Generalized Linear Mixed Models for Time Trends in Hawk Migrations

M.S. Plan B Project Report

August 12th, 2013

Marie Helbach

M.S. Applied and Computational Mathematics Candidate

Dr. Ronald Regal

Advisor

University of Minnesota Duluth

Department of Mathematics and Statistics

Table of Contents

i. Acknowledgments

I'd like to take this opportunity to thank Dr. Ronald Regal for all of the guidance he's given me over the past year and a half. My ability to apply what I've learned in real world situations is due to the material we've worked on together.

I'd also like to thank Dr. Richard Green, Dr. Xuan Li, and Dr. Gerald Neimi for serving on my degree committee, reviewing my paper and providing useful suggestions. Dr. Li, thank you for joining at the last minute.

I'd also like to thank Hawk Ridge Bird Observatory for providing the dataset in this project.

Thank you to the UMD Mathematics Department for giving me the opportunity to test my potential.

ii. Abstract

Hawk Ridge in Duluth, MN is known as one of the nation's largest fall migrating raptor observation sites, second only to Hawk Mountain in Pennsylvania. Hawk Ridge Bird Observatory has been conducting annual fall surveys since before 1972, in an effort to track annual population trends in over fifteen species of raptors. Testing annual trends in the broad-winged hawk population was the focus of my project. The data available for such analysis were in the form of counts. Over-dispersion, a high number of zero counts, lack of independence, and inconsistent survey durations were all issues with these data. Ideally a nonlinear model with random effects and the consideration of the Poisson, negative binomial, and zero-inflated distributions would be most appropriate. SAS has the capability to fit some models of this type but often encounters convergence issues. To avoid convergence issues, my approach made use of a set of generalized linear and generalized linear mixed models. Based on my results, the counts of broad-winged hawks do not show a significant decline or increase since 1972.

1 Introduction

Hawk Ridge in Duluth, MN, is one of the major sites in the nation for observing/surveying fall migrating raptors such as the bald eagle, broad-winged hawk, and peregrine falcon. Part-time counting efforts began in the 1950's, motivated in part by the then common practice of using the birds for target practice. It wasn't until 1972 that a complete and consistent full-time count began at Hawk Ridge, facilitated by the Hawk Ridge Bird Observatory. Protecting birds of prey in the Western Lake Superior Region is one of the missions of the observatory. Estimating and testing annual trends in fall migration counts contributes to this mission and serves as the focus of this project.

In particular, the goal of this project is to test the annual trend in fall migrating broad-winged hawks (BW) at Hawk Ridge. The data were provided by the Hawk Ridge Bird Observatory and contain counts of fifteen species of fall migrating raptors, collected from 1972 to 2011 between the months of August and December and between the hours of 0500 and 2000. The counts are best described as "hourly" although the actual duration of each survey ranges from 0 – 630 minutes. The most common duration was 60 minutes, accounting for roughly 90% of the surveys. At times however, counts were only recorded for the entire day; this is true for all of 2005 and 2006.

The figures below show trends of completeness of the surveys over time. Figure 1 was generated using all recorded surveys, displaying that over time the days in which the surveys occur expanded, particularly to later in the season. Recently, the surveys have ended on November 30th, which is 120 days since August 1st.

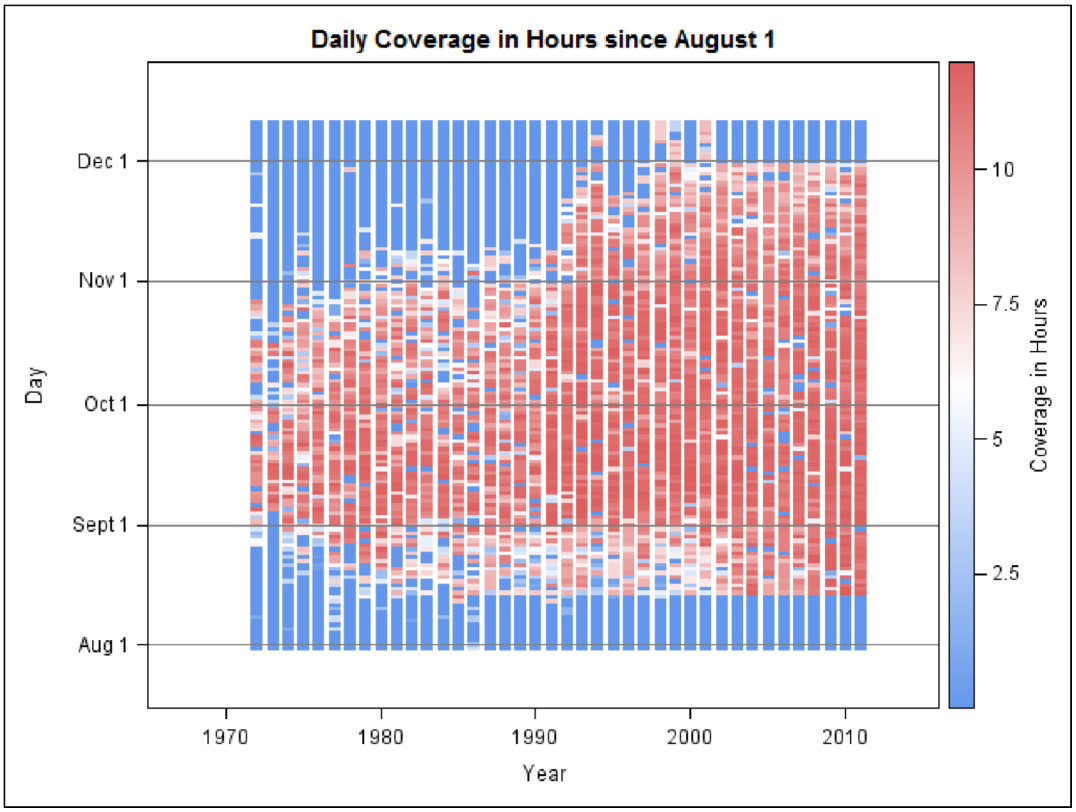

Figure 1: Daily Coverage in Hours since August 1st

Figure 2 was generated using only the surveys that occurred within the BW seasonal window (August 23rd to October 1st, explained in Section 2). In comparison to Figure 1, the surveys in Figure 2 are more consistently complete over time.

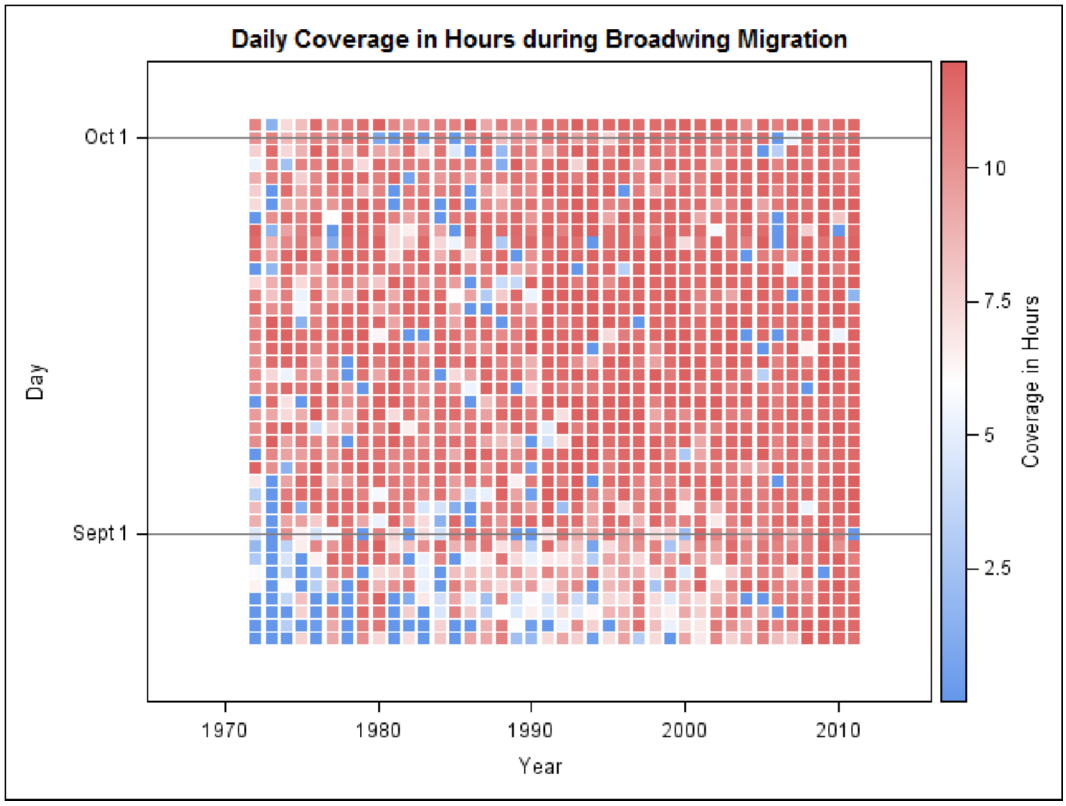

Figure 2: Daily Coverage in Hours since August 1st during BW Migration Window

Again using all of the surveys, Figure 3 reveals that most of the surveying occurred between the hours of 0800 and 1500. In later years there is a slight shift to include more representation earlier and later in the day.

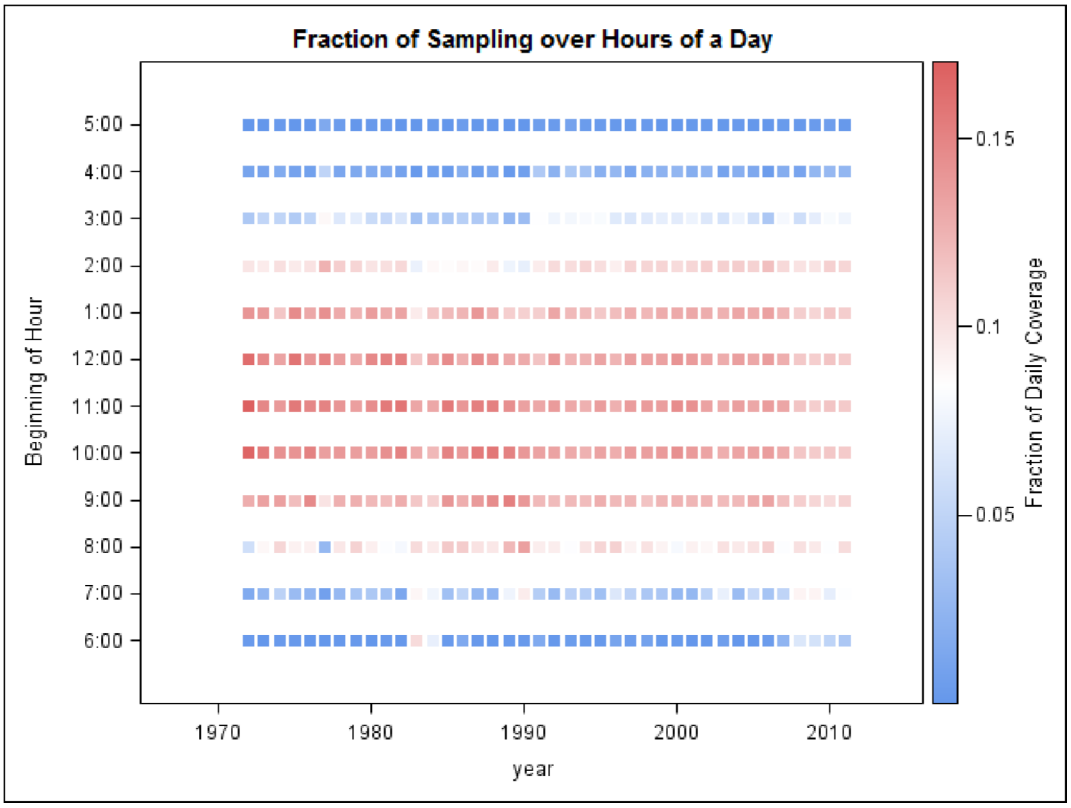

Figure 3: Fraction of Sampling over Hours of a Day

Figure 4 shows the range of days in which certain percentages of the year's BWs were seen. The days in which the BWs are seen remains fairly constant over time and the survey period (August – December) covers the migration completely.

