EcoStat, Inc.
PO Box 425
Mebane, NC 27302

Ph/Fx: (919) 304-6029 billwh@mindspring.com

September 9, 2009

To: Renee Culver, Joanie Stewart, Loan Tran, Emre Ergas

From: William Warren-Hicks, Song Qian

Subject: A Probability-Based Approach for Estimating Avian Mortality

Below is a document describing a probability model for estimating avian mortality at the Altamont. The model formally addresses many of the issues that are under discussion during the SRC phone calls including survey error, the probability of a true zero, and the true probability of mortality at any sampling unit. The model is consistent with the underlying concepts of the baseline and current survey designs, and the model can be adapted to include variables that may be important indicators of avian mortality including MW capacity, operating minutes, rotor swept area, and others.

Please call with any questions.

A Probability-Based Approach for Estimating Avian Mortality

The following discussion presents a probability-based formal model for estimating avian mortality at the Altamont. Components of the model have been tested with a prior data set and found to work well. A few of the key issues described below have not been evaluated using the current Altamont data, but will be implemented in the near future. The following discussion provides details on a formal mathematical approach that is statistically tenable, and could be used to estimate Altamont-wide avian mortality. The model is internally consistent, an important issue in probability-based model building. We may, after further evaluation of the model formulation, make changes to the probability structure as required to maximize the relationship between the probability formulation and a conceptual model of avian mortality at the Altamont. Advantages to the probability model described below follow:

1. The mathematical structure is transparent and the underlying probability structure and associated assumptions are provided in detail.

2. The probability model is consistent with the types of data collected in both the baseline and current study data sets, and the model has the ability to utilize many of the variables in the data set to estimate avian mortality.

3. The model structure is flexible, and will allow investigators the ability to test and evaluate relationships between the observed avian mortality and associated covariables like MW capacity, operating minutes, number of turbines, etc.

4. The probability model formally incorporates the many sources of sampling error directly into the probability framework, therefore eliminating the need to utilize data from outside sources. Sample error estimates resulting from the model can be compared to estimates currently used in the MT report.

5. Uncertainty estimates of the model parameters, as well as model predictions, are created during the process of implementing the model.

6. Outputs from the model can be shown in graphical form, therefore increasing communication the modeling results.

7. The model can be run on data subsets (e.g., different years of information), therefore allowing comparison and analysis of both temporal-specific and spatial-specific estimates of mortality.

8. The model can be setup to reflect the original survey design, with the concept of sampling unit directly specified in the data used to parameterize the model.

Introduction

Uncertainty in post-construction monitoring data arises from multiple sources, including sampling uncertainty. A number of approaches to estimating sampling uncertainty based on sampling error exist in the literature, and these approaches are usually directly linked with methods generating estimates of uncertainty in the population average or total. However, without a formal survey design that is implemented consistently during the survey timeframe, the resulting estimation methods associated with common sampling error models may be misleading.

For any given survey date during a post-construction monitoring program, the chance that the field survey crew finds a dead bird is low. That is, the total number of dead birds observed relative to the number of searches may be low. Indeed, even under the assumption of high true mortality a small local avian population will often result in a potentially small number of observed fatalities. Over time, the resulting data set generated as a result of the survey is dominated by zeros (defined as a search resulting in zero bird finds), as is the case with the MT current survey data set. The MT survey is challenging for two reasons. First, some of the zeros are true, indicating no mortality associated with the sampling unit (i.e., a turbine string or possibly a plot). However, due to a large number of possible causes (including observer error, scavenging during the survey period, weathering, etc.) there is a chance that the observations incorrectly reflect the true mortality at the location of interest during the time of the survey. Second, sampling strings are located over a wide and diverse geographic area, and the sampling unit characteristics are variable. For these two complicating factors, we believe that a model-based estimation approach may be more appropriate than a classical approach which is directly tied to a formal survey design. We are not discounting the use of formal survey design estimating equations, however. But, given the current confusion about the survey design itself, the following probability-based model may result estimates with a reasonably high degree of belief.

