Evaluating Time-Series Smoothing Algorithms for Multi-Temporal Land Cover Classification

Brandon M. Wheeler

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Geography

Yang Shao, Chair

John S. Iiames

James B. Campbell

April 28, 2015

Blacksburg, VA

Keywords: remote sensing, smoothing, MODIS, time-series, land cover classification

Evaluating Time-Series Smoothing Algorithms for Multi-Temporal Land Cover Classification

Brandon M. Wheeler

ABSTRACT

In this study we applied the asymmetric Gaussian, double-logistic, and Savitzky-Golay filters to MODIS time-series NDVI data to compare the capability of smoothing algorithms in noise reduction for improving land cover classification in the Great Lakes Basin, and providing groundwork to support cyanobacteria and cyanotoxin monitoring efforts. We used inter-class separability and intra-class variability, at varying levels of pixel homogeneity, to evaluate the effectiveness of three smoothing algorithms. Based on these initial tests, the algorithm which returned the best results was used to analyze how image stratification by ecoregion can affect filter performance.

MODIS 16-day 250m NDVI imagery of the Great Lakes Basin from 2001-2013 were used in conjunction with National Land Cover Database (NLCD) 2006 & 2011 data, and Cropland Data Layers (CDL) from 2008 to 2013 to conduct these evaluations. Inter-class separability was measured by Jeffries-Matusita (JM) distances between selected land cover classes (both general classes and specific crops), and intra-class variability was measured by calculating simple Euclidean distance for samples within a land cover class. Within the study area, it was found that the application of a smoothing algorithm can significantly reduce image noise, improving both inter-class separability and intra-class variability when compared to the raw data. Of the three filters examined, the asymmetric Gaussian filter consistently returned the highest values of inter-class separability, while all three filters performed very similarly for within-class variability. The ecoregion analysis based on the asymmetric Gaussian dataset indicated that the scale of study area can heavily impact within-class separability. The criteria we established have potential for furthering our understanding of the strengths and weaknesses of different smoothing algorithms, thereby improving pre-processing decisions for land cover classification using time-series data.

Acknowledgements

I would like to thank Dr. Shao, Dr. Iiames, and Dr. Campbell, for having the patience to work with me through the trials and tribulations of the past 2 years. It has been a difficult time in my life, but an experience I wouldn’t trade for anything.

I thank my family and friends for providing the very necessary support and love a task such as this required.

I thank the Virginia Tech Department of Geography faculty and staff for providing the education, experience, and encouragement required to enter and succeed in graduate school.

I thank the United States Environmental Protection Agency for providing a means for me to continue attending graduate school, and placing their trust in me to develop research useful to future projects at a greater scale.

Table of Contents

List of Figures

List of Tables

Introduction

Harmful algal blooms (HABs) have been estimated to deteriorate freshwater supplies at a cost of $2.2 billion annually within the United States. These damages negatively affect recreational water use, water front real estate value, biodiversity, and drinking water treatment costs (Dodds et al. 2009). The rapid spread of cyanobacteria and cyanotoxins that result from HABs holds particular consequence: cyanotoxins are included in the Safe Drinking Water Act's Contaminant Candidate List, and are correlated with the degradation of potable water, human respiratory irritation, and other human illnesses resulting from ingestion or skin exposure. The severity of complications which result from HABs warrant the establishment of a regularly updated cyanobacteria assessment network, which is currently being considered by the US Environmental Protection Agency (EPA Internal Report, 2014). Algal blooms result from a combination of factors including excess nutrients (Michalak et al., 2013), environmental conditions related to temperature, light, and stratification (Paerl and Huisman, 2008), and changes in land use practices related to urbanization and modern agricultural processes related to sediment and nutrient buildup in watersheds (Lunetta et al., 2010). A method for determining trends in cyanobacteria outbreaks utilizing land cover (LC) data to determine correlations could be valuable. However, existing references may be inadequate due to issues in temporal resolution. For example, the usability of National Land Cover Database (NLCD) reference data for understanding and monitoring HAB trends is suspect due to the 5-year lapse in time between each product. An approach using time-series Moderate Resolution Imaging Spectroradiometer (MODIS) Normalized Difference Vegetation Index (NDVI) data, available in 16 day composites, for LC characterization may augment research capabilities by filling in previous gaps.

Time-series NDVI data have been proven to be useful for LC characterization (Xiao et al., 2006) and change detection analysis (Lunetta et al., 2006). For LC characterization, monthly NDVI composite data derived from the Advanced Very High Resolution Radiometer (AVHRR) sensor have been used as a primary input to generate the 1km resolution global LC database (Loveland et al., 2000). More recently, global and regional LC mapping efforts have focused on using time-series NDVI or other vegetation indices from MODIS (Friedl et al., 2002; Knight et al., 2006). Compared to the traditional single image 'snap-shot' classification approach, the use of time-series remote sensing data or multi-temporal image classification often improves classification accuracy by incorporating both spectral and temporal profiles (Shao and Lunetta, 2011). Vegetative phenological differences observed in time-series data can be distinguished using separability measures initially designed for optical remote sensing data. LC derived from these separability measures provide unique opportunities to improve classification performance. For example, MODIS time-series data have been successfully used for crop-specific mapping in which corn and soybean can be delineated by examining NDVI temporal profiles from the crop growing season (Wardlow et al., 2007; Lunetta et al., 2010).

One of the main challenges in time-series remote sensing data analysis is dealing with image noise such as pseudo-hikes and pseudo-lows caused by cloud and shadow issues, weather impacts, and sensor-introduced noises (Goward et al., 1991). The presence of such outliers may add uncertainty in LC mapping and change detection efforts, therefore necessitating data cleaning, the process of filtering and smoothing anomalous time-series data (Lunetta et al., 2006). AVHRR and MODIS vegetation index products alleviate cloud/shadow noise through maximum value compositing (MVC), a common practice where only the maximum value within the composite period (e.g., 8 or 16 days) for a pixel is retained. The MVC algorithm, when applied to 16-day compositing window, can substantially reduce impacts from cloud cover. However, cloud contamination and atmospheric variability cannot be entirely removed through MVC, thus time-series vegetation index data may still have outliers that need to be removed by additional data smoothing or filtering (Holben, 1986; Lunetta et al., 2006).

A large number of data smoothing algorithms have been developed to reduce noises in remote sensing time-series data. For AVHRR NDVI data, Ma and Veroustraete (2006) developed a smoothing method using a mean-value iteration filter. Li and Kafatos (2000) utilized weighted regression smoothing. Viovy et al. (1992) offered an alternative to MVC for pre-processing using Best Index Slope Extraction. Swets et al. (1999) provides example of smoothing using a weighted least-squares approach. For MODIS NDVI data, Sakamoto et al. (2005) developed smoothing methods using wavelet and Fourier transforms. Bruce and Mathur (2006) developed a new wavelet-based feature extraction technique. Chen et al. (2004) modified the Savitzky–Golay filter to adapt to the upper envelope of the vegetation index data, an adaptation especially important because NDVI signals are often negatively biased. Lunetta et al. (2006) applied inverse Fourier transformation to estimate new NDVI values for outliers, while retaining the original NDVI values for cloud-free and good quality pixels. Additionally, Jonsson & Eklundh (2002, 2004) integrated Savitzky-Golay, asymmetric Gaussian, and double-logistic algorithms into a TIMESAT package which can be applied for smoothing time-series NDVI data from a variety of sensors, including AVHRR, MODIS, VEGETATION, and MERIS.

Although a variety of algorithms have been examined and implemented for smoothing time-series data, comparisons of the relative effectiveness of each algorithm is difficult due to lack of *in-situ* reference data (Hird and McDermid, 2009) and standard statistical measures (Atzberger and Eilers, 2011). Most published studies compare different smoothing algorithms for how well they derive phenological metrics such as start of season (SOS) and end of season (EOS) (e.g., Beck et al. 2006, Hird and McDermid, 2009; Atkinson et al., 2012). Little research has been conducted for the purpose of multi-temporal image classification. In a recent paper, Atzberger and Eilers (2011) proposed possible measures for evaluating the effectiveness of smoothing algorithms: 1) increased inter-class separability; and 2) reduced within-class variability of pseudo-invariant targets. For example, Jeffries–Matusita distance measures can be used to estimate inter-class separability, and Euclidean distance measures can be used to estimate within-class variability. Atzberger and Eilers (2011), however, did not conduct actual comparison of different smoothing algorithms because the main purpose of their paper was to highlight potential evaluation methods.