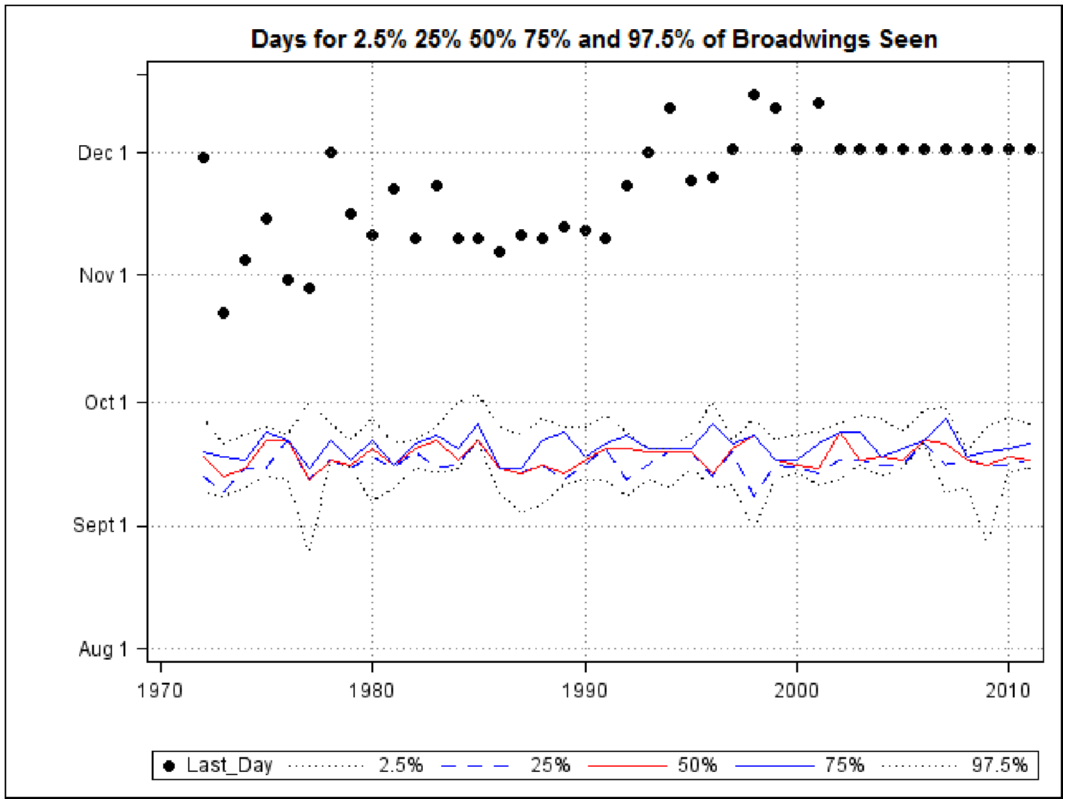

Figure 4: Days in which a certain % of BWs were seen.

In contrast, as seen in Figure 5, the bald eagle migration extends up to day 120 or November 30, the last day of recent surveys. The apparent trend of later arrivals in the plot is biased by surveys occurring later in the season in later years.

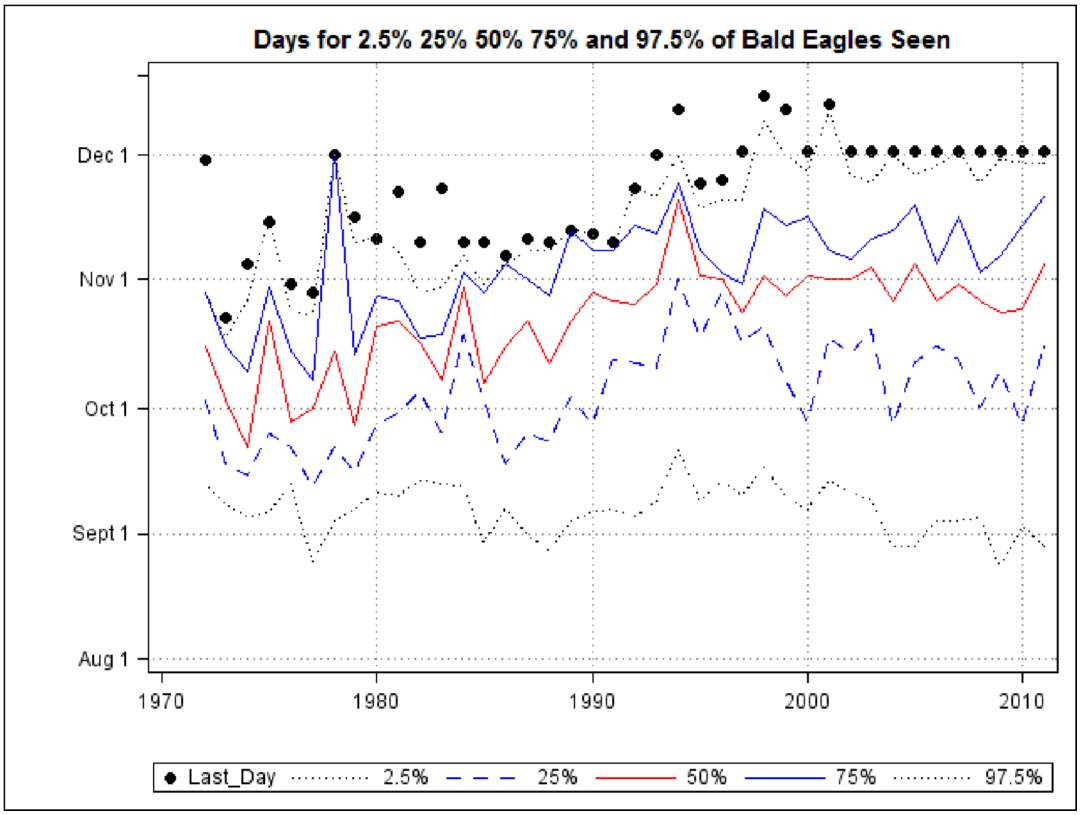

Figure 5: Days in which a certain % of bald eagles were seen.

Given the above observations, estimating the annual bald eagle population would require extrapolation of trends beyond the surveyed dates and consideration given to the possibility of the seasonal trend depending on years and day of the season by year interaction. Assessing whether the migration has trended later in the year would require extrapolation of counts in early years to later periods of the year where surveys were not conducted. Would bald eagles have arrived in the later periods that weren't surveyed? These issues are mentioned to highlight a complexity that doesn't exist in the broad-winged data. Complete surveying of the BW migration window and no apparent interaction between day of the season and year permits simpler models.

To estimate and test annual trends, an annual BW population index was derived. The index represents the sum of the year's counts, accompanied by an adjusted duration. The adjusted

duration represents the adjusted number of days in which the year's counts were gathered. In pursuit of this index, daily counts within each year accompanied by their own adjusted duration, in hours, were utilized. Estimates of the yearly and daily adjusted durations were obtained with the use of generalized linear models. Testing for trends in the annual index included the use of generalized linear mixed models. These models were chosen in part due to the potential non-normality of the data as well as the potential presence of additional random variability contributed by each year.

This paper continues with an in-depth description of the data, including reasons behind certain data cleaning decisions and the implementation of constraints, a general introduction to the statistical models used/investigated, and an in-depth description of specific models used. The paper concludes with the results of the tests for annual trends in the BW counts. SAS software was used extensively in the investigative and analysis process, code for such is included in the Appendix.

2 The Data

As previously mentioned, the data received from the Hawk Ridge Bird Observatory contains counts collected from 1972 to 2011 between the months of August and December and between the hours of 0500 and 2000. Accompanying each reported count are the following pieces of information: *date*, *start* (time the survey began), *end* (time the survey ended), and *duration* (reported duration of the survey, in minutes). The counts are best described as "hourly" although the actual *duration* accompanying each ranges from 0 – 630 minutes. Initial investigation of the data revealed the need for cleaning. The cleaning process is described in detail and can be found in the Appendix. In addition, in order to minimize the number of zero counts, hourly and seasonal windows were established.

2.1 Hourly and Seasonal Windows

The hourly and seasonal windows were established similar to Farmer et al. (2007). Using the cleaned data set, the hourly window was determined to be 0600 – 1800, representing the union of the yearly hourly windows. Each yearly hourly window includes at least 95% of the BWs counted that year. In terms of the variables in the data set, the window 0600 – 1800 refers to a *start* time ≥ 0600 and an *end* time ≤ 1800. It should be noted that the hourly window was influenced by days when the counts were recorded only for the entire day and not broken down into hourly counts. For example on September 29th, 2005, one survey was reported of which included a count of13 BWs, a *start* time of 0615, and an *end* time of 1545. The seasonal window was determined in the same fashion, being the union of the yearly seasonal windows, resulting in days ranging from August 23rd to October 1st (i.e. 22 to 61 days since August 1st). Counts obtained outside these two windows were excluded from the data used in the analyses.

3 Statistical Framework

3.1 Distributions Considered

Counts are often assumed to originate from a Poisson distribution. A random variable, *Y*, is said to have the Poisson distribution if it has a discrete probability density function (pdf) of the form

$$P(Y = y) = f(y) = e^{-\mu}\frac{(\mu)^y}{y!} \text{ where } \mu > 0 \text{ and } \mathrm{y} = 0, 1, 2, \ldots$$

A property of this distribution is the equality of the mean and variance: $E(Y) = Var(Y) = \mu$. Often with real count data, this property is not met. When the sample variance is larger than the sample mean, it is called over-dispersion. A common way to account for this possibility is to assume the random variable originates from a negative binomial distribution. Under this assumption, the mean and variance are no longer required to be equal: $E(Y) = \mu$ and $Var(Y) = \mu + k\mu^2$where k is referred to as the dispersion parameter. A random variable is said to have the negative binomial distribution if it has a pdf of the form

$$P(Y = y) = f(y) = \binom{\mathrm{y} + \alpha - 1}{\alpha - 1} p^y (1 - p)^{\alpha}$$

For the purpose of fitting generalized linear models, the above pdf is often re-parameterized in terms ofμ, the mean, and the dispersion parameter (SAS/Stat 9.3 User's Guide, 2013)

$$\frac{\Gamma\left(\mathrm{y}+\frac{1}{k}\right)}{\Gamma(\mathrm{y}+1)\Gamma\left(\frac{1}{k}\right)}\frac{(\mathrm{k}\mu)^{\mathrm{y}}}{(1+\mathrm{k}\mu)^{\mathrm{y}+\frac{1}{k}}}$$

The negative binomial distribution is a generalization of the Poisson, the Poisson the resulting distribution as the dispersion parameter approaches zero.

The negative binomial distribution is also a Gamma-Poisson Mixture. This Gamma-Poisson Mixture can be explained within the context of our BW counting process: let $\{Y(t), t \geq 0\}$ represent our BW counting process such that the number of BWs counted in any interval of length *t*, $Y(t)$, conditional on some observed value of a positive random variable,Λ, is a conditional or mixture Poisson process with rate λ (i.e. $Y(t)|\Lambda = \lambda \sim Poi(\lambda t)$ for $\lambda > 0$). If Λ follows a gamma distribution, the resultant distribution of $Y(t)$ is negative binomial (Ross, 2010). For example, suppose we are interested in the distribution of $Y(1)$ and $\Lambda \sim GAM(\theta, \alpha)$ such that $= \frac{p}{1-p}$, then

$$\begin{aligned}
P(Y(1)=y) = f(y) &= \int_0^\infty f(y|\lambda) f(\lambda) d\lambda \\
&= \int_0^\infty e^{-\lambda}\frac{(\lambda)^{\mathrm{y}}}{\mathrm{y!}}\frac{1}{\theta^{\alpha}\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\frac{\lambda}{\theta}}\,\mathrm{d}\lambda \\
&= \frac{1}{\mathrm{y!}\,\theta^{\alpha}\Gamma(\alpha)}\int_0^\infty (\lambda)^{\mathrm{y}+\alpha-1}\,e^{-\left(\lambda+\frac{\lambda}{\theta}\right)}\mathrm{d}\lambda \\
&= \frac{\left(\frac{\theta}{\theta+1}\right)^{y+\alpha}\Gamma(y+\alpha)}{\mathrm{y!}\,\theta^{\alpha}\Gamma(\alpha)}\int_0^\infty \frac{1}{\left(\frac{\theta}{\theta+1}\right)^{y+\alpha}\Gamma(y+\alpha)}\,e^{-\lambda\left(\frac{\theta+1}{\theta}\right)}\,(\lambda)^{\mathrm{y}+\alpha-1}\mathrm{d}\lambda \\
&= \frac{\left(\frac{\theta}{\theta+1}\right)^{y+\alpha}\Gamma(y+\alpha)}{\mathrm{y!}\,\theta^{\alpha}\Gamma(\alpha)} \\
&= \frac{(y+\alpha-1)!}{\mathrm{y!}\,(\alpha-1)!}\left(\frac{\theta}{\theta+1}\right)^{y}\left(\frac{1}{\theta+1}\right)^{\alpha}
\end{aligned}$$

$$= \binom{\mathrm{y} + \alpha - 1}{\alpha - 1} p^{y}(1-p)^{\alpha}$$

It is critical to note that the definition of a Poisson process includes the requirement that the increments are independent (e.g. the number of BWs counted between 0800 and 0900 on September 16, 2011 are independent of the number of BWs counted between 0900 and 1000 on the same day). Common sense suggests this is not the case. It is more likely that on certain days there will be a wave of BWs flying over Hawk Ridge. Take for example, September 15, 2003 when a total of 93,104 BWs were counted. Over 20,000 were counted each hour from 13001700. Treating these hourly counts as independent observations would result in estimates of trends that have more precision than warranted, resulting in claimed significance of a time trend when no such trend exists. This lack of independence motivates the investigation and use of models that include random effects, explained in later sections.

Another common occurrence with count data is an excessive amount of zeros (i.e. a higher incidence of zero counts than expected for the underlying distribution). When this is the case, zero-inflated distributions are commonly employed. Zero-inflated distributions assume the random variable originates from one of two possible data generation processes. The first process generates a random variable from either a Poisson or negative binomial distribution (i.e. the underlying distribution). The second process always produces a zero count. Letting $\lambda =$ the mean of random variable from the underlying distribution with parameter vector $\boldsymbol{\theta}$, and pdf = $g(y;\boldsymbol{\theta})$ and ω (called the zero inflation probability) the probability of being from the second process, the probability distribution of Y has the following form (SAS/Stat 9.3 User's Guide, 2013)

$$f(y;\boldsymbol{\theta},\omega) = \begin{cases} \omega + (1-\omega)g(0;\boldsymbol{\theta}) \text{for } y = 0 \\ (1-\omega)g(y;\boldsymbol{\theta}) \quad \text{for } y = 1, 2, \ldots \end{cases}$$

The $E(Y) = (1-\omega)\lambda$, regardless of the underlying distribution. The variance of Y depends on the underlying distribution

$$Var(Y) = \begin{cases} \mu + \left(\frac{\omega}{1-\omega}\right)\mu^2 & \text{if Poisson} \\ \mu + \left(\frac{\omega}{1-\omega} + \frac{k}{1-\omega}\right)\mu^2 & \text{if negative binomial} \end{cases}$$

Assuming the counts originate from any of the four distributions described above, the general linear model, with the assumption of a normally distributed response, falls short of being an appropriate tool in the analysis process. The generalized linear model allows for non-normal data and thus will serve essential. As previously stated, common sense suggests the counts are not independent. Generalized linear models assume the response variables are independent, thus motivating the investigation and potential use of generalized linear mixed models which allow not only for non-normal data but also correlated data.

