where the prediction corresponds to the filtering probability by taking into account only the observations until the current time t, for year ℓ.

Blind Predictions

All the above types of predictors (A)-(C) use the data of a specific year in the testing set to predict the marginal distribution of the hidden stages. Since historical data through the CPR are available, they can be used directly to estimate the unknown probabilities $\mathbb{P}(X_{\ell,t} = i)$, without taking into account the data $\{y_{\ell,t}\}_{\ell,t}$. In particular,

$$\hat{w}^{(k)}_{\ell,t,i} = \frac{\sum_{\ell \in T_k} w_{\ell,t,i}}{7}, \quad \forall \ell \in P_k, \quad \forall t = 1, \ldots, 35, \quad \forall i = 1, \ldots, 8. \tag{3.37}$$

We refer to these predictions as blind predictions, since the data are ignored.

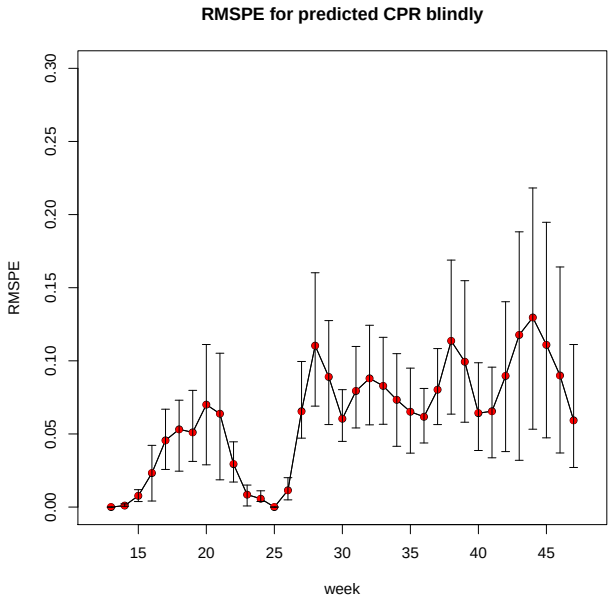


FIGURE 3.5: Results with the Blind Predictions. The dots correspond to the mean RMSPE computed for all 120 scenaria and the bars to 95% empirical confidence intervals.

In Figure 3.5 we illustrate the RMSPE values derived only from historical data by taking into account at each scenario only the CPR data from the 7 years of the training set and associated predicted values given by (3.37). Note that the prediction errors can be separated into 2 phases. The first concerns the data from week 13 to week 25, where the prediction error is relatively small with a peak at week 20, while it vanishes at week 25. The second one concerns the data from week 26 to week 47, where the prediction error is generally higher and more variable than that of the first phase.

The goal of this thesis is to compare the blind predictions with the predictions obtained by different modeling approaches, estimators and predictors, such as the HMM, the IMM and the types

of estimators and predictors described in this section. A modeling approach will be successful if it succeeds in "beating" the prediction errors obtained by the blind predictions.

6 Full Data Model

In this section we give the results that we obtained using some of the previously described models/estimators/predictors, when the observed data consist of both the NDVI and the AGDD values.

Maximum Likelihood Estimator

In this part of the thesis we show and discuss the results that we obtained using the MLE, both in the case of an IMM and in the case of an HMM. First, in Figure 3.6 we present the results from the IMM method, where both the estimator and the predictor are of type (A). Then, in Figure 3.7 we present the results from the HMM, where both the estimation and the prediction are performed through filtering and not smoothing.

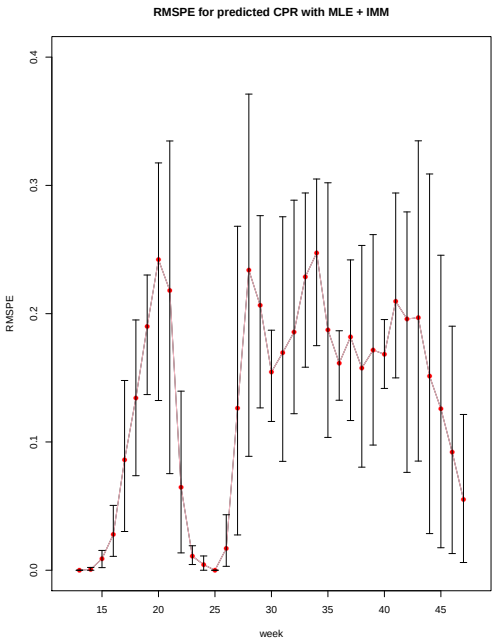


FIGURE 3.6: Mean RMSPE and associated 95% empirical confidence intervals obtained from all 120 scenaria with the IMM, under the MLE estimator and the predictor of type (A)

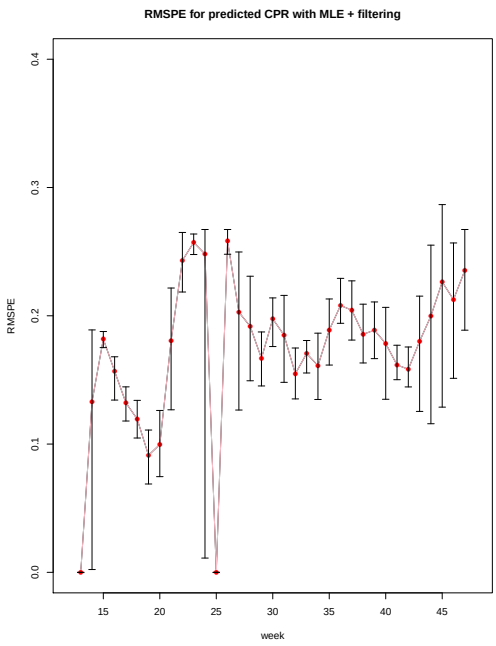


FIGURE 3.7: Mean RMSPE and associated 95% empirical confidence intervals obtained from all 120 scenaria with the HMM, under the MLE estimator, and weight updates through filtering (type (C))

From Figure 3.6, we observe that the RMSPE increases gradually, and reaches the first maximum around the 20_{th} week, and then it decreases gradually until the 25_{th} week. In this period the proportion of the 'emerged stage' increases gradually. In this week, results are less affected by overlaps of stages, since the only appeared stage is the "Emerged" phenological stage (see Chapter 1). After the emerged stage, the RMSPE reaches another maximum around the 28_{th} week, which is likely caused by progress stage overlaps. In Figure 3.7 the prediction error is more variable and the values around the weeks 26-47 are higher than those from the first weeks.

In Figure 3.8 we can see the RMSPE results when the weights are computed with filtering and the predictor is of type (A). The results show that predicting as an IMM decreases considerably the prediction errors.

In Figure 3.9 we compare the results of Figure 3.8 with those obtained by the blind predictions. The comparison shows that even if the former combination is the best one in the case of the MLE approach, the predictions are generally worse than the blind predictions, with the exception of some weeks at the end of the growing season (37-39, 44-47).

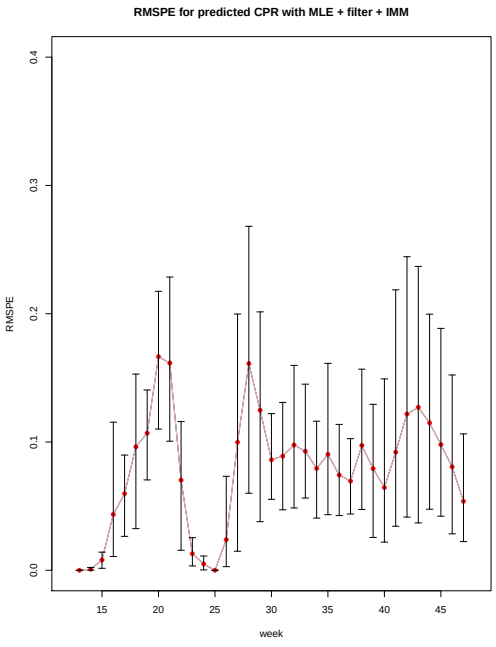


FIGURE 3.8: Mean RMSPE and associated 95% empirical confidence intervals obtained from all 120 scenaria with the HMM, under the MLE estimator and the predictor of type (A).

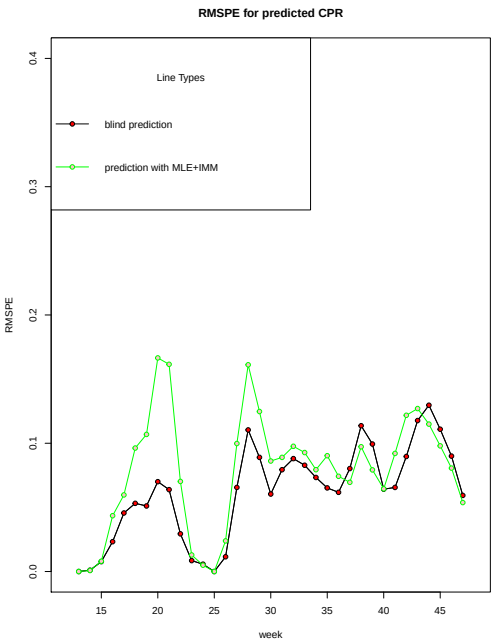


FIGURE 3.9: Blind Prediction vs Figure 3.8

We conclude that this model is not able to beat the predictions based only on historical data. But, we can observe that when we estimate the weights through filtering and use the predictor of type (A) we get lower RMSPE values which are closer to the values that we want to surpass. The other combinations neither have low RMSPE values (so they don't estimate the observed values well enough) nor are able to beat the Blind Prediction RMSPE values.

Moment Estimator

In an attempt to find a better estimator than that based on the historical data, we have experimented with a simpler estimator, the Moment Estimator. In this subsection we provide and discuss the results obtained from the ME, when the estimation is performed through filtering and the predictor is of type (A).

Since predicting as an IMM reduces the prediction errors, which means that the RMSPE values are lower, in Figure 3.10 we compare the curves from MLE and ME. The former gives us better RMSPE values, closer to zero and to those of the blind predictions. However, neither with MLE nor with ME we can beat predictions based only on historical data.

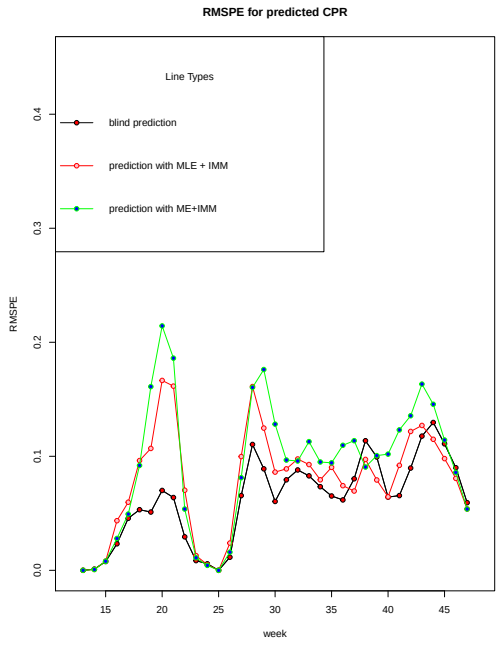


FIGURE 3.10: Comparison for the use of MLE and ME

To develop effective methods, in trying to defeat predictions based on historical data, we must follow a series of experiments. A start is to design a model which will include only the feature NDVI as an observed variable. The idea is to create a simpler model which may have similar behaviour to the more complex model. Below we describe the results analytically.