The main objective of this study was to compare three smoothing algorithms for multi-temporal land cover mapping applications in the Great Lakes Basin (GLB) region (Unites States section only). The three algorithms chosen were the Savitzky-Golay filter, the double-logistic filter, and the asymmetric Gaussian filter. Hird and McDermid (2009) used modeled NDVI data and simulated noise to compare six smoothing algorithms for the purpose of establishing phenological metrics, and found that the Savitzky-Golay, double-logistic, and asymmetric Gaussian filter were each very competitive options for restoring NDVI points to their true values. Their research on the strength of these filters indicates that they may be useful for establishing new criteria with which to evaluate smoothing algorithms. Previous research has successfully been conducted by the EPA utilizing the Fourier transformation (Lunetta et al., 2010), but while establishing criteria with which to compare smoothing algorithms the three filters elected were also chosen due to their comparative ease of implementation over the Fourier transformation. The specific research tasks were: (1) to evaluate within-class variability for stable deciduous and evergreen forests; (2) to evaluate inter-class separability for pairs of land cover types. Both general LC types (e.g., cropland and wetland) and crop-specific (e.g., corn and soybean) types were considered; (3) to evaluate how smoothing algorithms and pixel homogeneity/heterogeneity interact to affect within-class variability and inter-class separability; and (4) to further evaluate inter-class separability by including ecoregion stratification. A 250m 16-day MODIS NDVI dataset for the GLB from 2001-2013 was used for this analysis. Understanding multi-temporal MODIS smoothing algorithm performance for different LC types within the GLB may provide a better understanding of land use and land change within the region, potentially assisting further research of ecosystem services.

Study Area

The GLB consists of eight states in the US as well as part of the Province of Ontario, Canada (Figure 1.1). The GLB is an important region in North America from both an economic and ecological standpoint. The ecosystem supported by the Great Lakes features a range of habitats: forests, grasslands, and prairies, as well as dunes, marshes, and wetlands along the lakeshores. Lake Superior, Michigan, Huron, Ontario, and Erie together make up the largest total surface area of freshwater on Earth, providing drinking water for over 40 million people as well as 56 billion gallons of water each day for industrial, municipal, and agricultural use. Almost a third of the area of the Great Lakes Basin is devoted to farmland, producing 7% of the US and 25% of the Canadian total crop yield, respectively (EPA, 2008).

Methods

Data

Primary Data:

MODIS MOD13Q1 vegetation index data were obtained from NASA Reverb website (http://reverb.echo.nasa.gov/) for the GLB. The MOD13Q1 product includes 250m resolution Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and quality assessment (QA) information for both VI sets. MOD13Q1 data is calculated from atmospherically-corrected surface reflectance values and delivered as a 16-day composite image. For each composite period, the initial data consisted of four separate MODIS scenes due to the large study area.

Reference Data:

To segment crop from non-crop LC, National Landcover Database (NLCD) 2006 and 2011 data were acquired from the Multi-Resolution Land Characteristics Consortium (http://www.mrlc.gov/). Both the NLCD 2006 and 2011 datasets, both at a 30m spatial resolution, are developed with Landsat data as the primary input, using a 16-class land cover classification scheme, and a decision-tree model for classification. The user accuracy of NLCD 2006 approaches was found to be 78% both overall and within the GLB (Wickham et al., 2013). The accuracy of NLCD 2011 is currently being assessed.

To provide a crop validation layer, we obtained a Cropland Data Layer (CDLs) from CropScape (http://nassgeodata.gmu.edu/CropScape/), developed by the United States Department of Agriculture National Agricultural Statistics Service (USDA-NASS). Complete coverage for the US is available from 2008 to 2014. The spatial resolution of the CDL varies from 30m to 56m, depending on the imagery source (Boryan et al., 2011). Specific crop types such as corn, soybean, and wheat are mapped at an annual interval, with classification accuracies exceeding 80% for corn and soybean in most states.

Ecoregion Data:

An Omernik Level III (OL3) Ecoregion reference was downloaded from the Environmental Protection Agency (http://www.epa.gov/wed/pages/ecoregions/level_iii_iv.htm). The Omernik Ecoregion Classification system uses delineations are based on criteria such as climate, geology, soils, vegetation, wildlife, and hydrology and has been used to develop four levels within the United States (Omemik, 1987). OL3 is the second finest scale of ecoregion, and is suitable for analysis at the regional level. The GLB makes up 12 of 85 OL3 ecoregions within the United States.

Although the GLB consists of parts of the US as well as Canada, our study focused on the US portion of GLB, due to limited reference data available for Canada.

MODIS data pre-processing and land cover data preparation

The MODIS Reprojection Tool (MRT) was used to generate NDVI mosaics spanning the four MODIS scenes for each 16-day composite. The resultant mosaics were re-projected to an Alber's Equal Area Conic (AEA) projection. Each set of reference/supplemental data (NLCD 2006, NLCD 2011, CDL 2008-2014, Omernik L III Ecoregions) were also re-projected to the AEA projection. For NLCD data, each LC of interest (e.g., deciduous forest or evergreen forest) was extracted as binary from the reference data, and then degraded as land cover proportions within 250m MODIS NDVI grids. A threshold value of 50% proportional cover was used to identify pixels with at least one dominant LC type within the MODIS grid. A threshold value of 90% proportional land cover was used to identify homogenous pixels. The same procedure was applied to CDLs for 2008-2014 to derive corn and soybean pixels. For CDLs, we focused on corn and soybean because these two summer crops have a similar crop calendar, thus they are difficult to classify even with multi-temporal MODIS data (Shao et al., 2010). Creating this diverse library of reference data allowed us to evaluate each smoothing algorithm for both within-class and inter-class variability, for both general and specific LC types, for varying temporal resolutions, and with varying levels of pixel purity.

Time Series data smoothing

We evaluated three smoothing algorithms through the TIMESAT software package: (1) the adaptive Savitzky-Golay filter, (2) the asymmetric Gaussian, and (3) the double-logistic function (Jonsson & Eklundh, 2002). The Savitzky-Golay filter applies a moving window to a given time-series dataset. Within a moving window (e.g., 2n+1 points, n is a user defined window width), a quadratic polynomial function is used to fit all points and then the value of the central point is replaced by the fitted value. The adaptive Savitzky-Golay filter considers that noises from cloud/cloud shadow typically reduce the original NDVI value. Thus, the weight for each point ~~thus~~ can be re-assigned to favor points located above the initial polynomial fit. A new quadratic polynomial function can then be applied to derive NDVI values adapted to the upper envelope of the time-series data.

The asymmetric Gaussian algorithm relies mainly on five parameters to fit time-series data including the time of the minimum or maximum NDVI, the width and flatness of the right side of the function, and the width and flatness of the left side of the function. The double logistic function estimates four parameters to determine the left inflection point, the right inflection point, and rates of changes at two inflection points. Both the asymmetric Gaussian algorithm and the double logistic function are modifications of local model functions, which have been proven to be effective in capturing phenological cycle events defined by NDVI curves (Atkinson et al., 2011).

We first used TIMESAT in the MATLAB environment and visually compared raw NDVI data and smoothed curves (Figures 2.1-2.4). For example, for the adaptive Savitzky-Golay filter, three different window sizes of 9, 11, and 13 were examined to adjust the degree of smoothness of the fitted curve. TIMESAT also provides the option to use QA metadata during processing to check flags and weigh points appropriately based on level of quality. To adapt for the negative bias of noise in NDVI data, the most egregious data points based on QA information were removed. For all algorithms, we examined the number of envelope iterations and adaptation strength to fine-tune the upper envelope fitting. The final configurations were saved as TIMESAT setting files that allow automated data smoothing for a large MODIS NDVI dataset.

The data volume for time-series smoothing was large due to the expansive geographic coverage of our study area and study period (13 year) examined. At this data volume, the application of the smoothing algorithms is computationally expensive. Therefore, we used the Virginia Tech Advanced Research Computing's ITHACA cluster for the image analysis. The initial time-series data set was divided into 10 smaller study regions in order to take advantage of the parallel processing offered by ITHACA. The TIMESAT setting files were automatically generated using Python programming, and submitted to ITHACA via Secure Shell Client. The resultant sets of smoothed files were then merged into a single time series file providing full datasets of the region.

Evaluate within-class variability using Euclidean distance analysis

For a given LC type (e.g., deciduous forest), we expected to see relatively similar NDVI temporal profiles from different locations. We used a simple Euclidean distance measure to characterize the similarity of two selected pixels in the spectral-temporal domain (Figure 2.5).

