3-2005

# Customization of Discriminant Function Analysis for Prediction of Solar Flares

Evelyn A. Schumer


# CUSTOMIZATION OF DISCRIMINANT FUNCTION ANALYSIS FOR PREDICTION OF SOLAR FLARES

THESIS

Evelyn A. Schumer, Captain, USAF

AFIT/GAP/ENP/05-07

**DEPARTMENT OF THE AIR FORCE**
**AIR UNIVERSITY**

**AIR FORCE INSTITUTE OF TECHNOLOGY**

**Wright-Patterson Air Force Base, Ohio**

AFIT/GAP/ENP/05-07

# CUSTOMIZATION OF DISCRIMINANT FUNCTION ANALYSIS FOR PREDICTION OF SOLAR FLARES

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

Evelyn A. Schumer, BS

Captain, USAF

March 2005

AFIT/GAP/ENP/05-07

# CUSTOMIZATION OF DISCRIMINANT FUNCTION ANALYSIS FOR PREDICTION OF SOLAR FLARES

Evelyn A. Schumer, BS
Captain, USAF

Approved:

//SIGNED//

_____ _____
Devin Della-Rose (Chairman) date

//SIGNED//

_____ _____
K.D. Leka (Member) date

//SIGNED//

_____ _____
Graham Barnes (Member) date

## Abstract

This research is an extension to the research conducted by K. Leka and G. Barnes of the Colorado Research Associates Division, Northwest Research Associates, Inc. in Boulder, Colorado (CORA) in which they found no single photospheric solar parameter they considered could sufficiently identify a flare-producing active region (AR). Their research then explored the possibility a linear combination of parameters used in a multivariable discriminant function (DF) could adequately predict solar activity.

The purpose of this research is to extend the DF research conducted by Leka and Barnes by refining the method of statistical discriminant analysis (DA) with the goal of selecting those photospheric magnetic parameters most capable of identifying flare-producing active regions in hopes of increasing the reliability of short term flare warnings and the understanding of flare production. The data for this research were photospheric vector magnetograms captured by the Imaging Vector Magnetograph (IVM) at the University of Hawai`i Mees Solar Observatory at Haleakala and provided by CORA. Increasing the data set size was an essential task for this research in order to have a more statistically significant training sample for DA. This research also modified current DF procedures to enable the customization of the costs of flare false alarms and flare misses. Work was also done to expand the binary DF results to produce flare probability forecasts. The selection of the optimum combination of photospheric magnetic parameters to be used as predictors in a linear DF began with the elimination of redundant parameters and those parameters least likely to contribute to flare production. The selection of parameters was governed by maximizing the Mahalanobis distance in a step-up method. The DF results show a pre-flaring active region may be characterized by larger magnetic flux, an active region with a larger area of magnetic shear angle greater than 80˚, larger current of heterogeneity, larger spatial vertical magnetic field gradient, and a larger kurtosis of the shear angle.

With the optimum combination of parameters, DF flare probability forecasts were compared to the daily forecasts produced by the National Oceanic and Atmospheric Administration, Space Environment Center (NOAA SEC). The Chi-Squared values of each forecast show the objective DF based flare probability forecasting method performs as well as the subjective forecasting method employed by the SEC.

*To my beloved husband and cherished parents*

*Thank you for your*

*support and encouragement*

**Acknowledgments**

I would sincerely like to thank my faculty advisor, Maj Devin Della-Rose, for his guidance, patience, and constant motivation throughout this process. His energy and enthusiasm were contagious, and his experience and knowledge were greatly appreciated. I would also like to thank the members of my committee, Dr. K.D. Leka and Dr. Graham Barnes, for their expertise and patience and for being so generous with their time. I truly am grateful for the privilege to work with and learn from such highly respected scientists and professionals.

Evelyn A. Schumer

## Table of Contents

Page

## List of Figures

List of Tables

CUSTOMIZATION OF DISCRIMINANT FUNCTION
ANALYSIS FOR PREDICTION OF SOLAR FLARES

# 1. Background

## 1.1 Introduction

With the dawn of modern technology came mankind's introduction to space weather and the ever-changing solar environment, and with each passing year, the variety of technologies affected by the products of solar activity increases. These technologies are critical components in many systems which provide services the majority of us rely on in our daily lives such as telecommunication, commercial airlines, electrical power, wireless services, and terrestrial weather tracking and forecasting. Government agencies and military operations have also seen a dramatic increase in their dependence on space-based systems. These systems are vital to activities such as search and rescue operations, air traffic control, navigation and guidance control, satellite attitude control, and homeland defense. Unfortunately, an increase in solar activity or a solar energetic event, such as a solar flare, can have disastrous effects on these systems and can hinder routine services, governmental procedures, and critical military operations.

New uses for space-based technologies are continually being discovered, and more and more technologies are moving to space-born platforms. In light of the accelerating space-based era, it is more important than ever we understand and can forecast and predict solar flare events whose effects can reach Earth in a matter of

minutes, with little to no warning. Presently, there are agencies which publish daily and short-term flare probability forecasts that rely on subjective visual interpretations of solar active region (AR) magnetic complexity and evolution and McIntosh classification. It is the purpose of this research to explore discriminant analysis as an objective method of flare prediction and forecasting using data derived from an AR's photospheric magnetic field.

## 1.2 Our History with Space Weather

One of the inventions which helped bring about the era of modern technology, the telegraph, is also the technology to introduce mankind to the reality that activity on the sun influences Earth's electromagnetic characteristics [Song, 2001]. As telegraph communication increased in the late 1840's, it often fell victim to anomalous currents that at times would disrupt telegraph communication completely. Also during this time, a solar observer, Richard Carrington, was tracking an exceptionally large sunspot group and extreme auroral displays were widely seen. One of the first to study the anomalous current, W. H. Barlow, wrote in 1849, "in every case which has come under my observation the telegraph needles have been deflected whenever aurora has been visible" [Song, 2001].

The same mechanism that produced the abnormal current in the telegraph system in the 1800's can wreak havoc on today's power, fuel, and telecommunication lines and finds its origin in solar activity. Enormous amounts of solar radiation are absorbed by the Earth's magnetosphere and ionosphere, greatly increasing near-Earth

current systems. The enhanced, highly dynamic magnetospheric and ionospheric current systems lead to large variations in the time rate of change of the magnetic field at the Earth's surface, inducing potential differences across large areas of the surface. Earth-bound power, fuel, and telecommunication lines grounded to the Earth provide an excellent path for current to flow between the induced potential differences. Such a sudden flow of current through lines can be ruinous to power supplies and communications.

A magnetic storm in February 1958 was responsible not only for disrupting voice communications over the first cross-Atlantic telecommunications cable, but was also responsible for rendering Toronto's power systems temporarily unavailable. In 1972 a magnetic storm resulted in an hour-long outage of a major continental telecommunications cable from Chicago to the west coast. The entire province of Quebec fell victim to a magnetic storm in March 1989 when a major transformer failed due to induced surface potentials and power was unavailable for an entire day. This same storm nearly destroyed the first trans-Atlantic fiber voice cable when potential differences were established between cable terminals in New Jersey and England [Song, 2001].

In May 1998 we witnessed how widespread the impact of solar activity can be on our technologies. The previous examples of the crippling effects of solar activity were regional or localized; however, an epoch of solar flares in 1998 greatly disrupted the space around Earth and affected the entire North American continent. The electromagnetic energy and high-energy particles spewed towards Earth by the solar flares rendered a communications satellite inoperable. The Galaxy IV satellite

provided over 90 percent of North America's paging service and relayed radio and television signals [Carlowicz, 2002]. During that time doctors and nurses across the nation could not receive their pages and Americans were forced into a world before television and radio.

Another area vulnerable to solar activity with extensive, and even global, reach is the airline industry. A solar event can adversely affect the entire spectrum of air travel technologies. The surge of solar activity in March and April 2001 caused disruption and blackouts of radio signals and led to more than 25 flights being diverted in order to avoid flying through polar regions where communication was nearly impossible and passengers could have been exposed to as much solar radiation as that of 100 chest x-rays. Radio signals used to identify aircraft were disrupted, and navigation via the Global Positioning System (GPS) was not reliable. Consequently, planes were incapable of landing in low visibility conditions [Carlowicz, 2002].

The military, police, and fire emergency agencies rely heavily on high frequency (HF) and transionospheric wireless communication. HF technology uses the ionosphere to reflect radio signals. However, solar activity can alter ionospheric reflective properties and, thus, alter the propagation of wireless signals. In 1979, at the peak of the 21$^{st}$ solar cycle, the Orange County, California fire department experienced such an effect. The department received a distress signal from a downed commuter plane and responded only to determine later the signal had originated in West Virginia [Song, 2001]. The military has also come to rely on wireless communication during operations for real-time decisions and intelligence.

Consequently, lives, outcomes of battles, and our national defense depend on the ability to know when HF communications are available.

Many of the events accompanying and products of solar flares, such as radiowaves, X-rays, and relativistic protons, travel at or close to the speed of light. Consequently, by the time such radiation is observed, their effects are already being felt near or on Earth, and the method of now-casting is inadequate for issuing the desired several-hours warning of the arrival of such radiation. With the diversity and numbers of users of space-based technologies that are susceptible to the effects of solar activity growing and with the speed at which some products of solar flares reach the Earth, the need for a robust method to predict solar flare events is evident. An objective and reliable flare prediction system is crucial for protecting satellites and astronauts from excess radiation, space-born hardware from shortened lifetimes, space-born communication and navigation systems from failure, and warning of possible high frequency radio transmission blackouts.

## 1.3 Active Region Evolution

A solar flare, a type of solar event that can greatly affect space weather, is a localized explosive release of energy from the sun's atmosphere in the form of electromagnetic radiation and energetic particles. It is generally accepted the energy to produce solar flares is the stored magnetic energy of active regions (ARs), areas where the solar magnetic field departs from the simple dipole model. In these regions field lines are concentrated, and the situation is highly unstable. If a "trigger" occurs

to “tap” this stored energy via relaxation of the magnetic field lines, a solar flare results.

Active regions are locations of intense concentrations of magnetic flux. Present theory on AR evolution assumes the undisturbed and initial solar global magnetic field is a weak poloidal field. The theory also assumes the solar plasma conductivities are large and solar magnetic field lines are frozen-in, meaning the solar plasma and field lines move together with the same velocity. Furthermore, the sun, as a gaseous body, experiences differential rotation. Plasma at the solar equator rotates with a period of roughly 25 days per revolution whereas plasma at the poles takes nearly 32 days to complete a rotation. Due to the frozen-in nature of the solar magnetic field lines, this differential rotation distorts the initial poloidal field and, over time, transforms the initial magnetic field into a toroidal field (see Figure 1.1).

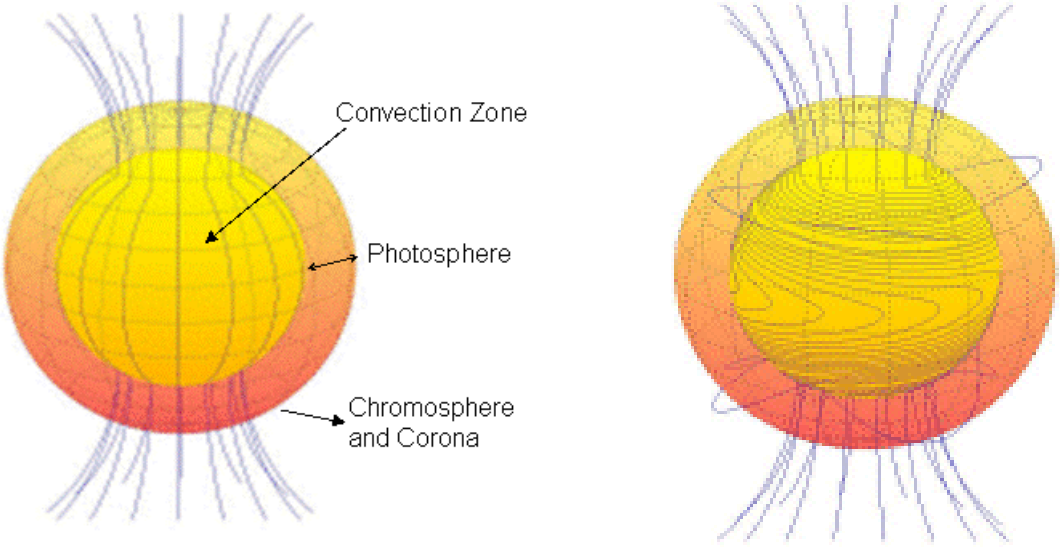


Figure 1.1 - Solar Magnetic Field Evolution from a Poloidal Field to a Toroidal Field, [http://solar.physics.montana.edu/hurlburt/]

As the differential rotation continues, the magnetic field lines become more twisted and coiled and magnetic flux tubes become buoyant and emerge above the photospheric surface (see Figure 1.2). Convective forces below the photosphere continue to act on the plasma and add further twist to the emerging flux tubes. The areas where the flux tubes break the photospheric surface correspond to observed sunspots within ARs.

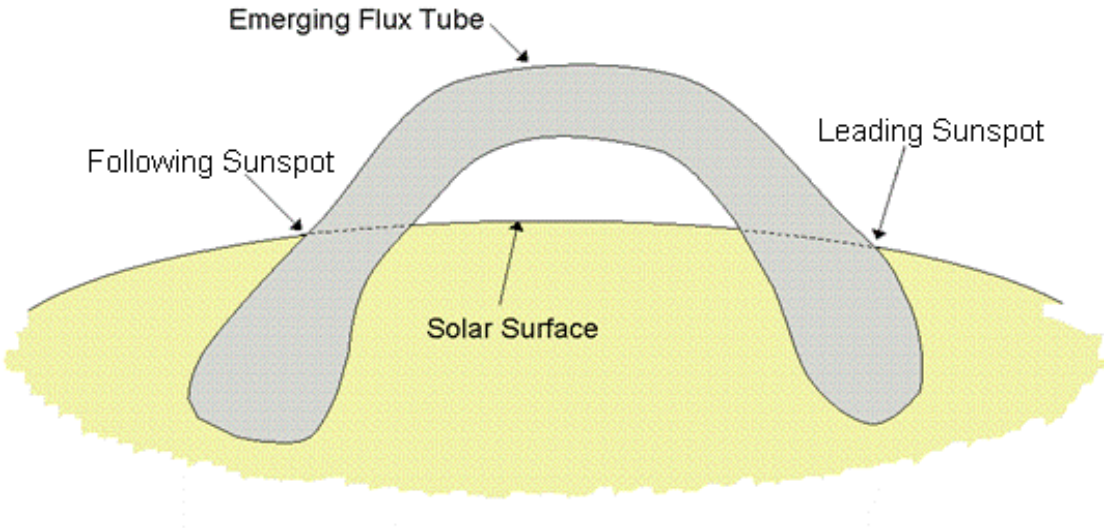


Figure 1.2 - Emerging Flux Tube and Sunspot Group within an AR

## 1.4 Present Solar Flare Theory

As plasma flow and convection continue to twist and stretch solar magnetic field lines, excess energy is built up in the field. Thus, the magnetic field deviates further from a potential field, which is the field configuration of lowest energy. The field lines become more and more stressed as the twisting and stretching continues. The

field lines can only withstand a finite amount of pressure and tension, and if field lines cross, a threshold is reached, or an instability arises, a solar flare may result.

A solar flare can result due to the relaxation of a complex, non-potential magnetic field. Relaxation of stressed magnetic field lines can explosively release enormous amounts of stored energy. During magnetic relaxation, stored magnetic energy is converted to particle kinetic energy, thermal energy, and electromagnetic energy of a flare. A type of relaxation that often occurs near AR neutral lines is reconnection and is shown in Figure 1.3. Neutral lines separate vertical magnetic field lines of opposite polarity in an AR and are areas where flares are frequently observed. Although reconnection occurs in the chromosphere, it is thought the stress mechanism of solar flares takes place in the photospheric magnetic field.

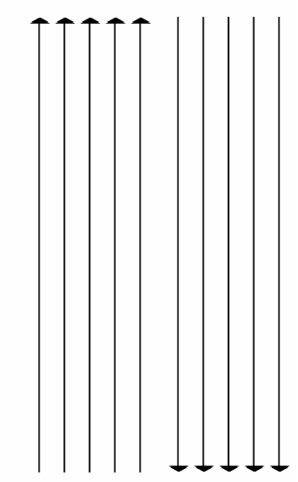

Figure 1.3a – Magnetic field lines of opposite polarity prior to reconnection

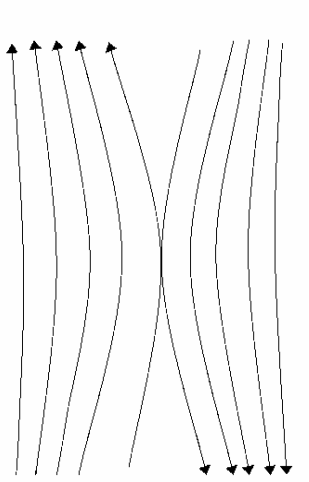

Figure 1.3b – Magnetic field lines of opposite polarity come into contact

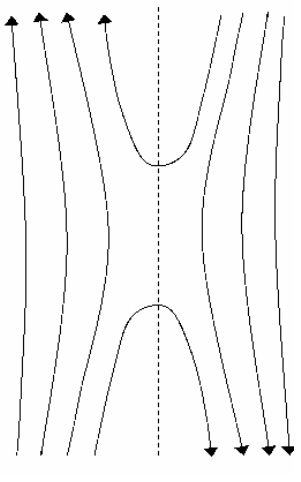

Figure 1.3c – Reconnection and relaxation of magnetic field lines, the dash line in the center represents zero magnetic field where two lines of opposing polarities came into contact and cancelled one another

The impulsive energy release of a solar flare can be close to $10^{25}$ Joules [Tascione, 1994]. Solar flares can last from a few minutes to a few hours, and the output radiation covers the spectrum from radio waves to gamma-rays.

## 1.5 Classification of Solar Flares

Solar flares can be classified according to area or size, their intensity in the visible wavelengths, and total X-ray emission. These characteristics are good indicators of the amount of energy released in the form of electromagnetic radiation and particle emissions [55th SWXS, 1997].

Importance is a measure of an optical flare's area or size at the time of maximum intensity in Hα. A unit often used to describe flare area is one millionth of the visible solar hemisphere, which is approximately equal to 3 million square kilometers. Another unit used is hemispheric square degree which is roughly equivalent to 48.5 hemispheric millionths. Table 1.1 summarizes Importance classification.

**Table 1.1 - Solar Flare Importance Classifications**

| | Flare Area | |
|---|---|---|
| **Importance Designator** | **Hemispheric Square Degrees** | **Millionths of Hemisphere** |
| 0 | 0 – 2.0 | 10 – 99 |
| 1 | 2.1 – 5.1 | 100 – 249 |
| 2 | 5.2 – 12.4 | 250 – 599 |
| 3 | 12.5 – 24.7 | 600 – 1199 |
| 4 | ≥ 24.8 | ≥ 1200 |

Another classification generally appended to the Importance numeral designator is Brightness (see Table 1.2). An optical flare's Brightness is a quantitative term describing the intensity of the flare at $\pm 0.4$Å, $\pm 0.6$Å, and $\pm 1.0$Å off the Hα line center as compared to background intensity. If the area does not brighten to at least 150% of the background, it is only considered to be a plage fluctuation [55th SWXS, 1997].

**Table 1.2 - Optical Flare Brightness Classifications**

| Brightness Designator | Brightness (% of Background) |
|---|---|
| F (Faint) | 150% - 259% |
| N (Normal) | 260% - 359% |
| B (Brilliant) | $\geq$ 360% |

Flares are also classified according to their peak X-ray flux within the 1-8Å band, as measured by Geostationary Operational Environmental Satellites (GOES). Sensors aboard GOES measure solar x-ray flux. See Figure 1.4 for an example of GOES data for M- and X-class flares on 12 July and 14 July 2000.

**Table 1.3 - X-ray Flare Classifications**

| Class | X-Ray Flux (watt/m$^2$) |
|---|---|
| C | $10^{-6} - 9\times10^{-6}$ |
| M | $10^{-5} - 9\times10^{-5}$ |
| X | $\geq 10^{-4}$ |

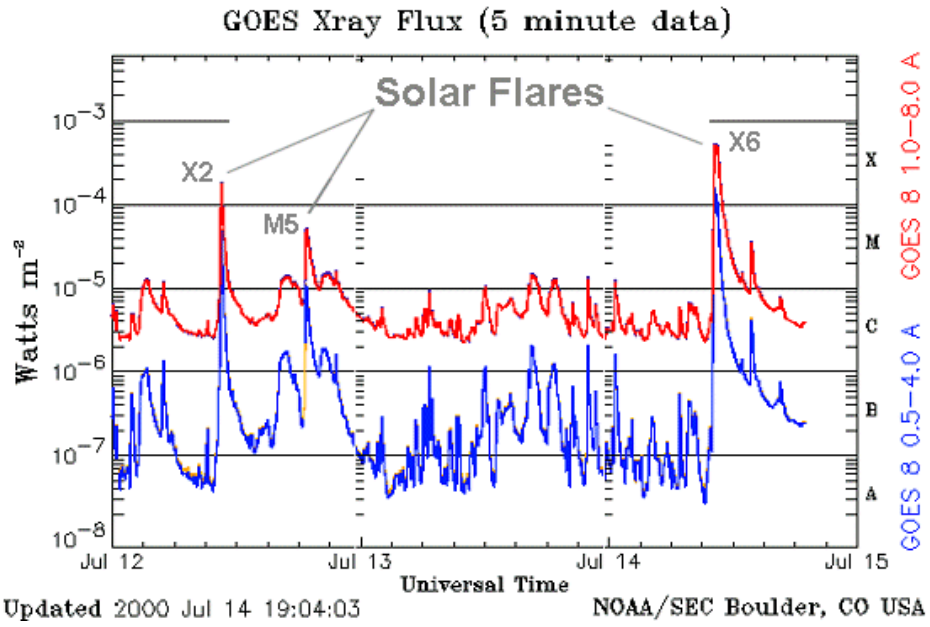


Figure 1.4 - GOES X-Ray Flux Data

## 1.6 McIntosh Classification Scheme

Sunspot classification schemes were developed in an attempt to identify those regions likely to produce flares. The original Zurich classification scheme categorized spot groups into nine classes based on visual characteristics and was developed by M. Waldmeier in 1938. In 1966, Patrick McIntosh built upon the Zurich Scheme and developed the McIntosh classification system which is used today. The McIntosh system assigns to an AR a three-letter designator. The first letter of the classification describes the group type or the unipolar and bipolar nature of the spot group. The second letter describes the penumbra of the largest spot in the group, and the third describes the compactness of the spots in the intermediate part of

the group. See Figures 1.5 and 1.6 and Tables 1.4 and 1.5 for examples and descriptions of the McIntosh classes.

