Draft version September 29, 2022
Typeset using LaTeX default style in AASTeX631



# Solar Flare Index Prediction Using SDO/HMI Vector Magnetic Data Products with Statistical and Machine Learning Methods

Hewei Zhang,[1,2] Qin Li,[3,2] Yanxing Yang,[4] Ju Jing,[3,2] Jason T.L. Wang,[5,2] Haimin Wang,[3,2] and Zuofeng Shang[1,2]

[1] *Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ, 07102-1982, USA*
[2] *Institute for Space Weather Sciences, New Jersey Institute of Technology, Newark, NJ, 07102-1982, USA*
[3] *Big Bear Solar Observatory, New Jersey Institute of Technology, 40386 North Shore Lane, Big Bear, CA 92314*
[4] *Department of Physics, New Jersey Institute of Technology, Newark, NJ, 07102-1982, USA*
[5] *Department of Computer Science, New Jersey Institute of Technology, Newark, NJ, 07102-1982, USA*

## ABSTRACT

Solar flares, especially the M- and X-class flares, are often associated with coronal mass ejections (CMEs). They are the most important sources of space weather effects, that can severely impact the near-Earth environment. Thus it is essential to forecast flares (especially the M-and X-class ones) to mitigate their destructive and hazardous consequences. Here, we introduce several statistical and Machine Learning approaches to the prediction of the AR's Flare Index (FI) that quantifies the flare productivity of an AR by taking into account the numbers of different class flares within a certain time interval. Specifically, our sample includes 563 ARs appeared on solar disk from May 2010 to Dec 2017. The 25 magnetic parameters, provided by the Space-weather HMI Active Region Patches (SHARP) from Helioseismic and Magnetic Imager (HMI) on board the Solar Dynamics Observatory (SDO), characterize coronal magnetic energy stored in ARs by proxy and are used as the predictors. We investigate the relationship between these SHARP parameters and the FI of ARs with a machine-learning algorithm (spline regression) and the resampling method (Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise, short by SMOGN). Based on the established relationship, we are able to predict the value of FIs for a given AR within the next 1-day period. Compared with other 4 popular machine learning algorithms, our methods improve the accuracy of FI prediction, especially for large FI. In addition, we sort the importance of SHARP parameters by Borda Count method calculated from the ranks that are rendered by 9 different machine learning methods.

*Keywords:* Power Transformation — Spline Regression — SMOGN — Feature Ranking — Sun Flares

## 1. INTRODUCTION

Solar flares are enhanced emissions across the electromagnetic spectrum that occurs on a minute-to-hour time frame (Benz 2017). The frequency of solar flare varies with the solar cycle, accounting for an increasing frequency every 11 years. Small flares usually only arouse an increase in intensities of visible light, soft X-ray and radio wave, releasing energy in a magnitude of $10^{28} \sim 10^{29}$ erg. However, a major (M-class or X-class) flare eruption can release energy as much as $4\times10^{32}$ erg, of which electromagnetic radiation only accounts for 1/4. The rest energy is released by high energy particles in the form of plasma cloud moving at a speed of 1500 km·s$^{-1}$. In particular, solar flares and the often associated coronal mass ejections (CMEs) are the drivers of "space weather". Geomagnetic storms caused by them can disrupt or damage spacecraft, harm astronauts and high-altitude pilots, interrupt communications and navigation systems, and even shut down portions of the electric transmission system when they reach Earth (Schwenn 2006). Therefore, reliable prediction of flares, especially large flares, is a very urgent task.

Solar flare usually originates from active region (AR) which is an area with enhanced magnetic fields on the Sun. Although the triggering mechanism of flares still remains elusive, it is certain that the free magnetic energy available to power a flare is stored in the host AR. Generally speaking, the more the AR's magnetic field deviates from a simple potential configuration, the more free magnetic energy is stored in the AR and the more likely it is that a flare will

occur in the AR. However, the coronal magnetic fields cannot be precisely measured for the time being and, most likely, in the foreseeable future. Therefore the flare forecasting efforts have been made almost exclusively with magnetic parameters which are derived from the photospheric magnetic fields (Toriumi & Wang 2019).

In the past few decades, traditional statistical methods have been extensively used to predict solar flares. Many statistical studies have focused on the distribution of solar flare intervals (Pearce et al. 1993; Boffetta et al. 1999; Lepreti et al. 2001; Kubo 2008; Wheatland 2010). Some statistical models are based on the morphological characteristics of flare (Gallagher et al. 2002; Wheatland 2004, 2005; Stepanov et al. 2004; Zharkov & Zharkova 2006; Barnes et al. 2007; Yu et al. 2009; Song et al. 2009; Bloomfield et al. 2012; McCloskey et al. 2016; Leka et al. 2018). For example, (Gallagher et al. 2002) and (Bloomfield et al. 2012) adopt the Poisson model to estimate the probability of solar flares. (Song et al. 2009) predict the probability of flares in each AR during the next 24 hrs by using the ordinal logistic regression method. However, the prediction accuracy of these traditional statistical methods for flares, especially large flares, is not satisfactory.

With the explosion of digital data, the human processing ability is being challenged. Machine learning methods, as a state-of-the-art tool, has attracted more and more attention. Several machine-learning algorithms have been applied to the flare prediction problem: Support Vector Machine (Li et al. 2007; Qahwaji & Colak 2007; Yuan et al. 2010; Bobra & Couvidat 2015a; Muranushi et al. 2015; Nishizuka et al. 2017), Random Forests (Liu et al. 2017; Florios et al. 2018; Wang et al. 2019a; Hazra et al. 2020), k-nearest neighbor (Li et al. 2008; Huang et al. 2013; Winter & Balasubramaniam 2015; Hazra et al. 2020), LASSO (Jonas et al. 2018; Benvenuto et al. 2018), Decision Trees (Yu et al. 2009, 2010), and LSTM (Liu et al. 2019a; Chen et al. 2019; Jiao et al. 2020; Wang et al. 2020; Sun et al. 2021), CNN (Park et al. 2018; Huang et al. 2018; Zheng et al. 2019; Zhu et al. 2020; Sun et al. 2021). These previous studies for forecasting solar eruptions using machine learning have yielded valuable results and have played an active role in promoting space weather forecasting. However, there are still significant shortcomings in the current research. The main problem in this field is that the data sample is imbalanced, which leads to poor prediction of large flares. Shin et al. (2016) adopt multiple linear regression and artificial neural network methods to forecast the maximum flare flux for strong flares. Kusano et al. (2020) present a physics-based model to predict large solar flares through a critical condition of magneto-hydrodynamic instability, triggered by magnetic reconnection. In this paper, we develop Spline Regression model (De Boor & Höllig 1982) to forecast the quantitative flare, and then compare and analyze prediction performance with other popular prediction models, i.e. Linear regression, Random Forest, LASSO and Gaussian Process Regression. In addition, to solve the problem of data imbalance, we adopt an advanced resampling method, SMOGN (Branco et al. 2017), to improve the prediction accuracy of large flares.

Since 2010 May, the *Helioseismic and Magnetic Imager* (HMI, Schou et al. 2012) on board the *Solar Dynamics Observatory* (SDO, Pesnell et al. 2011) has been continuously producing full-disk photospheric vector magnetograms at a 12-minute cadence. In particular, the Space-weather HMI Active Region Patches (SHARPs, Bobra et al. 2014; Bobra & Couvidat 2015b) released by the SDO-HMI team include automatically identified and tracked ARs in map patches and contain several essential magnetic parameters. Since then, these SHARP parameters, as the predictors, have been widely used in many flare prediction studies (Liu et al. 2017; Jonas et al. 2018; Liu et al. 2019a; Chen et al. 2019; Wang et al. 2019b; Jiao et al. 2020; Wang et al. 2020; Sun et al. 2021).

The goal of our research is to predict the flare productivity of a given AR, quantified by flare index (FIs, Jing et al. 2006; Song et al. 2009; Jiao et al. 2020), in the next 1-day period using 25 SHARPs parameters. To this end, we explore the relationship between FI and SHARP parameters as follows. First, we apply a popular family of power transformations for achieving approximate normality, i.e., Yeo-Johnson Power Transformations (Yeo & Johnson 2000). Then we test statistical significance and linearity for each SHARP parameter, and perform an exhaustive sieving method to exclude some highly correlated features from our model. Next, based on the feature selection result, we apply the spline regression technique to establish the relationship between the selected SHARP parameters and FI, from which we're able to predict the value of FI for a given AR. To address the data imbalance problem, we adopt an advanced resampling method, SMOGN, to improve the prediction accuracy for large FI. Finally, we compare the performance of four popular machine learning methods (i.e. Linear Reression, LASSO, Random Forest (RF) and Gaussian Process Regression (GPR)) with our method. In addition, to better understand the contribution of these SHARP parameters on the FI prediction, we sort the importance of SHARP parameters on FI prediction by Borda Count score calculated from the ranks that are rendered by 9 different machine learning methods.

The paper is organized in the following manner. Section 2 includes data preparation (Section 2.1) and feature selection (Section 2.3) for regression tasks. Section 3 introduces the prediction algorithm (Section 3.1) and the solu-

tion (Section 3.2) to the problem of data imbalance. Section 4 reports magnetic parameter ranking results and the corresponding physical explanation. The conclusions and future works are presented in the Section 5.

## 2. DATA PREPARATION

### 2.1. *Details of Data*

The overall flare productivity of a given AR has been quantified by soft X-ray (SXR) Flare Index (FI) (Antalova 1996; Abramenko 2005; Jing et al. 2006; Song et al. 2009; Jing et al. 2010). Specifically, flares, from weak to strong, are classified as B, C, M or X according to their peak SXR flux (of $10^{-7}$, $10^{-6}$, $10^{-5}$ and $10^{-4}$ W m$^{-2}$ magnitude order, respectively), as measured by the Geostationary Operational Environmental Satellite (GOES). FI is calculated by weighting the GOES SXR flares classes of B, C, M, and X as 0.1, 1, 10, 100, respectively, within a certain time window $\tau$, i.e.,

$$FI = \ 0.1 \times \sum_{\tau} I_{\rm B} + 1 \times \sum_{\tau} I_{\rm C} + 10 \times \sum_{\tau} I_{\rm M} + 100 \times \sum_{\tau} I_{\rm X} \ ,$$

where $\rm I_B$, $\rm I_C$, $\rm I_M$, and $I_X$ are GOES peak intensities of B-, C-, M-, and X-class flares produced by the given AR over the period $\tau$, and $\tau$ is selected to be 1 day in this study to account for the flare production generated from an AR on the solar disk during a day starting from 0:00 Universal Time (UT).

