COMPETING RISK MODELS FOR TURTLE NEST SURVIVAL IN THE BOLIVIAN AMAZON

by

RUOBING WANG

(Under the Direction of Jaxk Reeves)

ABSTRACT

Survival analysis is a very common method to model time-to-event data. In most studies, the probability of experiencing one event (such as failure or death) is investigated. Sometime, multiple events (competing risk events) are of interest. Competing risk events can be estimated crudely by separately modeling single events, but more complete models are necessary to perform complete competing risk modeling. In this thesis, a competing risk model is used to analyze *P.unifilis* (yellow-spotted Amazon River Turtle) nest survival based on data collected from 1910 nests observed in the Bolivian Amazon in 2005 and 2006. Under this scenario, turtle nests experience risk from animals, floods and humans. The results from competing risk models are evaluated to show the risk event and risk period for turtle nests.

INDEX WORDS: survival analysis censoring Kaplan-Meier estimate

competing risk model P.unifilis

COMPETING RISK MODELS FOR TURTLE NEST SURVIVAL IN THE BOLIVIAN AMAZON

by

RUOBING WANG

B.S., The South China Agricultural University, China, 1996

M.S., The East China Normal University, China, 2004

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2008

COMPETING RISK MODELS FOR TURTLE NEST SURVIVAL IN THE BOLIVIAN AMAZON

by

RUOBING WANG

Major Professor: Jaxk Reeves

Committee: William McCormick
Ron Carroll

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

COMPETING RISK MODELS FOR TURTLE NEST SURVIVAL IN THE BOLIVIAN AMAZON

by

RUOBING WANG

Major Professor: Jaxk Reeves

Committee: William McCormick
Ron Carroll

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Survival Function and Hazard Function

Survival analysis involves the modeling of "time-to-event" data (Collett, 2003). An event can be death of patients, heart attacks, failure of system, etc. Statisticians have expressed much interest in the distribution of "time-to-event". There are several equivalent definitions of "time-to- event". Two definitions below are commonly used to specify the survival distribution. The first or most traditional method is the survival function itself, S(t) =P {T>t}. Here "t" is the generic time, "T" is a random variable denoting the time of the event, and "P" stands for probability . Thus, S(t) is the probability that a randomly selected individual will survive to time t or later (Cantor, 2003). A second, equivalent, method is to specify the hazard function (denoted by h(t)). The hazard function is defined to be the event rate between t and $t+\Delta t$, conditional on survival until time t. It is defined as $h(t) = \lim_{\Delta t \to 0} P\,[t \le T < t+\Delta t \,|\, T \ge t]$. It can easily be shown that if the hazard function is known at all times t, then S(t) can be created. One advantage of using the hazard function, h(t), rather than the survival function, S(t), to compare distributions is that many survival functions look very similar over large ranges (S(0)=0 for all survival functions and $S(t) \to 0$ for larger t for all distributions), so that direct comparison of S(t) is sometimes difficult. Comparison of hazard functions (or log-hazard functions) makes it more apparent as to what time periods are risker for different conditions. This is illustrated in Figures 1.1 and 1.2 (Chen and Wang, 2000), where the two survival functions (Figure 1.1) are very similar over time periods, but for which it can be seen (Figure 1.2) that subjects under treatment have a higher risk before time 0.6, where those under placebo have lower initial risk, but higher risk after time 0.6.

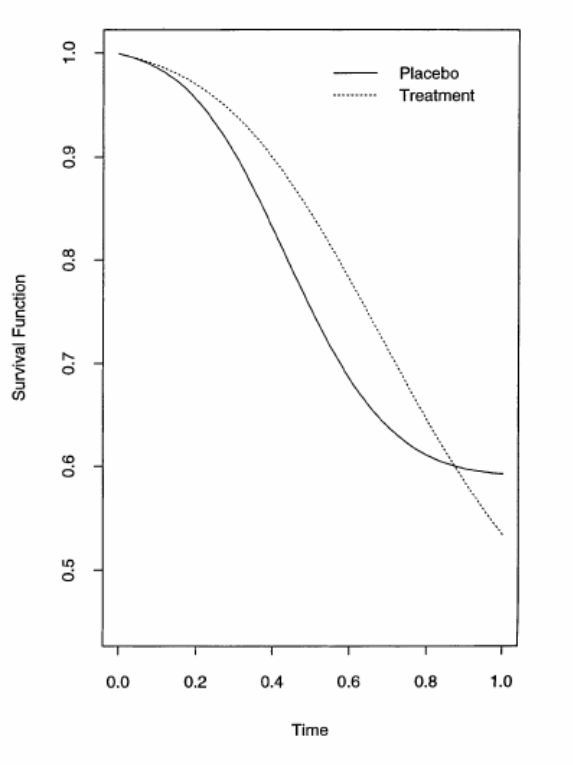

Figure 1.1 Survival Function

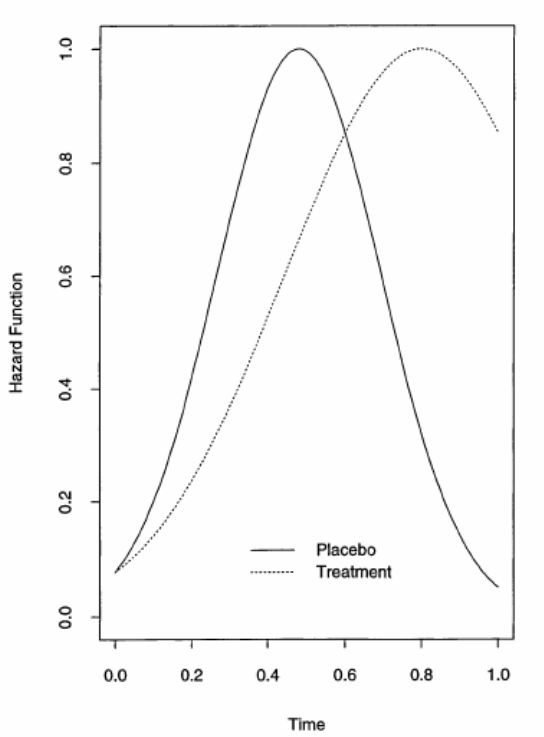

Figure 1.2 Hazard Function

1.2 Censoring

Censoring is a very common missing data problem in survival analysis. If the data are not censored, we know starting date and event date for all observations. When censoring occurs, we know the starting date, but we don't know when events occur. Censoring includes both right censoring and left censoring, but only right censoring is applicable to this research. If the true date of event is after some last observed date, this is called right censoring. For example, when the starting date is known but subjects are lost to follow-up or when a subject has not experienced any event by the conclusion of the study, right censoring occurs (Klein & Moeschberger, 2003). Some examples of right censoring are shown in Figure 1.3 and Table 1.1. In Figure 1.4, "Duration" represents the "time-to-event", while "IC" is the censoring indicator. Thus, those observations for which IC=0 (A, E) actually experienced the event of interest (symbol "X"). Those with IC=1 (B, C, D, F) are censored (symbol "O"). For Subjects B and D,

the censoring occurred because the trial ended after 12 months and subjects still had not experienced the event. For subjects C and E, the censoring occurred because the subjects dropped out after 6 and 9 months, respectively. During the observed months, the subjects didn't experience the event but we don't know what happened later. The key idea in survival analysis is to use all available information to estimate the survival distribution. Rather than discarding the censored observations because the actual event time is unknown, one uses any information which is available on survival during the period of observation.

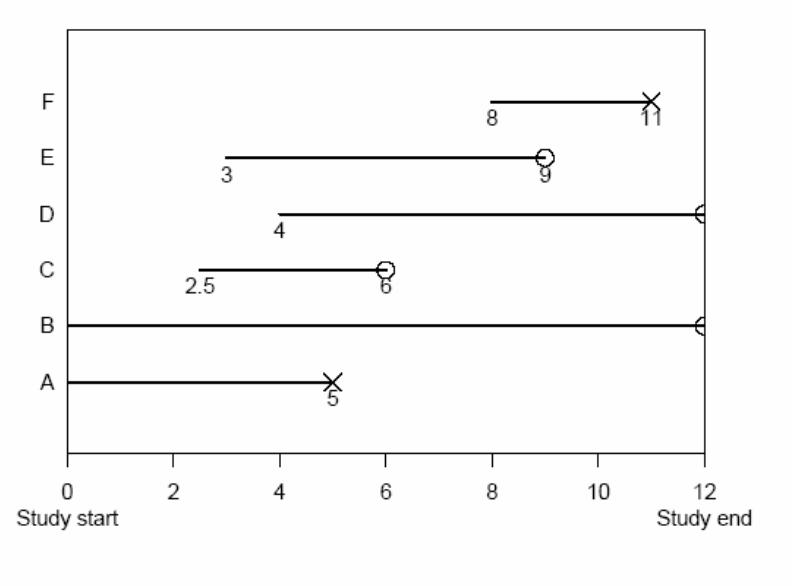

Figure 1.3 Right Censoring

Table 1.1 Right Censoring Table

Subject	Duration	IC
F	3	0
E	6	1
D	8	1
C	3.5	1
B	12	1
A	5	0

1.3 Model Types

There are two general kinds of models that can be used to model probability distributions for survival functions. One type is parametric survival models; the other is nonparametric survival models. Parametric models include exponential (a special case of Weibull), Weibull, log-normal, etc. Nonparametric survival models include Kaplan-Meier, and many proportional hazard models, etc. Different models may be appropriate in different conditions. For our purposes, proportional hazards models appear to be appropriate. These are actually semi-parametric models in that the baseline hazard function is nonparametric, but parametric modeling is uses to assess the effects of various other factors.

The Kaplan-Meier estimator uses the nonparametric maximum likelihood estimate method to calculate empirical survival curves. This method can take into account censored data. The formula is $S(t)= \prod_{t_i \prec t} \frac{n_i - d_i}{n_i}$. When there is no censoring, n_i is the number of survivors before time t_i. With censoring, n_i is the number of survivors before time t_i minus the number of censored cases. d_i is the number of subjects experiencing an event in the i th time period (Kaplan and Meier, 1958). Because censoring subjects reduces the sample size of subjects at risk, censoring affects the shape of the survival curve. The more subjects are censored the less reliable the survival curve is. Table 1.2 and Figure 1.4 demonstrate the steps needed to calculate the survival function using the Kaplan-Meier method for a particular data set. This simplified data set contains 6 survival times equal to 4, 5^+,6, 8^+, 9,12^+, where "+" represents censored survival times. One can note that this survival function never falls to zero. This occurs for the Kaplan-Meier estimator whenever any censored survival times are larger than the greatest observed event time.

Table 1.2 Examples of Kaplan-Meier Method Survival Function Calculations

interval (Start-end)	# of risk at start point	# of risk at end point	#censored data	# dead subjects	proportion survival	cumulative survival
[0—4)	6	6	0	0	6/6=1	1
[4—6)	6	6	0	1	5/6=0.83	1*0.83=0.83
[6—9)	5	4	1	1	3/4=0.75	0.83*0.75=0.623
[9—12)	3	2	1	1	1/2=0.5	0.623*0.5=0.31
[12- up)	2	2	1	0	2/2=1	0.31*1=0.31