Where p_i is NDVI value at time *i* for p location and q_i is NDVI value at time *i* for q location. The pair-wise Euclidean distance can be averaged if more randomly selected pixels are involved in the analysis. The averaged Euclidean distance measures were then used as a criteria to compare the similarity of NDVI temporal profiles from raw (i.e., unsmoothed), adaptive Savitzky-Golay filtered, asymmetric Gaussian fitted, and double-logistic function fitted NDVI products. Our hypothesis is that the smoothing algorithm providing the lowest Euclidean distance measurement is likely to be the most reliable option for the study area.

For within-class variability testing, we focused on deciduous and evergreen forest LC. The NLCD 2006 land cover data was used as the primary reference to identify the forest core pixels. Although NLCD has relatively high accuracy (e.g., ~80%) for most applications, edge pixels for LC patches are much more likely to be less pure than core pixels, as well as being more susceptible to change over time. Therefore, we isolated forest core pixels by: 1) removing all patches containing less than five MODIS pixels, 2) removing edge pixels from remaining patches, and 3) within each remaining forest patch, randomly selecting one pixel. Using these procedures, we generated a large number of pixels with a greater likelihood that they were pure and stable over 13-year study period. From these randomly selected pixels, we calculated pair-wise Euclidean distance and derived mean distance measure for each smoothing algorithm. A flowchart reference for the methods employed is depicted in Figure 2.6.

Evaluate inter-class separability for pairs of land cover types

Inter-class separability was evaluated using the Jeffries-Matusita (JM) distance measure. The JM distance has been widely used in remote sensing to measure the average distance between two class density functions (Richards and Jia, 2006; Wardlow et al., 2007). JM distance expresses separability between two classes where the lower-bound (0) indicates identical and impossible to separate two classes, and the upper-bound (2) represents that two classes can be perfectly separated. For this part of the research, we created several sets of JM distances measures. In the first set, we focused on five general land cover types in the GLB: deciduous forest, evergreen forest, pasture/hay, cultivated crops, and wetlands. JM distance measures were generated for all 10 possible land cover pairs for each NDVI input data, respectively. In the second set, we evaluated JM distance for corn-soybean separation based on CDL data, and increased temporal analysis from a single year to 7 years (2008-2014). Corn and soybean are two dominant crop types in the GLB and previous studies have shown that these two land cover types often have similar NDVI temporal profiles, thus, they are not easy to delineate (Shao et al., 2010; Wardlow et al., 2007).

Impacts from pixel homogeneity/heterogeneity

For determining how pixel purity interacts with intra-class variability, the analytical processes for determining within-class and inter-class separability were carried out using pixels with different purity indices generated in the pre-processing steps described above. For example, JM distances were calculated for corn and soybean between 2008-2014 for the raw, asymmetric Gaussian, double-logistic, and Savitzky-Golay NDVI datasets twice: once using the 50% threshold CDL data, and once more using the 90% threshold CDL data. By conducting analysis at varying levels of purity, we can compare scores at an additional level of complexity, and observe how each smoothing algorithm is affected.

Ecoregion stratification

For the data smoothing algorithm with best overall performance for the entire study area, ecoregion stratification was performed and ecoregion-based distance measures were conducted. Using the Omernik Level III Ecoregion as reference, the CDL-derived corn and soybean datasets for each year were further divided into 13 ecoregions present in the study area: Central Corn Belt Plains, Eastern Corn Belt Plains, Eastern Great Lakes Lowlands, Erie Drift Plain, Huron/Erie Lake Plains, North Central Appalachians, North Central Hardwood Forests, Northeastern Highlands, Northern Allegheny Plateau, Northern Lakes and Forests, Southeastern Wisconsin Till Plains, Southern Michigan/Northern Indiana Drift Plains (Figure 2.7). A flowchart reference for the methods employed is depicted in Figure 2.8.

Results

Several manually selected forest core pixels were used to calculate pairwise Euclidean distance values to compare how each smoothing algorithm compared to one another, as well as to the raw data (Tables 3.1 & 3.2) for the years of 2007-2009. For both deciduous and evergreen pixels, the smoothing algorithms generated more consistent results than the raw data. Average Euclidean distance values were ~30% smaller than the raw data. For the deciduous forest pixels in particular, the smoothed curves appeared to be similar with the difference of mean Euclidean distance measurements less than 2%. For the sampled evergreen data, the difference between algorithms was more pronounced. The mean Euclidean distance for Gaussian smoothing algorithm was ~10% greater than the other two algorithms. It is important to note that manually selected pixels provided some level of evidence that smoothed data were more consistent than raw data with the caveat that the sample size was very small, and only 3 years of the time-series data were examined.

The same Euclidean distance measures were repeated for 2750 randomly selected evergreen and deciduous forest pixels, respectively. With the much larger dataset, the overall trend observed in the initial sample data remained consistent (Table 3.3). For deciduous forest, the smoothed data produced very similar measurements in average Euclidean distance, all were substantially (~70%) smaller than the average Euclidean distance measurement generated from raw NDVI data.

The asymmetric Gaussian generated the highest JM distance values in all cases but cultivated crops - pasture/hay pair where double-logistic resulted in a higher measurement (Table 3.4). Although asymmetric Gaussian consistently performed best, the double-logistic function performed very similarly. The average JM distances across all categories were 1.34 and 1.32 for Gaussian and double-logistic algorithm, respectively. Additionally, the raw data and the Savitzky-Golay filter performed very close to one another, but at significantly lower average JM distances of 1.19 and 1.20 respectively.

Similar trends can also be observed in the LC specific JM measurements. For half of the pairings, all four sets of data returned similar results for separating pasture/hay from evergreen forest (ranging from 1.38-1.40), cultivated crops from evergreen forest (ranging from 1.40-1.41), cultivated crops from deciduous forest (ranging from 1.37-1.41), pasture/hay from wetlands (ranging from 1.33-1.37), and cultivated crops from wetlands (ranging from 1.38-1.41).

The results of JM distance calculations for corn and soybean are presented in Table 3.5 for pixel thresholds of 0.5 and 0.9 respectively. For the average JM scores for all years, we found that asymmetric Gaussian consistently scores highest, followed by double-logistic, raw, and Savitzky-Golay, respectively. When examining individual years, asymmetric Gaussian returns the highest score in every case but 2012. It is worth noting that overall, the difference between asymmetric Gaussian and double logistic scores were around 3-5%. Furthermore, it can be observed that increasing the pixel purity threshold from 50% to 90% for samples resulted in higher JM distance scores in every case by a significant amount: asymmetric Gaussian improved by 16.9%, double-logistic improved by 15.2%, raw improved by 24.3%, and Savitzky-Golay improved by 26.9%.

After evaluating the four datasets for JM distance, it was determined that in both the general LC and crop type analysis, the asymmetric Gaussian filter out-performed the other smoothing algorithms. We further analyzed the Gaussian filter's performance for individual ecoregions (Table 3.6). Although 13 ecoregions were originally chosen, three ecoregions returned results that were particularly suspicious, with JM measures indicating that the classes were perfectly separable. We evaluated the average patch sizes and total samples within those three ecoregions, and it was found that in the Driftless Area, North Central Appalachians, and North Eastern Highlands regions the JM scores were inflated due to very limited corn and soybean areas. These three ecoregions were thus removed from further analysis.

For the remaining 10 ecoregions, the Southern Michigan/Northern Indiana Drift Plains performed the worst for both the 0.5 and 0.9 purity thresholds, corresponding to average JM distance measurements of 1.018 and 1.417 respectively. The Southern Michigan/Northern Indiana Drift Plains have a large number of corn and soybean pixels. The region is described as consisting of many lakes and marshes as well as an assortment of landforms, soil types, soil textures, and land uses. Additionally, the region is noted as being less agriculturally focused than the neighboring Eastern Corn Belt Plains and Central Corn Belt Plains regions, having less nutrient rich soils than the Northern Lakes and Forests, and being better drained than the Huron/Erie Lake Plains (Omernik, 2013). The combination of these characteristics may contribute to a landscape where although there are large amounts of samples to compare, the fragmentation and variation of the landscape may contribute to a more difficult separation between corn and soybean. The best performing ecoregion is the Northern Allegheny Plateau. Average JM distance measures from 2008-2013 were 1.636 and 1.972 for 0.5 and 0.9 purity thresholds, respectively. These statistics are very high. We note that although the Northern Allegheny Plateau has an acceptable amount of corn samples, the region also features the lowest amounts of soybean samples and smallest average patch sizes for all ecoregions in the study area, at both thresholds. The Northern Allegheny Plateau is described as a mosaic of cropland, pastureland, and woodland (Omernik, 2013), so although there may not be a large amount of agriculture lands within the region, the clean physical separation of LC types in the ecoregion likely assists in the spectral separability of corn and soybean in terms of JM distances.