Count data are often modeled using Poisson or negative binomial regression (a form of overdispersed Poisson models). These models provide specific formulas for calculating the conditional probability of a 0 count, given the true number of dead birds. An implicit assumption of these models is that the true (unobserved) number of dead birds is larger than 0. When a 0 can be a true 0 [no dead bird(s)], the number of 0s in the data would be larger than the models can predict. To account for the inflated number of 0s, a mixed hierarchical model is proposed by Lambert (1992)[1]. In addition to the ambiguity associated with 0s, our monitoring data are also associated with uncertainty when the observed counts are positive but small (e.g., one observed dead bird may indicate more have been killed but not observed). The uncertainty associated with positive counts is a difficult statistical issue and can lead to biased estimates of model coefficients.

The following discussion presents two model forms for resolving these issues. The two model forms are based on different probability models and we plan to test and refine each approach. The results from each model form can be compared, which provides us with some degree of internal quality assurance checking of the prediction consistency.

We begin by generating a model that works is consistent with the current MT survey and baseline data sets. We then expand the model for use with the Altamont-wide list frame (i.e.,

[1] Lambert, D. 1992, Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, Technometrics, 34(1): 1-14.

those sampling units that have not been surveyed). In this approach, we can generate mortality estimates for the surveyed sampling units as well as forecast the potential mortality at sampling units that have not been surveyed.

The Zero-Inflated Poisson Regression Model

In the first stage, we use a zero-inflated binomial regression to estimate the probability of a "true 0", followed by the use of the zero-inflated Poisson regression model for estimating the underlying mortality rate for both currently surveyed sampling units and those with no survey information.

Below, we discuss the details of the proposed modeling approach. The objective of the following discussion is to provide an example of the models and approaches we will test using data collected during this project.

The most frequently used probabilistic model for a count variable is the Poisson distribution. A Poisson distribution is defined by one parameter, λ:

(eq. 1):

$$\Pr(Y=y) = e^{-\lambda}\lambda^{y} / y!$$

The parameter λ is the expected value of the count variable, Y. Given a non-negative λ, we can calculate the probability of $Y=0, 1, 2, \ldots, n$. The probability of observing a 0 (y=0) is $e^{(-\lambda)}$. That is, observing a zero is possible even when the expected value of the count variable is positive. However, this model assumes that there are some birds killed in the sampling area, and a 0 count can be seen as a form of sampling error. The Poisson model provides probabilistic assessment of the likelihood of reporting a 0 when there are on average λ dead birds. However, this model is potentially wrong. A zero can be a true 0 or a false zero. In the context of bird mortality, a true 0 suggests that there was no bird killed during the observation period. A false 0 suggests that the observed 0 may be in error. A false 0 occurs when the observer did not view the dead bird, or the dead bird was removed by a scavenger, or other reasons. When a 0 is observed, we are uncertain whether a 0 is a true 0, or not.

The zero-inflated Poisson model can be expressed in two parts. First, the probability of observing zero dead birds at sampling unit *i* is computed as the sum of the probability that there is no dead bird at the site (p_i) and the probability that the observer recorded a zero when there was at least one dead bird. Second, when the observed mortality is larger than 0, the observed number is modeled by the Poisson model:

eq. 2:

$$\pi_i(y_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & \text{if } y_i = 0 \\ (1-p_i)\left(e^{-\lambda_i}\lambda_i^{y_i} \big/ y_i!\right) & \text{if } y_i > 0 \end{cases}$$

where $\pi(y_i)$ is the density of the i_{th} observation(i.e., sampling unit), p_i is the probability that there is no dead bird at the site (probability of "true zero") where the i_{th} observation was made. The probability that there are dead birds at the site is then $(1-p_i)$. Assuming a Poisson model for the observed counts, when there are dead birds at the sampling unit, the number of observed dead birds follows a Poisson distribution. The probability of observing 0 is $e^{(-\lambda)}$, and λ_i is the expected number of dead birds, the parameter of interest.

We can also model the underlying Poisson parameter λ_i as a function of turbine string characteristics (x):

$$\log(\lambda_i) = c + dx_i$$

which would allow us to make forecasts for sites without observations but with known characteristics, x.

When the positive counts are uncertain, transforming the count data into a binary presence/absence variable can often reduce the uncertainty in the estimation of p_i because the Poisson model parameter is estimated based on observed counts, therefore any uncertainty associated with each observation will be propagated into the estimated model parameters. When transforming the count data into a binary variable, a positive count is converted to "presence" without ambiguity. As a result, uncertainty in the resulting binary variable is mostly associated with "absence." In this formulation of the probability model the estimation of p_i is relatively straight forward. Once p_i is well estimated, we can then plug it back into the original Poission likelihood function (equation 2).

