Jamie Riggs, PhD, IAA

A Generalized Linear Mixed Model for Enumerated Sunspots

Abstract

Sunspot count data from May, 2010 through June, 2017 were provided by the American Association of Variable Star Observers Solar Division to estimate monthly sunspot numbers. The data include sunspot counts for the sunspot cycle 24 minimum to minimum. Monthly estimates are determined from a mixed effects, loglinear model constructed specifically from these count data. The observer is treated as a random effect, and the observing condition and observer experience as fixed effects. This model differs in the treatment of the data distribution assumptions of the existing linear model developed by Shapley (1949) from Wald (1940), which models sunspot numbers by variance-stabilizing transformations prior to forming a weighted average of the monthly counts. The loglinear model methodology meets or exceeds the performance criteria set by Shapley, and provides a method for determining the relative sunspot number reported monthly by the American Association of Variable Star Observers Solar Division. Model improvements using additional explanatory fixed effects and further investigation of random effects probability distribution and link function combinations are discussed.

Keywords: Sunspots, Poisson distribution, gamma distribution, log link

1 Introduction

Sunspot counts data were provided to estimate monthly sunspot numbers from qualifying American Association of Variable Star Observers (AAVSO) members who submit observations on a monthly basis to the AAVSO Solar Division. The counts from each observer are checked for consistency and completeness, and then are combined such that the resulting sunspot numbers attempt to have minimal observational error. Effectively, the individual's monthly sunspot count is adjusted to the overall, weighted monthly average of all the qualifying numbers. As such, the overall, weighted monthly averages are assumed to be unbiased estimates of the true monthly sunspot numbers. As no sunspot number standard is available for this study, the corrections are therefore relative to the data provided. As such, "tuning" is required for each month in the study, as well as each successive month following the last month used in this study. This tuning is consistent with historical (Shapley, 1949; Taylor, 1985; Schaefer, 1993), and current (Clette et al., 2007) treatments.

Section 2 is an overview of the sunspot counting literature; Section 3 discusses the statistical model used by previous correction factor authors; Section 4 gives a brief tutorial on the mixed effects loglinear statistical model used in this paper; Section 5 details the analysis of the correction factors using a mixed effects loglinear model; Section 6 is a plan for future improvements; and Section 7 are the conclusions.

2 Background

The American sunspot number, R_a, is a relative index of daily and monthly sunspot activity, and the American Association of Variable Star Observers (AAVSO) Solar Division's program of

data-gathering and analysis has been active since its inception in 1944. Shapley (1949), Taylor (1985), and Schaefer (1993) provide descriptions of the method of sunspot numbers data reduction. Schaefer (1997) and Foster (1997) discuss enhancements and remedies to the data reduction needed for time-based changes in the reduction outcomes. These enhancements and remedies will not be discussed in this paper.

AAVSO observer raw data are submitted monthly to the Solar Division as sets of date- and time-stamped values which are converted to a sunspot number according to R. Wolf such that

$$R_i = 10g_i + s_i, \tag{1}$$

where i designates an individual observer for R_i, the adjusted sunspot number, g_i the number of sunspot groups reported, and s_i the number of reported spots. Taylor (1985) states that the grouping scheme is the evolutionary classification system outlined by M. Waldmeier (1961). Individual sunspots with an extent of 0.04 solar degrees and larger are counted.

Mid-month following the month of observation, after (usually) thirty or more reports have been received and initially processed, the computation of provisional sunspot numbers proceeds through application of the relation from Shapley (1949),

$$R_A = \frac{\sum_{i=1}^{N} w_i k_i R_i}{\sum_{i=1}^{N} w_i}, \tag{2}$$

for each day of the computational month. R_i is the daily sunspot count for individual contributor i, and R_A is the relative sunspot number after reducing all contributor counts, R_i. The parameter k_i adjusts for observer i counting conditions, and w_i is a derived weight that measures how well k_i adjusts to a standard for observer i.

According to Taylor (1985), the number of observers per day is expected to exceed eighteen. However, this number depends upon the phase of the sunspot cycle, on prevailing local weather conditions, and on observer confidence, especially during periods of minimal sunspot activity. The monthly American sunspot number, R_a, is formed by averaging the daily sunspot numbers (R_A) across the number of days in the month of interest.

Final American sunspot numbers are obtained when observer reports have been received. The monthly mean of the final values allows the calculation of the statistic, R_{sm}, the smoothed mean relative sunspot number. This number is computed from Waldmeier (1961) and is reproduced in Taylor as:

$$R_{sm} = \frac{1}{24}\left(N_{i-6} + N_{i+6} + 2\sum_{-5}^{5} N_i\right). \tag{3}$$

In Equation 3, $N_{\pm 6}$ is set equal to the sunspot mean number 6 months prior to the month for which R_{sm} is being calculated, and to the sunspot mean number 6 months after the month for which R_{sm} is being calculated, respectively. The intermediate successive month's mean values are taken under the summation. Thus, this moving average sunspot number lags six months behind the most recent month's sunspot determination.

3 Derivation of the Parameters k and w

We now examine how the correction parameters k_i and w_i are derived. Taylor (1985), with reference to a statement by Shapley (1949), reports that the data reduction method for determining the correction parameters comes from Wald (1940). Wald developed a model relating two random variables, where a random variable consists of observations sampled from a larger population. His method differs from simple linear regression in that simple linear regression assumes the independent variable can be measured without error, whereas Wald's method makes no such assumption.

3.1 Problem Formulation

In the American sunspot number calculations, it is assumed that no two observers will report the same Wolf number. However, the true Wolf number is considered to be at or close to the mean value of all the submitted counts. Shapley (1949) shows the expected value (mean) to be a reasonable estimate after data processing. The problem is to find parameters that adjust each observers daily submitted numbers to the expected value. Shapley uses a data reduction method developed by Wald (1940) to find values of the parameter k_i, where i indicates the ith observer's k value.

Wolf numbers are derived from counts of sunspots and sunspot groups. Counts data do not follow a Gaussian probability distribution function (PDF), which violates a Wald condition of using data that follow a Gaussian distribution. The sunspot numbers therefore are tranfromed using a natural logarithm. As counts of zero connot be transformed, Shapley drops them, which is considered poor practice by statisticians. We will drop zeros for this discussion, and incorporate zeros in the generalized linear model sections below.

We wish to fit a straight line to two variables R_i and R_A, defined above, each with uncorrelated errors according to Wald (1940). Madansky (1959) compared least squares estimation methods, two grouping methods, variance components analyses, and estimation by cumulants for a specific data set. The grouping method of Wald was shown to be one of the best straight line estimation methods among the 15 variants studied. Regardless, the fit from data grouping will be subject to the following two conditions in addition to R_i and R_A following a Gaussian PDF:

1. The fitted straight line relating y to x can be determined without making a priori assumptions on independence of the observed values of the x and y pairs relative to the standard deviations of the errors of x and y.

2. The unknown standard deviations of the two variable's errors can be well estimated from the observed values of x and y. The precision of the estimates increases with the sample size of the variable pairs and give asymptotically exact values with very large sample size.

In the context of sunspot numbers, consider two sets of daily random vectors with the random variable elements $\boldsymbol{X} = \{X_1, \ldots, X_N\}$, $(X_j = \log(R_{ij}))$ and $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$, $(Y_j = \log(R_{Aj}))$, where i indexes the observer and j indexes the day. Let the expected values of the elements of the $\boldsymbol{X}$ and $\boldsymbol{Y}$ vectors be denoted as vectors with elements $\boldsymbol{\mu_X} = \{\mu_{X1},\ \mu_{X2},\ \ldots,\ \mu_{XN}\}$ and $\boldsymbol{\mu_Y} = \{\mu_{Y1},\ \mu_{Y2},\ \ldots,\ \mu_{YN}\}$. The expected values are the true but unknown values of random

variables of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Let

$$\epsilon_j = Y_j - \mu_{Yj} \quad \text{and} \quad \eta_j = X_j - \mu_{Xj}, \tag{4}$$

denote the respective errors of the Y_j and X_j, and in vector form as $\boldsymbol{\epsilon} = \{\epsilon_1,\ \epsilon_2,\ \ldots,\ \epsilon_N\}$ and $\boldsymbol{\eta} = \{\eta_1,\ \eta_2,\ \ldots,\ \eta_N\}$. We make the following assumptions in which $\mathcal{E}(\cdot)$ is denotes the expected value of the parenthetical argument, and $\text{Var}(\cdot)$ denotes the variance of the argument:

A1) The random variable elements of the random vector $\boldsymbol{\epsilon}$ are identically independently distributed (iid) such that $\mathcal{E}(\epsilon_j \epsilon_k) = 0,$ for all $j \neq k$ and $\text{Var}(\epsilon_j) < \infty$.