**Table 1.4 - The First Letter of the McIntosh Classification Scheme**

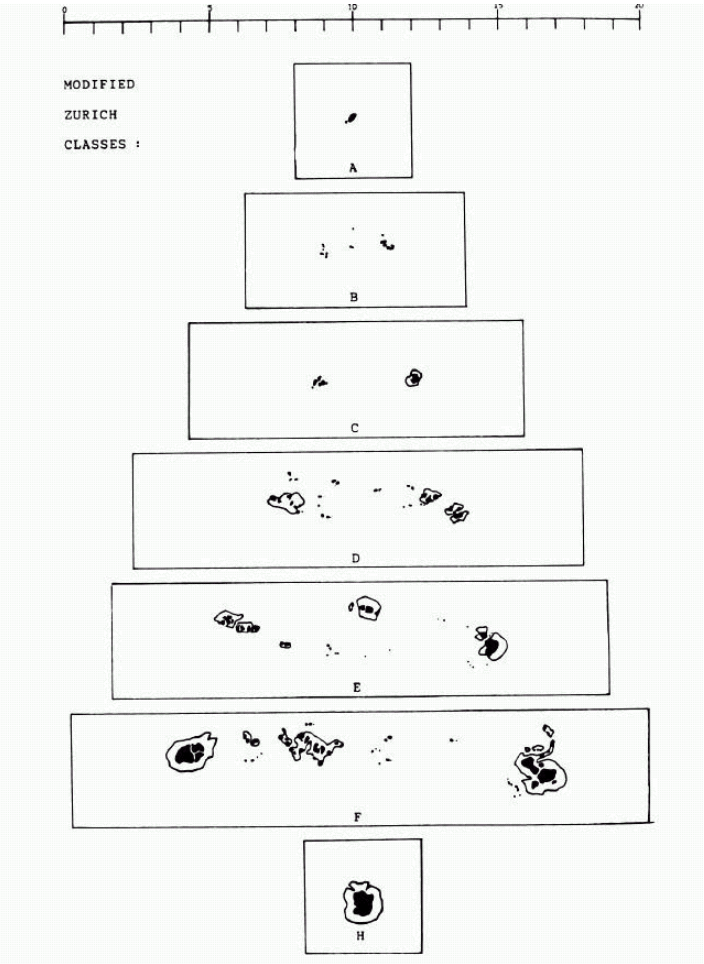


Figure 1.5 - Modified Zurich Classes

**Modified Zurich Classes**

**A** – a unipolar group with no penumbra

**B** – a bipolar group with no spots having a penumbra

**C** – a bipolar group with penumbra on one end of the group

**D** – a bipolar group of less than 10 degrees in length with penumbrae on spots at both ends of the group

**E** – a bipolar group of length 10-15 degrees with penumbrae on spots at both ends of the group

**F** – a bipolar group of greater than 15 degrees with penumbrae on spots at both ends of the group

**H** – a unipolar group with penumbra

* The First Letter of the McIntosh Classification Scheme is also the Modified Zurich Classes

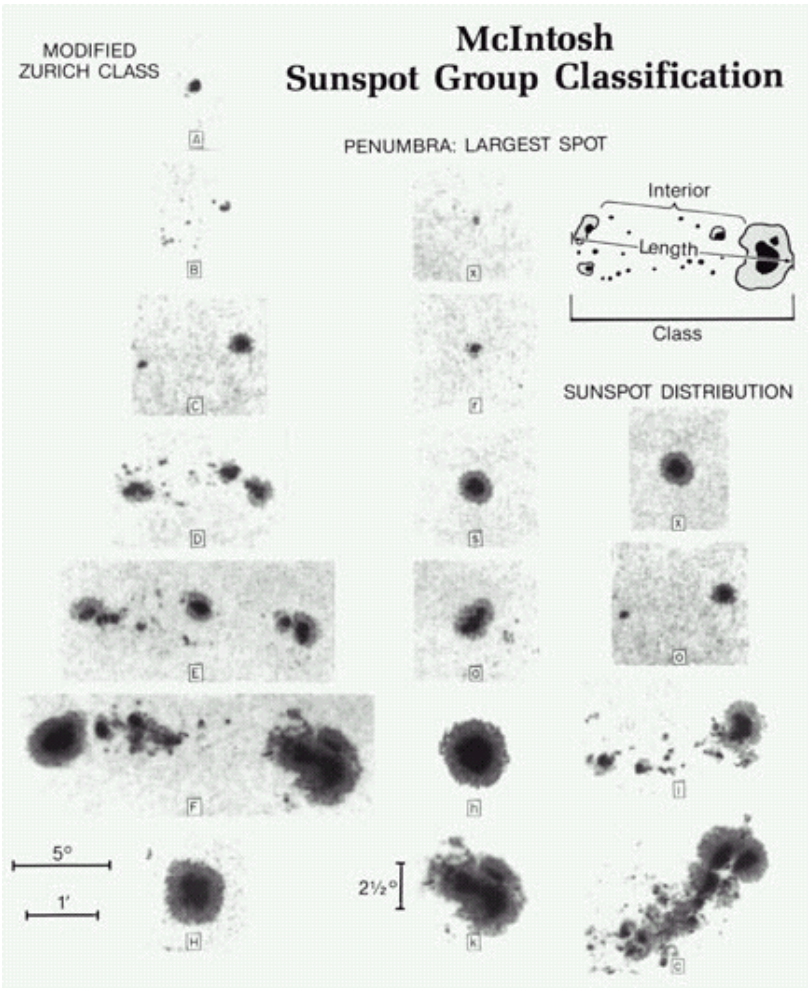

Figure 1.6 - McIntosh Sunspot Group Classification

**Table 1.5 – Second and Third Classes within McIntosh Classification Scheme**

| **Second Letter** | **Third Letter** |
|---|---|
| x – no penumbra, only used for classes A and B | **x** – single spot, unipolar group of Modified Zurich Classes A or H |
| **r** – rudimentary penumbra | **o** – open distribution with few, if any, small spots between the leading and following spots |
| **s** – small and symmetric spot | **i** – intermediate distribution with numerous umbral spots between the leading and following spots |
| **a** – small and asymmetric spot of diameter 2.5 degrees or less with irregular or separated penumbra | **c** – compact distribution with many spots between the leading and following spots, at least one of the intergroup spots has penumbra |
| **h** – large symmetric spot of diameter greater than 2.5 degrees | |
| **k** – large asymmetric spot of diameter greater than 2.5 degrees | |

## 1.7 SEC's Flare Forecasting Method

The National Oceanic and Atmospheric Administration Space Environment Center (NOAA SEC) produces daily short-term flare probability forecasts based on the McIntosh Classification Scheme for users in government, industry, and the private sector [Gallagher, 2005]. The SEC assigns separate probability forecasts for M- and X-class flares and for the time intervals of 24, 48, and 72 hours. The forecasts are computed based on many factors including an active region's McIntosh Classification, the region's previous activity, and its present evolution.

At the SEC, publishing a flare probability forecast begins with comparing the most complex and largest ARs present on the solar disk to previous active regions of the same McIntosh Classification. The SEC database of past active regions to which the current ARs are compared spans the dates of November 1988 to June 1996. An initial flare probability forecast is obtained based on the fraction of similarly classified ARs in the database that produced at least one flare. The probability equation used in SEC's flare prediction method is governed by previous studies of flare production rates and Poisson statistics [Gallagher, 2005].

Poisson statistics can be applied to counting experiments in which independent, random events are observed at a definite average rate [Taylor, 1997]. Previous studies have shown the nature of flare occurrences and flare rate distributions with respect to a given peak flux value can be modeled according to Poisson statistics. SEC's forecasting method relies on previous research that suggests the rate of flare production on the solar disk varies with time and the rates of flaring can be modeled as a piecewise Poisson process [Wheatland, 2001]. SEC's forecasting method also

relies on previous work done by Hudson [1991] which suggests the distribution of flare production rate versus peak flux value for the solar disk obeys the power law with an index $\xi \approx 1.8$. According to Wheatland [2001], the distribution of flare production rate versus peak flux given by

$$N(\Phi) = \lambda_o \, \Phi_o^{(\xi-1)} \, \Phi^{-\xi} \, (\xi - 1) \qquad (1.1).$$

With the known observation of $N_o$ flares with a peak flux of at least $\Phi_o$, $\lambda_o$ is the flare production rate equal to ($N_o/t$), where $t$ is the period of observation [Wheatland, 2001].

Wheatland [2001] then suggests that the current rate of flaring can be determined from the time history of observed flare production. From equation 1.1, if the current rate of flare production $\lambda_o$ above a threshold peak flux of $\Phi_o$ is determined, then the rate of flaring $\lambda_1$ above the peak flux of $\Phi_1$ is given by

$$\lambda_1 = \lambda_o \left( \frac{\Phi_1}{\Phi_o} \right)^{-\xi+1} \qquad (1.2)$$

[Wheatland, 2001]. Assuming flare production can be modeled as a piecewise Poisson process and the most current flare production rate $\lambda_x$ can be determined from the recent history flare observations, the probability of observing at least one flare with a peak flux greater than $\Phi_x$ within the time interval $\Delta t$ is given by

$$P_x(\Delta t) = 1 - e^{-\lambda_x \Delta t} \qquad (1.3)$$

[Wheatland, 2001]. It is from equation 1.3 that SEC derives the probability equation used in its flare forecasting method [Gallagher, 2004].

SEC's forecasting method assumes flare events are independent and obey the equations above. For example, let us assume an AR of McIntosh Classification, Eai, is observed and is the largest and most complex AR on the solar disk. According to the SEC database, 302 ARs between the dates of November 1988 and June 1996 were of the same class, Eai, and produced a total of 62 M-class events. Thus, the fraction of Eai regions that produce M-class flares is 62/302 or 0.205, and the corresponding flare production rate is $\lambda$=0.205 flares per observed Eai region. The probability the Eai region will produce at least one flare in the following 24 hours is $P = 1 - Exp[-0.205]$ or $P = 0.19$, where $\Delta t = 1\,\text{day}$ [Gallagher, 2005].

The SEC forecaster would then further refine the initial quantitative 19% flare probability forecast according to his/her previous experiences and visual interpretations of the structure and status of the AR. Taken into account would be the region's current evolution and past history of producing flares [Wheatland, 2004].

As a way to follow how well the forecasts compare to observed flare activity, SEC maintains a flare forecast verification plot for both M- and X-class flares (Figure 1.7).

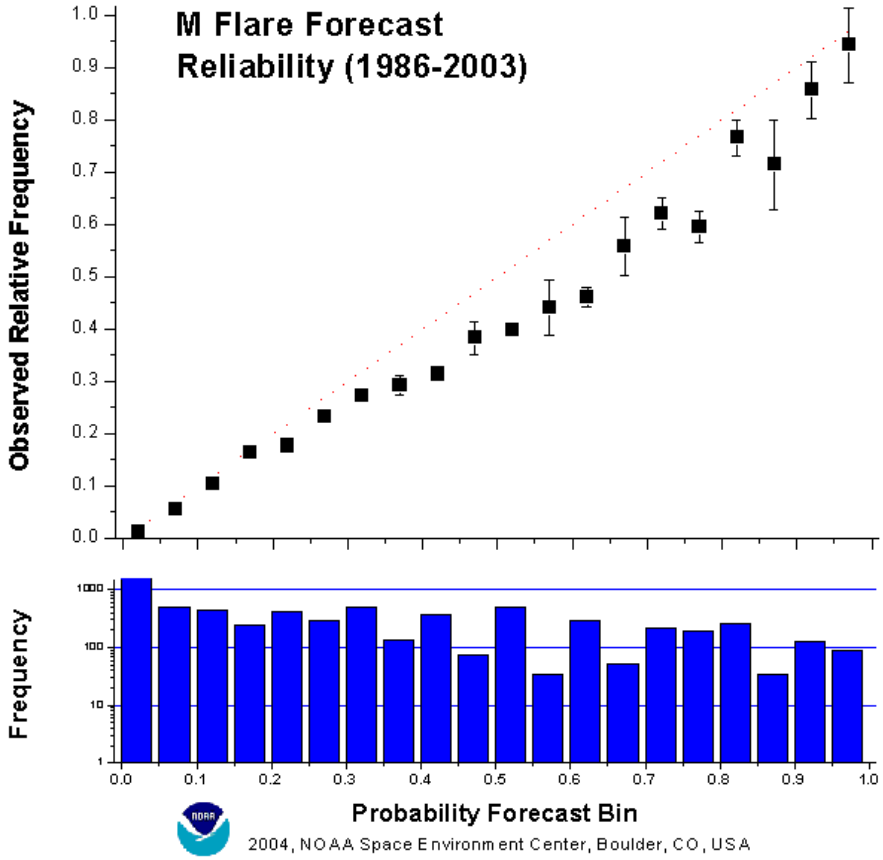


Figure 1.7 – SEC Flare Probability Forecast Verification Plot

SEC's verification plot in Figure 1.7 shows the relative frequency of days on which at least one M-class flare occurred with respect to forecasted flare probability. The 45˚ dashed line marks 100% forecast accuracy. SEC uses a modified version of standard error to determine the error associated with the observed relative frequency in its verification plot. From Figure 1.7, it appears that SEC's forecasting method is accurate for forecasts between 0% and 20%; however, SEC tends to over-forecast for probabilities larger than 20%. In other words, for flare probability forecasts greater than 20%, there are fewer days observed when M-class flares occur than SEC's forecasts would suggest.

## 1.8 The Chi-Squared Value

### 1.8.1. General $\chi^2$ Value

A means to quantify the accuracy of a forecast method is to measure the deviation of its forecast verification plot from the 45˚ line marking perfect accuracy. Calculating the chi-squared ($\chi^2$) value is one method to quantify this deviation and the statistical value we used in this research to quantify the accuracy of flare forecasting methods. In general, the $\chi^2$ value is used to determine whether or not an observed distribution of measurements is consistent with the expected theoretical distribution and is used to quantify the extent observed values deviate from the expected values.

For a general discussion of the $\chi^2$ value, suppose we have an experiment in which we measure *Y*, a continuous quantity, *N* number of times, giving us the measurements $y_1, y_2, \ldots y_N$ . Furthermore, we have reason to believe the distribution of our measurements is governed by the Gaussian distribution. We want to determine whether our hypothesis of a Gaussian distribution is valid for the actual distribution of our measurements. We begin be calculating the mean and standard deviation of our measured values [Taylor, 1997].

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} \tag{1.4}$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \bar{y})^2}{N-1}} \tag{1.5}$$

The next step is to compute the expected distribution of our $N$ measurements if our hypothesis is true. We must keep in mind $Y$ is a continuous quantity and does not take on discrete values; thus, we cannot speak of expected values of $y$ equal to any one value. Instead, we must consider how many measurements we expect to be within a given interval $a < y < b$. To do this, we divide the range of possible values into *bins* such that all bins have a least several measurements. Given the number of bins equals $\eta$ , we calculate the expected number of measurements, $E_k$ , that would fall into each bin $k$ if our hypothesis is true. For a Gaussian distribution, $E_k$ is found using the results of equations 1.4 and 1.5. Then we count how the number of actual measurements, $O_k$ , we observe within each bin [Taylor, 1997].

If our hypothesis is true, we would expect the deviations, $(O_k - E_k)$, to be small. Conversely, if our hypothesis was invalid, we would expect the deviations to be large. To quantify large and small deviations, we first calculate the expected outcome of our measurements if we were to repeat the experiment many times. The expected results for many different $O_k$ should have an average value of $E_k$ and a standard deviation of $\sqrt{E_k}$ . Thus, by considering the value of the ratio,

$$\frac{O_k - E_k}{\sqrt{E_k}} \tag{1.6}$$

we are able to quantify large and small deviations from the expected distribution. If our hypothesis is valid, the ratio for most bins should be of order one or smaller. We then square the ratio to avoid negative values and sum over all bins to calculate the $\chi^2$ value.

$$\chi^2 = \sum_{k=1}^{\eta} \frac{(O_k - E_k)^2}{E_k} \tag{1.7}$$

If the $\chi^2$ value equals zero, the observed values are in perfect agreement with the expected values. In general, if the hypothesis of the distribution is valid and the individual terms in equation 1.7 are of order one or smaller, $\chi^2$ will be of order $\eta$ or smaller. However, if $\chi^2 >> \eta$ , then we have good reason to suspect our measurements are not governed by the expected distribution [Taylor, 1997].

### 1.8.2. $\chi^2$ Value Applied to Flare Probability Forecasts

To calculate a modified $\chi^2$ value for a set of flare probability forecasts, such as those produced by SEC in Figure 1.7, the distribution of possible probabilities (0% to 100%) is broken up into $\eta$ number of probability bins or ranges, so that each bin contains at least several datapoints. The number of daily forecasts assigned a flare probability corresponding to each bin $k$ is $W_k$. $O_k$ is the percentage of flare-active days in bin $k$. The expected percentage of flare-active days assigned to each bin $k$ is equal to the bin midpoint and denoted by $E_k$.

Furthermore, the square root of the number of daily forecasts assigned to an individual bin, $\sqrt{W_k}$ , is used as a weighting factor. $\sqrt{W_k}$ was chosen as a weighting factor in order to weight those bins with potentially smaller errors more heavily. We defined the error for the values of the observed relative frequency of flares as

$$\delta_k = \frac{1}{\sqrt{W_k}} \ , \tag{1.8}$$

and our definition of error is modeled after the equation used to determine error for SEC's forecast verification plot. SEC modeled its error definition after that for standard error [Doggett, 2004]. For $N$ measurements of the same quantity $x$, with a standard deviation equal to $\sigma_x$ , standard error is given by

$$\delta_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} \ . \tag{1.9}$$

Since flare forecast verification is concerned with counting the relative occurrences of flare-producing and flare-quiet ARs, a modified definition of the standard error which neglects the standard deviation of measurements had to be used.

If a method of producing flare probability forecasts was not valid, we would expect the deviation of $O_k$ from $E_k$ and the total $\chi^2$ value to be large. However, if a method for producing flare forecasts is good, we would expect the total $\chi^2$ value to be small.

$$\chi^2 = \sum_{k=1}^{\eta} \left( (O_k - E_k)^2 \cdot \sqrt{W_k} \right) \tag{1.10}$$

For an example, see Figure 1.7 and notice the $k$=16 bin ranging from flare forecasts of 75% to 80%. The relative frequency of flare-active days given daily forecasts assigned a flare probability between 75% and 80% is $O_{16} \approx 0.60$. If close to 77.5% of the 200 days assigned a flare probability between 75% and 80% were flare-active, then there would be excellent forecast accuracy and a $\chi^2$ value close to zero for bin-16. However, we see for bin-16 the forecasts overestimated flare production, and actual flare production was around 60%. Consequently, the $\chi^2$ value for bin-16 is

$$\chi^2_{\kappa=16} = (0.60 - 0.775)^2 \cdot \sqrt{200} = 0.433 \tag{1.11}.$$

If equation 1.7 was solved for bin-16, we would see the resulting term to be less than one, and we would have no reason to doubt the validity of the hypothesis. Thus, we conclude the $\chi^2$ value for equation 1.11 of 0.433 is relatively small and does not give us reason to question the hypothesis.

## 1.9 Zeeman Effect and Vector Magnetographs

The McIntosh Classification Scheme and SEC's flare probability forecasting method rely on subjective interpretations of an AR's visual characteristics. However, the purpose of this research and of the previous research discussed below in §1.11 was to explore a statistical method for producing objective flare forecasts and predictions. From SEC's forecast verification plot in Figure 1.7, we see flare forecasts published by SEC based on an AR's visual characteristics compare relatively well with observed flare production. Given that the visual characteristics of an AR are governed by the state and evolution of the local solar magnetic field, information derived from magnetic field parameters may provide an objective means of predicting and forecasting solar flare activity. The solar layer for which past and present magnetic field data is available is the photosphere. Little to no magnetic data is currently available for other solar layers that may or may not provide better indicators of flare activity, such as the chromosphere and corona. The photosphere is also the solar layer in which ARs are observed in white light. Thus, it is the photospheric magnetic field we are concerned with for this research.

The instrument used to measure the direction and strength of the photospheric magnetic field is a vector magnetograph. This instrument relies on the Zeeman splitting of Fraunhofer lines and the polarization properties of sunlight to determine the magnitude of both the longitudinal (line-of-sight) and transverse magnetic field components [Phillips, 1995]. The most widely used spectra for photospheric magnetograms is light from iron of wavelength 5250Å.

The Zeeman effect is named after the Dutch physicist, Pieter Zeeman, who in August 1886, observed the spectral lines from a sodium flame were broadened and even split into two and three lines when the flame was placed between magnets. He further noted the amount of splitting is linearly proportional to the strength of the magnetic field through which the light passes. Zeeman determined the relationship between the magnitude of the external magnetic field and the wavelengths of the components of the split spectral line is

$$\Delta\lambda = \frac{eB\lambda^2}{4\pi m_e c} \ . \tag{1.12}$$

Here $B$ is the magnitude of the external magnetic field in units of Gauss, and $\lambda$ is the wavelength of the zero magnetic field spectral line [Radel and Navidi, 1994].

The Zeeman effect is due to the interaction between an external magnetic field and the magnetic dipole moment associated with the electron's orbital angular momentum. An in-depth discussion of the Zeeman effect is beyond the scope of this paper. We will, however, briefly discuss the Zeeman effect and apply it to the simple model of the hydrogen atom.

Due to the quantanization of energy, an atom can only absorb discrete amounts of energy corresponding to the allowed energy levels of its electron orbits. Thus, as light passes through a cool gaseous material capable of absorbing radiation of wavelengths, $\lambda$, dark bands or absorption lines, also known as Fraunhofer lines, will appear in the light's spectra at the given wavelengths, $\lambda$. See Figure 1.8.

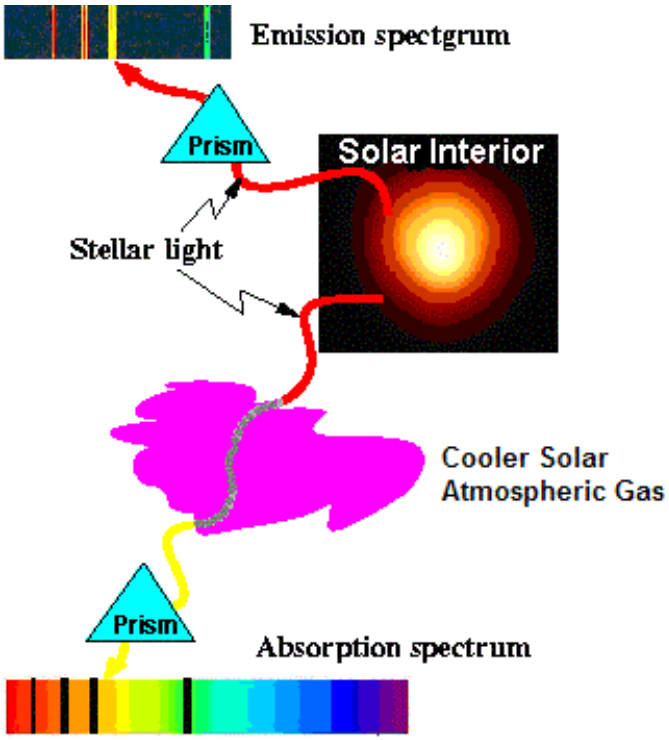


Figure 1.8 - Emission and Absorption Spectra
[http://phyun5.ucr.edu/~wudka/Physics7/Notes_www/node107.html]

In the absence of perturbing factors, such as an external magnetic field, many of the quantum states of an atom can have identical energies and are referred to as degenerate energy levels. However, the presence of a strong magnetic field can breakdown the degeneracy of atomic energy levels.