7 Single Feature Model and Comparisons

7.1 NDVI model

In this section, we discuss the results that we obtained using the same methods as in Section 6, when the observed data only consist of the NDVI.

Maximum Likelihood Estimation

In this subsection we provide and discuss the results that we obtained using MLE. In Figure 3.11 we illustrate the method in which the estimator and the predictor are both of type (A). Nearby, in Figure 3.12 we present the case where the estimation is performed through filtering but the predictor is of type (A).

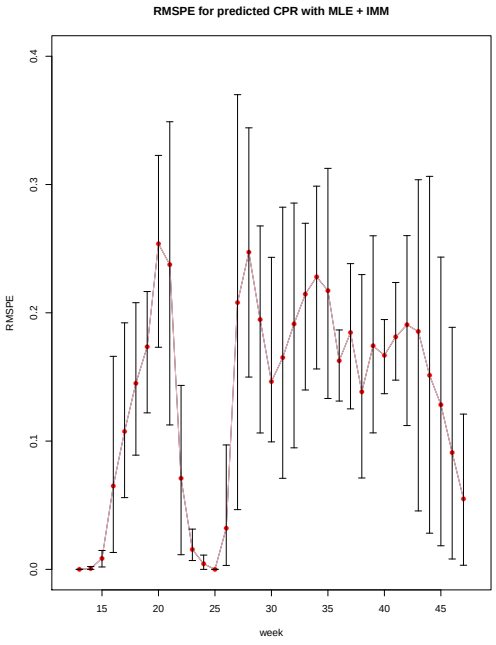


FIGURE 3.11: Use of MLE, estimator and predictor of type (A)

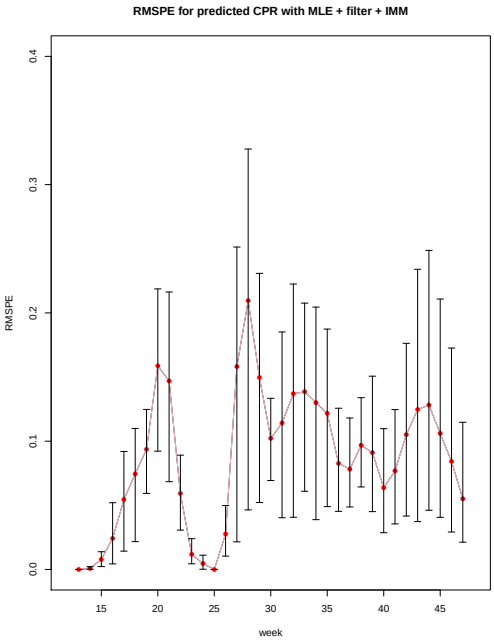


FIGURE 3.12: Use of MLE, estimation through filtering and predictor of type (A)

We observe that the RMSPE values in the latter case are lower than those from the first method, so in Figure 3.13 we compare the Figure 3.12 with this from Blind Predictions. We note that predicting as an IMM amounts to substantially lower values of the prediction errors just as in the case of the full data model.

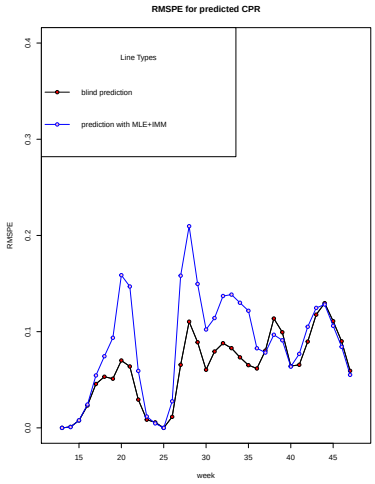


FIGURE 3.13: Blind Prediction vs Figure 3.12

Below, we compare the full data model with the one that includes only the NDVI as observed data. The interesting point here is that for a large number of weeks the exclusion of AGDD does not seem to play a significant role, but for weeks 27-37 the inclusion of AGDD seems to improve the prediction errors considerably. This could be explained by the fact that the thermal time substantially increases during the summer period and it becomes the most important predictive factor exactly at that period.

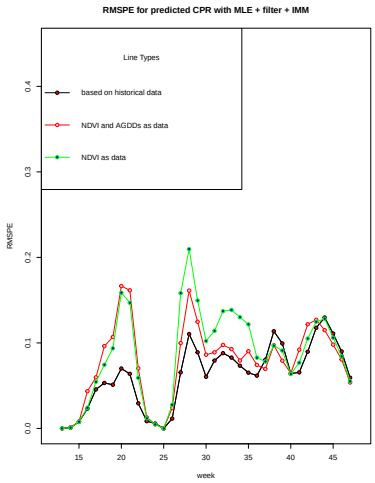


FIGURE 3.14: Comparison of the Figures 3.12, 3.8 and 3.5

Moment Estimator

In this subsection we make some experiments based on the results that stemmed from the above methods and from the use of the ME. From Figure 3.15 we observe that in the full data model the RMSPE values are closer to zero and to the blind predictions' respective RMSPE values.

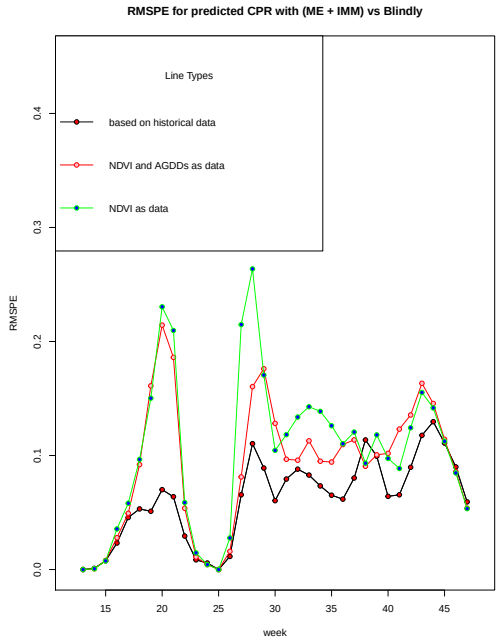


FIGURE 3.15: Comparison between the model based on historical data, the full data model and the single variable model, using the ME and with the estimators and predictors being of type (A).

Afterwards in Figure 3.16 we make a comparison between the full data model and the single variable model, when the estimation is performed through filtering and the predictor is of type (A). As we can see, the full model works better in this case since its RMSPE values are closer to zero than those of the NDVI-model. Then, in Figure 3.17 we stick to the single variable model and we compare the curves obtained from the use of the MLE or the ME. In this combination, the MLE works better and gives us lower RMSE values than the ME. However, the difference between the values is small, so a simpler estimator gives us almost the same values with the more complex estimator except for a small number of weeks (27, 28, 37).

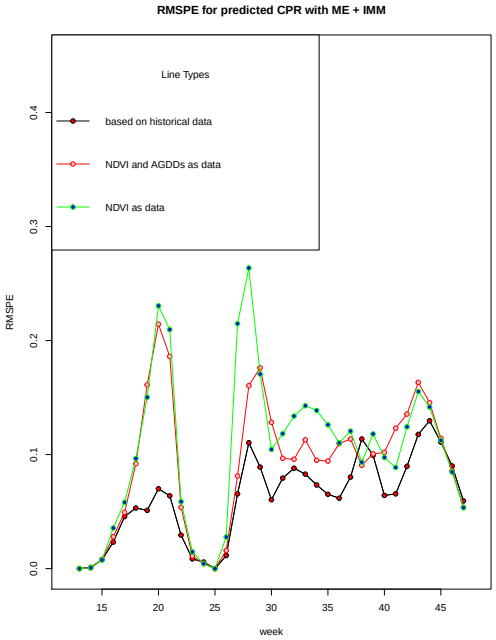


FIGURE 3.16: Comparison of full model with NDVI model. Use of ME, estimate like type (C) and predict like type (A)

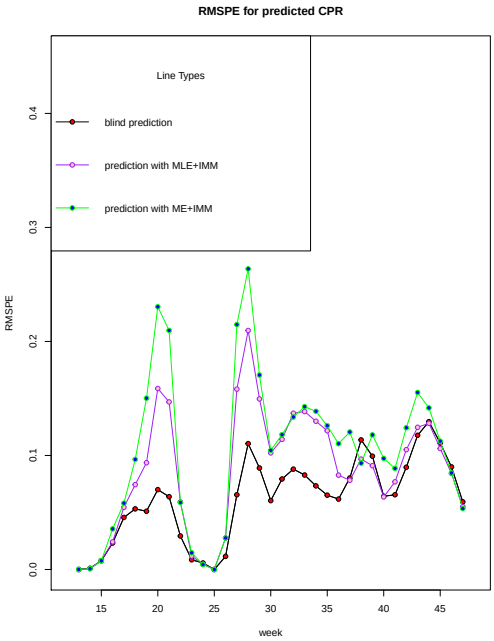


FIGURE 3.17: Comparison for the use of MLE or ME, estimate like type (C) and predict like type (A)

The above experiment was performed with the aim to find an appropriate model that can beat the predictions based on historical data. In the experiments we considered the full data model and the reduced data model. Unfortunately neither the full model nor the reduced one managed to achieve better results than the Blind Predictions. It is interesting to note that a simpler estimator, the ME, gave us results comparable to those of the MLE but not good enough to achieve our target.

7.2 AGDD model

In this section, we show the results that we obtained using the same methods as in Section 6, based on the AGDD as a single observed variable.

Maximum Likelihood Estimation

Here we present and discuss the results that we obtained using the MLE. In Figure 3.18 we illustrate the results from the method in which the estimator and the predictor are both of type (A) and we compare them with those from the Blind Predictions. In Figure 3.19 we present the results from the

method in which the estimation is performed through filtering but the predictor is of type (A) and we make the comparison with Blind Predictions.

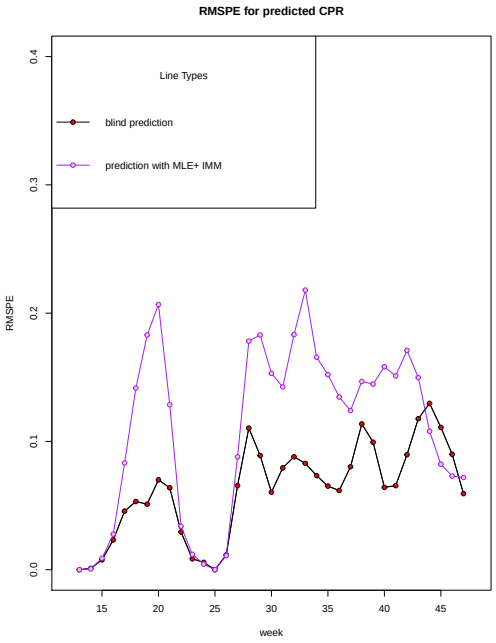


FIGURE 3.18: RMSPE under MLE, estimator and predictor of type (A). Comparison with Blind Predictions.