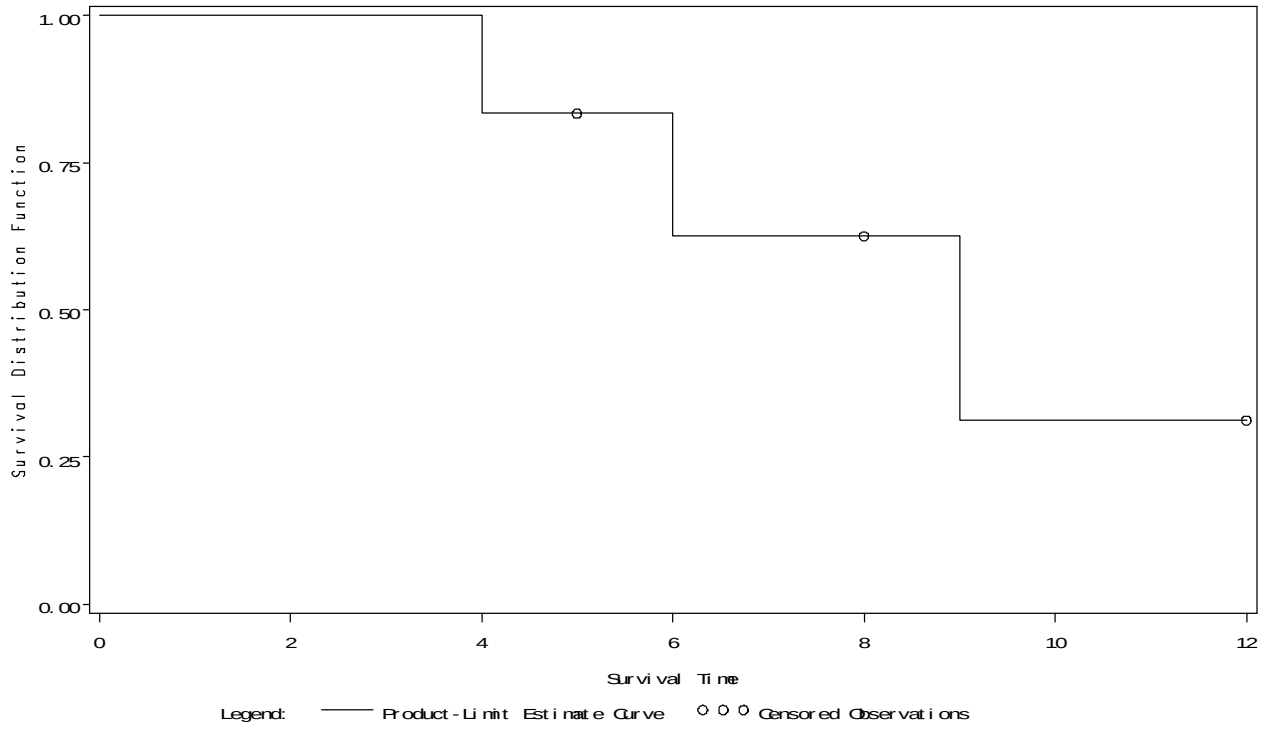

Figure 1.4 Plot of Kaplan-Meier Method Survival Function

The proportional hazards model is sometimes called a semi-parametric model. Proportional hazards models consists of two parts: the nonparametric part is the hazard function, which describe how hazard (risk) changes over time, and the parametric part contains the effect parameters, which describe how hazard relates to other factors - such as the choice of treatment, age, gender. It can model and test many inferences about survival without making any specific assumptions about the form of the life distribution model (Cox, 1972). The conditional hazard function, given the covariate vector X_i, is assumed to be of the form $h(t, X_i) = h_0(t) * \exp(X_i \beta)$, where, $h_0(t)$ denotes the baseline hazard function. No particular distributional form is assumed for the baseline hazard for $h_0(t)$; it is estimated nonparametrically. If $X_i = \mathbf{0}$ then the hazard function for the i th individual is the baseline hazard function. So at every value of t, the i th individual's hazard function is a constant proportion of the baseline hazard. The proportional hazards model has been widely used in survival analysis to estimate the effects of different covariates influencing the time to event data. This model puts minimal restrictions on the survival function itself, and allows one to more easily estimate the effect of treatment differences. Typically, one models the log-hazard function so that: $\ln(h(t, X_i)) = \ln(h_0(t)) + X_i \beta$ and uses regression techniques to estimate the β's ,with the log baseline hazard function playing the role of the intercept.

In some situations, we model a single type of time-to-event data (such as failure). But, in other cases, we are interested in multiple types of events. In our case, we are interested in risk of animal predation, flooding, and human intervention on turtle nests. Independent competing risks models provide a method to analyze multiple independent events. There are several steps to construct competing risk models. For the first step, we can estimate a multiple event survival model by estimating all single events separately, with a single model for each competing risk. In order to

estimate a given event hazard, we treat the subject as censored if it does not experience the event of interest (Satagopan et al 2004). For the second step, we must calculate the survival function by accounting for the presence of competing risks.

CHAPTER 2

DATA SET AND GENERAL PROBLEMS

2.1 Data Description

All data used in this thesis were collected from a conservation project conducted by Ph.D student Alison Lipman and her major professor Dr. Ron Carroll of the UGA Ecology department. I helped to analyze part of this data as my project in the statistics consulting class STAT 8000 in the Summer 2007, supervised by Dr. Jaxk Reeves. In this thesis, I will concentrate on a particular aspect of this data set, the survival distribution for nests of *P. unifilis* turtles.

Allison's conservation project focused on two declining species of South American River turtles: *P. unifilis* (yellow-spotted Amazon river turtle) and *P. expansa* (giant South American river turtle) in Noel Kempff Mercado National Park (a World Heritage Site in Bolivia) for 2005 and 2006. The primary concern of the project is whether humans' activities affect the decline of the population of the two species. Although the larger turtle (*P.expansa*) is of more ecological importance, there are relatively few nests of this type observations, so this thesis concentrates on the *P.unifilis* nests, for which more data is available. Human threat is one major factor for turtle population's decline. During turtle nesting season, female turtles and eggs are a reliable and readily available source of protein for local people. Local people prefer turtle meats over other meats because turtles are easy to hunt, transport and maintain. Researchers categorized beaches into different types such as *high human impact*, *low human impact* and *protected* beaches according to effect of human activities. Approximately 3 beaches of each impact status type were selected on two major rivers (Itenez and Paragua) in the region, and each selected beach was staffed by researchers. Allison hired and trained local residents as research assistants to collect

all project data. There are several stages to collecting data. In the initial stage, research assistants visited beaches every night during turtle nesting season (August through October). When they spotted a turtle, assistants waited to see if she nested. If she nested, they would wait until she covered her nest and catch her before she returned to the water. The following variables were then recorded for each mother turtle caught: location (beach, river), nest date and other response variables related to the mother. In the second stage, research assistants daily revisited all nests to see if any disturbances had occurred (Lipman, 2008). These disturbances could be due to several cases: Animal predation (A), Human poaching (H), Inundation by floods (I), or Natural hatching (N). Most often, on any particular day, there was no disturbance - the nest appeared just as it had when visited the previous day. However, if a disturbance was noted, the research assistant recorded the date of the disturbance and the type of event (A, H, I or N). Generally, once a nest event was recorded, it affected the entire nest, but there were a few cases where some eggs in the nests suffered another event, such as "I" followed two weeks later by "N". This occurred for only 161 out of the 1910 nests examined. When it did occur, we used the latter occurring event as the event of record. I analyzed much of the data collected in 2005-2006 as part of my STAT 8000 report. However, nest survival was examined only superficially in that report. In this thesis, I will focus on survival analysis of *P.Unifilis* turtle nests under different disturbance risks.

The main purpose of my thesis is to determine the risk that a *P.unifilis* nest will experience from the time it is created due to various threats. The hoped-for event is Natural hatching (N), but competing events are nest destruction by Animals (A), Humans (H), or Inundation (I). To perform survival analysis, we need data on three key response variables: nest date, disturbance type, and event date. From these we can calculate duration from duration=event date-nest date. For convenience, we have decided to use 31-July-05 and 31-July-06 as "day zero" for 2005 and

2006, respectively. So the variables 'nestdate' and 'eventdate' used in my analyses are actually the number of days after July 31st. This is a natural scale to use, since the earliest nesting dates for *P. unifilis* are in early August. The latest event date observed in this dataset is December 20 (day 142). This "zero" is for the "nestdate" or "eventdate" variables only – duration itself is measured in days from nest creation.

2.2 Summary of Data

Researchers observed and recorded a total 2086 *P. unifilis* female turtles, 1113 and 937 for 2005 and 2006, respectively. Only 1007 and 903 turtles nested , as show in Table 2.1 below. In my thesis, I would like to use the all 1910 nests in my analyses.

Table 2.1 Numbers of *P.Unifilis* Nests for 2005 and 2006

	2005	2006	
	P. unifilis	*P. unifilis*	Total
Nested	1007	903	1910
Dug or walked	106	70	176
Total	1113	973	2086

Because this conservation project is wild field research, there is much missing data. Some data are missing because the researcher couldn't locate the nests. Others are missing because the researcher wasn't able to record what happened to created nests. Table 2.2 shows how many nests have missing key responses and which key responses are missing. In that table, "Y" means that data are available, while "-" mean that data are missing. From Table 2.2, we see that although there were 1910 *P. unifilis* nests, only 1218 nests have all three response variables of primary interest (nest date, disturbance type, event date) recorded. This is only about 64% of the

total nests. If we omit the remaining 36% of the nests, we are losing much data and may have biased results. Thus, as explained in section 2.3, we will attempt an imputation to deal with cases where two of the three key variables are present and one is missing. These three patterns, representing 267, 93, 2 and nests, respectively, are noted by "*" in Table 2.2. If our imputation method is successful, this will allow us to use data from 1218+362=1580 nests, or about 83% of all nests.

Table 2.2 Categories of Missing Data Patterns for Nests

Nestdate	Disturbance type	Eventdate	Number of nests
Y	Y	Y	1218*
Y	Y	-	267*
-	Y	Y	93*
Y	-	Y	2*
Y	-	-	287
-	-	Y	1
-	Y	-	28
-	-	-	14
		Total	1910

Next, using the 1218 'complete information' nests, I compute the mean and standard deviation (SD) of the three numeric responses variable to provide some information. The mean and standard deviation of nestdate, eventdate and duration are computed for both year 2005 and 2006 and combined. From Table 2.3, Table 2.4 and Table 2.5, for the nestdate variable, in 2005, one observes that there is not much difference in means between the four disturbance classes. In 2006, Animal, Human and Natural hatching have similar means to each other (and to 2005), around day 45 (September 14th), but the inundated nests were built significantly earlier (day 32,

around September 1). This is a somewhat curious result, since there is no obvious reason why earlier built nests should have higher inundation risks than later built nests. For eventdate, there is tremendous variation between the four disturbance categories, with the pattern $\overline{Y}_H < \overline{Y}_A < \overline{Y}_I < \overline{Y}_N$ in both years, with observed sample means of 49, 65, 102 and 120, corresponding to September 18, October 4, November 10, and November 28, respectively. One also notes that the variation in eventdate for Animal is much larger than for Human and Inundation, which are not nearly as stable as for Natural hatching. That is, turtles tend to hatch within ± 2 weeks of November 28th, no matter when nests are created. The duration data follows from the nestdate and eventdate results, since duration=eventdate-nestdate. The most striking result is that Human poaching, if it occurs, will likely occur within the first two weeks after the nest is created. If the nest can survive two weeks, it is much less likely to suffer Human poaching than during the first two weeks. Animal predation risk is also much higher earlier than later, but not as extremely concentrated as Human poaching risk. Inundation and Natural hatching tend to occur at larger durations, although as previously noted, they seem to occur more because of "time of year" than duration. Since floods happen at particular season times, inundations tend to happen relatively later than human and animal destruction. If nothing happens to a nest, it can hatch successfully and this tends to occur about November 28 (day 120) in 2005 and 2006.

Table 2.3 Summary Statistics for Complete-Data Response Variables 2005

		Nest Date		Event Date		Duration	
Nests	Disturbance	Mean	SD	Mean	SD	Mean	SD
311	Animal	47.66	17.27	56.92	31.44	9.26	29.82
53	Human	40.68	20.69	44.89	23.36	4.21	13.84
22	Inundation	47.26	17.75	118.47	14.99	71.22	21.13
151	Natural hatching	43.60	14.83	120.18	11.34	76.59	15.64
537	Total	45.79	17.15	75.47	37.63	29.68	37.55

Table 2.4 Summary Statistics for Complete-Data Response Variables 2006

		Nest Date		Event Date		Duration	
Nests	Disturbance	Mean	SD	Mean	SD	Mean	SD
198	Animal	46.92	12.76	78.67	27.62	31.74	25.54
70	Human	43.57	13.16	52.96	18.10	9.39	12.79
91	Inundation	32.13	9.98	98.24	14.44	66.09	14.45
322	Natural hatching	44.43	11.37	119.86	7.31	75.43	7.23
681	Total	43.43	12.64	98.12	29.17	54.69	29.05

Table 2.5 Summary Statistics for Complete-Data Response Variables Combined 2005-2006

		Nest Date		Event Date		Duration	
Nests	Disturbance	Mean	SD	Mean	SD	Mean	SD
509	Animal	47.37	15.66	65.38	31.81	17.29	26.06
123	Human	42.33	16.80	49.48	20.93	7.15	13.44
113	Inundation	35.20	13.28	102.32	16.63	67.12	16.04
473	Natural hatching	44.16	12.56	119.97	8.79	75.80	10.66
1218	Total	44.47	14.84	88.13	35.01	43.67	35.31

Histograms of nestdate, eventdate and duration can also give us some information about the distribution of these variables. From Figure 2.1, we find that nestdate is distributed relatively evenly through nesting seasons. The histograms of eventdate (Figure 2.2) and duration (Figure 2.3) confirm the fact that Animal and Human disturbances tend to happen very early compared to Inundation and Natural hatching type, with Human poaching occurring, on average, even earlier than Animal predation.