For example, the principle quantum number, *n*, determines the energy of allowed states of the hydrogen atom. However, an electron may move in a number of orbits within the hydrogen atom for a given allowed energy. The orbits are designated by the orbital quantum number, $\ell$, and can take on the values of $0\ldots(n\text{-}1)$.

The magnitude of an orbit's magnetic dipole moment is proportional to its angular momentum, **L**, and is given by

$$\mu = \frac{-e}{2m_e} L \quad , \tag{1.13}$$

Since we have no loss of generality, we can align the z-axis of our coordinate system with the external magnetic field, and now $L = m_\ell \hbar$ and $\mu = \frac{-e}{2m_e} m_\ell \hbar$. $m_\ell$ is the orbital magnetic quantum number and can take on values from $-\ell$ to $+\ell$. An external magnetic field will exert a torque on the magnetic dipole, and the resulting magnetic potential energy associated with a magnetic dipole subject to an external magnetic field in the $\hat{z}$ directions is

$$\begin{aligned} U &= -\boldsymbol{\mu} \cdot \mathbf{B}_z \\ U &= \frac{e}{2m_e} L_z B_z = \frac{e\hbar}{2m_e} m_\ell B_z \end{aligned} \tag{1.14}$$

[Ohanian, 1995].

For a given quantum state in our example of the hydrogen atom, if the magnetic dipole moment is positive (parallel to the external magnetic field), a previously degenerate state is now available at an energy of

$$E_{\sigma+} = E_\lambda - U \tag{1.15}$$

where $E_\lambda$ is the energy of the zero external magnetic field spectral line. See Figure 1.9. The atom can now absorb a photon of wavelength

$$\sigma_+ = \lambda + \Delta\lambda \tag{1.16}$$

The orbital energy and wavelength available for absorption for the case of a negative magnetic dipole (oriented anti-parallel to the external magnetic field) is equal to

$$E_{\sigma-} = E_\lambda + U$$

$$\sigma_- = \lambda - \Delta\lambda \quad . \tag{1.17}$$

Thus, for the example of a transition between quantum states within the hydrogen atom from quantum state 2p ($n$=2 & $\ell$=1) to quantum state 1s ($n$=1 & $\ell$=0), Zeeman splitting of degenerate quantum states, $m_\ell$ = -1, 0, +1 , results in a triplet of energy levels and spectral lines.

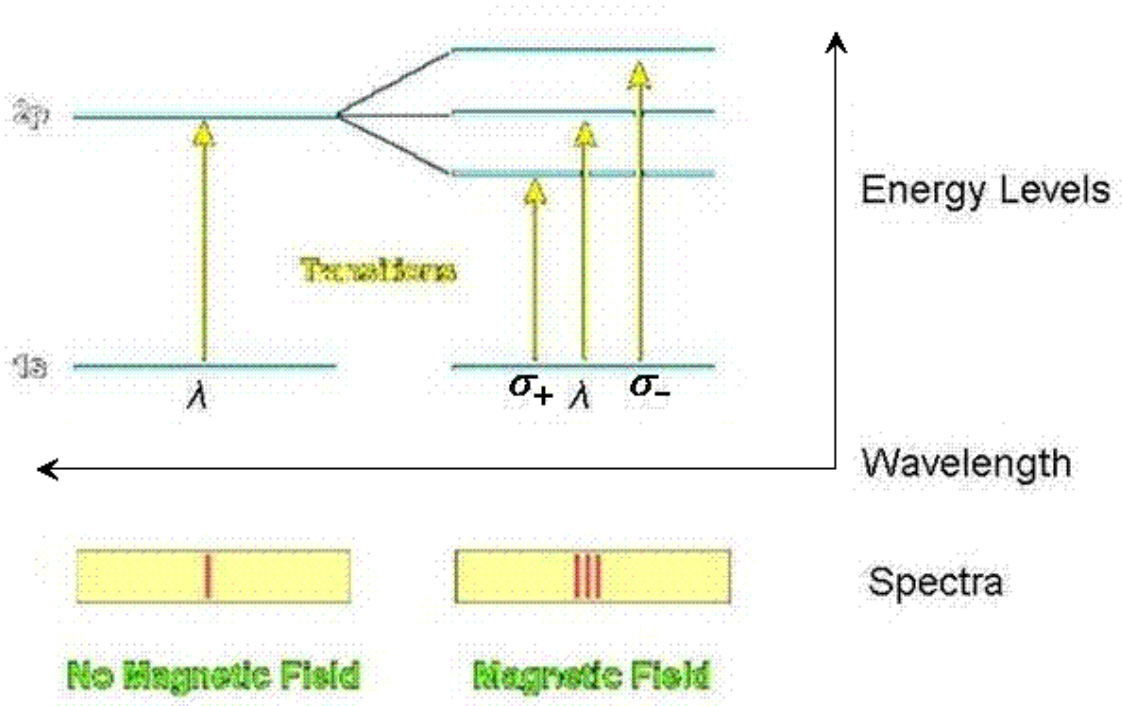


Figure 1.9 - Zeeman Splitting of Spectral Lines in Hydrogen Atom

In 1908 George Ellery Hale linked the Zeeman effect to solar spectra of sunspots. He observed no splitting of Fraunhofer lines and only broadening of the lines occurred for the spectra from solar regions void of sunspots. However, when the spectrograph slit admitted light from a region that included a sunspot, Zeeman splitting was observed. Further investigation by Hale led to the discovery the emissions from spectral lines created by Zeeman splitting were polarized.

The polarized components of sunlight yield information about the transverse and longitudinal components of the solar magnetic field while the amount of splitting observed in the spectra of sunlight is proportional to the strength of the magnetic field. The transverse solar magnetic field results in the linearly polarized component of sunlight, and the longitudinal component of the solar magnetic field is responsible

for the circularly polarized components. The circularly polarized light can then be broken down into two plane polarizations at 90 degrees to each other. The two plane polarized components can then be analyzed with filters to produce images of each direction of the circularly polarized components. The difference of the images yields the longitudinal component of the solar magnetic field [Phillips, 1995]. See Figure 1.10 for an example of a photospheric vector magnetogram.

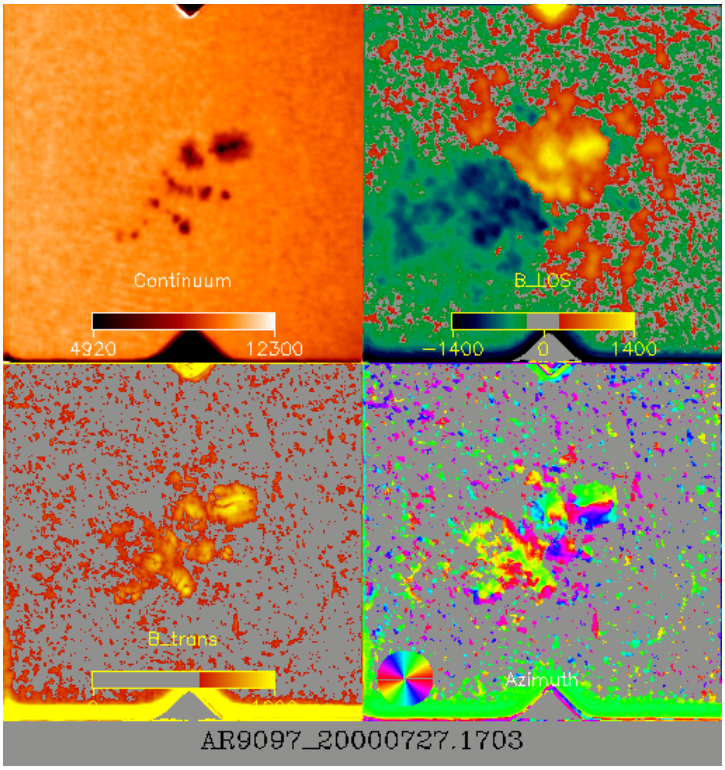


Figure 1.10 – Photospheric Vector Magnetogram,
(http://www.solar.ifa.hawaii.edu/IVM/Archive)

It is from the spectropolarimetric raw images of an active region that the Stokes polarization vector, [I, Q, U, V], is derived. See Table 1.6 for the components of the Stokes vector [Rees, 2001].

**Table 1.6 - Stokes Vector Components**

| | |
|---|---|
| [1, 0, 0, 0] | Random Polarization |
| [1, 1, 0, 0] | x-Linearly Polarized |
| [1,-1, 0, 0] | y-Linearly Polarized |
| [1, 0, 1, 0] | +45° Linearly Polarized |
| [1 0,-1, 0] | -45° Linearly Polarized |
| [1, 0, 0, 1] | Right-hand Circularly Polarized |
| [1, 0, 0,-1] | Left-hand Circularly Polarized |

The amplitudes of the components of the Stokes vector are calculated for each pixel in the image [Leka and Barnes, 2003a]. However, before the vertical and horizontal solar magnetic fields can be determined, the 180º transverse field ambiguity must be resolved. An ambiguity of 180º in the direction of the transverse field is due to the restriction to a single plane of observations of the electric field oscillation due to the transverse field. Thus, observed polarization effects of transverse field components that are parallel and anti-parallel yield identical linear results. See Figure 1.11.

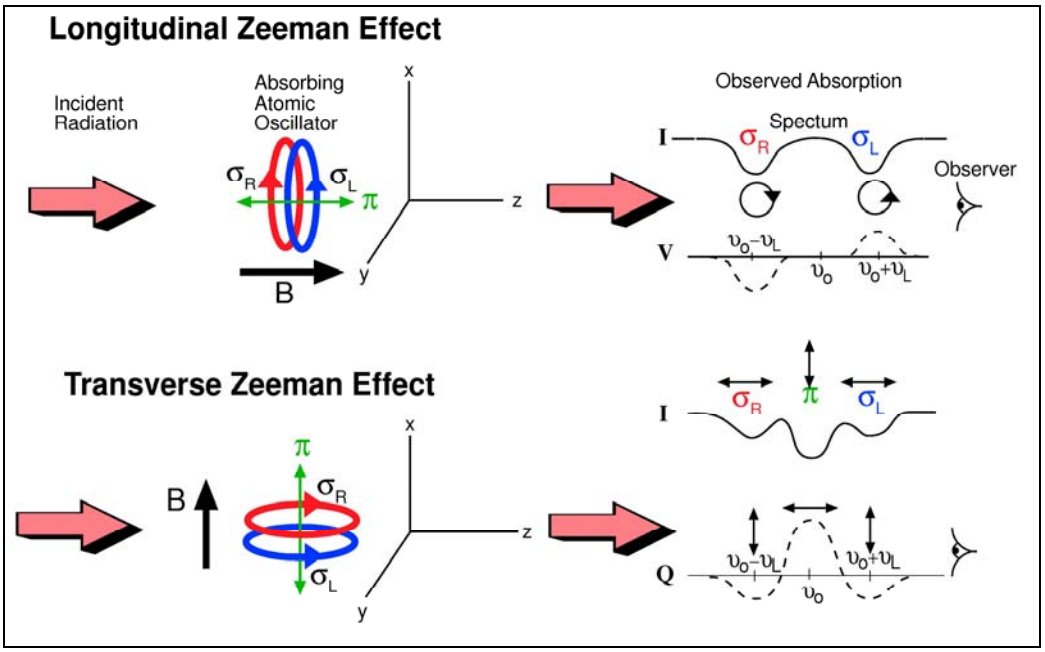


Figure 1.11 - Polarizations of Incident Radiation for Longitudinal and Transverse Magnetic Fields, http://www.hao.ucar.edu/public/research/spmf/smv_b.html

The method that was used to resolve the 180º ambiguity for the data used in this research begins by requiring the direction of the transverse field to be such that it forms the smallest angle with the transverse component of the potential field. The next step is then to convert the 2-demensional image from the image plane to the heliographic plane and the heliospheric coordinate system. After the coordinate system conversion, a second criterion is enforced and requires the orientation of the transverse field be such that it best matches the configuration of the computed force-free field. The 180º ambiguity is further resolved by minimizing the angle between neighboring field vectors. In regions of strong magnetic field, the final step is to select the orientation that minimizes the divergence of the magnetic field, $|\nabla \cdot \mathbf{B}|$. For weak magnetic field regions, the final step is to choose the orientation of the transverse component that minimizes electric current [Canfield et al, 1993].

## 1.10 Solar Magnetic Field Parameters

Insight into an AR's future flare production may be gained by understanding the state and evolution of the local magnetic field. Numerous parameters that contain information on the photosphere and solar magnetic field can be derived from vector magnetograms, and their spatial distributions can be parameterized using moment analysis. For this research, the first four moments of the parameter distributions were used,

Mean: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.18)$$

Standard Deviation: $$\sigma(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1.19)$$

Skew: $$\varsigma(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{\sigma}\right]^3 \quad (1.20)$$

Kurtosis: $$\kappa(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{\sigma}\right]^4 - 3.0 \quad (1.21)$$

Described below are a few of the parameters that may be potential indicators of solar activity and/or have been researched previously in other studies. Data for this research, as discussed below, and the previous research discussed in §1.11 were also derived from the following photospheric magnetic parameters.

### 1.10.1 Magnetic Field Vector (**B**)

Since ARs are associated with concentrations of magnetic flux, total magnetic flux is a quantity widely studied as an indicator of energetic events and can also be used as a measure of AR size. Historically, larger ARs, regions with large values of total flux, have been more likely to produce flares. Total magnetic flux is equal to the total $|B_z|$ for the entire AR or field of view, while the net flux is equal to the net vertical component of the magnetic field, $B_z$, for the entire AR or field of view. The distribution of $B_z$ may also give clues as to the likelihood a solar flare will occur. The evolution of the horizontal field component, $B_h$, also reflects the evolution of the local field. A decrease in $B_h$ may indicate emerging flux or an evolution towards a

more vertical field, whereas an increase in $B_h$ may be a sign of disappearing flux or an evolution towards a more horizontal magnetic field [Leka and Barnes, 2003a].

### 1.10.2 Spatial Magnetic Field Gradients ($\nabla \mathbf{B}$)

The emergence of new flux can lead to areas of strong spatial field gradients and shearing. Spatial magnetic field gradients are a quantitative measure of the magnetic complexity of ARs and the compactness or distribution of flux concentrations. This AR characteristic is reflected in the third designator of the McIntosh Classification Scheme (§1.6) [Leka and Barnes, 2003a].

### 1.10.3 Magnetic Shear ($\psi$)

Another parameter widely studied and linked to flare production is shear angle. Shear angle is a measure of the deviation of an AR's observed field from the potential field. Thus, it may also be a good indication as to the amount of energy stored in the local magnetic field prior to a solar flare event. In general, magnetic shear is the difference between the 3-dimensional observed magnetic field vector and the 3D potential field vector. Magnetic shear can arise from plasma motions within an AR.

In previous research, several different applications and components of shear angle have been studied. Research continues on determining which measure of shear is the most appropriate indicator of energy storage and flare productivity. Horizontal magnetic shear is defined as the difference between the observed horizontal magnetic field component and the horizontal component of the computed potential field [Li et

al., 2000]. There have also been studies of the appropriateness of restricting measurements of magnetic shear to regions near the neutral line versus over the entire AR [Leka and Barnes, 2003a; Smith, 1996]. Historically, highly stressed neutral lines have been a good indicator of imminent flare-production. Thus, by focusing on areas near neutral lines, the amount of free magnetic energy may be established.

Some researchers have proposed strong shearing in an AR is a necessary condition for flaring; however, studies investigating shear as a flare predictor or flare trigger [Smith, 1996; Li, 2000] indicate shearing alone is not an adequate factor for flare production. It has also been shown little to no loss in flare prediction accuracy is observed when other key parameters, such as total magnetic flux and persistence (past and present flare activity), are considered in the place of shear as a flare predictor [Smith et al, 1996].

### 1.10.4 Vertical Current Density ($J_z$)

The presence of strong currents is also an indication of a non-potential field. The vertical current density can be calculated from the curl of the horizontal magnetic field component.

$$J_Z(s) = (\nabla \times \mathbf{B})_h \tag{1.22}$$

Changes in the moments of $J_Z$ may indicate emerging flux. The total current density can also be broken down into two components, the current of chirality and the current of heterogeneity, where the current of heterogeneity is perpendicular to **B**. The ratio of the components of current may reflect whether or not the region is force-free. For

situations where the current of heterogeneity is greater than the current of chirality, the majority of the current flow is perpendicular to B, resulting in a Lorentz force which may add energy to the system. However, the field may be force free when the opposite is true, i.e. when current of chirality is greater than the current of heterogeneity [Leka and Barnes, 2003a]. In this situation we have less current flow perpendicular to **B**.

### 1.10.5 Twist Parameter (α)

Another measure of the amount of stress present in the solar magnetic field is the twist parameter. The twist parameter quantifies how tightly the field lines are wrapped about a flux tube [Leka and Barnes, 2003a; Holder et al, 2004; Sturrock et al, 1986]. The equation,

$$\nabla \times \mathbf{B} = \alpha \mathbf{B}, \tag{1.23}$$

describes the relationship between $\alpha$ and **B** when the region is assumed to be force-free. Since α is a measure of the stress and strain of magnetic field lines, the twist parameter may serve as a flare indicator. A field pattern with a large twist parameter may be far from its potential configuration and more likely to produce a flare.

### 1.10.6 Helicity ($h_m$)

Magnetic helicity is another measure of an AR's field complexity, twist, and deviation from the potential field, and is given by

$$H_m = \int_V \mathbf{A} \cdot \mathbf{B}\, dV \quad . \tag{1.24}$$

In ideal magnetohydrodynamics (MHD), magnetic helicity is one of the few global quantities conserved [Demoulin, et al, 2002]. Even in resistive MHD, magnetic helicity is conserved on time scales shorter than the global diffusion time scale. Current helicity, $H_c$, is given by equation 1.25, and from the temporal variations of current helicity, we can solve for the magnetic helicity.

$$H_c = \int_V \mu_o \mathbf{J} \cdot \mathbf{B}\, dV \tag{1.25}$$

[Leka and Barnes, 2003a].

### 1.10.7 Inclination Angle ($\gamma$)

The inclination angle, $\gamma$, is a parameter characterizing the magnetic field's orientation. Specifically, it is defined as

$$\gamma(s) = Tan^{-1}\left(\frac{|B_z|}{B_h}\right) \quad . \tag{1.26}$$

Values of $\gamma$ approaching 90˚ indicate a more vertical field, while values of $\gamma$ approaching 0˚ represent more horizontal fields [Leka and Barnes, 2003a].

### 1.10.8 Excess Magnetic Energy Density ($\varepsilon$)

The magnetic excess energy density is another measure of the non-potentiality of the photospheric magnetic field. A measure pertaining to the energy difference between the observed and potential fields is found by integrating $\varepsilon$ over the entire AR

and is given by equation 1.27. Since the integration is over only area and not volume, equation 1.27 does not represent the actual total energy difference.

$$E_{excess} = \frac{1}{8\pi}\int \varepsilon \, dA = \frac{1}{8\pi}\int (B_{pot} - B_{obs})^2 \, dA \qquad (1.27)$$

where $B_{pot}$ and $B_{obs}$ are the magnitudes of the potential and observed magnetic fields respectively [Leka and Barnes, 2003a].

## 1.11 **Previous Research**

Previous research has shown individual solar parameters are not by themselves good indicators of solar flare activity. Instead, it has been shown multivariable combinations of these parameters may have the capability to distinguish between flare-producing and flare-quiet active regions [Leka and Barnes, 2003a]. Research has also been conducted to apply discriminant analysis (DA) to the problem of solar flare prediction [Leka and Barnes, 2003b]. Multivariable DA is a promising tool given the distinct nature of the populations of flare prediction, flaring ARs and flare-quiet ARs.

Leka and Barnes [2003a] studied time series of the vector magnetograms of three ARs in hopes of identifying those characteristics unique to preflaring regions. The data they employed were derived from the Imaging Vector Magnetograph (IVM) data from the University of Hawai`i Mees Solar Observatory at Haleakala. The ARs included one which produced one M1 flare (AR8636), another region which produced an X3 and M1 flares (AR0030), and a flare-quiet region similar in size and

complexity as the flaring regions (AR8891). All three ARs were forecasted by SEC to flare.

Leka and Barnes studied those parameters that contained information about the state of the magnetic photosphere and/or have been researched previously in other solar activity studies (§1.10). In order to analyze the state and evolution of the photospheric magnetic field, the temporal variation and behavior of the field parameters (time averages, derivatives with respect to time) and the first four moments of the spatial distributions of the field parameters (spatial mean, standard deviation σ, skew ς, and kurtosis κ) were used as variables in the study. The field parameters included those derived from the solar magnetic field vector **B**, magnetic flux, inclination angle $\gamma$, horizontal spatial gradients of **B**, vertical current density $J_z$, twist parameter $\alpha$, current helicity density $h_c$, excess magnetic field energy density $\rho_e$, and shear angle $\psi$. See Table 1.8 for a list of the parameters.

Most of the parameters behaved similarly during both flare-quiet and flaring epochs or showed inconsistent results prior to a flare. Consequently, Leka and Barnes found no signatures unique to flaring ARs when considering parameters derived from the distribution of B and $J_z$, nor were flare-event signatures found in the higher moments of the magnetic field spatial gradients. Furthermore, a few parameters previously thought to be indicators of solar activity did not perform well in this study. The magnitudes of the total and net currents were similar for both flare-quiet and flaring epochs. Also, $\overline{|\nabla_h B|}$ and $\overline{|\nabla_h B_h|}$ showed no behavior consistent with a pre-flare signature.

Their results yielded a few parameter characteristics unique to the flare-quiet epochs, namely a larger $\sigma(\gamma)$, an evolution towards a more vertical $\gamma$, a larger $\varsigma(\rho_e)$, $\kappa(\rho_e)$, $\sigma(B_z)$, $\sigma(B_h)$, $\overline{|\nabla_h B_z|}$, $\sigma(h_c)$ and total $h_m$. Some parameters only showed slight trends prior to an event; there was a small rise in $\overline{|\nabla_h B|}$, $\overline{|\nabla_h B_z|}$, and $\kappa(J_z)$ and a possible decrease in $\sigma(J_z)$ before the flares. There were several parameters that showed strong behavior specific to the flaring epochs, such as larger $\alpha$, $\psi$, $\sigma(\psi)$, $\overline{h_c}$, and $\left|H_c^{net}\right|$.

From their initial research, Leka and Barnes found no single parameter that was an adequate predictor of a flare event. Magnitudes and evolution of certain parameters that were previously suggested to be good preflare signatures, such as magnetic flux, magnetic twist, and current flux, were nulled on account of similar behavior in the flare-quiet AR 8891. Leka and Barnes [2003a] propose no one parameter was sufficient to produce a flare. Instead, the best candidate for distinguishing an active region as flare-producing or flare-quiet may be to consider a combination of several key parameters.