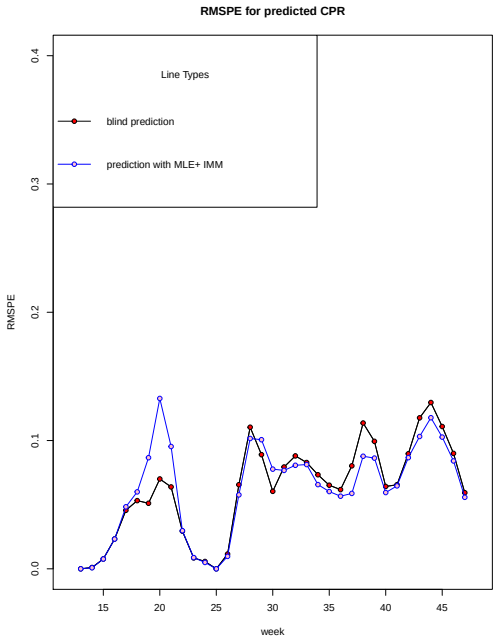


FIGURE 3.19: RMSPE under MLE, estimation through filtering and predictor of type (A). Comparison with Blind Predictions.

From Figure 3.18 we observe that, except of the weeks 46, 45 and 44 where the RMSPE values are lower than those from blind predictions, for all other weeks the prediction error is higher than that of the predictions based on the historical data. On the other hand, with the implementation of the second method we achieve better results as shown in Figure 3.19. The values, especially after week 22, are lower than the values that we want to surpass. This could be explained by the fact that the Thermal Time becomes more important over these weeks.

Below, we compare the NDVI model with the one including only the AGDD data. As we can see, the values from the latter model are lower than those from the NDVI model. This could be explained by the fact that the thermal time is more significant in corn's life cycle and, as it increases in the summer, it becomes the most important predictive factor.

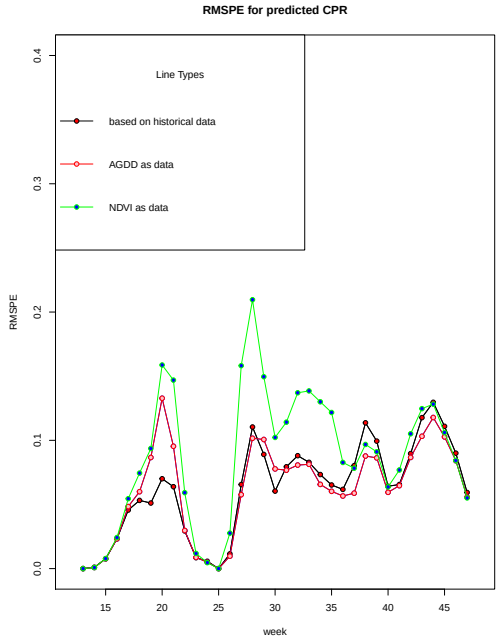


FIGURE 3.20: Comparison between Figures 3.12, 3.19 and 3.5

Moment Estimator

In this subsection we provide and discuss the results obtained with the use of the ME. We will illustrate the method in which the estimation is performed through filtering and the predictor is of type (A), since in this case we took the better results.

In Figure 3.21 we can see that during the first weeks the prediction error increases, until week 20, and then it decreases until week 25. From this week until the end of the growing season, the RMSPE values are close to those from blind predictions and especially from week 37 to week 47 the values surpass the predictions based on historical data. Nearby, in Figure 3.22 we compare the use of the MLE with the use of the ME when we follow the above procedure. We note that the MLE gives us better RMSPE values except for some weeks at the end of the growing season where the values from the ME surpass those stemmed from MLE.

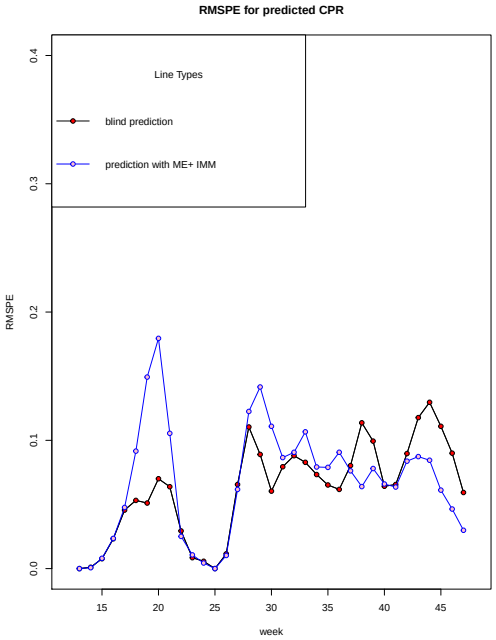


FIGURE 3.21: RMSPE under ME, estimation through filtering and predictor of type (A). Comparison with Blind Predictions

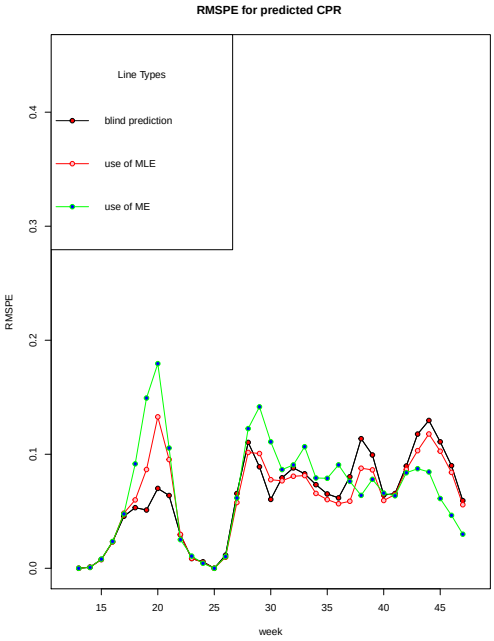


FIGURE 3.22: Comparison between Figures 3.19, 3.21 and 3.5

8 Differences

The previous modeling approaches did not succeed in beating the blind predictions. For this reason, we tried to improve the performance of the obtained predictions by considering data transformations, in particular by taking the differences between successive observations. The respective results showed that the NDVI is the least informative feature. We start by presenting the results that we obtained when we applied the differences only to the NDVI. More specifically, if we denote the data in a specific year by Y_t , $t = 1, 2, ..., 35$ then the new data will be the variable $D_t = Y_{t+1} - Y_t$, $t = 1, 2, ..., 34$. Below, we give the results that we obtained under some of the previously described models/estimators/predictors.

Maximum Likelihood Estimator

In this part of the thesis we present and discuss the results obtained under MLE. First, in Figure 3.23 we present the results from the IMM method, where the estimator and the predictor are both of type (A). Then, in Figure 3.24 we present the method in which the estimation is performed through filtering but the predictor is of type (A).

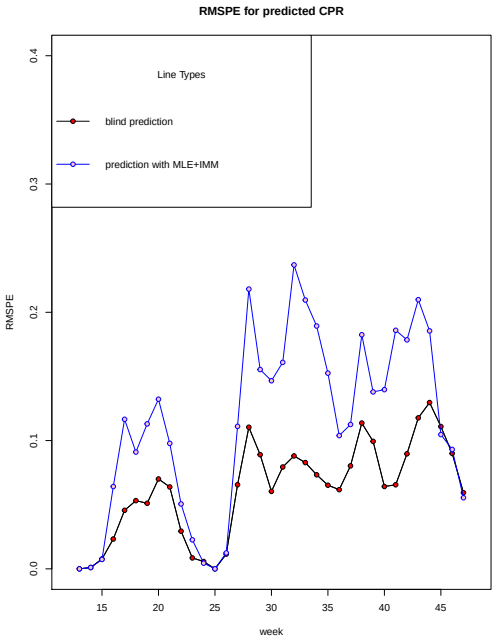


FIGURE 3.23: Mean RMSPE with the IMM, under the MLE estimator and the predictor of type (A). Comparison with Blind Predictions.

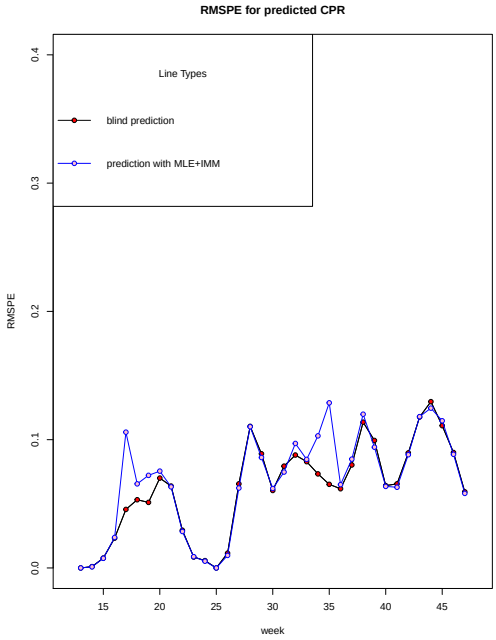


FIGURE 3.24: Mean RMSPE with the HMM, under the MLE estimator, estimate through filtering and the predictor of type (A). Comparison with Blind Predictions.

From Figure 3.23 we observe that the prediction error from week 13 until week 25 is close to that of the blind predictions with some of the values at the start of the growing season being the same with those of the blinds. In contrast, from week 26 until week 43 the RMSPE values increase and are higher than the values from the first period. Finally, in the last weeks of the growing season the RMSPE values are almost the same with those from the blind predictions.

In Figure 3.24 the results are quite satisfying. The prediction errors are almost identical to those of the blind predictions, except for some weeks (17, 18, 19, 32, 34 and 35) where the values are slightly higher than those obtained from the blind predictions.

We carried out the same experiment with the use of the ME but the results showed that the prediction error was higher than that under MLE.

Below, we compare the method which produced the best results in the case of the NDVI model (that is, when the estimation is performed through filtering but the predictor is of type (A)), with the model witch includes the differences from AGDDs and with this from the differences of both features.

As we may conclude from the very low RMSPE values, the differences work better for all models. The prediction error in both cases is close to this at the blind prediction. Furthermore, we note that the differences from the NDVI approach better our target. As we can see in Figure 3.29 the prediction error from the model derived only from the differences of NDVI is quite close to this at the blind predictions and lower than those from the other two models.

Finally, in Figure 3.30 we make the comparison between the full data model and the model which includes the AGDD and the differences from NDVI. We observe that the latter model gives us better RMSPE values, closer to those from the blind predictions. In particular, in weeks 23, 31, 32, 33 and 43 the values from the latter model surpass the predictions based on historical data.