At a 0.5 purity threshold it was observed that the Central Corn Belt Plains, Erie Drift Plains, Northern Lakes and Forests, North Central Hardwood Forests, and Northern Allegheny Plateau all returned very high JM distance scores (over 1.5 on average). From these ecoregions, only the Central Corn Belt Plains and Erie Drift Plains are noted as being agriculturally focused (Omernik, 2013). This indicates separability scores are influenced by factors beyond LC distribution- localized meteorological patterns and phenomena, as well as physical topography may also greatly influence the performance of a smoothing algorithm.

Discussion

There are many factors and criteria involved in our MODIS NDVI data smoothing analysis. The algorithms chosen for comparison were from TIMESAT package, but the methods developed may be applied to any smoothing algorithms for multi-temporal imagery datasets. Now that criteria for evaluation have been established, additional filters to be researched include the

Whittaker filter, and Fourier analysis. The Whittaker filter has shown promise in recent studies for the purposes of removing noise (Atkinson et al., 2012) and is noteworthy for ease of implementation and fast processing times (Atzberger and Eilers, 2011). Fourier analysis has been successfully applied in both the Great Lakes and Albemarle-Pamlico Basins by the EPA (Lunetta et al. 2010), however Chen et al. (2004) found that Fourier-based methods were inferior to other smoothing algorithms over asymmetrical phenological profiles (e.g. agriculture LC types). A logical next step in research would be to compare the effectiveness of the Whittaker filter, Fourier analysis, and asymmetric Gaussian. Doing so would provide insight into how these algorithms compare to one another, while also allowing researchers to determine if the more difficult filters to utilize provide significant benefit over the simpler methods.

The results of within-class variability using Euclidean distance as a measure provided evidence that in the GLB, smoothing can result in significantly more consistent NDVI profiles when dealing with temporally invariant LC such as core deciduous and evergreen forest. This information, although it does not reveal actual improvement in terms of classification accuracy statistics, does show that pre-processing can potentially improve ‘cleanness’ of NDVI time-series data. Furthermore, this measurement does not provide significant value beyond the examination of pseudo-invariant LC types, as it relies on NDVI profiles that are consistent from year to year. For this reason it is suitable for use in comparing smoothing algorithms for core forest pixels, but it is likely not suitable for dynamic agricultural lands where crops may change every year.

JM distances appear to be useful for the separability of one LC class from another. However, these statistics are critically dependent on the accuracy of the LC data used as reference (in this experiment, NLCD and CDL). This analysis demonstrated clear advantages to using the asymmetric Gaussian and double-logistic filters compared to Savitzky-Golay filtered and raw data when comparing inter-class separability, but additional research should be implemented to assess whether those results carry over to actual LC classification. A possible venue for analysis would be to use the filtered NDVI time-series data produced by this analysis in conjunction with multiple methods of LC classification and comparing the resultant maps to reference LC products (NLCD, CDL, etc.). This would allow a more robust assessment of the smoothing filters examined in this research, while also allowing a degree of examination on how the observed JM distances correlate to LC classification.

Some parallels to previous research regarding smoothing algorithms can be noted. Hird and McDermid (2009) found that when determining phenological metrics, the specific metric being observed maintained the greatest influence on filter performance. This finding was reflected through our research, in that differences between each filter were more apparent when measuring inter-class separability than measuring within-class variability. The user must understand the goals at hand when determining which smoothing algorithm to use, because each algorithm performs differently for each criteria.. Furthermore, Hird and McDermid (2009) found that the double-logistic and asymmetric Gaussian filter performed best overall in terms of noise reduction and maintaining phenological metrics under a variety of conditions. In our own research these two algorithms significantly out-performed the Savitzky-Golay filter in terms of inter-class separability, returning similar scores with a small advantage found for the asymmetric Gaussian filter. Finally, both Hird and McDermid (2009) and our own research found that the application of any smoother for removing noise is typically better than using the raw data.

In addition to the methods examined here, there are other methods with which smoothing algorithms can be compared. A study combining the methods examined here with the methods explored by Hird and McDermid (2009) using simulated data and noise could further validate the use of JM distances, Euclidean distance, and other possible criteria for evaluating smoothing algorithms by allowing researchers to determine if there are positive correlations between those measures to statistics such as root mean square error, or phenological metrics.

Conclusion

We designed this analysis to evaluate how smoothing algorithms for NDVI time-series data compare to one another, to increase the understanding of filter selection for pre-processing and potentially to improve the accuracy of multi-temporal LC classification efforts. We examined 13 years of MODIS NDVI data in the GLB using a variety of smoothing algorithms, LC reference datasets, pixel purities, and image stratification. We established criteria with which smoothing algorithms can be effectively evaluated for multi-temporal land cover classification, in lieu of in-situ data or phenological metrics. In all cases our analysis indicated that in the GLB, the application of a smoothing algorithm greatly stabilized within-class variability, providing evidence that smoothing is an important pre-processing step when working with time-series data. The filtered data returned significantly increased inter-class separability, indicating that the application of a data smoother can improve land-cover classification. In our study area, and from the three algorithms examined, the asymmetric Gaussian filter performed the best, scoring very similarly to the double-logistic and Savitzky-Golay filters for within-class variability results, and outperforming them both in terms inter-class separability for both general LC and specific crop types. However, it must be emphasized that this information is only valid for the study area, and does not indicate that the asymmetric Gaussian filter will outperform either the Savitzky-Golay or double-logistic filter in all scenarios.

References

Atkinson, P. M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sensing of Environment*, 123, 400-417.

Atzberger, C., Eilers, P. H. C., 2011, Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. *International Journal of Remote Sensing*, 32(13), 3689–3709.

Beck, P. S. A., Atzberger, C., Kjell, A. H., Johansen, B., Skidmore, A. K., 2006, Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. *Remote Sensing of Environment*, 100, 321-334.

Bruce, L.M., Mathur, A., Byrd, Jr., J.D., 2006. Denoising and wavelet-based feature extraction of MODIS multi-temporal vegetation signatures. *GIScience & Remote Sensing*, 43(1), 67-77.

Chen, J., Jonsson, P., Tamura, M., Gu, Z., Matsuhita, B., Eklundh, L. 2004, A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. *Remote Sensing of Environment*, 91, 332-344.

Dodds, W. K, Bouska, W. W., Eitzmann, J. L., Pilger, T. J., Pitts, K. L., Riley, A. J., Schloesser, J. T., Thornbrugh, D. J., 2009. Eutrophication of U.S. Freshwaters: Analysis of Potential Economic Damages. *Environmental Science & Technology*, 43(1), 12-19.

EPA Internal Report. 2014. Cyanobacteria Assessment Network (CyAN) for freshwater systems: an early warning indicator for toxic and nuisance blooms using ocean color satellites, 2014, Proposal, written communique

EPA. “Great Lakes Monitoring” Environmental Protection Agency http://www.epa.gov/glnpo/monitoring/great_minds_great_lakes/social_studies/without.html

Friedl, M. A., McIver, D.K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., Schaaf, C., 2002. Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment*, 83, 287-302.

Goward, S.N., Markham, B., Dye, D.G., Dulaney, W. Yang, J., 1991, Normalized difference vegetation index measurements from the advanced very high resolution radiometer. *Remote Sensing of Environment*, 35, 257–277.

Hird, J.H., McDermid, G.J., 2009. Noise reduction of NDVI time series: an empirical comparison of selected techniques. *Remote Sensing of Environment*, 113, 248–258.

Holben, B.N., 1986, Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing*, 7, 1417–1434.

Jonsson, P., Eklundh, L., 2002, Seasonality Extraction by Function Fitting to Time-Series of Satellite Sensor Data. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8), 1824-1831.

Knight, J. F., Lunetta, R. S., Ediriwickrema, J., Khorram, S., 2006. Regional Scale Land Cover Characterization Using MODIS-NDVI 250 M Multi-Temporal Imagery: A Phenology-Based Approach. *GIScience & Remote Sensing*, 43(1), 1-23.