3.2 Generalized Linear Model

The basic assumptions of the generalized linear model are that the set of independent random variables, $Y_1, Y_2, \ldots, Y_n$, are from the same distribution and that distribution is a member of the exponential family of distributions. A distribution is a member of the exponential family if the density function can be expressed in the following form (Montgomery, 2006)

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + h(y,\phi)\right\}$$

where θis the location parameter and ϕ the scale parameter. The basic idea of the generalized linear model is that some monotonic differentiable function, called the link function, of the expected value of the response variable, Y_i, can be expressed as a linear function of some set of parameters,$\beta_1, \beta_2, \ldots \beta_p$, and explanatory variables, $x_{i1}, x_{i2}, \ldots, x_{ip}$:

$$g[E(Y_i)] = g(\mu_i) = \boldsymbol{x_i'\beta} = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$\Rightarrow \mu_i = g^{-1}(\boldsymbol{x_i'\beta})$$

For Poisson or negative binomial distributions, the usual link function is the natural log. In most statistical literature this is written as "log" rather than "ln". For a Poisson model with log link we have

$$log(\mu_i) = \sum_{j=1}^{p} x_{ij}\beta_j$$

$$\Rightarrow \mu_i = exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$$

$$\Rightarrow f(y_i) = e^{-\mu_i}\frac{{\mu_i}^{y_i}}{y_i!} = \frac{exp\left(-exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\right)exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)^{y_i}}{y_i!}$$

To estimate the β parameters, we need to maximize the log likelihood with respect to the βs:

$$log\big(L(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} -\mu_i + \sum_{i=1}^{n} y_i log(\mu_i) - \sum_{i=1}^{n} log\ (y_i!)$$

$$= \sum_{i=1}^{n} -exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) + \sum_{i=1}^{n} y_i \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{i=1}^{n} log\ (y_i!)$$

Unlike regression with the assumed distribution being normal, there is no closed form expression for the maximum likelihood estimates of $\beta_1,\ \beta_2, \dots, \beta_p$. Rather, iterative numerical methods must be used to find the maximum likelihood parameter estimates. Negative binomial distributions are estimated similarly except that an extra parameter, the dispersion, needs to be estimated. The zero-inflated distribution requires the estimation of an extra set of β parameters due to the second process from which the random variable may have originated. In this case, some function of the zero inflation probability, ω_i, is expressed as a linear function of some additional parameters $\beta_{p+1}, \beta_{p+2}, \dots \beta_{p+p}$, and explanatory variables, $x_{i1}, x_{i2}, \dots, x_{ip}$. With a logit link

$$log\left(\frac{\omega_i}{1-\omega_i}\right) = \sum_{j=1}^{p} x_{ij}\beta_{p+j}$$

$$\Rightarrow \omega_i = \frac{exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}{1 + exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}$$

Remaining within the context of a zero-inflated distribution and further assuming the underlying distribution is Poisson, the log likelihood is then

$$log\big(L(\boldsymbol{\mu}, \boldsymbol{\omega})\big) = \sum_{y_i=0} log[\omega_i + (1-\omega_i)e^{-\mu_i}] + \sum_{y_i\neq 0} log\left[(1-\omega_i)\frac{{\mu_i}^{y_i}e^{-\mu_i}}{y_i!}\right]$$

$$= \sum_{y_i=0} log[\omega_i + (1-\omega_i)\,exp(-\mu_i)] + \sum_{y_i\neq 0} log(1-\omega_i) - \sum_{y_i\neq 0}\mu_i + \sum_{y_i\neq 0} y_i log\,(\mu_i) - \sum_{y_i\neq 0} log\,(y_i!)$$

$$= \sum_{y_i=0} log\left[\frac{exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}{1 + exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)} + \left(1 - \frac{exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}{1 + exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}\right) exp\left(-exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\right)\right]$$

$$+ \sum_{y_i\neq 0} log\left(1 - \frac{exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}{1 + exp\left(\sum_{j=1}^{p} x_{ij}\beta_{p+j}\right)}\right) + \sum_{y_i\neq 0} -exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) + \sum_{y_i\neq 0} y_i \sum_{j=1}^{p} x_{ij}\beta_j - \sum_{y_i\neq 0} log\,(y_i!)$$

A model assuming the underlying distribution is negative binomial would follow similarly except with negative binomial probabilities. Maximizing this likelihood can be particularly unstable (e.g. numerical methods have a harder time finding the max) if the data can be fit nearly equally well without extra zeros or a small chance of extra zeros with extra parameters$\beta_{p+1}, \beta_{p+2}, \ldots, \beta_{p+p}$.

3.3 Random Effects and Mixed Models

As previously stated, hourly BW counts within the same day are almost surely not independent observations. Two observations recorded on the same day are likely more correlated than two observations from different days of the same year. The same can be said about daily counts; daily counts for days closer to each other in time are potentially more correlated than daily counts from different years. Mixed models with random effects are used to account for such correlations. These correlations affect the parameter estimates for the fixed effects (the effects that are not random) and the p-values for tests of fixed effects such as tests for trends over time.

The next two sections demonstrate how mixed models work in the simplest situation, where all random variables originate from normal distributions. With a conceptual framework in place, the models are then extended to the more complex generalized linear mixed models.

3.3.1 Mixed Models with Normal Distributions

3.3.1.1 A Simplified Example with Means.

This simplified example will start a description of how mixed models work and how they allow for optimal weighting of surveys and account for correlated data. Imagine that we have three days of surveys in a given year and the counts of BWs in particular hours of the day are as follows:

Day	Hour	Y	Mean
1	1	Y_{11}	
1	2	Y_{12}	
1	3	Y_{13}	
1	4	Y_{14}	
1	5	Y_{15}	$\bar{Y}_1$
2	1	Y_{21}	
2	2	Y_{22}	
2	3	Y_{23}	$\bar{Y}_2$
3	1	Y_{31}	$\bar{Y}_3$

To estimate the average count over all days, μ, one must ask "How should we weight the means for the three days?" Consider two extreme cases. If all the observations on the same day are always the same, BWs are streaming by at a perfectly constant rate; one observation on a day is as good as 5 observations. In essence, we would really only have 3 observations and they should be weighted equally. On the other hand, if all days have the same mean (i.e. no random day effects) and the only variability is random variability from hour to hour then all 8

measurements are independent and we should average the eight values or equivalently weight the means with weights 5, 3, and 1.

In a more general case, along with the usual random variability from count to count with varianceσ^2, suppose there is also random variability from day to day with variance σ_D^2. In a usual mixed model the random effect due to the i^{th} day, D_i, and errors, ε_{ij}, are assumed to be normally distributed. The model in this situation would be

$$Y_{ij} = \mu + D_i + \varepsilon_{ij}$$

where $D_i \sim N(0, \sigma_D^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$ and all D_i and ε_{ij} are independent. In this model, values from the same day are not independent unless there is no day to day variation:

$$Corr(Y_{i1}, Y_{i2}) = \frac{Cov(Y_{i1}, Y_{i2})}{\sqrt{Var(Y_{i1})Var(Y_{i2},)}} = \frac{Cov(D_i, D_i)}{\sigma_D^2 + \sigma^2} = \frac{\sigma_D^2}{\sigma_D^2 + \sigma^2}$$

In estimating μ, we would use some linear combination of $\bar{Y}_1, \bar{Y}_2$, ..., each having an expected value of μ where

$$\bar{Y}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$$

A general fact is that to find the linear combination with smallest variance, one should weight the means inversely proportional to their variances (i.e. the higher the variability of $\bar{Y}_i$, the lower the weight, w_i). Since

$$Var(\bar{Y}_i) = \ Var(\mu + D_i + \bar{\varepsilon}_i) = Var(\mu) + Var(D_i) + Var(\bar{\varepsilon}_i) = \ \sigma_D^2 + \frac{\sigma^2}{n_i}$$

$$w_i = \left(\sigma_D^2 + \frac{\sigma^2}{n_i}\right)^{-1} = \left(\frac{\sigma_D^2}{\sigma^2} + \frac{1}{n_i}\right)^{-1}$$

$$\Rightarrow \hat{\mu} = \frac{\sum_i w_i \bar{Y}_i}{\sum_i w_i} = \frac{\sum_i \left(\frac{\sigma_D^2}{\sigma^2} + \frac{1}{n_i}\right)^{-1} \bar{Y}_i}{\sum_i \left(\frac{\sigma_D^2}{\sigma^2} + \frac{1}{n_i}\right)^{-1}}$$

One can see that the weight given to $\bar{Y}_i$ depends on the relative magnitudes of σ_D^2 and σ^2 (i.e. $\frac{\sigma_D^2}{\sigma^2}$). For example, suppose we have the following data

Day	Hour	Y	Mean
1	1	140	
1	2	145	
1	3	150	
1	4	155	
1	5	160	150
2	1	150	
2	2	155	
2	3	160	155
3	1	165	165

Using the SAS MIXED procedure, the variance estimates are

$$\hat{\sigma}_D^2 = 16.66 \qquad \hat{\sigma}^2 = 52.95$$

The estimated correlation between values on the same day is

$$\frac{\hat{\sigma}_D^2}{\hat{\sigma}_D^2 + \hat{\sigma}^2} = \frac{16.66}{16.66 + 52.95} = 0.24$$

The mean estimate is

$$\hat{\mu} = \frac{\sum \left(\frac{\sigma_D^2}{\sigma^2} + \frac{1}{n_i}\right)^{-1} \bar{Y}_i}{\sum \left(\frac{\sigma_D^2}{\sigma^2} + \frac{1}{n_i}\right)^{-1}} = \frac{1.94 * 150 + 1.54 * 155 + 0.76 * 165}{1.94 + 1.54 + 0.76}$$

$$\hat{\mu} = 0.46 * 150 + 0.36 * 155 + 0.18 * 165 = 154.50$$

The equation for the mean yearly count estimate displays the weighting decision. The days with more values are weighted more but not proportional to their sample sizes which would be the case if the nine observations were treated as independent.

3.3.1.2 A Simplified Regression Example

As indicated above, random effects models allow appropriate weighting of correlated data.

Additionally, by taking into account these correlations, random effect models also provide appropriate p-values, for example when wanting to test for trends. Consider the following example

Year	Day	Count	Mean
2000	1	100	100
2001	1	220	
2001	2	230	
2001	3	240	230
2002	1	230	
2002	2	240	
2002	3	250	
2002	4	260	
2002	5	270	250

If we were interested in estimating a yearly trend and if we assumed the counts were independent the following model would be appropriate

$$Y_{ij} = \beta_0 + \beta_1 x_{year} + \varepsilon_{ij}$$

This time *i* designates the year and *j* the day. Running the data through SAS using the REG procedure resulted in a p-value of 0.005 with 7 degrees of freedom, implying the existence of a significant yearly trend.

If we wanted to allow for the possibility of correlation between the counts occurring in the same year, the above model would be inappropriate. The more appropriate model would be

$$Y_{ij} = \beta_0 + \beta_1 x_{year} + D_i + \epsilon_{ij}$$

where $D_i \sim N(0, \sigma_D^2)$ represents the random year effect. Running the data through SAS this time requires the use of the MIXED procedure. The resulting variance estimates being

$$\hat{\sigma}_D^2 = 1932.22 \qquad \hat{\sigma}^2 = 200.00$$

The estimated correlation between values in the same year being

$$\frac{\hat{\sigma}_D^2}{\hat{\sigma}_D^2 + \hat{\sigma}^2} = \frac{1932.22}{1932.22 + 200.00} = 0.91$$

A correlation of 0.91 implies the counts within the same year are far from independent. That being the case the results from this model when testing for a significant yearly trend are more accurate. With a p-value of 0.25 and approximately 1.03 degrees of freedom, we would conclude more appropriately that there is no significant yearly trend.

In summary, mixed models are needed for appropriate weighting of correlated values and p-values. The description above was for the simpler case (assuming the data are normally distributed). However, as previously mentioned, BW counts are generally not normal. As a result, the normal mixed models are not appropriate. The next section considers the more appropriate model, generalized linear models with random effects, a specific subgroup of the all-encompassing generalized linear mixed models.