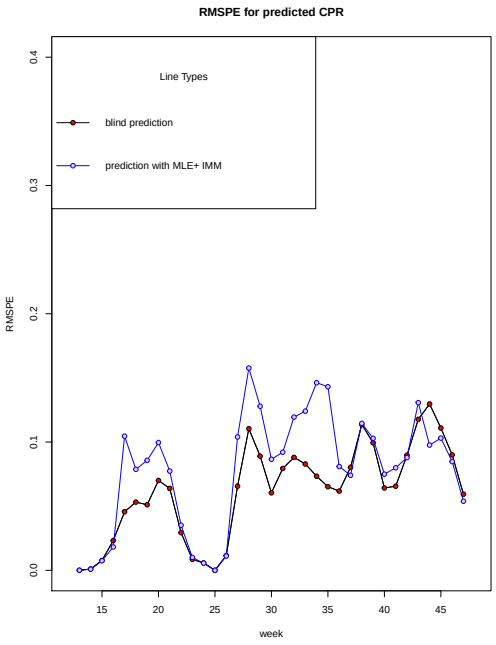


FIGURE 3.25: Model based on differences in both features vs Model based on historical data

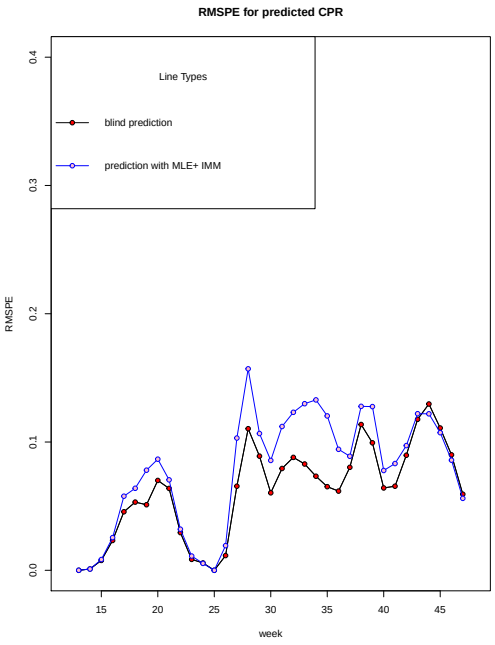


FIGURE 3.26: Model based on differences from AGDD vs Model based on historical data

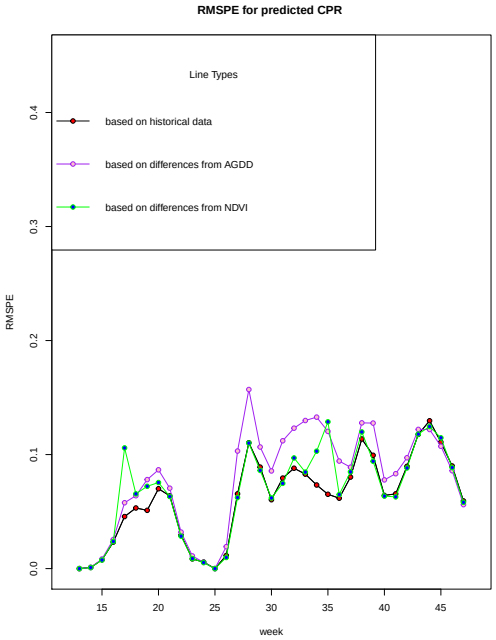


FIGURE 3.27: Comparison between Figures 3.24, 3.26 and 3.5

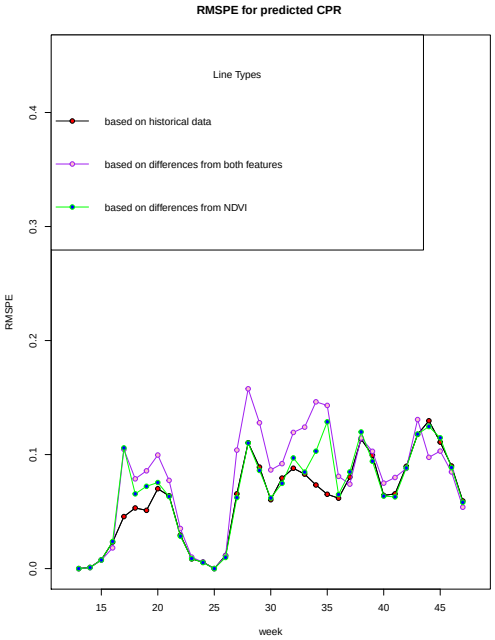


FIGURE 3.28: Comparison between Figures 3.24, 3.25 and 3.5

FIGURE 3.29: Comparison between differences of NDVI and the two other models

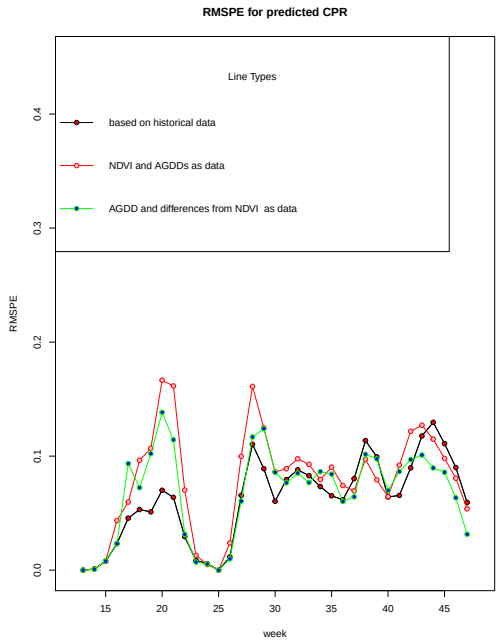


FIGURE 3.30: Comparison between the full data model, the model consists of the AGDDs and the differences of the NDVI and the model based only on historical data

9 Data From Substates

As we mentioned in Chapter 1, the states Nebraska, Illinois and Iowa, have different numbers of meteorological stations. We can profit from these stations by inducing substates from each state via the Thiessen Polygon (see Figures 1.23 , 1.24 and 1.25). This technique will considerably increase the available datasets with the hope to further reduce the prediction errors. So, first we create the new data for our research. Now, we have observations from 10 years, 35 weeks and 37 substates for Nebraska, 33 for Illinois and 23 for Iowa. We will focus on Nebraska and its 37 meteorological stations. We have already worked with Nebraska's Thiessen Polygon, so we know exactly the number of pixels inside each polygon. The procedure of creating the new augmented data set is the same as before, now performed for each substate separately. The new data are denoted by $\{y_{\ell,t,s}\}$, where $\ell = 1,\ldots,10$, $t = 1,\ldots,35$ and $s = 1,\ldots,37$, since s is an additional index added for the substates of Nebraska which correspond to the 37 meteorological stations selected.

Moment Estimator

We give here the new empirical estimators with the addition of the substates.

$$\tilde{\mu}_{i,s} = \frac{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i} y_{\ell,t,s}}{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i}}, \tag{3.38}$$

$$\tilde{\Sigma}^2_{i,s} = \frac{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i}(y_{\ell,t,s}-\tilde{\mu}_{i,s})^\top (y_{\ell,t,s}-\tilde{\mu}_{i,s})}{\sum_{\ell\in T}\sum_{t=1}^{35} w_{\ell,t,i}}, \tag{3.39}$$

where T denotes the training set. Here, in the case of an IMM, $w_{\ell,t,i}$ is estimated as follows:

$$\hat{w}_{\ell,t,i} = \sum_{s=1}^{37}\left(\frac{\sum_{\ell\in T} n_{\ell,s}}{\sum_{\ell\in T}\sum_{s=1}^{37} n_{\ell,s}}\right)\frac{\hat{w}_{t,i} f_i(y_{\ell,t,s};\tilde{\phi}_{i,s})}{\sum_{j=1}^{8}\hat{w}_{j,t} f_j(y_{\ell,t,s};\tilde{\phi}_{j,s})}, \tag{3.40}$$

where $\tilde{\phi}_{i,s} = (\tilde{\mu}_{i,s}, \tilde{\Sigma}^2_{i,s})$ and $n_{\ell,s}$ corresponds to the number of pixels which included in every substate for the year ℓ and the state s.

The coefficient $\frac{\sum_{\ell\in T} n_{\ell,s}}{\sum_{\ell\in T}\sum_{s=1}^{37} n_{\ell,s}}$ corresponds to the percentage of total pixels which can be found inside the s-substate and determines the contribution (weight) of the s-substate to the overall prediction. For the final computation of the RMSPE the same procedure is followed as in Chapter 3.

In order to illustrate the proposed methodology we compare the results from the model with substates, with and without the inclusion of the differences (Figures 3.31 and 3.32 respectively) in the case of the NDVI dataset.

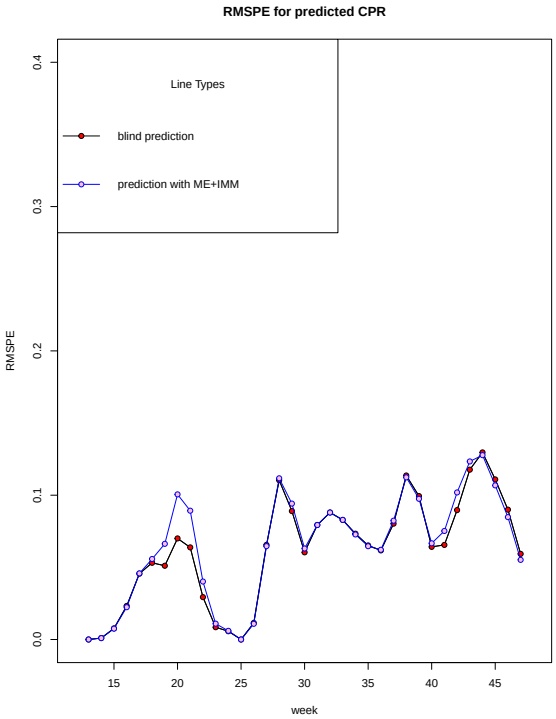


FIGURE 3.31: Mean RMSPE from the model derived from NDVI with substates as data, estimator ME and predictor of type (A). Comparison with Blind Prediction.

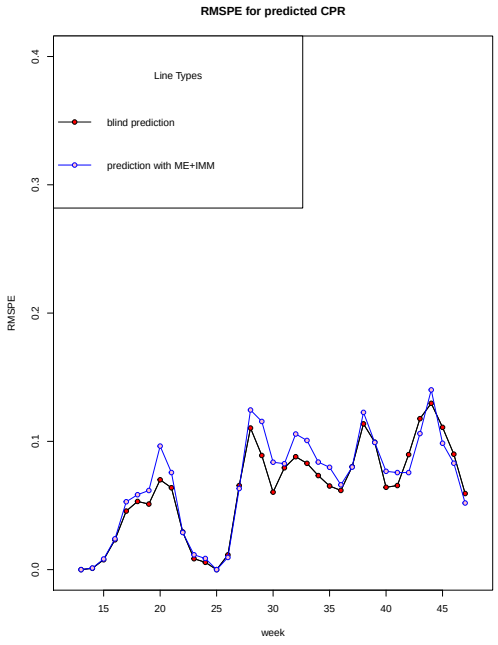


FIGURE 3.32: Mean RMSPE from the model derived from differences of NDVI with substates, estimator ME and predictor of type (A). Comparison with Blind Prediction.

In Figure 3.31 we observe that the RMSPE increases gradually, and reaches its first maximum around week 20, and then decreases until week 25. Then, the prediction error is almost identical to that from the blind prediction.

In Figure 3.32 we observe that the RMSPE values are close and in many cases equal those of the blind predictions. In the previous experiments it was obvious that when we used differences the results were closer to our target which is to beat the predictions based on the historical data. Here, differences give us a good approximation to blind predictions but the RMSPE values are higher than those from the NDVI data.