Li, Z., Kafatos, M., 2000, Interannual Variability of Vegetation in the United States and its Relation to El Nino/Southern Oscillation. *Remote Sensing of Environment*, 71, 239-247.

Lunetta, R., J. Knight, J. Ediriwickrema, J. Lyon, L. Worthy., 2006. Land-cover Change Detection Using Multi-temporal MODIS NDVI Data. *Remote Sensing of Environment*, 105(2), 142-54.

Lunetta, R. S., Shao, Y., Ediriwickrema, J., Lyon, J. G., 2010. Monitoring agricultural cropping patterns across the Laurentian Great Lakes Basin using MODIS-NDVI data. International Journal of Applied Earth Observation and Geoinformation, 12(2), 81-88

Ma, M., Veroustraete, F., 2006. Reconstructing pathfinder AVHRR land NDVI time-series data for the Northwest of China. *Advances in Space Research*, 37(4), 835-840..

Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N. S., Bridgeman, T. B., Chaffin, J. D, et al. 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meerological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences of the USA*, 110, 6448-6452.

NASA. “Vegetation Indices 16-Day L3 Global 250m” Land Processes Distributed Active Archive Center. https://lpdaac.usgs.gov/products/modis_products_table/mod13q1

NASA. “About Reverb” EOSDIS. http://www.echo.nasa.gov/reverb/about_reverb.htm

Omernik, J., 1987. Ecoregions of the Conterminous United States. *Annals of the Association of American Geographers*, 77, 118-125.

Paerl, H. W., Huisman, J., 2008. Blooms Like It Hot. *Science*, 320. 57.

Sakamoto, T., Yokozawa. M., Toritani, H., Shibayama, M., Ishitsuka, N., Ohno, H., 2005. A Crop Phenology Detection Method Using Time-series MODIS Data. *Remote Sensing of Environment*, 96(3-4), 366-74.

Shao, Y., Lunetta, R., Ediriwickrema, J., Iiames, J., 2010. Mapping cropland and major crop types across the Great Lakes Basin using MODIS-NDVI data. *Photogrammetric Engineering and Remote Sensing*, 75(1), 73-84.

Shao, Y., Lunetta, R., 2010. Sub-pixel mapping of tree canopy, impervious surfaces and cropland in the Laurentian Great Lakes Basin using MODIS time-series data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 4(2), 336-347.

Shao, Y., Lunetta, R., 2012. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. ISPRS Journal of Photogrammetry and Remote Sensing, 70, 78-87.

Swets, D. L., Reed, B. C., Rowland, J. D., & Marko, S. E., 1999. A Weighted Least-Squares Approach to Temporal NDVI Smoothing. 1999 ASPRS Annual Conference, From Image to Information, Portland, Oregon, May 17 – 21, 1999, Proceedings: Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing.

Viovy, N., Arino, O., Belward, A.S., 1992. The Best Index Slope Extraction (BISE): A method for reducing noise in NDVI time-series. *International Journal of Remote Sensing*, 13, 1585-1590.

Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J., Salas, W., & Mooreiii. B., 2006. Mapping Paddy Rice Agriculture in South and Southeast Asia Using Multi-temporal MODIS Images. *Remote Sensing of Environment*, 100(1), 95-113.

Appendix

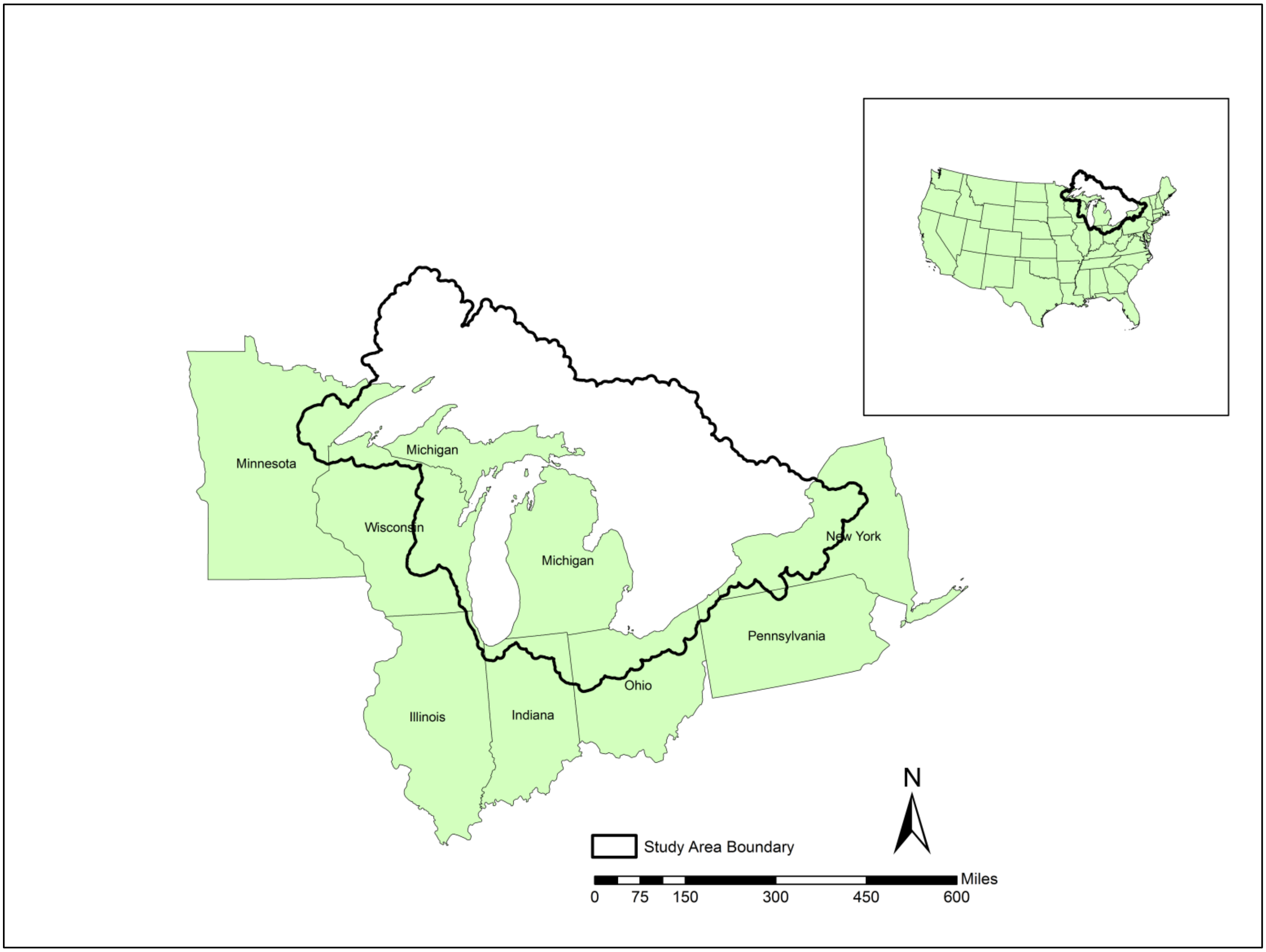

Figure 1.1: Great Lake Basin Study Area

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

Figure 2.1: Euclidean Distance Formula

Tables:

Pixel IDs	**Gaussian**	**Savitzky-Golay**	**Double Logistic**	**Raw**
1 to 2	1.915	1.889	1.920	2.063
1 to 3	0.953	1.046	0.981	1.282
1 to 4	0.950	0.962	0.945	1.254
2 to 3	0.413	0.494	0.372	1.050
2 to 4	1.357	1.359	1.382	1.667
3 to 4	1.299	1.310	1.302	1.733
Mean	**1.148**	**1.177**	**1.150**	**1.508**

Table 3.1: Deciduous Forest Pairwise Euclidean Distance, Selected Samples 2007-2009

Pixel IDs	**Gaussian**	**Savitzky-Golay**	**Double Logistic**	**Raw**
1 to 2	1.113	0.923	1.039	1.309
1 to 3	1.337	1.248	1.282	1.771
1 to 4	1.408	1.162	1.204	1.679
2 to 3	1.976	1.801	1.788	2.219
2 to 4	1.149	1.170	1.172	1.479
3 to 4	1.452	1.329	1.326	1.706
Mean	**1.406**	**1.272**	**1.302**	**1.694**

Table 3.2: Evergreen Forest Pairwise Euclidean Distance, Selected Samples 2007-2009

