AUTOMATED SUNSPOT DETECTION AND CLASSIFICATION USING SOHO/MDI IMAGERY

THESIS

Samantha R. Howard, 1st Lieutenant, USAF

AFIT-ENP-MS-15-M-078

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENP-MS-15-M-078

AUTOMATED SUNSPOT DETECTION AND CLASSIFICATION USING SOHO/MDI IMAGERY

THESIS

Presented to the Faculty

Department of Engineering Physics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Applied Physics

Samantha R. Howard, B.S.

1st Lieutenant, USAF

March 2015

AFIT-ENP-MS-15-M-078

AUTOMATED SUNSPOT DETECTION AND CLASSIFICATION USING SOHO/MDI IMAGERY

Samantha R. Howard, B.S.
1st Lieutenant, USAF

Committee Membership:

Dr. William F. Bailey, PhD

Chair

Lt Col Kevin M. Bartlett, PhD

Member

Robert D. Loper, Jr., PhD

Member

AFIT-ENP-MS-15-M-078

Abstract

Sunspots and their group classifications are one of the most generally accepted indicators of solar activity, but these data are currently generated by individual observers and are often based on hand drawings. This research modifies and expands previous work by Spahr in 2014 to automatically identify and classify sunspot groups in satellite images. The algorithm used produces consistent and accurate classifications. Data from the Solar and Heliospheric Observatory (SOHO) are analyzed to produce a database of sunspot information. In order to apply the algorithms, SOHO images are processed to correct for sensor sensitivities as well as changes in exposure and window degradation that vary with time. The resulting database improves on the data currently available from the National Oceanic and Atmospheric Administration's (NOAA) Solar Region Summaries (SRS) in that it does not change with the biases of individual solar observers. Results of the algorithm on SOHO/MDI data correlate well with NOAA's reported data for region properties summed over an observation. Linear regression models applied to the two data sets have R^2 values greater than 0.75, but the slopes of the models are less than 1. Properties of the regions are proportional, but not on a 1:1 scale In particular, the results of analyzing SOHO data report less than 25% of the spots reported by NOAA. Comparing individual regions, the location and area are statistically similar, but the length, number of spots, are not. By considering a test case comparison with an SDO observation, resolution is likely the main factor in detection discrepancies.

Acknowledgements

I extend thanks to my advisor Dr Bailey and committee members, Lt Col Bartlett, and Dr Loper, as well as my former advisor Dr Acebal. I also have sincere appreciation for the comments and reviews provided by Capt Jeff Graham and 1st Lt Michael Seery, which helped me immensely in my revisions.

Finally, I'm grateful to my cats for keeping me company through the long days and nights of work spent on this thesis.

Samantha R. Howard

Contents

Page

List of Figures

AUTOMATED SUNSPOT DETECTION AND CLASSIFICATION USING SOHO/MDI IMAGERY

I. Introduction

1.1 Impact of Solar Activity

The active sun goes through an eleven year cycle of magnetic activity, with sunspots as the most easily observed indicator of increased activity. Sunspots' relatively cool temperature (compared to their surroundings) causes a decrease in the emitted radiation that is easily detected in the visible wavelengths. Increased numbers, size, and complexity of sunspots and sunspot groups are positively correlated with an increase in solar activity, which in turn is indicative of a higher probability of solar flares and coronal mass ejections (CMEs). These intense solar emissions can negatively impact Air Force missions, including disruption of satellite communication, damage to satellite hardware, and power grid failures on Earth.

Classification of sunspots allows for consistent documentation of the state of solar activity. However, though current classification methods are based on the quantitatively defined McIntosh system, a level of subjectiveness is unavoidable due to their dependence on fallible human observers. Such errors are detrimental to the long term study of solar activity and may negatively impact prediction of flares and CMEs.

1.2 Current Operational Technique

Current sunspot observation is entirely ground-based and most are performed by hand. This process involves the projection of a white light image onto a sheet of

paper, which the observer traces with a pencil. Imprecision is unavoidable, including the thickness of the pencil lead in comparison to the fine structure of the images and the observer's personal drawing ability.

The Air Force Solar Observing Optical Network (SOON) follows a set of standardized procedures prescribed by the Air Force Weather Agency [1]. An abbreviated, but sufficient, sequential description of the procedures follows:

1. Observations are timed to coincide as closely as possible with 1500 UTC, taking into consideration visual observations of the turbulence in the atmostphere.

2. Observing worksheet is manually aligned for the current solar P-angle (orientiation of the Sun's North with respect to the geocentric North).

3. Focus and size of the solar disk is adjusted to fit an 18 cm diameter circle on the worksheet.

4. Analyst hand draws the sunspots in pencil.

5. Groups are identified and labeled. This includes comparing current observations with past observations.

6. Locations, areas, and lengths are determined using clear plastic overlays and mathematical approximations on a handheld calculator.

7. Magnetic maps are compared with white light image.

8. Groups are classified according to the McIntosh system (for a more detailed explanation of the McIntosh system, see Section 2.3)

Each step of this process is prejudiced by the individual analyst, depending on such factors as their drawing ability, precision, and judgement of measurements. Even at

the same location with analysts trained by the same person, there will be disagreements on classifications. This process is repeated at multiple locations worldwide, inviting more variations and potential error into the results. Similar processes are typical of current sunspot reportings, including that of the National Oceanic and Atmospheric Administration's (NOAA) Space Weather Prediction Center (SWPC).

In addition to the errors possible from human observers, the nature of ground-based observing adds further difficulties. The most obvious problem is that the Sun is obscured by the Earth itself during the night, which severely restricts observing opportunities. Similarly, issues may arise due to weather, particularly clouds [1], and aircraft occulting the solar disk.

Aside from these direct obstructions, the most influential factor in ground-based observations is the Earth's atmosphere. The signal from the Sun is attenuated due to scattering and absorption. Variations in both density and temperature of the atmosphere cause the refractive index to vary [1], thus causing distortion in the image as the light rays forming the image take different optical paths. This can cause problems in determining location, area, or existence of a sunspot.

1.3 Recent Research

1.3.1 Automated Sunspot Detection and Classification.

Much research has been done in the area of automated sunspot classification, but such methods have been slow to be accepted and integrated into official solar observations and predictions. The most notable exception, solarmonitor.org, displayed its own automatically detected magnetic active regions for several years using data from Solar and Heliospheric Observatory's (SOHO) Michelson Doppler Imager (MDI). Many techniques have been explored in the search for a reliable algorithm that is not computationally expensive.

A popular route of exploration has been the use of artificial intelligence, particularly neural networks that attempt to mimic the human brain [3]. The objective is to train the network to analyze sunspot groups like an expert [2, 4]. As observers often have their own quirks and nuances in sunspot analysis, these algorithms are sensitive to whom they are based on and what data sets are used in the training process. These techniques are not within the scope of the this project since the goal is a consistent and reliable output standardized to the McIntosh system as specified (not as practiced).

Techniques that have been used on sunspot identification and classification include thresholding, edge-detection, watershed filling, and region growing [5], [6], which have also been applied to active regions [7], as well as erosion and dilation [8]. Such methods have a defined and consistent output for a particular input [9], making them excellent choices for an objective algorithm. Spahr, on whose work this project is based, utilized thresholding [10].

1.3.2 Basis of This Research.

In 2013 and 2014, Spahr created a MATLAB code to analyze and classify data from the Solar Dynamics Observatory's (SDO) Helioseismic and Magnetic Imager (HMI), which demonstrated good correlation with data from the Air Force's Solar Observing Optical Network (SOON) at Holloman AFB and the NOAA's SWPC [10]. HMI has only been in operation since 2010 which limits the dates for which data is available. Therefore modification of the code to accept images from the older SOHO/MDI which operated from 1996 to 2011 provides more data for analysis and comparison. Additionally, the current implementation of grouping and classification of sunspots is computationally expensive [10], therefore improving efficiency and decreasing execution time for the code is an objective. Finally, alternative image processing techniques

will be explored to improve classification accuracy.

II. Background

2.1 Solar Cycle

The Sun experiences an approximately 22-year magnetic cycle, where the magnetic poles reverse every 11 years — causing an 11-year cycle in the variation in sunspot occurrences. The most widely accepted explanation of this is the Babcock Model [11]. This model takes a solar minimum as the start of the cycle, when the Sun's magnetic field takes the form of a simple dipole with field lines directly North-South, extending from just below the photosphere at the equator through the corona near the poles and connecting at several radii above the equator. Within the solar plasma, magnetic flux is entrained within the plasma it traverses [11]. Differential rotation, where the plasma at the equator of the Sun rotates faster than at the poles, therefore causes the field lines to be dragged along with the plasma. The divergence of a magnetic field is zero, so the lines stretch rather than break. As time progresses, the lines become nearly horizontal with the equator and the associated magnetic field becomes strong and complex. This is the point of solar maximum. The complexities produce magnetic loops that rise through the Sun and form active regions that encompass multiple layers. Regions generally begin as bipolar and eventually decay into separate polarities. Due to the attraction of unlike poles, these region remants individually migrate towards the opposite pole, over time neutralizing the magnetic field back the simple dipole — now with the reverse alignment, South-North. The Sun has returned to solar minimum. [12].

2.2 Sunspots

2.2.1 Evolution of Sunspot Groups.

Individual flux tubes generated by magnetic loops have a lower pressure than the surrounding plasma, with a difference relating to the square of the magnetic field strength. Due to the lower internal pressure, and associated lower plasma density, the flux rises through the layers of the Sun and eventually emerges at the photosphere. In addition to lower density, the lower pressure also leads to the tubes having a lower temperature than the environment. Assuming the Sun radiates as a blackbody, with the intensity as a function of temperature given by Equation 1, the flux tube will radiate with a lower intensity than its surroundings [11].

$$I_\lambda = \frac{2\pi hc^2}{\lambda^5} \frac{1}{exp(^{hc}/_{k\lambda T}) - 1} \tag{1}$$

This becomes particularly evident when a tube reaches the photosphere, where the majority of the Sun's visible light is produced. The flux tubes appear as darker areas on the solar disk. The temperature gradients between the tubes and the surrounding plasma generate a convective downdraft that exerts a force both horizontally and vertically. The vertical component is counteracted by the buoyancy (pressure gradient), leaving the horizontal force that can cause the tubes to cluster. If enough tubes coalesce together, the convection sustains the arrangement which appears as a sunspot [13].

2.2.2 Structure.

The central, darkest area of a sunspot is called the umbra and the outer ring is called the penumbra [11], as seen in Figure 1 . Within the umbra the magnetic field is nearly vertical, while the penumbra has a partially horizontal field. The spot forms

a depression in the photospheric plasma due to its lower density; the flat bottom is the umbra and slanted sides are the penumbra. This Wilson depression is evident when observing a spot near the limb of the Sun, because the portion of the penumbra farthest from the limb appears to vanish while the portion closest to the limb grows larger.

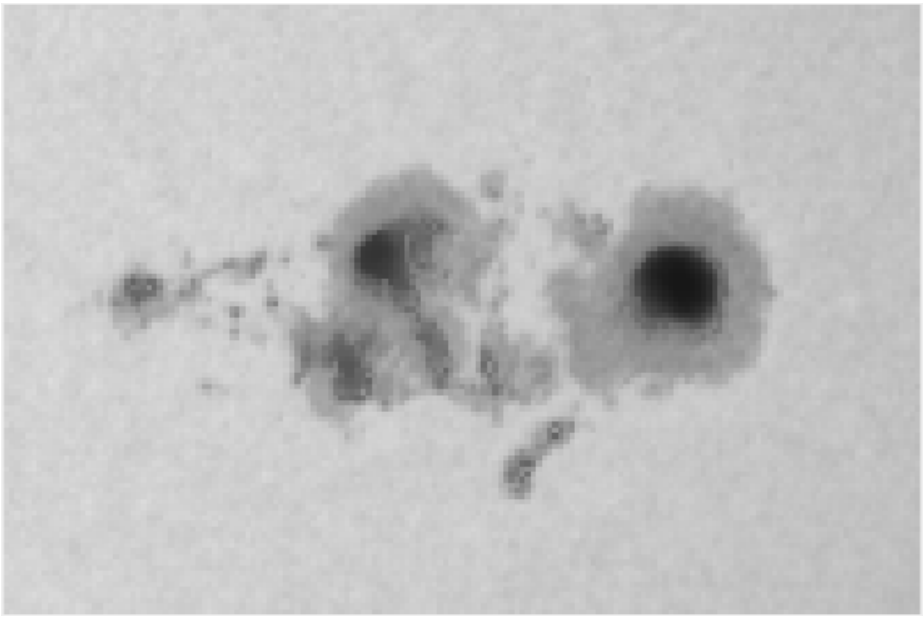

Figure 1. Sunspot in a SOHO/MDI continuum image, with the umbra and penumbra clearly evident.

Sunspots tend to grow larger and develop into groups as time progresses, for much the same reason as they form originally. Due to the net zero divergence in a magnetic field, sunspots often appear near others of opposing polarity. It is rare to see a large spot or group of spots with a single polarity. When groups decay, the poles separate and drift towards the opposing solar magnetic pole as discussed in Section 2.1.

2.2.3 Indicator of Solar Activity.

Given the ease of their observation and dependence on the Sun's magnetic field, sunspots are used as an indicator of overall solar activity. The sunspot number is

recorded by many organizations and is the primary means of tracking the solar cycle. Individual sunspot groups with particular features have been linked to solar flares [2]. This has been a focus of previous work including purely McIntosh classifications (discussed in Section 2.3) [2], as well as Poisson statistics [14] and sequential information about the evolution of the region [15].

2.3 McIntosh Classification of Sunspots

In 1966, an updated sunspot classification system was introduced that incorporated more features than previous systems with particular interest in improving correlation with solar flares [2]. This McIntosh system is still the standard today, used at both SWPC and Air Force installations, among many other international observatories. It consists of three components, explained in the following subsections, written in the form *Zpc*, where Z is the modified Zurich class, p is the penumbra on the largest spot, and c is the sunspot distribution within the group. The McIntosh system was originally designed to only require an intensity image for classification and is therefore largely based on visible features.

2.3.1 Modified Zurich.

Based on the Zurich system, this component describes the polarity of the group, the presence and location of penumbra, and the length of the group in heliographic degrees (Table 1). Zurich classifications were originally a self-contained classification system, describing the typical evolution of a sunspot group. The later in the alphabet, the older and more complex the group. Groups with only one polarity can only be assigned an 'A' or an 'H', indicating they are either at the beginning of their development and have not yet developed spots on both polarities or at the end, when the group is deteriorating into separate poles. Penumbra develop and increase in

size and occurrence across the group as it becomes more complex. The group length generally increases as the group evolves. The original definitions were designed to only require white light observations [2], but current techniques make use of readily available magnetic information (magnetic maps, magnetographs, or magnetograms) for better discernment [1].

Table 1. Modified Zurich classifications of the McIntosh system as currently used by Air Force Weather. Adapted from [1]

Class	Polarity	Penumbra	Size
A	unipolar	none	
B	bipolar	none	
C	bipolar	on spots of one polarity	
D	bipolar	on spots of both polarities	$\leq 10^{\circ}$
E	bipolar	on spots of both polarities	$> 10^{\circ}$ and $\leq 15^{\circ}$
F	bipolar	on spots of both polarities	$\geq 15^{\circ}$
H	unipolar	yes	

2.3.2 Penumbra.

The penumbra component describes the presence, maturity, symmetry, and diameter of the penumbra on the largest spot in the sunspot group (Table 2) [2]. The maturity of penumbra is how completely the penumbra surrounds the umbra, the more mature being the more complete. Symmetry describes how close the penumbra's shape is to a circle. The size is measured from North to South, in order to minimize foreshortening of the measurement.

Table 2. Penumbra classifications defined by the McIntosh system. Adapted from [2]

Class	Presence & Maturity	Diameter	Symmetry
x	none		
r	rudimentary, incomplete		
s	mature	small ($\leq 2.5°$)	symmetric
a	mature	small ($\leq 2.5°$)	asymmetric
h	mature	large ($> 2.5°$)	symmetric
k	mature	large ($> 2.5°$)	asymmetric

2.3.3 Sunspot Distribution.

The sunspot distribution component (Table 3) describes the "spottedness of the group's interior" [2]. This describes the number and strength of spots between the leader and follower. This is the most subjective component, with the differentiation between a "few" spots and "numerous" spots not defined quantitatively. The observation of at least one mature penumbra determines the classification definitively.

Table 3. Spot distribution classifications defined by the McIntosh system. Adapted from [2]

Class	Description
x	undefined for unipolar groups
o	open; few, if any, spots between leader and follower
i	intermediate; numerous spots between leader and follower, but no mature penumbra
c	compact; many strong spots, at least one with mature penumbra

2.4 Solar and Heliospheric Observatory: Michelson Doppler Imager

Launched in 1995 to an orbit about the first Lagrange point L_1, the Solar and Heliospheric Observatory (SOHO) carried the Michelson Doppler Imager (MDI) amongst its payload. As the name suggests, MDI has two Michelson interferometers that produce a small (94mÅ) bandpass tunable about the Ni I line of 6768Å. A single image centered on a particular wavelength is called a filtergram, and combinations of five individual images spaced at 75mÅ intervals across MDI's range are used to produce the typical observables [16].

2.4.1 White Light Approximation.

Sunspots are typically defined by their visibility in white light images of the Sun. However, SOHO's MDI instrument is not capable of taking a true white light observation. Instead, the continuum intensity, I_c, is approximated by using the particular filtergrams F_0, F_1, F_2, F_3, F_4 as in Equation 2.

$$I_c = 2F_0 + \sqrt{2((F_1 - F_3)^2 + (F_2 - F_4)^2)}/2 + \frac{F_1 + F_2 + F_3 + F_4}{4} \tag{2}$$

Intensitygrams provide the intensity in arbitrary units. This not an unusual occurrence in astronomical observing, as calibration is tedious and often unnecessary for analysis.

2.4.2 Magnetograms.

The Zeeman effect allows for the determination of a magnetic field strength by observing the difference in left- and right-circularly polarized light produced. This can be accomplished by taking the difference of Dopplergrams in each polarization [16]. Individual Dopplergrams are calculated via Equation 3. Calibrated magnetograms

provide the magnetic field in Gauss.

$$\alpha = \begin{cases} (F_1 + F_2 - F_3 - F_4)/(F_1 - F_3) & \text{if numerator} > 0 \\ (F_1 + F_2 - F_3 - F_4)/(F_4 - F_2) & \text{if numerator} \leq 0 \end{cases} \tag{3}$$

2.5 Imaging the Sun

Solar imaging can be quite complex. For ground-based observations, including those used operationally by the Air Force (Section 1.2), difficulties are numerous. Satellite-based observation avoids two of the most troubling issues, as it is outside the Earth's atmosphere and no longer subject to the diurnal occulting of the Sun due to its placement at L_1. Problems are still present, including the limb darkening that occurs for any observations in visible light and the electronics involved in capturing data. Charge-coupled devices (CCDs) require correction for sensitivity differences between pixels, thermal noise, amplifier noise, and readout noise, among other factors [17].