Given the results of their initial research and the nature of flare prediction, Leka and Barnes then investigated the method of DA as a means of selecting an appropriate combination of photospheric magnetic parameters for prediction of solar flares [Leka and Barnes, 2003b]. For their DA research, they again used time series of vector magnetograms from the University of Hawai`i Mees Solar Observatory. The time series were then divided into epochs ending with a GOES event, an hour of

continuous data capture, or a data-gap. The resulting epochs from seven ARs included 10 flaring and 14 flare-quiet.

The research began by comparing single-variable discriminant functions (DFs). DFs are combinations of variables with the goal of classifying observations or measurements into pre-determined, exclusive groups. Discriminant function analysis is discussed further in §2.1. The variables having the highest probability the flaring and flare-quiet epochs were from different populations did not perform much better than would a random variable. However, certain variables, such as the mean of $\sigma(\psi)$, that do not perform well as a single-variable DF, are present in some of the best multivariable DFs.

Leka and Barnes then demonstrated the need to consider a multivariate combination of parameters by looking at two-variable DFs. Lower error rates were achieved, and much higher probabilities the sample data were from different populations were attained. Even lower error rates were possible when two-variable pairs were used to form four-variable DFs. Surprisingly, when the variables of two poorly performing two-variable DFs were used together to form a four-variable DF, the resulting DF performed much better than the four-variable DF created from the best two-variable DF. Leka and Barnes concluded a DF's classification ability is determined by the proper combination of variables more than by individual variables.

In an attempt to determine the ideal combination of variables, a DF was constructed for every four-variable combination and was then ranked according to the resulting probability the samples were from different populations and the classification error rate. The variables appearing most frequently in the best four-

variable DFs were then identified. The same ranking of DFs and variables was done for two-variable combinations. According to Leka and Barnes, those variables appearing most frequently in the top 20 four-variable DFs should have the greatest predictive power, while the variables appearing most in the 20 worst four-variable DFs should have little predictive power. To support this, a 10-variable DF was created from the 5 most frequently appearing variables in the top 20 DFs and the 5 most frequently occurring variables in the worst DFs (see Table 1.7). The variables appearing in the 10-variable DF were then put into standard form. When DF variables are standardized, they are modified to have a mean of zero and a standard deviation equal to one. The magnitudes of the DF coefficients then reflect the predictive powers of corresponding DF parameters. The resulting standardized DF coefficients for the 10-variable DF verified Leka and Barnes's method. The variables chosen due to their occurrence in the best four-variable DF also had the largest standardized coefficients in the 10-variable DF. The worst variables, likewise, had the smallest standardized coefficients.

**Table 1.7 - 10-Variable DF**

| Parameter | Standardized Coefficient | Frequency in Best 4-Variable DF | Frequency in Worst 4-Variable DF |
|---|---|---|---|
| $\frac{d}{dt}(\kappa(\alpha))$ | 2.444 | 244 | 0 |
| $\sigma(\alpha)$ | 1.964 | 209 | 0 |
| $\kappa(B_h)$ | 1.575 | 158 | 0 |
| $\sigma(B_h)$ | -1.326 | 79 | 0 |
| $\sigma(\psi)$ | -0.520 | 164 | 152 |
| $\frac{d}{dt}(\sigma(\varepsilon))$ | 0.492 | 1 | 154 |
| $\frac{d}{dt}(\overline{B_z})$ | -0.370 | 5 | 178 |
| $\frac{d}{dt}(A(\psi > 80°))$ | 0.352 | 6 | 188 |
| $\frac{d}{dt}(\overline{B_h})$ | -0.258 | 2 | 187 |
| $\kappa(\psi)$ | 0.204 | 3 | 176 |

[Leka and Barnes, 2003b]

By comparing the flaring and flare-quiet epochs and considering the variables found in their research to have the best predictive power, Leka and Barnes found ARs may produce flares if they experience a twist parameter with an increasing kurtosis and larger standard deviation, a smaller $\sigma(\psi)$, and a horizontal field with a smaller standard deviation and larger kurtosis. They also stress that while they show better predictions are obtained when larger numbers of parameters are considered, uncertainties arise in their method due to the sample size being much smaller than the

list of candidate variables. Their recommendation of increasing the DF sample size in order to better represent flare-quiet and flaring ARs was a task taken on by the research for this paper and is discussed below (§ 2.2).

## Table 1.8 - Photospheric Magnetic Parameters

PARAMETERS USED IN THE DISCRIMINANT ANALYSIS

| Description | Formula | Variable |
|---|---|---|
| **Atmospheric Seeing** | | |
| Median of the granulation contrast | $s = \text{median}(\Delta I)$ | $s$ |
| **Distribution of Magnetic Fields** | | |
| Moments of vertical magnetic field | $B_z = \boldsymbol{B} \cdot \boldsymbol{e}_z$ | $\mathcal{M}(B_z)$ |
| Total unsigned flux | $\Phi_{\rm tot} = \sum \lvert B_z \rvert \, dA$ | $\Phi_{\rm tot}$ |
| Absolute value of the net flux | $\lvert \Phi_{\rm net} \rvert = \lvert \sum B_z \, dA \rvert$ | $\lvert \Phi_{\rm net} \rvert$ |
| Moments of horizontal magnetic field | $B_h = \sqrt{B_x^2 + B_y^2}$ | $\mathcal{M}(B_h)$ |
| **Distribution of Inclination Angle** | | |
| Moments of inclination angle | $\gamma = \tan^{-1}(B_z / B_h)$ | $\mathcal{M}(\gamma)$ |
| **Distribution of the Magnitude of the Horizontal Gradients of the Magnetic Fields** | | |
| Moments of total field gradients | $\lvert \nabla_h B \rvert = \sqrt{(\partial B/\partial x)^2 + (\partial B/\partial y)^2}$ | $\mathcal{M}(\lvert \nabla_h B \rvert)$ |
| Moments of vertical field gradients | $\lvert \nabla_h B_z \rvert = \sqrt{(\partial B_z/\partial x)^2 + (\partial B_z/\partial y)^2}$ | $\mathcal{M}(\lvert \nabla_h B_z \rvert)$ |
| Moments of horizontal field gradients | $\lvert \nabla_h B_h \rvert = \sqrt{(\partial B_h/\partial x)^2 + (\partial B_h/\partial y)^2}$ | $\mathcal{M}(\lvert \nabla_h B_h \rvert)$ |
| **Distribution of Vertical Current Density** | | |
| Moments of vertical current density | $J_z = (\partial B_y/\partial x - \partial B_x/\partial y)/\mu_0$ | $\mathcal{M}(J_z)$ |
| Total unsigned vertical current | $I_{\rm tot} = \sum \lvert J_z \rvert \, dA$ | $I_{\rm tot}$ |
| Absolute value of the net vertical current | $\lvert I_{\rm net} \rvert = \lvert \sum J_z \, dA \rvert$ | $\lvert I_{\rm net} \rvert$ |
| Sum of absolute value of net currents in each polarity | $\lvert I_{\rm net}^B \rvert = \lvert \sum J_z (B_z > 0) \, dA \rvert + \lvert \sum J_z (B_z < 0) \, dA \rvert$ | $\lvert I_{\rm net}^B \rvert$ |
| Moments of vertical heterogeneity current density[a] | $J_z^h = (b_y \partial B_x/\partial y - b_x \partial B_y/\partial x)/\mu_0$ | $\mathcal{M}(J_z^h)$ |
| Total unsigned vertical heterogeneity current | $I_{\rm tot}^h = \sum \lvert J_z^h \rvert \, dA$ | $I_{\rm tot}^h$ |
| Absolute value of net vertical heterogeneity current | $\lvert I_{\rm net}^h \rvert = \lvert \sum J_z^h \, dA \rvert$ | $\lvert I_{\rm net}^h \rvert$ |
| **Distribution of Twist Parameter** | | |
| Moments of twist parameter[b] | $\alpha = C J_z / B_z$ | $\mathcal{M}(\alpha)$ |
| Best-fit force-free twist parameter[b] | $\boldsymbol{B} = \alpha_{\rm ff} \nabla \times \boldsymbol{B}$ | $\lvert \alpha_{\rm ff} \rvert$ |
| **Distribution of Current Helicity** | | |
| Moments of current helicity[c] | $h_c = C B_z (\partial B_y/\partial x - \partial B_x/\partial y)$ | $\mathcal{M}(h_c)$ |
| Total unsigned current helicity | $H_c^{\rm tot} = \sum \lvert h_c \rvert \, dA$ | $H_c^{\rm tot}$ |
| Absolute value of net current helicity | $\lvert H_c^{\rm net} \rvert = \lvert \sum h_c \, dA \rvert$ | $\lvert H_c^{\rm net} \rvert$ |
| **Distribution of Shear Angles** | | |
| Moments of three-dimensional shear angle[d] | $\Psi = \cos^{-1}(\boldsymbol{B}^p \cdot \boldsymbol{B}^o / B^p B^o)$ | $\mathcal{M}(\Psi)$ |
| Area with shear $>\Psi_0$, $\Psi_0 = 45°$, $80°$ | $A(\Psi > \Psi_0) = \sum_{\Psi > \Psi_0} dA$ | $A(\Psi > 45°)$, $A(\Psi > 80°)$ |
| Moments of three-dimensional neutral-line shear angle | $\Psi_{\rm NL} = \cos^{-1}(\boldsymbol{B}^p_{\rm NL} \cdot \boldsymbol{B}^o_{\rm NL} / B^p_{\rm NL} B^o_{\rm NL})$ | $\mathcal{M}(\Psi_{\rm NL})$ |
| Length of neutral line with shear $>\Psi_0$ | $L(\Psi_{\rm NL} > \Psi_0) = \sum_{\Psi_{\rm NL} > \Psi_0} dL$ | $L(\Psi_{\rm NL} > 45°)$, $L(\Psi_{\rm NL} > 80°)$ |
| Moments of horizontal shear angle[e] | $\psi = \cos^{-1}(\boldsymbol{B}^p_h \cdot \boldsymbol{B}^o_h / B^p_h B^o_h)$ | $\mathcal{M}(\psi)$ |
| Area with horizontal shear $>\psi_0$ | $A(\psi > \psi_0) = \sum_{\psi > \psi_0} dA$ | $A(\psi > 45°)$, $A(\psi > 80°)$ |
| Moments of horizontal neutral-line shear angle | $\psi_{\rm NL} = \cos^{-1}(\boldsymbol{B}^p_{h,NL} \cdot \boldsymbol{B}^o_{h,NL} / B^p_{h,NL} B^o_{h,NL})$ | $\mathcal{M}(\psi_{\rm NL})$ |
| Length of neutral line with horizontal shear $>\psi_0$ | $L(\psi_{\rm NL} > \psi_0) = \sum_{\psi_{\rm NL} > \psi_0} dL$ | $L(\psi_{\rm NL} > 45°)$, $L(\psi_{\rm NL} > 80°)$ |
| **Distribution of Photospheric Excess Magnetic Energy Density** | | |
| Moments of photospheric excess magnetic energy density[d] | $\rho_e = (\boldsymbol{B}^p - \boldsymbol{B}^o)^2 / 8\pi$ | $\mathcal{M}(\rho_e)$ |
| Total photospheric excess magnetic energy | $E_e = \sum \rho_e \, dA$ | $E_e$ |

NOTES.—$\mathcal{M}(x)$ denotes taking the first four moments of the distribution of the variable $x$: the mean, $\bar{x}$, the standard deviation, $\sigma(x)$, the skew, $\varsigma(x)$, and the kurtosis, $\kappa(x)$. For each of these variables, we consider the mean value for an epoch, denoted by $\langle\rangle$ and the slope of a regression line, denoted by $d/dt$.

[Leka and Barnes, 2003b].

# 2. Methodology

## 2.1 Discriminant Analysis Applied to Solar Flare Prediction

### 2.1.1 Discriminant Function Analysis

The method of flare forecasting and prediction, explored in this research, is based on linear discriminant analysis (DA). DA is a multivariate statistical tool with the goal of classifying objects into predetermined exclusive groups based on a combination of selected parameters. For this research, measurements of new objects to be classified are compared to the linear combination which characterizes the groups. The comparison is then used to determine to which group the objects belong [Dillon and Goldstein, 1984].

DA begins with a selection of $k$ independent variables. Based on a preexisting set of data or training sample, a discriminant function (DF) is created from the $k$ variables that is best able (given the selected variables) to distinguish between the two populations or groups that constitute the training sample. For the case of flare prediction, a linear combination is created from the selected $k$ parameters that best separates, using the chosen parameters, the flare-producing and flare-quiet populations. The training sample is used to determine how heavily each variable should be weighted within the DF. The resulting weighting coefficients can be used as a measure of a variable's contribution to the DF when in standard form. Standardized variables have means of zero and standard deviations equal to one. The direction of the discriminant vector (equation 2.1) is the direction of maximum

separation in the *k*-dimensional space created by the data parameters [Wilks, 1995]. Hence, the direction of the discriminant vector is such that it minimizes in-group variance while maximizing the between-group variance [Dillon and Goldstein, 1984].

$$\mathbf{D} = \mathbf{C}^{-1} \cdot (\boldsymbol{\mu}_f - \boldsymbol{\mu}_q) \tag{2.1}$$

The vector, $\boldsymbol{\mu}_f$, contains the mean parameter values for the flare-producing AR population, and $\boldsymbol{\mu}_q$ contains the means for the flare-quiet AR population. $\mathbf{C}$ is the total covariance matrix given by

$$\mathbf{C} = \frac{(n_f - 1) \cdot \mathbf{C}_f + (n_q - 1) \cdot \mathbf{C}_q}{n_f + n_q - 2}. \tag{2.2}$$

For the purposed of this research, the number of ARs in the flare-producing sample population, $n_f$, is assumed to equal the number of ARs in the flare-quiet sample population, $n_q$. Furthermore, the dispersion of the data in both groups is assumed to be Gaussian and equal, so that $\mathbf{C}_f$ is equal to $\mathbf{C}_q$. DA performance relies on the training sample populations being good representations of the true flare-producing and flare-quiet AR populations, and a statistically significant training sample will increase the likelihood the sample populations are adequate representations of the actual AR populations.

The discriminant vector serves as a one-dimensional tool for classifying the *k*-dimensional data. The vectors describing the means for each population are projected onto the discriminant vector. The midpoint, $\lambda$, between the projections of the mean vectors onto the discriminant vector is given by

$$\lambda = \frac{\mathbf{D}^T \cdot (\boldsymbol{\mu}_f + \boldsymbol{\mu}_q)}{2}. \qquad (2.3)$$

The left-hand side of equation 2.4 is the discriminant function (DF), and a new AR is classified based on its data vector and the value of its DF relative to zero.

Classify AR as Flare-Producing if $\mathbf{D}^T \cdot \mathbf{X} - \lambda \geq 0$ (2.4a)

Classify AR as Flare-Quiet if $\mathbf{D}^T \cdot \mathbf{X} - \lambda < 0$ (2.4b)

$\mathbf{X}$ is the data vector describing the new AR to be classified. If the value of the projection of $\mathbf{X}$ onto the discriminant function, minus the value of the midpoint, $\lambda$, is greater than or equal to zero, then the AR is predicted to be flare-producing. If the value is less than zero, the AR is predicted to remain flare-quiet. The classification boundary where the DF equals zero can be defined by the plane with a normal vector parallel to the discriminant vector passing through $\lambda$. See Figure 2.1.

Figure 2.1 is a 2-dimensional example of DA given misclassification costs and prior probabilities are equal and the parameters are uncorrelated; unequal misclassification costs and prior probabilities are addressed in the following sections. The classification boundary is equidistant to both populations' means. The values of the parameters of the new AR, $\mathbf{X}$, place it to the side of the classification boundary corresponding to the flare-producing population. According to the slope of the

discriminant vector, Parameter B is better able to determine classification and, as a result, is weighted more heavily than Parameter A.

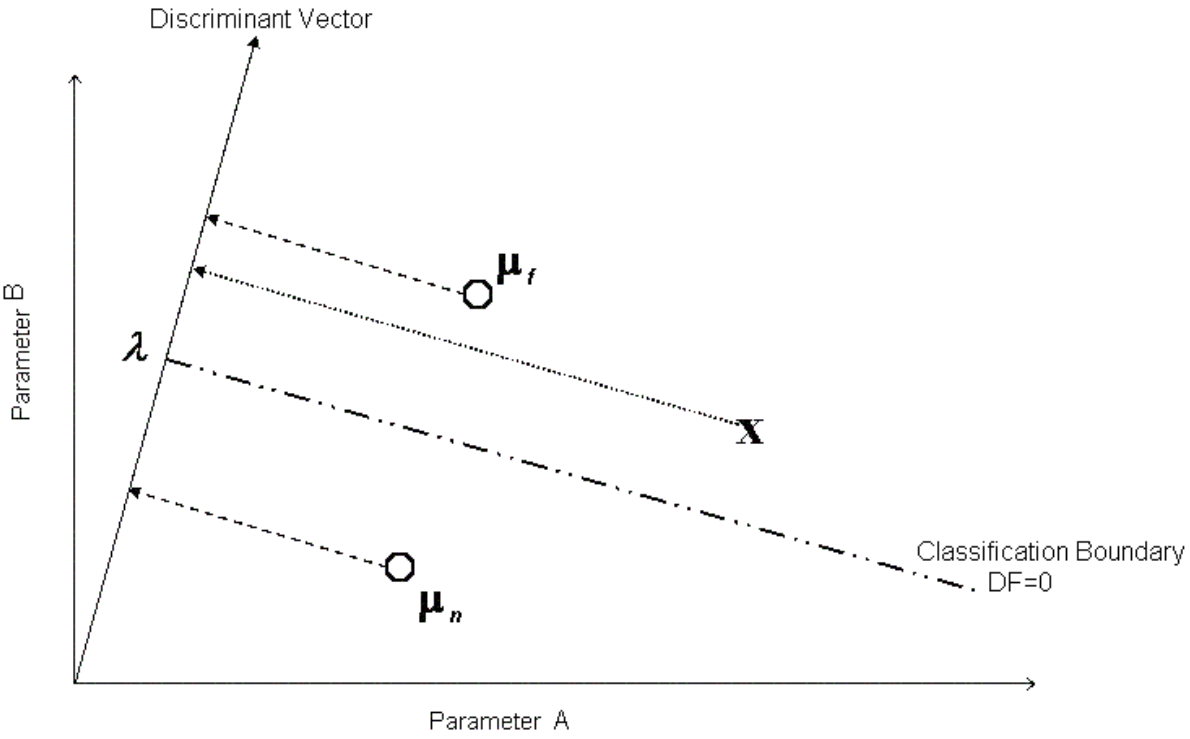


Figure 2.1 - 2-Dimensional Discriminant Analysis

### 2.1.2 Unequal Costs of Misclassification

Leka and Barnes's initial DA method did not address unequal costs associated with misclassification of an AR and was modified to allow for customized flare forecasts through the incorporation of unequal misclassification costs. The precautions taken by users of space-based systems in light of a possible solar flare can be quite costly. System shutdowns result in lost data and disruptions in communication and surveillance. It is also costly to maneuver a satellite out of

harm's way or to execute procedures to harden the satellite. Such maneuvers may also take the satellite away from its desired area of surveillance. Although damage to space-based systems due to energetic solar events can be extensive, the costs of hardening or shutting down systems in light of a solar threat cannot be ignored. Thus, an objective of this research was to allow users to customize the DF classification rules by specifying the cost associated with a miss and the cost associated with a false alarm. A miss is the misclassification of a flare-producing AR as a flare-quiet AR, and a false-alarm is the misclassification of a flare-quiet AR as a flare-producing AR.

Let the cost associated with a miss equal $L_{f,q}$ and the cost associated with a false alarm equal $L_{q,f}$. When $L_{f,q}$ and $L_{q,f}$ are equal, the classification rules are given by equations 2.4a and 2.4b. However, when $L_{f,q}$ and $L_{q,f}$ are not equal, the classification rules in equations 2.4a and 2.4b must be modified. For unequal misclassification costs, the classification rules become

Classify Observation as Flare-Producing if $$\mathbf{D}^T \cdot \mathbf{X} - \lambda - \ln\left(\frac{L_{q,f}}{L_{f,q}}\right) \geq 0 \qquad (2.5a)$$

Classify Observation as Flare-Quiet if $$\mathbf{D}^T \cdot \mathbf{X} - \lambda - \ln\left(\frac{L_{q,f}}{L_{f,q}}\right) < 0 \qquad (2.5b)$$

The term, $\ln\left(\frac{L_{q,f}}{L_{f,q}}\right)$, in equations 2.5a and 2.5b effectively moves the classification boundary according to the given costs. For example, if $L_{f,q}$ is larger than $L_{q,f}$, their

ratio would be less than one, and $\ln\left(\frac{L_{q,f}}{L_{f,q}}\right)$, would be negative. Thus, the classification boundary would shift towards the mean of the flare-quiet population. As a result, a new data vector would have a higher probability of being classified as a flare-producing AR. The occurrences of misses would then decrease; although, the rate of false alarms would increase [Wilks, 1995]. Figure 2.2 is an example of the shift of the classification boundary in response to a miss being twice as costly as a false-alarm.

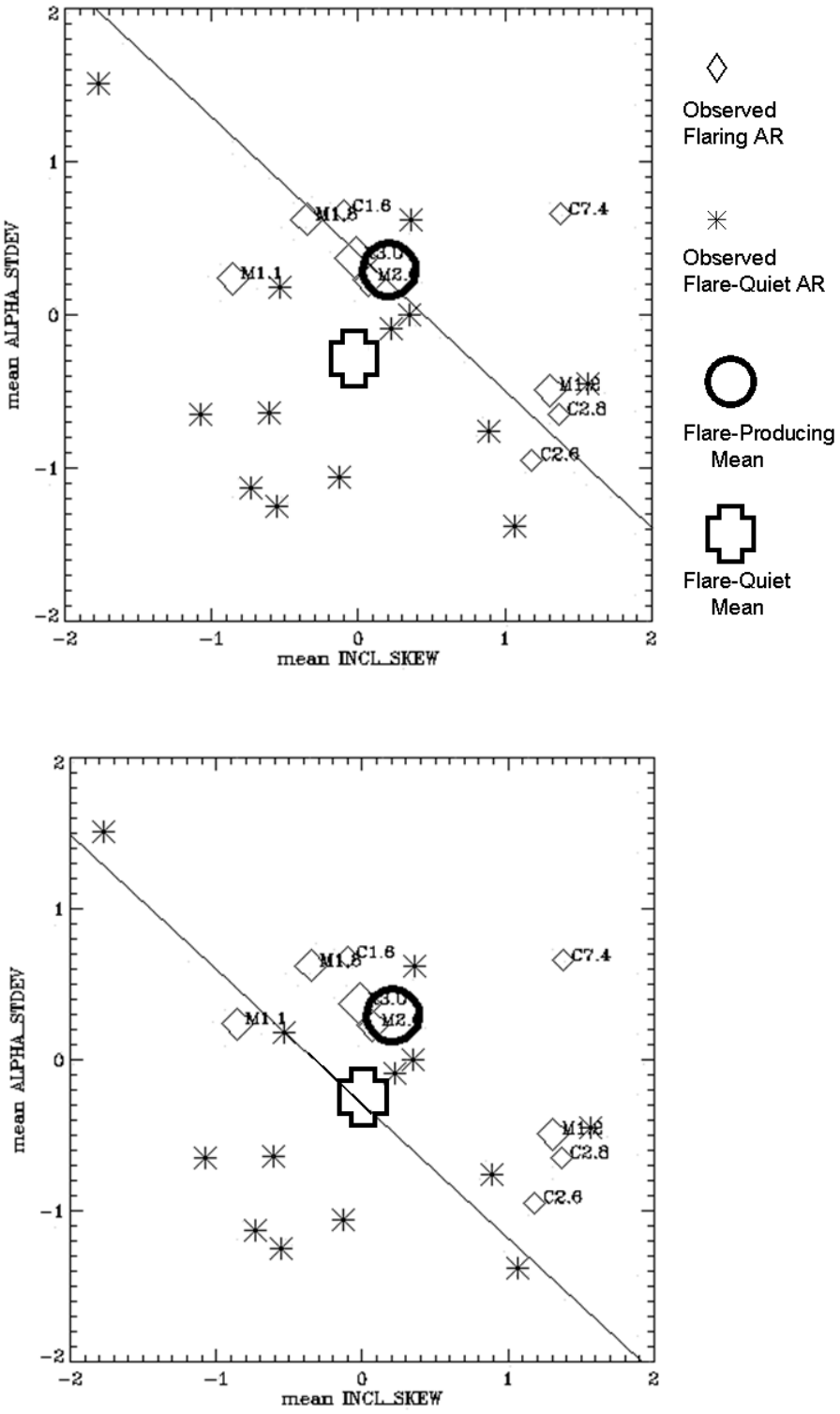


Figure 2.2 - DF and Unequal Misclassification Costs,
The plots show the results of the DF for the two parameters, the skew of the inclination angle $\varsigma(\gamma)$ and the standard deviation of the twist parameter $\sigma(\alpha)$. The top panel represents the classification boundary for equal misclassification costs. The bottom panel shows how the classification boundary moves towards the flare-quiet mean in response to the cost associated with a miss being twice the cost associated with a false-alarm.