Our sample includes 563 ARs from May 2010 to Dec 2017. For each AR, we calculate the FI of these ARs for each day during their disk passage. Thus a total of 2504 FIs of 563 ARs are acquired. As shown in Table 1, the data are categorized as four groups according to the magnitude of FI, i.e., $FI < 1; 1 \leq FI < 10; 10 \leq FI < 100; FI \geq 100$, equivalent to a daily average of a B-class or less, a C-class, a M-class and a X-class flare, respectively. Forecasting strong flares (M- or X-class) is particularly important because of their space weather effects. In our data set, 19% FIs are greater than 10, and only 2% FIs are greater than 100. In this work, we emphasize to improving the predictive accuracy of FI greater than 10.

SDO/HMI team has been releasing Space-weather HMI Active Region Patches (SHARP) since 2012 (Bobra et al. 2014), which can be found at the Joint Science Operations Center (JSOC) website. The 25 SHARP parameters of ARs are used as the FI predictors in this work, and are listed in Table 2. These SHARP parameters characterize physical properties of the AR, and are generally classified as intensive (spatial averages, e.g., MEANPOT), or extensive (summations or integrations, e.g., TOTPOT) measures (Welsch et al. 2009).

**Table 1.** The number of Flare Indices of 4 groups From 1 May 2010 to 28 Dec 2017 in the sample

| Group/Year | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | Total |
|---|---|---|---|---|---|---|---|---|---|
| FI<1 | 91 | 100 | 85 | 91 | 22 | 74 | 144 | 77 | 684 |
| 1<=FI<10 | 51 | 202 | 234 | 262 | 231 | 214 | 93 | 60 | 1347 |
| 10<=FI<100 | 5 | 60 | 71 | 82 | 95 | 79 | 12 | 12 | 416 |
| FI>=100 | 0 | 8 | 11 | 6 | 19 | 7 | 0 | 6 | 57 |
| Total | 147 | 370 | 401 | 441 | 367 | 374 | 249 | 255 | 2504 |

### 2.2. *Data Transformation*

Classic spline regression technique requires model error being Gaussian (Wahba 1990). The genuine data on solar flares, on the other hand, shows a lot of non-Gaussianity. As a result, data must be preprocessed before using the spline regression approach. In this paper, we apply the power transformation method (Box & Cox 1964), which is widely used in the statistical literature. SHARP parameters do not exhibit any explicit FI dependence in the original data — see Figure 1 (blue), whereas certain correlations between them emerge after the processing by power transformation — see Figure 1 (red), providing the possibility to predict solar flares by SHARP parameters.

One-parameter Box-Cox (Box & Cox 1964) and two-parameter Yeo-Johnson transformations (Yeo & Johnson 2000) are the two branches that are widely used in the implementation of power transformation. Particularly, Box-Cox family includes two different types of transformation based on the selection of $\lambda$:

$$y^{(\lambda)} = \begin{cases} \left(y^{\lambda} - 1\right)/\lambda & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \tag{1}$$

**Table 2.** 25 SDO/HMI Magnetic Parameters

| Keyword | Description |
|---|---|
| TOTUSJH | Total unsigned current helicity |
| TOTBSQ | Total magnitude of Lorentz force |
| TOTPOT | Total photospheric magnetic free energy density |
| TOTUSJZ | Total unsigned vertical current |
| ABSNJZH | Absolute value of the net current helicity |
| SAVNCPP | Sum of the modulus of the net current per polarity |
| USFLUX | Total unsigned flux |
| AREA_ACR | Area of strong field pixels in the active region |
| MEANPOT | Mean photospheric magnetic free energy |
| R_VALUE | Sum of flux near polarity inversion line |
| SHRGT45 | Fraction of area with shear $> 45^{\circ}$ |
| MEANSHR | Mean shear angle |
| MEANGAM | Mean angle of field from radial |
| MEANGBT | Mean gradient of total field |
| MEANGBZ | Mean gradient of vertical field |
| MEANGBH | Mean gradient of horizontal field |
| MEANJZH | Mean current helicity |
| MEANJZD | Mean vertical current density |
| MEANALP | Mean characteristic twist parameter, $\alpha$ |
| TOTFX | Sum of x-component of Lorentz force |
| TOTFY | Sum of y-component of Lorentz force |
| TOTFZ | Sum of z-component of Lorentz force |
| EPSX | Sum of x-component of normalized Lorentz force |
| EPSY | Sum of y-component of normalized Lorentz force |
| EPSZ | Sum of z-component of normalized Lorentz force |

where $y$ is a set of strictly positive numbers and $\lambda$ the power parameter. The logarithmic form in equation 1 is a special case only when the $\lambda$ is selected to be zero. Hence, the existence of non-positive numbers in the SHARP parameters are forbidden as the values of variable, making the Box-Cox transformation not applicable in this study. Yeo-Johnson transformation was proposed in 2000, preserving many good features of Box-Cox power family without the restriction that the values of $y$ must be strictly positive:

$$\psi(\lambda, y) = \begin{cases} \left((y+1)^{\lambda} - 1\right)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0 \\ -\left[(-y+1)^{2-\lambda} - 1\right)\right]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y+1) & \text{if } \lambda = 2, y < 0 \end{cases} \tag{2}$$

wherein the transformation has the same form as that of Box-Cox with only the replacing of $y$ by $y+1$ for strictly positive $y$, and the replacing of $y$ and $\lambda$ by $-y+1$ and $2-\lambda$ for strictly negative $y$, respectively.

### 2.3. *Feature Selection*

We describe our procedure for selecting the SHARP parameters. Our approach is based on the intuition that significant SHARP parameters should have strong marginal association with the observed flare index, for which a marginal nonparametric regression is fitted between the flare index and each individual SHARP parameter. Based on this, we coarsely select 18 SHARP parameters. Moreover, we propose a new method called *exhaustive sieve* to refine the result.

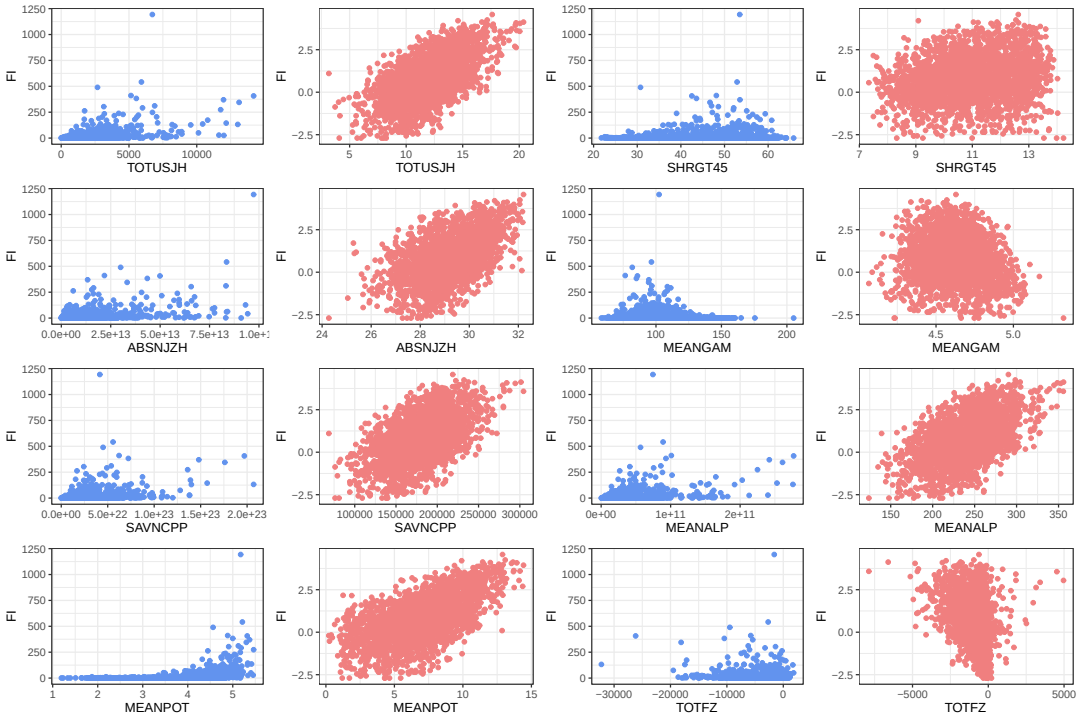


**Figure 1.** Evolution of FI with respect to 8 SHARP parameters, i.e., TOTUSJH, ABSNJZH, SAVNCPP, MEANPOT (1st and 2nd columns) and SHRGT45, MEANGAM, MEANALP, TOTFZ (3rd and 4th columns). The effect of power transformation is characterized by plotting the original data (blue) as well as the corresponding transformed data (red) in the same row.

#### 2.3.1. *Coarse Screening*

First, a non-parametric model for FI and SHARP parameters is developed:

$$y = f_j(X_j) + \sigma\varepsilon,\ \ j = 1, \ldots, 25, \tag{3}$$

where $y$ is the response variable, $X_1, \ldots, X_{25} \in [0, 1]$ are feature variables, $\epsilon$ is zero-mean continuous random variables with finite standard deviation $\sigma$, $f_j$ is an unknown function belonging to an $m$-order periodic Sobolev space $S^m$ on $[0, 1]$ which characterizes the marginal association between $y$ and $X_j$. One can regard $y$ and $X_j$ as the observed value of FIs and the 25 transformed SHARP parameters, and so, $f_j$ represents their functional relationship. We are interested in making valid statistical inferences regarding $f_i$ using techniques such as estimation or hypothesis testing.

Here, we propose a nonparametric hypothesis testing approach for testing the significance of each feature $X_1$, $X_2$,..., $X_{25}$ and exploring the correlation between the response variable and the feature variables. Our approach is motivated from the nonparametric inferential literature (Shang & Cheng 2017, 2013; Cheng & Shang 2015; Shang & Cheng 2015; Liu et al. 2019b; Yang et al. 2020; Liu et al. 2020; Shang 2010; Liu et al. 2021a,b). In this section, we test the marginal effect of each variable. Correlation analysis will be conducted in section 2.3.2.

The hypothesis for testing the significance of each individual SHARP parameter are

$$H_0 : f_j \text{ is constant } \textit{versus } H_1 : f_j \text{ is non-constant.} \tag{4}$$

Specifically, our test statistics for (4) is $T = \|\widehat{f}_j - \bar{y}\|^2_{L^2}$ in which $\widehat{f}_j$ is a smoothing spline estimator of $f_j$ based on model (3) and $\bar{y}$ is the averaged FIs. We reject $H_0$ at 1% significance level if $T > 2.576\sigma^2$ with $\sigma^2$ the theoretical variance of $T$ under $H_0$ (see Shang & Cheng (2017); Yang et al. (2020)). Rejection of $H_0$ implies that $X_j$ contributes significantly to the model. If $H_0 : f_j = \textit{constant}$ is not rejected, then this indicates that $X_j$ can be removed from the model.