A2) The random variable elements of the random vector $\boldsymbol{\eta}$ are iid such that $\mathcal{E}(\eta_j \eta_k) = 0,$ for all $j \neq k$ and $\text{Var}(\eta_k) < \infty$.

A3) The random variables ϵ_j and η_j are uncorrelated such that $\mathcal{E}(\epsilon_j \eta_j) = 0$ for all $j = 1, \ldots, N$.

A4) A simple linear relation holds between the true values of Y_j and X_j as

$$\mu_{Yj} = \alpha \mu_{Xj} + \beta, \quad , j = 1, \ldots, N. \tag{5}$$

where α is the slope and β is the y-intercept of the fitted straight line.

Let ϵ_l follow an iid PDF which we may denote as $f_\epsilon(\epsilon_j)$ and η_j follow an iid PDF which we may denote as $f_\eta(\eta_j)$ both for all $j = 1, \ldots, N$. The problem of fitting a straight line to two error-prone random variables may be formulated as follows:

1. x_j are the realizations of X_j
   y_j are the realizations of Y_j
2. The true values of $\mu_{Xj},\ \mu_{Yj}\ (j = 1, \ldots, N), \alpha$, and β are unknown.
3. From the realizations x_j and y_j we estimate
   (a) α and β
   (b) the standard deviation of $\boldsymbol{\epsilon}$ denoted as σ_ϵ
   (c) the standard deviation of $\boldsymbol{\eta}$ denoted as σ_η

3.2 Parameter Estimates

The basis of the parameter estimation for a straight line fit between two variables with respective associated errors is the grouping of the variable point pairs (x_j, y_j). For purposes of discussion, the x_j are considered to have no apparent clustering along the x-axis. These N data pair values are divided into two groups. Let

$$a_1 = \frac{1}{N}\left[\sum_{j=1}^{m} x_j - \sum_{j=m+1}^{N} x_j\right] \quad \text{and} \quad a_2 = \frac{1}{N}\left[\sum_{j=1}^{m} y_j - \sum_{j=m+1}^{N} y_j\right], \tag{6}$$

for $m = \lfloor N/2 \rceil$, where $\lfloor \cdot \rceil$ denotes the nearest integer. Then we estimate the slope α as

$$\hat{\alpha} = \frac{a_2}{a_1}. \tag{7}$$

It can be shown that $\hat{\alpha}$ is a consistent estimator of α; i.e., as $N \to \infty,\ \hat{\alpha} \to \alpha$. Shapley (1949) sets $\hat{\alpha} = 1$ so only the intercept need be estimated.

The intercept β of the regression line is estimated as

$$\hat{\beta} = \hat{\mu}_y - \alpha\hat{\mu}_x \equiv \hat{\mu}_y - \hat{\mu}_x \tag{8}$$

where

$$\hat{\mu}_x = \frac{1}{N}\sum_{i=1}^{N} x_j \quad \text{and} \quad \hat{\mu}_y = \frac{1}{N}\sum_{j=1}^{N} y_j. \tag{9}$$

The estimate $\hat{\beta}$ can be shown to be a consistent estimator of β.

Consistent estimators of the variances for σ_ϵ^2 and σ_η^2 use the following relationships:

$$\sigma_\eta^2 = \left[\hat{\sigma}_x^2 - \hat{\sigma}_{xy}\right] \frac{N}{N-1} \tag{10}$$

$$\sigma_\epsilon^2 = \left[\hat{\sigma}_y^2 - \hat{\sigma}_{xy}\right] \frac{N}{N-1} \tag{11}$$

where

$$\hat{\sigma}_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu}_x)^2} \tag{12}$$

$$\hat{\sigma}_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{\mu}_y)^2} \tag{13}$$

$$\hat{\sigma}_{xy} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y) \tag{14}$$

3.3 Confidence Interval for β

For $\hat{\beta} = \hat{\mu}_y - \hat{\mu}_x$, we have that

$$\hat{\beta} - \beta = (\hat{\mu}_y - \mu_y) - (\hat{\mu}_x - \mu_x) = \hat{\eta} - \hat{\epsilon}. \tag{15}$$

Let t_0 be the critical value of the t-statistic for a prespecified confidence level, a CI for β then is

$$\left[(\hat{\mu}_y - \hat{\mu}_x) - t_0\sqrt{\frac{(\sigma'_y)^2 + (\sigma'_x)^2 - \sigma'_{xy}}{N-2}},\ \ (\hat{\mu}_y - \hat{\mu}_x) + t_0\sqrt{\frac{(\sigma'_y)^2 + (\sigma'_x)^2 - \sigma'_{xy}}{N-2}}\right] \tag{16}$$

where all the terms are defined as above.

3.4 The k Value

Wald's parameter estimates in Equations 8 and 9 are converted into the familiar k-factors as follows:

$$\begin{aligned}\hat{\beta} &= \hat{\mu}_y - \hat{\mu}_x \\ &= \frac{1}{N}\sum_{j=1}^{N} y_j - \frac{1}{N}\sum_{i=1}^{N} x_j \\ &= \frac{1}{N}\left(\sum_{j=1}^{N} y_j - \sum_{i=1}^{N} x_j\right) \\ &= \frac{1}{N}\left(\sum_{j=1}^{N} \log R_{Aj} - \sum_{i=1}^{N} \log R_{ij}\right) \end{aligned} \tag{17}$$

Taking the antilog of $\hat{\beta}$, we obtain k_i as

$$k_i = e^{\hat{\beta}} = \left[\prod_{j=1}^{N} \exp\left(\frac{R_{Aj}}{R_{ij}}\right)\right]^{1/N}, \tag{18}$$

where $j = 1, 2, \ldots, N$ is the number of days chosen to determine k_i.

3.5 The w Value

Shapley (1949) defines the weighting factor w_i as

$$w_i = \frac{N-1}{\sum_{j=1}^{N} (\log R_{Aj} - \log R_{ij})^2 - N\hat{\beta}^2}. \tag{19}$$

Recall that w_i is a measure of how well k_i adjusts to the standard of daily averages. It is the inverse of the variance (square of the standard deviation) of deviation of the observed daily counts from the daily average from the Shapley (using Wald) fitted deviation of the daily observered counts from the daily average counts.

The current implementation of the AAVSO is to set w_i to 1 for each vetted observer. All others have a zero weight.

4 Statistical Models for Counts Data

Models used by statisticians for counts data are part of a broad class of models called generalized linear models (see McCullagh and Nelder (1989)). Generalized Linear Models (GLMs) are specified (Agresti, 1998) by three components: a random component, which identifies the response variable probability distribution, in this case the Poisson PDF or the negative binomial PDF for sunspot numbers; a systematic component, which for our case includes a matrix of observer designators, the

date and time of the observations, the seeing conditions, and the experience level; and a function, called a link, that specifies the relationship between the expected value of the sunspot number random variable and the systematic component, e.g., a natural log transformation.

The modeling method used by Shapley approximates the sunspot number response with a Gaussian distribution by log-transforming the count data. This transformation is intended to stabilize the variance, as the counts distribution expected value changes with the variance. That is, $E(Y) = Var(Y)$, where $E(Y)$ denotes the expected value of the Poisson-distributed sunspot numbers, Y, and $Var(Y)$ denotes the variance (the square root of which is the standard deviation) of the sunspot numbers. However, log transformation of a Poisson distribution does not always guarantee a resulting normal distribution particularly with low count values. With GLMs, if the link results in additive predictor variables, it is not necessary to also stabilize the variance or produce normality as with normal linear regression.

Consider sunspot numbers submitted by the various observers to be treated as independent Poisson random variables. Let y_i denote the sunspot numbers of the ith observer, and $\mu_i = \mathcal{E}(Y_i)$ denote the expected value of the ith observer's sunspot count, $i = 1, 2. \ldots, N$. The Poisson probability distribution function is

$$f(y_i; \mu_i) = \frac{\mathrm{e}^{-\mu_i} \mu_i^{y_i}}{y_i!} = \mathrm{e}^{-\mu_i} \frac{1}{y_i!} \mathrm{e}^{y_i \log(\mu_i)} \tag{20}$$

for nonnegative integer values of y_i. For the Poisson distribution, a GLM links a monotone function of μ_i to explanatory variables (e.g., observers) through a linear model. The canonical link function is the log link such that

$$\log(\mu_{ij}) = \beta_0 + \sum_i \beta_i x_{ij}, \quad i = 1, \ldots, N, \ j = 1, 2, \ldots, n_i. \tag{21}$$

In this GLM, x_{ij} are the observer sunspot counts $j = 1, 2, \ldots, n_i$ of the $i = 1, 2, \ldots, N$ observers. Model 21 is called a loglinear model.