2.5.1 Limb Darkening.

The photosphere is the source of the majority of the visible light radiation from the Sun, as that layer's Planckian radiation (given by Equation 1) peaks in the visible due to its temperature of approximately 6000K. From the Earth's perspective the radiation is not uniformly distributed across the solar disk, but rather there is a distinct drop in intensity at the edges.

Since the Sun is a gaseous object, Beer's Law (given in Equation 4) for the transmission of light can be applied. If a constant source function S_λ is assumed, then the solution does not result in a drop in intensity at the edges [11]. But if a linear source function $S_\lambda = S_\lambda(0) + b\cos(\theta)$ is assumed instead, then the result for the intensity at Earth (Equation 5) equates the cosine of the angle θ off of line-of-sight with the

optical depth. For a large θ, as would occur at the edge of the Sun, $cos(\theta) \to 0$. Therefore the optical depth, τ, goes to zero and the intensity also goes to zero — precisely what is observed.

$$\cos(\theta)\frac{dI_\lambda}{d\tau_\lambda} = I_\lambda - S_\lambda \tag{4}$$

$$I_\lambda(0, \cos(\theta)) = S_\lambda(\tau_\lambda) \tag{5}$$

To ensure the correct conclusions are made in observations, limb darkening must be taken into account.

2.5.2 Image Processing Correction.

CCDs determine the intensity of a light signal based on the current generated by the photoelectric effect. These devices are therefore sensitive to stray radiation, such as cosmic rays, which can produce signals not associated with the object of interest. Additionally, the signal is highly dependent on the sensitivity of each pixel in the array, any thermal noise, and imperfect grounding present in the system. These are corrected for through a series of images that quantify each type of noise, which can then be removed from the final image.

The highest level of data processing available on SOHO/MDI images, which is used in this project, includes corrections for voltage bias, pixel sensitivity, scaling due to focus changes, and strong cosmic rays [18]. Some of the corrections are known to be inaccurate, particularly the flat-fielding, which is the correction for pixel sensitivity. Additionally, the instrument experienced window degradation that was only partially accounted for with periodic increases in exposure time [19]. It is necessary to rectify these issues before using the data.

III. Methodology

3.1 Image Acquisition

The data from SOHO/MDI are available to the public through interfaces on the the NASA website or Stanford University's Joint Science Operations Center (JSOC). The NASA interfaces have an approximately 15 minute delay for system queuing and processing, while JSOC provides access as soon as the data can be retrieved from the database, so JSOC was utilized for data download.

Data required for the sunspot algorithm includes both a continuum intensity and magnetogram file. To minimize variation in solar activity within the observations, only records for which the continuum intensity and magnetogram date and time match to within the minute were used. This was implemented in the download stage to minimize the required disc space. Sets of observations were selected by obtaining a list of available records at 60 minute intervals over the lifetime of SOHO for both the continuum intensity and magnetograms and comparing the list contents.

Unlike the SDO data used in previous observations, which are available in JPEG format, SOHO data are only available in Flexible Image Transport System (FITS). Bundled archives (such as .tar) were not available at the time of access, so MATLAB was used to automate the downloading process. A detailed account of series selection, specific keywords, and MATLAB download code can be found in the Appendices.

3.2 Data Processing

The data must be processed to a level at which the algorithm can be applied. The original MATLAB code was written specifically for SDO data that was ingested as JPEGs already corrected for instrument variations, so the SOHO data must be additionally processed from the FITS files to a compatible format. Additionally, the

SOHO/MDI data is known to have inaccurate flat-fielding and correction for window degradation [19], so these must be corrected before applying the algorithm.

3.2.1 Data Type Conversion.

In order for the algorithm to be applied to the SOHO/MDI data, the data must be in the format of an unsigned 8-bit integer (uint8). This is because the algorithm was originally designed to run on SDO JPEG images, which are read by the MATLAB `imread` function as uint8 matrices. Many of the underlying functions in the algorithm require uint8 format as well, as they were written specifically for use with images. The SOHO/MDI data is originally stored as a signed 16-bit integer (int16), therefore requiring a conversion.

The range of the initial distribution of the SOHO/MDI data (example histogram in Figure 2), from -32768 to about 15000 (maximum values vary), is too large to be directly converted to a uint8 with a range of 0 to 255. A scaling must be applied before the values can be converted. First, it is noted that the value -32768 is an isolated peak well below the range of the rest of the data. This is the lowest value possible for int16 and the only negative value; it is assigned to all (and only) pixels off the solar disk. Therefore the only information it carries is that those pixels reside off the solar disk. There are no pixels with a zero value, so this large negative value can therefore be reassigned to zero with no loss of information, while significantly reducing the range of values to include in the scaling and conversion.

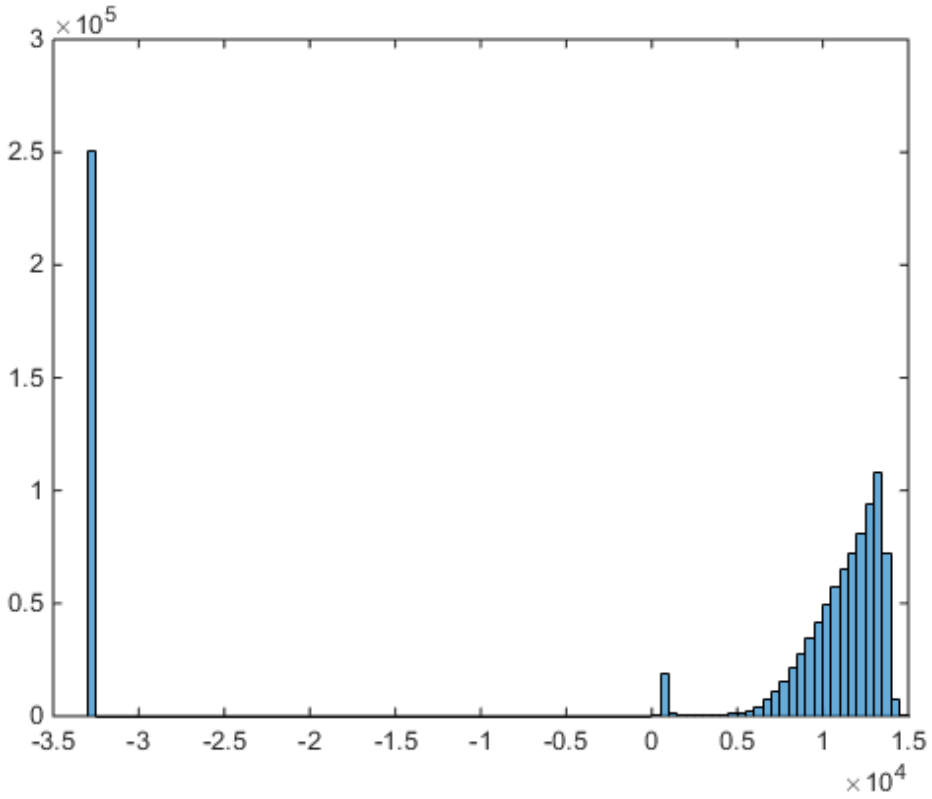

Figure 2. Histogram of the pixel values for the raw (16-bit integer) continuum intensity on January 13, 2003.

The maximum pixel value varies from approximately 13000 to 15000 depending on the individual image, so 15000 is arbitrarily selected as the maximum value for scaling to ensure consistency. The matrix is converted to double precision, values greater than 15000 are reset to 15000 (minimal data is lost by this process and high intensity areas are preserved), the matrix is scaled by $(\frac{255}{15000})$, and then finally converted to uint8 (example histogram in Figure 3).

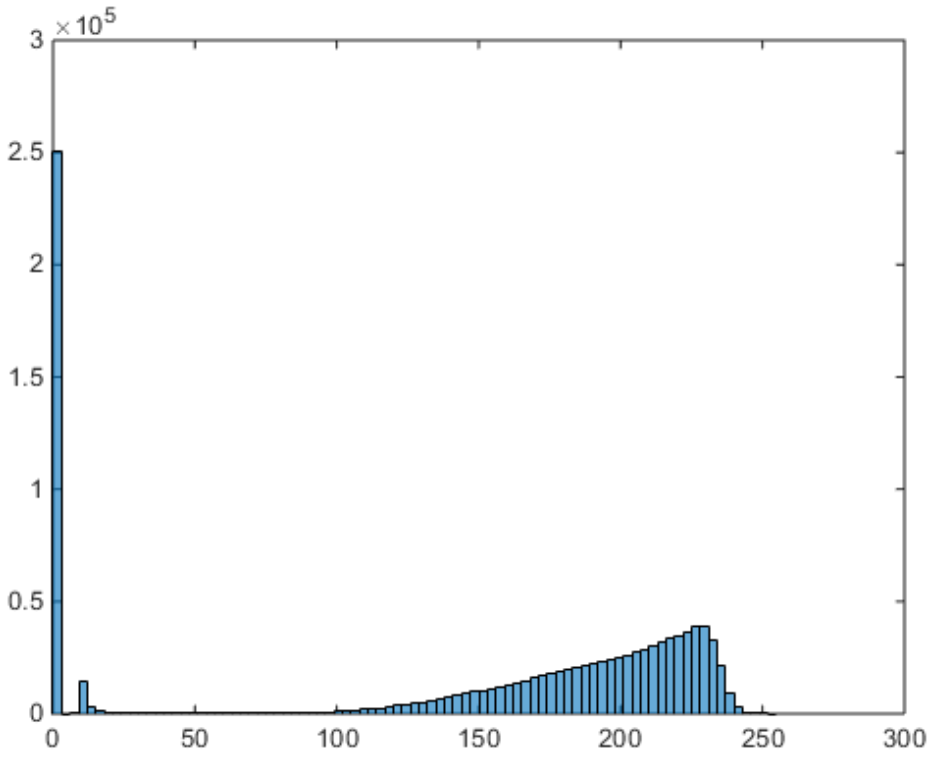

Figure 3. Histogram of the pixel values for the scaled and converted (8-bit integer) continuum intensity on January 13, 2003.

3.3 Image Processing

After the data is in a format compatible with the algorithm, further corrections can be applied. The SOHO/MDI data are known to have erroneous corrections for flat-fielding and window degradation. In order to apply a correction across multiple images, the solar disk needs to be standardized to a certain radius and center location.

3.3.1 Center and Radius Detection.

The center and radius of the solar disk are used to calculate not only the location of sunspots via spherical trigonometry, but also for resizing and centering images in the application of a correction matrix. The built-in function `imfindcircles` is used to identify the radius and center of circles with radii near the input value via

circular Hough Transform. The transformation is provided a reasonable estimate of the radius, whether via calculations based on the location of the SDO satellite or by reading information from the SOHO FITS header, and applied to a simplified continuum intensity image. However, the `imfindcircles` requires a simplified image of the solar disk, otherwise multiple circles will be detected.

In the previous method, which was developed for SDO images that were properly corrected, the simplification was accomplished by applying a Canny edge detection algorithm and flood filling the result. However, due to the erroneous flat-fielding and window degradation correction, this method has poor results when applied to SOHO images. This can be seen in the the many lines that appear that are not associated with the limb of the solar disk (Figure 4). The center and radius are prerequisites to complete an image correction, so a different technique had to be utilized. An effective method was found to be inversion and conversion to binary at a threshold of 0.9. This removes the darkest 10% of the image, which is associated with the area outside the solar disk, producing a circle the same size as the solar disk. The difference in the results of these methods are on the order of 1-2 pixels, sufficient to cause errors in the subsequent calculations. Figure 5 shows a segment of the solar limb, where the green curve is the edge of the circle identified by the inversion method while the red curve is that of the Canny method. The Canny method consistently underestimates the radius of the solar disk on SOHO images.

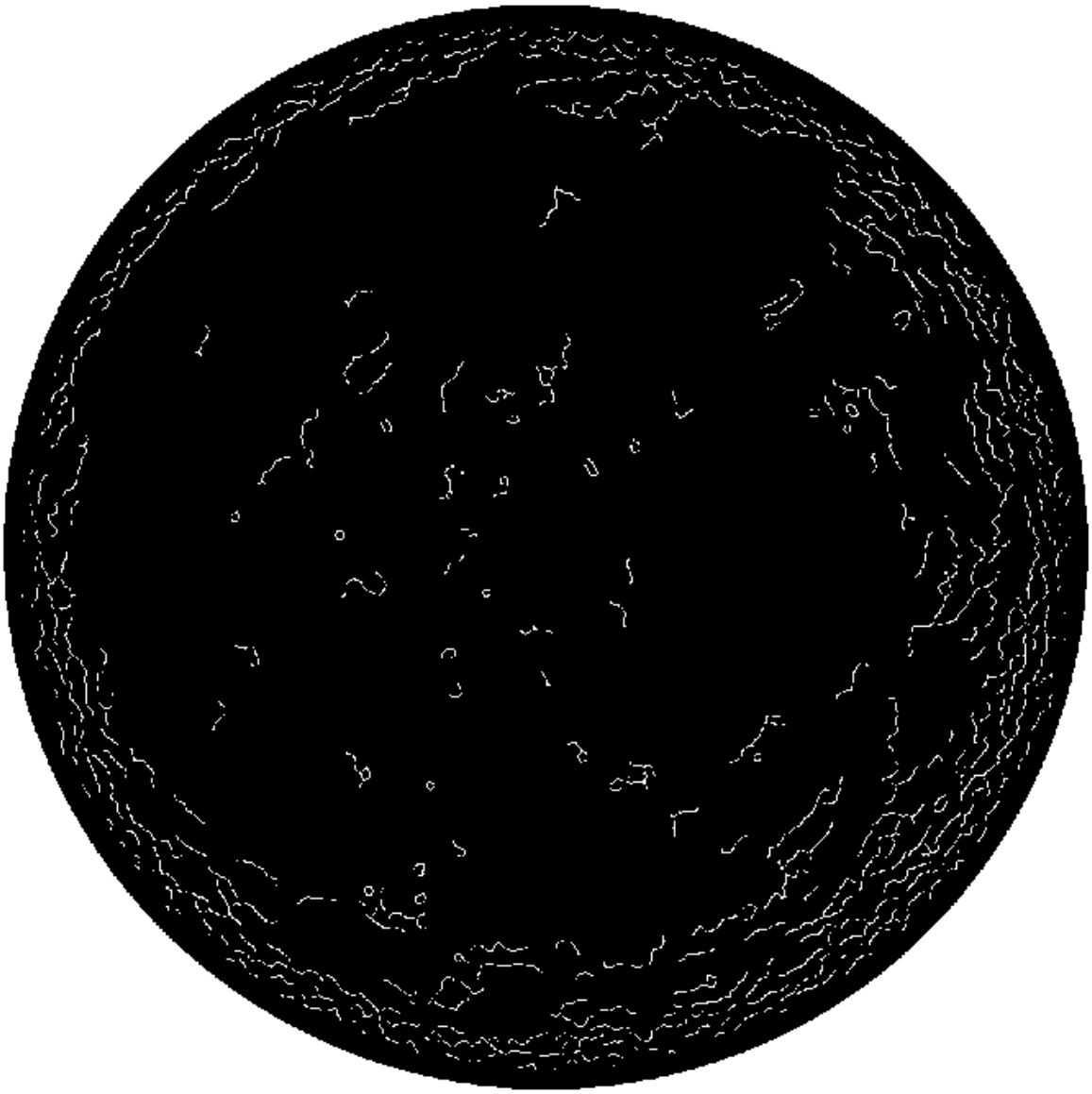

Figure 4. Flood-filled result of Canny edge detection. Note the many lines present that are not associated with the limb of the solar disk.

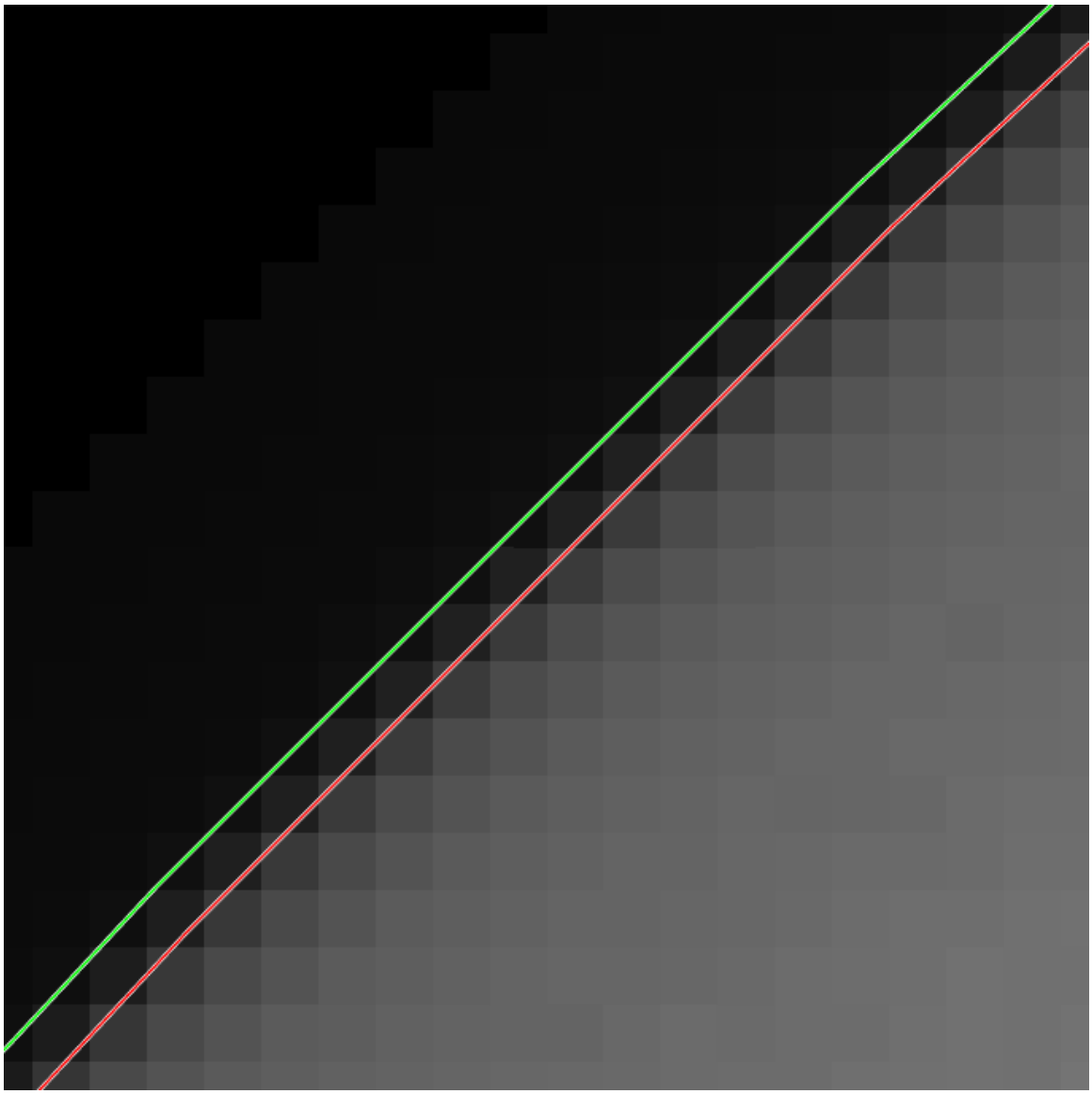

Figure 5. Comparison of results for Canny edge detection method (red) and binary thresholding method (green) determining the solar radius in a SOHO intensity image.