### 2.1.3 Unequal Prior Probabilities of Membership

If prior probability of membership to the flare-producing population, $P_f$, is not equal to prior probability of membership to the flare-quiet population, $P_q$, then the classification rules are given in equations 2.6a and 2.6b.

Classify Observation as Flare-Producing if $$\mathbf{D}^T \cdot \mathbf{X} - \lambda - \ln\left(\frac{L_{q,f}}{L_{f,q}} \cdot \frac{P_q}{P_f}\right) \geq 0 \qquad (2.6a)$$

Classify Observation as Flare-Quiet if $$\mathbf{D}^T \cdot \mathbf{X} - \lambda - \ln\left(\frac{L_{q,f}}{L_{f,q}} \cdot \frac{P_q}{P_f}\right) < 0 \qquad (2.6b)$$

If, for example, $P_q$ is greater than $P_f$, the classification boundary would move towards the flare-producing population mean as to allow for a greater number of future ARs to be classified as flare-quiet [Wilks, 1995]. See Figure 2.3.

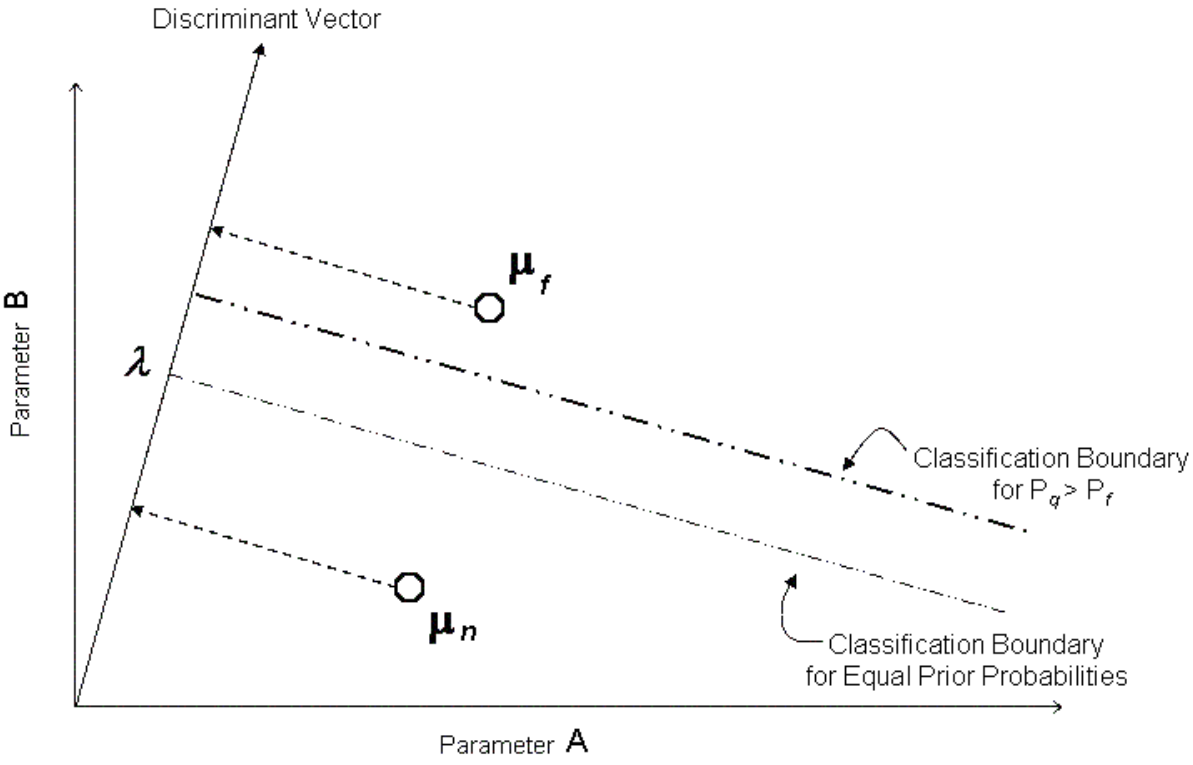


Figure 2.3 - 2-Dimensional DA for the Case of Greater Prior Probability of Membership to the Flare-Quiet Population

Misclassification costs and prior probabilities can be customized in order to satisfy one of the following misclassification criteria.

1) Minimize total instances of misclassification (false-alarms plus misses)

2) Require misclassification rate of misses be proportional to the flare-producing population size, and require misclassification rate of false-alarms be proportional to the flare-quiet population size

3) Require misclassification rates of misses and false-alarms be equal

For the DA flare prediction method developed in this research and in the preceding research conducted by Leka and Barnes [2003b] (§1.11), the goal was to minimize the overall misclassification errors.

### 2.1.4 The Mahalanobis Distance

A statistical measure of the separation between the two populations' mean vectors is the Mahalanobis distance. Unlike the squared Euclidean distance between two vectors, the Mahalanobis distance takes into account the variances of the variables as well as their covariances. See equations 2.7 and 2.8. For DA and statistical purposes, the Euclidean distance may not be very informative [Rencher, 2002]. The Mahalanobis distance relates the distance between the two vectors to how many standard deviations separated them.

$$\text{Euclidean Distance} = (\boldsymbol{\mu}_f - \boldsymbol{\mu}_q)^{\mathrm{T}}(\boldsymbol{\mu}_f - \boldsymbol{\mu}_q) \tag{2.7}$$

$$\text{Mahalanobis Distance} = (\boldsymbol{\mu}_f - \boldsymbol{\mu}_q)^{\mathrm{T}}\,\mathbf{C}^{-1}\,(\boldsymbol{\mu}_f - \boldsymbol{\mu}_q) \tag{2.8}$$

The presence of the inverse of the covariance matrix in the definition of the Mahalanobis distance standardizes all variables to the same variance and reduces correlation among variables. Thus, parameters with larger variances or are highly correlated are weighted less when computing the Mahalanobis distance [Rencher, 2002].

From the Mahalanobis distance, a measure of how probable it is that observations are from the same population can be derived. As the Mahalanobis distance increases, certainty in the existence of two discrete populations increases. However, an increase

in the Mahalanobis distance does not necessarily lead to lower misclassification rates given the variance within a population may be large. The Mahalanobis distance is one of the tools we used to analyze and select parameters for the DF-based flare forecasting method used in this research and discussed below.

## 2.2 Improving Statistical Significance of Sample Size

In the previous research conducted by Leka and Barnes [2003b] the number of photospheric parameters considered greatly outnumbered the number of data points in the training sample. Furthermore, the data used for their research were composed of time series of magnetograms and were used to investigate DA as a means of producing hourly flare forecasts. However, the purpose of this research was to explore DA as a tool for producing daily flare forecasts, and the data used here were "snapshots" of daily magnetograms of ARs present on the solar disk. Consequently, time derivatives of parameters were not possible. A priority of this research was also to improve the confidence of the DA results and to allow for more solar magnetic parameters to be considered by increasing the size of the training sample.

Statistically sound DA results rely on the population of the training sample to be much larger than the number of parameters to be consider for DF variables. The datapoints of a statistically significant training sample adequately describe the true flare-producing and true flare-quiet populations; thus, as the training sample size increases, confidence in the results also increases. Presently, the training sample has 1182 data points, well over the 147 photospheric magnetic parameters to be

considered. The dataset includes magnetograms from January 2001 to February 2003, 263 of which are associated with flaring ARs.

An AR was classified as flare-producing if at least one flare occurred within 24 hours after a magnetogram was taken of the region and only if the flare was reported as a GOES event of C-class or larger. If an AR did not produce at least one flare within 24 hours of a magnetogram being taken of it, then the AR was classified as flare-quiet. Only magnetograms with an observation angle of at most 50˚ from disk center and containing a solar magnetic field measurement of at least 500G were used.

The process of adding to the training data begins with the date and time stamp for a given magnetogram. From the date and time stamp, the AR number corresponding to the magnetogram can be found on the University of Hawai`i Mees Solar Observatory website (http://www.solar.ifa.hawaii.edu/IVM/archive.html). Soft X-ray and optical flare event information from NOAA SEC's website can then be associated with the appropriate AR (http://www.sec.noaa.gov/ftpmenu/indices.html).

## 2.3 Flare Probability Forecasts

In an attempt to compare how well DA-based flare prediction method does with respect to SEC's flare forecasting method, daily flare probabilities are produced along with the DF classification of an AR as flare-producing or flare-quiet. Effectively, the binary DF classification of ARs corresponds to a 50% forecast. The discriminant analysis development for this research assumes the probability distributions of flare-quiet and flare-producing AR populations are Gaussian and have equal covariance

matrices. These assumptions make it possible to revise Leka and Barnes's initial DA approach in order to produce flare probability forecasts based on the Gaussian distribution.

Probability forecasts were created by comparing parameter values of a candidate AR to the mean parameter values of the flare-quiet and flare-producing populations.

$$f(y) = \frac{e^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \tag{2.9}$$

Equation 2.9 represents a normalized, single variable Gaussian probability distribution. For a given population, μ is the parameter's mean value, *y* is the measured value, and σ is the population standard deviation. If a population is described by a set of *k* parameters, then the multivariate probability distribution normalized to unity is given by equation 2.10.

$$f(\mathbf{Y}) = \frac{1}{(2\pi)^{k/2}}\left|\mathbf{\Sigma}\right|^{-1/2} Exp\left[\frac{-1}{2}(\mathbf{Y}-\boldsymbol{\mu})'(\mathbf{\Sigma}^{-1})(\mathbf{Y}-\boldsymbol{\mu})\right] \tag{2.10}$$

where $\mathbf{\Sigma}$ is the population covariance matrix, $\mathbf{Y}$ is the vector of parameter values for the new AR to be classified, and f($\mathbf{Y}$) is the probability of observing $\mathbf{Y}$.

In order to produce a flare probability forecast, Gaussian probability distributions for both flare-quiet and flare-producing populations are calculated. Each population's distribution is then weighted by the total number of its members, $n_j$, to take into account unequal population sizes. Equation 2.11 now represents Gaussian probability distributions normalized to $n_j$.

$$f(\mathbf{Y})_j = \left(\frac{1}{(2\pi)^{k/2}}\left|\mathbf{\Sigma}\right|^{-1/2} Exp\left[\frac{-1}{2}(\mathbf{Y}-\boldsymbol{\mu})'(\mathbf{\Sigma}^{-1})(\mathbf{Y}-\boldsymbol{\mu})\right]\right)n_j \tag{2.11}$$

Thus, the value, $f(\mathbf{Y})_{flare-producing}$, corresponds to the number of flare-producing active regions described by the given measurement, **Y**. Likewise, $f(\mathbf{Y})_{flare-quiet}$ is the number of active regions sharing parameter values with the measurement, **Y**, that did not produce a flare. Our flare probability is then given by equation 2.12.

$$P(\mathbf{Y}) = 100\% * \frac{f(\mathbf{Y})_{flare-producing}}{f(\mathbf{Y})_{flare-quiet} + f(\mathbf{Y})_{flare-producing}} \tag{2.12}$$

The need to weight the distributions by population size can be seen in Figures 2.4 and 2.5. When both distributions are normalized to one and are not weighted by population size, as in Figure 2.4, a measurement of **Y** misleadingly looks to have a greater probability of having membership in Population 1. However, when the distributions are weighted by population size, we see there is actually a greater probability an observation of **Y** is from Population 2. See Figure 2.5. Thus, it is clear we are unable to compare the relative probabilities of an observation unless population sizes are taken into account.

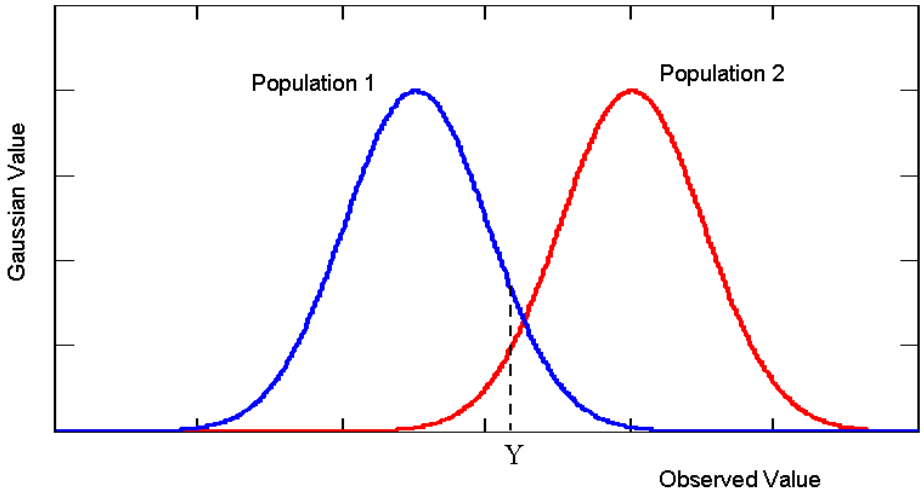

Figure 2.4 - Gaussian Distributions Normalized to One

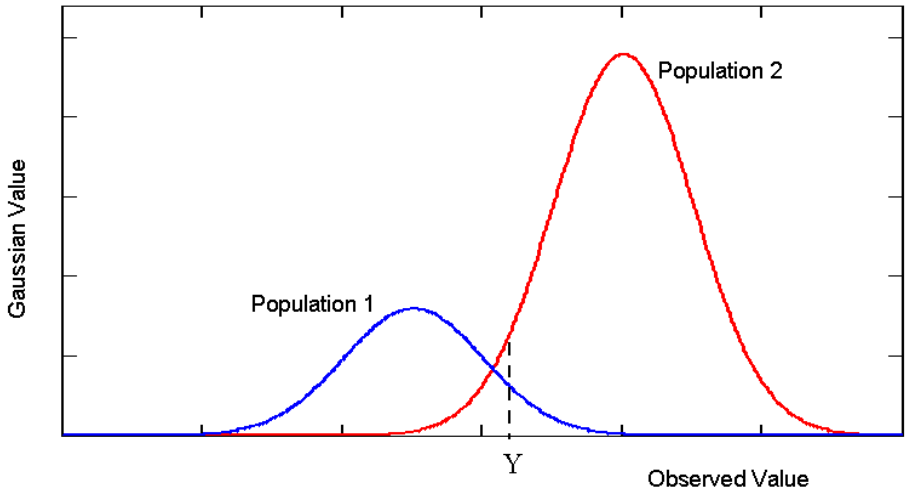


Figure 2.5 - Gaussian Distributions Weighted by Population Sizes and Normalized to $n_j$

## 2.4 Highly Correlated Variables and Shear Measure Selection

### 2.4.1 Shear Measure Selection Method

Many of the variables considered as DF parameters are highly correlated. Highly correlated parameters used in a DF may yield misleading results. Since the predictive power of the individual parameters would be shared, a situation could occur where one parameter, that alone is a good flare predictor, would in effect share its predictive power with another correlated variable. As a consequence, neither parameter would appear as a good indicator of a flare, or the opposite may also occur and both parameters could deceptively appear to be very good flare predictors.

To investigate a method of selecting among highly correlated variables, this research turned to the list of 40 parameters associated with the four different measures of shear angle (§1.10.3). Once an adequate test for selecting variables from a set of correlated parameters is identified, the goal is to apply the test to other groups

of highly correlated variables. See Tables 2.1a and 2.1b for the definitions of the shear measures and a list of the parameters derived from the moments and variables associated with each shear measure. Each measure of shear brought much of the same information to the DA. By selecting only one measure of shear to be included in the DF, process time was greatly reduced, predictive power was less diluted among the shear variables, and a candidate for the best measure of magnetic shear angle was identified.

**Table 2.1a - Shear Measures**

| | |
|---|---|
| 3D Shear Angle | Angle between observed **B** and the potential field vector, calculated over the entire AR |
| 3D Neutral Line Shear Angle | Angle between observed **B** and the potential field vector, calculated only in areas near neutral lines |
| Horizontal Shear Angle | Angle between the horizontal component of observed **B** and the horizontal component of the potential field vector, calculated over the entire AR |
| Horizontal Neutral Line Shear Angle | Angle between the horizontal component of observed **B** and the horizontal component of the potential field vector, calculated only in areas near neutral lines |

**Table 2.1b - Shear Parameters**

1. $\mu(\psi)$
2. $\sigma(\psi)$
3. $\varsigma(\psi)$
4. $\kappa(\psi)$
5. Total Area within AR with $\psi \geq 45^{\circ}$
6. Total Area within AR with $\psi \geq 80^{\circ}$
7. Fraction of Neutral Line with $\psi \geq 45^{\circ}$
8. Fraction of Neutral Line with $\psi \geq 80^{\circ}$
9. $\mu(\psi)$ Weighted by $|\mathbf{B}_h|$
10. $\sigma(\psi)$ Weighted by $|\mathbf{B}_h|$

We began by creating a 10-variable DF for each shear measure in order to evaluate how well each measure performs as a flare predictor and contributes to an accurate flare probability forecast. For each of the flare forecasts created by the shear measures, verification plots similar to Figure 1.7 were created and the $\chi^2$ value and Mahalanobis distance were computed. See Figures 2.6 through 2.9 and Tables 2.3 though 2.6 for a summary of the shear parameters results. The shear measures were then ranked according to the forecasts' $\chi^2$ values and Mahalanobis distances.

A probability bin boundary configuration for the verification plots was initially chosen to maintain statistical significance across all bins. Since the $\chi^2$ value is weighted by bin population, other forecast bin boundary configurations were selected to analyze the $\chi^2$ dependence on boundary selection. For example, one configuration

of bin boundaries was established so that the bin populations within a given verification plot were all approximately equal. The relative rankings of the shear measures as flare predictors, when sorted in relation to the $\chi^2$ value of their forecasts, proved to be highly dependent on bin boundary placement, so the $\chi^2$ value was not a reliable method by which to rank the shear measures. However, the rankings were consistent when sorting was determined by the Mahalanobis distance. See Table 2.2.

**Table 2.2 - Shear Measure Forecast Ranking with respect to Mahalanobis Distance**

| RANKING | MEASURE |
|---|---|
| 1 | Horizontal Shear |
| 2 | 3D Neutral Line Shear |
| 3 | Horizontal Neutral Line Shear |
| 4 | 3D Shear |

Another approach was taken to investigate the appropriateness of using the $\chi^2$ value as a means of ranking variables. Working with one measure of shear at a time, a "step-up" method was employed to test whether or not the $\chi^2$ value monotonically decreased as parameters were added to the DF. The expected behavior of DA, if all assumptions are valid, is an improvement in classification as the number of discriminant parameters increases. The step-up approach begins with a list of the 10 parameters derived from a single measure of shear. The procedure then selects the first variable from the list that returns the smallest $\chi^2$ value for its single-variable DF.

Once the first variable is selected, the step-up procedure cycles through the remaining nine variables and selects the variable, when coupled to the first variable, returns the smallest $\chi^2$ value for the two-variable DF. The process is repeated until all 10 variables have been selected and a 10-variable DF is created. Against expectations, the $\chi^2$ value did not monotonically decrease as variables were added to the DF; instead, the error rate fluctuated and showed no consistent behavior.

The step-up approach was repeated for each measure of shear with the Mahalanobis distance as the criterion for variable selection. When the Mahalanobis distance was used, there was a monotonic improvement in the error rate as parameters were added to the DF. Due to its dependence on bin population and bin boundaries and its inconsistent behavior, the $\chi^2$ value was ruled out as a criterion for shear measure and variable selection in favor of a selection rule based on the Mahalanobis distance. Furthermore, the Mahalanobis distance is not affected by misclassification costs or prior probability changes in the DF, increasing its robustness as a selection rule. Due to its consistent top ranking with respect to Mahalanobis distance, horizontal shear angle, the measure of shear angle defined as the difference between the horizontal components of the observed and potential field, is the measure distinguished as the best gauge of shear in this research.