3.4 Generalized Linear Mixed Model

Generalized linear mixed models allow for non-normal correlated data. Like the generalized linear model, the assumed distribution must be a member of the exponential family of distributions. The difference is that the distribution is that of the response, Y_i, conditional on some set of normally distributed random effects. Again, similar to the generalized linear model, the basic idea is that some monotonic differentiable function, called the link function, of the expected value of Y_i conditional on a set of random effects can be expressed as a linear function of a set of parameters and explanatory variables. In matrix notation (Littell, 2006)

$$g[E(\boldsymbol{Y}|\boldsymbol{D})] = \boldsymbol{X\beta} + \boldsymbol{ZD}$$
$$\Rightarrow E(\boldsymbol{Y}|\boldsymbol{D}) = g^{-1}(\boldsymbol{X\beta} + \boldsymbol{ZD})$$

where $\boldsymbol{X}$ represents the matrix of explanatory variables accompanying the vector of fixed effects parameters,$\boldsymbol{\beta}$, and $\boldsymbol{Z}$ the matrix of explanatory variables accompanying the vector of random effects parameters,$\boldsymbol{D}$.

Consider the following "simple" example where Y_{ij} represents the j^{th} count (for $j = 1,2, \dots n_i$) on the i^{th} day (for $i = 1,2,3$), $x_{i1}, x_{i2}, \dots, x_{in_i}$ independent explanatory variables, and D_i the i^{th} random day effect. Suppose the counts conditional on the random day effect are independent and follow a Poisson distribution

$$Y_{ij}|D_i \sim Poisson(\mu_{ij})$$

$$\Rightarrow f(y_{ij}|D_i) = e^{-\mu_{ij}} \frac{{\mu_{ij}}^{y_{ij}}}{y_{ij}!}$$

The following generalized linear mixed model would be appropriate

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} + D_i \; where \;\; D_i \sim N(0, \sigma_D^2)$$

$$\Rightarrow \mu_{ij} = exp(\beta_0 + \beta_1 x_{ij} + D_i)$$

One of the complexities of the generalized linear mixed model is encountered as one begins the process of finding parameter estimates; as one attempts to derive an expression for the marginal likelihood function

$$f(y_{11}, y_{12}, \dots y_{3n_3}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(y_{11}, y_{12}, \dots y_{3n_3}, D_1, D_2, D_3)\, dD_1\, dD_2\, dD_3$$

In this simple example, assuming the conditional random variables are independent makes things slightly easier

$$f(y_{11}, y_{12}, \dots y_{3n_3}, D_1, D_2, D_3) = \prod_{i=1}^{3} \prod_{j=1}^{n_i} f(y_{ij}|D_i)\, f(D_1) f(D_2) f(D_3)$$

$$= \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma_D^2}} e^{-0.5 * D_i^2/\sigma_D^2} \prod_{j=1}^{n_i} e^{-\mu_{ij}} \frac{{\mu_{ij}}^{y_{ij}}}{y_{ij}!}$$

$$= \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma_D^2}} e^{-0.5 * D_i^2/\sigma_D^2} \prod_{j=1}^{n_i} e^{-e^{\beta_0 + \beta_1 x_{ij} + D_i}} \frac{(e^{\beta_0 + \beta_1 x_{ij} + D_i})^{y_{ij}}}{y_{ij}!}$$

In pursuit of the parameter estimates, the marginal log likelihood (i.e. the log of the triple integral above) needs to be maximized with respect to β_0, β_1, and σ_D^2. These particular integrals

can be simplified some, but the functions can quickly become more complex, for example assuming a zero inflated distribution rather than Poisson. Software such as SAS is not written to identify particular simplifications but to handle all cases in a standard way. To maximize the likelihoods in SAS using the GLIMMIX procedure, the integrals are approximated with either Laplace or Gaussian quadrature numerical methods. Unfortunately, not uncommon for more complex models, these methods fail to converge.

In addition, there is another complication due to the variability in durations accompanying the survey. For example some surveys reflect the number of BWs counted in one hour where as other surveys reflect the number of BWs counted in eight hours. Suppose in an oversimplified example that μ_{ij}, the mean count for the j^{th} survey of the i^{th} day, depends on the hour of the day, x_{ij}, in a linear fashion with no random effects

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} \text{ for } j = 1,2, \dots 8$$
$$\Rightarrow \mu_{ij} = e^{\beta_0 + \beta_1 x_{ij}}$$

Suppose further that for $i = 1$ we have 8 surveys, $Y_{11}, Y_{22}, \dots, Y_{28}$, recorded for all 8 hours $x_{1j} = 8, 9, \dots, 15$ and for $i = 2$ we have only one survey, Y_{2+}, which represents the number of BWs counted for all 8 hours $x_{1j} = 8, 9, \dots, 15$. As a result, the expected value of Y_{2+} is

$$\mu_{2+} = \sum_{j=1}^{8} \mu_{ij} = \sum_{j=1}^{8} e^{\beta_0 + \beta_1 x_{2j}}.$$
$$\Rightarrow \log(\mu_{2+}) = log\left(\sum_{j=1}^{8} e^{\beta_0 + \beta_1 x_{2j}}\right)$$

which is a **nonlinear** function of β_0 and β_1, thus violating a critical assumption of the generalized **linear** mixed model. SAS does provide a procedure that handles nonlinear models, NLMIXED, but it is much more restrictive is what it allows. For example we cannot (without great pains) have random effects. If we had consistently reported daily totals, we wouldn't need to include the hourly effects in order to adjust for which hours were recorded on which days. In which case we wouldn't have to use a nonlinear model; the more usual linear models could be used.

4 The Analysis

As indicated above, ideally one would use a nonlinear mixed model for the hourly data. With different days having different numbers of hourly observations, this type of model would allow us to objectively and appropriately weight these daily counts. However, initial attempts to fit these types of models often resulted in convergence issues. Because of this, a simplified strategy was taken with the thought that the resulting trends could be used for comparisons to more complex models fit by others. This simplified strategy also excludes the consideration of weather effects. Again, the thought is that the results can be used for comparison to models that include weather effects fit by others in order to assess how much the precision of trend estimates is improved by including weather effects in the models.

It should also be noted that if the yearly counts were consistent in terms of which days were surveyed and the hours surveyed each day, we would naturally be interested in testing for time trends in these yearly counts. If there were no seasonal or daily effects, we could accommodate differing number of hours of observation in each year with an "offset". For example if we have a Poisson model where the expected number of hawks per hour is

$$\mu_k = exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$$

the number of hawks in duration $=h$ hours is Poisson with a mean

$$h * exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$$

The linear model for $\log(\mu_k)$ would be

$$\log(\mu_k) = \log(h) + \sum_{j=1}^{p} x_{ij}\beta_j$$

The "offset", $\log(h)$, is a term in the model having its coefficient defined to be 1. However, the arrival of BWs is greater during certain times of the day and certain days of the season. Observations for two hours starting at 6 am are not the same degree of daily coverage as two hours starting at 11 am. An effective duration of daily sampling would be the fraction of hawks we expect to see in the hours actually observed.

As a result of the above observations and choice of a simplified strategy, the analysis was conducted in a three step process: 1) Determine an effective duration for each day (i.e. adjusted daily duration), 2) Use daily counts and associated adjusted durations to determine effective durations for each year (i.e. adjusted yearly durations), 3) Use yearly counts and associated adjusted durations to test the significance of yearly trends. Note that using daily counts in step two allows us to use the data from 2005 and 2006 where the only counts available are daily totals.

4.1 Daily Counts with Adjusted Durations

4.1.1 Formula for the Adjusted Durations

Let the BW daily count, y_{ij+}, represent the sum of the hourly counts for each day within each year,

$$y_{ij+.} = \sum_{k=6}^{17} y_{ijk}$$

where y_{ijk} is the k^{th} hourly count ($k = 6, 7, \ldots, 17$ corresponding to the BW's hourly window) from the j^{th} day since August 1st ($j = 22, 23, \ldots, 61$ corresponding to the BW's seasonal window) in the i^{th}year ($i = 1, 2, \ldots, 40$ where for example $i = 1$ represents 1972). Let $\hat{f}_k$ represent the estimated fraction of BWs expected in the k^{th} hour of the day. The adjusted daily duration, in hours, accompanying each daily count,

$$d_{ij} = 12 * \sum_{k=6}^{17} \hat{f}_k * \frac{d_{ijk}}{60}$$

where d_{ijk} represents the duration, in minutes, accompanying the count, y_{ijk}. If an entire 12 hour day is sampled, then $d_{ij} = 12$. For partial days, the adjusted duration, d_{ij}, is 12 times the fraction of the day's BWs expected to be counted during the actual observation period.

To determine values for $\hat{f}_k$, only counts accompanied by a duration, d_{ijk}, of sixty minutes were used. These hourly counts represented 90% of the recorded observations. With a very large number of hourly counts, the estimation of the hourly trend is quite precise. Thus, using truly hourly data, estimates of the mean number of counts expected in each hour, $\hat{\mu}_k$, were used to determine

$$\hat{f}_k = \frac{\hat{\mu}_k}{\sum_{k=6}^{17} \hat{\mu}_k}$$

It was in the estimation of μ_k that the first generalized linear model was used.

4.1.2 Model Parameters and Form of the Linear Predictor

Included in the model were hourly and seasonal trends as well as fixed year effects. The seasonal trends and year effects were included to avoid bias in hourly trend estimates. As far as the form of the linear predictor, quadratic splines were fit. The following explanation introduces the basic form of a quadratic spline and the origin of notation used throughout the rest of the paper.

A quadratic spline is a piecewise polynomial function with internal knots being the place(s) at which the pieces of the function join (Montgomery 2006). By definition, a quadratic spline is a continuous function with a continuous first derivative. For example, a quadratic spline with two internal knots, $x = k_1$ and $x = k_2$, can be written

$$y = spline(x) = b_0 + b_1 * x + b_2 * x^2 + b_3 * [(x - k_1)^+]^2 + b_4 * [(x - k_2)^+]^2$$

where

$$(x - k_i)^+ = \begin{cases} 0 & for\ x < k_i \\ x - k_i & for\ x \geq k_i \end{cases}$$

It can be readily verified that this function and first derivatives are continuous at k_i. Thus for the different intervals of x

$$spline(x, k_1, k_2) = \begin{cases} b_0 + b_1 * x + b_2 * x^2 & x < k_1 \\ b_0 + b_1 * x + b_2 * x^2 + b_3 * (x - k_1)^2 & k_1 \leq x < k_2 \\ b_0 + b_1 * x + b_1 * x^2 + b_3 * (x - k_1)^2 + b_4 * (x - k_2)^2 & x \geq k_2 \end{cases}$$

Two quadratic splines, each having two internal knots, were used for the hourly and seasonal trends. For numerical stability, the hours and days were "centered": $x_{hour} = (Hour - 11)$ and $x_{day} = (Day - 40)$ where day is days since August 1st. The knots were place at $k_{Hr_1} = -2$ and $k_{Hr_2} = 4$ for the hourly spline and $k_{Day_1} = 0,\ k_{Day_2} = 10$ for the daily spline.

4.1.3 First Generalized Linear Model

Regardless of the assumed distribution, the generalized linear model used to estimate $\mu_k$was of the following form

$$\log(\mu_{ijk}) = \beta_0 + \sum_{i=1}^{40} \beta_i Yr_i + spline(x_{hour}, k_{Hr_1}, k_{Hr_2}) + spline(x_{day}, k_{Day_1}, k_{Day_2})$$

$$\Rightarrow \mu_{ijk} = \exp\left\{\beta_0 + \sum_{i=1}^{40} \beta_i Yr_i + spline(x_{hour}, k_{Hr_1}, k_{Hr_2}) + spline(x_{day}, k_{Day_1}, k_{Day_2})\right\}$$

The predictor Yr_i is a dummy variable, equaling one when the year = *i*, otherwise zero. Because the aim is to obtain the estimated number of BWs counted in the k^{th}hour of the j^{th} day in the i^{th} year, as previously mentioned, only hourly counts accompanied by a duration = 60 were used. As a result, counts obtained in the years 1972, 1973, 2005, and 2006 (i.e. $i = 1, 2, 34, \text{and } 35$) were excluded due to all or a majority of the counts being daily counts.

4.1.4 Determining the Distribution

In deciding whether to use Poisson, zero-inflated Poisson, negative binomial, or zero-inflated negative binomial distributions, we can't rely on goodness of fit measures such as the Akaike Information Criterion [AIC]. To base goodness of fit on independent pieces of information, I

used the method of cross-validation. Cross-validation is a model validation technique employed when the predictive accuracy of the model is of primary interest. It entails splitting the data into two parts: estimation, "training", data used to build the model and validation data used to the study the predictive ability of the model (it should be noted that the validation data are independent of the training data). Specifically, for each distribution, the above model was run 36 times (corresponding to the number of years represented in the data set). For each run, the validation data set represented the hourly counts from a specified year (not restricted to duration = 60), the estimation data set the hourly counts for all other years where duration = 60. The predicted hourly counts from the run were then averaged over the years. The resultant averages were used as the predicted hourly counts for the excluded year of which were eventually compared with the hourly counts in the validation data set after adjusting for duration differences.

For example, in one specific model run assuming some distribution, counts from the year 1974 represented the validation data set (i.e. y_{3jk}) and the counts from the remaining years (i.e. y_{ijk} where $i = 4, 5, \dots, 33, 36, \dots, 40$ and duration = 60) represented the estimation data set. The predicted hourly counts for 1974 were derived from model run that did not include predictions for 1974. Let I_{-3} represent the set of years of hourly data not including year *i*=3 (1974), JK_3 represent the set of days and hours (j, k) observed in 1974, and $\hat{y}_{ijk}$ represent the predicted value for the j^{th} day since August 1st and k^{th} hour in the i^{th} year. The predicted hourly counts for the year 1974, $\hat{y}_{3jk}$, were derived as follows:

$$\hat{y}_{3jk} = \frac{\sum_{i \epsilon I_{-3}} \hat{y}_{ijk}}{35} (j, k) \epsilon JK_3$$

Thus, under an assumed distribution, after a total of 36 runs, a set of predicted hourly counts, $\hat{y}_{ijk}$, were obtained for $k = 6, 7, \dots, 17$, $j = 22, 23, \dots, 61$, and $i = 1, 2, 3, \dots, 40$. Note that even though the years 1972, 1973, 2005 and 2006 are missing hourly data, we still find predicted hourly values.