Therefore, we convert the count data from each search into a binary variable (1=presence, and 0=absence), and model the resulting binary variable using a binomial model. In this new formulation, let y be the number of searches associated with a sampling unit that resulted in a positive (found at least one dead bird) result in N total searches, the basic model is

eq. 3:

$$y \sim \text{Binomial}(p_{obs}, N)$$

where p_{obs} is the probability of observing a positive outcome in a given search. This probability is the chance of actually observing a dead bird, not the probability of a bird kill at a site. The underlying model should consider two additional complicating factors, when observing a negative outcome (no dead bird).

1. There is a chance that there was no dead bird, and
2. There is a sampling error leading to a false negative result.

For the first factor, we can introduce a probability of a bird being killed at a given site (p_{kill}), and a sampling error (p_{sample}) representing the conditional probability of reporting a dead bird when the dead bird is present, and revise the basic model into a model with three steps:

$$
\begin{aligned}
&y_i \sim \text{Binomial}\ (p_{obsi}, N_i)\\
&p_{obsi} \sim kill_i * p_{sample\mathrm{i}}\\
&kill_i \sim \text{Binomial}(p_{killi}, 1)
\end{aligned}
$$

The model can be read as follows. For each sampling unit, the underlying probability of a bird being killed is p_{killi}, an unknown parameter to be estimated. The probability 1-p_{kill} is probability of "true zero" (or p_{i} in equation 2). The variable $kill_i$ is binary a taking value 0 (no kill) or 1 (kill). The parameter $p_{sample\mathrm{i}}$ is a conditional probability because it is used only when there is a kill, or $kill_i = 1$. So, we interpret p_{sample} as the probability of reporting a dead bird when there is at least one kill. The probability of reporting a kill (without knowing the true status) p_i is either 0 or p_{obs}.

When predictor variables are available, the mean parameters can be modeled using the generalized linear models:

$$
\begin{aligned}
&\text{logit}(p_{killi}) \sim X\beta\\
&\text{logit}(p_{samplei}) \sim Z\gamma
\end{aligned}
$$

For example, we could build a model using the following two predictor variables: (1) number of searches associated with each sampling unit (N_i), and (2) number of wind turbines ($Turbs_i$) in the sampling unit (assumed here to be a string):

$$
\begin{aligned}
&\text{logit}(p_{killi}) &= a + b \log(\text{Turbs}_{\text{i}})\\
\text{eq. 3} \quad &\text{logit}(p_{samplei}) &= \alpha + \beta \log(\text{N}_{\text{i}})
\end{aligned}
$$

With this binary response model, we can easily estimate the probability of true zero and apply it to the zero-inflated Poisson model. Overdispersion can be accounted for by introducing random error terms to equation 3.

The following narrative is based on the above model. Note that we can (and plan) to try different models for estimating pkill and psample. Other models could incorporate MW, total number of operating minutes, total rotor diameter, etc.

Computation

A Markov chain Monte Carlo simulation (MCMC) method is used for estimating model parameters (a, b, α, β). The WinBugs language is the most commonly used programming environment for MCMC. In a Bayesian model, prior information is readily incorporated into the computation method and the results can be presented in terms of the posterior distributions of model parameters. For this problem, vague flat prior distributions for the four model parameters were used reflecting the fact that there is no information on these parameters prior to the collection of the survey data.

Example WinBugs code for this type of model follows:

```
model
{
  for( i in 1 : n.str) {
    Hits[i] ~ dbin(P[i],N[i])
    Kills[i] ~ dbern(Pturb[i])
    P[i] <- Kills[i] * Psample[i]
    logit(Pturb[i]) <- a + b*log(N.turb[i])
    logit(Psample[i]) <- alpha+beta*log(N[i])
  }
  a ~ dnorm(0, 0.0001)
  b ~ dnorm(0, 0.0001)
  alpha ~ dnorm(0, 0.0001)
  beta ~ dnorm(0, 0.0001)
}
```

Estimating Bird Mortality Rate

In summary, we can use the zero-inflated Poisson regression model for estimating the underlying mortality rate based on the MT current survey data. We introduced the zero-inflated binomial model to estimate the probability of a true 0. When the probability of true 0 is estimable, the zero-inflated Poisson model can be used to estimate the true mortality rate using the maximum likelihood estimator. The sampling error probability (p_{sample}) is also estimated in the zero inflated binomial regression model. Because p_{kill} and p_{sample} are included in the same model, the estimated probability of bird kill (p_{kill}) includes the effect of sampling error. When applying the zero-inflated Poisson model for estimating the underlying mortality (λ), only p_{kill} is needed.