Maximum Likelihood Estimator

The update equations of the EM algorithm are identical for both types of models, the IMM and the HMM. The resulting MLEs differ since the weights $w_{\ell,t,i}^{(m)}$ are computed differently in these models. For a training set T, the update equations are given by :

$$\mu_{i,s}^{(m+1)} = \frac{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{m} y_{\ell,t,s}}{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{m}}, \tag{3.41}$$

$$\Sigma_{i,s}^{(m+1)} = \frac{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{(m)} (y_{\ell,t,s} - \mu_{i,s}^{(m+1)})^{\top} (y_{\ell,t,s} - \mu_{i,s}^{(m+1)})}{\sum_{\ell \in T} \sum_{t=1}^{35} w_{\ell,t,i}^{(m)}}, \tag{3.42}$$

where T is the training set, t denotes the weeks of our interest $(t = 1, 2, ..., 35)$, i is for the phenological stages $(i = 1, 2, ..., 8)$ and s denotes Nebraska's substates $(s = 1, 2, ..., 37)$. In order to have the recursive evaluation of the parameter estimates, we have to compute the quantities $w_{\ell,t,i}^{(m)}$. There are two different ways for the computation. In the case of an IMM,

$$w_{\ell,t,i}^{(m)} = \sum_{s=1}^{37} \tilde{n_s} \frac{\hat{w}_{t,i} f_i(y_{\ell,t,s}; \phi_{i,s}^{(m)})}{\sum_{j=1}^{8} \hat{w}_{t,j} f_j(y_{\ell,t,j}; \phi_{j,s}^{(m)})}, \tag{3.43}$$

where, $\phi_{i,s} = (\mu_{i,s}^{(m)}, \Sigma_{i,s}^{(m)})$ and $\tilde{n_s} = \frac{\sum_{\ell \in T} n_{\ell,s}}{\sum_{\ell \in T} \sum_{s=1}^{37} n_{\ell,s}}$. In the case of an HMM,

$$w_{\ell,t,i}^{(m)} = \mathbb{P}_{\theta^{(m)}}(X_{\ell,t,s} = i \mid y_{\ell,1:t,s}). \tag{3.44}$$

This is easily solved with the Baum Welch Algorithm and the Forward-Backward equations (see Section 2.4) from Chapter 2.

For the prediction stage we follow the same procedure as we saw analytically in Section 5 of Chapter 3. Below we present the results from the HMM method, where the estimation is performed through filtering and the predictor is of type (A). This method is applied in the NDVI-model.

In Figure 3.33 we compare the results of the above method with those obtained by the Blind Predictions. The comparison shows that the predictions are similar to the blind predictions, except for some weeks at the start of the growing season (18, 19, 20 and 21).

In an attempt to find an effective model that is able to beat the blind predictions we carry out the same experiment in the model which includes the differences from NDVI. In Figure 3.34 we show the comparison between the blind predictions and the predictions under the MLE, estimation through filtering and predictor of type (A). The results are similar to those of the blind predictions and for many weeks are identical. Unfortunately, they cannot surpass the predictions from the historical data.

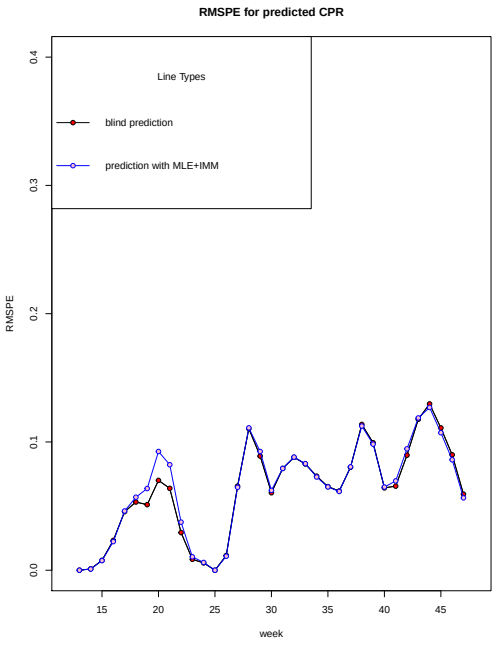


FIGURE 3.33: Mean RMSPE for all 120 scenaria with the HMM, estimator the MLE and predictor of type (A). Comparison with Blind Predictions.

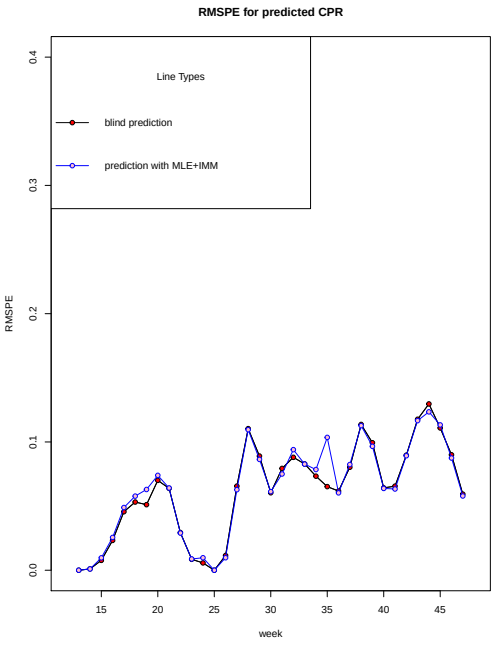


FIGURE 3.34: Mean RMSPE for all 120 scenaria with the HMM, estimator the MLE and predictor of type (A). Comparison with Blind Predictions.

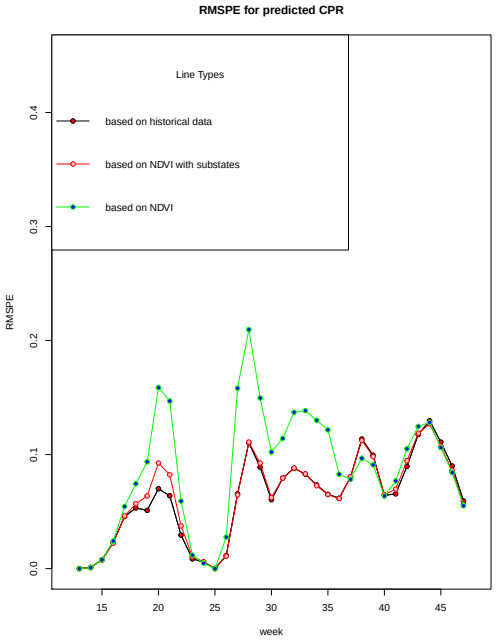


FIGURE 3.35: Comparison between Figures 3.12, 3.33 and 3.5

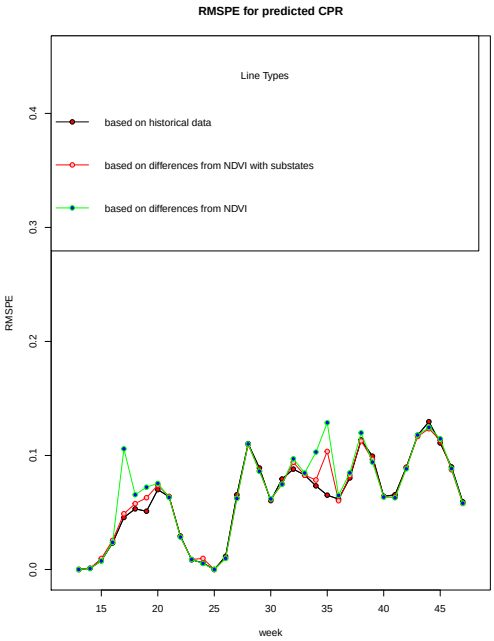


FIGURE 3.36: Comparison between 3.24, 3.34 and 3.5

In Figures 3.35 and 3.36 we compare the results without and with the inclusion of the substates in the case of the NDVI dataset and the differences from NDVI, respectively. We conclude that, for most weeks, the substates achieve lower RMSPE values than the other models, but in some weeks the prediction error is higher. In the next chapter we will use a different approach based on linear regression in order to propose a more effective method to beat the blind predictions.

Chapter 4

Regression

The earliest form of regression was the method of least squares, which was published by Legendre in 1805 [42], and by Gauss in 1809 [35]. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821 [36], including a version of the Gauss–Markov theorem.

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average. A phenomenon renowned as "regression toward the mean" ([54], [34]). For Galton, regression had only this biological meaning ([32], [33]), but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context ([65], [47]). In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925 ([30], [31], [14]). Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

In the 1950s and 1960s, economists used electromechanical desk "calculators" to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression [49].

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

Modeling refers to the development of mathematical expressions that describe in some sense the

behavior of a random variable of interest. This variable may be the price of wheat in the world market, the number of deaths from lung cancer, the rate of growth of a particular type of tumor etc. In all cases, this variable is called the **dependent variable** and usually denoted by **Y**. A subscript on Y identifies the particular unit from which the observation was taken, the time at which the price was recorded, the county in which the deaths were recorded, the experimental unit on which the tumor growth was recorded, and so forth. Most commonly, modeling is aimed at describing how the mean of the dependent variable E(Y) changes with changing conditions; the variance of the dependent variable is assumed to be unaffected by the changing conditions.

Other variables which are thought to provide information on the behavior of the dependent variable are incorporated into the model as predictor or explanatory variables. These variables are called the independent variables and are usually denoted by **X** with subscripts as needed to identify different **independent variables**. Additional subscripts denote the observational unit from which the data were taken. In classical regression X_s are assumed to be known constants.[50].

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

1 The linear Regression Model

Given a data set $\{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ϵ — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form :

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

where ^{T} denote the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in matrix notation as :

$$y = X\beta + \epsilon, \tag{4.2}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The random errors ϵ_i have zero mean and are assumed to have common variance σ^2 and to be mutually independent. Since the only random element in the model is ϵ_i, these assumptions imply that the y_i also have common variance σ^2 and are mutually independent. For purposes of inference, the random errors are usually assumed to be normally distributed, which implies that the y_i are also normally distributed. The random error assumptions are frequently stated as :

$$\epsilon_i \sim NID(0, \sigma^2), \tag{4.3}$$

where NID stands for "normally and independently distributed". The quantities in parentheses denote the mean and the variance, respectively, of the normal distribution.

2 A simple regression approach

We start by considering predictions from a simple regression approach. For each fixed week t, the model is trained from the dataset $\{W_{\ell,t,i}\}_{\ell=1}^{L}$ by using all the available CPR data from different years in the training set corresponding to week t and state i. We define as $N_{\ell,t}$ the data from the NDVI feature and as $T_{\ell,t}$ the data from the feature AGDD, in year ℓ and week t. For $\ell = 1,\ldots,L$ and $i = 1,\ldots,8$ we initially assume that

$$W_{\ell,t,i} = \alpha_{t,i} + \beta_{t,i}N_{\ell,t} + \gamma_{t,i}T_{\ell,t} + \epsilon_{\ell,t,i}, \tag{4.4}$$

where for each fixed t ($t = 1,\ldots,35$) and i, $\{\epsilon_{\ell,t,i}\}_{\ell=1}^{L}$ are independent normal errors with common variance $\sigma_{t,i}^2$. If $\widehat{\alpha}_{t,i}$, $\widehat{\beta}_{t,i}$, $\widehat{\gamma}_{t,i}$ are the estimated coefficients in the regression model given by (4.4), then the predicted values in the testing set $\widehat{W}_{\ell,t,i}$ are given by

$$\widehat{W}_{\ell,t,i} = \widehat{\alpha}_{t,i} + \widehat{\beta}_{t,i}N_{\ell,t} + \widehat{\gamma}_{t,i}T_{\ell,t}. \tag{4.5}$$

Since the weights take values in $[0,1]$ and their sum for all different i, equals unity some restrictions should be taken into account. Our first approach to this problem was the simplest one possible. In particular, we projected to zero, any predicted weight which attains a negative value and normalised to one all the positively predicted weights. Many other strategies, including transformations and multivariate considerations of the weights will be considered in the future, so as to obtain a more effective method of weights' prediction.