The matrix form of Equation 21 is

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}^*, \tag{22}$$

where $\boldsymbol{\mu_f}$ is the vector of mean sunspot numbers for each observer. $\boldsymbol{X}$ is the matrix of observer identifiers, the date and time of each observer's counts, the seeing conditions, and the experience level at the time of each observer's count. $\boldsymbol{\beta}$ is the parameter vector that is determined from maximum likelihood estimation.

The $\boldsymbol{X}$ matrix contains two types of predictor effects: random effects, which are a sample of all possible levels from the populations of these effects; and fixed effects, which are all possible levels of these effects. We can, therefore, partition $\boldsymbol{X}$ into a random effects component matrix $\boldsymbol{X_1}$, and a fixed effects component matrix $\boldsymbol{X_2}$. We now have a mixed loglinear model, a Generalized Linear Mixed Model (GLMM), that may be written as

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tag{23}$$

where $\boldsymbol{\mu}$ is the vector of mean sunspot numbers for each observer, $\boldsymbol{X}$ is the matrix of date and time indicators, seeing conditions, and the level of experience of the observers, and the $\boldsymbol{\beta}$ is the vector of fixed effects parameters, $\boldsymbol{Z}$ is the random effects matrix of observer identifiers, $\boldsymbol{u}$ is the random effects parameter vector to be estimated.

The Shapley approach to modeling sunspot numbers depends upon a sufficient transformation of a counts data distribution to follow a Gaussian distribution. The transformation is required to force homogeneous variance, though there are usually so-called outliers which, using a counts distriution, usually are not outliers. Bias is introduced when Gaussian outliers are assumed, and may need to be removed to obtain a stable transformation. Thus, information contained by the outliers is lost to the analysis. In addition, biased model parameters under-estimate the residual error of the model which often assigns significance to model coefficients which would be otherwise benign.

The GLMM is specifically designed to model counts data that follow a counts distribution. No data need be removed as the counts distribution is skewed in the direction of larger counts. The error structure used to model the counts distributions class accounts for overdispersion of residuals, when the mean of the counts is not equal to the variance of the counts. Hence, no information is eliminated due to the thick, right-tailed behavior of the counts distribution. Further, observer is a random effect that GLMM partitions into counts variance and observer variance. The observer random effect variance often follows a probability distribution that is different than that of the counts response, and GLMM allows for these different variance distributions in the same model. The Wald method must force all variance structures into Gaussian distributions, which can result in biased estimates, and may need adjustments over time. Dissimilar variance structure modeling in GLMMs leads to correct determination of unbiased model parameter estimation and significance. Modern GLMM construction produces monthly sunspot number estimates that are more efficient and consistent than the Shapley method for assumed Gaussian-distributed data. See Riggs and Lalonde (2017) for further information on counts models.

5 Generalized Linear Mixed Model Construction

We now examine the AAVSO sunspot count data. The analysis includes a description of the data set, parameter estimation via a quasi-Poisson-distributed loglinear mixed effects model, and an assessment of the validity of the model. The data spans from May 2010 through June, 2017. They were submitted to the AAVSO Solar Division by AAVSO members. The numbers of submissions by each member vary for each month, as does which members make the submissions. For this reason, observer is treated as a random effect in the modeling process. The quasi-Poisson distribution makes a linear adjustment to the Poisson variance to account for the sunspot numbers having variance larger than the mean:

$$\text{mean} = \mu, \ \text{variance} = \mu + \phi\mu, \tag{24}$$

in which ϕ is a multiplier to inflate the variance over the value of the mean. When $\phi = 0$, the quasi-Poisson distribution degenerates to the equi-dispersion Poisson distribution.

5.1 Sunspot Numbers Data Set

We begin the examination of the sunspot data set with an exploratory analysis. This analysis provides an overview of the data set contents that are important for constructing a sufficient model of sunspot numbers. We determine the probability distribution characteristics of the submitted counts and the observer random effect.

A summary of the data set in Tables 1 and 2 lists the number of cases by observer (the obs column), and by experience level (the r column) though not for all observers and experience. The number of cases of the observer seeing conditions (the see column) are given for all four levels of "E" for excellent, "F" for fair, "G" for good, and "P" for poor. The estimates of the minimum and maximum (Min and Max), the first and third (1st and 3rd) quartiles, the median (Median) and the mean (Mean) are given for year (the year column), month (the mon column), day (the day column), and Wolf number (the w column). A possible indicator that w does not follow a normal (Gaussian) distribution is shown by the large absolute difference (49 - 58.92 = 9.92) between the median and the means of the respective variables. This difference is greater than 15% of the interquartile range of 89 - 23 = 66. When the median and mean are approximately equal, the distribution is likely to be symmetric, which is a characteristic of a normal distribution. We test the assumption that w is normally distributed below.

Table 1: 201706 Summary of Sunspot Numbers

obs	jd	year	mon	day
ARAG : 2553	Min. :1721096	Min. :2010	Min. : 1.00	Min. : 1.00
CHAG : 2340	1st Qu.:2456060	1st Qu.:2012	1st Qu.: 4.00	1st Qu.: 8.00
BRAB : 2335	Median :2456646	Median :2013	Median : 7.00	Median :16.00
BROB : 2065	Mean :2456364	Mean :2013	Mean : 6.57	Mean :15.72
KNJS : 1950	3rd Qu.:2457281	3rd Qu.:2015	3rd Qu.: 9.00	3rd Qu.:23.00
HOWR : 1938	Max. :2457935	Max. :2017	Max. :12.00	Max. :31.00
(Other):46664				

Table 2: Summary of Sunspot Numbers

see	g	s	w	r	silso
E:10935	Min. : 0.000	Min. : 0.00	Min. : 0.00	0000A :24894	Min. :0.0000
F:18423	1st Qu.: 2.000	1st Qu.: 8.00	1st Qu.: 32.00	3000F : 9764	1st Qu.:0.0000
G:25556	Median : 4.000	Median : 19.00	Median : 60.00	2500E : 7766	Median :0.0000
P: 4931	Mean : 4.105	Mean : 25.45	Mean : 66.51	3500G : 4618	Mean :0.3288
	3rd Qu.: 6.000	3rd Qu.: 37.00	3rd Qu.: 95.00	1000B : 4228	3rd Qu.:1.0000
	Max. :18.000	Max. :204.00	Max. :293.00	1500C : 3059	Max. :1.0000
				(Other): 5516	

A example of the range of counts submitted across all the observers in any one month of data is shown in Figure 1. It is a plot of the minimum, maximum, and average of the daily submitted

counts for June, 2017. This type of plot is produced for each month of submitted sunspot counts as they are obtained from the observers.

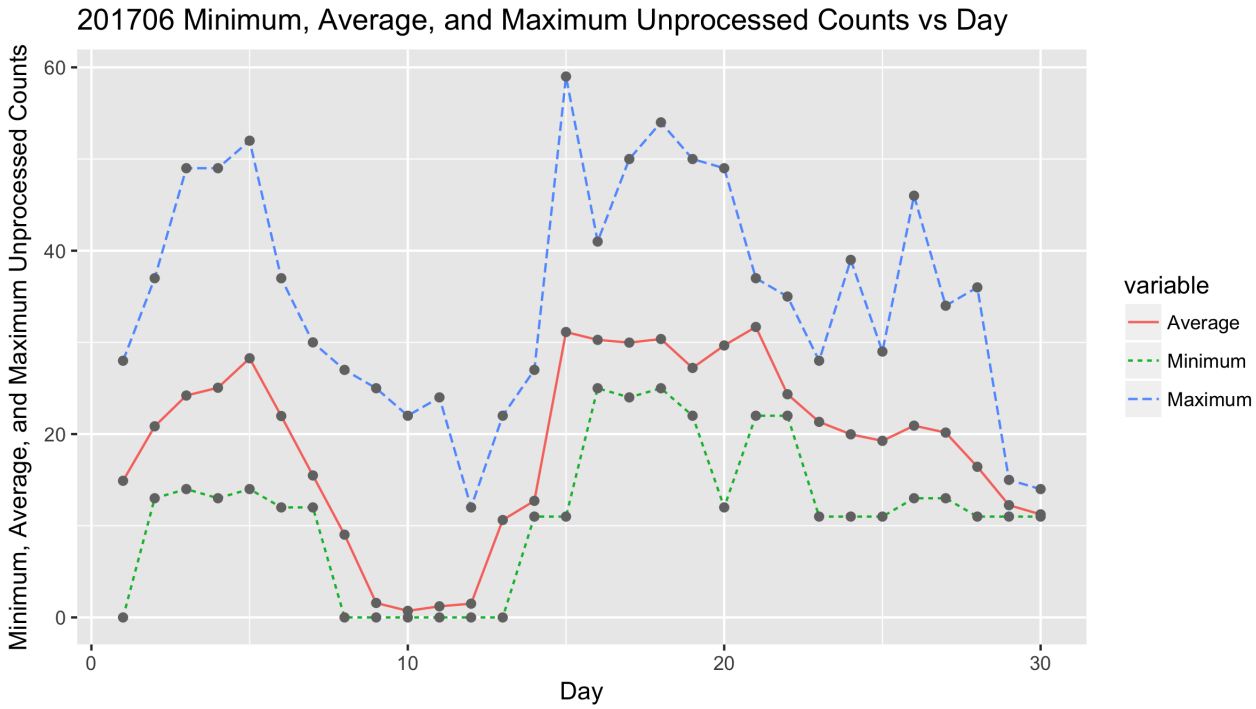

Figure 1: Raw average sunspot count by day of the month.