3.3.2 Flat-Field and Window Degradation Correction.

As mentioned, it is known that bad flat-field corrections and window degradation corrections have been applied [19]. Flat-fielding is a typical astronomical correction that adjusts the raw output based on known variations in pixel sensitivities [17], so that if all pixels were hit with the same intensity of light, the output would be similarly

consistent. Window degradation is a problem specific to SOHO/MDI and is a time-dependent reduction in the instrument sensitivity [20]; with no correction, images taken in 1997 would be brighter than those taken on 2011. Increases in exposure time were adopted to decrease this problem, but are a stepwise solution to a continuous issue. In order to produce calibrated and usable observations, both of these issues must be taken into account.

Table 4. Exposure changes were periodically enacted to counter time-dependent degradation of the instrument sensitivity. [21]

Dates of Exposure Changes
January 19, 2001 at 1938
January 17, 2003 at 2008
June 11, 2004 at 1935
March 21, 2006 at 2012
June 17, 2008 at 2004

A true flat-field correction requires a specially exposed image of a completely even light source. It is impossible to produce truly correct flat-fields at a later time due the time-dependent nature of the correction, so an alternative method must be used. A complex method involving masking active regions similar to that used by Potts and Diver on high resolution SOHO/MDI images [22] was initially examined, but a simpler solution was determined to be sufficient.

SOHO/MDI observations were split into six month increments, with additional splits at times of exposure change [21]. The images were placed into three dimensional matrices and the median was taken along the time dimension, resulting in a two dimensional image where each pixel is the median value of all the values of that pixel in the time frame. Due to the kinetic nature of sunspots, it is unlikely a specific pixel

will be darkened by activity over a sufficient period of time to cause the median to be a value other than that of the quiescent sun (Figure 6).

Figure 6. Example SOHO median intensity image.

A similar median image is produced for carefully selected SDO/HMI continuum intensity images, from spotless observations (see Appendices). Since the algorithm was designed to act on such images, which have previously been corrected, this median image can be considered the correct image of the quiescent sun. Therefore a correction image is formed by subtracting the SOHO median (Figure 6) from the SDO median (Figure 7), producing a correction matrix that can be simply added to the uncorrected SOHO images, seen in Figure 8. The pixel intensity distributions in Figure 9 illustrate the differences between the SOHO and SDO intensity distributions, for which the correction image compensates.

Figure 7. Example SDO median intensity image.

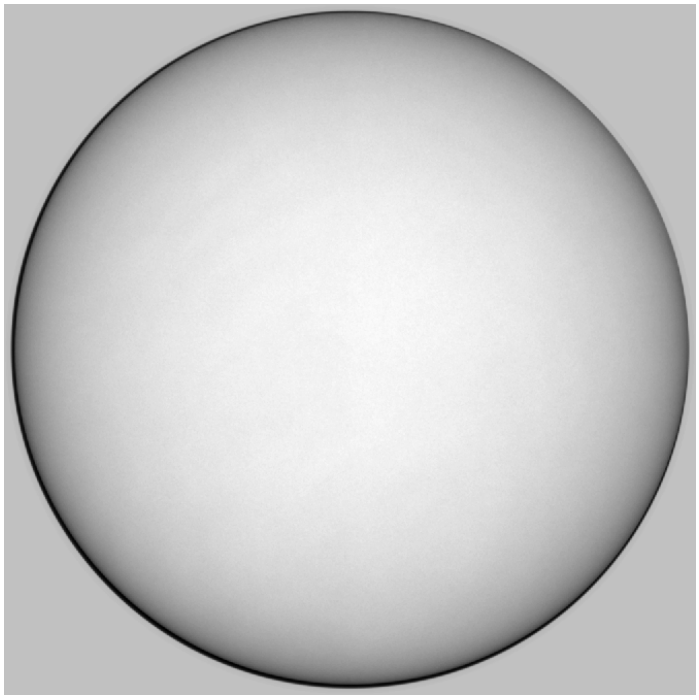

Figure 8. Example correction image, which is the difference between SDO and SOHO medians.

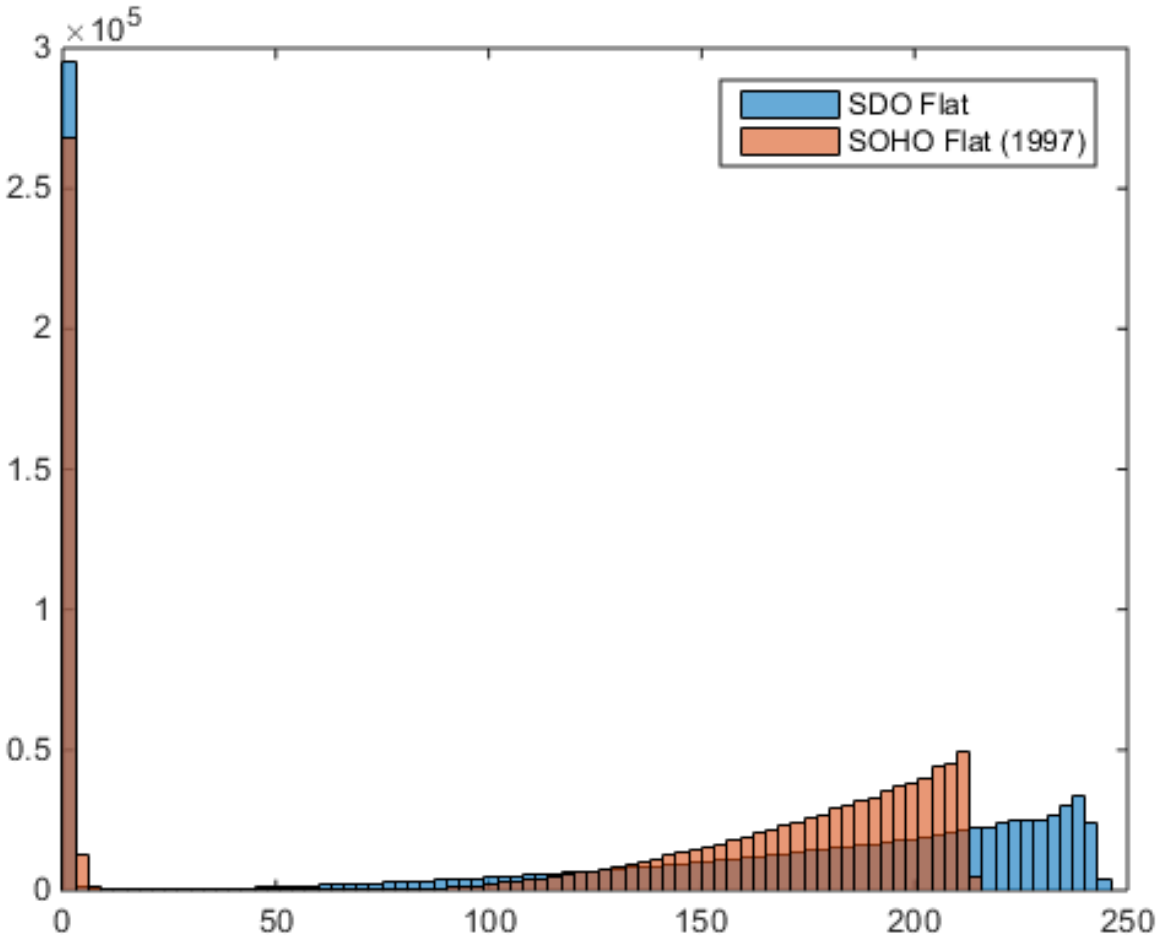

Figure 9. Comparison of histograms for SDO (red) and SOHO (blue) medians (also called flats). Note the different values for the peak counts, which is adjusted for by the application of the correction image.

Due to the variation in pixel size of the observed solar disk over time, it is necessary to resize and center images so the correction is aligned consistently. The center was selected to be the center of matrix [512.5,512.5] and the radius was selected to be 493, based on a radii observed empirical testing.

3.3.3 Limb Darkening Correction.

After the SOHO images have been corrected for flat-fielding and window degradation, there is still an additional correction for limb darkening that must be applied.

Due to the effect of Beer's Law discussed in chapter II, the solar emissions in the visible wavelengths are less intense towards the edge of the sun; particularly in a method that determines spots based on variations in intensity, it is necessary to flatten the intensity across the disk. Otherwise, the limb will appear dark like a sunspot.

This correction is achieved by applying a mask (adding a matrix) according to the linear version of Eddington's approximation, following $I = 1 - \frac{1+\sqrt{3}\rho}{1+\sqrt{3}}$. A visual represenation of that matrix can be seen in Figure 10 and example result of applying this matrix to a SOHO continuum intensity image can be see in Figure 11. Though the correction reduces the effect of limb darkening, it does not remove it entirely.

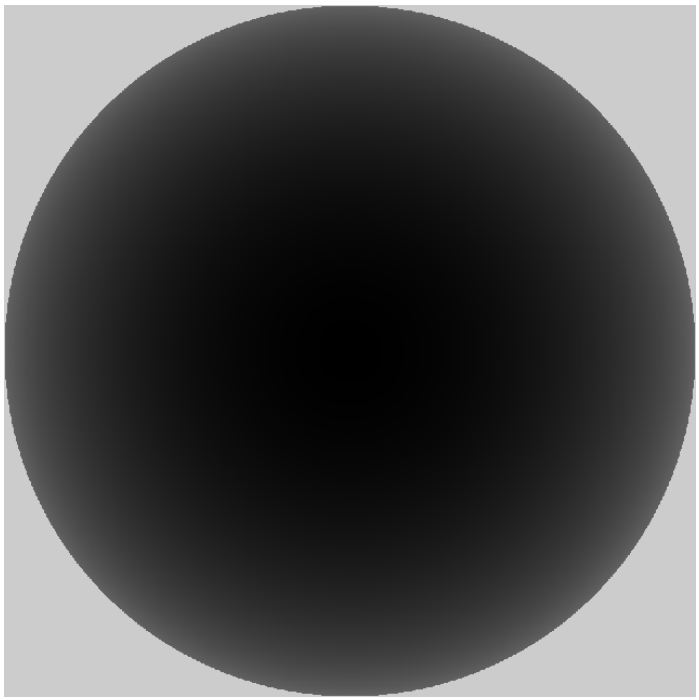

Figure 10. Limb darkening correction matrix.

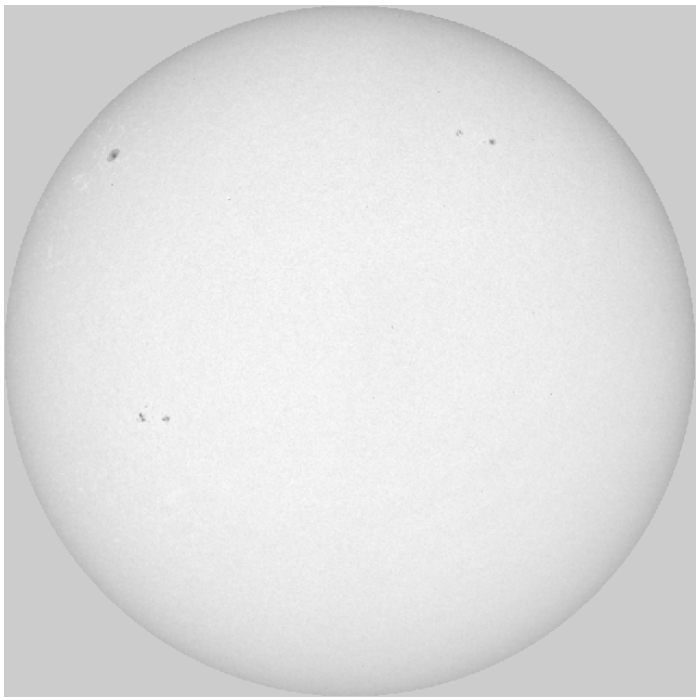

Figure 11. Example resulting limb darkening corrected intensity image.

3.3.4 Removal of Limb Edge.

SOHO observations have a low resolution and signal-to-noise ratio compared to SDO. Due to the very low signal (low intensity) at the edge of the solar limb, the noise overcomes the signal in such a way as to prevent the thresholding method from executing properly (discussed later in Subsection 3.4.1). Therefore, it is necessary to remove this outer limb. Data is lost, but it would not be reliable due to the high noise and drastic observation angle.

Pixels within the outermost 4% of the solar disk are set to the same value as the section of the image outside the solar disk. Necessarily, center and radius determinations cannot be performed after this adjustment, as they would result in incorrectly small values.

3.4 Identification and Grouping of Sunspots

After the image has been processed, the procedure for identifying sunspots, grouping them into active regions, and classifying those regions according to the McIntosh system can be applied.

3.4.1 Thresholding.

Binary Conversion. The sunspot detection is a fairly simple algorithm. The continuum intensity image is inverted, such that pixel values that were low (dark) on the original image become high (bright), as seen in Figure 12. This inverted image is converted to a binary matrix via the MATLAB function `im2bw`, where values below a specified threshold are set to 0 (false) and those above are set to 1 (true). Since sunspots are dark in the visible spectrum, the pixels they cover will be the brightest regions on the inverted image, and therefore 'true" in the binary matrix.

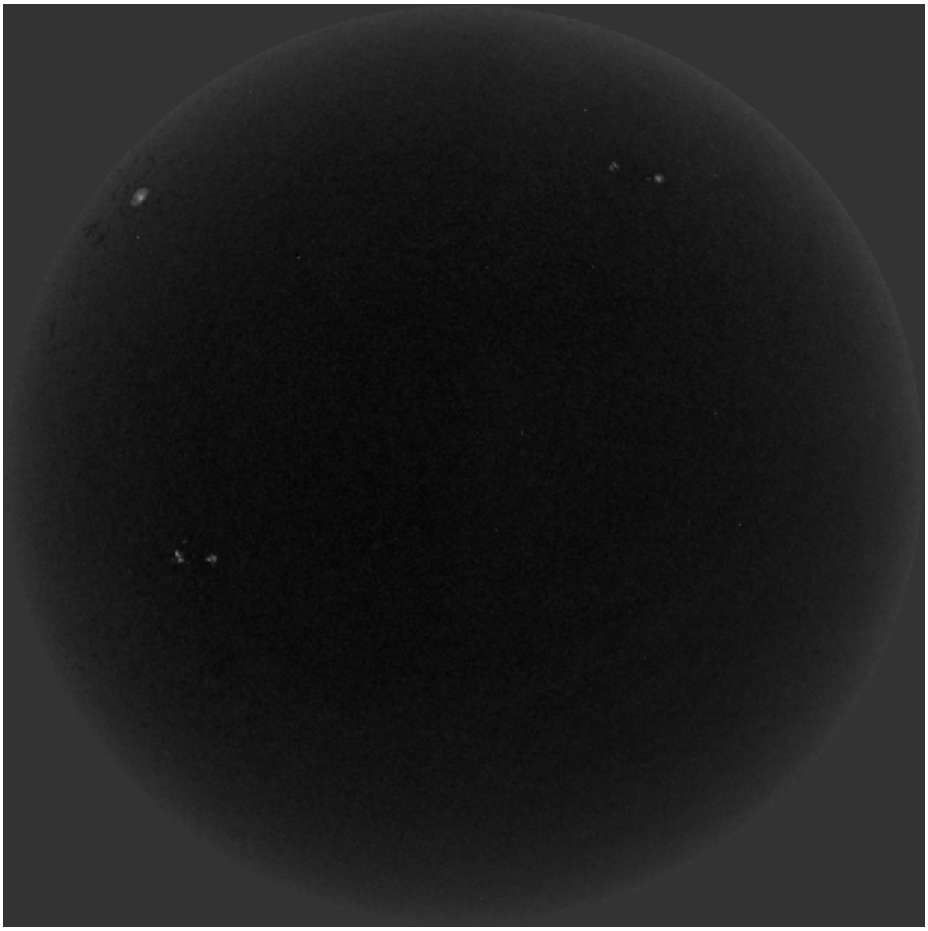

Figure 12. SOHO continuum intensity image inverted for thresholding. Note the active regions now appear as bright spots.

Counting of Sunspots and Iteration. The MATLAB function `bwlabel` is applied to the binary image, which identifies and labels 8-connected components (includes adjoining faces and diagonals) within the matrix. This separates "true" pixels into separate spots, such that each component is one penumbra. The total number of spots (components) is reported.

The threshold is initially set at 0.4, a value empirically determined by visually examining the values typical of penumbra. However, this is just an informed guess and the conversion to binary is repeated for varying thresholds. This iterative process is continued until the number of detected spots increases logarithmically, indicating

the noise level in the image has been reached. The threshold is then reverted to the previous value and the resulting binary image used for further analysis.

Noise Filtering. Once the spots have been identified, it is necessary to clean out noise from the results, such as that in Figure 13. This is accomplished by first applying a simple median filter on the continuum intensity image; this sets the value of a pixel to the median value of the pixels in a 2x2 neighborhood. This eliminates unusually high or low values in single pixels, which are typical of noise. For a binary image, this removes single values that are unlike the surrounding values, such as a single pixel "true" amidst a neighborhood of "false" values. Due to the low resolution of the SOHO observations, the neighborhood was reduced to 2x2 from the 3x3 used with SDO data.

To this filtered image, a morphological opening is performed. This utilizes a diamond shaped structuring element that removes are remaining single pixels with extreme values, preventing a stray pixel from appearing as a spot.

Figure 13. Initial result of thresholding with noise present throughout the image. A crescent of noise, seen in the upper right, has been misidentified as spots.

The magnetic information is then incorporated by converting the magnetogram to binary, where values greater than 205.5 or less than 50 are "true" (the magnetogram data is shifted from being centered at zero, so these values are equivalent to very positive and very negative Gauss measurements) and less extreme values are "false". Figure 14 is an example of the binary image of magnetic extremes. Sunspots only occur in areas of high magnetic activity, so any spots detected in regions without extreme values for magnetic activity are likely to be errors.

Figure 14. Binary image of the magnetogram extreme values, greater than 205.5 and less than 50.

The binary, filtered, opened continuum intensity image is finally multiplied by the binary magnetogram image, resulting in detections with minimized noise and in regions of high magnetic activity. This can be seen in Figure 15. The noise previously observed (Figure 13) is no longer present.

Figure 15. Final thresholding result after filtering, after the noise has been filtered out.

The classification of sunspot groups requires both penumbra and umbra measurements, therefore a second image is produced for a threshold 0.2 greater (a difference determined empirically). This image is the umbras, the darkest region of a sunspot.

3.4.2 Group Definition.

Individual sunspots must be grouped in regions for classification; spots tend to appear in clusters around high magnetic fields. Once determined, the properties of these regions are then evaluated compared to the specifications of the McIntosh system.

Spherical Trigonometry. The image of the solar disk in the SOHO observations is a two dimensional projection of the three dimensional solar hemisphere visible to the satellite. Therefore, a simple measurement of distance in pixels is not sufficient for locating sunspots; a pixel at the center of the solar disk covers less area than one at the limb. Making use of information available in the FITS header, including the B-angle, Carrington longitude of the observer (here the observer is the SOHO satellite), and the solar diameter in radians, accurate heliographic latitude and longitude locations can be determined. These angles describe the geometric relationship between the satellite and the sun.