By following the above procedure, for each scenario corresponding to the chosen training years, we have derived the prediction of $W_{\ell,t,i}$, in year ℓ, week t and state i. Then, we use the $\hat{W}_{\ell,t,i}$ in order to compute the RMSPE, in the same way as in Equations (3.23) and (3.24) of Chapter 3. Below, we present the results that stemmed from three different regressions. First, in Figure 4.1, we use as explanatory variable only the NDVI. Then, in Figure 4.2, we use as explanatory variable only the AGDD and finally, in Figure 4.3, we use both of these features as explanatory variables and we compare this curve with those from Figures 4.2 and 3.5. Also, note that, if no feature is included in the model, then fitting only with the constant corresponds to the blind predictions, based on the means from the historical data.

Since the NDVI differences performed better than the raw NDVI values in the case of the IMM and HMM, we have also considered the combination of AGDD and NDVI differences in the linear regression setting (see Figure 4.4).

From these Figures we can conclude that Regression works and achieves our target to beat the

predictions based only on historical data. The results show, especially in the case of the AGDD being used as a single explanatory variable, that we finally have a model which can surpass the blind predictions. It is also interesting to notice that the inclusion of the NDVI only deteriorates predictions (see Figure 4.3). Many tests could be further performed to transform the NDVI in such a way that this feature becomes more informative.

We would like to carry out more experiments within the Regression approach. So far, we have achieved our initial target but we also know that we can still produce better results, maybe under a more complex regression model or with some transformations of the data. Because of time limitations regarding the completion of this master thesis, we did not explore further possibilities. We hope to achieve this goal in the future.

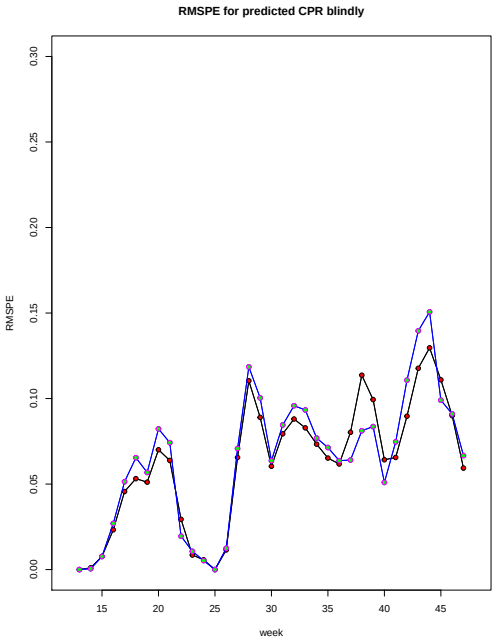


FIGURE 4.1: Mean RMSPE from simple linear regression with NDVI as explanatory variable.

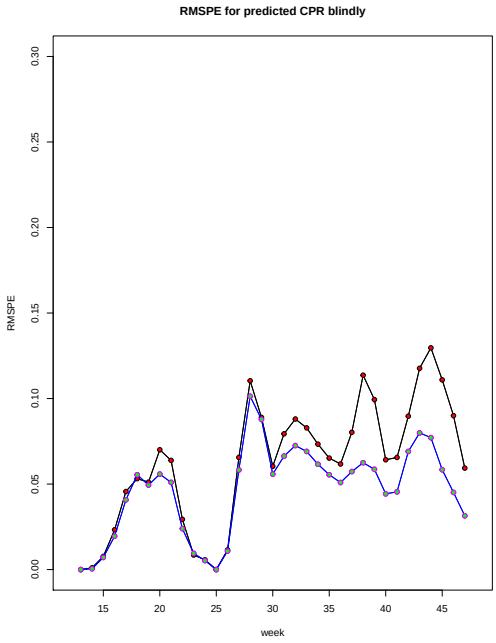


FIGURE 4.2: Mean RMSPE from simple linear regression with AGDD as explanatory variable.

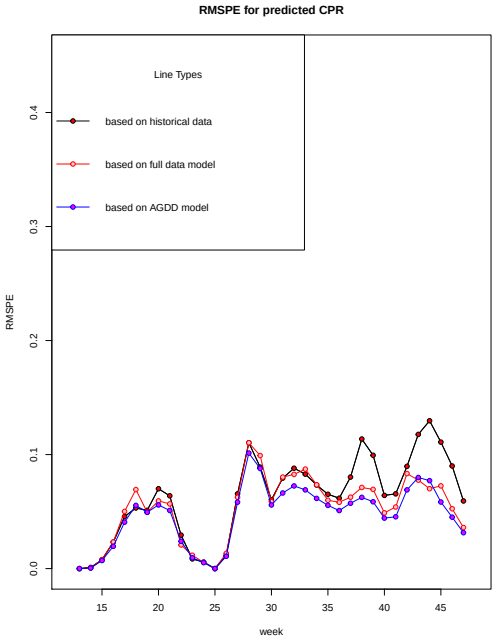


FIGURE 4.3: Mean RMSPE when both NDVI and AGDD as explanatory variables are included in the regression model. Comparison with blind predictions and with predictions from the AGDD model.

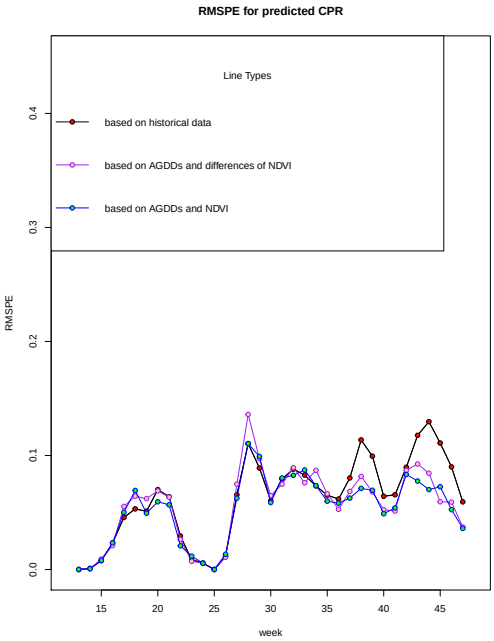


FIGURE 4.4: Comparison between the full data model, the model including AGDD and differences of NDVI as explanatory variables and the model based only on historical data

Bibliography

[1] . URL https://modis.gsfc.nasa.gov/.

[2] . URL https://terra.nasa.gov/about/terra-instruments/modis.

[3] . URL https://ladsweb.modaps.eosdis.nasa.gov/.

[4] . URL https://wiki.landscapetoolbox.org/doku.php/Home.

[5] . URL https://earthobservatory.nasa.gov/.

[6] . URL https://modis.gsfc.nasa.gov/about/design.php.

[7] . URL https://www.sentinel-hub.com/eoproducts/ndvi-normalized-difference-vegetation-index.

[8] . URL https://fieldcrops.cals.cornell.edu/corn/planting-techniques.

[9] . URL https://ipm.missouri.edu/IPCM/2012/5/Early-Corn-Root-Development.

[10] . URL https://www.pioneer.com/us/agronomy/staging_corn_growth.html.

[11] . URL https://ladsweb.modaps.eosdis.nasa.gov/about/purpose/.

[12] . URL https://www.darpa.mil/about-us/timeline/tiros.

[13] Jeff A. and Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. 4 1997.

[14] John Aldrich. Fisher and regression. *Statistical Science*, 20(4):401–417, November 2005. URL https://www.jstor.org/stable/20061201.

[15] Clement Atzberger. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote Sens*, 5:949–981, 2 2013. URL https://doi.org/10.3390/rs5020949.

[16] L.E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, pages 1–8, 1972.

[17] LE Baum and T Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Institute of Mathematical Statistics*, 33(6):1554–1563, December 1966.

[18] Leonard E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

[19] Leonard E. Baum and G. R. Sell. Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, page 211–227, 1968.

[20] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 44:164–171, 1970.

[21] T.A. Boden, T.R. Karl, C.N. Jr. Williams, and F.T. Quinlan. United states historical climatology network (hcn) serial temperature and precipitation data. 1 1987.

[22] T.A. Boden, T.R. Karl, C.N. Jr. Williams, and F.T. Quinlan. United states historical climatology network (hcn) serial temperature and precipitation data. 1 1990.

[23] Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. (8), 2 2011.

[24] Kurt E. Brassel and Douglas Reif. A procedure to generate thiessen polygons. 7 1979. URL https://doi.org/10.1111/j.1538-4632.1979.tb00695.x.

[25] et al Chang, J.C. Corn and soybean mapping in the united states using modis timeseries data sets. *Agronomy Journal*, 99:1654—1664, 2007.

[26] K.M. De Beurs and G.M Henebry. Spatio-temporal statistical methods for modelling land surface phenology. *Phenological Research: Methods for Environmental and Climate Change Analysis*, 43: 177–208, 9 2009.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–22, 1977.

[28] D. R. Easterling, E.H. Mason T. R. Karl, P. Y. Hughes, and D. P. Bowman. United states historical climatology network (u.s. hcn), monthly temperature and precipitation data. 1996.

[29] D. R. Easterling, T. R. Karl, J. H. Lawrimore, and S. A. Del Greco. United states historical climatology network daily temperature, precipitation, and snow data for 1871-1997. 1999.

[30] Ronald A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, June 1922. URL https://www.jstor.org/stable/2341124.

[31] Ronald A. Fisher. Statistical methods for research workers. 1925.

[32] Francis Galton. Typical laws of heredity. *Nature*, 15:492–5, 512–4, 532–3, April 1877.

[33] Francis Galton. Presidential address, section h, anthropology. *Report of the British Association for the Advancement of Science*, 55:1206–14, 1885.

[34] Francis Galton. Kinship and correlation. 4(2):5, 1989.

[35] Carolus Friedricus Gauss. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. 1809.

[36] Carolus Friedricus Gauss. Theoria combinationis observationum erroribus minimis obnoxiae. 1821/1823.

[37] Gil Gullickson. Emergence sets the stage for bumper corn yields next fall. *Agronomy Insider*, 2015.

[38] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. 4(3):223–296, 2010.

[39] Brent N. HOLBEN. Characteristics of maximum-value composite images from temporal avhrr data. *International Journal of Remote Sensing*, 7(11):1417–1434, 4 2007. URL https://doi.org/10.1080/01431168608948945.

[40] P. Y. Hughes, E. H. Mason, T. R. Karl, and W. A. Brower. United states historical climatology network daily temperature and precipitation data. 10 1992.

[41] Erick Larson. Identifying corn reproductive growth stages and management implications. *Mississippi Crop Situation*, 2018.

[42] Adrien Marie Legendre. Nouvelles méthodes pour la détermination des orbites des comètes. 1805.

[43] Gu Leon. Em and hmm.

[44] Susan K. MAXWELL, Jaymie R. Meliker, and Pierre Goovartes. Use of land surface remotely sensed satellite and airborne data for environmental exposure assessment in cancer research. *Journal of Exposure Science and Environmental Epidemiology*, 20(25):176–185, 2 2010.

[45] Andrew McCallum. Hidden markov models baum welch algorithm. 3 2004.