Figure 2 shows the pairwise scatter plots of the group number g, the sunspot count s, and the Wolf number w including a LOESS fit (solid red curve). The g versus s plot (top row, second panel) shows that as the number of sunspots s increases from 0 to 200, the number of groups g increases from 0 to 20, with several large group numbers of just a few sunspots. The scatter shape is increasing for both the groups and the sunspot numbers. The LOESS curve shows an increasing, convex bivariate relationship. The w versus s plot (last row, second panel) depicts a strong association for increasing w as s increases. The scatter appears nearly constant throughout the range of w. The LOESS fit is nearly linear. The w versus g (last row, first panel) also has a LOESS fit that appears linear and increasing w with increasing g. The covariance is non-constant throughout the range of w, first increasing at small to medium values of g and w, and then decreasing for medium to large values of g and w. The nonhomogeneous variance of w must be accounted for in a sunspot model.

Because sunspot numbers are cyclical, and because the counts data extend from a lower boundary of a quiescent Sun to an increasingly active Sun, ithen returning to a quiescent Sun, the month-over-month variability in sunspot numbers is first increasing and then decreasing. This may be seen in Figure 3. The median sunspot numbers (red dot in the boxes) show a slightly oscillating increasing trend beginning with the May 2010 box plot then decreasing to June 2017box plot. Fig-

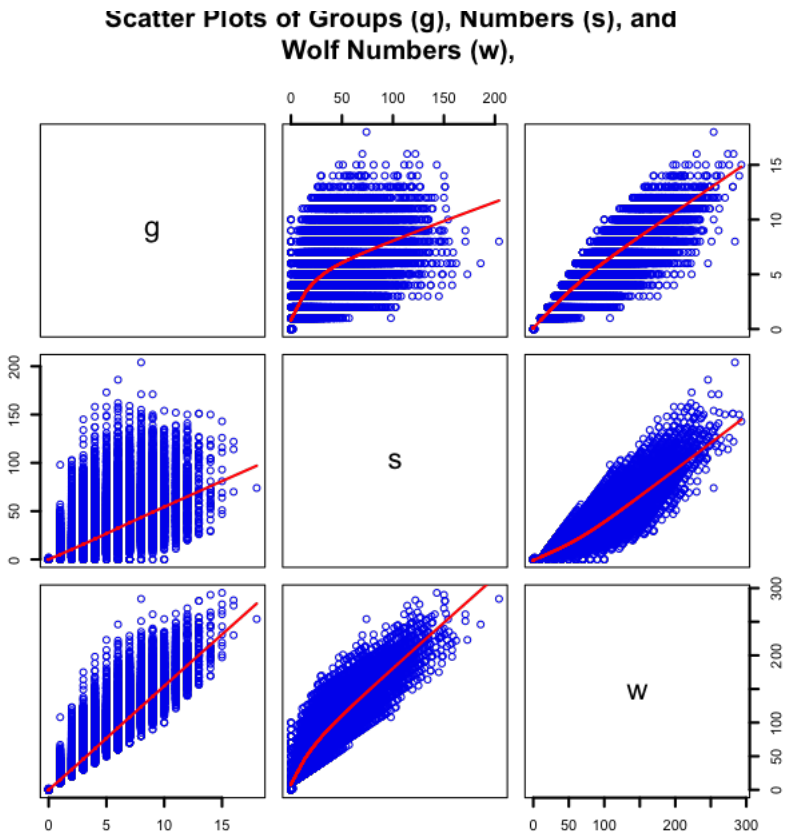

Figure 2: Scatter plots of sunspot groups g, numbers s, and Wolf numbers w with LOESS fits.

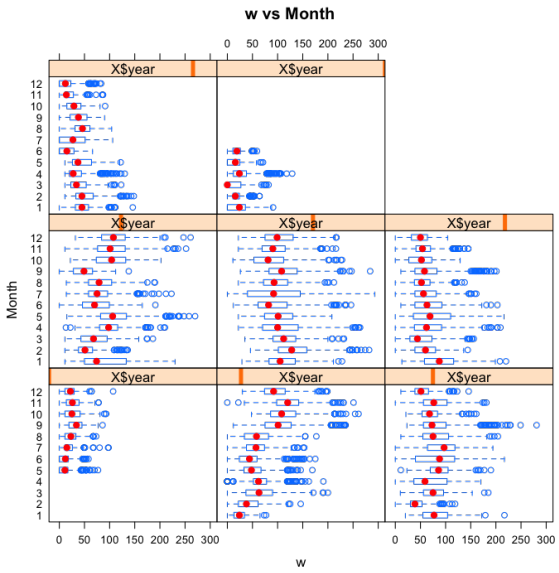

Figure 3: Box plots of raw Wolf number (w) by month and year.

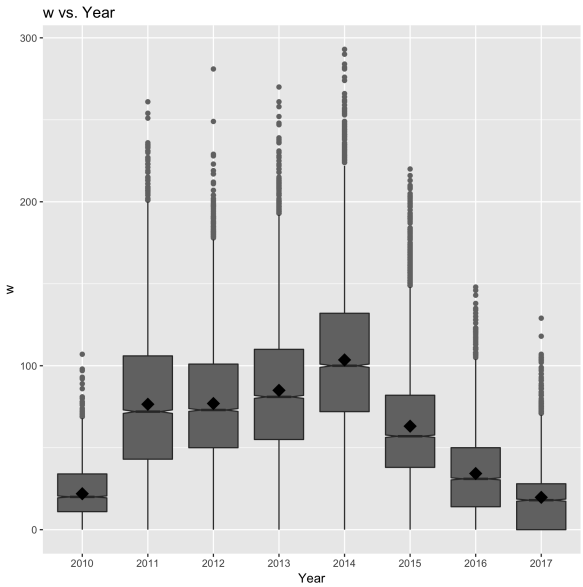

Figure 4: Box plots of raw Wolf number (w) by year.

ure 4 shows a clear difference among the years from 2010 and 2017. This information is important for accounting for the variance of the sunspot count model for time effects.

Figure 5 has box plots of the four levels of seeing conditions reported by the observers for each submitted sunspot count. The box plots show the skewness in the right tail of the distribution as indicated by the extended upper whiskers when compared with the lower whisker length. The 95% confidence-sized notches at the horizontal median bars indicate that the medians of the E, F, and G seeing levels aresimilar. These differ significantly from that of the median of the P level, which is expected.

The ranks in Figure 6 shows the nine differeing levels of experience of the AAVSO observers. Of the ranked observers, 1000B suggests an overall lower count. The model examined the significance of the rankings in earlier versions of the model. Until further examination, observer rank is removed from the current model. The unlabeld rank is of no consequence with rank removed from the model.

We now examine the distribution characteristics of the Wolf number w. Figure **??** shows a histogram of the w data. The bar heights are not symmetrical and skewed to the right. This is consistent with the normal Q-Q plot result that the w data are not normally distributed. TFigure 8 shows a normal quantile-quantile (Q-Q) plot. The black points of the w data do not follow the black solid line, which marks a normal distribution pattern, on the left side of the plot, w does not follow a normal distribution. However, we shall use the quasi-Poisson distribution in the construction of the counts model.

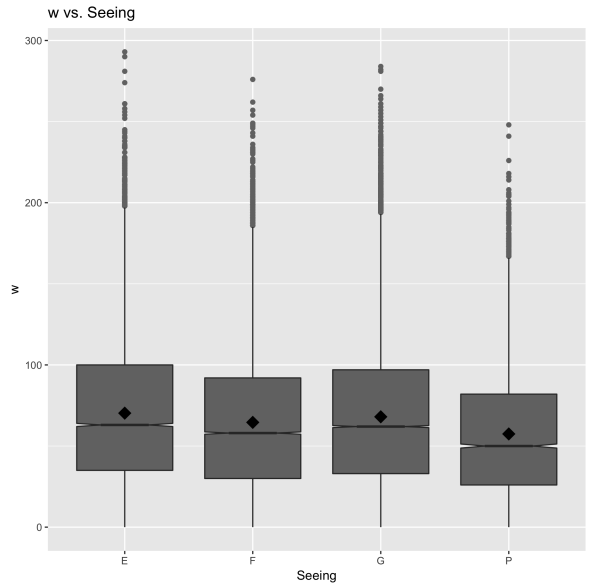

Figure 5: Box plots of raw Wolf number (w) by month and year.