The method adheres to the likelihood principle, which requires that statistical inference be made based only on the observed data. The basis for estimating the true mortality is the zero-inflated Poisson model (equation 2).

Maximum likelihood estimator of λ

Based on the zero-inflated Poisson model (equation 2), the likelihood of observing, say, k 0s and r non-zero counts is:

eq. 4: $$[1-p_{kill}+p_{kill}e^{-\lambda}]^k \Pi_{i=1}{}^r(p_{kill}e^{-\lambda}\lambda^{y_i}/y_i!)$$

Using the estimated p_{kill}, we can find the MLE of λ, which is the value of λ that maximizes eq. 4. Using the estimated λ, we can make probabilistic inference about the likely number of dead birds. For example, if a site is searched 10 times and returns 8 0s; one 1 and one 2, and the estimated p_{kill} is 0.8 (representing a gross exaggeration for the purpose of illustration), the likelihood function is

eq. 5:
$$L = (0.2 + 0.8\, e^{-\lambda})^8 (0.8e^{-\lambda}\lambda)(0.8e^{-\lambda}\lambda^2/2)$$

To maximize the likelihood L (solving λ by setting the partial derivative of L with respect to λ to be 0), the λ value is about 0.4 (Figure 1). That is, the maximum likelihood estimate of λ, unadjusted for the chance of a true zero, is 0.4. Using the zero-inflated model, we can draw the following probability distribution of the number of dead birds without adjusting for the possibility of a true zero:

Y	0	1	2	3	4	5
Pr	0.74	0.21	0.043	0.0057	0.0006	0.00004

This distribution shows that for any search, there is a high chance of finding no birds, and a very small chance of finding more than 1 bird. In the above formulation, λ is the expected number of birds given that the observed outcome is a positive value conditional on the available data. However, some zeros may be true zeros. Adjusting for the chance that a zero is a true event results in the the expected number of birds of 0.32 (= 0.4 *0.8, i.e, with pkill = 80% there is a 20% chance that the observed zeros are true zeros). Therefore, over a search interval of 10 events the expected number of dead birds is 3.2 birds (10*0.32). Uncertainty about the estimated λ can be estimated using the likelihood profile (Figure 1).