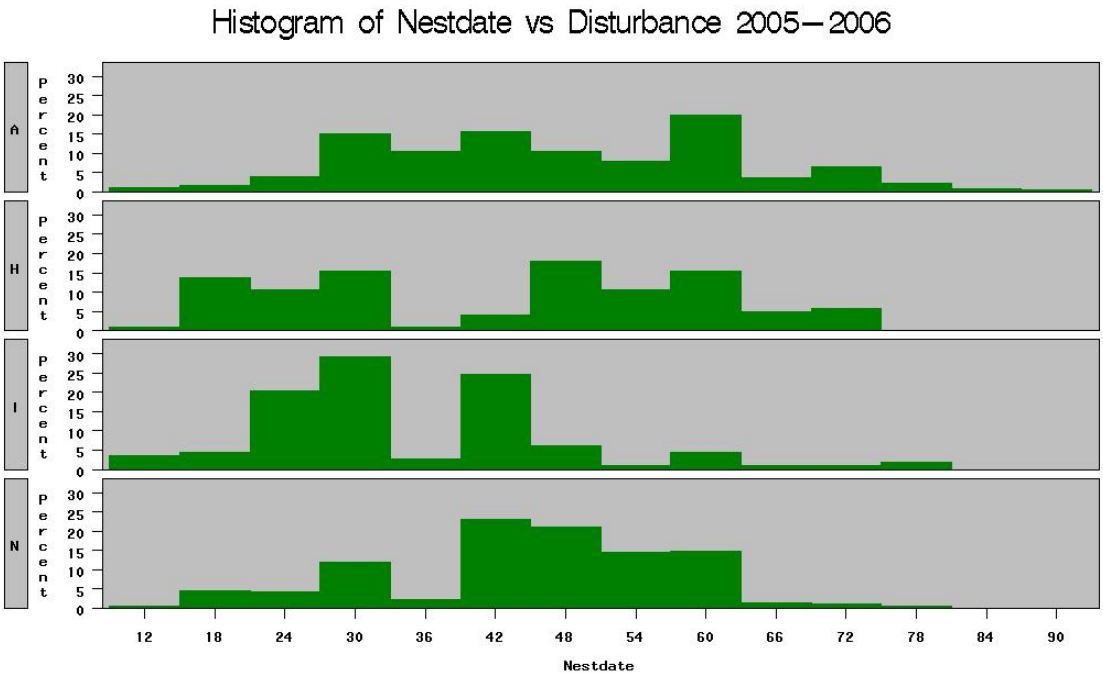

Figure 2.1 Histogram of Nestdate vs Disturbance Types 2005-2006

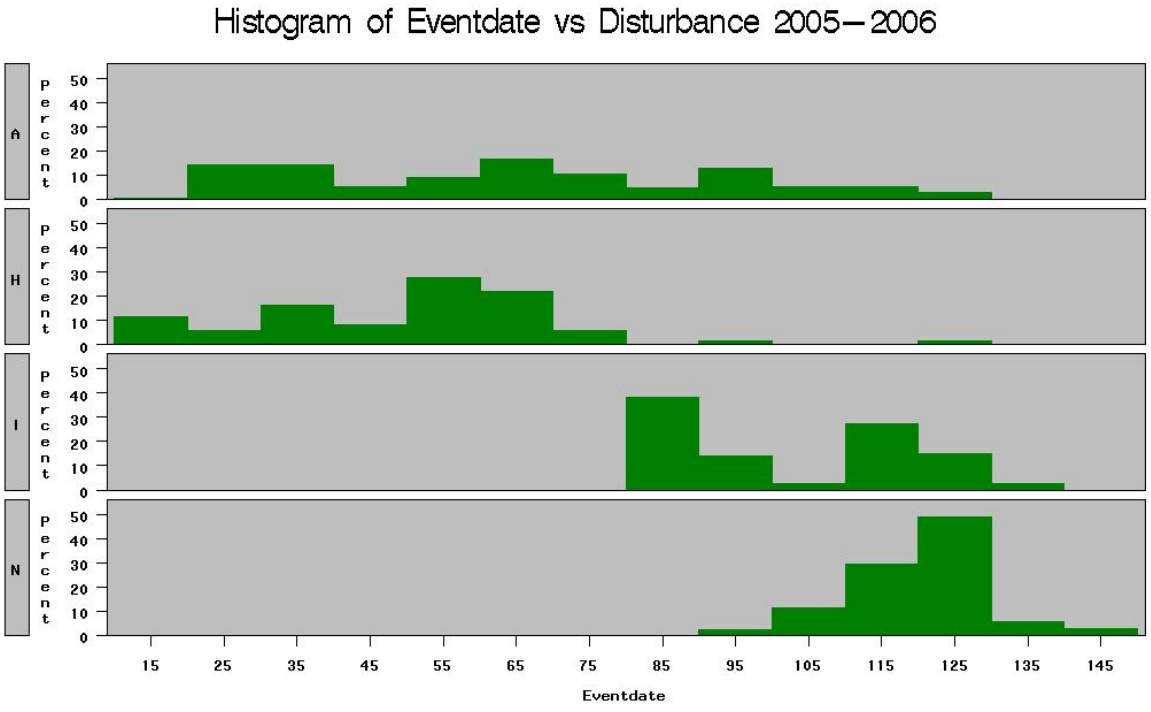

Figure 2.2 Histogram of Eventdate vs Disturbance Types 2005-2006

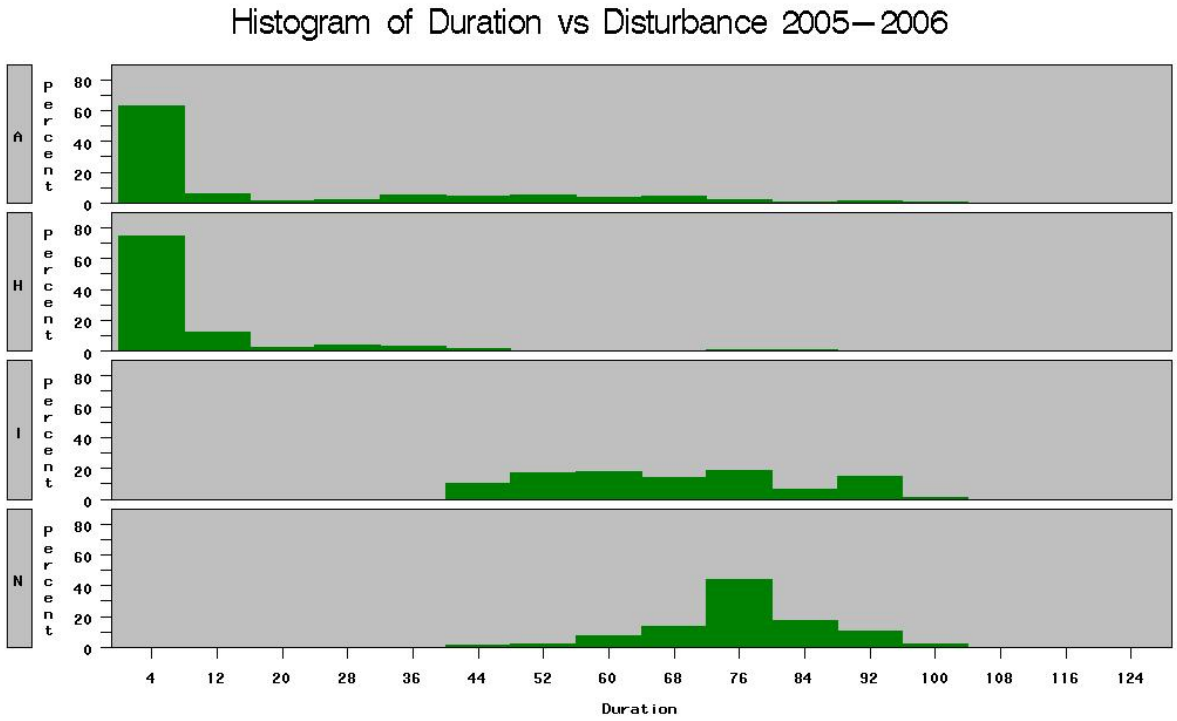

Figure 2.3 Histogram of Duration vs Disturbance Types 2005-2006

Scatterplots of the data for 2005 and 2006 are presented below (Figure 2.4 and Figure 2.5).We plot the 1218 complete response variable data for both 2005 and 2006 in separate plots. The event types are color coded: Animal=Green, Human=Blue, Inundation=Black, and Natural hatching=Pink. The X-axis is nestdate and the Y-axis is duration. The four diagonal lines in the scatter plot from lowest to highest represent event dates: September 1, October 1, November 1 and December 1 for each year. From Natural hatched nests, we observe something very interesting. Turtles tend to hatch at a particular time of the year, no matter when the eggs are laid. Baby turtles tend to wait for rain, and then all come out at the same time, around November 28 ± 2 weeks. This is quite different from humans, where the natural gestation period is about 280 days, no matter what time of year a baby is conceived. Of course, the duration results here are a bit skewed in that "nest date" is not conception date. In actuality, all *P.unifilis* eggs are

fertilized around the same time of year and have about the same time from conception until hatching. The variability in duration noted above is due to the fact that most eggs will hatch around the same time (November 28) no matter whether the mother turtle deposits them on the beaches early (August) or later (October).