Including an irrelevant variable in the model might actually raise the mean square error, reducing the model's effectiveness. This test assists in determining the value of each regression variable contained inside the regression model. Seven SHARP parameters failed to present significance in our study. The P-values in Figure 2 reveal that

the p values for these seven parameters (MEANSHR, MEANGBT, MEANGBZ, MEANGBH, MEANJZD, EPSY and EPSZ) are greater than 0.01, indicating that these parameters can be removed due to lack of statistical significance.

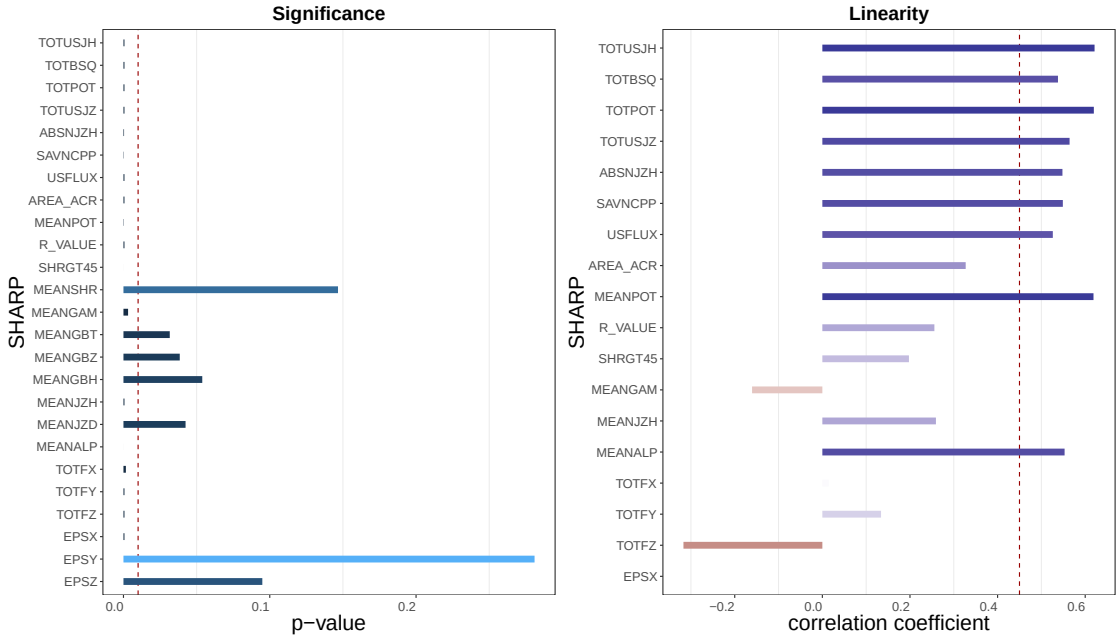


**Figure 2.** The histogram on the left shows p-values of 25 SHARP parameters with respect to the significance test. The histogram on the right gives the correlation coefficient value of the rest 18 SHARP variables. The vertical dashed lines highlight critical values used to define parameters that are statistically significant (p-value = 0.01 in left panel) and linearly correlated ($r = 0.5$ in right panel).

The second step is to determine whether the relationship between FI and the remaining SHARP variables is linear. Here, we apply the 'correlation coefficient' (Pearson 1896) to assess a possible linear association between FI and each SHARPs. Correlation coefficients (Equation 5) range from -1 to +1, with +/-1 denoting an ideal positive/negative linear relationship and 0 denoting the absence of a linear relationship.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]}} \tag{5}$$

A condition that is necessary for a perfect correlation is that the shapes of the individual X data and the individual Y data must be the identical (Ratner 2009). Additionally, correlations must be shown to be statistically significant before the correlation coefficient is considered meaningful (Taylor 1990). Thus, power transformation and significance testing provide a robust theoretical foundation for correlation analysis in our work. The correlation coefficients between SHARPs and FI with a cutoff of 0.5 are shown in Figure 2 under the assumption of a large sample size ($n > 100$, See Taylor 1990). 9 SHARP parameters (TOTUSJH, TOTBSQ, TOTPOT, TOTUSJZ, ABSNJZH, SAVNCPP, USFLUX, MEANPOT, MEANALP) represent large correlation coefficients ($r > 0.5$), indicating that the linear relationships between these 9 SHARP variables and the FI are statistically significant.

Therefore, based on the above significance test and linear correlation analysis, we establish the following preliminary model:

$$\begin{aligned} FI = {} & \beta_1 * TOTUSJH + \beta_2 * TOTBSQ + \beta_3 * TOTPOT + \beta_4 * TOTUSJZ + \beta_5 * ABSNJZH + \beta_6 * SAVNCPP + \\ & \beta_7 * USFLUX + \beta_8 * MEANPOT + \beta_9 * MEANALP + f_1(AREA\_ACR) + f_2(R\_VALUE) + f_3(SHRGT45) + \\ & f_4(MEANGAM) + f_5(MEANJZH) + f_6(TOTFX) + f_7(TOTFY) + f_8(TOTFZ) + f_9(EPSX) + \varepsilon_i, \ i = 1, \ldots, 18 \end{aligned} \tag{6}$$

where $\beta_i, i = 1, \ldots, 9$ are the coefficients for the 9 linear SHARP parameters. And $f_i$, $i = 1, \ldots, 9$, are the functions of the 9 nonlinear SHARP parameters.

### 2.3.2. *Exhaustive Sieve*

While we exclude seven SHARP parameters based on their importance in the hypothesis test, the resulting model remains relatively complicated and may be impacted by the high correlation between several components. Figure 3 gives the symmetric correlation matrix for the 25 SHARP parameters with color coding. A correlation matrix is a table that displays the correlation coefficients between sets of variables. Each cell in the table represents the correlation between two SHARP parameters. Green and yellow represent positive and negative correlation coefficient values, respectively. The darker the color, the closer the absolute value of the correlation coefficient is to 1, indicating a stronger linear relationship between variables. We recognize that some SHARP parameters are substantially correlated based on this correlation matrix. Notably, excessive correlation can diminish the precision of estimated coefficients, leading to skewed or misleading findings, which weakens the statistical power of the regression model. To address this problem, we perform an exhaustive sieving method to exclude some highly correlated features. The 3 steps are described as follows.

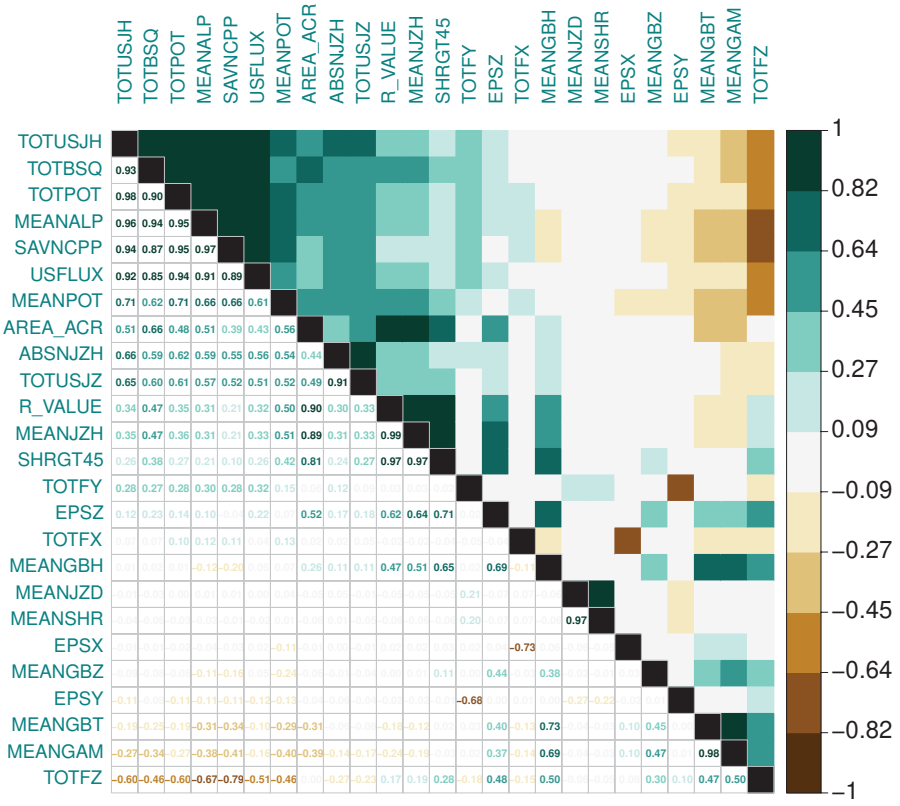


**Figure 3.** Correlation matrix that displays correlation coefficients for 25 SHARP parameters. The upper and lower triangles are correspondingly filled with colors and numeric values. Positive correlations are displayed in green and negative correlations in yellow. Color intensity of the cell is proportional to the correlation coefficients.

1. We begin by grouping the highly correlated features according to their correlation coefficients. Group1: TOTUSJH, TOTBSQ, TOTPOT, MEANALP, SAVNCPP, USFLUX; Group2: MEANGAM, MEANGBT; Group3: R_VALUE, MEANJZH, SHRGT45; Group4: TOTUSJZ, ABSNJZH; Group5: MEANSHR, MEANJZD

2. Eliminate one or more parameters from each group using the exhaustive sieve approach and compare the performance of each combination of various groups.

3. Finally, obtain the optimal combination and construct the model.

The exhaustive sieved model is as follows:

$$FI = \beta_1 * TOTUSJH + \beta_2 * MEANALP + \beta_3 * USFLUX + \beta_4 * TOTPOT + \beta_5 * SAVNCPP + \beta_6 * MEANPOT + f_1(SHRGT45) + f_2(MEANGAM) + f_3(AREA\_ACR) + f_4(EPSX) + f_5(TOTFZ) + \varepsilon_i,\ i = 1, \ldots, 11 \tag{7}$$

## 3. IMPLEMENTATION OF FLARE INDEX PREDICTION

### 3.1. *Multivariate Spline Regression*

We develop the FI prediction model using the multivariate spline regression technique in accordance with the exhaustive sieved model above. Spline regression is one of the most important nonlinear regression techniques. Rather than creating a single model for the entire data set, spline regression divides it into k continuous bins via knots and fits each bin with a separate model, such as a linear function or low-order polynomial function (such as quadratic or cubic multinomial, etc.). Splines are continuous, piece-wise polynomial functions. To fit a piecewise function, it is self-evident that the more knots in the model, the more flexible it is. While spline regression can be considered piecewise regression, it is not straightforward piecewise regression; rather, it is piecewise regression with constraints, which needs continuity at each knot. In this study, we perform regression using natural cubic spline.