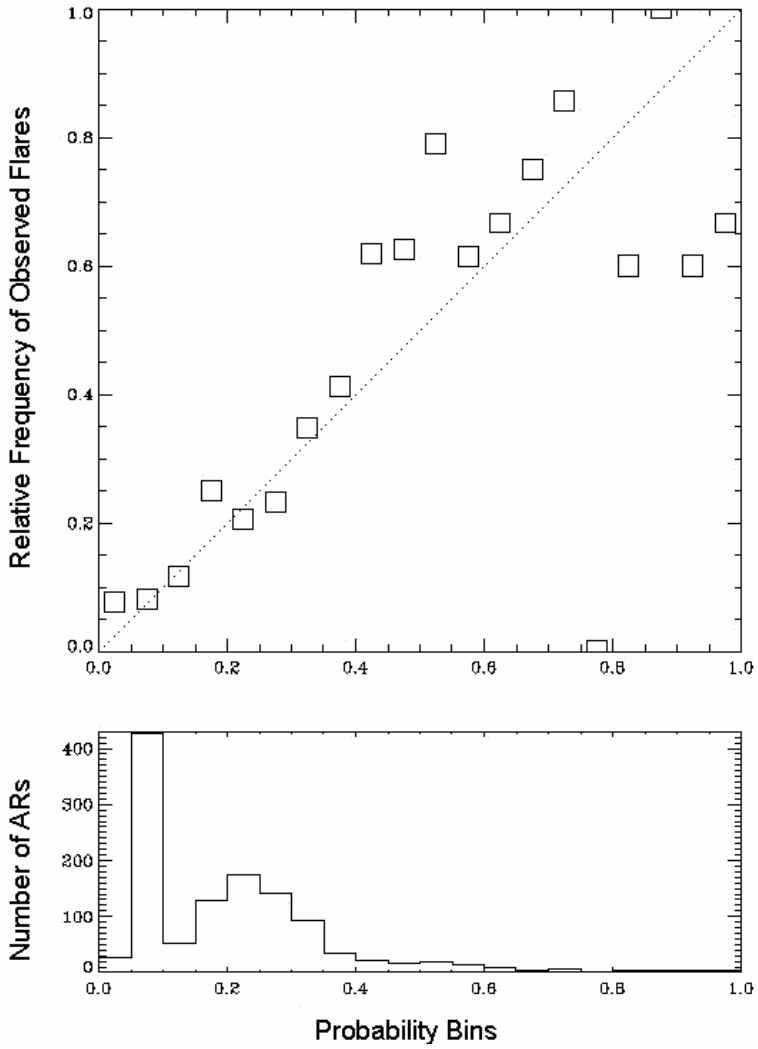


Figure 2.6 - Horizontal Shear Angle Verification Plot

**Table 2.3a - Horizontal Shear Angle Classification Table**

| | | Predicted | |
|---|---|---|---|
| | | Flare | Flare-Quiet |
| Observed | Flare | 5 | 212 |
| | Flare-Quiet | 18 | 901 |

**Table 2.3b - Horizontal Shear Angle**

| |
|---|
| Mahalanobis Distance……..1.0283 |
| $\chi^2$ Value…….....1.0163 |
| Rate of Correct Classification…..0.8037 |

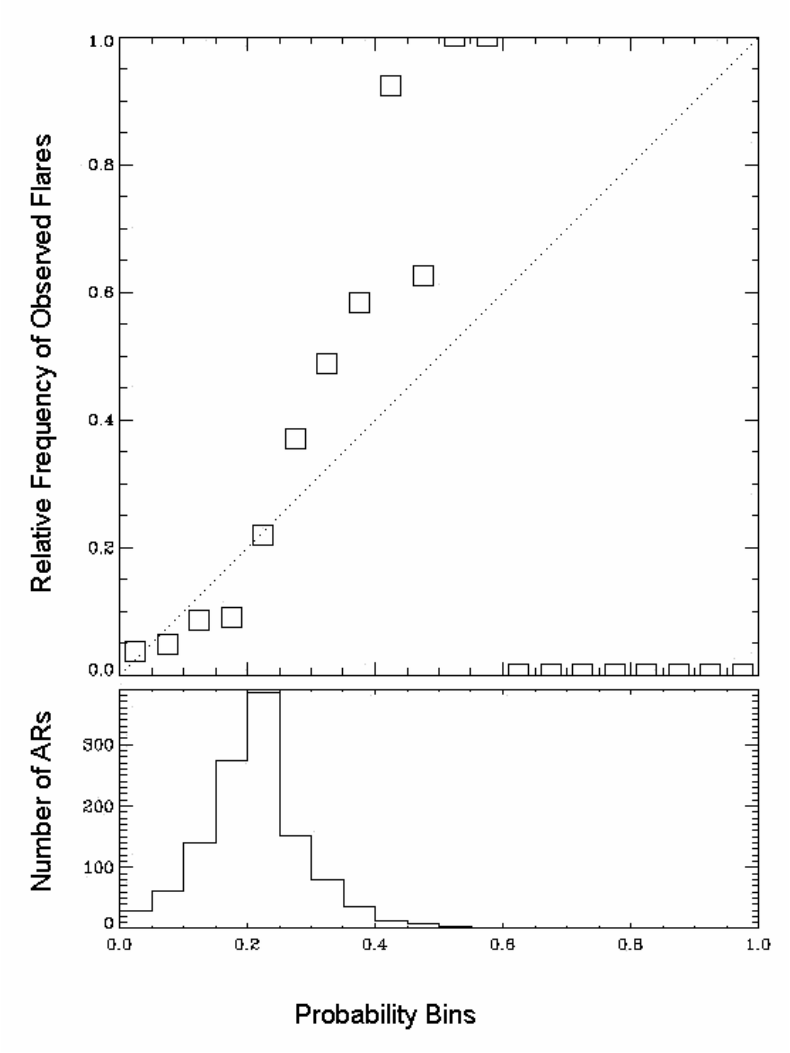


Figure 2.7 - 3D Shear Angle Verification Plot

**Table 2.4a - 3D Shear Angle Classification Table**

| | | Predicted | |
|---|---|---|---|
| | | Flare | Flare-Quiet |
| Observed | Flare | 5 | 258 |
| | Flare-Quiet | 2 | 917 |

**Table 2.4b - 3D Shear Angle**

| |
|---|
| Mahalanobis Distance……..0.4655 |
| $\chi^2$ Value……..0.6802 |
| Rate of Correct Classification…..0.7800 |

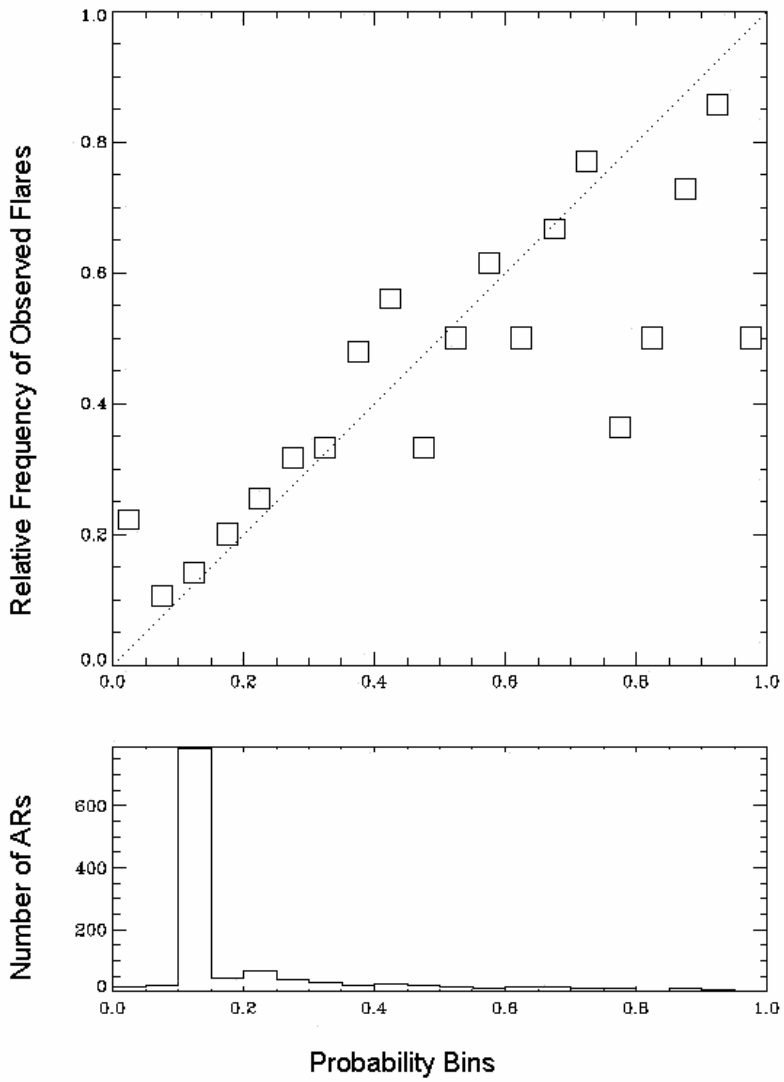


Figure 2.8 - Horizontal Neutral Line Shear Angle Verification Plot

**Table 2.5a - Horizontal Neutral Line Shear Angle Classification Table**

| | | Predicted | |
|---|---|---|---|
| | | Flare | Flare-Quiet |
| Observed | Flare | 69 | 194 |
| | Flare-Quiet | 46 | 873 |

**Table 2.5b - Horizontal Neutral Line Shear Angle**

| |
|---|
| Mahalanobis Distance……....0.9266 |
| $\chi^2$ Value……..0.5572 |
| Rate of Correct Classification…..0.7970 |

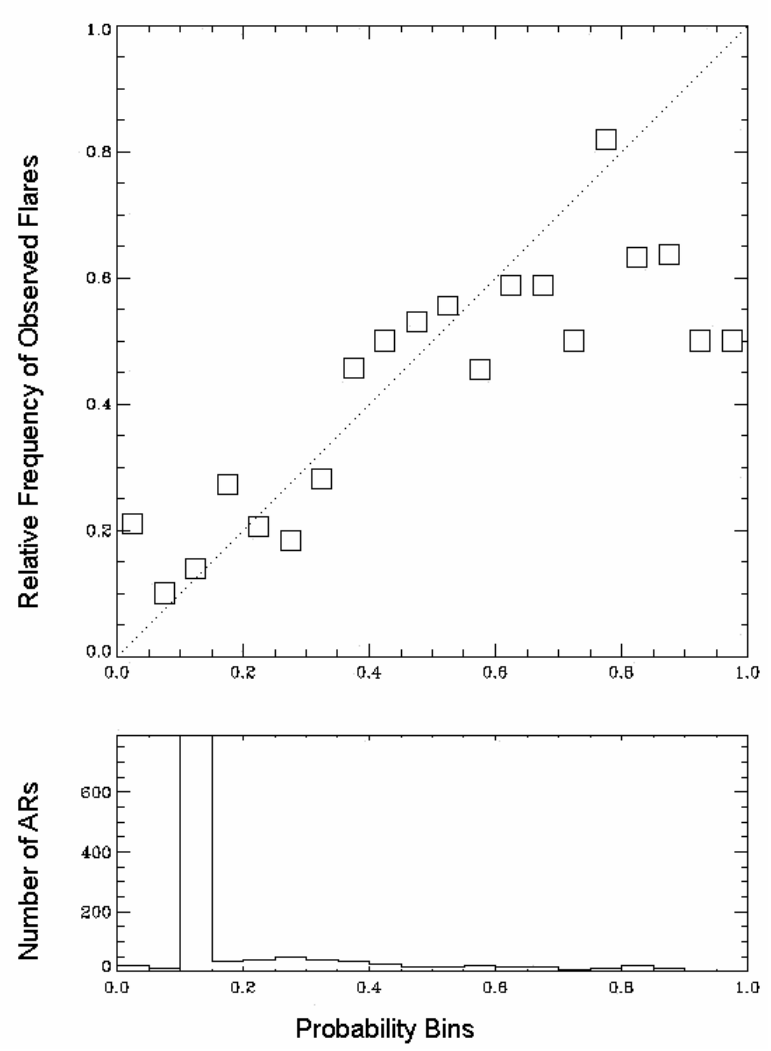


Figure 2.9 - 3D Neutral Line Shear Angle Verification Plot

**Table 2.6a - 3D Neutral Line Shear Angle Classification Table**

| | | Predicted | |
|---|---|---|---|
| | | Flare | Flare-Quiet |
| Observed | Flare | 79 | 184 |
| | Flare-Quiet | 51 | 868 |

Table 2.6b - 3D Neutral Line Shear Angle

| | |
|---|---|
| Mahalanobis Distance…….....0.9328 | |
| $\chi^2$ Value……..0.9214 | |
| Rate of Correct Classification…..0.8012 | |

### 2.4.2 Shear Measure Probability Distributions

The forecast distributions of the 10-variable DF for each of the four shear measures can be seen in the bottom frames of Figures 2.6 through 2.9 in §2.4.1. An interesting thing to note is how the probability distributions differ for the different measures of shear.

Both calculations of the shear angle restricted to areas near the neutral line have very similar distributions, peaking with most of their population in the 10-14% probability bin and having very few forecasts in the higher probabilities.

The probability distribution of the 3D shear measure is extremely different and has more of a bell-shape. It also peaks at a higher probability bin, 20-24%, but does not assign a forecast of flaring greater than 55% to any AR. Looking at the forecast verification plot for 3D shear (Figure 2.7) we see the parameter under-forecasted flare occurrences for all bins between 25% and 60%. Furthermore, the 10-variable DF of the 3D shear parameters only predicted 5 ARs to flare out of the 263 flare-producing ARs. Its poor ability as a discriminator is captured in its relatively small value for the Mahalanobis distance. However, compared to the other shear measures, the 3D shear forecast's $\chi^2$ value is one of the smallest, but that is due to the absence of ARs assigned to the higher probability bins. Recall from §1.8, each $k^{th}$ term in the calculation of the $\chi^2$ value is weighted by the number of ARs assigned to the $k^{th}$ bin. The higher probability bins, where agreement between prediction and observation was poor, did not contain any datapoints and, thus, did not factor into the calculation of the $\chi^2$ value. There are a few possible explanations for the poor performance of the 3D shear parameters. 1) The 3D shear measure may not be a good indication of the

magnetic shearing, and as a consequence, the 3D shear parameters just are not good predictors and cannot be used to discriminate between flare-producing ARs and flare-quiet ARs. 2) The distributions of the 3D shear parameters do not resemble a Gaussian distribution, a violation of the DA assumptions. 3) The much greater flare-quiet population size might have overshadowed the little predictive power possessed by the 3D shear variables. From §2.3, the probability forecast method developed in this research accounted for unequal population sizes, and the number of flare-quiet ARs is about 350% larger than the number of flaring ARs. Thus, unequal population sizes coupled with the small Mahalanobis distance would contribute to most ARs being categorized as flare-quiet. 4) Some researchers have suggested limiting the calculation of magnetic shear angle to only the areas near AR neutral lines, and our results of the parameters derived from the measures of 3D shear and measures restricted to the neutral line initially seem to support this. If in fact it is the shearing in areas near the neutral lines that most contribute to solar flares, then calculating the 3D shear for the entire AR may be diluting the flare-specific information.

The measure of shear selected to be used in the step-up procedure, the horizontal shear measure, shares probability distribution characteristics with the measures that are restricted to the neutral line and with the 3D shear measure. Its probability distribution is similarly peaked as the neutral line measures, but at the 5-9% probability bin, and it has a small almost bell-like feature between the probabilities of 10% and 49%.

The shear measure probability distributions tell us the 3D shear measure may not be an adequate measure of shear since it predicts no ARs to flare. The measure of

horizontal shear angle which is calculated over the entire AR and the measure of horizontal shear angle restricted to the neutral line produce accurate forecasts when bins contained many datapoints. It may be in the horizontal field where we should look for shear thresholds for flare production.

## 2.5 Discriminant Function Variable Selection

### 2.5.1 Selecting DF Variables

How well the DF performs is directly related to the selected variables' ability to predict solar flares. Due to time constraints of this research and the processing time of the DF procedures used, the list of initial 147 possible photospheric magnetic and AR parameters had to be reduced. A shortened list of parameters was chosen by reducing correlation and redundancy among the parameters and by eliminating those parameters most unlikely to contribute to flare production. Once a reduced list of parameters was chosen, the step-up method, described in §2.4.1, was applied to determine the best subset of parameters from which to create the DF. See Appendix A for a list of the 147 possible parameters and those chosen for the final list used in the step-up procedure.

From among the AR parameters, we selected seeing and observation angle to include in the final, shortened variable list. Although these two variables give no insight into flare production, they were included as control variables. If the step-up procedure had selected either parameter as having strong predictive abilities, we

would have been hesitant to trust our results. Fortunately, our procedures did not identify either parameter as being a good predictor of flares.

Redundancy of information and high correlation among the parameters was also a concern. By identifying redundancy among variables, we were able to eliminate many parameters. The total vertical current density $\mathbf{J}_z$, for example, is the sum of the current of chirality $\mathbf{J}_c$ and the current of heterogeneity $\mathbf{J}_h$. Thus, all the information available for $\mathbf{J}_z$ can be contained in two of the three parameters, and we chose to exclude those parameters derived from the current of chirality. Current of heterogeneity is normal to **B** and may shed light on any forcing that may be present. Redundancy within the $\mathbf{J}_z$ and $\mathbf{J}_h$ parameters was further reduced by excluding the parameters describing the absolute values and signed values of the positive and negative currents. This information is incorporated in the parameters for the total currents $\mathbf{I}_{tot}$ and $\mathbf{I}_{h\_tot}$, the absolute value of the net currents $\mathbf{I}_{net}$ and $\mathbf{I}_{h\_net}$, and the mean values for $\overline{\mathbf{J}_z}$ and $\overline{\mathbf{I}_h}$.

$$\mathbf{I}_{\mathbf{tot}} = \mathbf{I}_+ + \left| \mathbf{I}_- \right| \tag{2.13}$$

$$\mathbf{I}_{net} = \left| \mathbf{I}_+ + \mathbf{I}_- \right| \tag{2.14}$$

$$\mathbf{I}_{h\_\mathbf{tot}} = \mathbf{I}_{h+} + \left| \mathbf{I}_{h-} \right| \tag{2.15}$$

$$\mathbf{I}_{h\_net} = \left| \mathbf{I}_{h+} + \mathbf{I}_{h-} \right| \tag{2.16}$$

where $\mathbf{I}_+$ and $\mathbf{I}_-$ are the currents associated with each sign.

Since the direction of the net current may yield information about flare production, we included the signed first four moments of $\mathbf{J}_z$. We did not include the

parameters describing the vertical currents emanating from each magnetic polarity. Due to hemispheric biases and solar cycle effects on solar magnetic polarity, currents specific to a magnetic polarity may not contribute to a general picture of a preflaring state.

The total magnetic field vector parameters is another set of variables that can be reduced by including only those parameters describing elements that may contribute to solar flares and are not redundant. The first four signed moments of the horizontal and total magnetic field vectors are included. However, only the moments of the absolute value of the vertical field and the absolute value of the net flux $|\Phi_{net}|$ are used in order to avoid hemispheric biases. The magnitude should be a better indicator of solar activity than the direction of the vertical magnetic field and magnetic flux. The parameters for the flux associated with each magnetic polarity are not used since the information is included in the parameter, total flux $\Phi_{tot}$, which is a parameter included in the step-up procedure.

$$|\Phi_{net}| = |\Phi_{+} + \Phi_{-}| \tag{2.17}$$

$$\Phi_{tot} = |\Phi_{+}| + |\Phi_{-}| \tag{2.18}$$

Furthermore, the absolute values of the moments of the twist parameter and of current helicity are also used along with the absolute value of the net current helicity. The signed values are not included since it is the amount, and not the direction, of twist and helicity present in an AR that is an indication of the complexity of the magnetic field and of flare production. Due to their ability of quantifying the magnetic complexity of an AR, we also include in the step-up procedure the first four

moments of the inclination angle and excess energy; total excess energy; the signed gradients of the vertical, horizontal, and total magnetic field; and the 10 horizontal shear parameters.

To complete the selection of the most appropriate DF parameters, we applied the step-up procedure to the final list of 69 photospheric magnetic parameters using the Mahalanobis distance as the selection rule. We addressed the problem of how many parameters to include in the construction of the final DF by noting in the step-up procedure the amount of improvement in the Mahalanobis distance with the addition of the each parameter. Given our set of 69 parameters, there was little improvement after the fifth and sixth parameters were added. The top six parameters selected, in order, were the total magnetic flux, the total area of the AR with horizontal shear angle greater than 80°, the mean of the gradient of the vertical magnetic field, the total current of heterogeneity, the kurtosis of the horizontal shear angle, and the standard deviation of the horizontal shear angle.

Since there was little improvement in the Mahalanobis distance with the addition of the sixth variable and the amount of improvement progressively decreased and leveled off, we evaluated the performance of the DF created from all of the top six parameters and compared the results to those of the DF created from only the top five parameters. See Table 2.7.

**Table 2.7 – 6-Variable versus 5-Variable DF**

| | 6-Variable DF | 5-Variable DF |
|---|---|---|
| **Mahalanobis Distance** | 2.135 | 2.075 |
| **$\chi^2$ Forecast Value** | 1.595 | 1.118 |
| **Rate of Correct Classification** | 0.827 | 0.829 |

Even though the 6-variable combination resulted in a larger Mahalanobis distance, we concluded only five parameters were needed for the DF due to the better probability forecast and lower error rates obtained with the 5-variable DF.

### 2.5.2 5-Variable DF Results

See Table 2.8 and equation 2.15 for the results of the DF constructed from the five selected parameters.

**Table 2.8 - Top 5 Discriminant Function Variables**

| Parameter* | Flare-Quiet AR Mean | Flare-Producing AR Mean | DF Standardized Coefficient |
|---|---|---|---|
| $\Phi_{tot}$ | $9.734\times10^{21}$ Mx | $2.315\times10^{22}$ Mx | 1.6270 |
| Total Area of $\psi \geq 80^{\circ}$ | 69 $Mm^2$ | 78 $Mm^2$ | -1.2051 |
| $\overline{\nabla B_z}$ | 90.609 G $Mm^{-1}$ | 101.901 G $Mm^{-1}$ | 0.4045 |
| $I_h$ | $3.55\times10^{11}$ A | $1.219\times10^{12}$ A | 0.5502 |
| $\kappa(\psi)$ | 4.346 | 4.636 | -0.3149 |
| DF Constant | --- | --- | 1.8269 |

* Parameters are listed in order of importance according to their corresponding standardized coefficients

$$f(\mathbf{X}) = 1.8269 + 1.6270(\Phi_{tot}) - 1.2051(\psi_{A>80^{\circ}}) + 0.4045(\overline{\nabla B_z}) + 0.5502(I_h) - 0.3149(\kappa(\psi)) \quad (2.19)$$

As discussed previously, flares historically are produced in regions of highly concentrated magnetic flux and increased field complexity. Total flux and magnetic shear are two of the most researched parameters for solar flare production. Thus, it was expected these two parameters would be among the top flare predictors. From Table 2.8 we see the average total flux for a flaring AR is over twice as large as that of a flare-quiet region. Active regions are by definition areas of concentrated magnetic flux, and total magnetic flux has often been a parameter used to describe an AR's size. Larger ARs in which the solar magnetic field has been highly concentrated have historically been flare-productive. Also, previous research has identified emerging magnetic flux as a possible flare trigger. Flux emergence may be

identified within an AR by an increase in the total magnetic flux. Thus, a larger total magnetic flux value may indicate an approach to a magnetic threshold and an AR more likely to produce a flare.

As expected, a shear parameter is also among the top predictors. The parameter is a specific measure of the total area within the AR with a horizontal magnetic shear angle greater than 80˚. Again the value of the parameter is larger for flaring ARs, which is consistent with flare theory. The relation made in previous research between larger areas of intense shearing and flare production is supported here; however, we see strong shearing is not a sufficient condition for flares as has been previously proposed. This parameter of shear alone is not a good predictor of solar flares. By itself, the parameter for the area of horizontal shear angle at least 80˚ did not predict a single AR to flare. However, it was the second parameter to be chosen in the step-up procedure, and when used in conjunction with other parameters, it had the ability to reduce the misclassification error rates and to increase the statistical separation between the flaring and flare-quiet populations.

We also see a larger kurtosis of the horizontal shear angle is an indication of a preflare state. This suggests the shear angle distribution for flaring ARs is more peaked near the population mean and the values of the shear angle for flare-producing ARs are concentrated near the population mean. Thus, there is less of a spread in the values of the shear angle among flaring ARs than among flare-quiet ARs. Furthermore, the kurtosis of a purely Gaussian distribution is equal to 0; thus, the value of its kurtosis may indicate the horizontal shear parameter distribution may be non-Gaussian.