For the original application to SDO, longitude was reported in the system Central Median Distance (CMD). This is a stationary longitude system only applicable to the face of the Sun visible to Earth [1] and is more commonly referred to as the Stonyhurst longitude [23]. This is considered appropriate [24] given SDO's geosynchronous orbit provides it with very nearly the same perspective as an observer on Earth [25] and sufficient for comparison to Air Force observations.

A second longitude measurement is the Carrington longitude, which is a system that rotates with the sun [1] and is reported on NOAA's solar region summary in addition to the Stonyhurst longitude. SOHO maintains an orbit around the first Lagrange point (reference Chapter II) which causes its location in relation to Earth to vary. Therefore a correction is necessary for an accurate longitude in either system to be determined. The Carrington latitude of the SOHO satellite is available in the FITS header keywords, while the Stonyhurst longitude is not [23], therefore eliminating the Stonyhurst longitude as an option.

The latitude and longitude calculation was modified for the SOHO satellite in light of this difference. The Carrington longitude is found by adding the observer's Carrington longitude (L_0) to the longitude as calculated for the visible face of the Sun,

which accounts for the difference in perspective. Based on derivations in Smart [26], the latitude and longitude are calculated from the following equations, which now include the Carrington longitude of SOHO at the time of observation:

$$\begin{aligned} B &= \arcsin(\sin B_0 \cos\rho + \cos B_0 \sin\rho \cos\theta) \\ L &= \arcsin(\sin\rho \sin\theta \sec B) + L_0 \\ \text{where:} \rho &= \arcsin(\tfrac{r}{r_0}) + SD(\tfrac{r}{r_0}) \end{aligned} \tag{6}$$

Table 5. Relation between the angles at the observer-sun system, as used in Equation 6.

	From the perspective of the observer (SOHO)
B_0	latitude of the center of the solar disk
L_0	longitude of the center of the solar disk
SD	angle the solar diameter subtends

Magnetic Polarity. The magnetic field of the component spots is important in the classification of a sunspot group; bipolar groups recieve different classifications than unipolar groups. The polarity of a spot is determined by taking the average of the values in the magnetogram over the pixels of interest; if the average is greater than 127.5 (effectively 0 Gauss), then it is considered positive. If the average is less than 127.5, it is considered negative.

Acceptance as Group Member. For a new spot to be added to a region, it must satisfy one of two conditions:

1. If it is the same polarity as other spots already in the region, it must be within 3 heliospheric degrees of another spot in the region.

2. If it is the opposite polarity, it must be within 15 heliospheric degrees of another spot in the region.

During this process, each region is noted as being either unipolar (includes only a spot or spots of one polarity) or bipolar (includes spots of both polarities), which is utilized in the classification step.

3.4.3 Classification.

Once the sunspots have been grouped into regions, region properties are calculated and a flowchart is applied to determine the classification. This process produces the same classifications each time it is run, meaning the results are consistent.

Region Properties. The MATLAB function `regionprops` is used to determine geometric information about each region.

1. Area: the number of pixels in the region

2. Centroid: the pixel location [x,y] of the region's center of mass

3. Eccentricity: the eccentricity of the region

4. Extrema: pixel locations [x,y] of the extreme points of the region

These data are used to determine information for either the classification or reporting of region information. Area is converted to millionths of a solar hemisphere

(MoSH), centroid is converted to latitude and longitude, and extrema are converted to latitude and longitude and used to determine the length of the region in degrees. Eccentricity is used as-is to determine the symmetry of the region, where a value less than 0.5 is considered symmetrical.

Decision Tree. From the data collected in the grouping function and `regionprops`, all the necessary information is available to classify the regions according to the McIntosh system. A separate function determines each component of the classification, following the decision trees in Figures 16, 17, and 18, which are also described in detail in Chapter II.

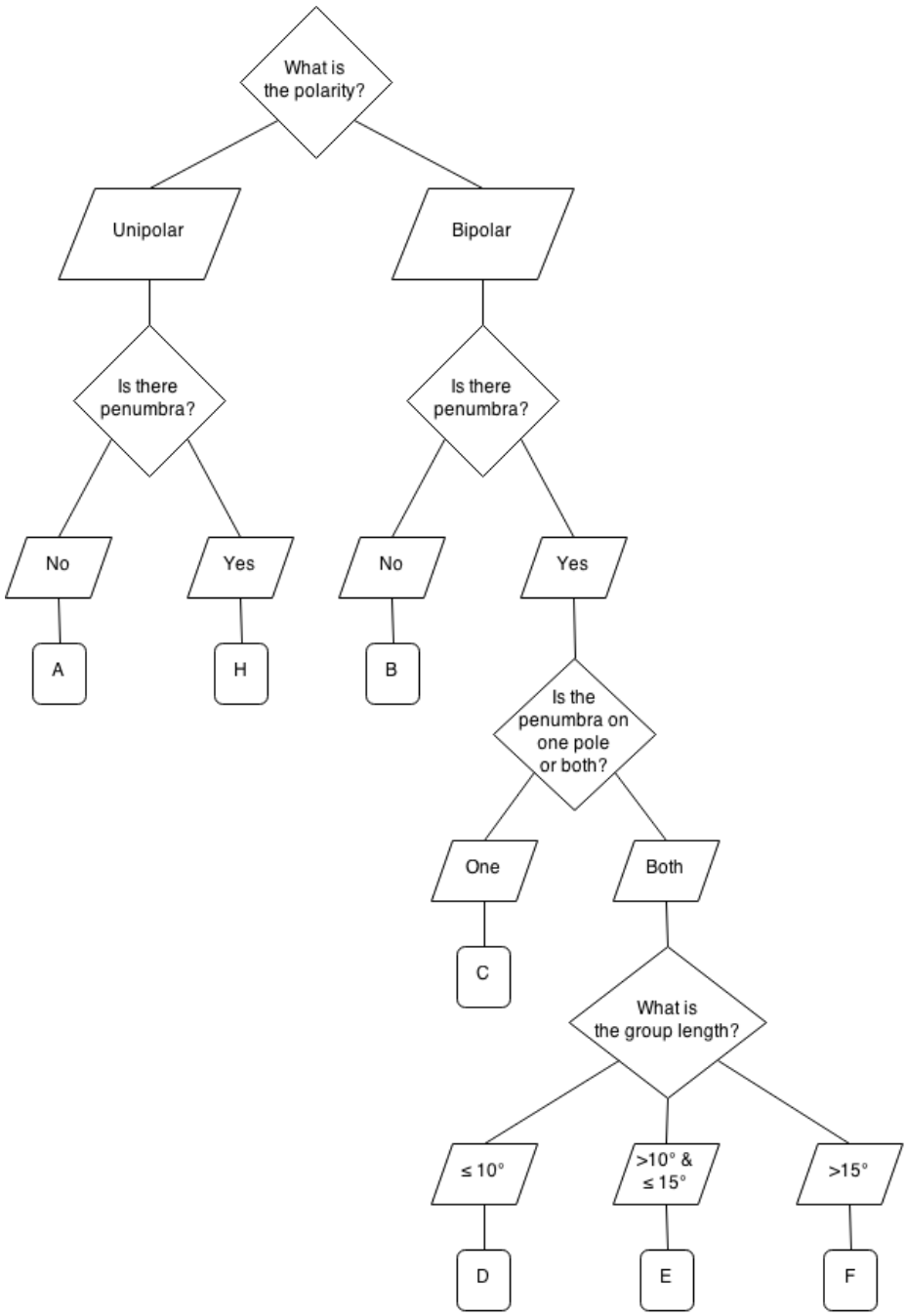

Figure 16. Flow chart for deciding modified Zurich class component of McIntosh classification, based on Air Force Weather techniques. Adapted from [1]

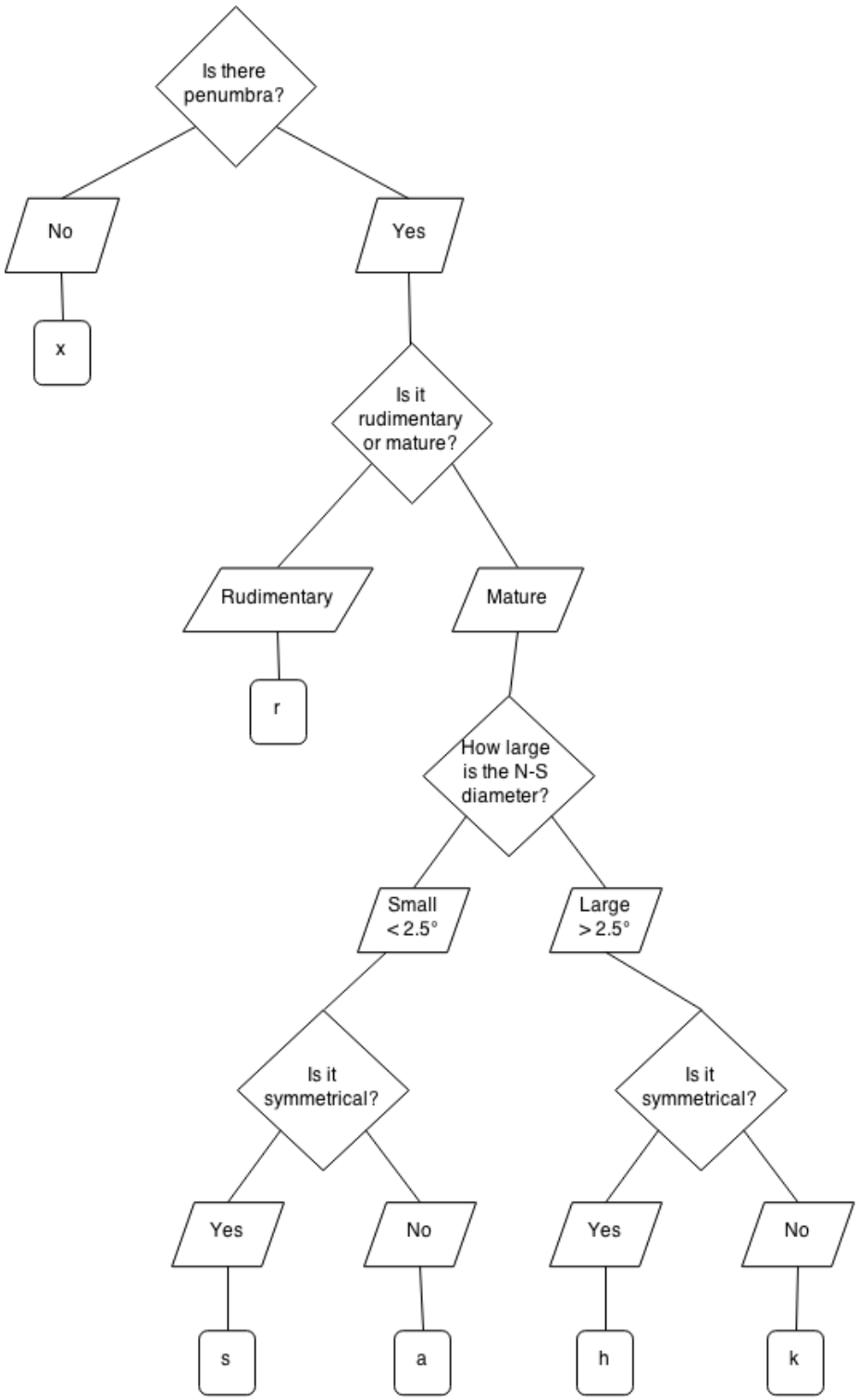

Figure 17. Flow chart for deciding penumbra component of McIntosh classification. Adapted from [2]

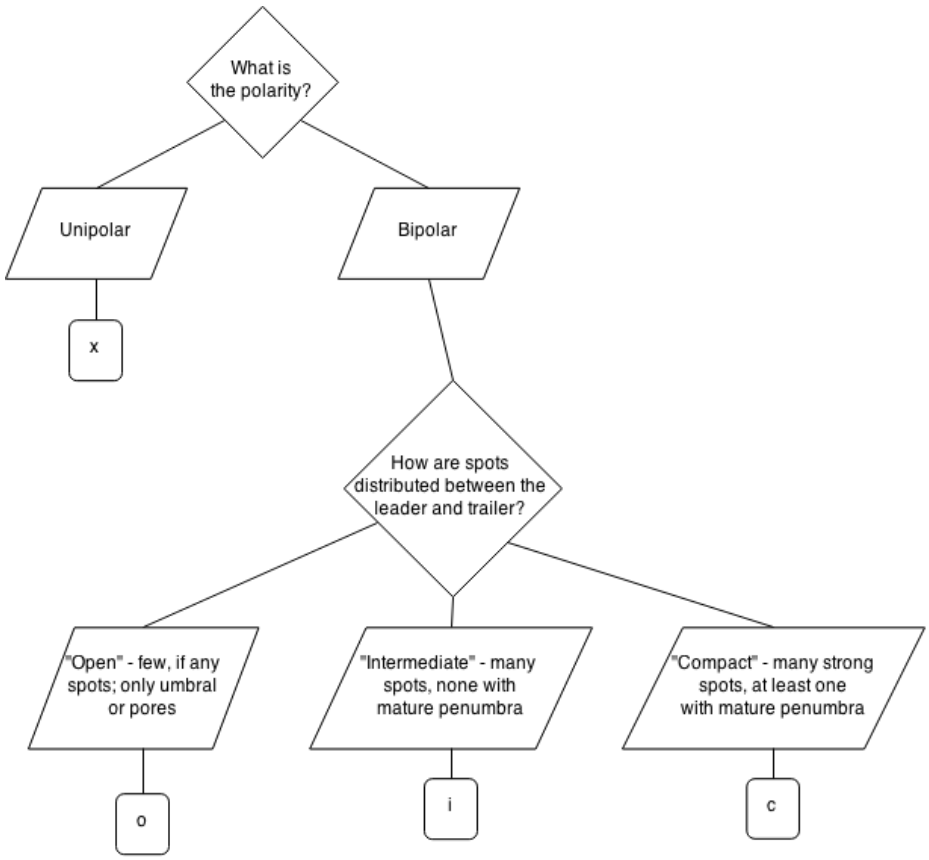

Figure 18. Flow chart for deciding spot distribution component of McIntosh classification. Adapted from [2]

3.5 Text Output

For each observation, a text file is generated with a filename that includes the observation date and time. Within the body of the file, the date the code was run is included, followed by the important identifying details (latitude, longitude, length, area, and number of umbras) and McIntosh classification for each region as shown in Table 6.

Table 6. Output data for each region and the associated units. Note that for latitude and longitude, a negative value indicates South and West locations, respectively.

Latitude	Longitude	Length	Area	Umbras	McIntosh
Degrees	Degrees	Degrees	Millionths of a Solar Hemisphere	Scalar	3 Characters

3.5.1 Removal of Incomplete Images.

During generation of the database, an unexpected issue was encountered. A portion of the available data were not sufficient for detection of sunspots; specifically, it appears that the CCD readout was incomplete. This can be seen in Figure 19, in which the lower portion of the image contains no information and therefore displays as black.

This was rectified by performing a check before processing and applying the algorithm. The portions of the image that are not properly read out have the same value as outside the solar disk on a typical image (appearing at -32678 before scaling, as discussed in Subsection 3.2.1). The percent of the image that is outside the solar disk on a correct image is known to be less than 25%, therefore if an image has greater than that percentage of pixel values equal to -32678, it is incomplete.

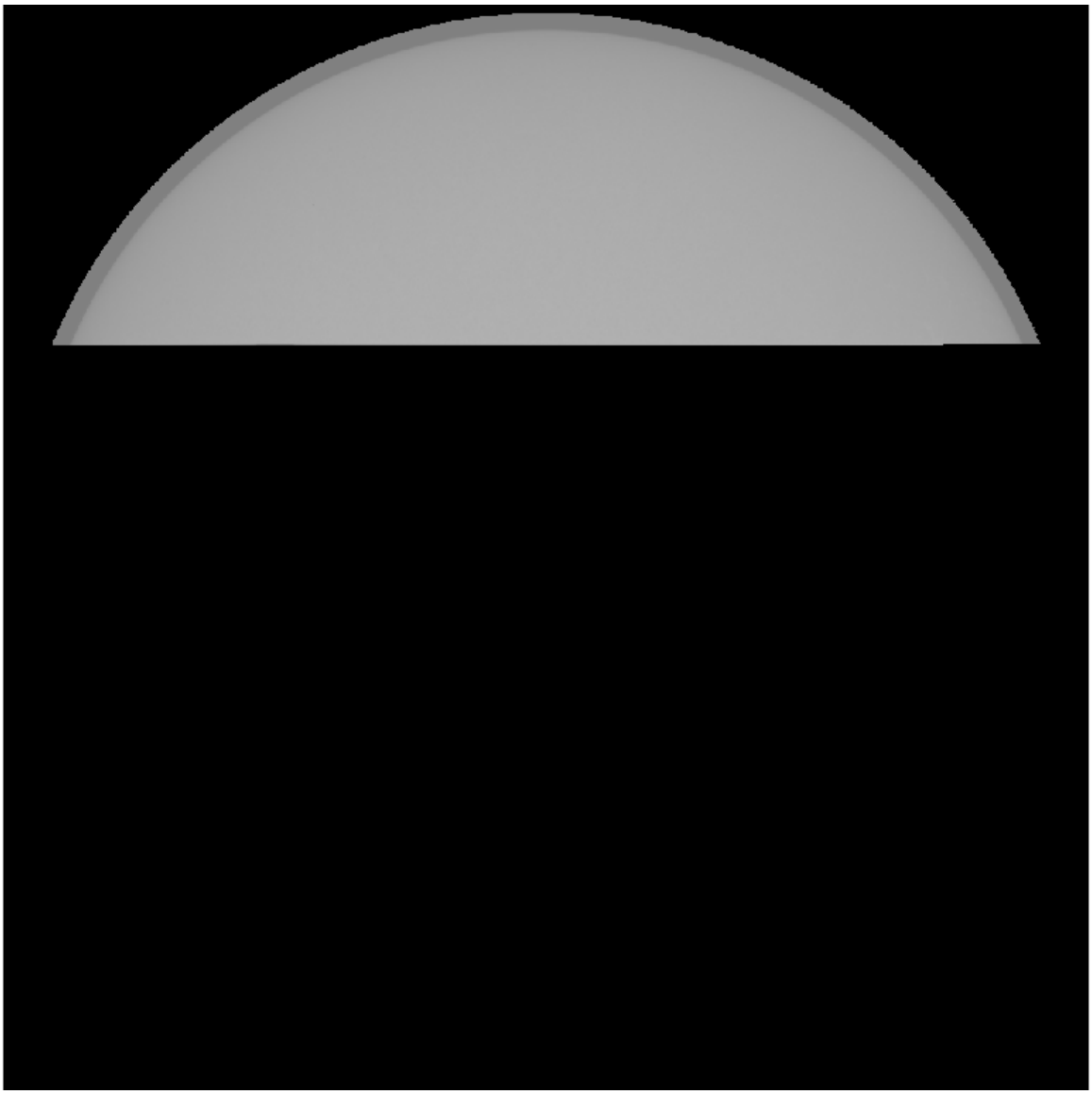

Figure 19. Sample continuum intensity image dated 16 April 2000 at 0800 demonstrating incomplete CCD readout.