B-spline or basis spline is a commonly used spline basis of spline functions: any spline function of a given degree can be represented as a linear combination of B-splines of the same degree. A B-spline curve is defined as the following:

$$f(t) = \sum_{i=0}^{n} B_{i,d}(t)\mathbf{C}_i \tag{8}$$

where $B_{i,d}(t)$ represents the B-spline basis functions at the scalar $t_i$ with degree of d. Note that, the sequence of $t_i$ are referred to be nondecreasing knots where $0 \leq i \leq n + d + 1$. $\mathbf{C}_i$ denotes control points for which the collection of n + 1 control points may be considered column vectors $\hat{\mathbf{C}}$:

$$\hat{\mathbf{C}} = (\mathbf{C}_0, \mathbf{C}_1, \ldots, \mathbf{C}_n)^\top \tag{9}$$

Consider, $\hat{\mathbf{S}}$ to be the collection of the sample data $\mathbf{S}_p$, where

$$\hat{\mathbf{S}} = (\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_p)^\top \tag{10}$$

The least-squares error function between the B-spline curve and the sample points can then be represented as:

$$E(\hat{\mathbf{C}}) = \frac{1}{2}\sum_{p=0}^{m}\left|\sum_{j=0}^{n} B_{j,d}\left(t_p\right)\mathbf{C}_j - \mathbf{S}_p\right|^2 \tag{11}$$

By selecting the control points, this error function may be minimized to the greatest extent possible.

B-splines can be used to represent any spline (of any degree). Degree = 3 is a common choice. The spline is referred to as a cubic spline in this situation. Typically, the second derivative of each third-order polynomial is set to zero at the endpoints in order to assure smoothness at the data points. This is known as natural cubic spline, which has a lower tendency for data points to fluctuate.

In our model, we employ both cubic and natural cubic B-splines. The tool we utilized here is *bs* and *ns* from *splines* R package. We collect data samples from 2010 to 2016 as training sets, and those in 2017 as testing sets. Given the comparatively large FI in 2017, utilizing 2017 data to assess our model's predictive power for large FI is an effective way to validate our model's predictive power for large FI. The number of knots associated with each parameter is determined by the predicted RMSE result, that is, the number of knots that minimizes the obtained RMSE. Figure 4 compares the model performance of spline regression to that of four popular machine learning methods (i.e. Linear Reression, LASSO, Random Forest (RF) and Gaussian Process Regression (GPR)). Since we pre-process the data by power transformation, we should inverse the results after the prediction is complete. The outer and inner panels, respectively, display the prediction results before and after the inverse transformation. Due to the fact that R squared is not appropriate to compare linear model and nonlinear model (Kvålseth 1985), RMSE (Equation 12)

and MAE (Equation 13) are selected as metrics for evaluating the predictive effect of FI. As illustrated in Figure 4, the RMSE and MAE of spline regression both exhibit the lowest values among the five tested algorithms, illustrating the advantage of our model. The graph in the lower-right corner of figure 4 is a combination of spline regression and SMOGN, which will be discussed in the following section. Here, we implement RF using the **randomForest** function (Liaw et al. 2002) in **R** package **randomForest**. Furthermore, we fine tune the model by defining a grid of algorithm with the following functions: **trainControl** from the **caret** package to conduct repeated 5-fold cross-validation; and **expand.grid** from the **base** package to get all combinations of hyperparameters, i.e., *mtry* is selected between 1 and 10, *maxnode* and *nodesize* are selected between 2 and 20, and *ntree* is selected from 100, 250, 300, 350, 400, 450, 500, 550, 600, 800, 1000, 2000, 3000. Following parameters are included in the final optimized model: $mtry = 3, maxnode = 17, nodesize = 6, and ntree = 100$. We construct GPR using the **gausspr** function (Williams & Barber 1998) in the **kernlab R** package with type = "regression" and kernel = "polydot". The kernel parameter is selected with 5-fold cross-validation from the eight most common kernel functions in the package; For LASSO, we conduct 5-fold cross-validation using the **cv.glmnet R** function (Simon et al. 2011) and get the lasso tuning parameter = lambda.min which minimizes the cross-validation error. Noting that RF is not robust to single outliers, the RMSE of RF is high due to the existence of a large FI in 2017. However, it outperformed models other than spline regression when test data sets are more balanced. Therefore, we will continue to examine this model as we add additional data in the future.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{12}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{13}$$

In the knots selection process of spline regression, if the training set and test set are quite specialized, it is easy to run into over-fitting difficulties. To verify the validity of our model, we generate up to 500 different training and testing folds of our data sample. To be specific, 90% of data samples from 2010 to 2016 were randomly selected as the training set, while 90% of the data samples from 2017 were randomly selected as the test set. Then fold them together. Thus, each fold represents a unique combination of training and test data that can be used to estimate the performance of our algorithm. The averages of RMSE and MAE are 18.9858 and 4.1932, respectively.

The prediction of strong flares is of significance in mitigating space weather effects, whereas weak flares allow us to assess the reliability of the methodology in this study. Thus, we conduct additional 7 predictions using the data from the year 2010 to 2016 as the test sets, respectively. Each of these 7 predictions is produced by training the models using the data from the rest years. The aggregated performance is represented as boxplots—see Figure 6a-c. The Spline Regression model exhibits the best performance in the predictions of strong, weak and all FIs.

**Table 3.** 9 feature rankings algorithms, and respective R package used.

| Learner | Parameter Variants |
|---|---|
| Wald Test | **rms** (Harrell Jr 2021) |
| LASSO | **glmnet** (Simon et al. 2011) |
| Random Forest (RF) | **caret** (Kuhn 2021) |
| Bagging Multivariate Adaptive Regression Splines (BMARS) | **caret** (Kuhn 2021) |
| Fisher Score | **Rdimtools** (You 2021) |
| DALEX | **DALEX** (Biecek 2018) |
| Boruta | **Boruta** (Kursa & Rudnicki 2010) |
| Stepwise Forward and Backward Selection | **stats** (R Core Team 2021) |
| Relative Importance from Linear Regression (RILR) | **relaimpo** (Grömping 2006) |

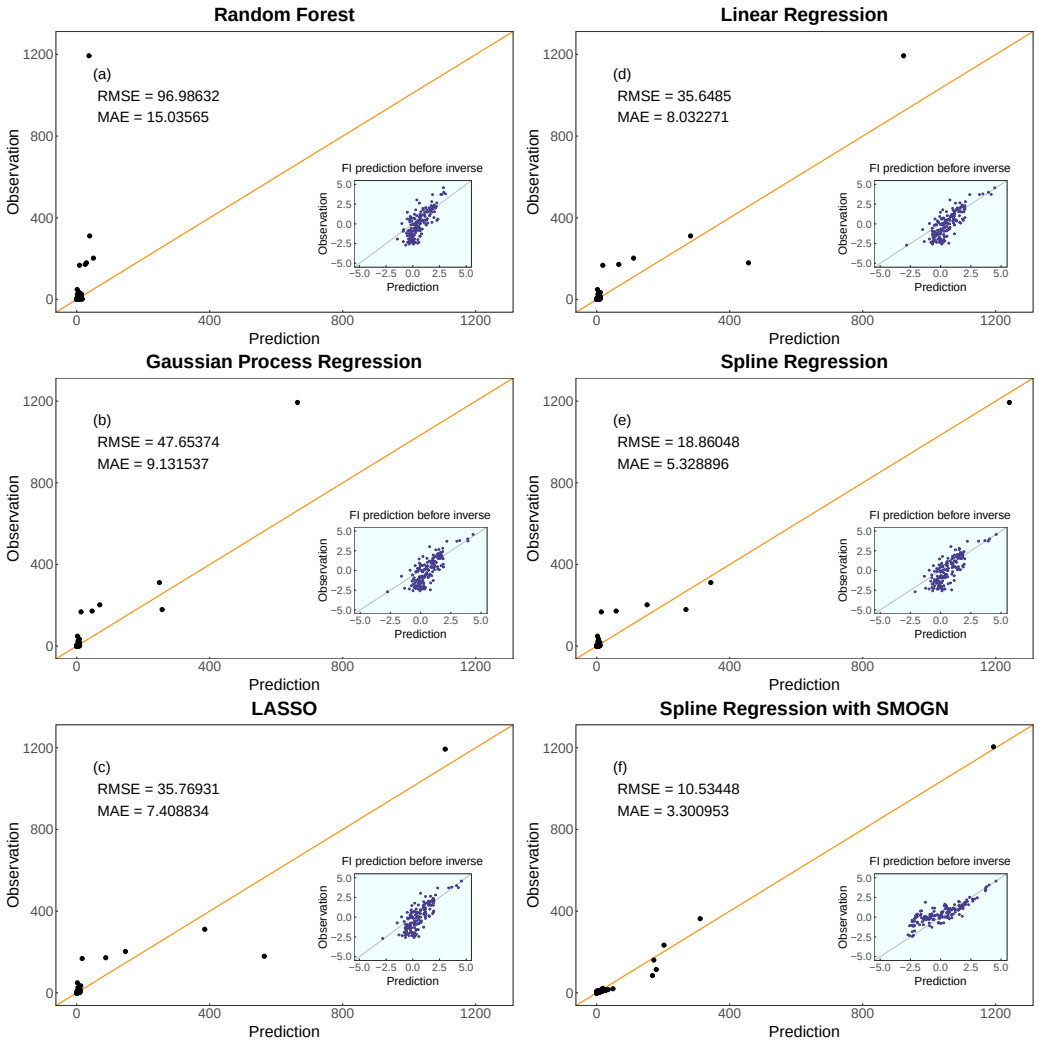


**Figure 4.** Performance of FI prediction using different models, i.e., Random Forest (a), LASSO (b), Gaussian Process Regression (c), Linear Regression (d), Spline Regression (e), and Spline Regression with SMOGN (f), is estimated by plotting dots at coordinates (predicted FI, observed FI). The reference line (yellow) highlights the ideal case wherein the prediction is strictly the same as the observation. Inner panels show the predicted FI without inverse processing vs. observed FI by power transformation.

### 3.2. *Data Imbalance Problem*

In almost all flares prediction studies, the prediction accuracy of weak flares is always higher than that of strong flares. The main reason is that the data is imbalanced. The number of FIs bigger than ten accounts for just 19% of the total in our data set. There are only 2% of FIs with a total value greater than 100.

Several techniques have been proposed to deal with imbalanced classification tasks (He & Garcia 2009; López et al. 2013). However, the problem of imbalanced domains in regression task is more complicated. One of the reasons is that the target variable is continuous. In comparison to fixed categorical variables, continuous variables may have an infinite number of possible values. In addition, the definition of more or less relevant values of the target is ambiguous.

In this study, we employ a pre-processing approach called Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (Branco et al. 2017), or SMOGN for short, to deal with imbalanced regression. This technique is effective for skew distribution-affected machine learning regression models. The SMOGN algorithm, in particular, when combined with spline regression, can achieve excellent performance (Branco et al. 2017). Essentially, it is a combination of random under-sampling (Torgo et al. 2013, 2015) with two over-sampling techniques: SMOTER (Torgo et al. 2013) and introduction of Gaussian Noise(Branco et al. 2016). By combining random under-sampling and over-sampling, it is possible to achieve a more balanced distribution of minority and majority cases while reducing their bias.