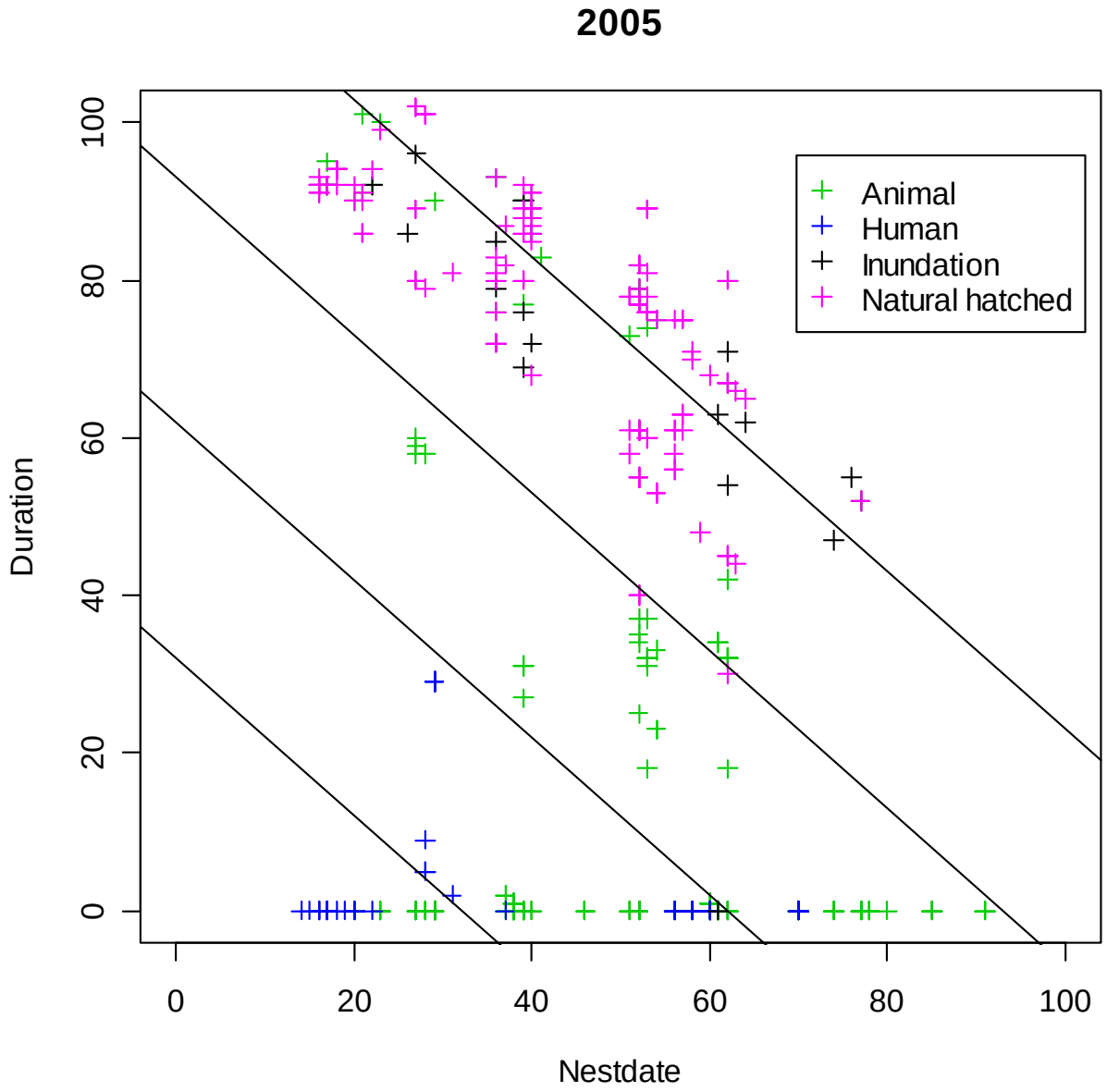

Figure 2.4 Scatterplot of Durations vs Nestdate for 2005

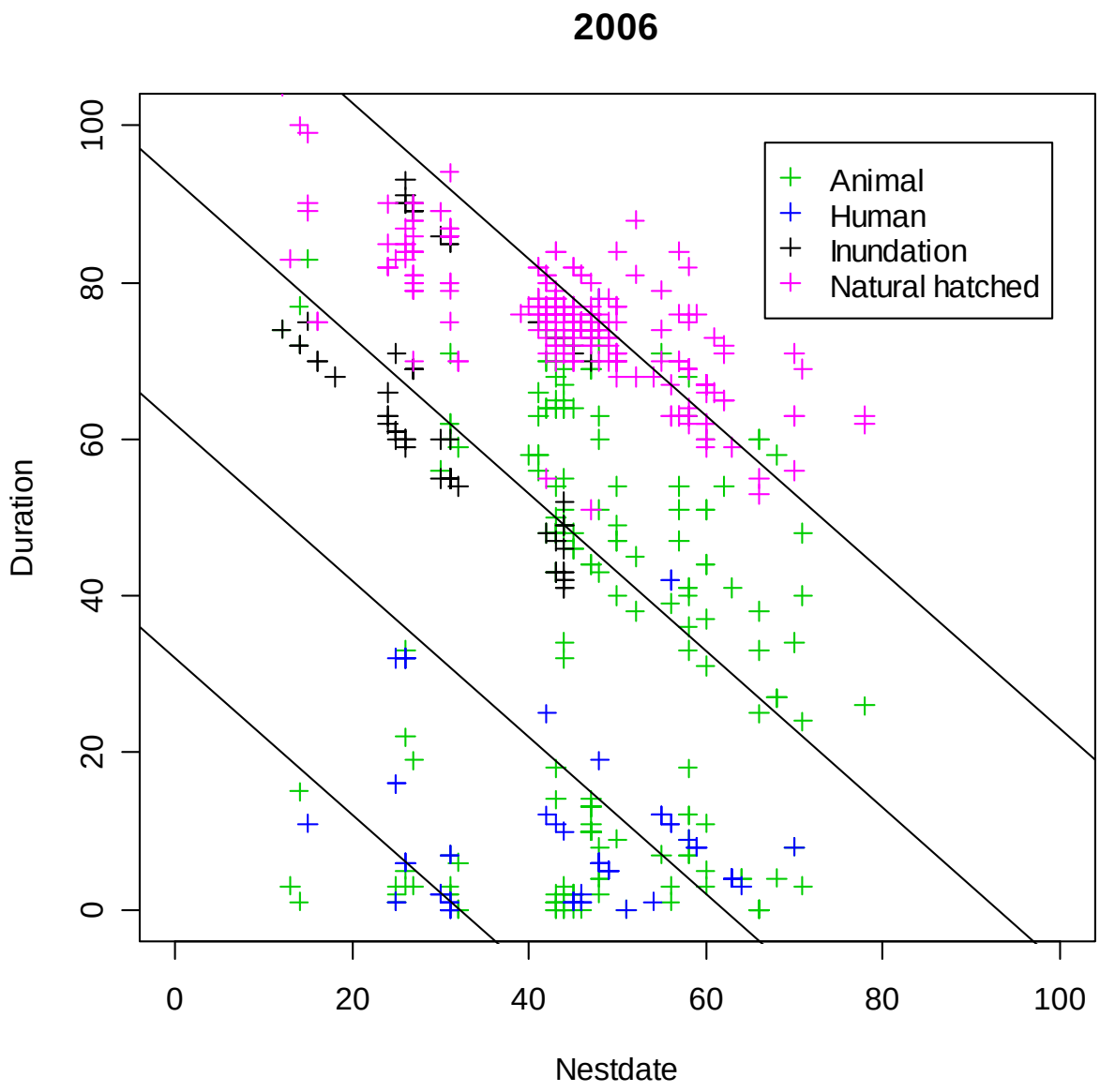

Figure 2.5 Scatterplot for Duration vs Nestdate 2006

2.3 Imputation of Missing Data

From Table 2.2, we see that 1218 out of 1910 nests have all three response variables of primary interest (nest date, disturbance type, event date) recorded. This is only about 64% of the total nests. If we omit the remaining 36% of nests, we are losing a lot of data and may have biased results. Thus, we will attempt an imputation to deal with cases where two of the three key variables are present and one is missing. If our imputation method is successful, this will allow us to use 1218+362=1580 nests, or about 83% of all nests. There is a subset of 267 nests for which we know only nestdate and disturbance type. In order to use these observations in an analysis, we must construct a model to estimate duration from these two available response variables. For another subset of 93 nests, we must use disturbance type and eventdate to estimate duration. Because the observed duration data are sometimes skewed, we should find an appropriate transformation to make these data approximately normally distributed. According to different disturbance types, different transformation methods are used. From examining the plots in Figures 2.1 to 2.3, we see that the distributions of duration times for both Inundation and Natural hatching distributions are roughly mound-shaped, so there is no particular need for transformation. However, the distributions for both Animal predation and Human looting are extremely skewed with many zeroes and small durations, but with long right-hand tails. In both cases, it appears that the true distribution is a zero-inflated process, which could be modeled as a mixture of a point-mass at zero and a skewed distribution. So, in both cases, we use a logistic model to estimate the probability that a "0" was recorded, and we then model the non-zero portion of the data by applying a general linear model to the square-root of duration (Animal) and log(duration) (Human). The general form of the imputation models used when nestdate or eventdate were missing is summarized in Table 2.6. Table 2.7 shows the number of imputations

performed for each of the 8 cases shown in Tables 2.6. For the two cases where Disturbance type was missing, but nest date and event data were known, we inputed the missing event to be "N" based on the information available.

Table 2.6 Model Forms Used for Imputations

Disturbance	Using nestdate to estimate duration (n=267)	Using eventdate to estimate duration (n=93)
Animal	Zero inflated process and GLM on Sqrt(duration)	Zero inflated process and GLM on Sqrt(duration)
Human	Zero inflated process and GLM on log(duration)	Zero inflated process and GLM on log(duration)
Inundation	GLM on duration	GLM on duration
Natural Hatching	GLM on duration	GLM on duration

Table 2.7 Numbers of Imputations Performed by Disturbance and Missing Classes

Disturbance	Data having complete Response Variables	Eventdate to be estimated	Nestdate to be estimated
Animal	509	122	10
Human	123	18	1
Inundation	113	103	6
Natural Hatching	473	24	76
Total	1218	267	93

Imputation Models

For both animal and human distributions, I modeled the probability of a "0" by a logistic model using River, Human Impact, Year, and nestdate (or eventdate, as appropriate) as predictor variables. For Inundation, Natural hatching and the non-zero portion of the Animal and Human

data, I used general linear models applied to the appropriately transformed durations. In all cases, I began with a full model which included Year, nestdate (eventdate) and their interactions, River, and Human Impact (HI). I then used parsimonious procedures to reduce the complexity of the models so that only statistically significant (at α=0.05) parameters were retained. Table 2.8 displays the parameter estimates for the four logistic models used to predict the probability of a "0" duration for human and animal distributions. Tables 2.9 and 2.10 display the parameter estimates for the GLM models for imputation from nestdate or eventdate, respectively. In Tables 2.8-2.10, I used different initials to indicate different level of variables. For river, IRI represents river Itenez, with river Paragua used as the baseline. For human impact, IHL represents Lower human impact beaches, IHH represents High human impact beaches and Protected beaches are the baseline. For year, IY5 represents the year 2005 and the year 2006 is the baseline.

Table 2.8 Logistic Model Parameter Estimators for P(zero) for Animal and Human Risks

	Prodictor	Models
Animal	Nestdate	ln(P/Q)=-1.55+1.83*IY5+0.023*nestdate +1.44*IHH-0.67*IHL
Animal	Eventdate	ln(P/Q= 2.74+1.68*IY5-0.05*eventdate +1.14*IHH-0.40*IHL
Human	Nestdate	ln(P/Q)=-3.52+2.46*IY5+0.072*nestdate +1.31*IRI
Human	Eventdate	ln(P/Q)= 1.28+2.11*IY5 -0.036*eventdate

Table 2.9 GLM Model Parameter Estimates Using Nestdate

Disturbance	R^2	Typical RMSE	Response	intercept	Nestdate	IHH	IHL	IRI	IY5	
Animal	0.05	28.07	Sqrt(duration)	8.69	-0.083					(2005)
			Sqrt(duration)	5.20	0.002					(2006)
Human	0.47	6.13	Log (duration)	7.92	-0.065	-2.40		-2.12		
Inundation	0.80	67.72	Duration	112.63	-0.84	1.09	21.96	-29.42	15.42	
Natural Hatching	0.56	75.8	Duration	111.30	-0.78	-1.83	3.20	-2.63		(2005)
			Duration	97.39	-0.50	-1.83	3.20	-2.63		(2006)

Table 2-10 GLM Model Parameter Estimates Using Eventdate

Disturbance	R^2	Typical RMSE	Response	intercept	Eventtdate	IHH	IHL	IRI	IY5	
Animal	0.89	9.33	Sqrt(duration)	-2.24	0.092	-0.95	-0.64	0.98		
Human	0.44	6.31	Log (duration)	-0.26					0.93	
Inundation	0.53	67.72	Duration	90.47	-0.24			13.6		(2005)
			Duration	-29.84	0.85			13.6		(2006)
Natural Hatching	0.38	75.8	Duration	-9.89	0.69	-6.81	-6.24	15.14		(2005)
			Duration	49.02	0.19	-6.81	-6.24	15.14		(2006)

Results of Imputation

Using the models shown in Tables 2.9-2.10, we imputed durations for the 362 observations noted in Table 2.7. The equations themselves allow us to predict an expected value. We then added random terms from the appropriate binomial (logistic) or normal (GLM) model, back-transformed (if necessary) and rounded the results to the nearest integer. We checked to make sure the results obtained were possible (i.e. no negative duration or imputed nestdates before the

hatching season, etc). It should be noted that the results which we obtained are random, so that slightly different results could have been obtained if different random values had been generated by the imputation simulations. One could perform multiple imputations if one was concerned about this, but since we simply wished to augment our date set with reasonable values so that complete survival analysis could be completed, we have performed only what is known as a single imputation method. As a check, we calculated the summary statistics for nestdate, eventdate, and duration by disturbance type for 2005, 2006 and combined for the augmented data set, as shown in Tables 2.11, 2.12, and 2.13 respectively. These can be compared with the original (complete response variable data set, n=1218) summary for these variables shown in Tables 2.3-2.5, respectively, to see that the imputation seems reasonable. Other verifications can be seen from Figures 2.6 to 2.8, where the original data are shown in black and the imputed data are shown in red. The final result is that we now have an augmented data set of 1580 observations with 'complete' records with respect to nestdate, eventdate and disturbance type, and we use these to perform the survival analyses of Chapter 3.