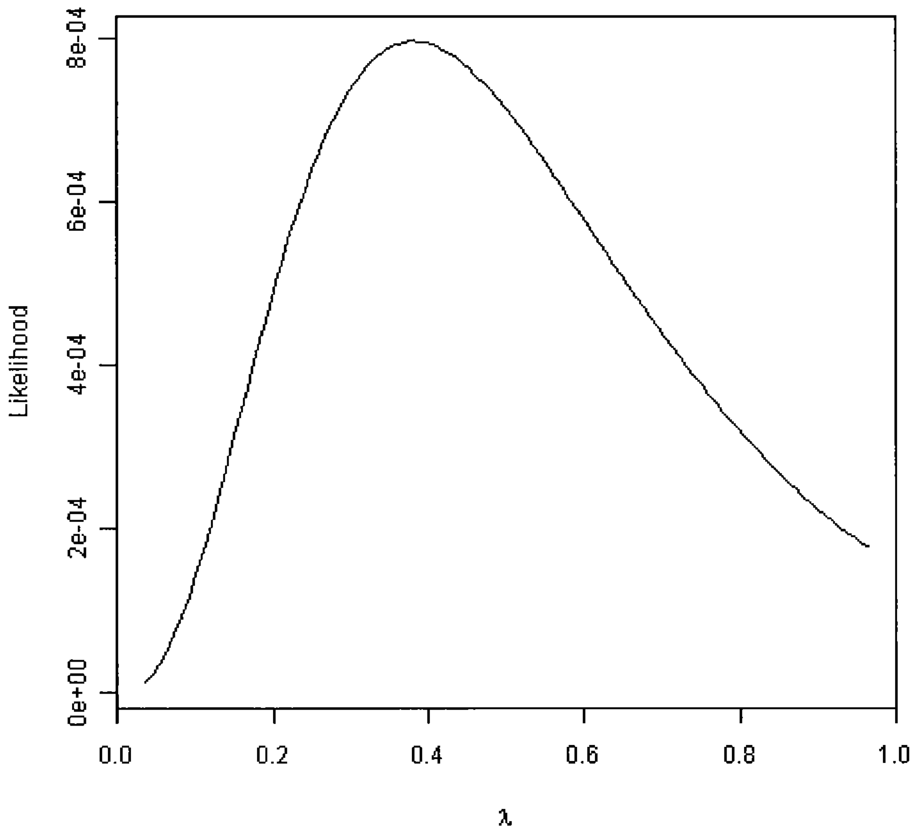

Figure 1: Likelihood profile (of a hypothetical data set) can be used for estimating the uncertainty on λ.

Predictions when observed data are all zeros

When all observed counts at a sampling unit are entirely 0s, the likelihood function is reduced to

$$[1-p_{kill}+p_{kill}e^{-\lambda}]^k$$

leading to the MLE of λ to be 0. In the above equation k is the number of observed zeros.

Figure 2 shows the likelihood profile using a p_{kill} of 0.8. The figure shows that when k is small, the likelihood profile is quite flat, suggesting a high level of uncertainty about the estimated λ. As k increases, the likelihood profile is increasingly concentrated at 0 suggesting our increased confidence about the estimated λ (of 0). Figure 2 represents a general process of learning through statistics. Specifically, if one search is conducted that results in a finding of no dead bird, we can infer that there was actually no mortality. But we are unable to put a high level of confidence on the conclusion. But if we conducted many searches and all returned 0 dead, our estimate of the true mortality is still 0, but we are more confident about the estimate when there are more observed 0s. The likelihood profiles in Figure 2 are rescaled such that the area under each curve is 1. These rescaled likelihood profiles are the same as the Bayesian posterior distributions of λ. We can use the 90th percentile of these posterior distributions as a conservative estimate of the upper bound of λ. The 90th percentile can be numerically estimated. They are 1.4, 0.3, 0.15, and 0.09 for k = 2, 10, 20, and 30, respectively. This result suggests that the more 0s we observe, the more likely the true bird mortality is 0.

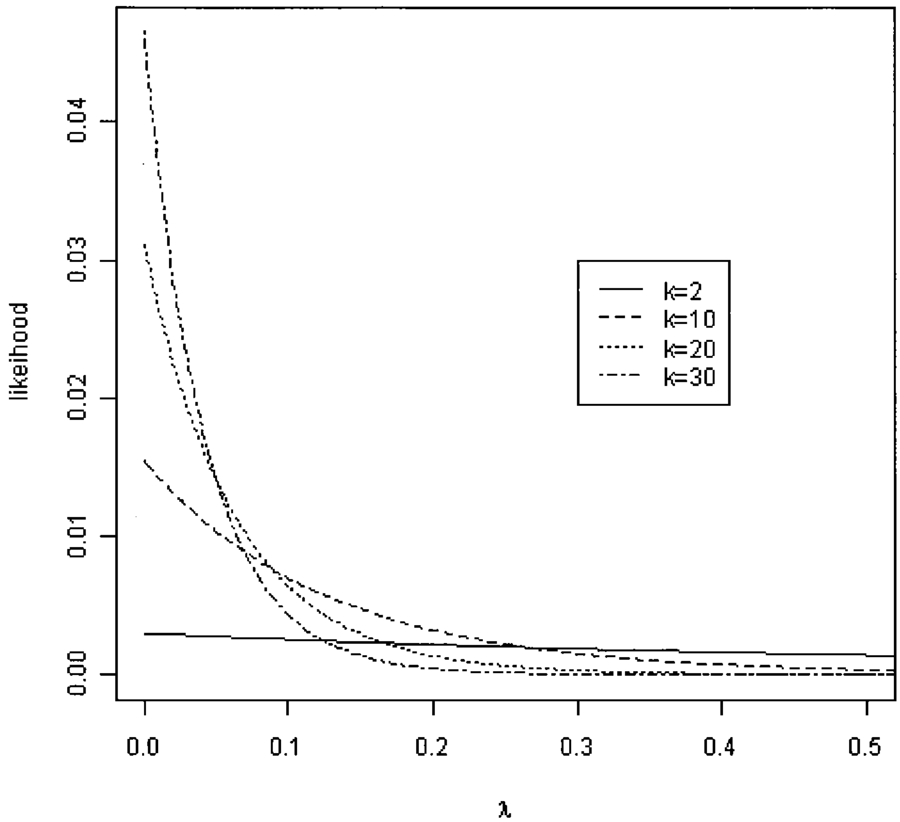

Figure 2. Likelihood profiles for λ of observing different numbers of 0s (k) are based on a true 0 probability of 0.2. As the number of 0s increases, we are increasingly confident about the conclusion that the likely mortality is 0 ($\lambda = 0$). The likelihood profile can be used to provide a conservative upper bound of λ.