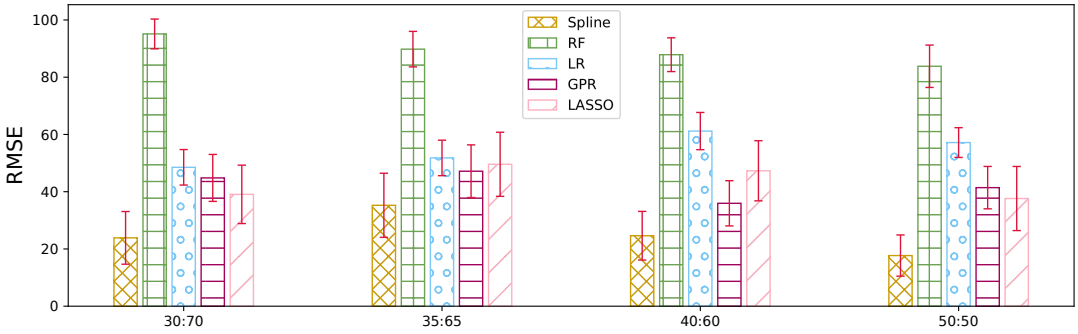


**Figure 5.** RMSE of predicted FI results from re-sampled data sets by SMOGN algorithm with fixed sampling ratios of minority to majority (i.e., 30:70, 35:75, 40:60 and 50:50), respectively. Standard deviations of RMSE are represented by red error bars.

First of all, the SMOGN algorithm begins by separating the training set into two parts: $P_N$ for majority(normal or less important) part, and $P_R$ for rare but important part. $t_E$ denotes the user-defined relevance threshold for defining the sets $P_N$ and $P_R$. $u\%$ and $o\%$ represent the percentage of under-sampling and over-sampling, respectively.

Random undersampling is a technique in which observations from the majority part $P_N$ are randomly removed. This sample is then combined with the observations from the minority part to create the final training set for the selected learning algorithm.

The over-sampling will use either SMOTER or the introduction of Gaussian Noise strategy to generate new cases for the new "more balanced" training set. One of the most widely used approach to synthesizing new examples for classification task is called the Synthetic Minority Oversampling Technique (SMOTE, Torgo et al. 2013). The SMOTER is a variant of SMOTE for addressing regression.

According to the distance between the seed example and all the remaining cases in $P_R$ under consideration, the user is required to set a threshold for determining whether the k-nearest neighbour is at a safe or unsafe distance and thus choose between SMOTER and the introduction of Gaussian Noise. SMOTER is used if the selected neighbor is deemed "safe". Otherwise, it is preferable to generate a new case by introducing Gaussian noise.

Similarly, we select data from 2010 to 2016 as the training set and data in 2017 as the testing set. The set of FIs greater than 10 was designated as the minority part $P_R$, and the remainder as the majority part $P_N$. In order to confirm the robustness of synthetic data, we produced data sets by applying SMOGN algorithm to original data with fixed sampling ratios of minority to majority (i.e., 30:70, 35:75, 40:60, and 50:50), respectively (Hasanin et al. 2019). A hundred data sets were generated for each resampling ratio. Here, to avoid the issues caused by high-dimensionality of the data, only 11 SHARP parameters obtained from section 2.3.2 are used. Meanwhile, we examined the boxplots of each sample, confirming that the synthetic data reflects the physical significance of solar flare. The **smogn Python** library (Branco et al. 2017) utilized here is available at https://github.com/nickkunz/smogn. Based on our experiments, SMOGN is safe to use when parameter k (the number of neighbors to consider for interpolation used in over-sampling) is set to 6 or 7. The results in figure 5 are obtained by implementing 5 ML algorithms (i.e., spline regression, RF, linear regression, GPR and LASSO) on these synthetic data sets. The combination of spline regression and SMOGN shows an optimal performance with $\text{RMSE}_{min} = 10.53448$, which is evidently smaller than the RMSE = 18.82912 by spline regression alone. Additionally, as shown in Figure 6d, SMOGN also enhances the performance in forecasting large FI.

## 4. MAGNETIC PARAMETER RANKING

As mentioned earlier in section 2.3.1, 18 SHARP parameters are evaluated to be statistically significant for the FI prediction. However, we cannot effectively rank the importance of these 18 SHARP parameters because almost all of them have a p value of zero. In this section, we sort the importance of these 18 parameters by Borda Count method calculated from the ranks that are rendered by 9 different machine learning methods (See table 5: Wald Test, LASSO, RF, Bagging MARS, Fisher Score, Boruta, DALEX, Stepwise Selection and Relative Importance from Linear Regression). Inspired from voting rules in electoral systems, the idea of Borda count method here is to combine independent feature ranking results into a more reliable feature ranking. For each ranking, the lowest-ranked feature

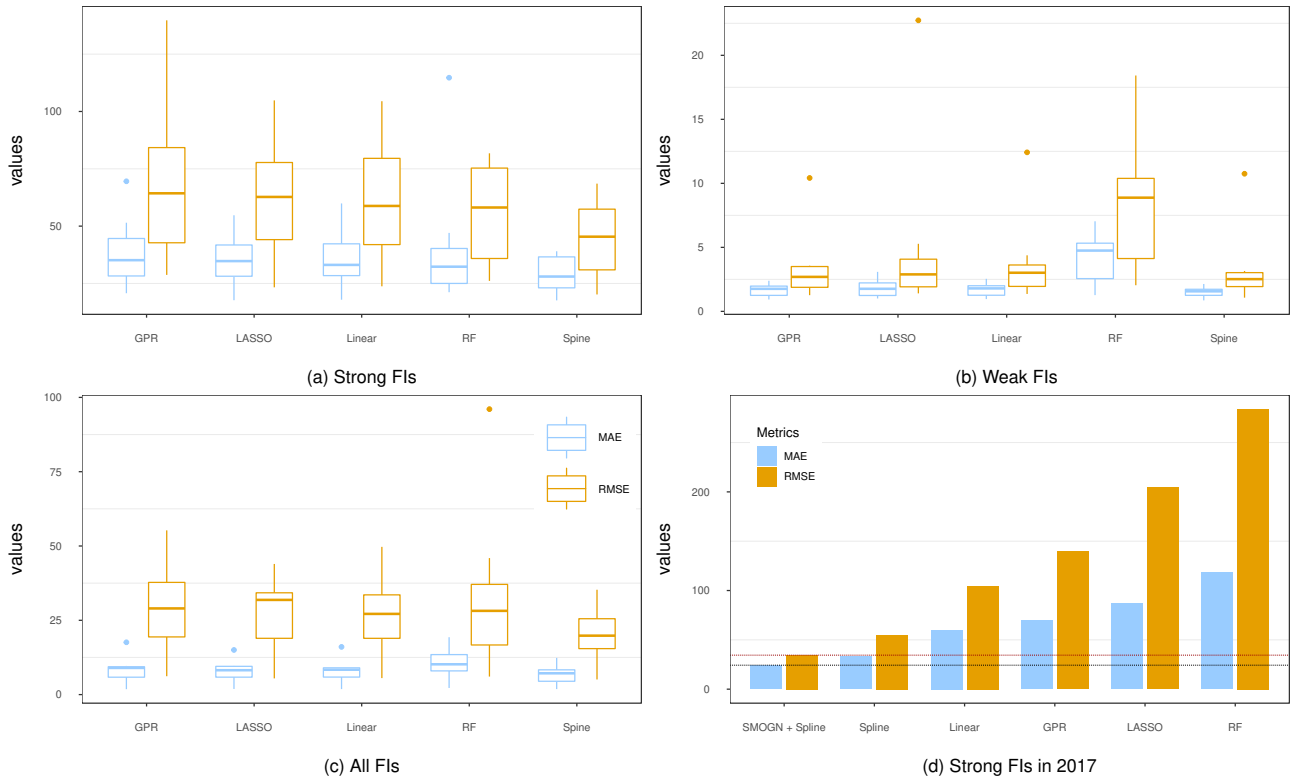


**Figure 6.** Boxplots of the RMSE (orange) and MAE (blue) of the predictions on the (a) strong FIs (FI > 10), (b) weak FIs, and (c) all FIs. (d) The RMSE and MAE of the predictions on the strong FIs in 2017. The vertical dashed lines highlight the minimum values of the RMSE and MAE.

gets 1 point, the second-lowest-ranked feature gets 2 points, and so on, until the highest-ranked feature gets points equal to the number of features in all 'votes', and the total score for each feature is called the Borda count. Borda count is sometimes referred to as a consensus-based voting method due to its ability to choose a more widely acceptable choice above the majority-supported one; see (Green-Armytage et al. 2016).

Table 4 and Figure 7 display the 9 ranking results and final scores obtained from Borda Count method, respectively. The top-10 SHARP parameters are TOTUSJH, MEANPOT, TOTUSJZ, TOTPOT, ABSNJZH, SAVNCPP, USFLUX, MEANGAM, TOTBSQ and TOTFZ. We compared our results with the results of similar previous studies (e.g., Bobra & Couvidat 2015a; Liu et al. 2017, 2019a; Chen et al. 2019; Wang et al. 2019a; Yeolekar et al. 2021). Due to inconsistencies in data and response variables, there are ranking differences in the top 10 SHARP features identified by different strategies. Notably, by comparing our customized rankings, FFS (Yeolekar et al. 2021) and Fisher-score (Bobra & Couvidat 2015a; Liu et al. 2017), we find that 9 of these 10 best-performing parameters are identical, indicating the robustness of our ranking results. In the rankings of Chen et al. (2019) and Liu et al. (2019a), the top 10 features change slightly, but the 10 best-performing features in our ranking continue to rank well in their rankings.

The top overall ranking here is TOTUSJH (total unsigned current helicity) with a Borda count of

$$TOTUSJH : 18(3) + 17(1) + 16(4) + 15(1) = 150 \tag{14}$$

TOTUSJH is calculated as $\sum |\mathbf{B}_z \cdot \mathbf{J}_z|$, where $\mathbf{B}_z$ and $\mathbf{J}_z$ are the normal component of magnetic field and electric current density only measured at the photospheric level of the Sun. TOTUSJH, as an approximation to the volume integral of $\mathbf{B} \cdot \mathbf{J}$ which characterizes the local linkage of elementary currents (Démoulin 2007), is an important proxy measure of an AR's magnetic non-potentiality (i.e., the degree of deviation from the potential state). It also outperforms other parameters in many previous studies of flare forecasting(e.g., Bobra & Couvidat 2015a; Liu et al. 2017; Chen et al. 2019; Liu et al. 2019a; Yeolekar et al. 2021).