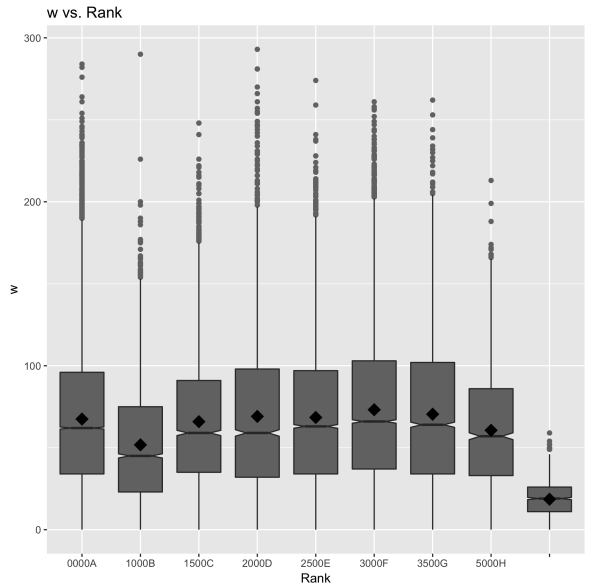

Figure 6: Box plots of raw Wolf number (w) by year.

5.2 Mixed Effects Loglinear Model Analysis

A generalized linear mixed model (GLMM) of the form in Equation 23 was fitted from the sunspot data. Unlike the Shapley method, all levels of observer experience are used as the model can discriminate the more experienced observers from novices. Also unlike the Shapley method, submitted zero counts are used as GLMM has no issues with zeros. The parameters of the GLMM are estimated using maximum likelihood estimation method (Laplace approximation) that is numerically equivalent to the numerical solution to the marginal density integral of y given the vectors $\boldsymbol{\beta}$ and $\boldsymbol{a}$. The GLMM reduces to:

$$y \propto \exp(\text{ year + month + seeing + observer }), \tag{25}$$

where μ is the model counts outcome, year (2010 and 2017) is the year the observations were submitted, month (1-12) is the month of the year the sunspots were counted, and seeing is the observing condition (poor, fair, good, excellent) reported by each observer. The observer is treated as a random effect, and the associated variance component is combined with the residual error to test the fixed effects. The variance component of the observer effect for the 59,845 observations is in Table 3.

The fixed effects analysis of the GLMM parameter estimates are in Table 4. Each of the levels of the fixed effects of year, month, and seeing condition, significantly accounts for a portion of the variability of the counts data. This is indicated by the less than 0.01 values of all the levels of the fixed effects in the $Pr(>|z|)$ column. The year and month significance is attributed in part to the cyclicity of the sunspot counts. Seeing conditions are significant for the obvious fact that

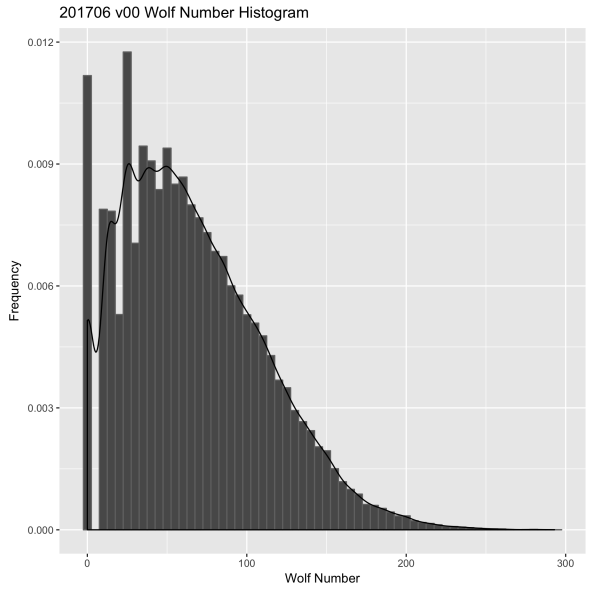

Figure 7: The non-Gaussian histigram for Wolf number w.

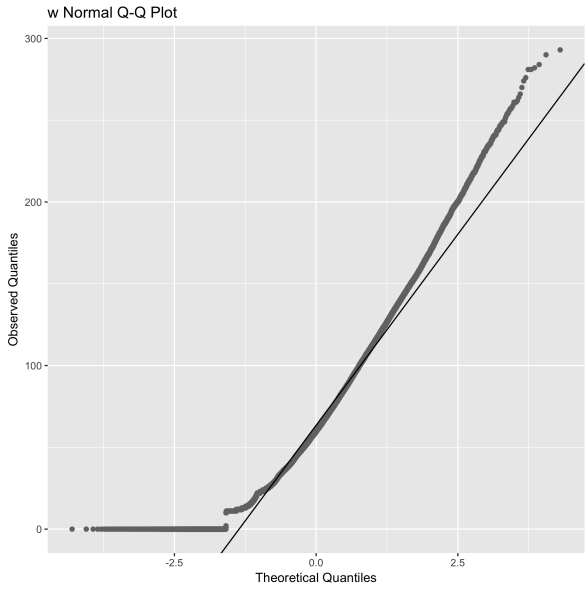

Figure 8: The normal Q-Q plot for Wolf number, w.

Table 3: Random effects values

Groups	Name	Variance	Std.Dev.
x1	(Intercept)	0.041895	0.20468

cloud cover obscuration, e.g., compromises the counting process. Experience level, i.e., the lifetime number of observer submissions did not significantly account for a portion of the variability in the sunspot counts in earlier models, so further work is needed to define observer experience levels.

The results discussed in the preceding paragraph dictate that the estimated sunspot numbers are determined from the the year, month, and seeing conditions that are reported daily. A discussion of the method used to determine these estimates is given in Section 5.4.

5.3 Mixed Effects Loglinear Model Diagnostics

Table 5 is a summary of the model fit statistics. The dev/df ratio of 13.10802 and the σ^2/μ ratio of 23.60726, both indicate the model is overdispersed even after compensating by using a quasi-Poisson PDF. This overdispersion may be due in part to incorrectly defined seeing condition levels, incorrect assumption on the distribution of the random effect (observer), or lack of additional explanatory variables such as type of equipment used. These possibilities will be explored in a future paper, including the use of a negative binomial PDF which provides more flexibility in defining the relationship between the mean and the vairance of the sunspot numbers.

Figure 9 shows four diagnostic plots that assess model fit. The Normal Q-Q plot (upper left

Table 4: 201706 Parameter Estimates

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	3.2323	0.0324	99.7725	0.0000
seeF	-0.1886	0.0071	-26.3757	0.0000
seeG	-0.1014	0.0062	-16.3149	0.0000
seeP	-0.2928	0.0105	-27.9321	0.0000
silso1	0.1058	0.0478	2.2118	0.0270
year2011	1.2053	0.0154	78.3659	0.0000
year2012	1.2234	0.0153	79.8128	0.0000
year2013	1.3184	0.0153	86.2540	0.0000
year2014	1.5070	0.0152	99.2556	0.0000
year2015	1.0089	0.0156	64.7658	0.0000
year2016	0.4038	0.0166	24.3746	0.0000
year2017	-0.1264	0.0207	-6.1011	0.0000
mon2	-0.1555	0.0118	-13.1558	0.0000
mon3	-0.0790	0.0109	-7.2487	0.0000
mon4	0.0311	0.0109	2.8562	0.0043
mon5	0.0333	0.0104	3.2129	0.0013
mon6	-0.1714	0.0109	-15.7266	0.0000
mon7	-0.0860	0.0106	-8.1068	0.0000
mon8	-0.0702	0.0104	-6.7206	0.0000
mon9	0.0607	0.0101	6.0288	0.0000
mon10	0.0081	0.0106	0.7586	0.4481
mon11	0.0463	0.0109	4.2478	0.0000
mon12	-0.0431	0.0116	-3.7200	0.0002

panel) of the residuals shows the upper tail of the residuals are concave (upward opening) over what is expected of a normal distribution. This is an indication that the residuals distribution is skewed to the right (in the direction of larger counts), which is in line with the overdispersion we saw earlier. So the distribution of the residuals is likely to follow a skewed distribution, commonly, a gamma PDF. The Residuals vs Fitted plot (lower left panel) shows a clear funnel shape indicating nonhomogeneous variance in the fitted values. The sharp truncation for small values of the residuals suggests the GLMM does not manage the inflation of the number of zero counts; a further feathure for future investigation using a zero-inflated PDF. The larger fitted values, $fitted(m)$, has more variance in the residuals $residual(m)$ than do the smaller fitted values. The model may be improved with an error structure defined by a gamma PDF.