[46] Perry Miller, Will Lanier, and Stu Brandt. Using growing degree days to predict plant stages.

[47] Karl Pearson, G. U. Yule, Norman Blanchard, and Alice Lee. The law of ancestral heredity. *Biometrica*, 2(2):211–236, February 1903.

[48] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, pages 257 – 286, February 1989.

[49] Rodney Ramcharan. Regressions: Why are economists obessessed with them? 43(1), March 2006.

[50] John O. Rawlings, Sastry G. Pantula, and David A. Dickey. Applied regression analysis: A research tool,second edition. 2001.

[51] Bradley C. Reed, Jesslyn F. Brown, Darrel VanderZee, Thomas R. Loveland, James W. Merchant, and Donald O. Ohlen. Measuring phenological variability from satellite imagery. (5):703–714, 10 1994. URL https://doi.org/10.2307/3235884.

[52] J. Ren, Z. Chen, and Q Zhou. Regional yield estimation for winter wheat with modis-ndvi data. *International Journal of Applied Earth Observation and Geoinformation*, 10(4):403–413, 12 2008. URL https://doi.org/10.1016/j.jag.2007.11.003.

[53] Steven W. Ritchie, John J. Hanway, and Garren 0. Benson. How a corn plant develops. *Iowa State University of Science and Technology*, 1993.

[54] Mogull Robert. Second-semester applied statistics. *Kendall/Hunt Publishing Company*, 2004.

[55] Jie Shan, Ejaz Hussain, Kyohyouk Kim, and Larry Biehl. Use of land surface remotely sensed satellite and airborne data for environmental exposure assessment in cancer research. *Photogrammetric Engineering Remote Sensing*, 76:102–104, 2 2010.

[56] Yonglin Shen, Liping Di, Lixin Wu, Genong Yu amd Hong Tang, and Guoxian Yu. Hidden markov models for corn progress percents estimation in multivariate time series. *IEEE*, 8 2012.

[57] Yonglin Shen, Lixin Wu, Liping Di, Genong Yu, Hong Tang, Guoxian Yu, and Yuanzheng Shao. Hidden markov models for real-time estimation of corn progress stages using modis and meteorological data. (8):1734–1753, 4 2013.

[58] Gregory S.McMaster and W.W.Wilhelm. Growing degree-days: one equation, two interpretations. *Agriculture And Forecast Meteorology*, 87:291–300, 12 1997.

[59] Dr. Puteri Suhaiza Sulaiman. Checking plant health through normalized difference vegetation index (ndvi) using mobile phone. *University Putra Malaysia*, 11 2017. URL https://doi.org/10.3390/rs5020949.

[60] Wanxiao Sun, Shunlin Liang, Gang Xu, Hongliang Fang, and Robert Dickinson. Mapping plant functional types from modis data using multisource evidential reasoning. *Remote Sensing of Environment*, 112:1010—1024, 2008.

[61] Usha Ramya Tatavarty. Implementation of numerically stable hidden markov model. 5 2011.

[62] D.L. Trudgill, A. Honek, D. Li, and N.M. van Straalen. Thermal time - concepts and utility. *Annals of Applied Biology*, 146:1–14, 2005.

[63] E. F. Vermote, S. Y. Kotchenova, and J. P. Ray. Modis land surface reflectance science computing facility. 2011.

[64] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11: 95–103, 1983.

[65] G. Udny Yule. On the theory of correlation. *Journal of the Royal Statistical Society*, 60(4):812–854, 1897.

Appendix A

AppendixA

MODISTsp: A Tool for Automatic Preprocessing of MODIS Time Series

1 Introduction

The development of MODISTsp started from modifications of the ModisDownload "R" script by Thomas Hengl (2010), and successive adaptations by Babak Naimi (2014).MODISTsp is a novel "R" package allowing to automatize the creation of time series of rasters derived from MODIS Land Products data. It allows to perform several preprocessing steps on MODIS data available within a given time period.The basic functionalities for download and preprocessing of MODIS datasets provided by these scripts were gradually incremented with the aim of developing a stand-alone application allowing to perform several preprocessing steps on all available MODIS land products by exploiting a powerful and user-friendly GUI front-end.Also,allowing the creation of time series of both MODIS original layers and additional Quality Indicators extracted from the aggregated bit-field QA layer.Finally, allowing the automatic calculation and creation of time series of several additional Spectral Indexes starting form MODIS surface reflectance products.

Required MODIS HDF files are automatically downloaded from NASA servers and resized, reprojected,resampled and processed according to user's choices. For each desired output layer, outputs are saved as single-band rasters corresponding to each acquisition date available for the selected MODIS product within the specified time period. "R" RasterStack objects with temporal information as well as Virtual raster files (GDAL vrt and ENVI META files) facilitating access to the entire time series can be also created.

2 Installation

First,install the stable version of **MODISTsp**,from CRAN：

install.packages(MODISTsp)

Note that if the “GTK+”library is not already installed on the system, installation may fail. In that case, install and load the **gWidgetsRGtk2** library beforehand：

install.packages(gWidgetsRGtk2)

library(gWidgetsRGtk2)

Upon loading **gWidgetsRGtk2** , an error window will probably appear. This signals that library “GTK+” is not yet installed on the system or is not on the PATH. To install it press “OK”. A new window dialog window will appear.Select “Install GTK” and then “OK” . Windows will download and install the “GTK+” library. When it finishes, the RSession should be restarted.

3 Running the tool in Interactive Mode: the MODIStsp GUI

The easiest way to use MODIStsp is to use its powerful GUI (Graphical User Interface) for selection of processing options, and then run the processing.

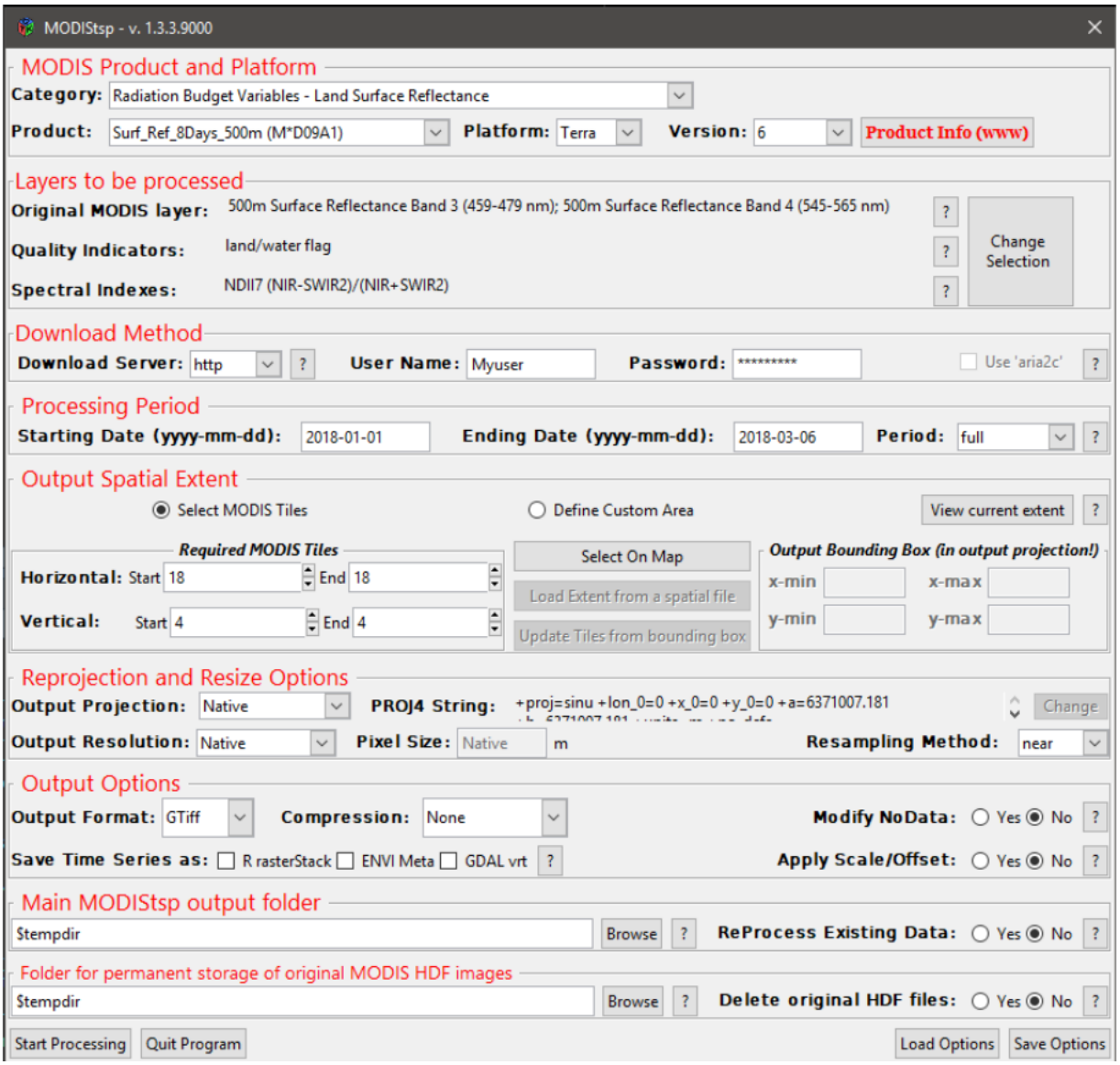
To open the GUI, load the package and launch the MODIStsp function, with no parameters：

library(MODIStsp)

MODIStsp()

This opens a **GUI** from which processing options can be specified and eventually saved (or loaded).At the first execution of MODIStsp, a Welcome screen will appear,signaling that MODIStsp is searching for a valid GDAL installation. Press “OK” and wait for GDAL to be found.

The GUI allows selecting all processing options required for the creation of the desired MODIS time series. The main available processing options are described in detail in the following.



4 Selecting Processing Parameters

4.1 MODIS Product,Platform and Layers

The top-most menus allow to specify details of the desired output time series：

1. Category and Product：

Selects the MODIS product of interest.In our case,select “Radiation Budget Variables–Land Surface Reflectance ”and “Serf_Ref _Daily _250m”(M∗D09GQ) respectively.

2. MODIS platform(s)：

Selects if only TERRA, only AQUA or Both MODIS platforms should be considered for download and creation of the time series.In that case select TERRA.

3. version：

Selects whether processing version 5 or 6 (when available) of MODIS products has to be processed.

After selecting the product and version, clicking the "Change Selection "button opens the Select Processing Layers GUI panel, from which the user must select which MODIS original layers and/or derived Quality Indexes (QI) and Spectral Indexes (SI) layers should be processed.
The left-hand frame allows to select which original MODIS layers should be processed.There are many choices but select：
250m Surface Reflectance Band 1 (620-670 nm)
250m Surface Reflectance Band 2 (841-876 nm)

The central frame allows to select which Quality Indicators should be extracted from the original MODIS Quality Assurance layers.Here,choose：
cloud state
atmospheric correction performed

For MODIS products containing surface reflectance data, the right-hand frame allows to select which additional Spectral Indexes should be computed.The index that we are interested for is：
NDVI (NIR-Red)/(NIR+Red)

Clicking "Done ! "returns to the main.