The step-up procedure also associated flaring ARs with a larger gradient of the vertical magnetic field. This result suggests a more complex magnetic field with a larger spatially varying vertical component. This also supports the notion of emerging flux as a flare trigger. Emerging flux can lead to large gradients in the vertical field. One example of how emerging flux can lead to large gradients within an AR is flux emergence that introduces magnetic field lines of a polarity opposite to that of the surrounding area, effectively creating an island of magnetic polarity. These situations would lead to a complex magnetic field structure and a large vertical field gradient. Furthermore, a larger vertical field gradient would be present along neutral lines where flares often occur. As mentioned in §1.4 it is at locations such as near neutral lines where magnetic flux of opposite polarity exist and reconnection is highly likely. Thus, a larger vertical magnetic field gradient may identify those ARs in which there are locations where magnetic conditions are primed for flare activity.

Another parameter selected was the total current of heterogeneity, the component of the total current perpendicular to **B**. We see from our results the mean magnitude of the current of heterogeneity for flaring ARs is an order of magnitude larger than in flare-quiet ARs. As discussed in §1.10.4, currents perpendicular to **B** result in a Lorentz force which can add energy to the system. This availability of additional energy can increase the probability of a flare, as our results support.

Our confidence in DA for flare prediction is also strengthened with the absence of the control parameters, seeing and observation angle, from the top parameters selected. Significant reduction to the DF error rate stopped after about the sixth

variable was added in the step-up procedure. Seeing was not added until the 16th parameter, and observation angle was included as the 31st parameter.

## 2.6 Comparison to SEC Forecasts

To compare how well the DF flare probability forecasting method performs against present flare warning systems, we compared the forecast verification plot created from the results of our final 5-variable DF discussed in §2.5 to SEC's forecast verification plot (§1.7). The verification plots are shown in Figures 2.10 and 2.11. The $\chi^2$ value was the tool used for the comparison. Since we did not have the numerical data supporting SEC's forecast verification plot, we had to visually deduce the probability bin populations and the difference between the observed flare frequency and the frequency expected for 100% forecast accuracy. In order to keep as many factors as possible identical throughout the comparison of the two forecast methods, a visual inspection of the DF probability forecast was also done to determine the bin populations and forecast deviations. We constructed the DF probability forecast verification plot with the same size and number of bins as SEC's plot. See Appendix B for the data used to calculate the $\chi^2$ values for both forecast methods.

In our initial calculations of the $\chi^2$ value for the DF probability forecasts, each term in the calculation was weighted by the square-root of the number of ARs assigned to the corresponding bin. See equation 1.10. However, a weighted $\chi^2$ value is not the most appropriate tool for comparing the DF-base probability forecasts to

SEC's forecasts due to the gross differences in database size and the fact that some of the bins in the DF forecast verification plot lack a statistically significant number of datapoints. SEC's database contains approximately 6500 datapoints, while our training sample only contains 1182 datapoints. Thus, we consider $\chi^2$ values normalized by the sum of the weighting factors for our comparison (equation 2.20)

$$\chi^2 = \frac{\sum_{k=1}^{\eta} \left( (O_k - E_k)^2 \cdot \sqrt{W_k} \right)}{\sum_{k=1}^{\eta} \sqrt{W_k}} \tag{2.20}$$

If the $\chi^2$ values are normalized by the sum of the weighting factors (equation 2.20) and we only look at the forecast deviations from perfect accuracy, we obtain the unweighted $\chi^2$ value of 1.53 for SEC and 1.67 for the DF forecasts. From these values, it looks as if SEC's method performs slightly better than the DF forecasts; however, the lack of a statistically significant number of datapoints in the higher probability bins may contribute to errors in the DF forecasts. With a bin containing few datapoints, we are unable to say whether the bias of the observed frequency of flaring is due to a bias in the forecasting method or due to the lack of a statistically significant sample. The datapoints may be outliers and not representative of the true AR population described by the probability bin.

To avoid the statistically insignificant bins, we then calculated the $\chi^2$ value for only the 4 bins with the most datapoints (see Appendix B). The $\chi^2$ value for SEC's forecasts is now 0.16, and the $\chi^2$ value for the DF forecasts is 0.12. When we only consider statistically significant bins, the DF probability forecasts seem to perform better. However, several differences between the two forecasting methods may

contribute to the disparities in the forecasting methods. The daily flare forecasts published by SEC are for the entire solar disk and assign a probability of at least one flare occurring somewhere on the solar disk; whereas, our DF method assigns a probability of flaring to each AR present on the solar disk. Thus, the datapoints for each verification plot are defined in a slightly different manner. Furthermore, SEC produces separate forecasts for M- and X-class flares. The plot in Figure 2.10, which was used in the $\chi^2$ calculation for SEC's forecasts, was for M-class flare forecasts only. Presently, the DF forecasting method does not distinguish between flare classes when assigning a probability of flaring. Consequently, the DF forecast verification plot incorporates C-, M-, and X-class flares.

Overall, the results show that an objective DF based method for producing flare probability forecasts may perform as well as the present subjective method employed by SEC. To further explore how the methods compare, a more rigorous comparison should be made in which as many factors are equal or consistent throughout the comparison as possible. For example, it may be insightful to limit the DF forecast verification plot to only M-class flares and then make the comparison with SEC. Also, a much larger DF sample size is needed to be able to compare the two methods across all probability bins.

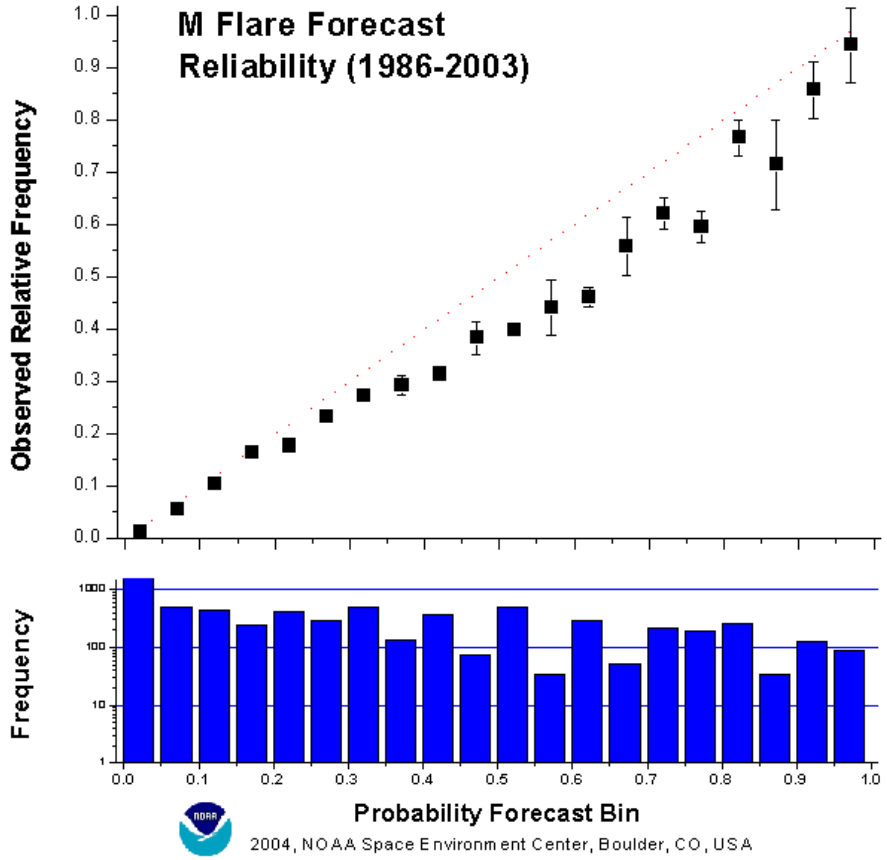


Figure 2.10 - SEC Flare Probability Forecasts Verification Plot
(http://www.sec.noaa.gov/forecast_verification/mFlare.html)

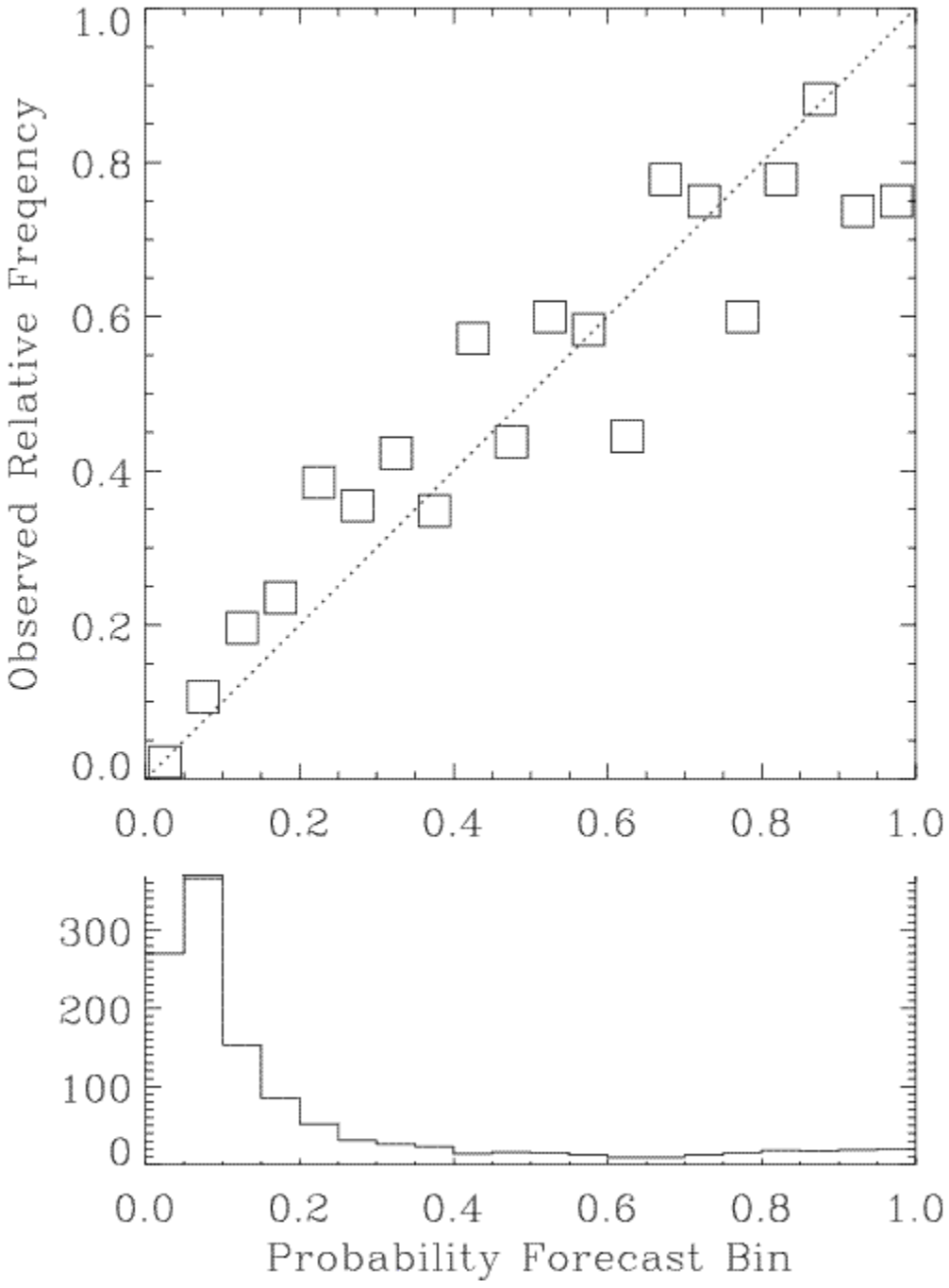


Figure 2.11 – 5-Variable DF Flare Probability Forecast Verification Plot

# 3. Discussion and Future Work

## 3.1 Forecast Versus Modeling Accuracy

In Sections 2.4 and 2.5, the final list of variables was selected to which the step-up procedure was applied in order to identify the optimum combination of parameters to use for the DF. The selection tool of choice used in the step-up procedure was the Mahalanobis distance due to its consistent results and its reproducibility. Even though we elected to use the Mahalanobis distance in place of the $\chi^2$ value, the $\chi^2$ value could have been used. The results would then need to be viewed in a different manner.

The Mahalanobis distance is a measure of the separation of the flaring and flare-quiet AR population means in *k*-space. Thus, this distance is a reflection of the discriminating power of the DF and the correctness of the AR population models and DA assumptions. The $\chi^2$ value, on the other hand, is a measure of flare probability forecast accuracy. If the goal is for forecast optimization and not DF optimization, then minimizing the $\chi^2$ value would be more useful to the forecaster since the quantity measures the deviation of the probability forecasts from actual observations. Variable selection using the $\chi^2$ value would then identify the combination of variables producing the most accurate probability forecast. This combination of variables would not necessarily result in the lowest AR misclassification rate or yield the most accurate DF. However, if the goal is to select variables that will optimize the DF and population discreteness, then maximizing the Mahalanobis distance may yield better

results. The DF alone produces only a binary flare forecast product, i.e. the AR will produce a flare or the AR will be flare-quiet. Increasing the relative separation of the flaring and flare-quiet AR populations may increase population discreteness and may decrease possible overlap in the populations in *k*-space if the in-group variances remain constant and do not increase. This in turn would increase the trustworthiness of the DF classification of ARs.

Another criterion for selecting variables during the step-up procedure that should be further investigated in future work is minimizing the rate of misclassification. The DF created from variables selected due to their ability to minimize instances of misclassification may or may not yield a forecast verification plot with the smallest $\chi^2$ chi-squared value. However, it is another method of selecting DF variables with the goal of optimizing forecast accuracy. We briefly explored this method of variable selection in our research, and the result of the step-up procedure with the criterion of minimizing the rate of misclassification was a DF composed of two variables, total magnetic flux and the kurtosis of the inclination angle. Due to time constraints and our greater confidence in the results of the step-up procedure when the criterion was minimizing the $\chi^2$ value or maximizing the Mahalanobis distance, we decided not to pursue minimizing misclassification rates as a criterion for the step-up procedure.

## 3.2 Variable Selection Methods

The most exhaustive and reliable method of identifying the combination of variables that would yield the most accurate DF is to calculate the DFs for every single-variable, 2-variable, 3-variable, 4-variable… all the way up to… every 147-variable permutation and select among those countless permutations the best combination of variables. However, even this method would not be ideal unless all of the variables were uncorrelated and all DA assumptions were valid. Nevertheless, the time demands of such an undertaking make it unrealistic for most research projects.

The task can be made more manageable, as we did in §2.4 and §2.5, by further reducing the variable list to only those parameters expected to be among the best predictors and to further reduce correlation among variables. One approach to reducing the variable list is to explore different measures of DF variable contribution, such as the F-Value, the Partial F-Value, and Discriminant Loadings [Dillon and Goldstein, 1984]. F-values are similar to the DF standardized weighting coefficients and ignore correlation among parameters; thus, they can be misleading. However, the Partial F-Value is less affected by variable interdependence. The Partial F-Value shows the separation provided by the variable of interest after adjusting for the other variables. Discriminant Loadings return a parameter's simple correlation with the DF in a univariate context. Thus, Discriminant Loadings reflect how a variable alone separates the groups disregarding the presence of the other variables. This is may not be desirable since it does not provide information on how variables perform jointly or how the performance of a single variable changes as other parameters are added or

dropped [Rencher, 2002]. As was shown previously, some variables, which alone may not be good predictors, can be present in some of the better performing multivariate DFs.

The variable contribution measures could be applied to a set of highly correlated parameters, such as the parameters derived from a given measure of shear, to see which measure performs best at selecting among interdependent variables. Once the contribution measure and best predictor from the subset is selected, the remaining list of variables or other subsets of interdependent variables could be subjected to the chosen contribution measure. This would, hopefully, reduce redundant variables and parameter correlation and eliminate poor predictors.

Another method of selecting DF parameters worth exploring is a step-down method, which is similar to the step-up method covered in §2.4.1. The step-down method would begin with all 147 available parameters and would then eliminate the one variable that either contributes least to the separation of the groups or to reducing the error rates. It would be interesting to see how the step-up and step-down results differ with respect to error rates and the ultimate predictors selected.

## 3.3 Parameter and Population Distributions

Two of the assumptions made for our development of the DA were 1)all candidate parameters had Gaussian distributions and 2)the flare-quiet and flare-producing populations had equal covariance matrixes. We know there are positive-

definite parameters that can never be truly Gaussian and are aware the covariance matrixes of flaring ARs and non-flaring ARs may not be equal. Better characterization of the distributions of the parameters and AR populations would lead to greater confidence in DF predictions and parameter selections.

An improvement in the model from which DF is created can only improve prediction accuracy. The present DF parameter selection may also change if all of the candidate variables are represented by the correct distribution. A parameter presently recognized as a poor predictor may have been categorized as such because its true distribution may be far from Gaussian. In the final parameter list used in the step-up procedure (§2.4), many of the parameters were restricted to absolute values; thus, their distributions were far from Gaussian. However, this does not mean the five variables selected for use in the DF are not good flare predictors. If non-Gaussian distributions are better able to describe parameter distributions and were applied to the appropriate variables during DA, we may see other parameters step up as better predictors of solar flares, and we may see a change in the number of parameters needed for an adequate DF increase or decrease. We were unable to discover a statistical test to gauge the uncertainty introduced to our results by the violation of the assumption that all of the DF parameters had Gaussian distributions. Identifying such a test would be helpful for future DA work.

## 3.4 Training Sample

The present data constituting the DA training sample covers the time frame of January 2001 to February 2003, which corresponds to the latter half of the 23rd solar cycle peak occurring at the end of 2000 and beginning of 2001. Future work is needed to expand the training sample data and to incorporate data from other parts of the solar cycle. This should be done to characterize solar magnetic parameters throughout a solar cycle and to see if there are solar cycle specific effects with respect to the parameter distributions. This will also identify any solar cycle forecast bias. Increasing the sample size will also improve the statistical significance of the probability forecast verification plots by providing additional data to those probability bins containing few datapoints.

## 3.5 Photosphere Versus Chromosphere

Thresholds for flare production have not been established nor has a unique combination of parameters able to specify an AR as flare-quiet or flare-producing been identified. The photosphere may not be force free and may not be the source for pre-flare magnetic signatures. The currents and magnetic field present in the photosphere may not reflect the magnetic conditions in the chromosphere or corona where reconnection and relaxation of the magnetic field lines occur and where we

know the magnetic fields are force-free. For flare-unique conditions, the chromosphere may be the next place to search.

## 3.6 Summary

We have demonstrated discriminant analysis as a means of predicting solar flares when applied to photospheric magnetic parameters. We have also shown the importance of a statistically significant training sample to the confidence in DA results. In order to reduce processing time and the interdependence of DF variables, we reduced the list of candidate photospheric magnetic parameters to those deemed most likely to contribute to solar flare production and eliminated redundant parameters and reduced the subset of highly correlated shear parameters. Although we have not identified a combination of parameters unique to flaring ARs, we have revealed those conditions necessary for flare production. Due to the population of flare-quiet ARs being much larger than the population of flare-producing ARs, a rate of correct classification of 0.777 can be achieved by predicting every AR to remain flare-quiet. However, the 5-variable DF shown in equation 2.19 obtains a rate of correct classification of 0.829.

We have also shown the performance of objective flare probability forecasts derived from a linear multivariable DF compares to that of the subjective flare forecasts produced by SEC. The performance of the DF-based probability forecasts could be improved with the use of more appropriate parameter and AR population distributions. The assumption of Gaussian parameter distributions and populations

with equal covariance matrixes is not always valid. Invalid assumptions may account for the limiting value of 0.829 for the rate of correct classification. Since less restrictive assumptions may lead to a non-linear DF, a quadratic DF may be better suited for the task of solar flare prediction. With a better model of solar parameters and AR populations it may be possible to identify those conditions and thresholds sufficient for flaring and to increase the rate of correct classification.