IV. Results and Analysis

4.1 Database

The main product of the algorithm is a database of all the detected regions and associated properties for the observations where the continuum intensity and magnetogram times match to within a few minutes (the difference is typically only 30 seconds). The observations span fifteen years from 23 August 1997 to 11 April 2011.

Generation of the database resulted in a total of over 8,000 output files. Each output file corresponds to an observation. The number of observations varies by year, depending on operations of both the SOHO satellite and MDI instrument, as well as the data available. Due to the instrument's observation cadence, the occurrence of time-matched intensity images and magnetograms was not as often as the desired once per hour. Observations in 1997 and 2011 are particularly few because those include the start and end of the instrument's functioning. 1998 has fewer observations than it might have, due to a loss of communication with the SOHO satellite from June to October. 2009 has fewer due to problems retrieving data sets from the JSOC database.

Table 7. Number of output files for each year of SOHO/MDI.

Year	Number
1997	242
1998	709
1999	822
2000	689
2001	744
2002	762
2003	588
2004	618
2005	676
2006	545
2007	590
2008	542
2009	381
2010	677
2011	259

4.2 Signal-to-Noise Ratio

To examine the quality of the algorithm output for the SOHO data (hereafter simply referred to as the SOHO results), the signal to noise ratio (SNR) is calculated. In this situation, the signal is the pixels associated with the sunspots and the noise is the rest of the image. However there are observations where there are no spots, which would result in an undefined SNR. Instead, an alternative definition common

in image processing [9] is used. This is the ratio of the mean of the pixel values to the standard deviation of the same (Equation 7).

$$SNR = \frac{\mu_{pixels}}{\sigma_{pixels}} \tag{7}$$

Plotting the SNR over time, a distinctive pattern emerges. Particularly evident in the SNR calculated for the raw images and after correcting flat-fielding and window degradation (Figure 20) is a sinusoidal variation that aligns with SOHO's orbital period of one year. The SNR actually decreases from an average of 3.477 to 3.329 after the correction has been applied, but this is not unexpected since the correction is designed for errors, not noise, and introduces its own noise to the data.

Comparing the SNR of the raw images to that of images with the limb darkening correction and after removal of the limb edge (Figure 21) reveals that removing limb darkening great improves the SNR, bringing the average from 3.477 to 18.333. This factor of five increase demonstrates the power of removing variations in the background. Removing the troublesome limb edge raises the SNR another 25% to an average of 22.556.

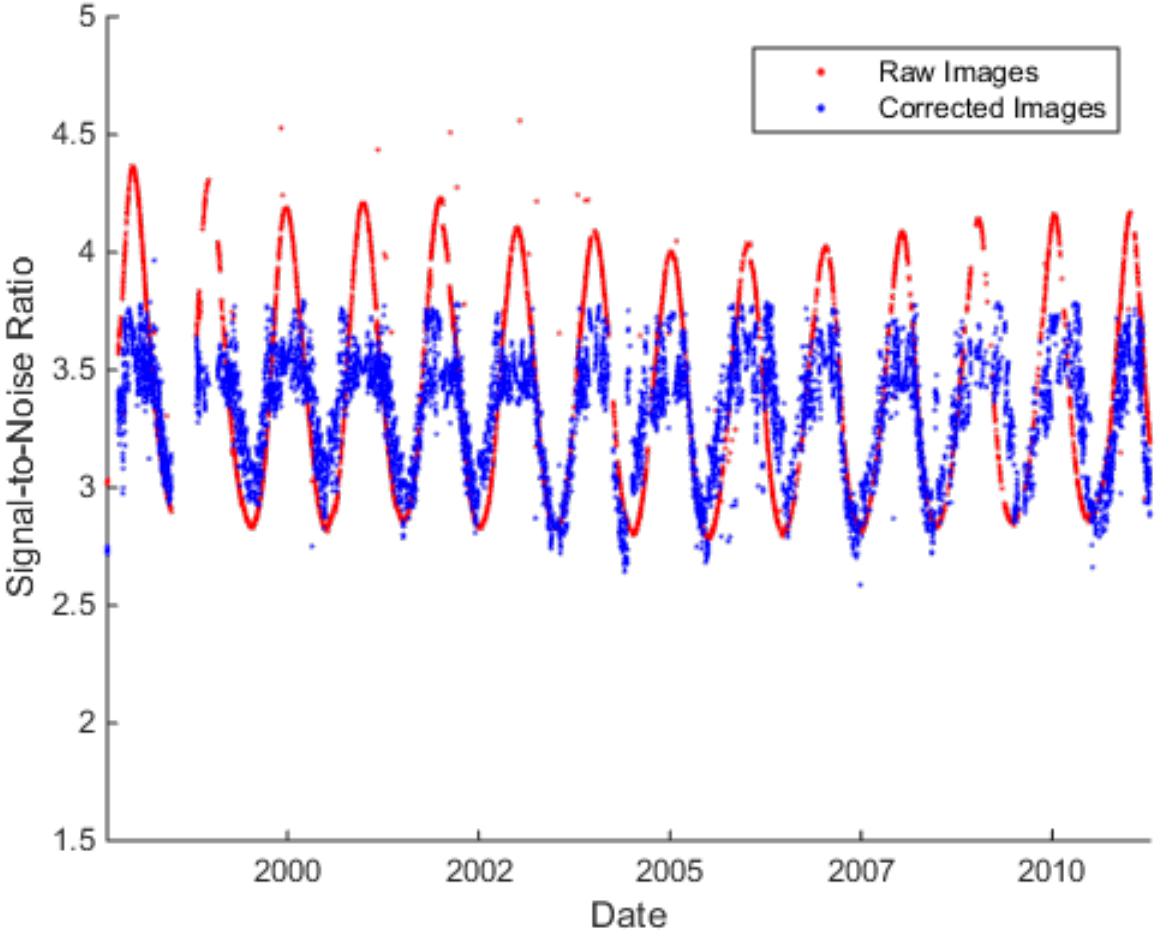

Figure 20. Signal-to-noise ratio over time for SOHO intensity images, both raw and corrected for flat-fielding and window degradation.

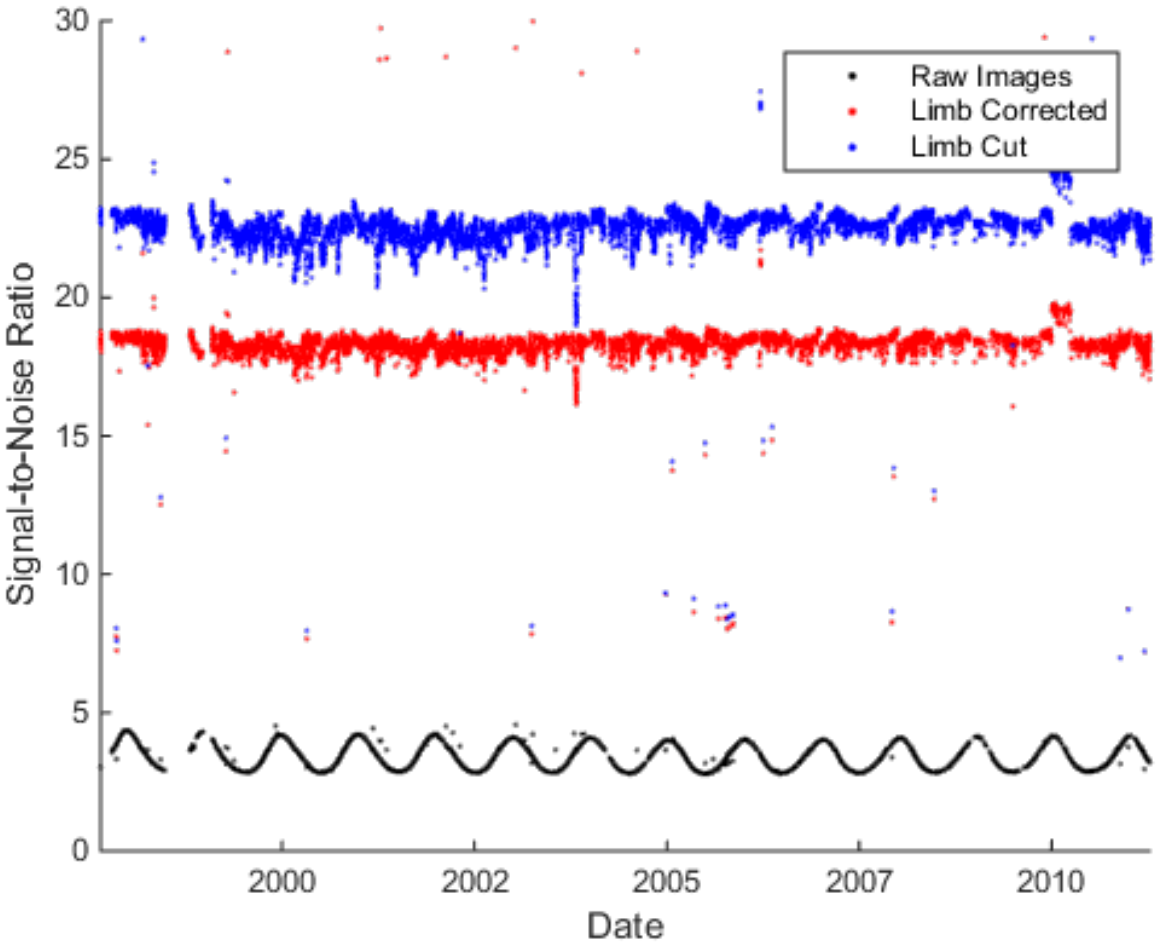

Figure 21. Signal-to-noise ratio over time for SOHO intensity images, including raw, corrected for limb darkening, and with the limb edge removed.

Examining the SNR shows that the corrections applied adjust the images from a poor quality (SNR 4) to a good, but not excellent, quality (SNR 20). It is cautiously concluded that the images are of sufficient quality to run the algorithm. Though the SNR well describes the relative amount of noise present in the image, it does not consider how much *information* (such as small sunspots) may be lost in the process.

4.3 Full Date Range

To get an overview of the SOHO results, consider all the observations over the full range of dates. Instead of examining each invidual region, a summation (total) of region properties over each observation is used for analysis. These properties include length (degrees), area (millionths of a solar hemisphere), number of spots, and number of regions. This allows large scale patterns to be discerned, such as those expected by the solar cycle.

4.3.1 Outlier Data.

The initial overview demonstrates that some outliers are present in the data. Particularly, there are a small number of data points in the summed region areas that are several orders of magnitude larger than the rest of the data (Figure 22). Data for summed length, number of spots, and number of regions have outliers that correlate to those of the summed region areas.

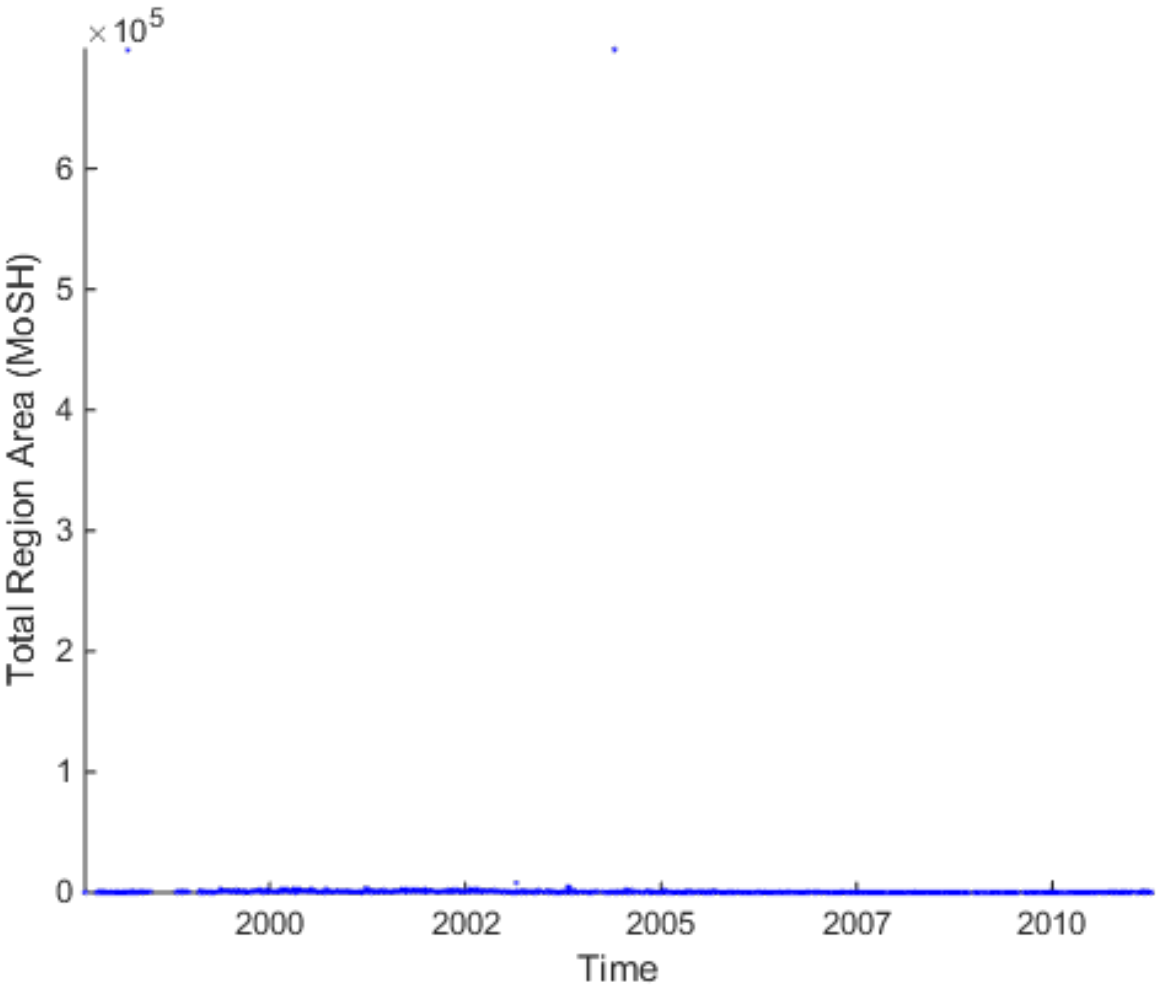

Figure 22. Summed region areas Outliers are present along the top edge of the plot.

The outlier values for the summed region area are on the order of 60% of the entire solar hemisphere, which is not physically possible. Therefore, these are removed from the data set before further analysis is completed. The resulting plot (Figure 23 is closer to the expected result.

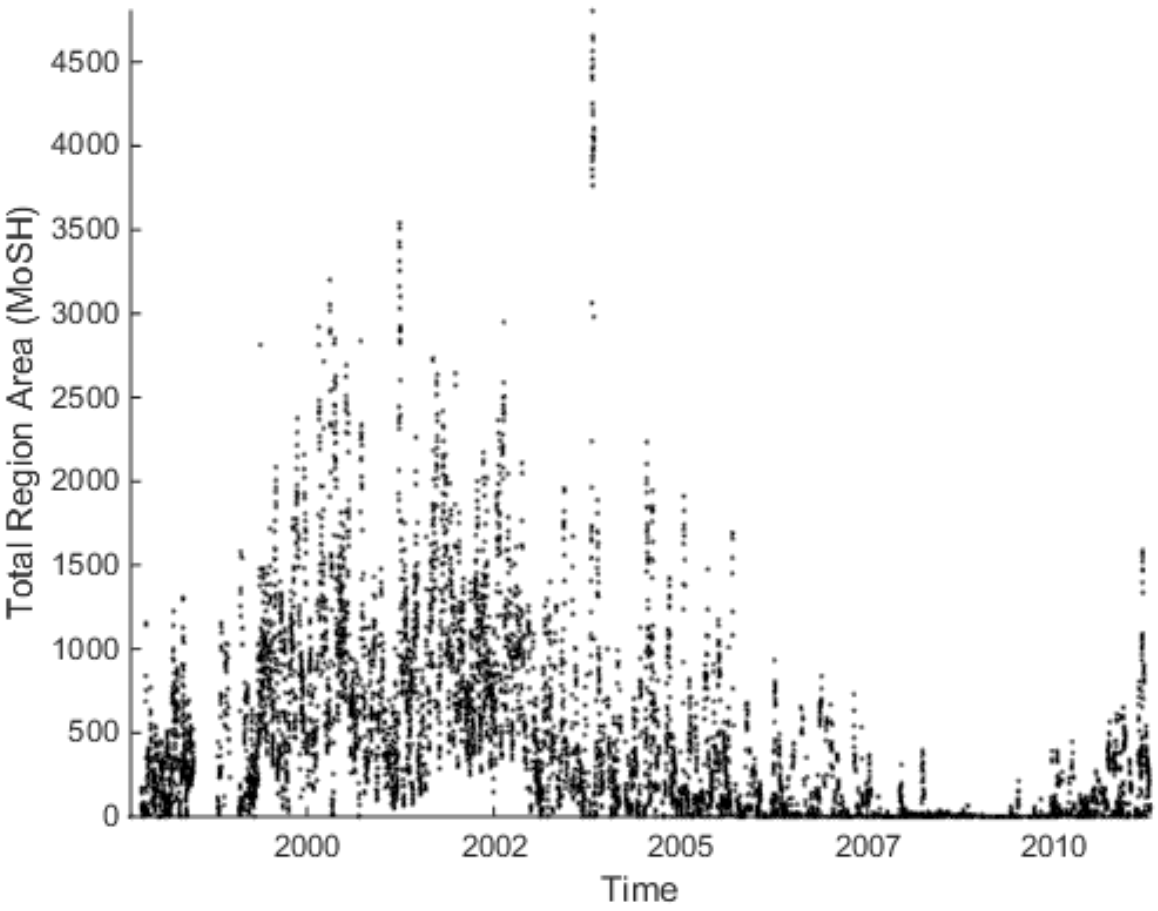

Figure 23. Summed region areas (millionths of a solar hemisphere) in each observation SOHO results after the removal of outliers.

4.3.2 Overview and Solar Cycle.

All of the summed region properties reflect the expected sinusoidal variations of the 11 year solar cycle. In Figure 24, vertical lines mark the maximum of solar cycle 23 in March 2000 (red) and the minimum starting solar cycle 24 on January 4, 2008 (blue). The properties all follow a very close pattern of behavior.