**Table 4.** 9 feature rankings for 18 SHARPS

| SHARP | Wald Test | LASSO | RF | BMARS | Fisher Score | DALEX | Boruta | Stepwise | RILR |
|---|---|---|---|---|---|---|---|---|---|
| TOTUSJH | 1 | 2 | 4 | 3 | 3 | 3 | 3 | 1 | 1 |
| TOTBSQ | 17 | 14 | 8 | 5 | 2 | 7 | 13 | 16 | 7 |
| TOTPOT | 2 | 9 | 6 | 2 | 5 | 6 | 9 | 4 | 2 |
| TOTUSJZ | 10 | 5 | 1 | 7 | 1 | 2 | 1 | 10 | 4 |
| ABSNJZH | 8 | 6 | 2 | 6 | 4 | 4 | 4 | 6 | 6 |
| SAVNCPP | 3 | 10 | 10 | 8 | 7 | 10 | 12 | 2 | 5 |
| USFLUX | 16 | 7 | 5 | 14 | 8 | 5 | 7 | 17 | 9 |
| AREA_ACR | 14 | 15 | 9 | 12 | 11 | 11 | 16 | 13 | 10 |
| MEANPOT | 6 | 4 | 3 | 1 | 10 | 1 | 2 | 9 | 3 |
| R_VALUE | 9 | 8 | 15 | 10 | 15 | 13 | 10 | 12 | 12 |
| SHRGT45 | 12 | 16 | 12 | 11 | 16 | 12 | 6 | 3 | 15 |
| MEANGAM | 4 | 1 | 14 | 9 | 17 | 14 | 11 | 5 | 13 |
| MEANJZH | 11 | 17 | 11 | 4 | 14 | 9 | 5 | 14 | 14 |
| MEANALP | 18 | 18 | 7 | 16 | 6 | 8 | 15 | 15 | 8 |
| TOTFX | 15 | 13 | 18 | 17 | 13 | 18 | 17 | 11 | 17 |
| TOTFY | 7 | 11 | 16 | 15 | 12 | 17 | 18 | 8 | 16 |
| TOTFZ | 5 | 12 | 17 | 13 | 9 | 15 | 8 | 7 | 11 |
| EPSX | 13 | 3 | 13 | 18 | 18 | 16 | 14 | 18 | 18 |

**Table 5.** 9 feature rankings algorithms, and respective R package used.

| Methods | R Packages |
|---|---|
| Wald Test | **rms** (Harrell Jr 2021) |
| LASSO | **glmnet** (Simon et al. 2011) |
| Random Forest (RF) | **caret** (Kuhn 2021) |
| Bagging Multivariate Adaptive Regression Splines (BMARS) | **caret** (Kuhn 2021) |
| Fisher Score | **Rdimtools** (You 2021) |
| DALEX | **DALEX** (Biecek 2018) |
| Boruta | **Boruta** (Kursa & Rudnicki 2010) |
| Stepwise Forward and Backward Selection | **stats** (R Core Team 2021) |
| Relative Importance from Linear Regression (RILR) | **relaimpo** (Grömping 2006) |

The next top ranking parameters are MEANPOT (mean photospheric magnetic free energy; $\frac{1}{N}\sum(\mathbf{B}^{Obs}-\mathbf{B}^{Pot})^2$), TOTUSJZ (total unsigned vertical current; $\sum|\mathbf{J}_z|dA$), TOTPOT (total photospheric magnetic free energy density; $\sum(\mathbf{B}^{Obs}-\mathbf{B}^{Pot})^2 dA$), and ABSNJZH (absolute value of the net current helicity; $|\sum \mathbf{B}_z\cdot\mathbf{J}_z|$). The top five parameters, with the exception of MEANPOT which is an average-type (intensive) parameter, are all extensive parameters that scale with AR's size. Of these top five parameters, two relate to current helicity (TOTUSJH and ABSNJZH), two to magnetic free energy (MEANPOT and TOTPOT), and the other one measures vertical electric current (TOTUSJZ). Despite their different physical definitions, these parameters are all taken as an important proxy for the magnetic non-potentiality of the AR. Their high ranking in predicting FI indicates that, compared to other magnetic parameters, current helicity, magnetic free energy and electric current are key factors in flare forecasting.

## 5. CONCLUSIONS

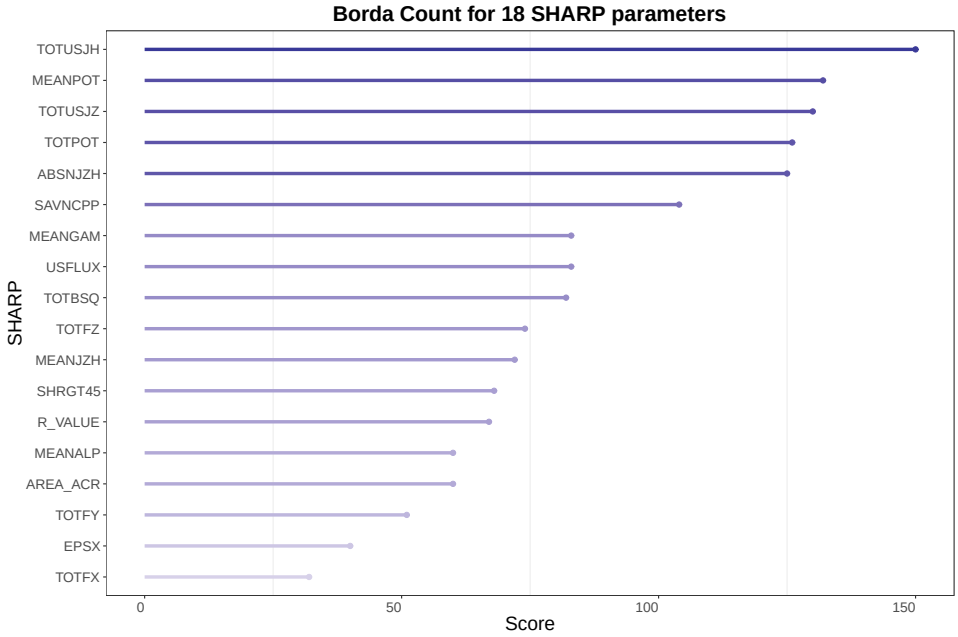


**Figure 7.** The corresponding Borda count for each SHARP parameter based on the 9 feature ranking results.

In this paper, we generate flare indices(1-day period) for each Active Region from May 2010 to Dec 2017. 25 SHARP parameters, which produced by the SDO/HMI team, are selected for our data samples for FI prediction. Yeo-Johnson transformation support spline regression models by normalizing data. Furthermore, certain correlations between SHARP and FI emerge after the processing by this transformation. Then, we test statistical significance of each SHARP parameter and the existence of a linear relationship between FI and SHARPs. Based on the results of hypothesis testing, we establish a preliminary model. Following that, according to the correlation matrix of 25 SHARP parameters, an exhaustive sieving procedure is utilized to eliminate some highly correlated features from the model. After feature selection, we adopt multivariate spline regression algorithm to predict FI and compare its performance to that of four popular machine learning methods (i.e. Linear Reression, LASSO, Random Forest (RF) and Gaussian Process Regression (GPR)). Here, the data sample from 2010 to 2016 is used for training, while the data sample from 2017 is used for testing. Finally, in order to improve the accuracy of strong flare index prediction, SMOGN, an advanced resampling method, is introduced to solve the problem of data imbalance.

Along with the FI prediction, we rank the importance of the SHARP parameter and derive a physical interpretation based on this ranking. The main results of this paper are summarized as follows.

1. Certain correlations between some SHARP parameters and FI emerge after the processing by Yeo-Johnson transformation. The linearity test in section 2.3.1 quantifies this fact through the correlation coefficient value. 9 SHARP parameters (TOTUSJH, TOTBSQ, TOTPOT, TOTUSJZ, ABSNJZH, SAVNCPP, USFLUX, MEANPOT, MEANALP) present large correlation coefficient value, illustrating the existence of linear relationship with FI. This is the first ever that the relationship between SHARPs and FI has been established.

2. 18 of the 25 SHARP parameters are demonstrated statistically significance in FI prediction by hypothesis testing in section 2.3.1. Meanwhile, TOTUSJH is shown to be the most important SHARP parameter in section 2.3. It also outperforms other parameters in many previous studies of flare forecasting. MEANPOT, TOTUSJZ, TOTPOT, and ABSNJZH are the following four ranking factors. This finding suggests that current helicity, magetnic free energy and electric current are the physical controlling factor of the flare productivity of ARs. Furthermore, certain parameters can still be omitted from the model due to the existing of highly correlation. Eventually, we select 11 parameters for the FI prediction model.

3. Spline regression is demonstrated to be the most effective method for FI prediction among the five tested algorithms (Random Forest, LASSO, Gaussian Process Regression, Linear Regression, Spine Regression), with an averages of RMSE and MAE of 18.9858 and 4.1932, respectively. On this premise, the combination of SMOGN

technique and spline regression considerably improves the accuracy of FI prediction in our study, particularly for large FI. Notably, there are several considerations to keep in mind when generating data, including the fact that the data's reliability, ratio, etc., will have a substantial influence on the results of the prediction.

On the basis of our findings, we conclude that utilizing SHARP parameters and the spline regression algorithm is a valid method for FI forecasting. It will be fascinating to identify additional effective parameters (such as time-dependent variables) that may be used to further improve the FI prediction.

We thank the team of SDO/HMI for producing vector magnetic field data products. We appreciate deeply the comments and suggestions made by the referee which improved the paper significantly. The work is supported by US NSF under grants AGS-1927578, AGS-1954737, AGS-2149748, AST-2204384 and AGS-2228996; NASA grants 80NSSC21K1671 and 80NSSC21K0003.