Figure 9 also shows the Poisson variance by mean plot (upper right panel) shows non-constant variance across the mean values. Possible remedies include the use of a PDF with more flexible mean-to-variance relationship. Finally, the Residuals vs Sequence plot (lower right panel) reveals unexplained cycling over time along with time-varying variance known as volatility. This cycling is not unexpected, and is currently being investigated to incorporate time-dependent periodicity.

Table 5: Model Fit by the Laplace Approximation Estimation Method

AIC	BIC	logLik	deviance	df	dev/df	groups: x1	σ^2/μ
161447	161625	-80700	161399	12313	13.10802	53	23.60726

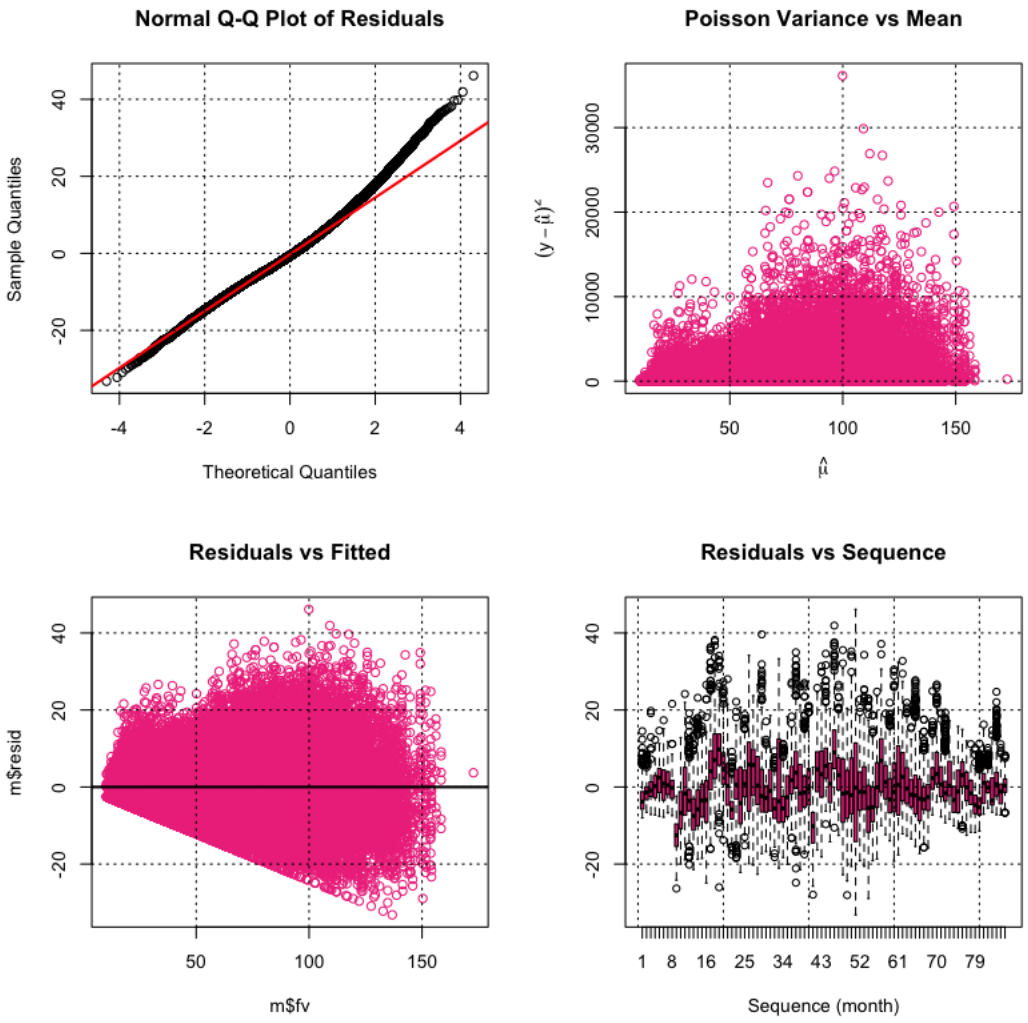

Figure 9: Generalized loglinear mixed model diagnostic plots.

5.4 R_a Estimation

We obtain the monthly GLMM-estimated means for sunspot numbers from the fixed effects. The variance of these estimates incorporates the observer random effect variance component. The monthly means are calculated by holding the seeing condition at the excellent level, then varying the year and month in the GLMM to produce the estimates. From Equation 22,

$$\begin{aligned} \boldsymbol{m} = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{*}) \Rightarrow \log(\boldsymbol{m}) &= \boldsymbol{X}\hat{\boldsymbol{\beta}}^{*} \\ &= \begin{bmatrix} \boldsymbol{j} & \boldsymbol{X_f} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} \\ &= \boldsymbol{X_f}\hat{\boldsymbol{\beta}} + \hat{\mu}\boldsymbol{j}, \end{aligned} \tag{26}$$

where $\boldsymbol{m}$ is the vector of model-fitted sunspot numbers; $\hat{\beta}^{*}$ is the vector of estimated model parameters that includes the overall mean parameter and the $p = 3$ coefficients of the x_1 to x_3 fixed effects of year, month, and seeing conditions, respectively, for the observers as random effects; and $\boldsymbol{j}$ is the $N \times 1$ vector of ones that is $N = \sum_{i=1}^{n} n_i,\ i = 1, 2, \ldots, n$ numbers of submissions for each of the n observers. Note that the $\hat{\boldsymbol{\beta}}^{*}$ vector has been partitioned into two parts, the overall mean $\hat{\mu}$ and the fixed effects parameter estimates $\hat{\beta}$. The design matrix $\boldsymbol{X}$ is the $N \times (p + 1)$ matrix of the indicators of the overall mean sunspot counts and the x_1 to x_3 fixed effects. Thus, noting that each column of the $\mathbf{X}$ matrix is labeled $\mu,\ x_1$ to x_3 for convenience, and the labels are not a part of the actual matrix,

$$\boldsymbol{X} = \begin{bmatrix} \mu & x_1 & x_2 & x_3 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \left[\boldsymbol{j}_{N\times 1}, \boldsymbol{X_f}_{N\times p}\right],$$

where the design matrix $\boldsymbol{X}$ is partitioned to be conformal with the partitioned parameter vector in Equation 26.

To obtain relative sunspot numbers, we solve Equation 26 for $\hat{\mu}$:

$$\begin{aligned} \hat{\mu}\boldsymbol{j} &= \log(\boldsymbol{m}) - \boldsymbol{X_f}\hat{\boldsymbol{\beta}} \\ \Rightarrow \mathrm{e}^{\hat{\mu}\boldsymbol{j}} &= \boldsymbol{m} - \mathrm{e}^{\boldsymbol{X_f}\hat{\boldsymbol{\beta}}} \\ \Rightarrow \theta\boldsymbol{j} &= \boldsymbol{\Delta} \end{aligned} \tag{27}$$

where $\theta = \mathrm{e}^{\hat{\mu}}$ and the vector $\boldsymbol{\Delta} = \boldsymbol{m} - \mathrm{e}^{\boldsymbol{X}\hat{\boldsymbol{\beta}}}$. Taking the means by month of the elements of the vectors on both sides of Equation 27 we have

$$R_A = \frac{1}{N}\sum_{i=1}^{N}\theta_i = \frac{1}{N}\sum_{i=1}^{N}\delta_i. \tag{28}$$

It is important to note that the variance of the random effect of observer has been accounted for by the model, and hence the variance of the monthly estimates incorporate the observer variance. The monthly mean sunspot numbers from the May 2010 through June 2017data set are given in Table 6.