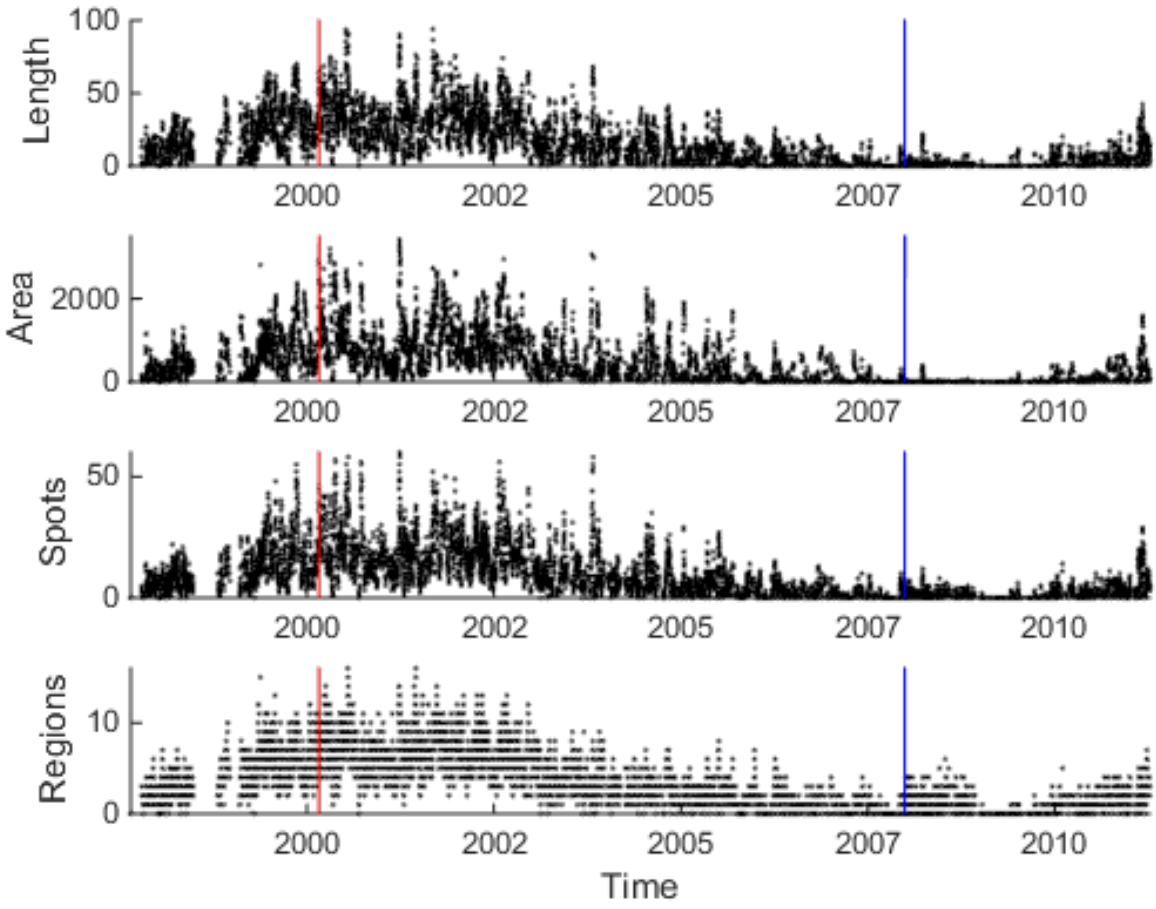

Figure 24. Summed region properties in each observation for the SOHO results over the full date range. From top to bottom: region length (degrees), region area (millionths of a solar hemisphere), number of spots, and number of regions. Vertical lines indicate the maximum of solar cycle 23 (red) and minimum starting solar cycle 24 (blue).

Since all demonstrate similar behavior, it is only necessary to consider one region property. The number of sunspots is a simple indicator of solar activity, so this is the selected property and can be seen in Figure 25. Peaks aside from the solar maximum are evident; these correspond with the Bastille day event on July 14, 2000, a CME ejection and X20 flare in early April 2001, and the Halloween storm at the end of October 2003.

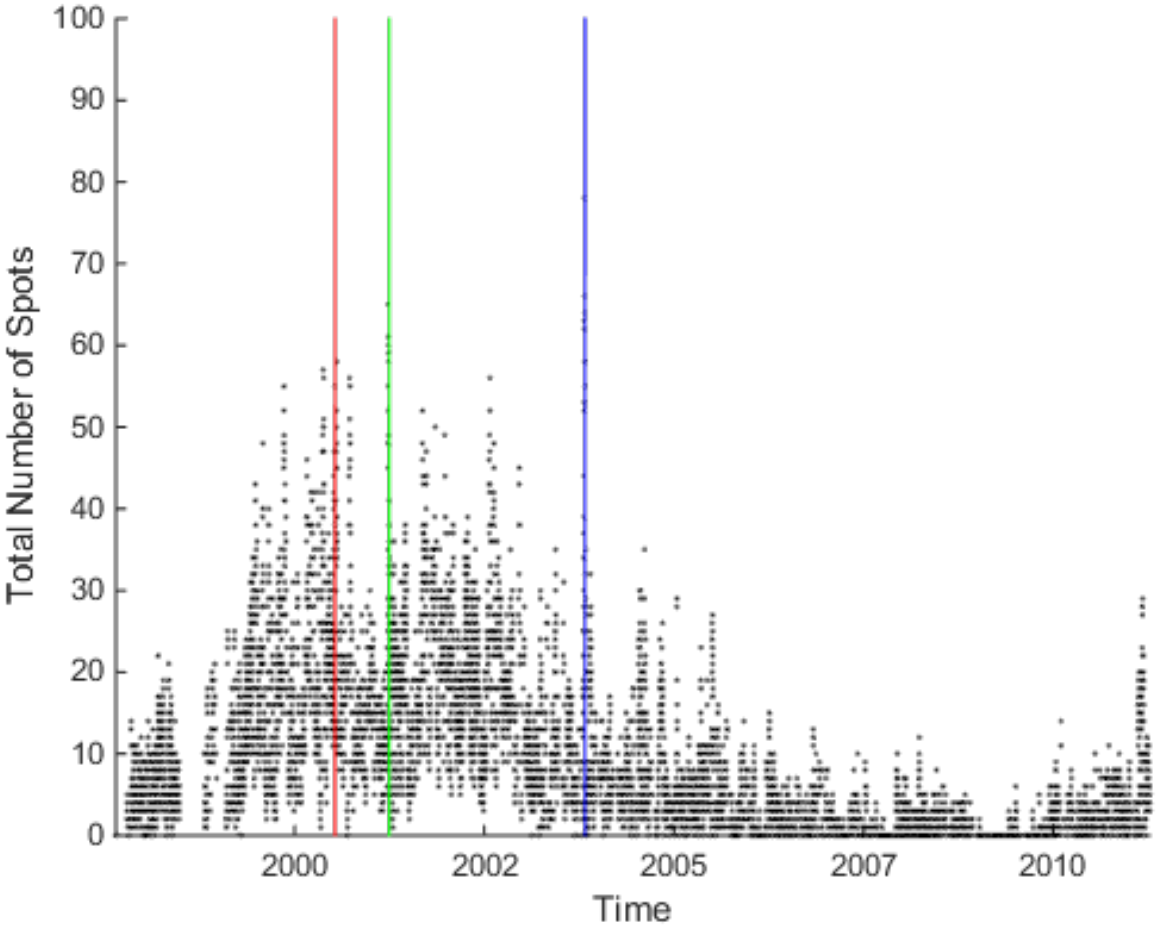

Figure 25. Total number of spots in each observation for the SOHO results over the full date range. Vertical lines indicate, from left to right: Bastille day event (red), CME and X20 flare (green), and Halloween storm (blue).

The current standard for solar activity data are NOAA's Solar Region Summaries

(NOAA SRS). Comparing the SOHO results with the NOAA SRS data indicates that the large scale pattern of the results are similar, though the magnitude of the SRS sunspot number is approximately four times higher than that of the SOHO results (Figure 26).

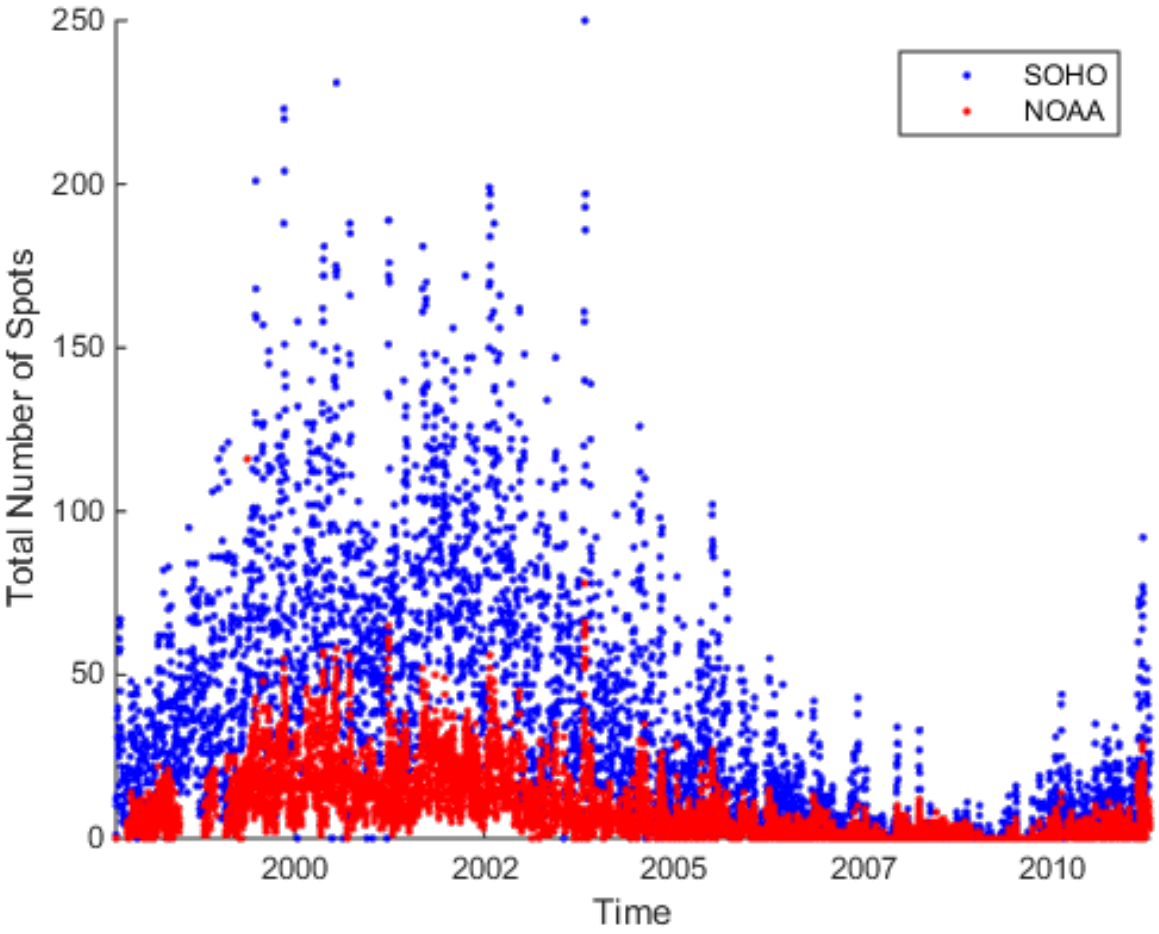

Figure 26. Number of spots in each observation for the SOHO results (red) and NOAA SRS (blue) over the full date range

4.4 Direct Comparison to NOAA Solar Region Summaries

For a closer examination of the accuracy of the SOHO results, comparison of individual observations is conducted. The NOAA SRS are published at 0030 UTC; SOHO results within ± one hour correspond sufficiently and minimize differences associated with changes in the solar activity with time.

4.4.1 Summed Region Properties.

Given more than 2,000 relevant results, a comparison for the full date range is best performed on the summed region properties rather each region. A more precise comparison with a smaller sample is considered in the Subsection 4.4.2.

Plotting the SOHO results as a function of NOAA SRS allows for a linear regression model to be fit to these data. Ideally, the SOHO results and NOAA SRS would be nearly identical, with an intercept of zero and a slope of one. The R^2 value would be 1, indicating the data aligns perfectly with the regression model. An exact match is actually not expected, due to the algorithm's lack of dependence on human observers. The results do indicate there is a strong relationship between the two ($R^2 > 0.75$ for all models), though it is not exact.

The total region areas (Figure 27) are particularly similar, with the highest values of slope and R^2 of all the considered properties. The slope value of 0.806 indicates the areas detected by the algorithm are approximately 80% off those reported by NOAA, a reasonable result given the more precise calculation of area by the computer. The R^2 value of 0.872 means the data are well correlated.

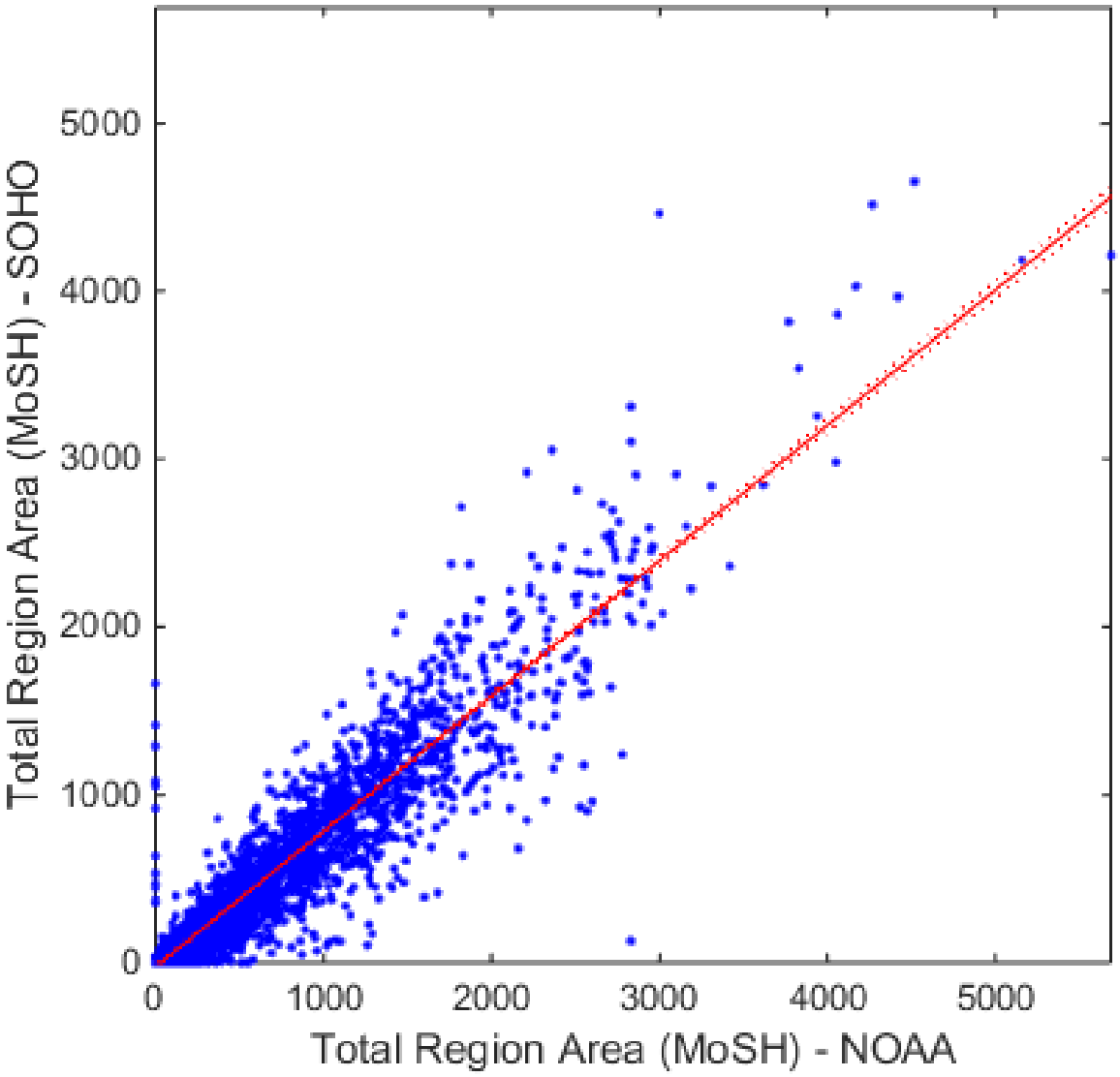

Figure 27. Summed region areas (millionths of a solar hemisphere) in each observation for time matched SOHO results and NOAA SRS.

The total region lengths (Figure 28) have only a slope of 0.577, suggesting the algorithm consistently measures less than 60% of the lengths reported by NOAA. This may also be due to the precision in calculation, but such a low value indicates there is possibly another factor related to other properties.

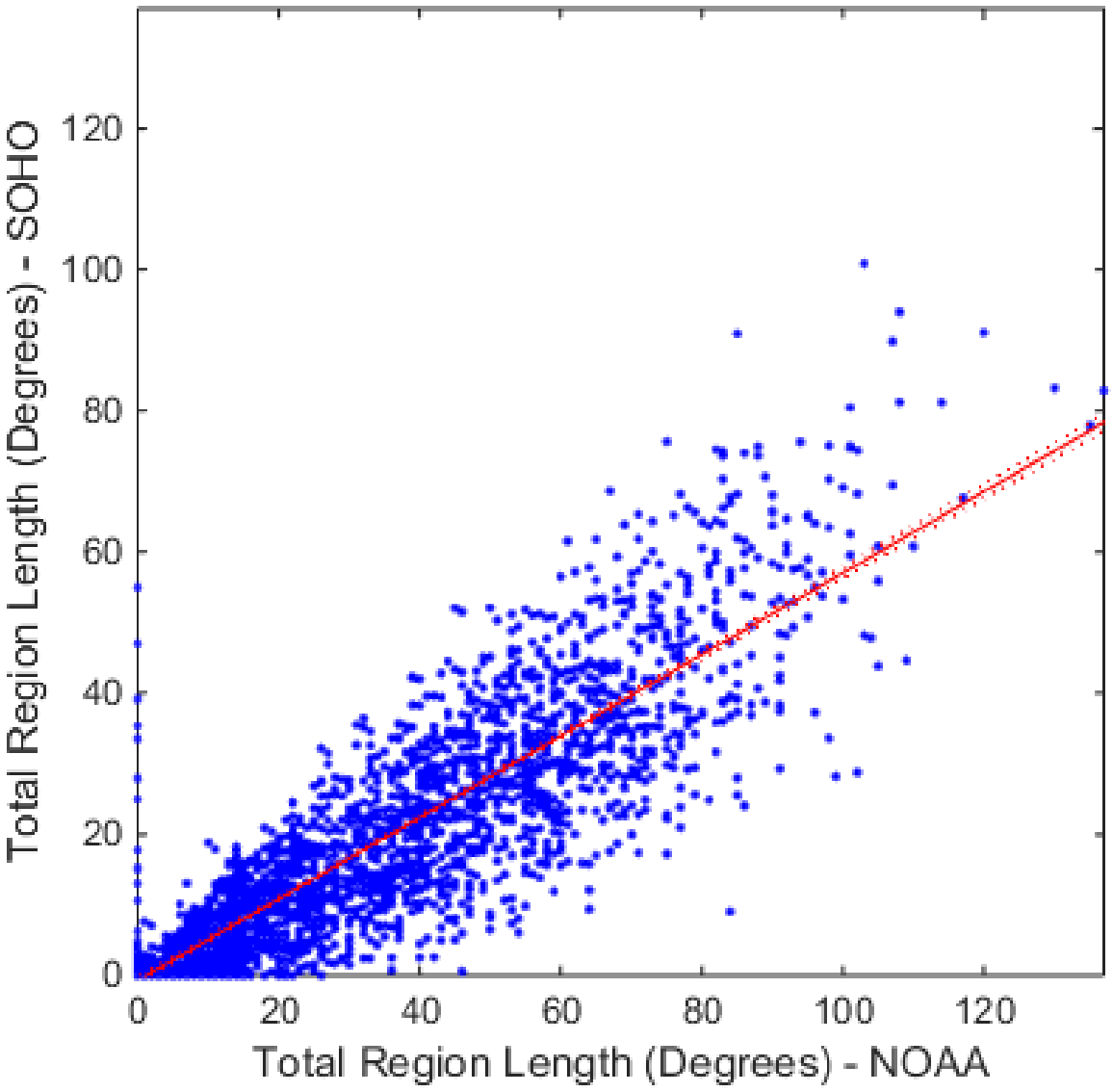

Figure 28. Summed region lengths (degrees) in each observation for time matched SOHO results and NOAA SRS.