Because the predicted hourly counts represent predictions associated with a duration = 60, before comparisons were made, the predictions were adjusted to reflect what would be expected based on the observed duration

$$adj\_\hat{y}_{ijk} = \hat{y}_{ijk} * \left(\frac{d_{ijk}}{60}\right)$$

To enable the use of all years in the goodness of fit test (i.e. to include the years 1972, 1973, 2005, and 2006) comparisons were made between the observed daily totals, y_{ij+} and expected daily totals

$$\hat{y}_{ij+} = \sum_{k \in K_{ij}} adj\_\hat{y}_{ijk}$$

where K_{ij} is the set of hours during which actual observations took place on the j^{th} day of the i^{th} year. The resultant distribution was chosen based on having the minimum chi-square value

$$\sum_i \sum_j \frac{\left(y_{ij+} - \hat{y}_{ij+}\right)^2}{\hat{y}_{ij+}}$$

4.1.5 SAS Results

The cross-validation results are displayed in the table below:

Table 5: Cross-Validation Results (SAS)

Model	df	Chi-square	Chisq./df
Negative Binomial	1267	7032694	5551
Poisson	1267	8060311	6362
Zero-Inflated Negative Binomial	1267	7478943	5903
Zero-Inflated Poisson	1267	7971802	6292

The negative binomial distribution produced the smallest chi-square value and thus was the assumed distribution in the model. Fitting the model to the data produced the following fractions, $\hat{f}_k$, of BWs observed over the hours of the day.

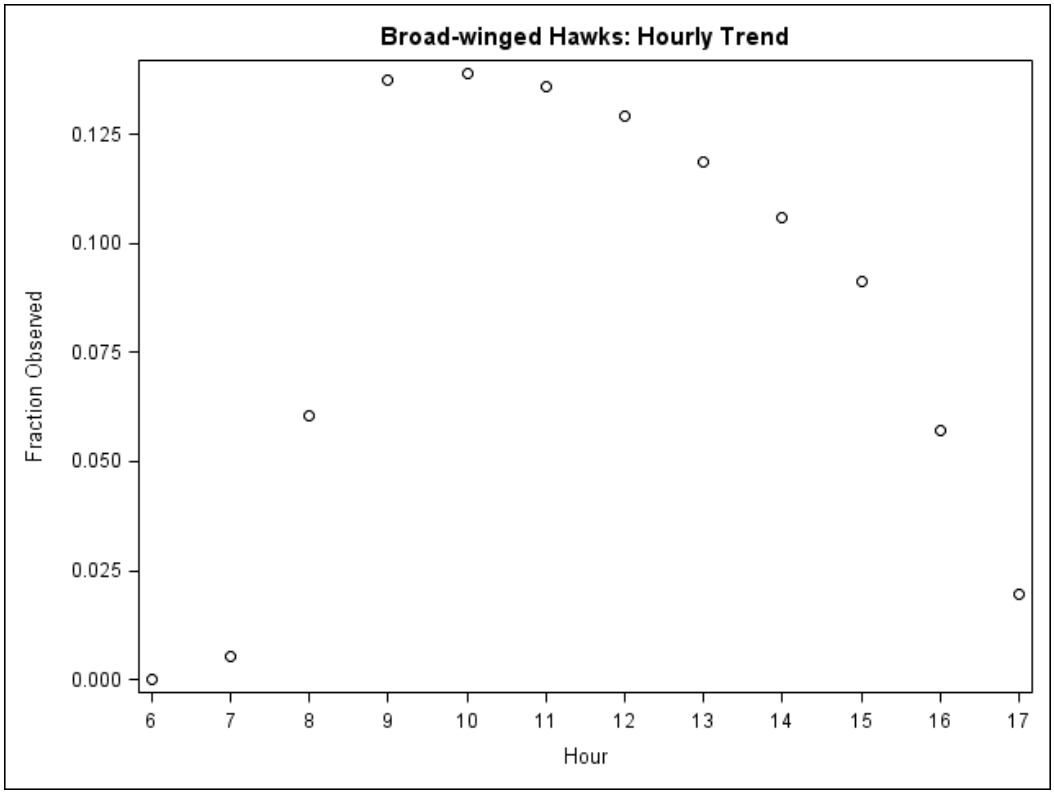

Figure 8: Fraction of the Day's BWs seen each hour.

Again to restate, the adjusted daily duration associated with a day's total is then calculated as

$$d_{ij} = 12 * \sum_{k=6}^{17} \hat{f}_k * \frac{d_{ijk}}{60}$$

4.2 Yearly Counts with Adjusted Durations

4.2.1 Formula for the Adjusted Durations

Steps in obtaining the adjusted yearly durations were similar to the steps used in obtaining the adjusted daily durations. Letting, y_{ij+}, represent the total count on a given day in a particular year, the yearly total count is the sum of the daily counts within each year,

$$y_{i++} = \sum_{j=22}^{61} y_{ij+}$$

The adjusted duration, d_i, in days, accompanying each yearly count was determined using the following equation:

$$d_i = 40 * \sum_{j=22}^{61} \hat{f}_j * \frac{d_{ij}}{12}$$

where $\hat{f}_j$ represents the estimated fraction of BWs expected on the j^{th}day since August 1st

$$\hat{f}_j = \frac{\hat{\mu}_j}{\sum_{j=22}^{61} \hat{\mu}_j}$$

The mean number of counts on the j^{th} day since August 1st, μ_j, was estimated with the use of a second generalized linear model, very similar to the first.

4.2.2 Second Generalized Linear Model

Again, regardless of the assumed distribution in the model, the mean function was of the following form

$$\log(\mu_{ij}) = log(d_{ij}) + (\beta_0 + \sum_{i=1}^{40} \beta_i Yr_i + spline(x_{day}, k_{Day_1}, k_{Day_2})$$

$$\Rightarrow \mu_{ij} = \exp\left\{\beta_0 + \sum_{i=1}^{40} \beta_i Yr_i + spline(x_{day}, k_{Day_1}, k_{Day_2})\right\}$$

with the knots and Yr_i being define as before, the term $log(d_{ij})$ necessary due to the varying number of hours accompanying each daily count.

4.2.3 Determining the Distribution

The assumed distribution was determined with the use of cross-validation. Different from the runs using the hourly data, because our data represent daily counts, counts from the years 1972, 1973, 2005, and 2006 were included in the validation runs as well as the final model. Again, the predicted daily counts from the validation runs were adjusted based on the observed daily duration

$$adj\_\hat{y}_{ij} = \hat{y}_{ij} * \left(\frac{d_{ij}}{12}\right)$$

The resultant distribution was chosen based on having the minimum chi-square value

$$\sum_i \sum_j \frac{\left(y_{ij+} - adj\_\hat{y}_{ij+}\right)^2}{adj\_\hat{y}_{ij+}}$$

4.2.4 SAS Results

The cross-validation results are displayed in the table below:

Table 6: Cross-Validation Results (SAS)

Model	df	Chi-square	Chisq./df
negative binomial	1600	8069480	5043
Poisson	1600	9086730	5679
Zero-inflated negative binomial	1600	7954510	4972
Zero-inflated Poisson	1600	9251337	5782

This time the zero-inflated negative binomial distribution produced the smallest chi-square value and thus was the assumed distribution. Fitting the model to the data produced the following estimates of the fractions, $\hat{f}_j$, of BWs seen on average on each day of the migration season

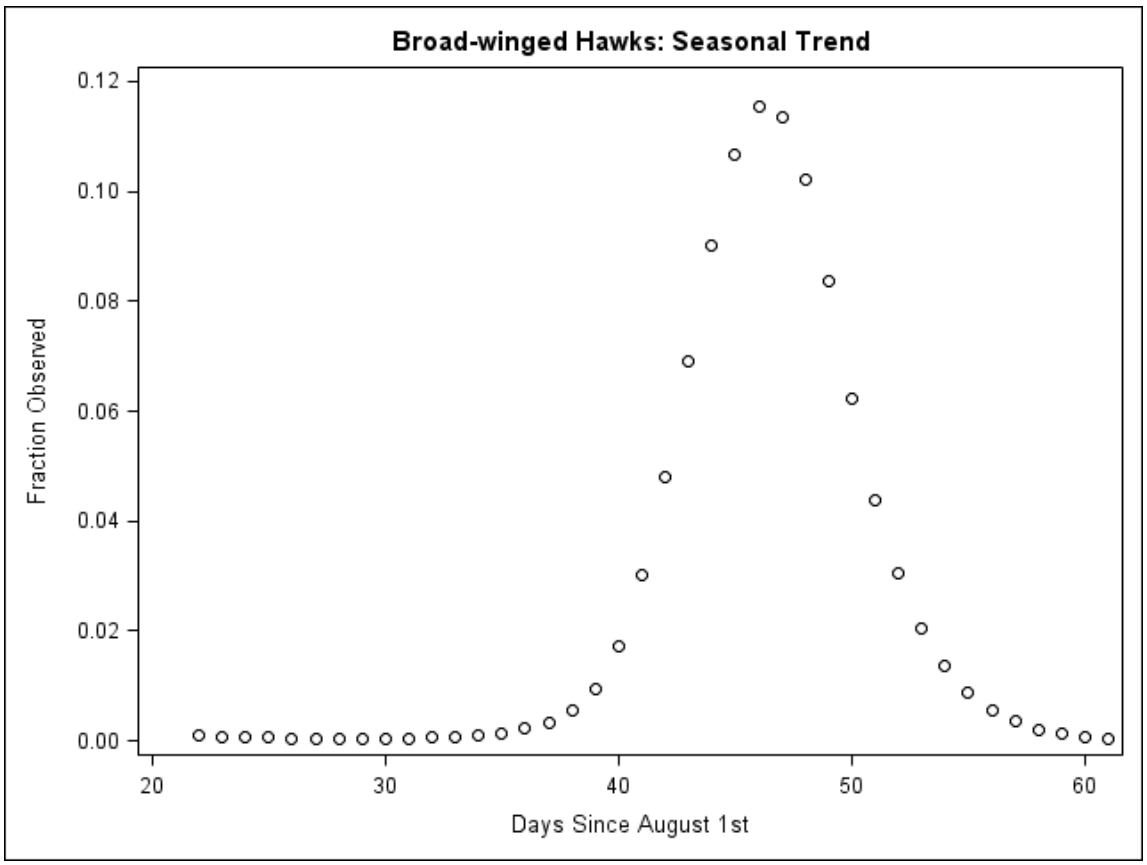

Figure 9: Fraction of the Year's BWs seen each day.

Again to restate, the adjusted yearly durations for each year were then calculated as

$$d_i = 40 * \sum_{j=22}^{61} \hat{f}_j * \frac{d_{ij}}{12}$$

4.3 Testing Annual Trends

Using the BW yearly counts, y_{i++} and the adjusted yearly durations, the plot below shows the annual trend in BWs per day adjusted for the year's adjusted duration of observation

$$BW\ per\ day = \frac{y_{i++}}{d_i}$$

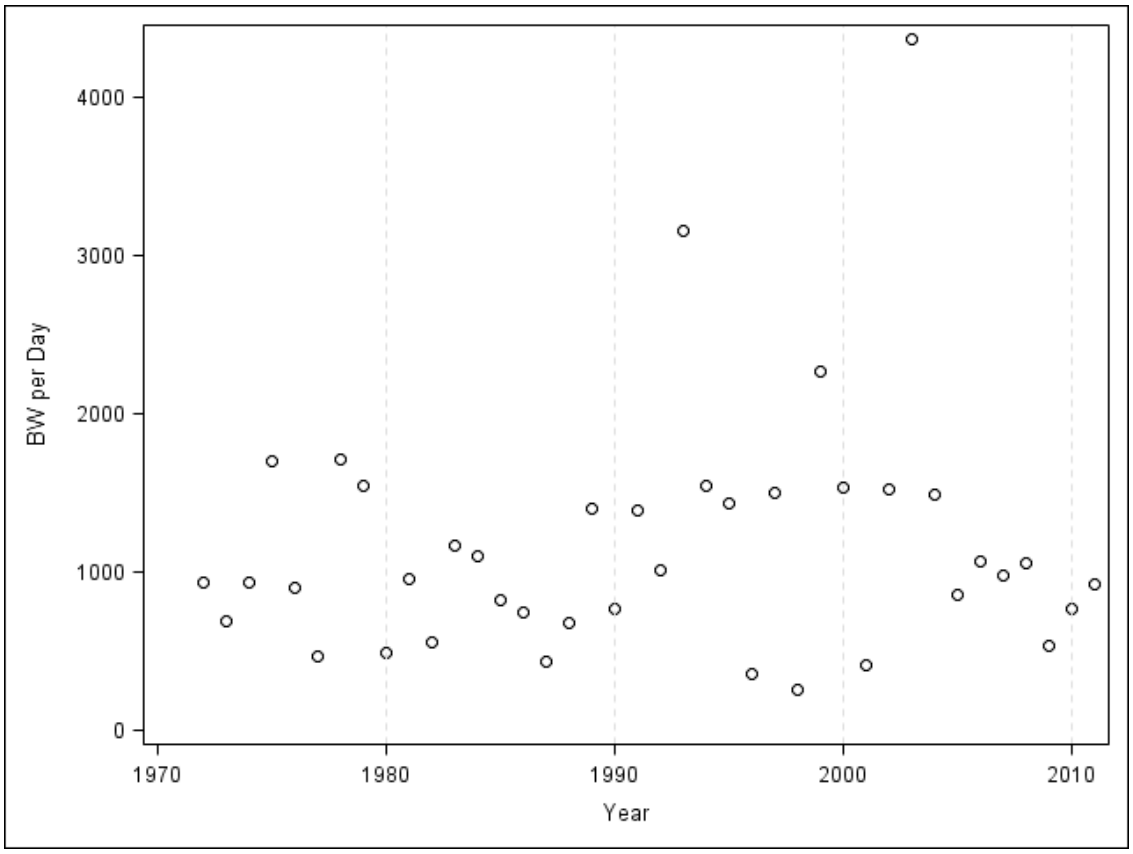

Figure 10: BW per Day against Year

With the exception of two or three big years, the plot suggests no major changes in the BW population over the years. When testing for an annual trend, the Poisson and negative binomial distributions as well as the potential for random yearly effects were considered. As a result, a total of four models were investigated: two generalized linear models (one for each distribution) and two generalized linear mixed models. The linear predictors are below

$$\textit{Generalized Linear Model}: \quad log(\mu_i) = log(d_i) + \beta_0 + \beta_1\, x_{year}$$

$$\textit{Generalized Linear Mixed Model}: \quad log(\mu_i) = log(d_i) + \beta_0 + \beta_1\, x_{year} + D_i$$

where $D_i \sim N(0, \sigma_D^2)$ represents the random year effect, $\mu_i = E(Y_{i++}|D_i)$ in the mixed model and $\mu_i = E(Y_{i++})$ in the generalized linear model. The marginal likelihood functions in the mixed models were approximated using Laplace's method. The table below displays the SAS results:

Table 7: Testing Annual BW Population Trend			
Model	**-2 Log Likelihood**	**Yearly Trend p-value**	**Number of Parameters in Distribution**
Poisson	538886	<0.0001	1 mean
Neg Bin	901.21	0.2843	2 mean, dispersion
Mixed Poisson	899.02	0.5875	2 mean, year variance
Mixed Neg Bin	899.07	0.5356	3 mean, dispersion, year variance

The pairs of models can be compared using the following test statistic

$$Test\ Statistic = \chi^2(1)$$
$$= -2\log(likelihood)_{Poisson} + 2\log(likelihood)_{NegBin}$$

If $\chi^2(1) > \chi^2_{0.95}(1) = 3.84$ then the negative binomial distribution should be the assumed distribution. It's clear from the table that the Poisson is a poor fit when compared to the negative binomial; $\chi^2(1) = 537984.79 > 3.84 = \chi^2_{0.95}(1)$. This conclusion supports what's previously been discussed: the model assumes all BWs arrive independently, which is far from the case, and the model does not pay attention to how consistently the data follow the trend line; there is no separate estimate of a variance. The two mixed models are essentially the same, with a $\chi^2(1) = 0.05$, the Mixed Poisson would be preferable. We are unable to compare the negative binomial and the mixed Poisson with a statistical test. But it's helpful to remember that the negative binomial model assumes the Poisson mean follows a gamma distribution (negative binomial is a Gamma Poisson Mix) and the mixed Poisson model assumes that the Ln(Poisson mean) follows a normal distribution. Thus one could say the two models are very similar, differing only in the assumption regarding in the random variability from year to year. Since the likelihood of the mixed Poisson model is a bit better than the negative binomial model, with a difference in 2*(Log likelihood) or AIC of 2.05, one could conclude that a lognormal model for mean Poisson counts fits better than a gamma model for the mean counts.

In any case, with the exception of the Poisson, the p-values support our initial observation: the absence of a significant yearly trend.

5 Conclusion

Testing for annual population trends in fall migrating raptors is of interest to the Hawk Ridge Bird Observatory. Data available for such analysis are in the form of counts. The potential for over-dispersion, the existence of a high number of zero counts, the lack of independence between the counts and the fact that some counts reflect hourly or nearly hourly totals where as others reflect daily totals all influenced the type of analyses considered and/or conducted.

Ideally a nonlinear model with random effects and the consideration of the Poisson, negative binomial, and zero-inflated distributions would be most appropriate. Although SAS has the potential to fit such models, getting the model to include random effects would require a concerted effort and convergence issues would be likely. As a result, an alternative simplified approach was decided upon with the thought that the results could later be compared to the results of more complex models investigated in a later project. This simplified approach did not include the consideration of weather effects. Again, the decision was based on minimizing the frequency of convergence issues and the thought that models fit with weather variables later on could be compared to such.

This simplified approach consisted of three major steps. First, to address the inconsistency of reporting (i.e. hourly totals and daily totals) a generalized linear model using only counts accompanied by a duration = 60 minutes was used to derive adjusted daily durations. Daily counts were calculated as a simple sum of the day's hourly counts. Second, a generalized linear model was fit to the daily totals, taking into account the adjusted daily durations, in order to derive an adjusted yearly duration. The second step was taken to avoid potential convergence issues as well as allowing us to fit zero-inflated models; the SAS GLIMMIX procedure for mixed models does not include zero-inflated models. The yearly count, the sum of the year's daily counts, accompanied by its adjusted yearly duration served as the annual population index. A scatter plot of the yearly count per day was plotted over the years and suggested the absence

of a significant yearly trend. Two generalized linear models and two generalized linear models with a random year effects were fit to the yearly data (Poisson and negative binomial distributions considered). The generalized linear model assuming a Poisson distribution performed poorly; trend results were not reliable. The other three models performed similarly, all three suggesting no significant yearly trend.

The next step in this process would be to run more complex models and compare their results with the results of this project. Including weather variables in the models would be a logical first type of model explored; common sense suggests it would be a more appropriate model due to the known fact that days with certain wind directions are known to produce higher counts. Including weather variables would start with the daily counts with adjusted durations as derived here along with daily weather summaries.

6 References

Delwiche, Lora D., and Susan J. Slaughter. *The Little SAS Book: A Prime.* 4th ed. Cary, NC: SAS Institute Inc., 2008. Print.

Dobson, Annette J. *An Introduction To Generalized Linear Models.* New York: Chapman and Hall, 1990. Print.

Farmer, Christopher J., David J. T. Hussell, and David Mizrahi. "Detecting Population Trends in Migratory Birds of Prey." *The Auk* 124.3 (2007): 1047-1062.

Hawk Ridge Bird Observatory Website: http://www.hawkridge.org/visit/migration.html.

Jung, Clarence S. "Weather Conditions Affecting Hawk Migrations." *Lore* 14 (1964): 134-144.

Littell, Ramon C., et al. *SAS for Mixed Models.* 2nd ed. Cary, NC: SAS Institute Inc., 2006. Print.

Montgomery, Douglas C., et al. *Introduction to Linear Regression Analysis.* 4th ed. Hoboken: John Wiley and Sons, Inc., 2006. Print.

Ross, Sheldon M. *Introduction to Probability Models.* 10th ed. Burlington: Elsevier Inc., 2010. Print.

7 Appendices

7.1 Data Cleaning

The cleaning process focused around two major inconsistencies in the data. The first due to the fact that each reported count is accompanied by two durations: the reported *duration* and the elapsed time between the *start* and *end* times. One would assume these durations would agree, unfortunately this was not always the case, raising the question: "Which duration should be used?" The second inconsistency was due to the impossibility of certain reported counts; some counts were accompanied by a *duration* and/or an elapsed time equaling zero. Thus raising the question: "Should these counts be included in the data set?" To facilitate the cleaning process, the following variables were created for each survey: *View duration* = elapsed time between *start* time and *end* time and *total birds* = sum of the counts of all species.

7.1.1 "Which duration should be used when *duration* ≠ *view duration*?"

It was decided that the reported *duration* was more reliable and thus was honored when *duration* ≠ *view duration* except when *duration* = 0 or 1. Under this exception, *view duration* was honored. Table1 highlights the number of surveys for which *duration* was honored, 2154, and the number of surveys for which *view duration* was honored, 18, when the two disagreed.

Table 1: Where Duration ≠ View Duration

Duration	View Duration								
Frequency	**15**	**30**	**45**	**60**	**75**	**540**	**570**	**1439**	**Total**
0	0	0	0	15	1	0	0	0	16
1	0	0	0	2	0	0	0	0	2
5	3	0	0	2	0	0	0	0	5
8	2	0	0	2	0	0	0	0	4
10	5	0	0	7	0	0	0	0	12
11	0	0	0	1	0	0	0	0	1
12	1	0	0	2	0	0	0	0	3
15	0	0	0	411	0	0	0	0	411
16	1	0	0	0	0	0	0	0	1
18	0	0	0	1	0	0	0	0	1
20	7	1	0	10	0	0	0	0	18

Table 1: Where Duration ≠ View Duration									
Duration	**View Duration**								
Frequency	**15**	**30**	**45**	**60**	**75**	**540**	**570**	**1439**	**Total**
22	1	0	0	2	0	0	0	0	3
23	0	1	0	0	0	0	0	0	1
24	0	1	0	1	0	0	0	0	2
25	0	2	0	4	0	0	0	0	6
27	0	1	0	0	0	0	0	0	1
28	0	3	0	2	0	0	0	0	5
30	0	0	1	1196	0	0	0	0	1197
32	0	1	0	0	0	0	0	0	1
33	0	2	0	2	0	0	0	0	4
34	0	2	0	1	0	0	0	0	3
35	0	2	0	5	0	0	0	0	7
37	0	2	0	3	0	0	0	0	5
38	0	0	0	1	0	0	0	0	1
39	0	0	1	0	0	0	0	0	1
40	0	0	9	12	0	0	0	0	21
41	0	0	1	0	0	0	0	0	1
42	0	0	0	2	0	0	0	0	2
44	0	0	1	0	0	0	0	0	1
45	0	0	0	387	0	0	0	0	387
48	0	0	1	0	0	0	0	0	1
49	0	0	1	0	0	0	0	0	1
50	0	0	2	7	0	0	0	0	9
51	0	0	0	2	0	0	0	0	2
52	0	0	0	1	0	0	0	0	1
54	0	0	0	1	0	0	0	0	1
55	0	0	0	6	0	0	0	0	6
56	0	0	0	1	0	0	0	0	1
59	0	0	0	1	0	0	0	0	1
65	0	0	0	1	0	0	0	0	1
90	0	0	0	3	0	0	0	0	3
105	0	0	0	1	0	0	0	0	1
180	0	0	0	0	0	0	0	1	1
480	0	0	0	0	0	1	1	0	2
Total	20	18	17	2095	1	1	1	1	2154

7.1.2 "Should the surveys be included when *duration= view duration = 0*?"

When *duration* = *view duration* = 0 and *total birds* = 0 the survey was automatically dropped from the data set. When *duration* = *view duration* = 0 and *total birds* ≠ 0, durations reported within a few days of the survey of interest were investigated in hopes of finding a pattern to use in making an educated guess at an appropriate duration modification. With the exception of one, the surveys were dropped from the data set due to the absence of a strong pattern. The exception was: when *duration* = *view duration* = 0 and *total birds* = 233 the survey was modified such that *duration* = *view duration*= 420, *start* = 0800, and *end* = 1500. Tables 2 and 3 highlight the changes mentioned above. In summary, a total of 74 + 8 surveys were removed from the data set based on the above conditions.

Table 2: Where Duration = View Duration = 0

Total Birds	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	74	89.16	74	89.16
1	2	2.41	76	91.57
3	1	1.20	77	92.77
4	1	1.20	78	93.98
8	1	1.20	79	95.18
9	1	1.20	80	96.39
19	1	1.20	81	97.59
24	1	1.20	82	98.80
233	1	1.20	83	100.00

Table 3: Where Duration = View Duration = 0 and Total Birds ≠ 0

Total birds	View Duration	Duration	Date	Start	End	Prev Date	Prev Start	Prev End	Next Date	Next Start	Next End
9	0	0	25-11-97	5:00	5:00	24-11-97	5:00	5:00	26-11-97	10:00	11:00
19	0	0	24-11-97	5:00	5:00	23-11-97	5:00	5:00	25-11-97	5:00	5:00

Table 3: Where Duration = View Duration = 0 and Total Birds ≠ 0											
Total birds	View Duration	Duration	Date	Start	End	Prev Date	Prev Start	Prev End	Next Date	Next Start	Next End
24	0	0	23-11-97	5:00	5:00	22-11-97	5:00	5:00	24-11-97	5:00	5:00
8	0	0	22-11-97	5:00	5:00	21-11-97	13:00	14:00	23-11-97	5:00	5:00
3	0	0	02-08-92	5:00	5:00	26-07-92	8:00	11:00	10-08-92	11:00	12:00
1	0	0	06-08-89	5:00	5:00	08-11-88	11:00	12:00	15-08-89	8:00	9:00
1	0	0	18-08-78	5:00	5:00	16-08-78	5:00	5:00	19-08-78	10:00	11:00
4	0	0	16-08-78	5:00	5:00	27-10-77	12:00	13:00	18-08-78	5:00	5:00
233	0	0	25-09-73	5:00	5:00	24-09-73	9:00	15:00	27-09-73	8:00	15:00

7.1.3 Minor Changes

The following changes occurred due to minor inconsistencies in reporting. The survey from September 26th, 1977, with a *start* time of 1600 honors *view duration* rather than *duration*, since the duration of 65 overlaps the next hour. The survey from August 15th, 1997, with a *start* time of 0000 was dropped from the data set. The survey from September 18th, 2011, with a *start* time of 1000 and a *duration* = 0 was changed to a *duration* = 75. The survey from September 18th, 2011, with a *start* time of 1100 was dropped from the data set, since this was included in the previous 75 minutes.