	Gaussian	**Savitzky-Golay**	**Double-Logistic**	**Raw Data**
Deciduous forest	1.740	1.779	1.782	2.494
Evergreen forest	2.683	2.528	2.473	3.434
Mean	**2.212**	**2.154**	**2.128**	**2.963**

Table 3.3: Euclidean Distance Measurements, 2750 Random Samples 2001-2013

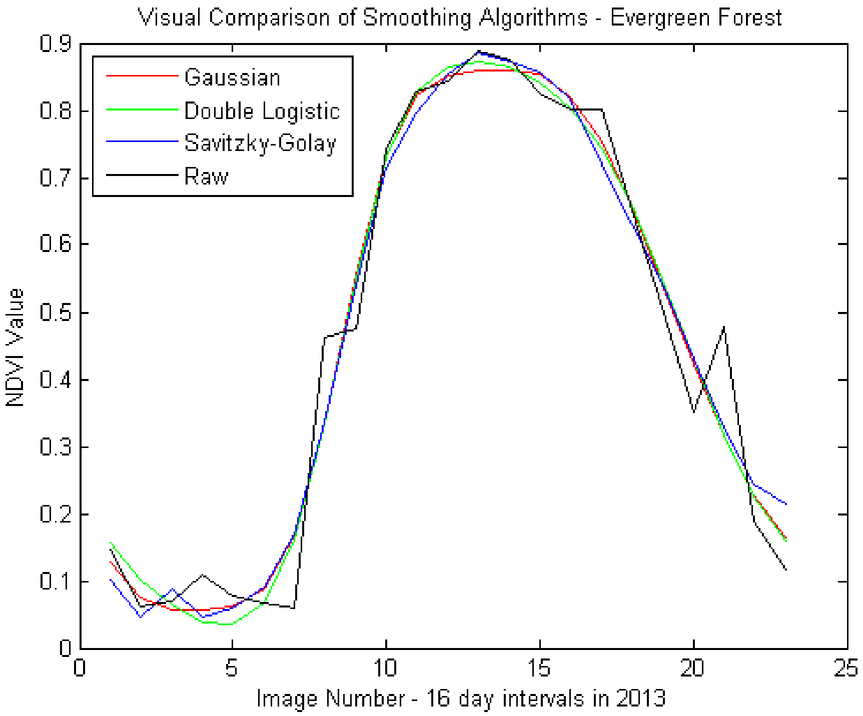

Figure 2.2: Visual Comparison of Smoothing Algorithm Effects, Evergreen Forest

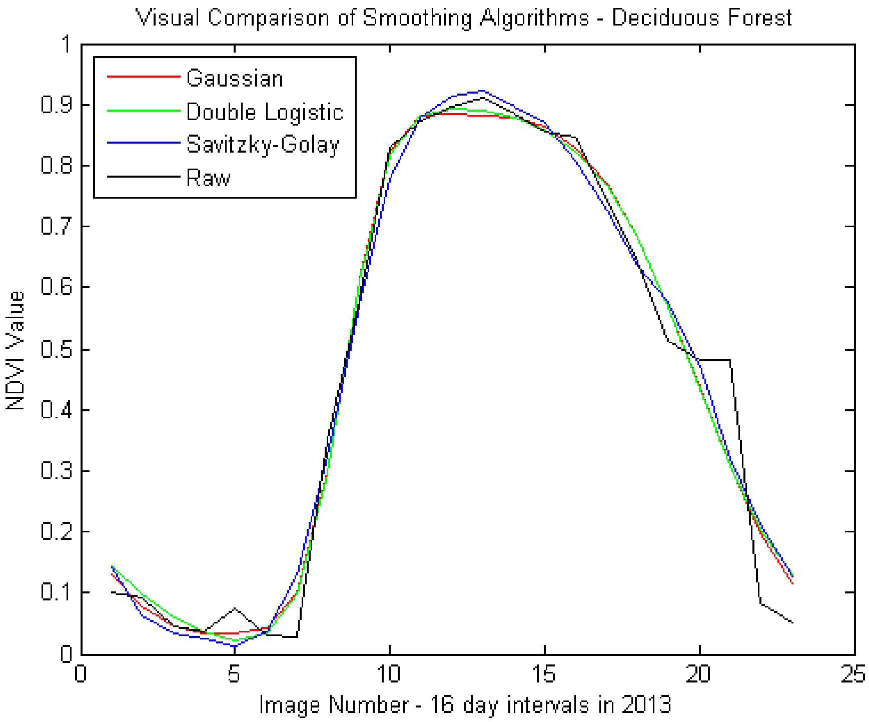

Figure 2.3: Visual Comparison of Smoothing Algorithm Effects, Deciduous Forest

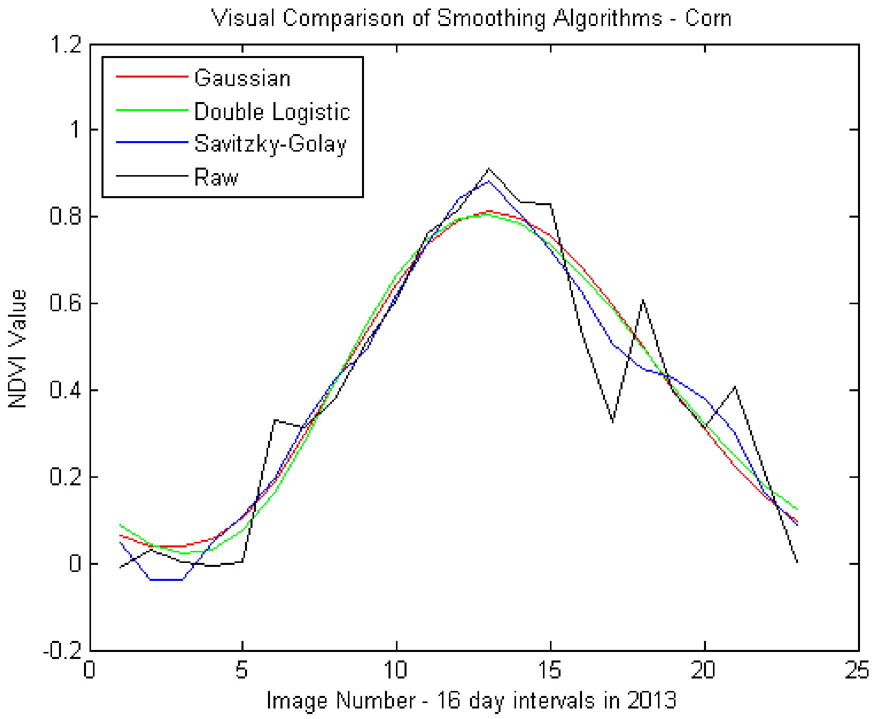

Figure 2.4: Visual Comparison of Smoothing Algorithm Effects, Corn

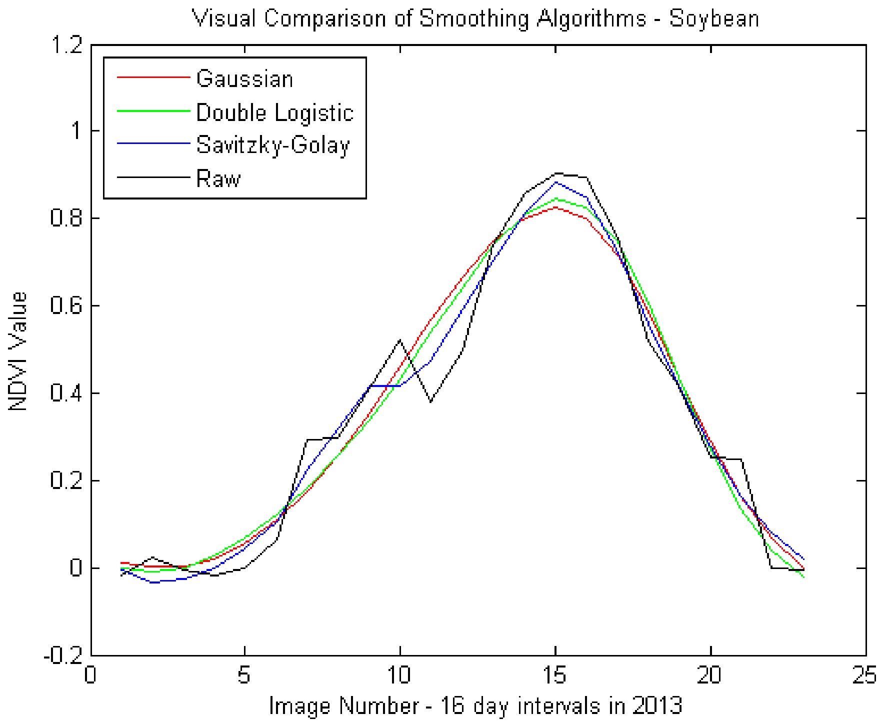

Figure 2.5: Visual Comparison of Smoothing Algorithm Effects, Soybean

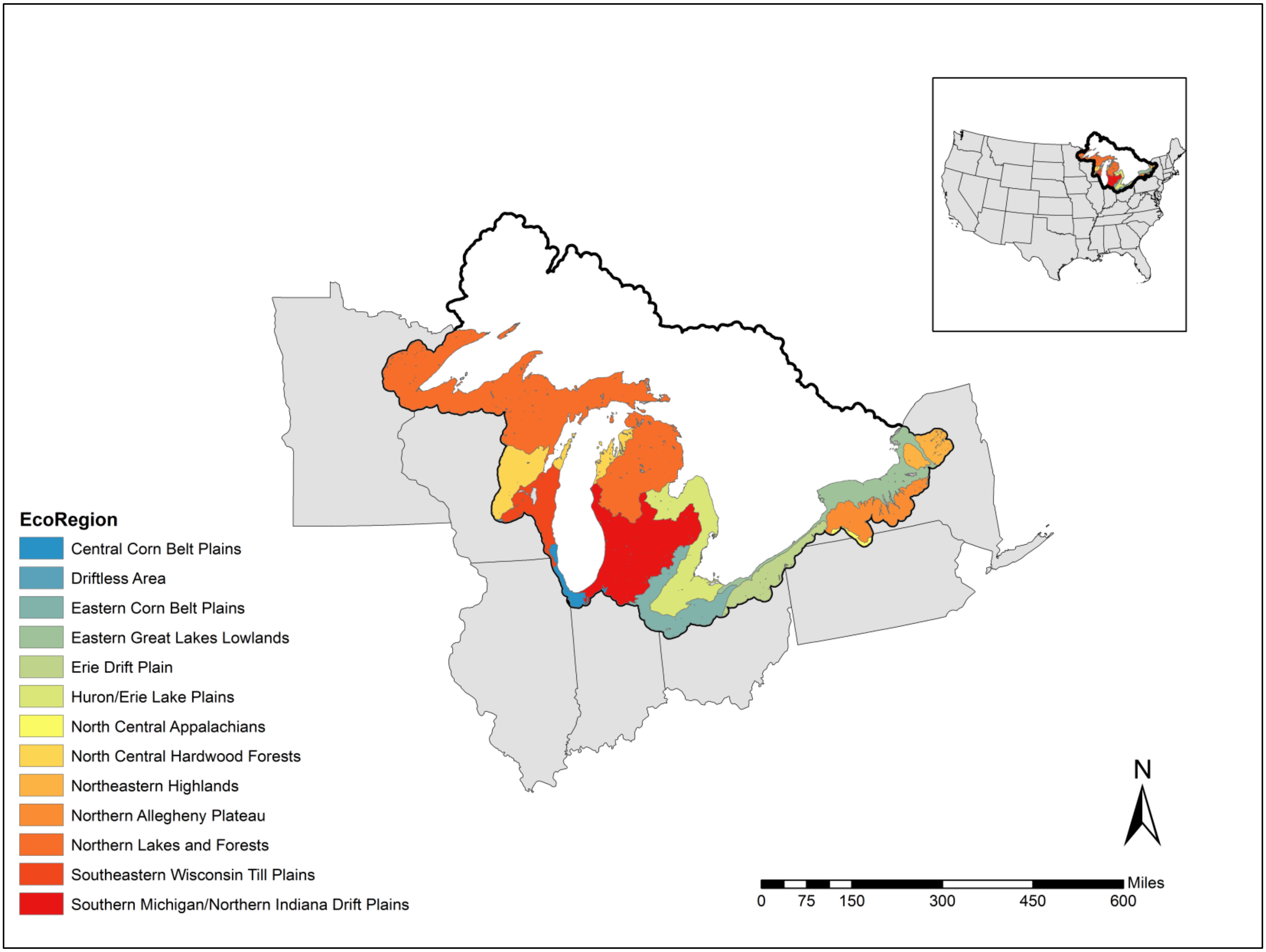
EcoRegion
Central Corn Belt Plains
Driftless Area
Eastern Corn Belt Plains
Eastern Great Lakes Lowlands
Erie Drift Plain
Huron/Erie Lake Plains
North Central Appalachians
North Central Hardwood Forests
Northeastern Highlands
Northern Allegheny Plateau
Northern Lakes and Forests
Southeastern Wisconsin Till Plains
Southern Michigan/Northern Indiana Drift Plains
N
Miles
0
75
150
300
450
600

Figure 2.7: Ecoregion Map

Raw Data	Deciduous Forest	Evergreen Forest	Pasture/Hay	Cultivated Crops
Deciduous Forest				
Evergreen Forest	1.21			
Pasture/Hay	1.30	1.40		
Cultivated Crops	1.37	1.41	0.88	
Wetlands	0.84	0.73	1.35	1.40

SG	Deciduous Forest	Evergreen Forest	Pasture/Hay	Cultivated Crops
Deciduous Forest				
Evergreen Forest	1.22			
Pasture/Hay	1.30	1.38		
Cultivated Crops	1.37	1.40	0.88	
Wetlands	0.91	0.83	1.33	1.38

GA	Deciduous Forest	Evergreen Forest	Pasture/Hay	Cultivated Crops
Deciduous Forest				
Evergreen Forest	1.39			
Pasture/Hay	1.39	1.40		
Cultivated Crops	1.41	1.41	1.35	
Wetlands	1.22	1.05	1.37	1.41

DL	Deciduous Forest	Evergreen Forest	Pasture/Hay	Cultivated Crops