Table 2.11 Summary Statistics for Augmented-Data Response Variables 2005

		Nestdate		Eventdate		Duration	
Nests	Disturbance	Mean	SD	Mean	SD	Mean	SD
430	Animal	47.52	17.00	54.95	24.27	7.43	20.4
72	Human	42.61	18.77	47.40	23.26	4.79	15.60
96	Inundation	50.28	14.18	122.06	9.86	72.03	13.47
174	Natural hatching	44.47	14.52	121.07	11.30	76.60	15.37
772	Total	46.77	16.45	77.81	38.66	31.04	37.47

Table 2.12 Summary Statistics for Augmented-Data Response Variables 2006

		Nest Date		Event Date		Duration	
Nests	Disturbance	Mean	SD	Mean	SD	Mean	SD
211	Animal	47.14	13.89	79.17	27.76	32.03	26.22
70	Human	43.57	13.16	52.96	18.10	9.39	12.79
126	Inundation	35.17	11.60	101.48	15.51	66.31	15.10
401	Natural hatching	45.37	13.20	120.17	8.39	74.80	8.11
808	Total	43.61	13.14	100.98	28.45	57.37	28.16

Table 2.13 Summary Statistics for Augmented-Data Response Variables 2005-2006

		Nest Date		Event Date		Duration	
Nests	Disturbance	Mean	SD	Mean	SD	Mean	SD
641	Animal	47.39	16.03	67.95	28.64	16.52	26.48
142	Human	43.08	16.19	53.84	25.47	6.37	12.74
222	Inundation	41.70	14.79	109.55	16.24	67.85	14.38
575	Natural hatching	45.10	13.60	120.44	9.36	75.34	10.83
1580	Total	45.15	14.93	89.66	35.75	44.50	35.56

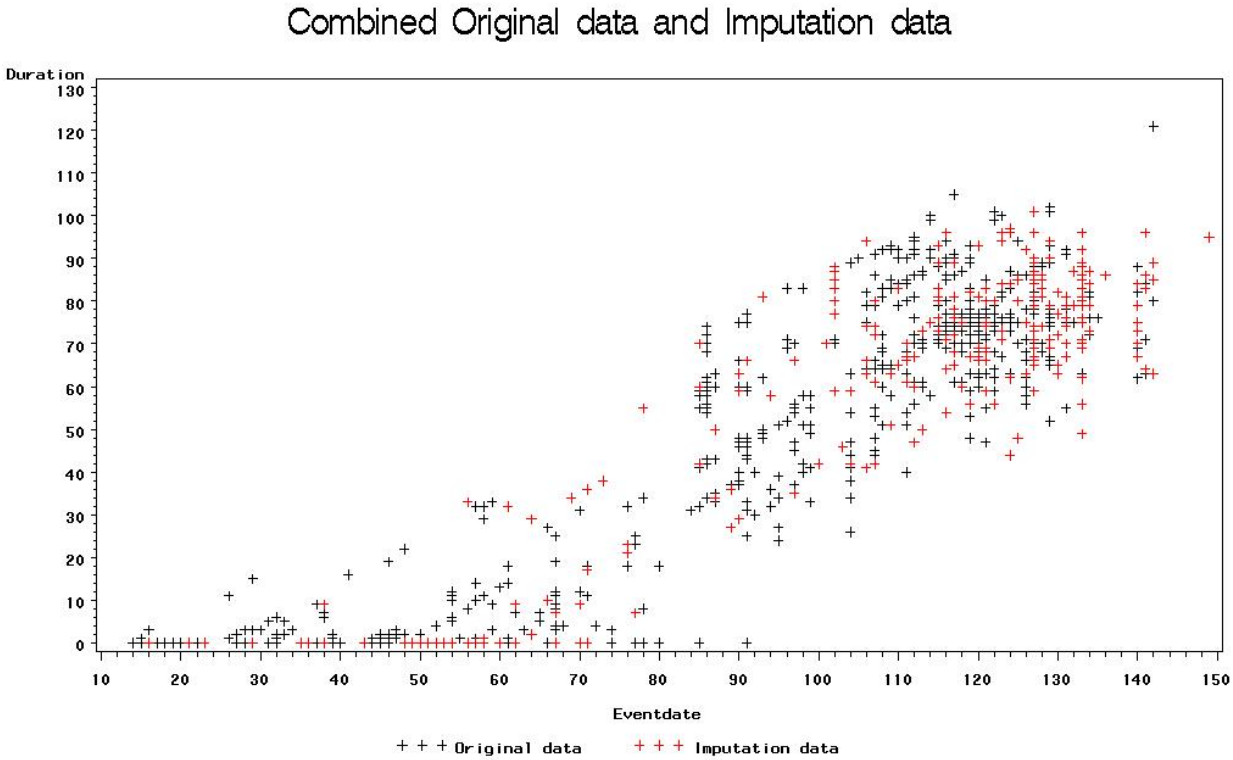

Figure 2.6 Plot of Duration vs Eventdate for Augmented Data

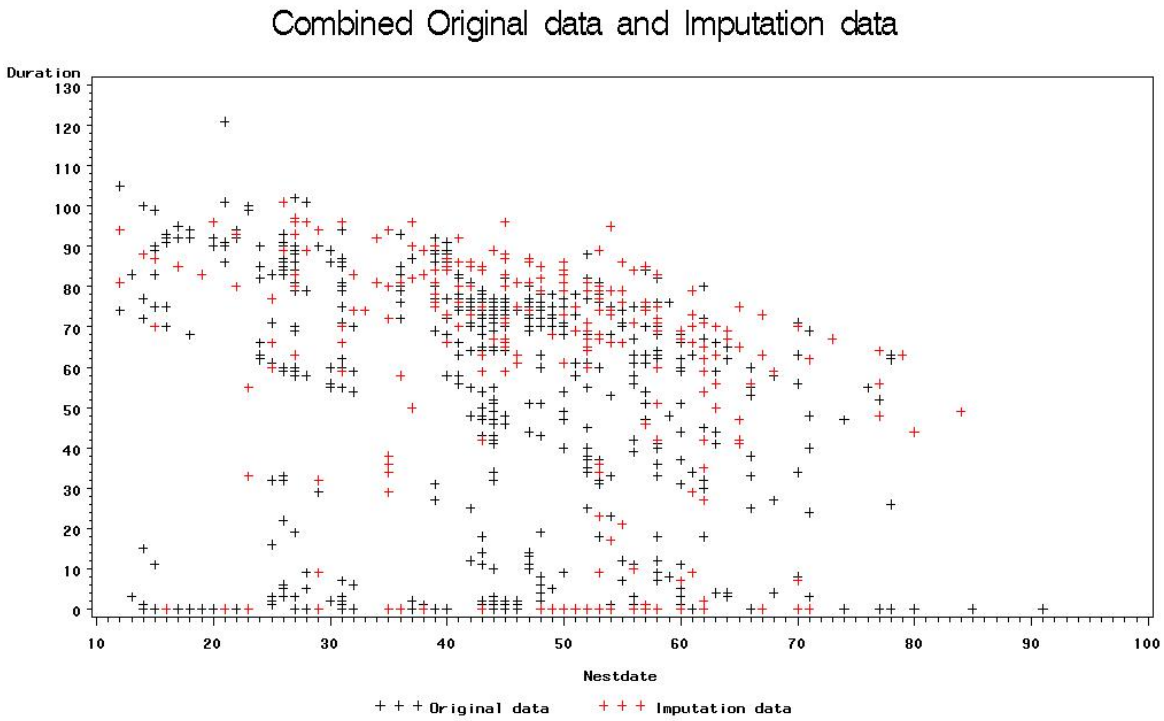

Figure 2.7 Plot of Duration vs Nestdate for Augmented Data

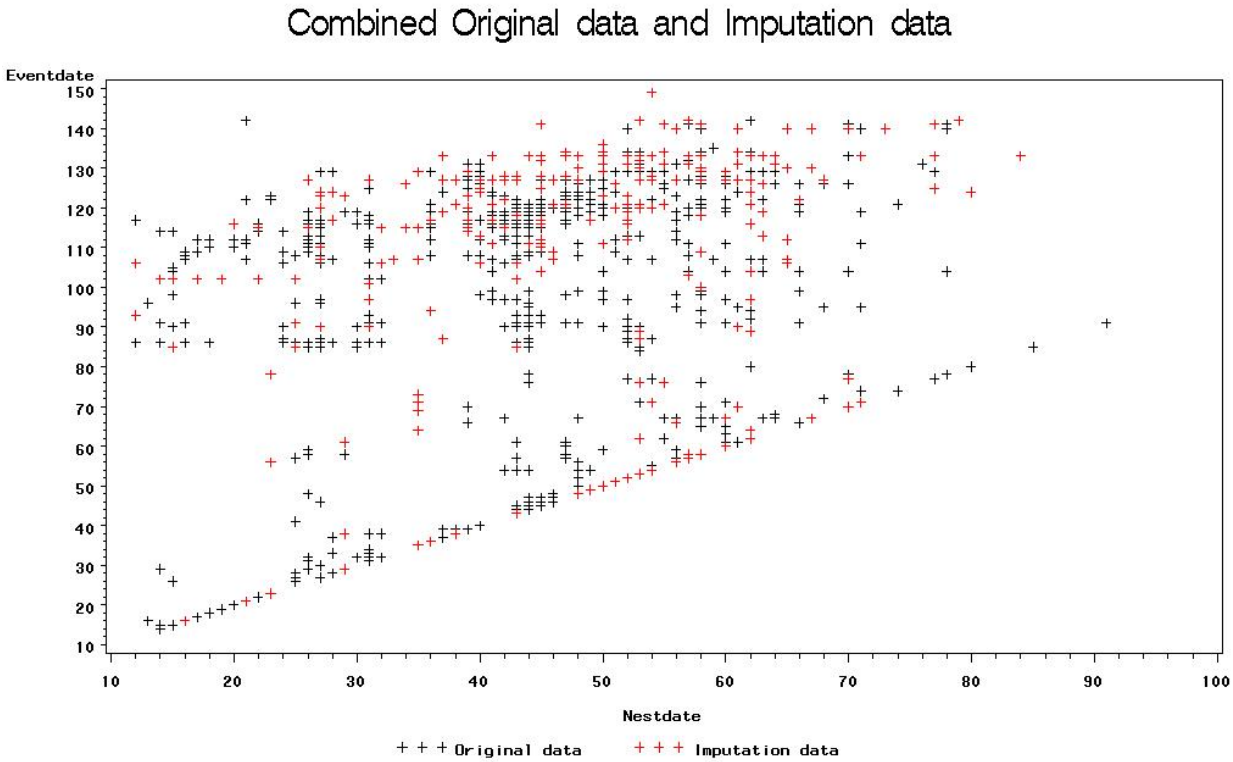

Figure 2.8 Plot of Eventdate vs Nestdate for Augmented Data

CHAPTER 3

ANALYSIS

3.1 Combined Events Analysis

For the 1580 nests for which complete or imputed complete results are available, we can create various empirical survival curves using the Kaplan-Meier procedure. The only difficulty is in deciding what an "event" is. If we use the most general definition, that any of the four outcomes (A, H, I or N) is an event, it is quite easy to calculate the Kaplan-Meier survival function, and, indeed, there is no censoring to deal with since all 1580 nests experienced exactly one of the events. This K-M survival function is shown in Figure 3.1. Note the very steep mortality near duration zero (entirely due to Animal predation and Human poaching), the relatively flat portion from duration 15 to duration 55 (when Animal predation is the most common risk), and the precipitous drop from duration=55 to duration=120, where both Inundation and Natural hatching tend to occur. Of course, one can't see separate risks from Figure 3.1, since all events are treated the same. We will return to this topic in section 3.2.

In addition to not controlling for the different types of risk, the Kaplan-Meier survival function is a function of duration only. However, we know that risk might depend on other factors. So, we considered a proportional hazards model for overall risk, where the class variables which we controlled for were River (Itenez, Paragua), Human Impact (High, Low, Protected), Year (2005, 2006), and the continuous covariate was *nestdate*. The results of this analysis, using indicators IRI (Itenez indicator), IHH (High Human Impact indicator), IHL (Low Human Impact indicator), and IY5 (year 2005 indicator) are shown in Table 3.1. From this output, we can see that the Itenez River is slightly, but significantly, less risky than the Paragua River, that there is no significant Year effect, that High Human Impact beaches are significantly

more risky than Low or Protected beaches (which are about equally risky), and that 'events' tend to happen more quickly as *nestdate* becomes later. All of these statements are true when one considers all types of nest disturbances to be 'events', but are actually quite misleading in the aggregate. More useful competing risk analyses are presented in the next section of this thesis.