Table 4: Minor Changes				
Date	Start	End	Duration	ViewDuration
26-09-77	9:00	10:00	45	60
26-09-77	10:00	11:00	60	60
26-09-77	11:00	12:00	60	60
26-09-77	12:00	13:00	60	60
26-09-77	13:00	14:00	60	60
26-09-77	14:00	15:00	60	60
26-09-77	15:00	16:00	60	60
26-09-77	16:00	17:00	65	60
26-09-77	17:00	18:00	60	60
15-08-97	9:00	10:00	60	60
15-08-97	10:00	11:00	60	60

Table 4: Minor Changes				
Date	**Start**	**End**	**Duration**	**ViewDuration**
15-08-97	11:00	12:00	60	60
15-08-97	0:00	23:59	180	1439
18-09-11	5:45	6:00	15	15
18-09-11	6:00	7:00	60	60
18-09-11	7:00	8:00	60	60
18-09-11	8:00	9:00	60	60
18-09-11	9:00	10:00	60	60
18-09-11	10:15	11:15	0	75
18-09-11	11:00	12:00	15	60

7.2 SAS Code

7.2.1 Daily Counts with Adjusted Durations

7.2.1.1 Breaking Counts into Hours

```
data counts;
  format start hhmm. end hhmm.;
  set sasdata.clean_counts;
  run;
proc sort data=counts;
  by date start;
  run;
data counts;
  set counts;
  format prev_end hhmm.;
  by date;
  prev_end = lag1(end);
  if first.date then prev_end = .;
  run;
proc print data=counts (obs=20);
  where prev_end>start;
  var date start end prev_end view_duration duration;
  run;
proc sort data=counts;
  by descending date descending start;
  run;
data rev_duration_counts;
  set counts;
  format next_start hhmm. next_date date9.;
```

```
  by descending date;
  next_date = lag1(date);
  next_start = lag1(start);
  if first.date then next_start = .;
  run;
proc sort data=rev_duration_counts;
  by date start;
  run;
data rev_duration_counts;
  set rev_duration_counts;
  by date;
  format prev_end hhmm. prev_date date9.;
  prev_date = lag1(date);
  prev_end = lag1(end);
  if first.date then prev_end = .;
  diff_min = round(view_duration - used_duration, 1);
  drop duration;
  run;
data rev_duration_counts;
  set rev_duration_counts;
  format old_start hhmm. old_end hhmm.;
  by date;
  if (diff_min ne 0) then do;
    if (last.date) or (start + 60*used_duration <  next_start) then do;
            old_end = end;
      end = start + 60*used_duration;  * won"t take care of all problems;
                                          * Check for problems;
      end;
    else if (first.date) or ( end - 60*used_duration > prev_end ) then do;
             old_start = start;
       start = end - 60*used_duration;
             end;
          else error_code = "Cannot fit duration";
  end;
  end_hr = hour(end);
  end_min = minute(end);
  start_hr =hour(start);
  start_min = minute(start);
  prev_end = lag(end);
  if first.date then prev_end = .;
  run;
proc sort data=rev_duration_counts;
  by descending date descending start;
```

```
  run;
data rev_duration_counts;
    set rev_duration_counts;
            format next_start hhmm. next_date date9.;
            by descending date;
            next_date = lag1(date);
            next_start = lag1(start);
            if first.date then next_start = .;
            run;
proc print data=rev_duration_counts;
  where error_code ne " ";
  run;
proc sort data=rev_duration_counts;
  by date start;
  run;
data checker;
    set rev_duration_counts;
            by date;
            if end_min = 0 then do;
      end_min = 60;
              end_hr = end_hr - 1;
              end;
            prev_end_min = lag1(end_min);
            if first.date then prev_end_min = .;
            run;
proc print data=checker (obs=5);
  var date prev_date next_date start start_hr start_min next_start end end_hr end_min prev_end;
  where year = 2011;
  run;
data into_hours;
  set checker;
  where start_hr ge 6 and end_hr le 17;
  by date;
  if first.date then start_h = 6;
    else start_h = start_hr;
  if last.date then end_h = 17;
    else end_h = hour(next_start) - 1;   * <== next;
  do h =  start_h to end_h; *0;
  if (h lt start_hr) then do; * 1;
    new_duration = 60;
    Cover = "Not Covered";
    step = 1; output;
  end; * 1;
```

```
if h = start_hr then do; * 2;
  if start_min ne 0 then do; *2.1;
    new_duration = start_min;
          Cover = "Not Covered";
    step = 2.1; output;
  end; * 2.1;
  if start_hr < end_hr then do; *2.2;
    new_duration = 60 - start_min;
          Cover = "Covered";
    step = 2.2; output;
          end; * 2.2;
  if (start_hr = end_hr) then do; *2.3;
    new_duration = end_min - start_min ;
            Cover = "Covered";
            step = 2.3; output;
    if (end_min < 60) then do; *2.4;
     new_duration = 60 - end_min;
            Cover = "Not Covered";
            step = 2.4; output;
    end; *2.4;
         end; *2.3;
end;  *2 ;
if h > start_hr then do; * 4;
  if h < end_hr then do; * 4.1;
    new_duration = 60;
            Cover = "Covered";
            step = 4.1; output;
        end;  *4.1 ;
  if h = end_hr then do; * 4.2;
          new_duration = end_min;
            Cover = "Covered";
            step = 4.21; output;
     new_duration = 60 - end_min;
                    Cover = "Not Covered";
                    step = 4.22; if new_duration gt 0 then output;
  end; *4.2;
  if h > end_hr then do; * 4.3;
    new_duration = 60;
            Cover = "Not Covered";
            step = 4.3; output;
          end; * 4.3;
end; * 4;
end; *0;
```

```
 run;
proc freq data=into_hours;
  table h;
  run;
proc sort data=into_hours;
  by date h cover;
  run;
proc means noprint data=into_hours;
  var new_duration;
  by date h cover;
  id  year month day ;
  output out=new_into_hours sum=new_duration;
  run;
%let vlist = TV OS BE NH SS CH NG RS BW RT RL GE AK ML PG;
options mprint spool;
%macro add_birds();
data bird_into_hours;
  set new_into_hours;
  file = 'bird';
  hour=h;
  %do i=1 %to 15;
     %scan(&vlist,&i) = .;
     %end;
  run;
%mend;
%add_birds()
data bird_into_hours;
  set bird_into_hours;
  drop h _type_ _freq_;
  days_since_8_1 = date - mdy(8, 1, year);
  *used_duration = new_duration;
  run;
data initialize;
 format init_day date9. date date9. cover $11.;
 Cover = 'Not Covered';
 file = 'init';
 New_Duration = 60;
 do year = 1972 to 2011;
     init_day = mdy(8, 1, year);
 do days_since_8_1 = 0 to 136;
    date = init_day + days_since_8_1;
           month = month(date);
           day = day(date);
```

```
 do hour = 6 to 17;
   output;
 end;
 end;
 end;
 run;
data break_into_hours;
  set bird_into_hours initialize;
  run;
proc sort data=break_into_hours;
  by year days_since_8_1 hour file;
  run;
data sasdata.break_into_hours;
  set break_into_hours;
  by year days_since_8_1 hour file;
  if file = 'init' and not (first.date or first.days_since_8_1 or first.hour) then delete;
  run;
```

7.2.1.2 Data Set used for Hourly Predictions

```
data others;
  match = 1;
  used_duration = 60;
  ln_duration = log(60);
  hour=11;
  days_since_8_1 = 45;
  year = 2011;
  run;
proc sort data=sasdata.clean_counts;
  by hour;
  run;
data temporary;
  set sasdata.clean_counts;
  where days_since_8_1 ge 22 and days_since_8_1 le 61 and hour ge 6 and hour le 17;
  keep hour;
  run;
data temporary;
  set temporary;
  format plot $25.;
  by hour;
  match = 1;
  if first.hour;
  plot = "hour";
  run;
```

```
data hour;
  merge others temporary;
  by match;
  drop match;
  run;
data sasdata.BW_with_plot;
  set  sasdata.clean_counts hour;
  where hour ge 6 and hour le 17 and days_since_8_1 ge 22 and days_since_8_1 le 61;
  run;
```

7.2.1.3 Cross-Validation

```
data plot;
  do year = 1972 to 2011;
  do days_since_8_1 = 22 to 61;
  do hour = 6 to 17;
  plot = 'Valid';
  day_center = days_since_8_1 - 40;
  if day_center ge 0 then d40 = (day_center - 0)**2;
    else d40 = 0;
  if day_center ge 10 then d50 = (day_center - 10)**2;
    else d50 = 0;
   hour_center = hour - 11;
     if hour_center ge -2 then h9 = (hour_center + 2)**2;
        else h9 = 0;
    if hour_center ge 4 then h15 = (hour_center – 4)**2;
    else h15 = 0;
  output;
  end;
  end;
  end;
  run;
data plot;
  set plot;
  if not(year in (1972 1973 2005 2006));
  run;
data counts;
  set sasdata.clean_counts;
  where hour ge 6 and hour le 17 and days_since_8_1 ge 22 and days_since_8_1 le 61;
  day_center = days_since_8_1 - 40;
  if day_center ge 0 then d40 = (day_center - 0)**2;
    else d40 = 0;
  if day_center ge 10 then d50 = (day_center - 10)**2;
    else d50 = 0;
```

```
 hour_center = hour - 11;
    if hour_center ge -2 then h9 = (hour_center + 2)**2;
       else h9 = 0;
   if hour_center ge 4 then h15 = (hour_center – 4)**2;
    else h15 = 0;
keep year days_since_8_1 month day hour hour_center h9 h15 day_center d40 d50
used_duration bw;
  run;
data about_60;
  set counts;
  where used_duration = 60 and not (year in (1972 1973 2005 2006) );
  run;
%macro cross(dist, z_dist);
%do year = 1972 %to 2011;
 data wo_year;
    set about_60;
      plot = 'Data ';
           where year ne &year;
           run;
  data plotting;
    set plot;
           where year ne &year;
    run;
  data both;
    set wo_year plotting;
           run;
proc genmod data=both;
    title "&year";
           class year;
    model bw = year
              hour_center|hour_center h9 h15
                     day_center|day_center d40 d50/dist=&dist;
    output out=genpred p=pred;
           ods output ConvergenceStatus=Convergence;
    run;
data Convergence;
  set Convergence;
  year = &year;
  model = "&dist";
  drop status;
  run;
proc append data=convergence base=all_convergence;
  run;
```

```
data genpred;
   format model $5.;
   set genpred;
   where plot = 'Valid';
   model =  "&dist";
bird = "bw";
   keep year days_since_8_1 hour bird model pred ;
   run;
proc sort data=genpred;
  by days_since_8_1 hour;
  run;
proc means noprint data=genpred;
  var pred;
  id year model bird;
  by days_since_8_1 hour;
  output out=out_gen mean=pred;
  run;
data out_gen;
  set out_gen;
  year = &year;
  drop _freq_  _type_;
  run;
proc append data=out_gen base=all_pred;
  run;
proc genmod data=both;
   title "&year";
         class year/param=ORTHPOLY;
   model bw = year
          hour_center|hour_center h9 h15
                              day_center|day_center d40 d50/dist=&z_dist;
   zeromodel year
          hour_center|hour_center h9 h15
                              day_center|day_center d40 d50;
   output out=z_genpred p=pred ;
   ods output ConvergenceStatus=Convergence;
   run;
data Convergence;
  set Convergence;
  year = &year;
  model = "&z_dist";
  drop status;
  run;
proc append data=convergence base=all_convergence;
```

```
  run;
data z_genpred;
    set z_genpred;
          format model $5.;
    where plot = "Valid";
    model =  "&z_dist";
          bird = "bw";
    keep year days_since_8_1 hour bird model pred;
    run;
proc sort data=z_genpred;
  by days_since_8_1 hour;
  run;
proc means noprint data=z_genpred;
  var pred;
  id year model bird;
  by days_since_8_1 hour;
  output out=z_out_gen mean=pred;
  run;
data z_out_gen;
  set z_out_gen;
  year = &year;
  drop _type_ _freq_;
  run;
proc append data=z_out_gen base=all_pred;
  run;
%end;
%mend;
data all_pred;
  format year 5.0  days_since_8_1 5.0 hour 5.0 model $5. bird $5. pred 15.8 ;
  run;
data all_convergence;
  format year 5.0 model $5. reason $205.;
  run;
*options mprint spool;
%cross(nb,zinb);
%cross(poi,zip);
proc print data=all_Convergence;
  run;
 *NO CONVERGENCE ISSUES;

data sasdata.all_pred_using_hourly;
  set all_pred;
  run;
```

```
proc print data=sasdata.all_pred_using_hourly (obs=10);
  run;
proc sort data=counts;
  by year days_since_8_1;
  run;
proc means noprint data = counts;
  var bw;
  by year days_since_8_1;
  where year not in (1983, 1984, 1992, 2004, 2007, 2010);
  output out=bird_sum sum=total_count;
  run;
data bird_sum;
  set bird_sum;
  drop _TYPE_ _FREQ_;
  run;
proc sort data=sasdata.break_into_hours;
  by year days_since_8_1 hour ;
  run;
data break_cover;
  set sasdata.break_into_hours;
  where cover = 'Covered' and hour ge 6 and hour le 17 and days_since_8_1
              ge 22 and days_since_8_1 le 61;
  keep year days_since_8_1 hour new_duration cover;
  run;
proc sort data=all_pred;
  by year days_since_8_1 hour ;
  run;
proc sort data=break_cover;
  by year days_since_8_1 hour ;
  run;
data with_pred;
  merge break_cover all_pred;
  by year days_since_8_1 hour;
  run;
data with_pred;
  set with_pred;
  where cover='Covered';
  fraction = new_duration/60;
  run;

proc sort data=with_pred;
  by model year days_since_8_1;
  run;
```

```
proc means noprint data=with_pred;
  where model ne ' ' and year not in (1983, 1984, 1992, 2004, 2007, 2010);
  var pred new_duration;
  weight fraction;
  by model year days_since_8_1;
  output out=pred_sum sum=total_pred;
  run;
proc sort data=pred_sum;
  by year days_since_8_1;
  run;
proc sort data=bird_sum;
  by year days_since_8_1 ;
  run;
data sums;
  merge pred_sum bird_sum;
  by year days_since_8_1;
  chisq = (total_count - total_pred)**2/total_pred;
  if total_count = . then delete;
  run;
proc sort data=sums;
  by model;
  run;
proc means noprint data=sums;
  var chisq;
  by model;
  output out=chisq sum=chisq;
  run;
data sasdata.chisq_daily;
  set chisq;
  chisq_div_df = chisq/_freq_;
  run;
proc print data=sasdata.chisq_daily;;
  run;
```