Deciduous Forest				
Evergreen Forest	1.34			
Pasture/Hay	1.39	1.38		
Cultivated Crops	1.41	1.41	1.37	
Wetlands	1.15	0.94	1.36	1.40

	Raw Data	SG	GA	DL
Average	1.19	1.20	1.34	1.32

Table 3.4: Generalized Landcover Class, Jeffries Matusita Scores

JM distance for corn and soybean separability (0.5 threshold)

year	Raw data	SG	GA	DL
2008	0.887	0.854	1.087	0.968
2009	0.900	0.878	1.078	1.057
2010	0.896	0.964	1.227	1.114
2011	0.875	0.845	1.075	1.061
2012	0.807	0.819	1.088	1.179
2013	0.812	0.719	1.063	1.046
2014	0.839	0.799	1.049	0.992
Average	0.859	0.840	1.095	1.060

JM distance for corn and soybean separability (0.9 threshold)

year	Raw	SG	GA	DL
2008	1.110	1.098	1.274	1.159
2009	1.095	1.092	1.253	1.214
2010	1.109	1.186	1.335	1.294
2011	1.108	1.082	1.413	1.195
2012	1.020	1.050	1.283	1.318
2013	1.019	0.937	1.255	1.232
2014	1.013	1.016	1.154	1.144
Average	1.068	1.066	1.281	1.222

Table 3.5: Crop Specific Landcover Class, Jeffries Matusita Scores, Multiple Thresholds

Asymmetric Gaussian dataset: JM distance for corn and soybean separability by ecoregion (0.5 threshold)

	2008	2009	2010	2011	2012	2013	Average
All Regions	1.149	1.649	1.225	1.273	1.323	1.205	1.304
Northern Lakes and Forests	1.447	1.973	1.507	1.523	1.620	1.282	1.559
North Central Hardwood Forests	1.791	1.850	1.662	1.200	1.539	1.295	1.556
Southeastern Wisconsin Till Plains	0.649	1.573	1.538	0.902	1.250	1.543	1.243
Central Corn Belt Plains	1.500	1.608	1.759	1.598	1.396	1.503	1.561
Eastern Corn Belt Plains	1.099	1.392	1.312	1.145	0.941	0.708	1.099
Southern Michigan/Northern Indiana Drift Plains	0.836	1.411	0.817	0.998	1.089	0.959	1.018
Huron/Erie Lake Plains	1.156	1.290	1.409	1.197	1.304	0.746	1.184
Northern Allegheny Plateau	1.334	1.969	1.776	1.431	1.751	1.554	1.636
Erie Drift Plain	1.219	1.635	1.534	1.831	1.431	1.568	1.536
Eastern Great Lakes Lowlands	1.300	1.654	1.147	1.224	1.591	1.472	1.398