The number of spots (Figure 29) detected has the lowest slope of the considered properties, only 0.235. Less than 25% of the spots reported by NOAA are also noted by the algorithm. This is significant and indicates the algorithm is missing a large number of sunspots. Yet the R^2 value of 0.820 suggests that the SOHO results still have a strong linear relationship with NOAA's values, though at a ratio closer to 1:4 than 1:1.

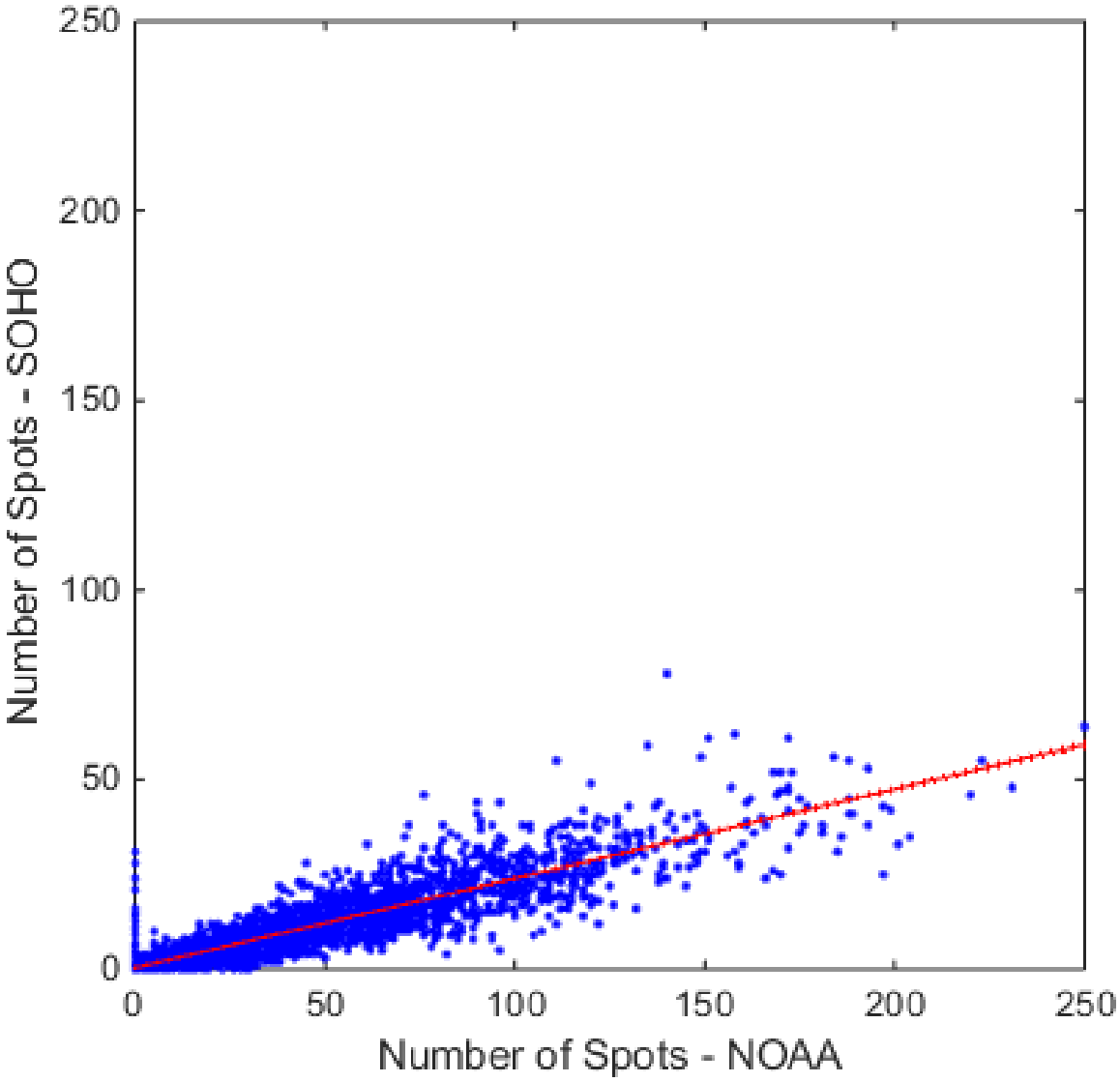

Figure 29. Number of spots in each observation for time matched SOHO results and NOAA SRS.

The number of regions (Figure 30) has a slope of 0.650 and an R^2 of 0.782; this is the lowest correlation and a middling relationship. The algorithm groups purely based on magnetic polarity and distance between spots, possibly leading to differences from the human-determined grouping of spots.

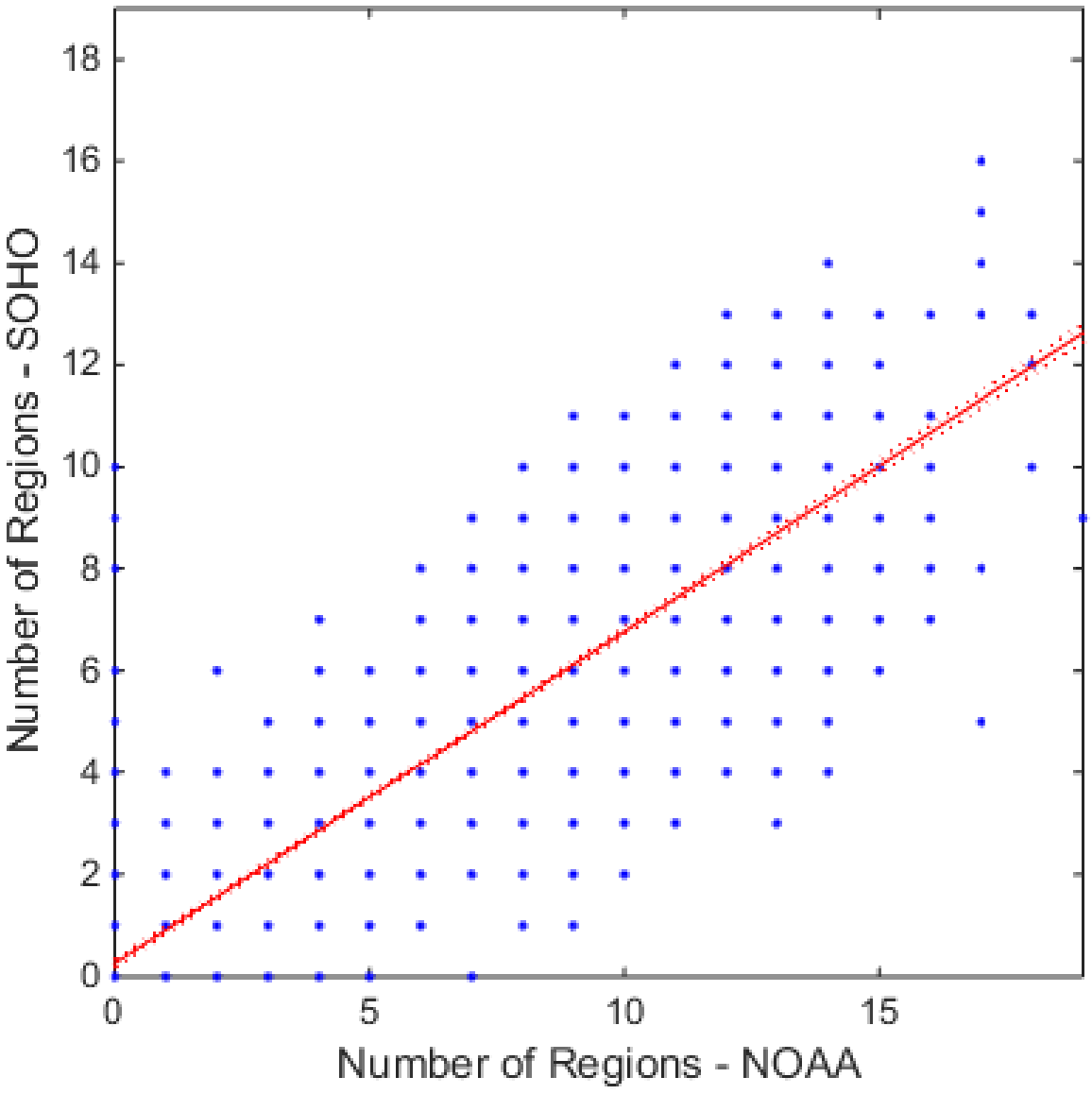

Figure 30. Number of regions in each observation for time matched SOHO results and NOAA SRS.

The regression models are summarized in Table 8.

Table 8. Results of fitting a linear regression model to each summed region property, SOHO as a function of NOAA.

	Intercept	Slope	R^2
Total Area (MoSH)	-22.907	0.806	0.872
Total Length (°)	-0.688	0.577	0.790
Number of Spots	0.466	0.235	0.820
Number of Regions	0.264	0.650	0.782

4.4.2 Region by Region Evaluation.

Sunspot group classification is on a region-by-region basis, so the most precise comparison is between individual regions within the time matched observations. The high number of relevant observations, greater than 2,000, prohibits a full examination of the entire set. A pseudo-randomly determined selection of date-matched observations, after outlier removal, were considered for such a comparison. Ten such dates were generated, listed in Table 9. Though purposely selecting dates of high activity would lead to more groups to compare, such a restriction would prohibit the application of statistical measures.

Table 9. Pseudo-randomly selected dates for region-to-region comparison.

Year	Month	Day
1999	September	26
	November	22
2002	June	14
2003	February	4
2007	February	18
	April	28
2009	May	30
2010	May	14
		21
	June	14

The observations for May 30, 2009, May 14, 2010, and June 27, 2010 have no detected regions for either the SOHO results or NOAA SRS data. This is excellent, demonstrating that the algorithm does not detect false positives for a quiet Sun.

May 21, 2010 has a single region detected by the algorithm, while none were reported by NOAA. Consulting the processed intensitygram for SOHO's observation at that time (Figure 31) shows that there is indeed a region, small as indicated by the SOHO classification of Axx and a calculated area of 4.05 MoSH. Apparently the NOAA observers missed this small region.

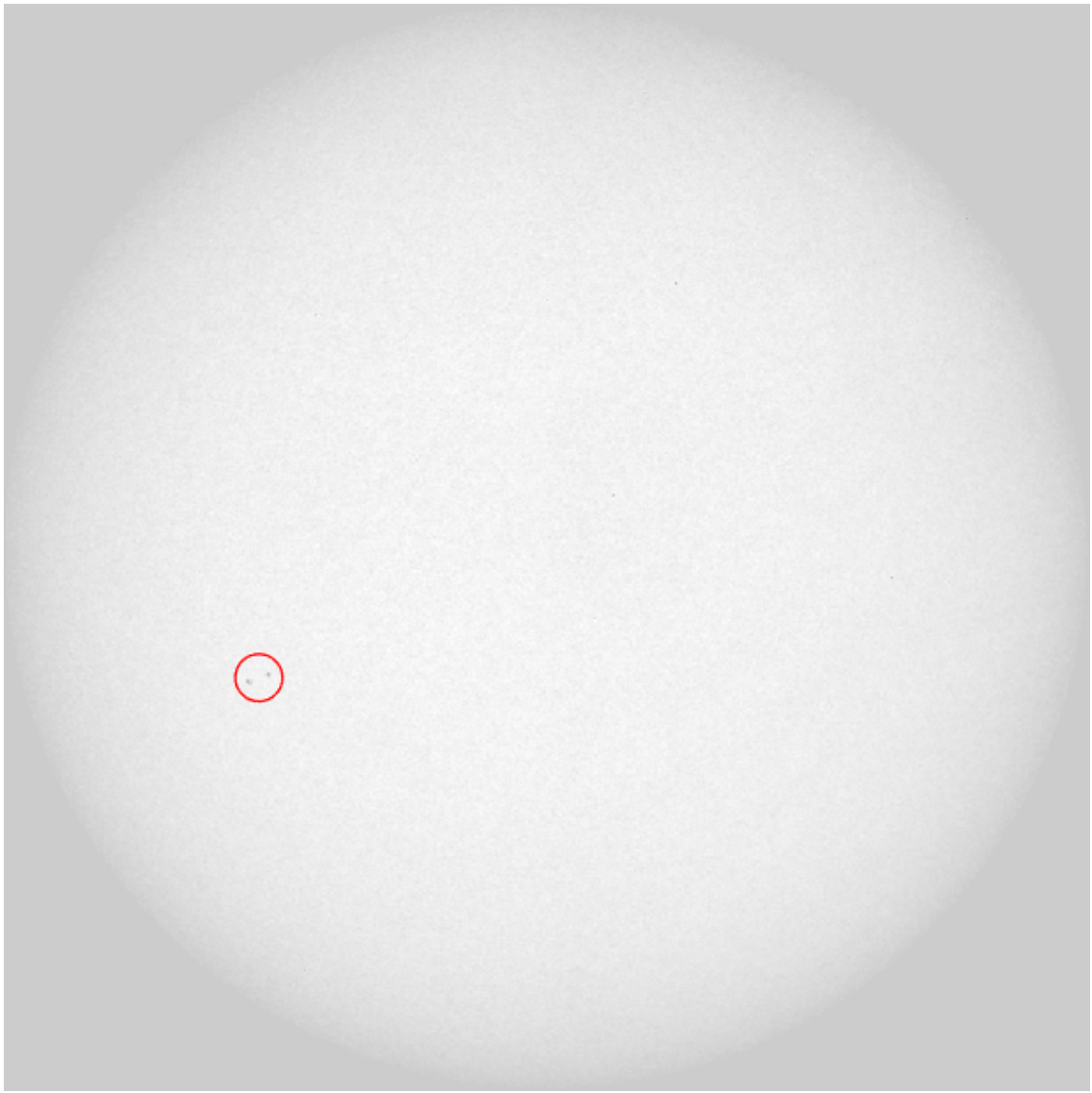

Figure 31. Processed SOHO intensitygram for May 21, 2010. The circle indicates the detected region.

The next simplest comparisons are for those with only one detected region in both data sets. This occurs for both February 18 and April 28, 2007. The region properties and classifications are detailed in Table 10. The numerical values of the region properties are very similar, except for the number of spots on April 28 where SOHO only detected 25% of the spots reported by NOAA. This is consistent with the regression results from the full analysis.

The classifications do not match, but are similar. The region on February 18 is classified as Axx by the algorithm but Hsx by NOAA. This means the algorithm did not detect a penumbra around the spot, while NOAA observers saw a symmetrical penumbra. For April 28, the classification by the algorithm is Cao and by NOAA Dkc. These variations are explained by difference in the detected penumbra, particularly the size, symmetry, and location on the magnetic polarities.

Table 10. Region properties and classifications for February 18 and April 28, 2007.

	Latitude (°)	Longitude (°)	Length (°)	Area (MoSH)	Spots	McIntosh
February 18, 2007						
SOHO	-10.65	132.84	0.92	25.64	1	Axx
NOAA	-11	132	1	20	1	Hsx
April 28, 2007						
SOHO	-9.97	307.62	4.52	416.03	2	Cao
NOAA	-10	308	7	500	8	Dkc

The observations from September 26, 1999 (Table 11) and February 4, 2003 (Table 12) each have two regions detected by the algorithm and three regions reported by NOAA. Properties for regions with close locations are very similar, except in general the number of spots in a region. The second region detected on September 26, 1999 is almost an exact match — the difference is only the symmetry of the penumbra.

The second region detected on February 4, 2003 is a perfect match of classification, despite the differences in numerical values for the area and number of spots. The regions not detected by the algorithm are small and have either rudimentary or no penumbra.

Table 11. Region properties and classifications for September 26, 1999.

	Latitude (°)	Longitude (°)	Length (°)	Area (MoSH)	Spots	McIntosh
SOHO	-20.93	252.27	2.44	129.23	1	Hax
NOAA	-21	245	6	110	3	Cso
SOHO	19.34	208.48	1.61	86.17	1	Hax
NOAA	20	207	2	90	1	Hsx
NOAA	-11	188	3	10	4	Bxo

Table 12. Region properties and classifications for February 4, 2003.

	Latitude (°)	Longitude (°)	Length (°)	Area (MoSH)	Spots	McIntosh
SOHO	-13.76	160.52	13.1	300.84	11	Bxo
NOAA	-14	160	12	280	25	Eai
SOHO	-5.9	224.04	5.77	97.87	3	Dso
NOAA	-5	223	7	130	13	Dso
NOAA	-4	301	1	40	1	Hrx

June 14, 2002 presents a more complicated comparison (Table 13), with three regions detected by the algorithm and nine reported by NOAA. The number of spots in the SOHO data are much less than those in the NOAA data, causing the classifications to be significantly different. This is particularly evident in the third region, where the algorithm detected only one spot with an area of 5.71 MoSH, compared to NOAA's seven spots encompassing an area 60 MoSH.

Referencing back to the SRS text file for this date, the region at latitude +18 and Carrington longitude 170 had a Stonyhurst longitude of West 95, placing it right on the solar limb. Therefore this region was likely missed due to the necessary data removal from the limb. For the other undetected regions, consider the processed intensitygram for SOHO's observation on this date (Figure 32). On visual inspection these regions cannot be seen, due to either resolution or noise in the image.

Table 13. Region properties and classifications for June 14, 2002.