The estimated probability of reporting a dead bird when present (p_{sample}) is often used to estimate the "true" magnitude of bird mortality by dividing the number of observed dead birds with this probability. This approach can be evaluated by comparison to the maximum likelihood estimator proposed here. When no dead bird is observed, either there was no bird kill or the dead bird is missed. The estimated p_{sample} is unlikely to provide much information in itself. Rather, p_{sample} is a vehicle for a better estimating and understanding the probability of a bird kill, p_{kill}, which is applied to the zero-inflated Poisson model for estimating the true mortality. The approach of treating a 0 count as a fraction of dead bird violates the basic statistical principle (the likelihood principle) and is counter intuitive. From a hypothesis testing perspective, an observed 0 is a piece of evidence supporting the null hypothesis that $\lambda = 0$, and a positive count is a piece of information against the null. The more 0s we observe, the stronger the evidence in favor of the null. The zero-inflated Poisson model follows this simple logic.

Expanding the model to predicting mortality for sites without observations

Because our model is fitted to covariates of site characteristics, we are able to predict both p_{sample} and p_{kill} for those sites without observations. These predicted probability values will provide information on the likelihood of bird kill and the likelihood of sampling error.

The zero-inflated Poisson regression can be also be used to develop models for λ (e.g., a simple log-linear model, $\log(\lambda_i) = c+dx_i$). This step requires that we combine the zero-inflated Poisson and zero-inflated binomial model together and link λ with covariates describing site characteristics. This step is a simple extension of the zero-inflated binomial regression model. With estimated p_{kill}, coefficients c and d (hence λ), we can estimate the exact number of kills for sites without observations. This process can be implemented in WinBUGS (suppose that we use the number of turbines as a covariate for λ):

```
model
{
  for( i in 1 : n.str) {
    counts[i]  ~ dpois(mu[i])
    mu[i] <- Pkill[i] * lambda[i]
    log(lambda[i]) <- c + d * log(N.turb[i])
    Hits[i] ~ dbin(P[i],N[i])
    Kills[i] ~ dbern(Pkill[i])
    P[i] <- Kills[i] * Psample[i]
    logit(Pkill[i]) <- a + b*log(N.turb[i])
    logit(Psample[i]) <- alpha+beta*log(N[i])
  }
  a ~ dnorm(0, 0.0001)
  b ~ dnorm(0, 0.0001)
  c ~ dnorm(0, 0.0001)
  d ~ dnorm(0, 0.0001)
  alpha ~ dnorm(0, 0.0001)
  beta ~ dnorm(0, 0.0001)
}
```

Ideally, we don't have to run the binomial model described above, but only run the full zero-inflated Poisson regression model. The WinBUGS code shown above combines both model forms. That is, we should be able to generate a more accurate estimate of pkill, and a relationship between lambda and string-specific covariates. The binomial model option can be run for the survey data to ensure that an accurate estimate of pkill is generated. Because both pkill and λ are predicted by string characteristics, we can use the zero-inflated Poisson model for predicting sites without observations. For those with observations, we can (but don't have to) use the MLE described in the document to generate a second set of estimates for λ. This will give us a chance to compare models and assess the level of uncertainty in the positive counts. If we decided that those positive counts are accurate, we can remove the binomial component.

From a pure statistical perspective, including those sampling units without observations is of no consequence to the model, because they will not change the likelihood function. But, including them in the analysis data set will allow automatic estimation of λ within the Bayesian computational method.

Please note that we did not include in this document the mathematics for implementing or solving the conditional probability model within the Bayesian probability paradigm. The calculus for computing the conditional probabilities of the random parameters from the joint distribution is not presented. Our purpose with this document is to explain the underlying probability framework, we did not provide a full mathematical explanation of the MCMC method and associated procedures for generating marginal or conditional distributions of the random model parameters.