Table 6: Year Month (ym) Relative Sunspot Numbers with 99% CI

ym	Ra	lci99	uci99	aavso	silso
2010.05	23.6771	23.1373	24.2169	8.4000	8.7000
2010.06	18.3859	17.8968	18.8750	11.0000	13.6000
2010.07	20.6128	20.1583	21.0673	15.2000	16.1000
2010.08	20.3170	19.8237	20.8103	18.3000	19.6000
2010.09	23.9106	23.3989	24.4223	22.8000	25.2000
2010.10	22.7334	22.2450	23.2219	21.0000	23.5000
2010.11	23.4543	22.9273	23.9813	20.9000	21.6000
2010.12	22.4281	21.7819	23.0743	13.9000	14.5000
2011.01	76.1685	74.4550	77.8819	17.7000	18.7000
2011.02	66.4790	64.9822	67.9759	29.1000	29.6000
2011.03	71.3016	69.8293	72.7738	48.0000	55.8000
2011.04	78.6047	76.9263	80.2832	47.3000	54.4000
2011.05	79.8776	78.2865	81.4686	37.3000	41.5000
2011.06	65.3251	63.9522	66.6980	35.2000	37.0000
2011.07	71.3799	69.8104	72.9494	41.5000	43.8000
2011.08	73.7195	72.2756	75.1634	42.4000	50.5000
2011.09	83.9009	82.7985	85.0033	73.8000	78.0000
2011.10	79.1393	77.7815	80.4972	78.9000	88.0000
2011.11	80.3451	78.6221	82.0681	84.6000	96.7000
2011.12	74.6987	73.0532	76.3442	65.8000	73.0000
2012.01	78.2991	76.7451	79.8531	55.8000	58.2000
2012.02	66.1308	64.7239	67.5377	29.2000	33.1000

Continued on next page

Table 6: Year Month (ym) Relative Sunspot Numbers with 99% CI

ym	Ra	lci99	uci99	aavso	silso
2012.03	73.7753	72.4565	75.0940	53.1000	64.1000
2012.04	78.3510	76.0409	80.6611	51.4000	55.2000
2012.05	83.8057	82.3436	85.2679	61.8000	69.0000
2012.06	68.3582	67.1488	69.5676	59.7000	64.5000
2012.07	75.6738	74.4001	76.9476	64.2000	51.3000
2012.08	74.4590	73.2146	75.7033	57.7000	63.1000
2012.09	84.7912	83.3408	86.2417	57.7000	61.5000
2012.10	81.5436	80.0029	83.0842	48.3000	53.3000
2012.11	83.8646	82.1836	85.5456	56.7000	61.4000
2012.12	75.7323	74.1392	77.3254	37.4000	40.8000
2013.01	88.1836	86.5359	89.8314	63.8000	62.9000
2013.02	76.2369	74.7503	77.7235	37.8000	38.0000
2013.03	80.9833	79.4589	82.5076	50.6000	57.9000
2013.04	91.2002	89.6655	92.7349	70.6000	72.4000
2013.05	91.4373	89.8621	93.0125	77.4000	78.7000
2013.06	75.2104	73.8749	76.5459	51.0000	52.5000
2013.07	81.2398	79.9760	82.5036	57.0000	57.0000
2013.08	82.0495	80.7687	83.3302	60.0000	66.0000
2013.09	92.5473	90.9477	94.1470	34.6000	36.9000
2013.10	87.4447	85.8942	88.9952	74.5000	85.6000
2013.11	90.0570	88.1851	91.9289	73.9000	77.6000
2013.12	83.3973	81.7211	85.0734	77.8000	90.3000
2014.01	104.7289	102.5265	106.9314	77.4000	82.0000
2014.02	90.5046	88.7851	92.2240	93.9000	102.8000
2014.03	99.6098	97.9287	101.2908	80.9000	92.2000
2014.04	110.8564	108.9752	112.7376	76.9000	84.7000
2014.05	110.7137	108.9429	112.4845	72.3000	75.2000
2014.06	91.0392	89.5636	92.5147	67.2000	71.0000
2014.07	99.6661	98.0408	101.2914	72.5000	72.5000
2014.08	100.1774	98.6732	101.6816	71.2000	74.7000
2014.09	114.1735	112.3314	116.0155	83.2000	87.6000
2014.10	107.6668	105.8658	109.4677	59.5000	60.6000
2014.11	111.4176	109.3160	113.5191	65.8000	71.1000
2014.12	100.8565	98.6682	103.0449	75.8000	78.0000
2015.01	63.9101	62.6783	65.1419	65.9000	67.0000
2015.02	55.2072	53.9184	56.4960	42.4000	44.8000
2015.03	59.5511	58.4580	60.6442	38.0000	38.4000
2015.04	67.0902	65.9001	68.2802	49.0000	54.4000
2015.05	66.6091	65.5393	67.6789	56.3000	58.8000

Continued on next page

Table 6: Year Month (ym) Relative Sunspot Numbers with 99% CI

ym	Ra	lci99	uci99	aavso	silso
2015.06	55.2559	54.3140	56.1979	50.2000	68.3000
2015.07	59.3031	58.2899	60.3162	47.9000	65.8000
2015.08	61.0534	60.0527	62.0542	39.5000	57.2000
2015.09	69.3058	68.1774	70.4342	49.2000	72.1000
2015.10	65.3241	64.2106	66.4375	39.3000	48.3000
2015.11	68.1038	67.1819	69.0257	39.6000	55.9000
2015.12	61.1754	59.9432	62.4076	36.4000	44.8000
2016.01	35.6519	35.0186	36.2852	33.7000	43.3000
2016.02	30.1633	29.5652	30.7614	38.3000	46.8000
2016.03	32.4183	31.8219	33.0146	30.5000	38.9000
2016.04	35.9003	35.2624	36.5383	26.6000	30.9000
2016.05	36.6282	36.0022	37.2542	33.7000	48.4000
2016.06	30.0087	29.5326	30.4849	13.1000	19.5000
2016.07	32.8929	32.3849	33.4010	21.2000	27.5000
2016.08	33.5176	32.9657	34.0696	33.0000	47.9000
2016.09	37.8122	37.1848	38.4396	27.7000	37.1000
2016.10	35.8743	35.2526	36.4959	22.7000	31.7000
2016.11	36.9875	36.3505	37.6245	14.0000	22.2000
2016.12	33.5318	32.8592	34.2044	11.1000	20.0000
2017.01	20.9517	20.5529	21.3505	18.4000	26.2000
2017.02	17.5624	17.2182	17.9066	14.4000	20.6000
2017.03	19.2990	18.9575	19.6404	11.3000	15.5000
2017.04	21.7835	21.4286	22.1384	21.6000	33.2000
2017.05	21.6594	21.3064	22.0124	12.5000	18.1000
2017.06	17.8449	17.6146	18.0751	15.5000	19.3000

Figure 10 is a plot of the monthly data used to construct the GLMM, the GLMM monthly mean sunspot numbers, and the Solar Influence Data Center (SIDC) estimates found on the National Oceanic and Atmospheric Administration's (NOAA) web site for the NOAA National Geophysical Data Center SIDC values. Each box and whisker plot is the summarized the monthly data used to construct the loglinear mixed model. The solid bar near the center of each box is the median of the data. The boxes themselves represent the interquartile range, IQR, (25^{th} to 75^{th} percentiles) of the data. The whiskers are terminated by short horizontal bars, and cover from the monthly data from 1.5*IRQ below the first quartile to 1.5*IQR above the third quartile. The points beyond the whiskers are values that contribute to the overdispersion of the residuals.

Figure 10 also shows a solid red curve that connects the boxes of the GLMM monthly sunspot fitted values when the seeing conditions are excellent. The solid green curve is the AAVSO monthly mean sunspot numbers. The AAVSO estimates are obtained using the Shapley/Wald model described in 2 above. The solid blue curve is the SIDC values. Both the AAVSO and SIDC numbers