4.2 Download Method

Select the method to be used for download. Available choices are：

http :

Download through ftp from NASA lpdaac http archive (http://e4ftl01.cr.usgs.gov). This requires providing a user name and password, which can be obtained by registering an account at the address https://urs.earthdata.nasa.gov/profile

offline :

This option allows to process/reprocess HDF files already available on the user's PC without downloading from NASA.Is useful if the user already has an archive of HDF images, or to reprocess data already downloaded via MODIStsp to create time series for an additional layer
Checking the **use_aria2c** option allows to accelerate the download from NASA archives.

My choice for download was http method.

4.3 Processing Period

Specify the starting and ending dates to be considered for the creation of the time in the series corresponding fields. Dates must be provided in the yyyy–mm–dd format.
To specify the dates,visit the VegScape (Vegetation Condition Explorer) website.The only thing you have to do,is to decide type,period,year and date.I was interested about type NDVI,weekly period,the years(2002-2011)and I ran the dates corresponding to the 13th to 47th week.
The **Period** drop-down menu allows to choose between two options :

1.full : all available images between the starting and ending dates are downloaded and processed

2.seasonal : data is downloaded only for one part of the year, but for multiple years.This allows to easily process data concerning a particular season of interest.

My choice was **full**.

4.4 Spatial Extent

Allows to define the area of interest for the processing,two main options are possible **Select MODIS Tiles** and **Define Custom Area**.
Define Custom Area : specify a custom spatial extent for the desired outputs either by :

1.Manually inserting the coordinates of the Upper Left and Lower Right corners of the area of interest in the **Bounding Box** frame.

2.pressing the "Load Extent from a Spatial File "and selecting a raster or vector spatial file.

3.pressing the "Select on Map "button.

The procedure I followed was to use the first of these methods and then press the "Update Tiles From Bounding Box "button. In this way, the prices in category **Select MODIS Tiles** were automatically adjusted.But before completing this procedure I added the next category and then returned to the completion of this.

4.5 Reprojection and Resize

Specify the options to be used for reprojecting and resizing the MODIS images.There are two options at the window :

- **Output Projection**: select either the Native MODIS projection (Default) or specify a user-defined one
- **Output Resolution , Pixel Size and Reprojection Method**: specify whether output images should inherit their spatial resolution from the original MODIS files, or be resampled to a user-defined resolution.

For the first option I did the following procedure. I chose **User Defined** , and then I inserted a valid "Proj4 "string for LATLON WGS84 in the pop-up window.For the second I chose the **Native** option.

4.6 Output Options

Several processing options can be set using check-boxes:

Output Files Format:Two of the most commonly formats used in remote sensing applications are available at the moment: "ENVI "binary and
"GeoTiff ".My selection was **Geotiff**.

Save Time Series as : Specify if virtual multitemporal files should be created. These virtual files allow access to the entire time series of images as a single file without the need of creating large multitemporal raster images. Available virtual files formats are "R"rasterStacks, ENVI meta-files and GDAL "vrt "files.The only one that I chose was "R "rasterStacks.

Modify No Data : Specify if NoData values of MODIS layers should be kept at their original values, or changed to those specified within the "MODIStsp_Products_Opts "XML file.

Apply Scale/Offset: Specify if scale and offset values of the different MODIS layers should be applied.

The last two choices were not selected.

4.7 Main MODIStsp Output Folder

Select the main folder where the pre-processed time series data will be stored.
The **Reprocess Existing Data** check-box allows to decide if images already available should be reprocessed if a new run of MODIStsp is launched with the same output folder. I set it to "No "which means that MODIStsp skips dates for which output files following the MODIStsp naming conventions are already present in the output folder. This allows to incrementally extend MODIS time series without reprocessing already available dates.

4.8 Folder for permanent storage of original MODIS HDF images

Select the folder where downloaded original MODIS HDF files downloaded from NASA servers will be stored.
The **delete original HDF files** check-box allows also to decide if the downloaded images should

be deleted from the file system at the end of the processing.To avoid accidental file deletion, this is always set to “No ”by default and I kept it in this option.

4.9 Saving and Loading Processing Options

Specified processing parameters can be saved to a JSON file for later use by clicking on the Save Options button.Previously saved options can be restored clicking on the Load Options button and navigating to the previously saved JSON file. I did not choose these options.

4.10 Starting the processing

Click on **Start Processing**. MODIStsp will start accessing NASA servers to download and process the MODIS data corresponding to our choices.For each date of the specified time period, MODIStp downloads and preprocesses all hdf images required to cover the desired spatial extent. Informative messages concerning the status of the processing are provided on the console, as well as on a self-updating progress window.The processed time series are saved in specific subfolders of the main selected output folder.

Appendix B

Appendix B

Table 1. Basic information of selected meteorological stations. Meteorological data include ID, name, and the geographic coordinate (i.e., latitude, longitude, and elevation) of each station. SA means US state abbreviations.

ID	SA	Name	Lat($^\circ N$)	Lat($^\circ W$)	Elev(m)
130112	IA	ALBIA 3 NNE	41.07	92.79	268.2
130133	IA	ALGONA 3 W	43.07	94.31	377.6 49
130600	IA	BELLE PLAINE	41.88	92.28	246.9
131402	IA	CHARLES CITY	43.08	92.67	309.1
131533	IA	CLARINDA	40.72	95.02	298.7
131635	IA	CLINTON 1	41.79	90.26	178.3
132724	IA	ESTHERVILLE 2 N	43.43	94.82	396.8
132789	IA	FAIRFIELD	41.02	91.96	225.6
132864	IA	FAYETTE	42.85	91.82	344.4
132977	IA	FOREST CITY 2 NNE	43.28	93.63	396.2
132999	IA	FORT DODGE 5NNW	42.58	94.2	347.5
134063	IA	INDIANOLA 2W	41.37	93.65	287.1
134142	IA	IOWA FALLS	42.52	93.25	344.4
134735	IA	LE MARS	42.78	96.15	364.2
134894	IA	LOGAN	41.64	95.79	301.8
135769	IA	MT AYR	40.71	94.24	359.7
135796	IA	MT PLEASANT 1 SSW	40.95	91.56	222.5
135952	IA	NEW HAMPTON	43.05	92.31	349.9

ID	SA	Name	Lat($^\circ N$)	Lat($^\circ W$)	Elev(m)
137147	IA	ROCK RAPIDS	43.43	96.17	411.5
137161	IA	ROCKWELL CITY	42.4	94.63	364.2
137979	IA	STORM LAKE 2 E	42.63	95.17	434.3
138296	IA	TOLEDO 3N	42.04	92.58	289.3
138688	IA	WASHINGTON	41.28	91.71	210.3
110072	IL	ALEDO	41.2	90.75	219.5
110187	IL	ANNA 2 NNE	37.48	89.23	195.1
110338	IL	AURORA	41.78	88.31	201.2
111280	IL	CARLINVILLE	39.29	89.87	189.3
111436	IL	CHARLESTON	39.48	88.17	198.1
112140	IL	DANVILLE	40.14	87.65	170.1
112193	IL	DECATUR WTP	39.83	88.95	189
112483	IL	DU QUOIN 4 SE	37.99	89.19	128
113335	IL	GALVA	41.17	90.04	246.9
113879	IL	HARRISBURG	37.74	88.52	111.3
114108	IL	HILLSBORO	39.15	89.48	192
114198	IL	HOOPESTON 1 NE	40.47	87.66	216.4
114442	IL	JACKSONVILLE 2E	39.73	90.2	185.9
114823	IL	LA HARPE	40.58	90.97	210.3
115079	IL	LINCOLN	40.15	89.34	177.7
115326	IL	MARENGO	42.29	88.65	248.4
115712	IL	MINONK	40.91	89.03	228.6
115768	IL	MONMOUTH	40.92	90.64	227.1
115833	IL	MORRISON	41.80	89.97	183.8
115901	IL	MT CARROLL	42.1	89.98	195.1
115943	IL	MT VERNON 3 NE	38.35	88.85	149.4
116446	IL	OLNEY 2S	38.7	88.08	146.3
116526	IL	OTTAWA 5SW	41.33	88.91	160
116558	IL	PALESTINE	39	87.62	140.2
116579	IL	PANA 3E	39.37	89.02	213.4
116610	IL	PARIS WTR WKS	39.64	87.69	207.3
116910	IL	PONTIAC	40.89	88.64	198.1
117551	IL	RUSHVILLE	40.12	90.56	201.2
118147	IL	SPARTA 1 W	38.12	89.72	163.1

ID	SA	Name	Lat($^\circ N$)	Lat($^\circ W$)	Elev(m)
118740	IL	URBANA	40.08	88.24	219.8
118916	IL	WALNUT	41.55	89.6	210.3
119241	IL	WHITE HALL 1 E	39.44	90.38	176.8
119354	IL	WINDSOR	39.44	88.6	210.3
250130	NE	ALLIANCE 1WNW	42.11	102.9	1,217.4
250375	NE	ASHLAND NO 2	41.04	96.38	326.1
250435	NE	AUBURN 5 ESE	40.37	95.75	283.5
250640	NE	BEAVER CITY	40.13	99.83	658.4
251145	NE	BRIDGEPORT	41.67	103.1	1,117.4
251200	NE	BROKEN BOW 2 W	41.41	99.68	762
252020	NE	CRETE	40.62	96.95	437.4
252100	NE	CURTIS 3NNE	40.67	100.49	829.4
252205	NE	DAVID CITY	41.25	97.13	490.7
252820	NE	FAIRBURY 5S	40.07	97.17	411.5
252840	NE	FAIRMONT	40.64	97.59	499.9
253175	NE	GENEVA	40.53	97.6	496.8
253185	NE	GENOA 2 W	41.45	97.76	484.6
253365	NE	GOTHENBURG	40.94	100.15	787.9
253615	NE	HARRISON	42.69	103.88	1,478.3
253630	NE	HARTINGTON	42.62	97.26	417.6
253660	NE	HASTINGS 4N	40.65	98.38	591.3
253735	NE	HEBRON	40.18	97.59	451.1
253910	NE	HOLDREGE	40.45	99.38	707.1
254110	NE	IMPERIAL	40.52	101.66	999.7
254440	NE	KIMBALL 2NE	41.25	103.63	1,435
254900	NE	LODGEPOLE	41.15	102.64	1,168
254985	NE	LOUP CITY	41.28	98.97	627.3
255080	NE	MADISON	41.83	97.45	481.6
255310	NE	MC COOK	40.22	100.62	796.1
255470	NE	MERRIMAN	42.92	101.71	986
255565	NE	MINDEN	40.52	98.95	658.4

ID	SA	Name	Lat($^\circ N$)	Lat($^\circ W$)	Elev(m)
256135	NE	OAKDALE	42.07	97.97	521.2
256570	NE	PAWNEE CITY	40.12	96.16	378
256970	NE	PURDUM	42.07	100.25	819.9
257070	NE	RED CLOUD	40.1	98.52	524.3
257515	NE	SAINT PAUL 4N	41.27	98.47	541
257715	NE	SEWARD	40.9	97.09	438.9
258395	NE	SYRACUSE	40.68	96.19	335.3
258465	NE	TECUMSEH 1S	40.35	96.19	338.3
258480	NE	TEKAMAH	41.79	96.23	338.3
258915	NE	WAKEFIELD	42.27	96.86	423.7