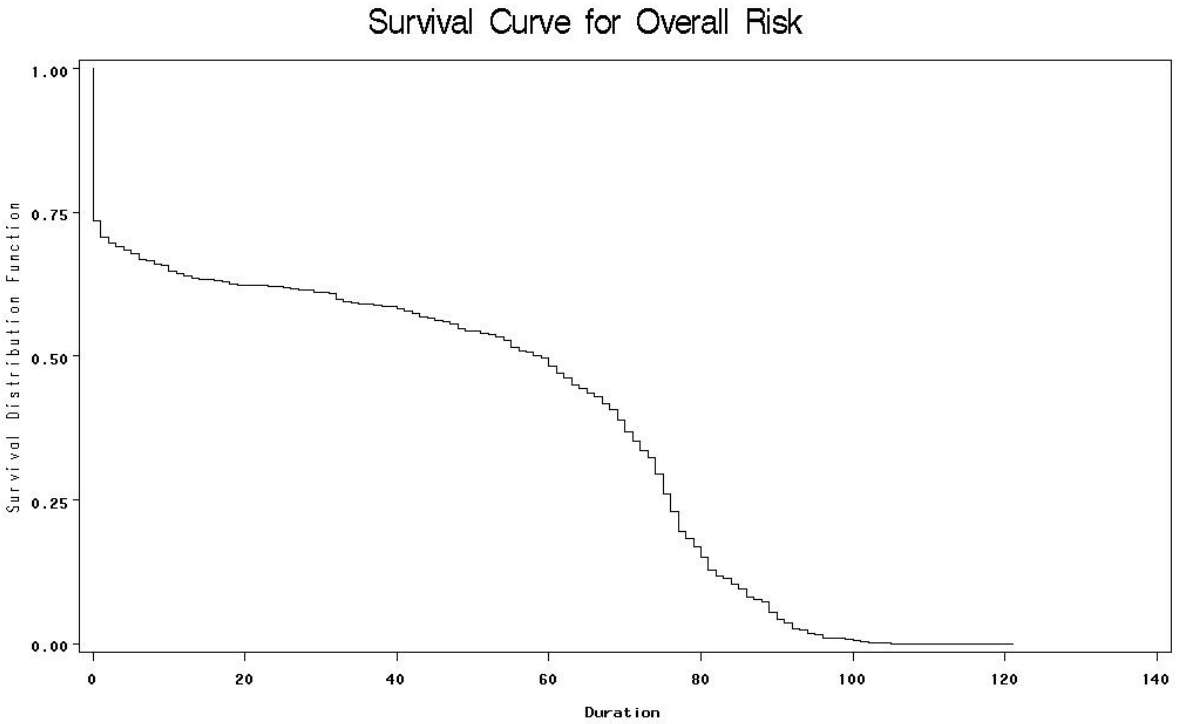

Figure 3.1 K-M Survival Curve for Combined Data

Table 3.1 Parameter Estimators from Proportional Hazards Model for Combined Events

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
IRI	1	-0.24933	0.07744	10.3669	0.0013	0.779
IHH	1	1.21357	0.06910	308.4110	<.0001	3.365
IHL	1	0.13286	0.07516	3.1250	0.0771	1.142
IY5	1	-0.03266	0.05359	0.3715	0.5422	0.968
nestdate	1	0.02543	0.00233	118.6404	<.0001	1.026

3.2 Separate Events Analyses

If one considered each event separately, regarding all other events as "censored", one will obtain the survival curves shown in Figures 3.2, 3.3, 3.4, and 3.5, for A, H, I, and N, respectively. Note that all of these except for N fail to fall to zero, since the largest event time is less than the largest censored time. These graphs have very different shapes from each other and from the combined survival function in Figure 3.1, since the risk as a function of duration is very different for these four types of events. Table 3.2 summarizes this disparity somewhat succinctly, where the table entries are the durations in days at which the indicated survival percentiles are achieved. Note that because of heavy relative censoring, neither the 'A' nor the 'H' survival function ever falls below 0.25; 'H' never even falls to 0.75.

Table 3.2 Comparison of Percentiles of K-M S(t) for Different Risks

Type	Events	Censored	S(t)=0.9	S(t)=0.75	S(t)=0.5	S(t)=0.25	S(t)=0.1
A	641	939	0	2	92	---	---
H	142	1438	78	---	---	---	---
I	222	1358	69	80	93	---	---
N	575	1005	66	74	77	87	96
All	1580	0	0	1	59	76	85