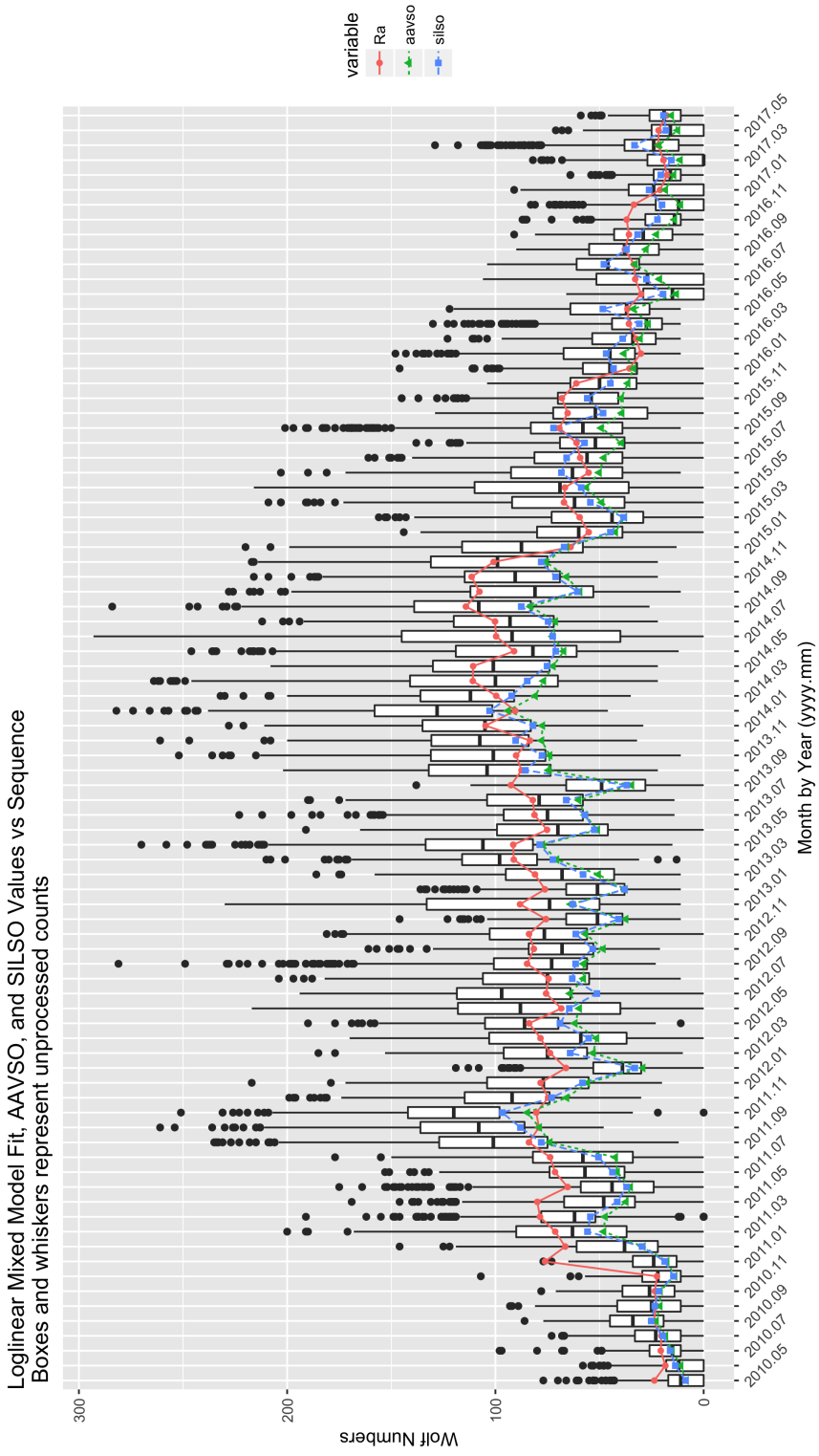

Figure 10: GLMM fitted data for R_a. AAVSO data: https://www.aavso.org/category/tags/solar-bulletin. SILSO data: WDC-SILSO, Royal Observatory of Belgium, Brussels

elimate counts from fair and poor seeing conditions. In addition, they force the counts data to approximately follow a Gaussian distribution and remove counts of zeros. These analyses bias the total amont of counts variance which explains in part the differences from the GLMM numbers.

Shapley (1949) removed zeros from his analysis. This is to avoid taking the natural logarithm of zero, which is not analytically tractable. Counts of zero are not a problem with models using the quasi-Poisson distribution in which zero is a valid count, and thus presents no analytical issues especially, as was seen in Figure **??**, these data are zero-inflated. Shapley also assumes that between-observer covariance is zero. This assumption is not required when treating the observer as a random effect, and the correlation structure is under investigation to further inprove the GLMM. Also, no minimum sunspot counts are needed from each observer to fit the GLMM.

The Shapley model forces the slop term to be one, which, as stated earlier, is generally an assumption that is strictly data set dependent. The GLMM makes no such assumption on the fixed effects slope parameters, and thereby make the GLMM more robust to changes in future reported sunspot counts. We see then, that as the GLMM is designed specifically for counts data, the conditions imposed by Shapley for the Wald-based model are not necessary for unbiased sunspot number estimation.

6 GLMM Improvements

There are three basic areas for model improvement. The first area for improvement is to test and incorporate additional viable fixed effects. The known fixed effects slated for test are observer experience, the image magnification, type of filter, and the method of observation of each observer. The correct magnification is required not just to assure all possible sunspots can be resolved, but also to maintain consistency with historical observations. Two filter types are in common use and are known as white filters and Hydrogen α filters. The two methods are projection and direct observation. Projection focuses the solar image onto a screen and direct observation is the solar image is focused on the retina. A new model will incorporate magnification as a continuous effect, and the filter and method will each be represented as a nominal effect.

A current and major area of investigation in GLMM's is the treatment of the random effects, which is the second area for model improvement. As was noted in Sections 5.2 and 5.3, the residuals are overdispersed. This is likely due to latent population heterogeneity in the sunspot counts. A possible source for this overdispersion is how the observer random effect is treated. This model assumes the random effect follows a normal distribution and are uncorrelated. Future models will explore the use of conjugate pairs for overdispersion reduction. It is apparent in the top half of Table 7 that the mean-variance ratio is reduced toward one as the treatment of the observer effect begins as fixed, analyzed as if normally distributed, with the best ratio for the Poisson-Gamma conjugate pair. If the observer counts have either or both within observer or across observer correlation, the correlation structure then can be accounted for in the error structure of the model, thus segregating known sources of variation from the ubiquitous random variation. This segregation will reduce the amount of random variation in the model residuals which is a desireable property for empirical models.

Finally, the third area for improvement is the identification of the cyclical behavior in the

Table 7: Improvements from Error Structure Changes

$\boldsymbol{\eta}\|\boldsymbol{u}$ Dist	Link $g(\boldsymbol{\mu})$	$\boldsymbol{u}$ Dist	Link $v(\boldsymbol{u})$	Method	$s^2/\bar{x}$
Poisson	log	fixed	NA	GLS	22.87
Poisson	log	Normal	identity	log-likelihood	21.66
Poisson	log	Gamma	log	h-likelihood	18.49
Poisson	log	Poisson	identity	h-likelihood	?
Gamma	log	Gamma	identity	h-likelihood	?
Gamma	inverse	inverse Gamma	inversey	h-likelihood	?

residuals for incorporation into the model. Sunspot cycles occur on both long term and short term periods. As the current model is only one solar cycle long (cycle number 24), only the short term periodicity can be studied and modeled. Sunspot number monthly periodicity is currently under investigation.

7 Conclusions

We began the sunspot modeling process with a description of the work preceding our use of generalized linear mixed models (GLMM) for estimating monthly sunspot numbers. The description included discussions on fitting straight lines to two variables each with measurement errors, and how these methods were used by Shapley (1949) from a technique developed by Wald (1940). This was followed by a short discourse on the applicability of the loglinear quasi-Poisson GLMM.

Data supplied by the American Association of Variable Star Observers Solar Division spanning the period from May 2010 through June 2017were used to fit a GLMM. We showed each predictor's role in the model, and how these predictors contributed to the fit of the model. We found the GLMM has overdispersion that needs further attention, and we discussed remedial measures and model improvements in general.

The method for determining monthly sunspot counts, adjusted for the variance due to the random effects, and after accounting for the variability in counts due to time in the solar cycle, seeing conditions, and experience was presented.

Finally, we evaluated the GLMM assumptions against the Shapley assumptions and conditions and found the GLMM to have desirable properties for making estimates of monthly sunspot numbers. The GLMM method may be applied in the presence of reference counts to improve sunspot number accuracy. This GLMM method, in combination with other solar activity measures, can be used to generate a solar activity index that may represent underlying solar physical processes.

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc, New York, 3rd edition, 1998.

Frederic Clette, David Berghmans, Petra Vanlommel, Ronald A.M. Van der Lindon, Andre Koeckelenbergh, and Laurence Wauters. From the wolf number to the international sunspot index: 25 years of sidc. *Advances in Space Research*, 40:919–928, 2007.

Grant Foster. Inflation of aavso sunspot counts. *Journal of the American Association of Variable Star Observers*, 26:50 – 58, 1997.

A. Madansky. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54(285):173–205, 1959. ISSN 01621459. URL `http://www.jstor.org/stable/2282145`.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, Florida, 1989.

J.D. Riggs and T.L. Lalonde. *Handbook for Applied Modeling: Non-Gaussian and Correlated Data*. Cambridge University Press, 2017. ISBN 9781316601051. URL `https://books.google.com/books?id=ZtnAnQAACAAJ`.

B. E. Schaefer. Visibility of sunspots. *Astrophysical Journal*, 411:909–919, July 1993.

Bradley E. Schaefer. Automatic inflation in the aavso sunspot number. *Journal of the American Association of Variable Star Observers*, 26(1):40–46, 1997.

A.H. Shapley. Reduction of sunspot-number observations. *Publication of the Astronomical Society of the Pacific*, 61(358):13–21, February 1949.

P.O Taylor. The Comptation of American Relative Sunspot Numbers. *Journal of the American Association of Variable Star Observers*, 14:28 – 32, 1985.

A. Wald. The fitting of straight lines if both variables are subject to error. *Annals Mathematical Statistics*, 11(3):284–300, 1940.

M. Waldmeier. *The sunspot-activity in the years 1610-1960*. Verlag Schulthess u. Co. AG., Zurich, 1961.