## REFERENCES

Abramenko, V. 2005, The Astrophysical Journal, 629, 1141, doi: /10.1086/431732

Antalova, A. 1996, Contributions of the Astronomical Observatory Skalnate Pleso, 26, 98

Barnes, G., Leka, K., Schumer, E., & Della-Rose, D. 2007, Space Weather, 5, doi: /10.1029/2007SW000317

Benvenuto, F., Piana, M., Campi, C., & Massone, A. M. 2018, The Astrophysical Journal, 853, 90, doi: 10.3847/1538-4357/aaa23c

Benz, A. O. 2017, Living Reviews in Solar Physics, 14, 1, doi: /10.1007/s41116-016-0004-3

Biecek, P. 2018, Journal of Machine Learning Research, 19, 1. https://jmlr.org/papers/v19/18-416.html

Bloomfield, D. S., Higgins, P. A., McAteer, R. J., & Gallagher, P. T. 2012, The Astrophysical Journal Letters, 747, L41, doi: /10.1088/2041-8205/747/2/L41

Bobra, M. G., & Couvidat, S. 2015a, The Astrophysical Journal, 798, 135, doi: /10.1088/0004-637X/798/2/135

—. 2015b, The Astrophysical Journal, 798, 135, doi: /10.1088/0004-637X/798/2/135

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, Solar Physics, 289, 3549, doi: /10.1007/s11207-014-0529-3

Boffetta, G., Carbone, V., Giuliani, P., Veltri, P., & Vulpiani, A. 1999, Physical review letters, 83, 4662, doi: http://dx.doi.org/10.1103/PhysRevLett.83.4662

Box, G. E., & Cox, D. R. 1964, Journal of the Royal Statistical Society: Series B (Methodological), 26, 211, doi: /10.1111/j.2517-6161.1964.tb00553.x

Branco, P., Ribeiro, R. P., & Torgo, L. 2016, arXiv preprint arXiv:1604.08079, doi: /10.48550/arXiv.1604.08079

Branco, P., Torgo, L., & Ribeiro, R. P. 2017, in First international workshop on learning with imbalanced domains: Theory and applications, PMLR, 36–50. https://proceedings.mlr.press/v74/branco17a.html

Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, Space Weather, 17, 1404, doi: /10.1029/2019SW002214

Cheng, G., & Shang, Z. 2015, Annals of Statistics, 43, 1351 , doi: 10.1214/15-AOS1313

De Boor, C., & Höllig, K. 1982, Journal d'analyse Mathématique, 42, 99, doi: /10.1007/BF02786872

Démoulin, P. 2007, Advances in Space Research, 39, 1674, doi: /10.1016/j.asr.2006.12.037

Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, Solar Physics, 293, 1, doi: /10.1007/s11207-018-1250-4

Gallagher, P. T., Moon, Y.-J., & Wang, H. 2002, Solar Physics, 209, 171, doi: /10.1023/A:1020950221179

Green-Armytage, J., Tideman, T. N., & Cosman, R. 2016, Social Choice and Welfare, 46, 183, doi: /10.1007/s00355-015-0909-0

Grömping, U. 2006, Journal of Statistical Software, 17, 1, doi: 10.18637/jss.v017.i01

Harrell Jr, F. E. 2021, rms: Regression Modeling Strategies. https://CRAN.R-project.org/package=rms

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. 2019, Journal of Big Data, 6, 1, doi: /10.1186/s40537-019-0274-4

Hazra, S., Sardar, G., & Chowdhury, P. 2020, Astronomy & Astrophysics, 639, A44, doi: /10.1051/0004-6361/201937426

He, H., & Garcia, E. A. 2009, IEEE Transactions on knowledge and data engineering, 21, 1263, doi: 10.1109/TKDE.2008.239

Huang, X., Wang, H., Xu, L., et al. 2018, The Astrophysical Journal, 856, 7, doi: 10.3847/1538-4357/aaae00

Huang, X., Zhang, L., Wang, H., & Li, L. 2013, Astronomy & Astrophysics, 549, A127, doi: /10.1051/0004-6361/201219742

Jiao, Z., Sun, H., Wang, X., et al. 2020, Space Weather, 18, e2020SW002440, doi: /10.1029/2020SW002440

Jing, J., Song, H., Abramenko, V., Tan, C., & Wang, H. 2006, The Astrophysical Journal, 644, 1273, doi: /10.1086/503895

Jing, J., Tan, C., Yuan, Y., et al. 2010, The Astrophysical Journal, 713, 440, doi: /10.1088/0004-637X/713/1/440

Jonas, E., Bobra, M., Shankar, V., Hoeksema, J. T., & Recht, B. 2018, Solar Physics, 293, 1, doi: /10.1007/s11207-018-1258-9

Kubo, Y. 2008, Solar Physics, 248, 85, doi: /10.1007/s11207-008-9135-6

Kuhn, M. 2021, caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret

Kursa, M. B., & Rudnicki, W. R. 2010, Journal of Statistical Software, 36, 1. http://www.jstatsoft.org/v36/i11/

Kusano, K., Iju, T., Bamba, Y., & Inoue, S. 2020, Science, 369, 587, doi: /10.1126/science.aaz2511

Kvålseth, T. O. 1985, The American Statistician, 39, 279, doi: /10.2307/2683704

Leka, K., Barnes, G., & Wagner, E. 2018, The NWRA classification infrastructure: Description and extension to the discriminant analysis flare forecasting system (daffs), EDP Sciences, doi: /10.1051/swsc/2018004

Lepreti, F., Carbone, V., & Veltri, P. 2001, The Astrophysical Journal Letters, 555, L133, doi: /10.1086/323178

Li, R., Cui, Y., He, H., & Wang, H. 2008, Advances in Space Research, 42, 1469, doi: /10.1016/j.asr.2007.12.015

Li, R., Wang, H.-N., He, H., Cui, Y.-M., & Du, Z.-L. 2007, Chinese Journal of Astronomy and Astrophysics, 7, 441, doi: /10.1088/1009-9271/7/3/15

Liaw, A., Wiener, M., et al. 2002, R news, 2, 18. http://CRAN.R-project.org/doc/Rnews/

Liu, C., Deng, N., Wang, J. T., & Wang, H. 2017, The Astrophysical Journal, 843, 104, doi: /10.3847/1538-4357/aa789b

Liu, H., Liu, C., Wang, J. T., & Wang, H. 2019a, The Astrophysical Journal, 877, 121, doi: /10.3847/1538-4357/ab1b3c

Liu, M., Shang, Z., & Cheng, G. 2019b, in 32nd Annual Conference on Learning Theory, ACM, 1–33. https://proceedings.mlr.press/v99/liu19a.html

Liu, M., Shang, Z., & Cheng, G. 2020, Electronic Journal of Statistics, 14, 3070, doi: /10.1214/20-EJS1733

Liu, M., Shang, Z., Yang, Y., & Cheng, G. 2021a, IEEE Transactions on Pattern Analysis and Machine Intelligence, forthcoming. https://ieeexplore.ieee.org/abstract/document/9369095

Liu, R., Boukai, B., & Shang, Z. 2021b, Journal of Mathematical Analysis and Applications, 505, 125561, doi: /10.1016/j.jmaa.2021.125561

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. 2013, Information sciences, 250, 113, doi: /10.1016/j.ins.2013.07.007

McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2016, Solar Physics, 291, 1711, doi: /10.1007/s11207-016-0933-y

Muranushi, T., Shibayama, T., Muranushi, Y. H., et al. 2015, Space Weather, 13, 778, doi: /10.1002/2015SW001257

Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, The Astrophysical Journal, 835, 156, doi: /10.3847/1538-4357/835/2/156

Park, E., Moon, Y.-J., Shin, S., et al. 2018, The Astrophysical Journal, 869, 91, doi: 10.3847/1538-4357/aaed40

Pearce, G., Rowe, A., & Yeung, J. 1993, Astrophysics and space science, 208, 99, doi: http://dx.doi.org/10.1007/BF00658137

Pearson, K. 1896, Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character, 253, doi: /10.1098/rsta.1896.0007

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. 2011, in The solar dynamics observatory (Springer), 3–15, doi: /10.1007/978-1-4614-3673-7_2

Qahwaji, R., & Colak, T. 2007, Solar Physics, 241, 195, doi: /10.1007/s11207-006-0272-5

R Core Team. 2021, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ratner, B. 2009, Journal of targeting, measurement and analysis for marketing, 17, 139, doi: /10.1057/jt.2009.5

Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, Solar Physics, 275, 229, doi: /10.1007/978-1-4614-3673-7_11

Schwenn, R. 2006, Living Reviews in Solar Physics, 3, 1, doi: /10.12942/lrsp-2006-2

Shang, Z. 2010, Electronic Journal of Statistics, 4, 1411, doi: 10.1214/10-EJS588

Shang, Z., & Cheng, G. 2013, Annals of Statistics, 41, 2608, doi: /10.1214/13-AOS1164

—. 2015, Annals of Statistics, 43, 1742, doi: 10.1214/15-AOS1322

—. 2017, Journal of Machine Learning Research, 18, 1. http://jmlr.org/papers/v18/16-289.html

Shin, S., Lee, J.-Y., Moon, Y.-J., Chu, H., & Park, J. 2016, Solar Physics, 291, 897, doi: /10.1007/s11207-016-0869-2

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. 2011, Journal of Statistical Software, 39, 1. https://www.jstatsoft.org/v39/i05/

Song, H., Tan, C., Jing, J., et al. 2009, Solar Physics, 254, 101, doi: /10.1007/s11207-008-9288-3

Stepanov, A. V., Benevolenskaya, E. E., Benevolenskaya, E. E., & Kosovichev, A. G. 2004, Multi-Wavelength Investigations of Solar Activity (IAU S223), Vol. 223 (Cambridge University Press), doi: /10.1017/S1743921304006118

Sun, Z., Bobra, M., Wang, X., et al. 2021, Earth and Space Science Open Archive, 32, doi: 10.1002/essoar.10508256.1

Taylor, R. 1990, Journal of diagnostic medical sonography, 6, 35, doi: /10.1177/875647939000600106

Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. 2015, Expert Systems, 32, 465, doi: /10.1111/exsy.12081

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. 2013, in Portuguese conference on artificial intelligence, Springer, 378–389, doi: /10.1007/978-3-642-40669-0_33

Toriumi, S., & Wang, H. 2019, Living Reviews in Solar Physics, 16, 1, doi: /10.1007/s41116-019-0019-7

Wahba, G. 1990, Spline models for observational data (SIAM), doi: /10.1137/1.9781611970128

Wang, J., Liu, S., Ao, X., et al. 2019a, The Astrophysical Journal, 884, 175, doi: /10.3847/1538-4357/ab441b

—. 2019b, The Astrophysical Journal, 884, 175, doi: /10.3847/1538-4357/ab441b

Wang, X., Chen, Y., Toth, G., et al. 2020, The Astrophysical Journal, 895, 3, doi: /10.3847/1538-4357/ab89ac

Welsch, B. T., Li, Y., Schuck, P. W., & Fisher, G. H. 2009, The Astrophysical Journal, 705, 821, doi: 10.1088/0004-637x/705/1/821

Wheatland, M. 2004, The Astrophysical Journal, 609, 1134, doi: /10.1086/421261

—. 2005, Space Weather, 3, doi: /10.1029/2004SW000131

—. 2010, The Astrophysical Journal, 710, 1324, doi: /10.1088/0004-637X/710/2/1324

Williams, C. K., & Barber, D. 1998, IEEE Transactions on pattern analysis and machine intelligence, 20, 1342, doi: /10.1109/34.735807

Winter, L. M., & Balasubramaniam, K. 2015, Space Weather, 13, 286, doi: /10.1002/2015SW001170

Yang, Y., Shang, Z., & Cheng, G. 2020, in 33rd Annual Conference on Learning Theory, ACM, 1–47. https://proceedings.mlr.press/v125/yang20a.html

Yeo, I.-K., & Johnson, R. A. 2000, Biometrika, 87, 954, doi: /10.1093/biomet/87.4.954

Yeolekar, A., Patel, S., Talla, S., et al. 2021, in 2021 International Conference on Data Mining Workshops (ICDMW), IEEE, 1067–1076, doi: /10.1109/ICDMW53433.2021.00138

You, K. 2021, Rdimtools: Dimension Reduction and Estimation Methods. https://CRAN.R-project.org/package=Rdimtools

Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, Solar Physics, 255, 91, doi: /10.1007/s11207-009-9318-9

Yu, D., Huang, X., Wang, H., et al. 2010, The Astrophysical Journal, 710, 869, doi: /10.1088/0004-637X/710/1/869

Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. 2010, Research in Astronomy and Astrophysics, 10, 785, doi: /10.1088/1674-4527/10/8/008

Zharkov, S., & Zharkova, V. 2006, Advances in Space Research, 38, 868, doi: /10.1016/j.asr.2006.03.035

Zheng, Y., Li, X., & Wang, X. 2019, The Astrophysical Journal, 885, 73, doi: /10.3847/1538-4357/ab46bd

Zhu, G., Lin, G., Wang, D., & Yang, X. 2020, The Astrophysical Journal, 902, 72, doi: /10.3847/1538-4357/abb2a0

# Reference sheet for `natbib` usage

(Describing version 7.1 from 2003/06/06)

*For a more detailed description of the* `natbib` *package, LaTeX the source file* `natbib.dtx`*.*

### Overview

The `natbib` package is a reimplementation of the LaTeX `\cite` command, to work with both author–year and numerical citations. It is compatible with the standard bibliographic style files, such as `plain.bst`, as well as with those for `harvard`, `apalike`, `chicago`, `astron`, `authordate`, and of course `natbib`.