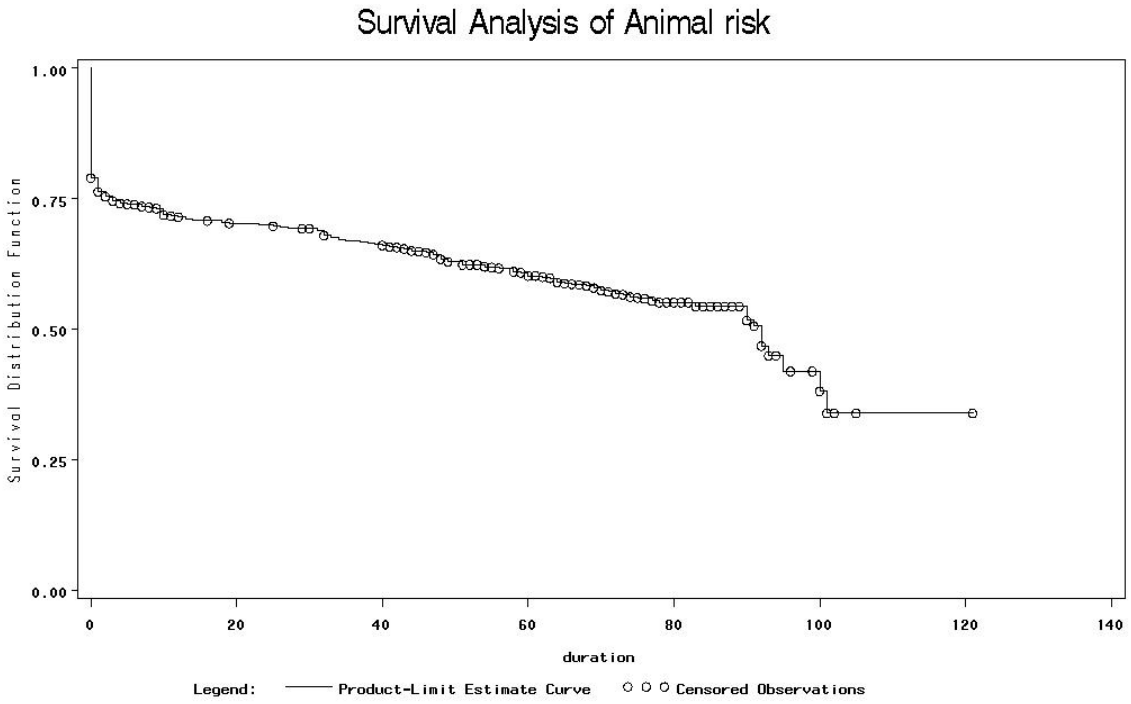

Figure 3.2 Survival Curve for Animal Risk

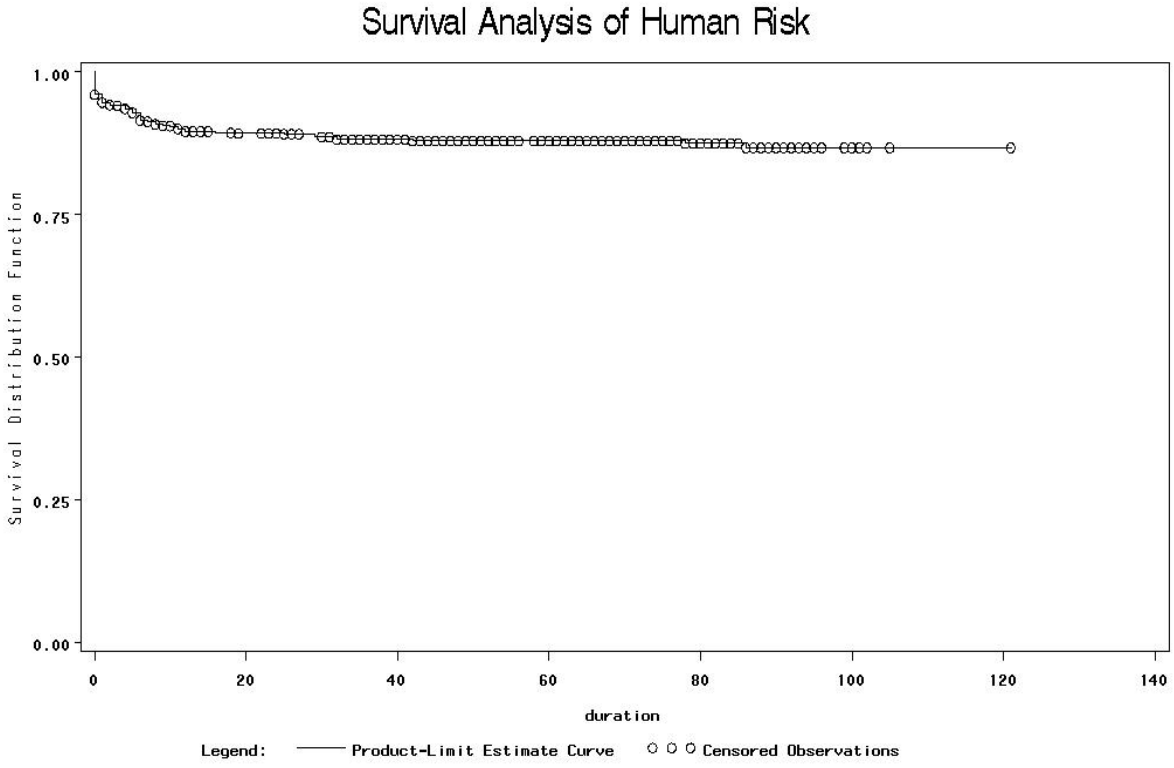

Figure 3.3 Survival Curve for Human Risk

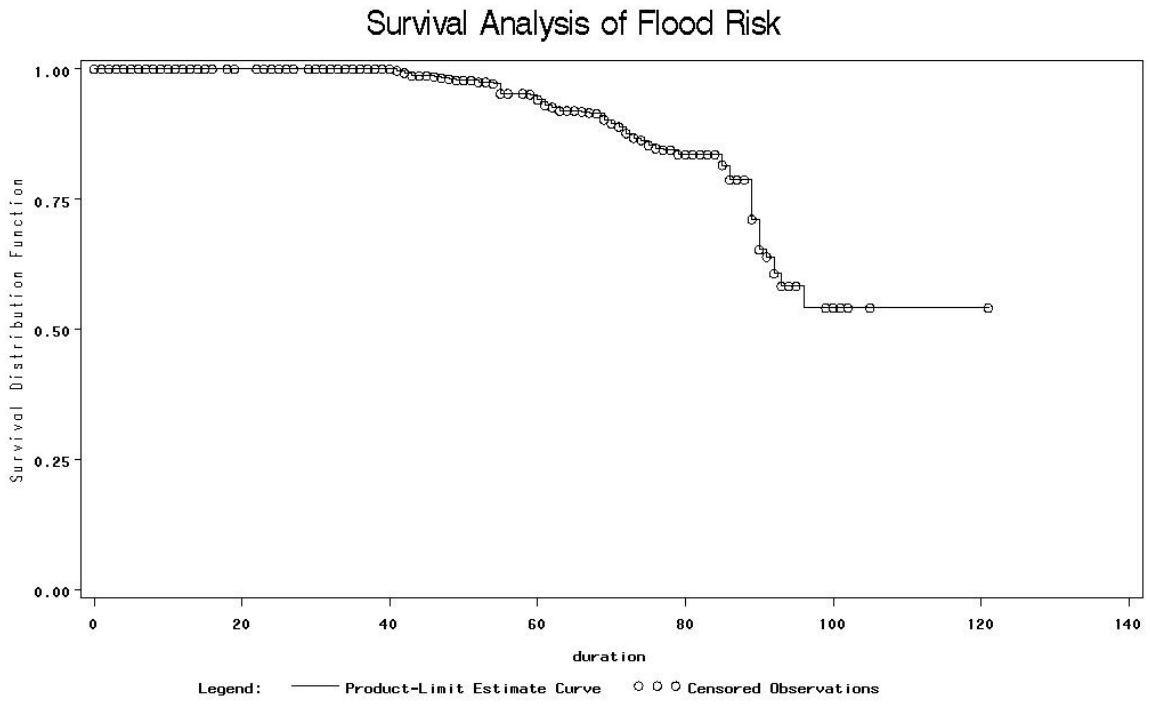

Figure 3.4 Survival Curve for Inundation risk

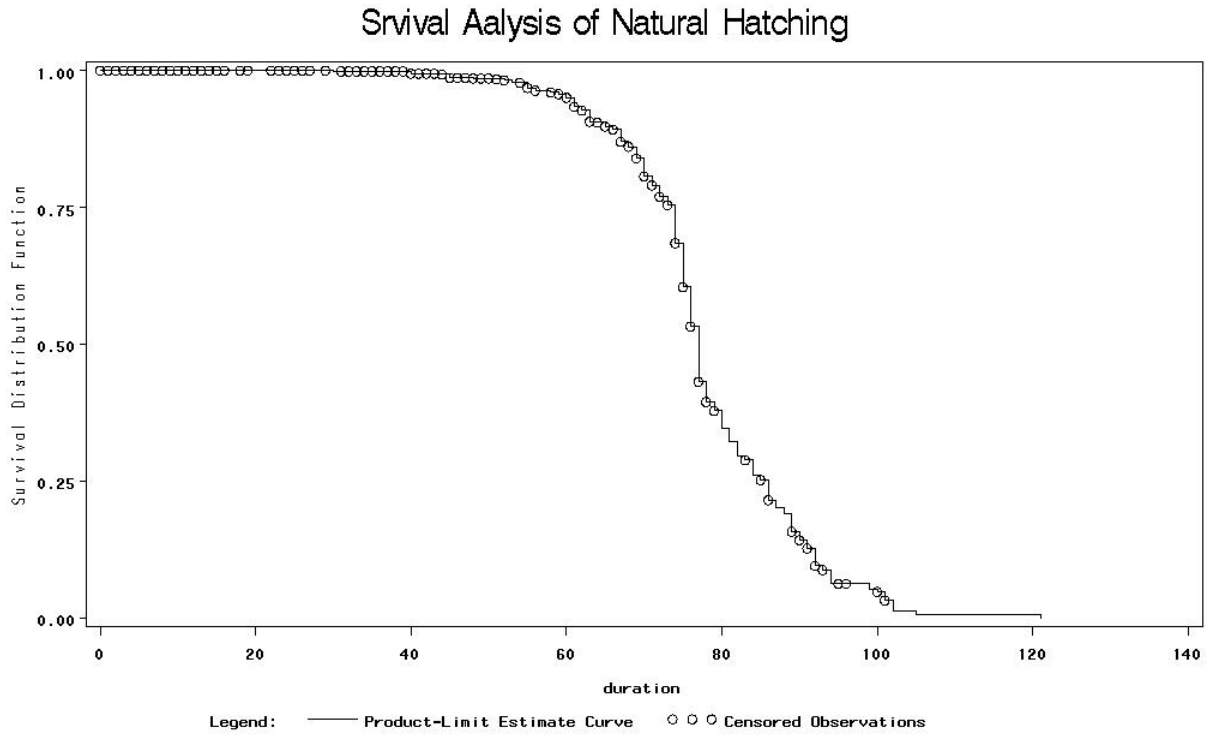

Figure 3.5 Survival Curve for Natural Hatching Risk

Figures 3.2 to 3.5 display the separate Kaplan-Merier survival functions for the four different disturbance types when each risk is considered separately. For these analyses, a nest which succumbed to a risk other than that of interest was considered censored, as exemplified by the survival function for (A) shown in Figure 3.2. Of course, what we really want is a combined S(t) displaying cumulative survival and failure due to various causes. This is displayed, in survival function form, in Figure 3.6. There, at any time *t*, the S(t) (black) curve displays the proportion of turtle nests which had not yet succumbed to any disturbance as of time *t*, while the Green, Yellow, Blue, and Red curves represent the cumulative proportion which had succumbed to A, H, I, and N, respectively. Of course, the black curve falls from 1 to 0 over the duration pictured, with the eventual height of the succumbing curves reflecting the overall probability of succumbing to each disturbance, agreeing with Table 3.2 (42% A, 39% N, 10% H, 9% I). This involves no particular statistical modeling. It is just a restatement of what is observed in the data, where duration and disturbance type are the only variables accounted for. In the next section, we will examine more sophistical competing risk models.

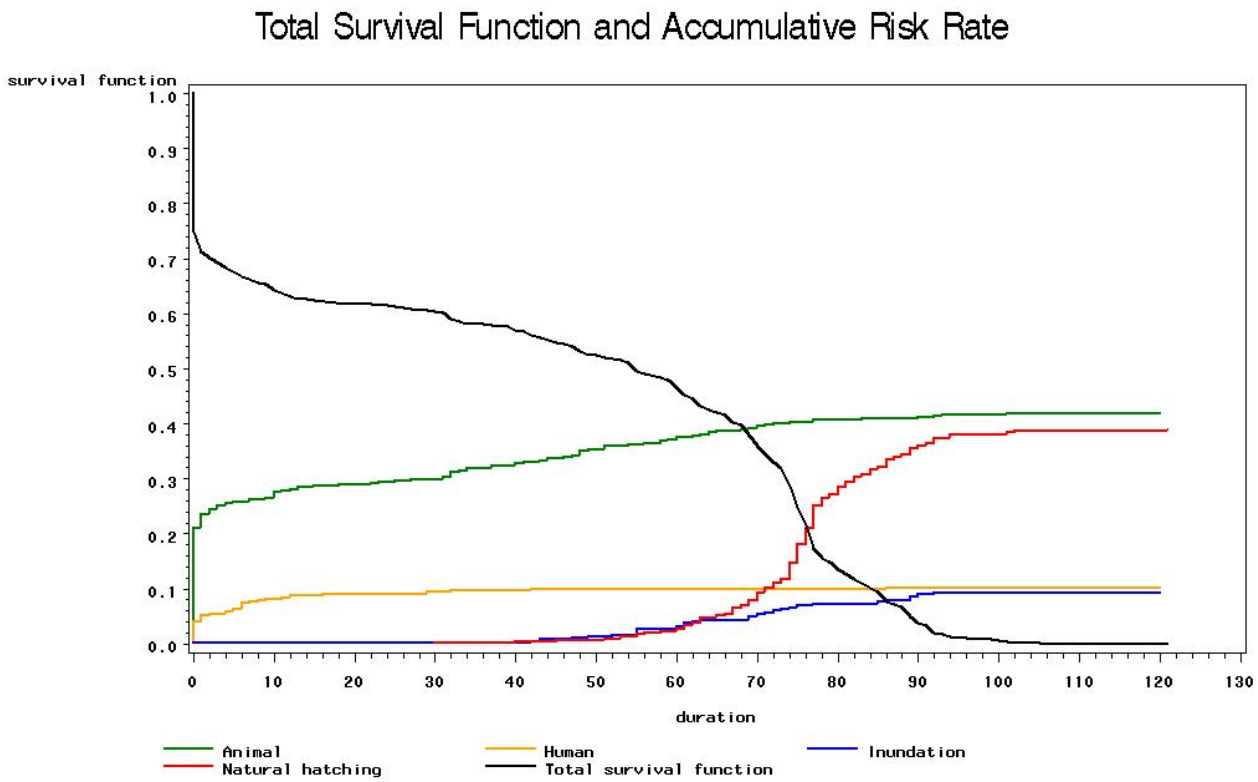

Figure 3.6 Total Survival Function and Accumulative Risk Rate

3.3 Competing Risk Models

The Kaplan-Meier method above provided us many useful statistical inferences. But, Kaplan-Meier is a descriptive procedure, which considers duration as the only salient variable. It doesn't evaluate the effects of covariates. If covariates are important in determining time to event, we must use other models such as the Cox proportional hazards model to analyze the data. Cox proportional hazards models allow us to include additional covariates (such as starting time, gender, age, etc). For this research, we wonder if River, Year, Human Impact and *nestdate* have any effect on nests' survival time for different events. I start with a full additive model including River (IRI represents Itenez; Paragua is the base line), Human Impact (IHL represents Lower human impact, IHH represents High human impact; Protected is the base line), Year (IY5

represents 2005; 2006 is the baseline), and *nestdate*. The parameter estimates for different models are presented in Table 3.3. In Table 3.3, "*" means the parameter is **not** significant and it can be dropped later.

Table 3.3 Full Additive Proportional Hazards Models for Separate and Combined Risks

	Nests	-2 log L	IRI	IHH	IHL	IY5	Nestdate
Animal	641	8666	-1.23	0.76	0.66	0.87	-0.0001*
Human	142	1562	0.66	5.99	1.14*	-0.19*	0.001*
Inundation	222	2550	2.19	0.16*	-1.23	-0.37	0.063
Natural	575	6119	0.56	0.19*	-0.44	-0.10	0.081
Total	1580	19763	-0.25	1.21	0.13	-0.03*	0.025

I dropped the insignificant parameters and built new models to fit the data. These final models are shown in Table 3.4.

Table 3.4 Final Proportional Hazards Model for Four Separate Risks

Risk	Final Model
Animal	$\log h(t) = \log h_{0A}(t) - 1.20*IRI + 0.76*IHH + 0.66*IHL + 0.87*IY5$
Human	$\log h(t) = \log h_{0H}(t) + 0.66*IRI + 5.25*IHH$
Inundation	$\log h(t) = \log h_{0I}(t) + 2.24*IRI - 1.29*IHL - 0.35*IY5 + 0.064*nestdate$
Natural hatching	$\log h(t) = \log h_{0N}(t) + 0.59*IRI - 0.48*IHL - 1.02*IY5 + 0.085*nestdate$

The final models tell us which covariates have significant effects on the survival function of time-to-event models. The high human impact beaches, not surprisingly, are very hazardous with respect to Human poaching. The protected beaches seem to guard against both Animal

predation and Human poaching, but have little effect on risks due to Inundation and Natural hatching. Inundation and Natural hatching risks are directly related to nestdate. The later the nests were built, the higher the chance the nests would experience Inundation and Natural hatching events. Inundation events are also highly related to river - the nests built in the Itenez river have a much higher chance to experience Inundation.

The models shown in Table 3.4 provide evidence of the relative effects of the covariates on different types of risks. However, much is masked in the $h_0(t)$ functions, which are not at all the same for the different risks. In Figure 3.7 is displayed the plot of log(-log(S(t))) vs log(duration) for the separate risks (A=Green, H=Blue, I=Red, N=Black) under the baseline condition (Paragua River, Protected beach, 2006), with nestdate set at its median value (day 45). This is the best graphical means by which to gauge the relative risks of the various events at different durations, although the log-scale for duration complicates comprehension. From this plot, we can see that Animal risk is much higher than Human risk by an almost constant amount (in log-scale) for all durations. Inundation and Natural Hatching risks are basically non-existent before log-duration=3.3 (which corresponds to 30 days after nesting, or day 75 for this baseline), at which time both Natural hatching and Inundation risks begin to increase sharply, overtaking Animal risks by duration 55 (day =99). Figures 3.8, 3.9, 3.10, and 3.11 show similar plots, where one baseline characteristic at a time is changed from those shown in Figure 3.7. In Figure 3.8, the river is changed from Paragua to Itenez, in Figure 3.9, beach is changed from Protected to High human impact, in Figure 3.10, year is changed from 2006 to 2005, and in Figure 3.11, nestdate is changed from 45 (September 14th) to 65 (October 4th).

In Figure 3.8, where River is changed from Paragua to Itenez, the plots are similar to those in 3.7, but now Inundation risk increases faster than Natural hatching risk over the region from

duration 33 to 55 and risk from Animals is decreased. Figure 3.9, for High Human Impact beaches, is strikingly different from the previous two figures in that Human risk is nearly the same as Animal risk during the early stages, although by duration 40, both Inundation and Natural Hatching pass these. From Figure 3.10, where Year is changed from 2006 to 2005, we observe a pattern similar to that observed in Figure 3.8, but the Animal risk is even worse elevated. Finally, in Figure 3.11, where nestdate is changed from 45 to 65, we see that there is no effect on Animal or Human risks (which are unaffected by nestdate), but that the risks for Inundation and Natural hatching become much larger than in the baseline graph of Figure 3.7. This makes sense, because if the nesting date becomes later, the time for flooding and Natural hatching becomes nearer, increasing their relative risks.