# List of Initial 147 Candidate Photospheric Magnetic Parameters

| | | | |
|---|---|---|---|
| Horizontal magnetic field | $\overline{B_h}$ | Excess magnetic energy density | $\overline{\varepsilon}$ |
| | $\sigma(B_h)$ | | $\sigma(\varepsilon)$ |
| | $\varsigma(B_h)$ | | $\varsigma(\varepsilon)$ |
| | $\kappa(B_h)$ | | $\kappa(\varepsilon)$ |
| | | Total excess magnetic energy | $E_{tot} = \sum \varepsilon \, dA$ |
| Vertical magnetic field | $\overline{B_z}$ | | |
| | $\sigma(B_z)$ | Horizontal gradient of the horizontal field | $\overline{\lvert\nabla_h B_h\rvert}$ |
| | $\varsigma(B_z)$ | | $\sigma(\lvert\nabla_h B_h\rvert)$ |
| | $\kappa(B_z)$ | | $\varsigma(\lvert\nabla_h B_h\rvert)$ |
| | $\overline{\lvert B_z\rvert}$ | | $\kappa(\lvert\nabla_h B_h\rvert)$ |
| | $\sigma(\lvert B_z\rvert)$ | Horizontal gradient of the vertical field | $\overline{\lvert\nabla_h B_z\rvert}$ |
| | $\varsigma(\lvert B_z\rvert)$ | | $\sigma(\lvert\nabla_h B_z\rvert)$ |
| | $\kappa(\lvert B_z\rvert)$ | | $\varsigma(\lvert\nabla_h B_z\rvert)$ |
| | | | $\kappa(\lvert\nabla_h B_z\rvert)$ |
| Total magnetic field | $\overline{B}$ | Horizontal gradient of the total field | $\overline{\lvert\nabla_h B\rvert}$ |
| | $\sigma(B)$ | | $\sigma(\lvert\nabla_h B\rvert)$ |
| | $\varsigma(B)$ | | $\varsigma(\lvert\nabla_h B\rvert)$ |
| | $\kappa(B)$ | | $\kappa(\lvert\nabla_h B\rvert)$ |

| | |
|---|---|
| Inclination angle | $\overline{\gamma}$ |
| | $\sigma(\gamma)$ |
| | $\varsigma(\gamma)$ |
| | $\kappa(\gamma)$ |
| Total magnetic flux associated with each magnetic polarity | $\Phi_{tot}^{+} = \sum B_z^{+}\, dA$ |
| | $\Phi_{tot}^{-} = \sum B_z^{-}\, dA$ |
| Total unsigned magnetic flux | $\Phi_{tot} = \sum \lvert B_z \rvert\, dA$ |
| Net magnetic flux | $\lvert \Phi_{net} \rvert = \lvert \sum B_z\, dA \rvert$ |
| | $\Phi_{net} = \sum B_z\, dA$ |
| Twist parameter | $\overline{\alpha}$ |
| | $\sigma(\alpha)$ |
| | $\varsigma(\alpha)$ |
| | $\kappa(\alpha)$ |
| | $\overline{\lvert \alpha \rvert}$ |
| | $\sigma(\lvert \alpha \rvert)$ |
| | $\varsigma(\lvert \alpha \rvert)$ |
| | $\kappa(\lvert \alpha \rvert)$ |
| Best-fit force-free twist parameter | $\overline{\lvert \alpha_{ff} \rvert}$ |
| | $\overline{\alpha_{ff}}$ |

| | |
|---|---|
| Helicity | $\overline{h_c}$ |
| | $\sigma(h_c)$ |
| | $\varsigma(h_c)$ |
| | $\kappa(h_c)$ |
| | $\overline{\lvert h_c \rvert}$ |
| | $\sigma(\lvert h_c \rvert)$ |
| | $\varsigma(\lvert h_c \rvert)$ |
| | $\kappa(\lvert h_c \rvert)$ |
| | $H_c^{tot} = \sum \lvert h_c \rvert\, dA$ |
| | $\lvert H_c^{net} \rvert = \lvert \sum h_c\, dA \rvert$ |
| | $H_c^{net} = \sum h_c\, dA$ |
| Vertical current of each sign | $I^{+} = \sum J_z^{+}\, dA$ |
| | $I^{-} = \sum J_z^{-}\, dA$ |
| | $I_{tot} = I^{+} + \lvert I^{-} \rvert$ |
| | $I_{net} = I^{+} + I^{-}$ |
| | $\lvert I_{net} \rvert = \lvert I^{+} + I^{-} \rvert$ |

Current of Chirality

$$\overline{J_c} = \overline{\left(\frac{B}{\mu_o}\nabla \times \hat{b}\right)}$$

$$\sigma(J_c)$$

$$\varsigma(J_c)$$

$$\kappa(J_c)$$

$$\overline{|J_c|}$$

$$\sigma(|J_c|)$$

$$\varsigma(|J_c|)$$

$$\kappa(J_c)$$

$$I_c^{tot} = \sum |J_c| \, dA$$

$$I_c^{net} = \sum J_c \, dA$$

$$\left|I_c^{net}\right| = \left|\sum J_c \, dA\right|$$

Vertical current density

$$\overline{J_z} = \overline{(J_h + J_c)}$$

$$\sigma(J_z)$$

$$\varsigma(J_z)$$

$$\kappa(J_z)$$

$$\overline{|J_z|}$$

$$\sigma(|J_z|)$$

$$\varsigma(|J_z|)$$

$$\kappa(|J_z|)$$

Vertical current associated with each magnetic polarity

$$I_{tot}^{B_z^+} = \sum \left|J_z^{B_z^+}\right| dA$$

$$I_{tot}^{B_z^-} = \sum \left|J_z^{B_z^-}\right| dA$$

$$I_{net}^{B_z^+} = \sum J_z^{B_z^+} \, dA$$

$$I_{net}^{B_z^-} = \sum J_z^{B_z^-} \, dA$$

$$\left|I_{net}^{B_z^+}\right| = \left|\sum J_z^{B_z^+} \, dA\right|$$

$$\left|I_{net}^{B_z^-}\right| = \left|\sum J_z^{B_z^-} \, dA\right|$$

$$I_{tot}^{B_z} = \left|I_{tot}^{B_z^+}\right| - \left|I_{tot}^{B_z^-}\right|$$

$$I_{net\_sum}^{B_z} = \left|I_{net}^{B_z^+}\right| + \left|I_{net}^{B_z^-}\right|$$

$$I_{net\_diff}^{B_z} = \left|I_{net}^{B_z^+}\right| - \left|I_{net}^{B_z^-}\right|$$

Current of heterogeneity

$$\overline{J_h} = \overline{\left(\frac{\nabla B}{\mu_o} \times \hat{b}\right)}$$

$$\sigma(J_h)$$

$$\varsigma(J_h)$$

$$\kappa(J_h)$$

$$\overline{|J_h|}$$

$$\sigma(|J_h|)$$

$$\varsigma(|J_h|)$$

$$\kappa(|J_h|)$$

$$I_h^{tot} = \sum |J_h| \, dA$$

$$I_h^{net} = \sum J_h \, dA$$

$$\left|I_h^{net}\right| = \left|\sum J_h \, dA\right|$$

| | | | |
|---|---|---|---|
| Horizontal shear angle | $\overline{\psi_H}$ | 3D shear angle restricted to neutral line | $\overline{\psi_{NL}}$ |
| | $\sigma(\psi_H)$ | | $\sigma(\psi_{NL})$ |
| | $\varsigma(\psi_H)$ | | $\varsigma(\psi_{NL})$ |
| | $\kappa(\psi_H)$ | | $\kappa(\psi_{NL})$ |
| | $\overline{\psi_H}$ weighted by $\lvert\mathbf{B}_h\rvert$ | | $\overline{\psi_{NL}}$ weighted by $\lvert\mathbf{B}_h\rvert$ |
| | $\sigma(\psi_H)$ weighted by $\lvert\mathbf{B}_h\rvert$ | | $\sigma(\psi_{NL})$ weighted by $\lvert\mathbf{B}_h\rvert$ |
| | Area within AR of $\psi_{3D} > 45°$ | | Area within AR of $\psi_{NL} > 45°$ |
| | Area within AR of $\psi_{3D} > 80°$ | | Area within AR of $\psi_{NL} > 80°$ |
| | Fraction of neutral line of $\psi_{3D} > 45°$ | | Fraction of neutral line of $\psi_{NL} > 45°$ |
| | Fraction of neutral line of $\psi_{3D} > 80°$ | | Fraction of neutral line of $\psi_{NL} > 80°$ |

| 3D Shear Angle | $\overline{\psi_{3D}}$ | Horizontal shear angle restricted to neutral line | $\overline{\psi_H}$ |
|---|---|---|---|
| | $\sigma(\psi_{3D})$ | | $\sigma(\psi_H)$ |
| | $\varsigma(\psi_{3D})$ | | $\varsigma(\psi_H)$ |
| | $\kappa(\psi_{3D})$ | | $\kappa(\psi_H)$ |
| | $\overline{\psi_{3D}}$ weighted by $\lvert\mathbf{B}_h\rvert$ | | $\overline{\psi_H}$ weighted by $\lvert\mathbf{B}_h\rvert$ |
| | $\sigma(\psi_{3D})$ weighted by $\lvert\mathbf{B}_h\rvert$ | | $\sigma(\psi_H)$ weighted by $\lvert\mathbf{B}_h\rvert$ |
| | Area within AR of $\psi_{3D} > 45°$ | | Area within AR of $\psi_{3D} > 45°$ |
| | Area within AR of $\psi_{3D} > 80°$ | | Area within AR of $\psi_{3D} > 80°$ |
| | Fraction of neutral line of $\psi_{3D} > 45°$ | | Fraction of neutral line of $\psi_{3D} > 45°$ |
| | Fraction of neutral line of $\psi_{3D} > 80°$ | | Fraction of neutral line of $\psi_{3D} > 80°$ |

- $\hat{b}$ is the unit vector in direction of the field

| Final Reduced List of Candidate DF Parameters | | |
|---|---|---|
| 1. Observation angle | 24. $\overline{\lvert\nabla_h B\rvert}$ | 47. $\overline{\lvert h_c\rvert}$ |
| 2. Seeing Parameter | 25. $\sigma(\lvert\nabla_h B\rvert)$ | 48. $\sigma(\lvert h_c\rvert)$ |
| 3. $\Phi_{tot}$ | 26. $\varsigma(\lvert\nabla_h B\rvert)$ | 49. $\varsigma(\lvert h_c\rvert)$ |
| 4. $\overline{B_h}$ | 27. $\kappa(\lvert\nabla_h B\rvert)$ | 50. $\kappa(\lvert h_c\rvert)$ |
| 5. $\sigma(B_h)$ | 28. $\overline{J_h}$ | 51. $\overline{\gamma}$ |
| 6. $\varsigma(B_h)$ | 29. $\sigma(J_h)$ | 52. $\sigma(\gamma)$ |
| 7. $\kappa(B_h)$ | 30. $\varsigma(J_h)$ | 53. $\varsigma(\gamma)$ |
| 8. $\overline{\lvert B_z\rvert}$ | 31. $\kappa(J_h)$ | 54. $\kappa(\gamma)$ |
| 9. $\sigma(\lvert B_z\rvert)$ | 32. $I_h^{tot}$ | 55. $\overline{\varepsilon}$ |
| 10. $\varsigma(\lvert B_z\rvert)$ | 33. $I_h^{net}$ | 56. $\sigma(\varepsilon)$ |
| 11. $\kappa(\lvert B_z\rvert)$ | 34. $\overline{J_z}$ | 57. $\varsigma(\varepsilon)$ |
| 12. $\overline{B}$ | 35. $\sigma(J_z)$ | 58. $\kappa(\varepsilon)$ |
| 13. $\sigma(B)$ | 36. $\varsigma(J_z)$ | 59. $E_{tot}$ |
| 14. $\varsigma(B)$ | 37. $\kappa(J_z)$ | 60. $\overline{\psi_H}$ |
| 15. $\kappa(B)$ | 38. $\mathrm{I_{tot}}$ | 61. $\sigma(\psi_H)$ |
| 16. $\overline{\lvert\nabla_h B_h\rvert}$ | 39. $\mathrm{I_{net}}$ | 62. $\varsigma(\psi_H)$ |
| 17. $\sigma(\lvert\nabla_h B_h\rvert)$ | 40. $\overline{\lvert\alpha\rvert}$ | 63. $\kappa(\psi_H)$ |
| 18. $\varsigma(\lvert\nabla_h B_h\rvert)$ | 41. $\sigma(\lvert\alpha\rvert)$ | 64. $\overline{\psi_H}$ weighted by $\lvert\mathbf{B}\rvert$ |
| 19. $\kappa(\lvert\nabla_h B_h\rvert)$ | 42. $\varsigma(\lvert\alpha\rvert)$ | 65. $\sigma(\psi_H)$ weighted by $\lvert\mathbf{B}\rvert$ |
| 20. $\overline{\lvert\nabla_h B_z\rvert}$ | 43. $\kappa(\lvert\alpha\rvert)$ | 66. Area within AR of $\psi_{3D} > 45°$ |
| 21. $\sigma(\lvert\nabla_h B_z\rvert)$ | 44. $\overline{\lvert\alpha_{ff}\rvert}$ | 67. Area within AR of $\psi_{3D} > 80°$ |
| 22. $\varsigma(\lvert\nabla_h B_z\rvert)$ | 45. $H_c^{tot} = \sum\lvert h_c\rvert dA$ | 68. Fraction of neutral line of $\psi_{3D} > 45°$ |
| 23. $\kappa(\lvert\nabla_h B_z\rvert)$ | 46. $\lvert H_c^{net}\rvert = \lvert\sum h_c\, dA\rvert$ | 69. Fraction of neutral line of $\psi_{3D} > 80°$ |

Appendix B: Flare Forecast Verification Visual $\chi^2$ Calculations Data

**SEC FORECASTS**

| Probability Bin | Bin Population | Expected minus Observed | Weighted Chi-Squared | |
|---|---|---|---|---|
| 0-5 | 1300 | 0 | 0.000E+00 | |
| 5-10 | 500 | 0 | 0.000E+00 | |
| 10-15 | 450 | 0 | 0.000E+00 | |
| 15-20 | 250 | 0 | 0.000E+00 | |
| 20-25 | 450 | 0.03 | 1.909E-02 | |
| 25-30 | 300 | 0.03 | 1.559E-02 | |
| 30-35 | 500 | 0.04 | 3.578E-02 | |
| 35-40 | 150 | 0.17 | 3.540E-01 | |
| 40-45 | 400 | 0.11 | 2.420E-01 | |
| 45-50 | 80 | 0.08 | 5.724E-02 | |
| 50-55 | 500 | 0.12 | 3.220E-01 | |
| 55-60 | 35 | 0.13 | 9.998E-02 | |
| 60-65 | 300 | 0.16 | 4.434E-01 | |
| 65-70 | 55 | 0.11 | 8.974E-02 | |
| 70-75 | 250 | 0.1 | 1.581E-01 | |
| 75-80 | 200 | 0.17 | 4.087E-01 | |
| 80-85 | 300 | 0.05 | 4.330E-02 | |
| 85-90 | 40 | 0.15 | 1.423E-01 | |
| 90-95 | 150 | 0.06 | 4.409E-02 | |
| 95-100 | 90 | 0.02 | 3.795E-03 | |
| | 6300 | 1.53 | 2.479079498 | <--- Sums from columns |

**SEC's 4 MOST POPULATED BINS**

| Probability Bin | Bin Population | Expected minus Observed | Weighted Chi-Squared |
|---|---|---|---|
| 0-5 | 1300 | 0 | 0.000E+00 |
| 5-10 | 500 | 0 | 0.000E+00 |
| 30-35 | 500 | 0.04 | 3.578E-02 |
| 50-55 | 500 | 0.12 | 3.220E-01 |
| Column Sums--> | 2800 | 0.16 | 3.578E-01 |

## DF PROBABILITY FORECASTS

| Probability Bin | Bin Population | Expected minus Observed | Weighted Chi-Squared | |
|---|---|---|---|---|
| 0-5 | 270 | 0 | 0.000E+00 | |
| 5-10 | 370 | 0.02 | 7.694E-03 | |
| 10-15 | 150 | 0.05 | 3.062E-02 | |
| 15-20 | 80 | 0.05 | 2.236E-02 | |
| 20-25 | 50 | 0.15 | 1.591E-01 | |
| 25-30 | 30 | 0.07 | 2.684E-02 | |
| 30-35 | 30 | 0.1 | 5.477E-02 | |
| 35-40 | 30 | 0.02 | 2.191E-03 | |
| 40-45 | 20 | 0.15 | 1.006E-01 | |
| 45-50 | 20 | 0.03 | 4.025E-03 | |
| 50-55 | 20 | 0.07 | 2.191E-02 | |
| 55-60 | 15 | 0 | 0.000E+00 | |
| 60-65 | 10 | 0.17 | 9.139E-02 | |
| 65-70 | 10 | 0.1 | 3.162E-02 | |
| 70-75 | 15 | 0.02 | 1.549E-03 | |
| 75-80 | 15 | 0.17 | 1.119E-01 | |
| 80-85 | 20 | 0.05 | 1.118E-02 | |
| 85-90 | 20 | 0 | 0.000E+00 | |
| 90-95 | 20 | 0.2 | 1.789E-01 | |
| 95-100 | 20 | 0.25 | 2.795E-01 | |
| | 1215 | 1.67 | 1.136200768 | <--- Sums from columns |

## DF's 4 MOST POPULATED BINS

| Probability Bin | Bin Population | Expected minus Observed | Weighted Chi-Squared |
|---|---|---|---|
| 0-5 | 270 | 0 | 0.000E+00 |
| 5-10 | 370 | 0.02 | 7.694E-03 |
| 10-15 | 150 | 0.05 | 3.062E-02 |
| 15-20 | 80 | 0.05 | 2.236E-02 |
| Column Sums--> | 870 | 0.12 | 6.067E-02 |

Bibliography

55th Space Weather Squadron (55th SWXS). "The Active Sun." Space Environment Forecaster Course, Air Force Space Command Course 15W3A-001, June 1997.

Bao, S.D., H.Q. Zhang, G.X. Ai, and M. Zhang. "A Survey of Flares and Current Helicity in Active Regions." *Astronomy and Astrophysics* 1999: 139: 311-320.

Canfield, Richard C., J. -F. de La Beaujardiere, Yuhong Fan, K.D. Leka, A.N. McClymont, Thomas R. Metcalf, Donald L. Mickey, Jean-Piere Wuelser, and Bruce W. Lites. "The Morphology of Flare Phenomena, Magnetic Fields, and Electric Currents in Active Regions. I: Introduction and Methods." *The Astrophysics Journal* 1993: 411: 362-369.

Carlowicz, Michael J., and Ramon E. Lopez. Storms from the Sun. Washington D.C.: The Joseph Henry Press, 2002.

Carroll, Bradley, and Dale Ostlie. An Introduction to Modern Astrophysics. New York: Weber State University, 1996.

Demoulin, P., C. H. Mandrini, L. Van Driel-Gesztelyi, M. C. Lopez Fuentes, and G. Aulanier. "The Magnetic Helicity Injected by Shearing Motions." *Solar Physics* 2002: 207: 87-110.

Dillon, William, and Matthew Goldstein. Multivariate Analysis: Methods and Applications. New York: The City University of New York, 1984.

Doggett, Kent A. "Flare-forecasting Questions." E-mail to KD Leka. 16 Nov 2004.

Gallagher, Peter. "Flare Prediction System." Excerpt from the NASA Goddard Space Flight Center's Solar Data Analysis Center Active Region Monitor 2.0 webpages. n.peg. http://beauty.nascom.nasa.gov/arm/latest/forecast.html. February 23, 2005.

Holder, Zachary, Richard C. Canfield, Rebecca A. McMullen, Dibyendu Nandy, Robert F. Howard, and Alexei A. Pevtsov. “On the Tilt and Twist of Solar Active Regions.” *The Astrophysics Journal* 20 Aug. 2004: 611: 1149-1155.

Hudson, H. S. “Solar Flares, Microflare, Nanoflares, and Coronal Heating.” *Solar Physics* 1991: 133: 357.

Leka, K.D., and G. Barnes. “Photospheric Magnetic Field Properties of Flaring Versus Flare-Quiet Active Regions I: Data, General Approach, and Sample Results.” *The Astrophysical Journal* 1 Oct. 2003a: 595: 1277-1295.

Leka, K.D., and G. Barnes. “Photospheric Magnetic Field Properties of Flaring Versus Flare-Quiet Active Regions II: Discriminant Analysis.” *The Astrophysical Journal* 1 Oct. 2003b: 595: 1296-1306.

Leka, K.D., and G. Barnes. “Photospheric Magnetic Field Properties of Flaring Versus Flare-Quiet Active Regions III: Discriminant Analysis of a Statistically Significant Database.” Private Communication: 2003c.

Li, Hui, Takashi Sakurai, Kiyoshi Ichimoto, and Satoru UeNe. “Magnetic Field Evolution Leading to Solar Flares II. Cases with Low Magnetic Shear and Flare-Related Shear Change.” *Publications of the Astronomical Society of Japan* 2000: 52: 483-497.

Ohanian, Hans C. Modern Physics: Second Edition. Up Saddle River: Prentice Hall, 1995.

Phillips, Kenneth J.H. Guide to the Sun. New York: Cambridge University Press, 1995.

Radel, Stanley R., and Marjorie H. Navidi. Chemistry: Second Edition. St. Paul: West Pubilishing Company, 1994.

Rees, W. G. Physical Principles of Remote Sensing: Second Edition. Cambridge: Cambridge University Press, 2001.

Rencher, Alvin C. Methods of Multivariate Analysis. New York: John Wiley and Sons, Inc., 2002.

Smith, Jesse B., Jr., Donald F. Neidig, Philip H. Wiborg, Edward A. West, Mona J. Hagyard, Mitzi Adams, and Paul H. Seagraves. "An Objective Test of Magnetic Shear as a Flare Predictor." Astronomical Society of the Pacific Conference Series: Solar Drivers of Interplanetary and Terrestrial Disturbances, Vol. 95: 1996.

Song, Paul, Howard Singer, and George Siscoe, ed. Space Weather. Washington D.C.: American Geophysical Union, 2001.

Sturrock, P.A., T.E. Holzer, D.M. Mihalas, and R.K. Ulrich. Physics of the Sun, Volume 1: The Solar Interior. Dordrecht: D. Riedel Publishing Company, 1986.

Tascione, Thomas F. Introduction to the Space Environment: Second Edition. Malabar: Krieger Publishing Company, 1994.

Timm, Neil H. Applied Multivariate Analysis. New York:Springer-Verlag, 2002.

Taylor, John R. An Introduction to Error Analysis. Sausalito: University Science Books, 1997.

Wheatland, M.S. "Rates of Flaring in Individual Active Regions." Solar Physics, Oct 2001, Vol 203: 87-106.

Wheatland, M.S. "A Bayesian Approach to Solar Flare Prediction." The Astrophysical Journal, July 2004, Vol 609: 1134-1139.

Zhang, Chang-Xi, G. B. Gelfreikh, and Jing-Xiu Wang. “Magnetic Field Strengths and Structures from Radio Observations of Solar Active Regions.” Chinese Journal of Astronomy and Astrophysics 2002: Vol 2: 266-276.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 074-0188*



| **1. REPORT DATE** *(DD-MM-YYYY)* | **2. REPORT TYPE** | **3. DATES COVERED** *(From – To)* |
|---|---|---|
| March 2005 | **Master's Thesis** | Jun 2004 – Feb 2005 |

**4. TITLE AND SUBTITLE**

CUSTOMIZATION OF DISCRIMINANT FUNCTION ANALYSIS FOR PREDICTION OF SOLAR FLARES

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Schumer, Evelyn A, Captain, USAF

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)**
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/GAP/ENP/05-07

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

**10. SPONSOR/MONITOR'S ACRONYM(S)**

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

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

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
This research is an extension to the research conducted by K. Leka and G. Barnes of the Colorado Research Associates Division, Northwest Research Associates, Inc. in Boulder, Colorado (CORA) in which they established no single photospheric solar parameter could sufficiently identify a flare-producing active region (AR). Their research then explored the possibility a linear combination of parameters used in a multivariable discriminant function (DF) could adequately predict solar activity.

The purpose of this research is to extend the DF research conducted by Leka and Barnes by refining the method of statistical discriminant analysis (DA) with the goal of selecting those photospheric magnetic parameters most capable of identifying flare-producing active regions in hopes of increasing the reliability of short term flare warnings and the understanding of flare production. The data for this research were photospheric vector magnetograms captured by the Imaging Vector Magnetograph (IVM) at the University of Hawai`i Mees Solar Observatory at Haleakala and provided by CORA. Increasing the data set size was an essential task for this research in order to have a more statistically significant training sample for DA. This research also modified current DF procedures to enable the customization of the costs of flare false alarms and flare misses. Work was also done to expand the binary DF results to produce flare probability forecasts. The selection of the optimum combination of photospheric magnetic parameters to be used as predictors in a linear DF began with the elimination of redundant parameters and those parameters least likely to contribute to flare production. The selection of parameters was governed by maximizing the Mahalanobis distance in a step-up method. The DF results show a pre-flaring active region may be characterized by a larger magnetic field, active region with a larger area of magnetic shear angle greater than 80˚, larger current of heterogeneity, larger spatial vertical magnetic field gradient, and a larger kurtosis of the shear angle.

With the optimum combination of parameters, DF flare probability forecasts were compared to the daily forecasts produced by the National Oceanic and Atmospheric Administration, Space Environment Center (NOAA SEC). The Chi-Squared values of each forecast show the objective DF based flare probability forecasting method performs as well as the subjective forecasting method employed by the SEC, and may perform better with a more statistically significant dataset.

**15. SUBJECT TERMS**
Solar Flares, Flare Prediction, Discriminant Analysis, Flare Forecasting, Flare Predictors, Probability Forecasting

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| **REPORT** | **ABSTRACT** | **c. THIS PAGE** | | | **Della-Rose, Devin, Major, USAF (ENP)** |
| **U** | **U** | **U** | **UU** | 110 | **19b. TELEPHONE NUMBER** *(Include area code)* (937)255-3636 x4514 devin.della-rose@afit.edu |