7.2.1.4 Model Run with Assumed Distribution

```
data BW_with_plot;
  set sasdata.BW_with_plot;
  where hour ge 6 and hour le 17;
  day_center = days_since_8_1 - 40;
  if day_center ge 0 then d40 = (day_center - 0)**2;
    else d40 = 0;
  if day_center ge 10 then d50 = (day_center - 10)**2;
    else d50 = 0;
```

```
  hour_center = hour - 11;
    if hour_center ge -2 then h9 = (hour_center + 2)**2;
       else h9 = 0;
   if hour_center ge 4 then h15 = (hour_center – 4)**2;
    else h15 = 0;
  run;
proc genmod data=BW_with_plot;
  class year;
  where hour ge 6 and hour le 17 and days_since_8_1 ge 22 and days_since_8_1 le 61 and
used_duration = 60 and (not (year in (1972 1973 2005 2006)));
  model bw = year
           hour_center|hour_center h9 h15
                              day_center|day_center d40 d50/dist=nb;
    output out=outpm p=predict;
  run;
data sasdata.hour_pred;
  set outpm;
  where plot='hour';
  merge = 1;
  keep hour predict merge;
  run;
proc means sum data=sasdata.hour_pred;;
  var predict;
  output out=total sum=total;
  run;
data total;
  set total;
  merge = 1;
  drop _freq_ _type_;
  run;
data sasdata.hours;
  merge sasdata.hour_pred total;
  by merge;
  fraction = predict/total;
  keep hour fraction;
  run;
data break_into_hours;
  set sasdata.break_into_hours;
  where cover = "Covered";
  run;
proc sort data=break_into_hours;
  by hour;
  run;
```

```
data with_fraction;
  merge break_into_hours hours;
  by hour;
  cover_fraction=(new_duration/60)*fraction;
  run;
proc sort data=with_fraction;
  by year month day;
  run;
proc means noprint data=with_fraction;
   var cover_fraction;
   by year month day;
   output out=cover_frac sum=cover_fraction;
   run;
data cover_frac;
  set cover_frac;
  adj_duration = cover_fraction*12;
  run;
data clean_counts;
  set sasdata.clean_counts;
  where days_since_8_1 ge 22 and days_since_8_1 le 61 and hour ge 6 and hour le 17;
  run;
proc sort data=clean_counts;
  by year month day;
  run;
proc means noprint data=clean_counts;  *only want sums within hourly and seasonal window!!!;
  var bw;
  by year month day;
  output out=out_bw_day_tot sum=bw_day_tot;
  run;
data sasdata.with_adj_duration;
  merge out_bw_day_tot cover_frac;
  by year month day ;
  days_since_8_1 = mdy(month,day,year)-mdy(8, 1, year);
  ln_adj_duration=log(adj_duration);
  run;
```

7.2.2 Yearly Counts with Adjusted Durations

7.2.2.1 Data Set used for Daily Predictions

```
data others;
  match = 1;
  adj_duration = 12;
  ln_adj_duration = log(12);
```

```
  year = 2011;
  run;
proc sort data = sasdata.with_adj_duration;;
  by days_since_8_1;
  run;
data temporary;
  set sasdata.with_adj_duration;;
  where days_since_8_1 ge 22 and days_since_8_1 le 61;
  keep days_since_8_1;
  run;
data temporary;
  set temporary;
  format plot $25.;
  by days_since_8_1;
  match = 1;
  if first.days_since_8_1;
  plot = "days_since_8_1";
  run;
data days_since_8_1;
  merge others temporary;
  by match;
  drop match;
  run;
proc print data=days_since_8_1;
  run;
data sasdata.BW_with_plot_2;
  set  sasdata.with_adj_duration days_since_8_1;
  where days_since_8_1 ge 22 and days_since_8_1 le 61;
  run;
```

7.2.2.2 Cross-Validation

```
data plot;
  do year = 1972 to 2011;
  do days_since_8_1 = 22 to 61;
  plot = 'Valid';
  adj_duration = 12;
  ln_adj_duration = log(adj_duration);
  day_center = days_since_8_1 - 40;
  if day_center ge 0 then d40 = (day_center - 0)**2;
    else d40 = 0;
  if day_center ge 10 then d50 = (day_center - 10)**2;
    else d50 = 0;
    output;
```

```
  end;
  end;
  run;
data counts;
  set sasdata.with_adj_duration;
  where days_since_8_1 ge 22 and days_since_8_1 le 61;
  day_center = days_since_8_1 - 40;
  if day_center ge 0 then d40 = (day_center - 0)**2;
    else d40 = 0;
  if day_center ge 10 then d50 = (day_center - 10)**2;
    else d50 = 0;
  run;
%macro cross(dist, z_dist);
%do year = 1972 %to 2011;
  data wo_year;
    set counts;
          plot = 'Data ';
          where year ne &year;
          run;
  data plotting;
       set plot;
          where year ne &year;
    run;
  data both;
    set wo_year plotting;
          run;
proc genmod data=both;
    title "&year";
          class year;
    model bw_day_tot = year day_center|day_center d40 d50/offset=ln_adj_duration dist=&dist;
    output out=genpred p=pred;
          ods output ConvergenceStatus = Convergence;
    run;
data Convergence;
  set Convergence;
  year = &year;
  model = "&dist";
  drop status;
  run;

proc append data=convergence base=all_convergence;
  run;
data genpred;
```

```
    format model $5.;
    set genpred;
          where plot = 'Valid';
    model =  "&dist";
          bird = "bw";
    keep year days_since_8_1  bird model pred ;
    run;
proc sort data=genpred;
  by days_since_8_1 ;
  run;
proc means noprint data=genpred;
  var pred;
  id year model bird ;
  by days_since_8_1;          *getting average daily count over all years to use as predicted values
for the year of "interest" in validation set;
  output out=out_gen mean=pred;
  run;
data out_gen;
  set out_gen;
  year = &year;
  drop _freq_  _type_;
  run;
proc append data=out_gen base=all_pred;
  run;
proc genmod data=both;
    title "&year";
          class year;
    model bw_day_tot = year day_center|day_center d40 d50/offset=ln_adj_duration
dist=&z_dist;
    zeromodel  day_center|day_center ;
    output out=z_genpred p=pred ;
    ods output ConvergenceStatus=Convergence;
    run;
data Convergence;
  set Convergence;
  year = &year;
  model = "&z_dist";
  drop status;
  run;

proc append data=convergence base=all_convergence;
  run;
data z_genpred;
```

```
    set z_genpred;
          format model $5.;
          where plot = 'Valid';
    model =  "&z_dist";
          bird = "bw";
    keep year days_since_8_1  bird model pred;
    run;
proc sort data=z_genpred;
  by days_since_8_1 ;
  run;
proc means noprint data=z_genpred;
  var pred;
  id year model bird ;
  by days_since_8_1 ;
  output out=z_out_gen mean=pred;
  run;
data z_out_gen;
  set z_out_gen;
  year = &year;
  drop _type_ _freq_;
  run;
proc append data=z_out_gen base=all_pred;
  run;
%end;
%mend;
data all_pred;
  format year 5.0  days_since_8_1 5.0 model $5. bird $5. pred 15.8 ;
  run;
data all_convergence;
  format year 5.0 model $5. reason $205.;
  run;
options mprint spool;
%cross(nb,zinb);
%cross(poi,zip);
proc print data=all_Convergence;
  run;
*NO CONVERGENCE ISSUES;
proc sort data=all_pred;
  by year days_since_8_1 ;
  run;
proc sort data=counts;
  by year days_since_8_1;
  run;
```

```
data with_pred;
  merge counts all_pred;
  by year days_since_8_1;
  pred_bw_day_tot = (adj_duration/12)*pred;
  chisq = (bw_day_tot - pred_bw_day_tot)**2/pred_bw_day_tot;
  if pred = . then delete;
  run;
proc sort data=with_pred;
  by model;
  run;
proc means noprint data=with_pred;
  var chisq;
  by model;
  output out=chisq sum=chisq;
  run;
data sasdata.chisq_yearly;
  set chisq;
  chisq_div_df = chisq/_freq_;
  run;
proc print data=sasdata.chisq_yearly;
  run;
```

7.2.2.3 Model Run with Assumed Distribution

```
data BW_with_plot_2;
  set sasdata.BW_with_plot_2;
  day_center = days_since_8_1 - 40;
  if day_center ge 0 then d40 = (day_center - 0)**2;
    else d40 = 0;
  if day_center ge 10 then d50 = (day_center - 10)**2;
    else d50 = 0;
  run;
proc genmod data=BW_with_plot_2;
  where days_since_8_1 ge 22 and days_since_8_1 le 61;
  class year;
  model bw_day_tot = year day_center day_center*day_center d40 d50  /dist=zinb
offset=ln_adj_duration;
  zeromodel day_center day_center*day_center;
  output out=outpm  p=predict pzero=zero_pred;
  run;
proc sgplot data=outpm;
  where plot="days_since_8_1";
  scatter y=predict x=days_since_8_1;
  run;
```

```
data sasdata.seasonal_pred;
  set outpm;
  where plot="days_since_8_1";
  keep days_since_8_1 predict merge;
  merge = 1;
  run;
proc means sum data=sasdata.seasonal_pred;
  var predict;
  output out=total sum=total;
  run;
data total;
  set total;
  merge = 1;
  run;
data sasdata.seasonal;
  merge sasdata.seasonal_pred total;
  by merge;
  fraction = predict/total;
  keep days_since_8_1 fraction;
  run;
proc print data=sasdata.seasonal;
  run;
*Check;
proc means sum data=seasonal;
  var fraction;
  run;
proc sort data=sasdata.with_adj_duration;
  by days_since_8_1;
  run;
data with_fraction;
  merge sasdata.with_adj_duration seasonal;
  by days_since_8_1;
  cover_fraction = (adj_duration/12)*fraction;
  run;
proc sort data=with_fraction;
  by year;
  run;
proc means noprint data=with_fraction;
   var cover_fraction;
   by year;
   output out=cover_frac sum=cover_fraction;
   run;
data cover_frac;
```

```
  set cover_frac;
  year_duration = cover_fraction*40;  * For days 22 to 61;
  run;
proc sort data=sasdata.with_adj_duration;
  by year;
  run;
proc means noprint data=sasdata.with_adj_duration;
  var bw_day_tot;
  by year;
  output out=out_bw_yr_tot sum=bw_yr_tot;
  run;
data sasdata.bw_yr_tot_with_dur;
  merge out_bw_yr_tot cover_frac;
  by year ;
  run;
data sasdata.bw_yr_tot_with_dur;
  set sasdata.bw_yr_tot_with_dur;
  year_class = year;
  bw_per_day = bw_yr_tot/year_duration;
  ln_yr_duration = log(year_duration);
  run;
```

7.2.3 Testing Annual Trends

```
data year;
  plot = 'year';
  ln_yr_duration = log(40);
  do year = 1972 to 2011 by 0.1;
    output;
end;
  run;
data with_plot;
  set sasdata.bw_yr_tot_with_dur year ;
  run;
proc glimmix data=with_plot method=laplace;
  parms 0;
  class year_class;
  model bw_yr_tot =  year/dist=poisson ddf=38 offset=ln_yr_duration;
  random year_class;
  run;
proc glimmix data=with_plot maxopt=1000 method=laplace;
  parms 0 3;
  nloptions technique=newrap gconv2=1.0e-03 ;
  class year_class;
```

```
  model bw_yr_tot =  year/dist=negbin ddf=39 offset=ln_yr_duration;
  random year_class;
  run;
proc glimmix data=with_plot method=laplace;
  model bw_yr_tot =  year/dist=poisson  offset=ln_yr_duration;
  run;
proc glimmix data=with_plot method=laplace;
  model bw_yr_tot =  year/dist=negbin  offset=ln_yr_duration;
  run;
```