### Loading

Load with `\usepackage[`*options*`]{natbib}`. See list of *options* at the end.

### Replacement bibliography styles

I provide three new `.bst` files to replace the standard LaTeX numerical ones:

`plainnat.bst` `abbrvnat.bst` `unsrtnat.bst`

### Basic commands

The `natbib` package has two basic citation commands, `\citet` and `\citep` for *textual* and *parenthetical* citations, respectively. There also exist the starred versions `\citet*` and `\citep*` that print the full author list, and not just the abbreviated one. All of these may take one or two optional arguments to add some text before and after the citation.

| | | |
|---|---|---|
| `\citet{jon90}` | ⇒ | Jones et al. (1990) |
| `\citet[chap.~2]{jon90}` | ⇒ | Jones et al. (1990, chap. 2) |
| `\citep{jon90}` | ⇒ | (Jones et al., 1990) |
| `\citep[chap.~2]{jon90}` | ⇒ | (Jones et al., 1990, chap. 2) |
| `\citep[see][]{jon90}` | ⇒ | (see Jones et al., 1990) |
| `\citep[see][chap.~2]{jon90}` | ⇒ | (see Jones et al., 1990, chap. 2) |
| `\citet*{jon90}` | ⇒ | Jones, Baker, and Williams (1990) |
| `\citep*{jon90}` | ⇒ | (Jones, Baker, and Williams, 1990) |

### Multiple citations

Multiple citations may be made by including more than one citation key in the `\cite` command argument.

| | | |
|---|---|---|
| `\citet{jon90,jam91}` | ⇒ | Jones et al. (1990); James et al. (1991) |
| `\citep{jon90,jam91}` | ⇒ | (Jones et al., 1990; James et al. 1991) |
| `\citep{jon90,jon91}` | ⇒ | (Jones et al., 1990, 1991) |
| `\citep{jon90a,jon90b}` | ⇒ | (Jones et al., 1990a,b) |

### Numerical mode

These examples are for author–year citation mode. In numerical mode, the results are different.

| | | |
|---|---|---|
| `\citet{jon90}` | ⇒ | Jones et al. [21] |
| `\citet[chap.~2]{jon90}` | ⇒ | Jones et al. [21, chap. 2] |
| `\citep{jon90}` | ⇒ | [21] |
| `\citep[chap.~2]{jon90}` | ⇒ | [21, chap. 2] |
| `\citep[see][]{jon90}` | ⇒ | [see 21] |
| `\citep[see][chap.~2]{jon90}` | ⇒ | [see 21, chap. 2] |
| `\citep{jon90a,jon90b}` | ⇒ | [21, 32] |

### Suppressed parentheses

As an alternative form of citation, `\citealt` is the same as `\citet` but *without parentheses*. Similarly, `\citealp` is `\citep` without parentheses. Multiple references, notes, and the starred variants also exist.

| | | |
|---|---|---|
| `\citealt{jon90}` | ⇒ | Jones et al. 1990 |
| `\citealt*{jon90}` | ⇒ | Jones, Baker, and Williams 1990 |
| `\citealp{jon90}` | ⇒ | Jones et al., 1990 |
| `\citealp*{jon90}` | ⇒ | Jones, Baker, and Williams, 1990 |
| `\citealp{jon90,jam91}` | ⇒ | Jones et al., 1990; James et al., 1991 |
| `\citealp[pg.~32]{jon90}` | ⇒ | Jones et al., 1990, pg. 32 |
| `\citetext{priv.\ comm.}` | ⇒ | (priv. comm.) |

The `\citetext` command allows arbitrary text to be placed in the current citation parentheses. This may be used in combination with `\citealp`.

### Partial citations

In author–year schemes, it is sometimes desirable to be able to refer to the authors without the year, or vice versa. This is provided with the extra commands

| | | |
|---|---|---|
| `\citeauthor{jon90}` | ⇒ | Jones et al. |
| `\citeauthor*{jon90}` | ⇒ | Jones, Baker, and Williams |
| `\citeyear{jon90}` | ⇒ | 1990 |
| `\citeyearpar{jon90}` | ⇒ | (1990) |

### Forcing upper cased names

If the first author's name contains a *von* part, such as "della Robbia", then `\citet{dRob98}` produces "della Robbia (1998)", even at the beginning of a sentence. One can force the first letter to be in upper case with the command `\Citet` instead. Other upper case commands also exist.

| | | | |
|---|---|---|---|
| when | `\citet{dRob98}` | ⇒ | della Robbia (1998) |
| then | `\Citet{dRob98}` | ⇒ | Della Robbia (1998) |
| | `\Citep{dRob98}` | ⇒ | (Della Robbia, 1998) |
| | `\Citealt{dRob98}` | ⇒ | Della Robbia 1998 |
| | `\Citealp{dRob98}` | ⇒ | Della Robbia, 1998 |
| | `\Citeauthor{dRob98}` | ⇒ | Della Robbia |

These commands also exist in starred versions for full author names.

### Citation aliasing

Sometimes one wants to refer to a reference with a special designation, rather than by the authors, i.e. as Paper I, Paper II. Such aliases can be defined and used, textual and/or parenthetical with:

```
\defcitealias{jon90}{Paper~I}
\citetalias{jon90}              ⇒   Paper I
\citepalias{jon90}              ⇒   (Paper I)
```

These citation commands function much like `\citet` and `\citep`: they may take multiple keys in the argument, may contain notes, and are marked as hyperlinks.

### Selecting citation style and punctuation

Use the command `\bibpunct` with one optional and 6 mandatory arguments:

1. the opening bracket symbol, default = (
2. the closing bracket symbol, default = )
3. the punctuation between multiple citations, default = ;
4. the letter 'n' for numerical style, or 's' for numerical superscript style, any other letter for author–year, default = author–year;
5. the punctuation that comes between the author names and the year
6. the punctuation that comes between years or numbers when common author lists are suppressed (default = ,);

The optional argument is the character preceding a post-note, default is a comma plus space. In redefining this character, one must include a space if one is wanted.

Example 1, `\bibpunct{[}{]}{,}{a}{}{;}` changes the output of

```
\citep{jon90,jon91,jam92}
```

into [Jones et al. 1990; 1991, James et al. 1992].

Example 2, `\bibpunct[; ]{(}{)}{,}{a}{}{;}` changes the output of

```
\citep[and references therein]{jon90}
```

into (Jones et al. 1990; and references therein).

### Other formatting options

Redefine `\bibsection` to the desired sectioning command for introducing the list of references. This is normally `\section*` or `\chapter*`.

Define `\bibpreamble` to be any text that is to be printed after the heading but before the actual list of references.

Define `\bibfont` to be a font declaration, e.g. `\small` to apply to the list of references.

Define `\citenumfont` to be a font declaration or command like `\itshape` or `\textit`.

Redefine `\bibnumfmt` as a command with an argument to format the numbers in the list of references. The default definition is `[#1]`.

The indentation after the first line of each reference is given by `\bibhang`; change this with the `\setlength` command.

The vertical spacing between references is set by `\bibsep`; change this with the `\setlength` command.

### Automatic indexing of citations

If one wishes to have the citations entered in the `.idx` indexing file, it is only necessary to issue `\citeindextrue` at any point in the document. All following `\cite` commands, of all variations, then insert the corresponding entry to that file. With `\citeindexfalse`, these entries will no longer be made.

### Use with `chapterbib` package

The `natbib` package is compatible with the `chapterbib` package which makes it possible to have several bibliographies in one document.

The package makes use of the `\include` command, and each `\included` file has its own bibliography.

The order in which the `chapterbib` and `natbib` packages are loaded is unimportant.

The `chapterbib` package provides an option `sectionbib` that puts the bibliography in a `\section*` instead of `\chapter*`, something that makes sense if there is a bibliography in each chapter. This option will not work when `natbib` is also loaded; instead, add the option to `natbib`.

Every `\included` file must contain its own `\bibliography` command where the bibliography is to appear. The database files listed as arguments to this command can be different in each file, of course. However, what is not so obvious, is that each file must also contain a `\bibliographystyle` command, *preferably with the same style argument.*

### Sorting and compressing citations

Do not use the `cite` package with `natbib`; rather use one of the options `sort` or `sort&compress`.

These also work with author–year citations, making multiple citations appear in their order in the reference list.

### Long author list on first citation

Use option `longnamesfirst` to have first citation automatically give the full list of authors.

Suppress this for certain citations with `\shortcites{`*key-list*`}`, given before the first citation.

### Local configuration

Any local recoding or definitions can be put in `natbib.cfg` which is read in after the main package file.

**Options that can be added to `\usepackage`**

`round` (default) for round parentheses;

`square` for square brackets;

`curly` for curly braces;

`angle` for angle brackets;

`colon` (default) to separate multiple citations with colons;

`comma` to use commas as separaters;

`authoryear` (default) for author–year citations;

`numbers` for numerical citations;

`super` for superscripted numerical citations, as in *Nature*;

`sort` orders multiple citations into the sequence in which they appear in the list of references;

`sort&compress` as `sort` but in addition multiple numerical citations are compressed if possible (as 3–6, 15);

`longnamesfirst` makes the first citation of any reference the equivalent of the starred variant (full author list) and subsequent citations normal (abbreviated list);

`sectionbib` redefines `\thebibliography` to issue `\section*` instead of `\chapter*`; valid only for classes with a `\chapter` command; to be used with the `chapterbib` package;

`nonamebreak` keeps all the authors' names in a citation on one line; causes overfull hboxes but helps with some `hyperref` problems.