	Latitude (°)	Longitude (°)	Length (°)	Area (MoSH)	Spots	McIntosh
SOHO	-21.03	84.39	2.94	295.54	1	Hax
NOAA	-20	84	7	310	7	Cho
SOHO	-15.07	143.65	1.6	33.83	2	Bxo
NOAA	-15	146	12	340	12	Eho
SOHO	-18.57	67.25	0.44	5.71	1	Axx
NOAA	-18	67	3	60	7	Dao
NOAA	18	170	5	60	1	Hsx
	12	116	3	20	4	Bxo
	14	16	1	10	1	Axx
	9	16	0	10	1	Axx
	6	149	1	10	2	Hrx
	-28	40	1	10	1	Hrx

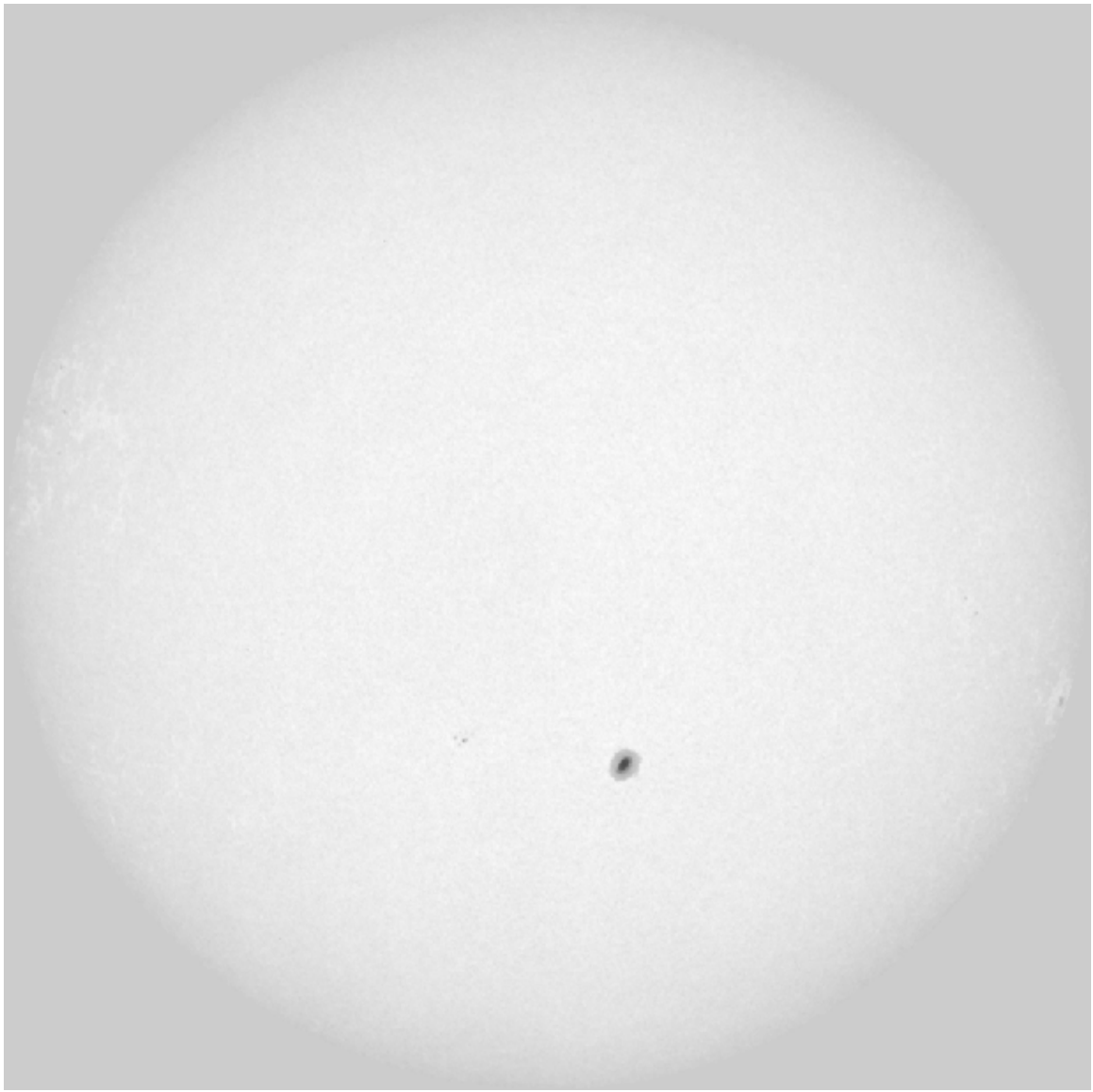

Figure 32. Processed SOHO intensitygram for June 14, 2002.

The randomly selected date with the most regions is November 22, 1999, with nine regions in both the SOHO results and the NOAA SRS data. All regions detected by the algorithm have a corresponding region reported by NOAA, but only one classification is an exact match. As in the previous comparisons the main variation is in the number of spots, which is expected from the results of linear regression model. This accounts for most of the differences in classifications, though the symmetry of the

penumbra is also a factor. The length and area values are close for all corresponding regions, within a factor of two for most and all within an order of magnitude. It is particularly important to note that the SOHO results significantly downgrade the classification for several sunspot groups, largely due to unobserved sunspots.

Table 14. Region properties and classifications for November 22, 1999.

	Latitude (°)	Longitude (°)	Length (°)	Area (MoSH)	Spots	McIntosh
SOHO	-13.78	235.61	13.08	1108.72	2	Hkx
NOAA	-12	235	12	790	25	Ekx
SOHO	-16.52	174.36	19.08	340.82	10	Fao
NOAA	-16	174	12	220	32	Eai
SOHO	-29.08	164.12	6.47	60.29	2	Dao
NOAA	-30	167	6	20	12	Dso
SOHO	14.84	237.28	1.49	33.35	1	Axx
NOAA	16	240	11	80	9	Cso
SOHO	17.53	211.2	10.25	82.93	6	Eao
NOAA	18	211	10	120	14	Dao
SOHO	-15.2	241.06	0.98	8.34	1	Axx
NOAA	-15	163	3	20	6	Bxo
SOHO	3.63	148.08	0.53	8.25	1	Axx
NOAA	4	146	5	20	8	Cso
SOHO	-12.35	111.36	0.76	7.64	1	Axx
NOAA	-14	113	4	10	2	Bxo
SOHO	-26.28	132.3	0.49	5.09	1	Axx
NOAA	-26	135	1	10	2	Axx

With the regions are now matched, the properties of each individual region can be compared using a paired t-test. This examines the distribution of the differences in the property values and tests the hypothesis that the difference is equal to zero. For latitude and longitude, this test can only be performed for corresponding regions, while the length, area, and number of spots for unmatched regions are compared to zero values. The null hypothesis of this test is $H_0 : \mu_{SOHO} = \mu_{NOAA}$. The resulting p-value is the lowest significance level needed to reject this hypothesis. A typical significance level is $\alpha = 0.05$.

The p-values of latitude, longitude, and region area are all greater than 0.2, a very high significance level. These properties are therefore statistically the same between the SOHO results and the NOAA SRS data. Yet the p-values for region length and number of spots are both less than 0.01, a low significance. Those two properties are not statistically the same between the two data sets, which is not unexpected. The low value of the slopes for region length and number of spots found by the linear regression models indicated the correlation between those calculations by the algorithm and the reports of NOAA. Exact p-values are reported in Table 15.

Region Property	P-Value
Latitude	0.425
Longitude	0.278
Length	0.009
Area	0.614
Spots	9E-5

Table 15. P-values for paired t-test on region properties in randomly selected sample of observations.

4.5 Test Case Comparison with SDO

To explore these apparent discrepancies further, a test case comparing code output for SDO and SOHO data from the same date and time was conducted, including an SDO image that was resized to the same resolution as SOHO (from 4096x4096 pixels down to 1024x1024). Due to the different systems for longitude (Stonyhurst versus Carrington), these could not be compared. The results in Table 16 suggest that resolution is a major factor in the low number of sunspots detected in the SOHO images. The resized SDO data only reports 27% of the spots and 75% of the regions reported by the full resolution data.

Table 16. Results for test case January 2, 2011 at 0000Z. Longitude has been omitted due to differences in the measurement system (Stonyhurst versus Carrington).

	Latitude (°)	Length (°)	Area (MoSH)	Spots	McIntosh
SOHO	31.99	3.31	229.20	1	Hax
	-13.21	6.19	71.51	2	Dao
	34.16	7.81	40.80	2	Dao
SDO	31.32	3.07	230.06	2	Hax
	-13.56	6.62	85.42	22	Dai
	34.03	8.75	59.35	16	Dai
	-28.68	0.16	0.63	1	Axx
SDO Resized	31.45	3.33	245.55	1	Hax
	-13.69	6.84	97.40	6	Dai
	34.08	8.57	70.51	4	Dai

This comparison points to resolution being the main factor in the low detection rate of sunspots in the SOHO results.

4.6 Discussion of Variations

Though the SOHO results for some region properties are fairly well correlated with the NOAA SRS data, others are not at all. Several factors are likely for the differences between the SOHO results and the NOAA SRS.

1. Low resolution: SOHO/MDI produces images only 1024x1024 pixels. Small or faint spots may not be resolved. Not only will this reduce the number of spots and regions detected, but may significantly affect the length and area of regions, particularly if spots at the edge of a group are missed. As examined in the test case (Section 4.5), the resolution is likely the main issue with the SOHO results.

2. High noise: any activity towards the limb is lost due to the noise. Since lengths and areas towards the limb appear smaller than those in the center, even losing a pixel may cause great variation in length and area. Entire spots or regions may be missed as well, such as in the data from June 14, 2002 in the region-by-region evaluation.

3. Instrument degradation: As mentioned previously (reference III), the light reaching the sensor was attenuated by window degradation, possibly causing the faintest of the sunspots to not be perceived.

4.7 Evaluation of Usability

Caution in further use of this data is recommended. The R^2 values for the summed region properties indicate a fairly good correlation between the results of the algorithm and the data reported by NOAA. However, the low p-values for region length and number of spots, as well as the rarely matching classifications point to the algorithm performing imperfectly on the SOHO/MDI data. As discussed in the previous section, this is most likely due to problems with the SOHO data itself.

The database produced here is sufficient for analysis of patterns and relative magnitudes, particularly involving those properties with higher p-values in the t-test (latitude, Carrington longitude, and region area). However, it is not suggested for such considerations as flare probabilities based on region classification or absolute magnitude of properties.

V. Conclusion

5.1 Summary of Results

Modification of the image processing in the algorithm was necessary in adapting it from use on SDO to SOHO. Particularly important are the flat-fielding and window degradation correction, which allow thresholding to apply properly to the image.

The database generated by the algorithm applied to SOHO/MDI contains individual observations. Compared to NOAA's Solar Region Summaries the data correlates well for region properties (length, area, number of spots, and number of regions) summed over an entire observation, though it is not a 1:1 relation. In specific region-by-region evaluation, location (latitude and Carrington longitude) and region area are statistically similar for regions matched between the database and the SRS data.

5.2 Future Work

Future work in this topic has three main branches: generation of a database using SDO observations, correlation of data to solar flare activity, and transition of this algorithm to operational use.

5.2.1 Generation of SDO Database.

SDO/HMI produces images of much higher resolution than those of SOHO/MDI and has not suffered from window degradation. It is still in operation and analysis of historical data could be directly applied to real time observations.

5.2.2 Solar Flare Correlations.

An analysis of the SOHO database correlations with solar flare activity may prove useful, despite the imperfections. Particularly, relative time variations in properties

show promise [27, 15]. The large range of dates that SOHO covers provides much historical information about solar activity.

A direct correlation of instantaneous region classification from SDO to solar flare activity could potentially provide a large improvement in current solar flare probability forecasting. Current statistics are based on the once daily SRS, though the region classification may have changed — possibly dramatically — between the last observation and the occurence of the solar flare.

5.2.3 Implementation of Algorithm for Operations.

The most important future work in this area is transitioning the algorithm to operational use. The speed and accuracy of the algorithm would allow multiple reports a day and free human analysts to conduct analyses which computers can not perform. Techniques have not changed since the 1970s despite large advances in technology. The Air Force's greatest resource is its personnel; operational use of the algorithm would allow better application of their time and effort.

Appendices

A Image Download and Processing

A.1 Download of Files.

In the Joint Science Operations Center (JSOC) database, the appropriate series are:

1. Full-Disk Continuum (Linearly Polarized), mdi.fd_Ic
2. Full-Disk Magnetograms, mdi.fd_M_lev182

The format for data requests is `[1999.01.01_00:00:00_TAI-1999.12.31_23:59:59_TAI@60m]`.

On the export page, selected method was `ftp` and the protocol was `FITS`. The filename format was modified to `cont.{T_REC:A}.{segment}` for continuum intensity data and `magn.{T_REC:A}.{segment}` for magnetogram data. This standardizes the filenames for easier ingestion into MATLAB.

The data location was copied to a .csv, which was read into the MATLAB and automatically downloaded using the function `urlwrite`.

Unfortunately, uncompressed FITS files are no longer available from the JSOC database.

A.2 FITS Header Information.

Within each FITS file is a header of information in addition to the data contents. This header includes a slope and intercept for the scaling of the raw data to appropriate units indicated in the file keywords, as identified in Table 17. Using the built in MATLAB function for reading FITS files (`fitsread`, specifying a 'raw' output) the content is read out in the same class it is stored, rather than automatically scaled and converted to double precision by MATLAB.

Table 17. FITS Information for Continuum Intensity and Magnetograms.

	Continuum Intensity	Magnetogram
Class (Raw)	signed 16-bit integer (int16)	signed 32-bit integer (int32)
Units (Scaled)	arbitrary intensity units	Gauss

Additionally, necessary angles for calculating locations on the solar disk are associated with particular keywords in the FITS header.

Table 18. Angles and their associated keywords in the FITS header.

Angle	Keyword
B_0	CRLT_OBS
L_0	CRLN_OBS
SD	RSUN_OBS

A.3 Indexing Correction.

An additional transformation must be applied to the results matrix of integers, due to a difference in indexing. SOHO/MDI FITS files identify the lower left-hand corner as the origin, whereas MATLAB indentifies the upper-left. Therefore, a simple row exchange is necessary; application of the MATLAB function `flipud` corrects for this discrepancy.

A.4 Correction for Instrument Rotation.

The SOHO satellite is not oriented directly with the sun's north pole. Therefore, using keyword information in the FITS header, both the intensity image and magnetogram are rotated using `imrotate` to align the top of the image with the solar north pole. This effectively sets the P-angle referenced in later calculations to zero

(see subsection 3.4.2).

A.5 SDO Median Image.

Images from the following dates were used to generate the SDO median image used in generating the correction image.

Table 19. Dates of SDO observations with no active regions, utilized for the SDO median image in the calculation of the correction matrix for SOHO.

Year	Month	Days
2010	August	21-24
	September	9
	October	6-7
	December	18-24
2011	August	17
2014	July	18

Bibliography

1. USAF, "Air Force Weather Agency Manual 15-1: Space Environmental Observations, Solar Optical Observing Techniques," (2010).

2. McIntosh, P. S., "The Classification of Sunspot Groups," *Solar Physics* **125**(2), 251–267 (1990).

3. Aschwanden, M. J., "Image Processing Techniques and Feature Recognition in Solar Physics," *Solar Physics* **262**(2), 235–275 (2010).

4. Colak, T. and Qahwaji, R., "Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images," *Solar Physics* **248**(2), 277–296 (2008).

5. Jewalikar, V. and Singh, S., "Automated Sunspot Extraction, Analysis and Classification," in *International Conference on Image and Video Processing and Computer Vision*, 151–157 (2010).

6. Zharkov, S., Zharkova, V., Ipson, S., and Benkhalil, A., "Technique for Automated Recognition of Sunspots on Full-Disk Solar Images," *EURASIP Journal on Applied Signal Processing* **15**, 2573–2584 (2005).

7. Benkhalil, A., Zharkova, V., Ipson, S., and Zharkov, S., *Automatic Detection of Active Regions on Solar Images*, 460–466, Knowledge-Based Intelligent Information and Engineering Systems, Springer (2004).

8. Curto, J. J., Blanca, M., and Martinez, E., "Automatic Sunspots Detection on Full-Disk Solar Images using Mathematical Morphology," *Solar Physics* **250**, 411–429 (2008).

9. Gonzalez, R. C., Woods, R. E., and Eddings, S. L., *Digital Image Processing Using Matlab*, Gatesmark, United States of America, 2nd ed. (2009).

10. Spahr, G. M., *Fully Automated Sunspot Detection and Classification using SDO HMI Imagery in MATLAB*, Master's thesis, Air Force Institute of Technology, Graduate School of Engineering and Management (March 2014). AFIT/ENP/14-M-34.

11. Foukal, P. V., *Solar Astrophysics*, Wiley-VCH (2008).

12. Babcock, H. W., "The Topology of the Sun's Magnetic Field and the 22-Year Cycle.," *Astrophysical Journal* **133**, 572–587 (1961).

13. Parker, E. N., "Sunspots and the Physics of Magnetic Flux Tubes. I. The General Nature of the Sunspot," *The Astrophysical Journal* **230**, 905–913 (1979).

14. Wheatland, M. S., "A Bayesian Approach to Solar Flare Prediction," *The Astrophysical Journal* **609**(2), 1134 (2004).

15. Li, R. and Zhu, J., "Solar flare forecasting based on sequential sunspot data," *Research in Astronomy and Astrophysics* **13**(9), 1118–1126 (2013).

16. Scherrer, P. H., Bogart, R. S., Bush, R. I., Hoeksema, J. T., Kosovichev, A. G., Schou, J., Rosenberg, W., Springer, L., Tarbell, T. D., Title, A., Wolfson, C. J., Zayer, I., and MDI Engineering Team, "The Solar Oscillations Investigation - Michelson Doppler Imager," *Solar Physics* **162**(1-2), 129–188 (1995).

17. Berry, R., *Choosing and Using a CCD Camera*, Willmann-Bell, Richmond, VA (1992).

18. Solar Oscillations Investigation Team, "Definitive MDI Full-Disk Magnetic Field Images: Level 1.8 MDI Magnetograms," (2010).

19. Solar Oscillations Investigation Team, "Known Problems in Calibration of MDI Intensity Data," (2000).

20. Bogart, R. S., "Correction of MDI Continuum Intensity for Instrument Degradation," (2000).

21. National Aeronautics and Space Administration, "Events Table Query Results: MDI Exposure Changes," (2009).

22. Potts, H. E. and Diver, D. A., "Post-doc Derivation of SOHO Michelson Doppler Imager flat fields," *Astronomy & Astrophysics* (2008).

23. Joint Science Operations Center, "JSOC Keywords for Metadata," (2013).

24. Thompson, W. T., "Coordinate systems for solar image data," *Astronomy & Astrophysics* **449**, 791–803.

25. National Aeronautics and Space Administration, "Solar Dynamics Observatory Exploring the Sun in High Definition," (2008).

26. Smart, W. M., *Text-Book on Spherical Astonomy*, Cambridge at the University Press (1949).

27. Georgoulis, M. K., "Toward an efficient prediction of solar flares: Which parameters, and how?," *Entropy* **15**, 5022–5052 (2013).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE (DD–MM–YYYY)	2. REPORT TYPE	3. DATES COVERED *(From — To)*
03-26-2015	Master's Thesis	Sep 2013 – Mar 2015

4. TITLE AND SUBTITLE
Automated Sunspot Detection and Classification Using SOHO/MDI Imagery

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Howard, Samantha R, 1st LT

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENP-MS-15-M-078

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally Left Blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A:
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This research modifies and expands previous work by Spahr in 2014 to automatically identify and classify sunspot groups in satellite images. Data from the Solar and Heliospheric Observatory (SOHO) are analyzed to produce a database of sunspot information that is not biased by individual solar observers. Results of the algorithm on SOHO/MDI data correlate well with NOAA's reported data for region properties with R2 values greater than 0.75, but with a ratio of less than one. In particular, the results of analyzing SOHO data report less than 25% of the spots reported by NOAA. By considering a test case comparison with an SDO observation, resolution is likely the main factor in detection discrepancies.

15. SUBJECT TERMS
sunspots,solar activity, MATLAB,image processing, classification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. William F. Bailey, AFIT/ENP
U	U	U	U	90	19b. TELEPHONE NUMBER (Include Area Code) (937) 255-3636, x4555; william.bailey@afit.edu

Standard Form 298 (Rev. 8–98)
Prescribed by ANSI Std. Z39.18