Asymmetric Gaussian dataset: JM distance for corn and soybean separability by ecoregion (0.9 threshold)

	2008	2009	2010	2011	2012	2013	Average
All Regions	1.585	1.659	1.562	1.611	1.667	1.480	1.594
Northern Lakes and Forests	1.976	1.994	1.945	1.824	1.807	1.892	1.906
North Central Hardwood Forests	1.982	1.928	1.963	1.712	1.938	1.866	1.898
Southeastern Wisconsin Till Plains	1.666	1.762	1.699	1.463	1.695	1.795	1.680
Central Corn Belt Plains	1.848	1.907	1.957	1.946	1.856	1.805	1.886
Eastern Corn Belt Plains	1.799	1.825	1.733	1.698	1.647	1.136	1.640
Southern Michigan/Northern Indiana Drift Plains	1.566	1.476	1.226	1.356	1.480	1.398	1.417
Huron/Erie Lake Plains	1.758	1.796	1.822	1.640	1.749	1.432	1.699
Northern Allegheny Plateau	1.921	1.985	2.000	1.957	1.999	1.970	1.972
Erie Drift Plain	1.730	1.953	1.882	1.924	1.913	1.898	1.883
Eastern Great Lakes Lowlands	1.589	1.913	1.541	1.799	1.881	1.812	1.756

Table 3.6: Ecoregion Jeffries Matusita Scores, Asymmetric Gaussian Corn and Soybean data

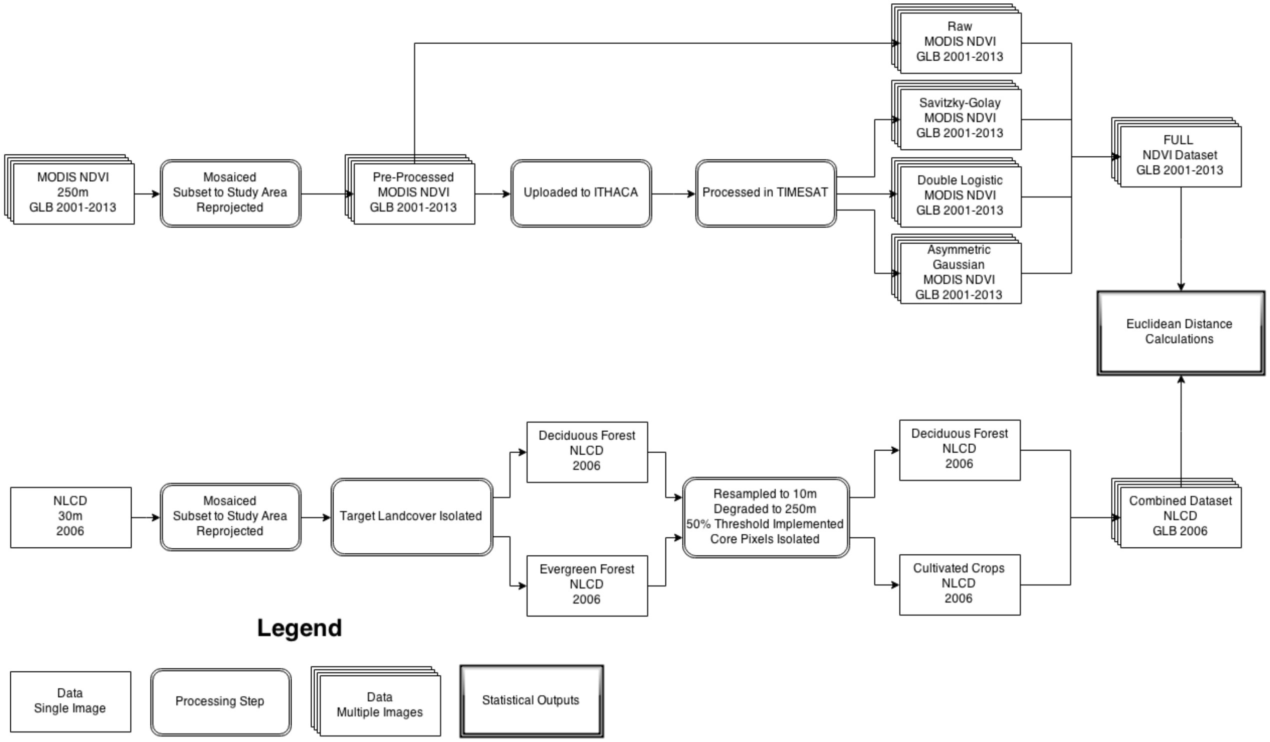

Figure 2.6: Euclidean Distance Processes

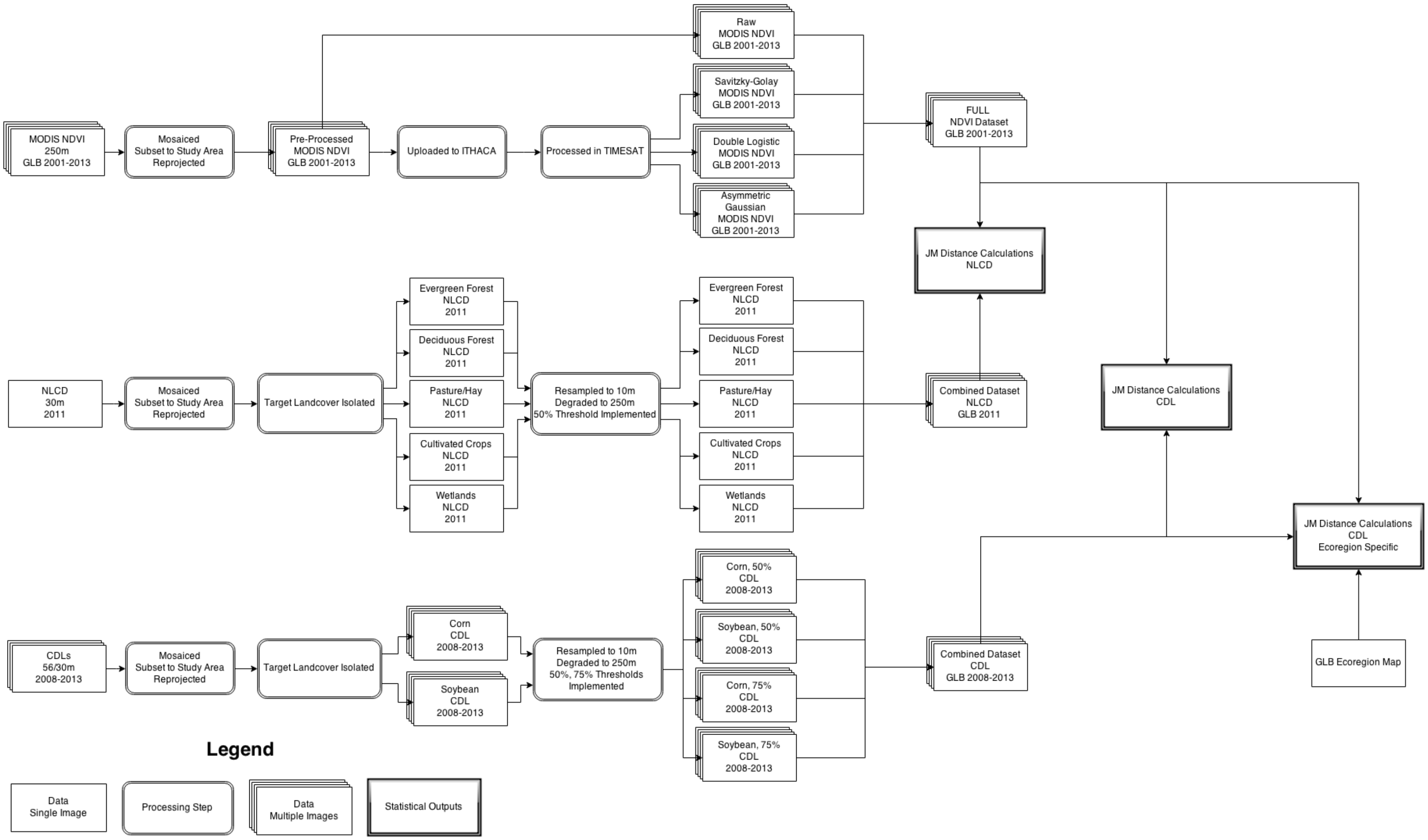

Figure 2.8: JM Distance Processes