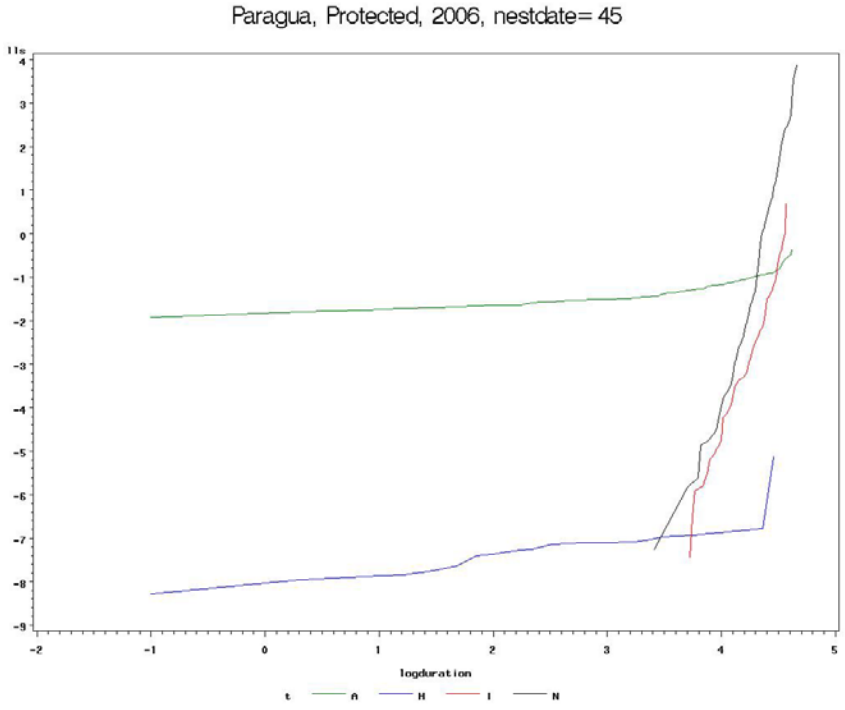

Figure 3.7 lls vs log(duration) for Beaches of Paragua, Protected, 2006

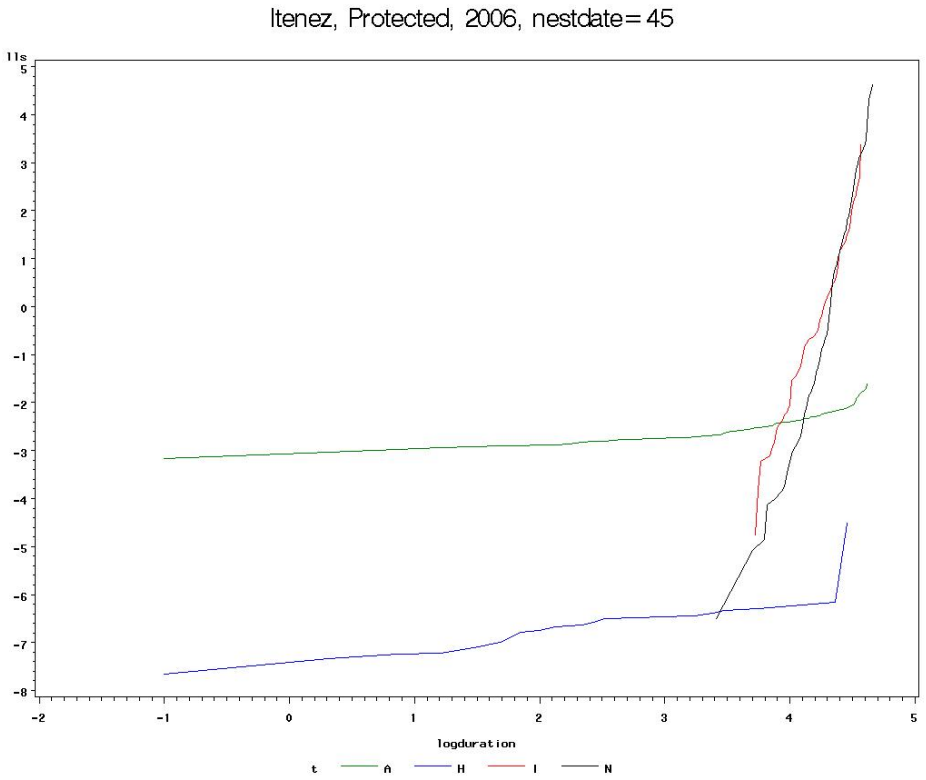

Figure 3.8 lls vs log(duration) for Beaches of Itenez, Protected, 2006

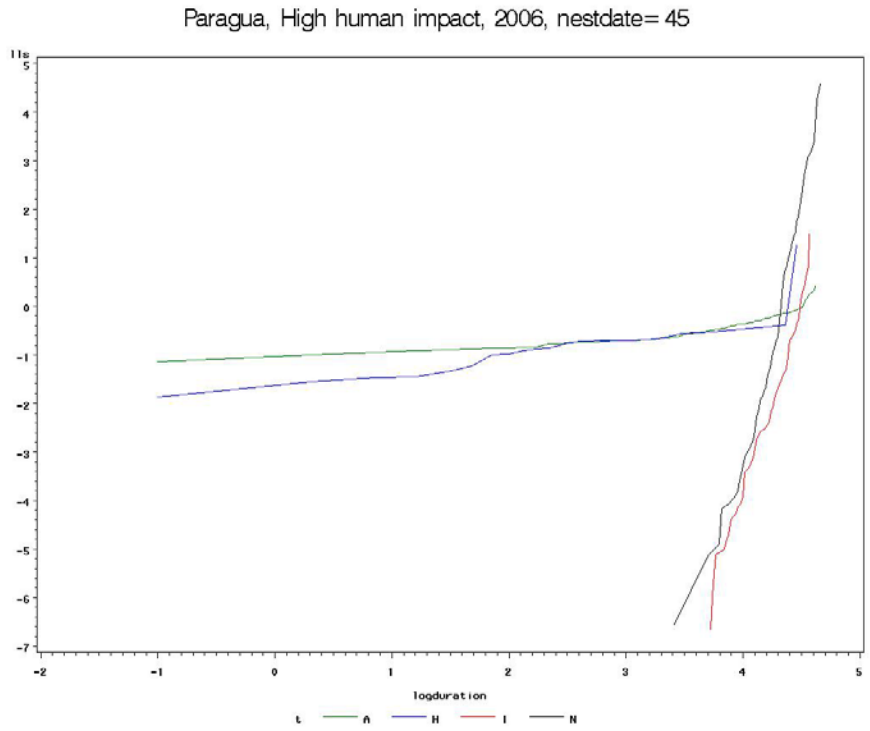

Figure 3.9 lls vs log(duration) for Beaches of Paragua, High human impact, 2006

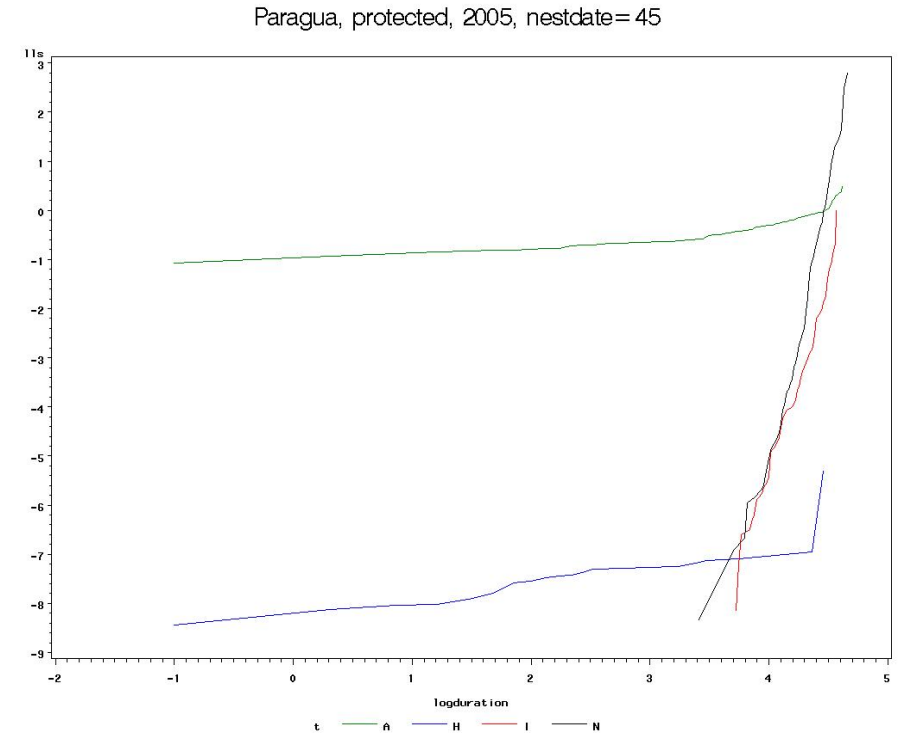

Figure 3.10 lls vs log(duration) for Beaches of Paragua, Protected, 2005

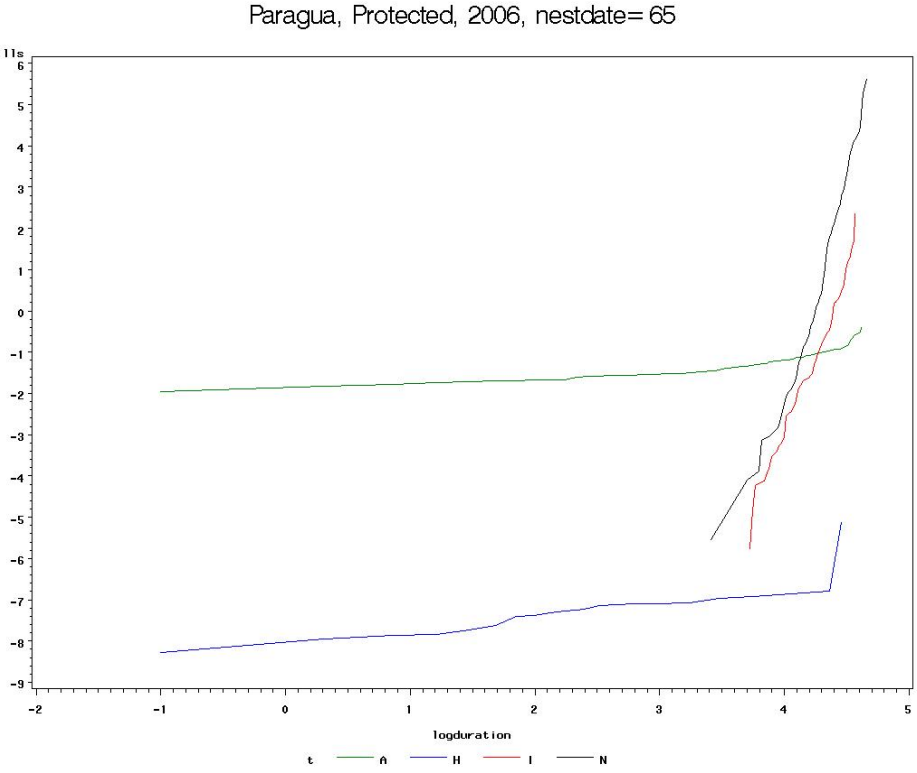

Figure 3.11 lls vs log(duration) for Baseline Beaches, Nestdate=65

CHAPTER 4

CONCLUSION

Survival analysis (also called time-to-event analysis) is a very useful technique to compare the risks for events. It uses tables and plots to explain survival functions for time-to-event data. In my thesis, the competing events are disturbances of nests by Animal, Human, Inundation and Natural hatching, respectively. Using simple survival analysis, I initially treated the four risks separately. Separate survival analyses provide insight into the shape of the survival function for each risk. Kaplan–Meier is a very simple but valuable method for estimating the survival curve. It is recommended that we look at the Kaplan-Meier curves for separate risks before we examine more complex models. Figures 3.2- 3.6 show plots of the cumulative survival and failure function. The X-axis is duration (time to effect); the Y-axis is cumulative survival or failure. The cumulative survival function at time *t* is the probability of survival to that time. These plots show that animal and human disturbance happen very early. There are about 10%, and 28% of nests disturbed by animal and human by the duration 10 days, respectively. From Figure 3.3, we see that less than 10% of all nests are disturbed by humans, and in addition, if a nest survives two weeks without being poached by humans, there is a 95% chance that it will not ever be poached by humans. On the other hand, nests disturbed by Inundation and Natural hatching tend to occur at later durations. None of the 1580 observed nests experienced a flood or natural hatching event before durations of 41 and 30 days, respectively, so the proportional hazards models can not assign any risk to this time period for these disturbance types.

A competing risks model is a model for multiple types of events for a given subject, where the subject is observed until it experiences the first event. We are interested in the probability of experiencing an event by a given time considering different covariates. In this thesis, we are

interested in Animal predation, Human poaching, Inundation, and Natural Hatching risks controlling for River, Human Impact status,Year, and nestdate covariates. We found that Animal risk is always higher than Human risk, except at the Higher Human Impact beaches, for which the Human risk is almost the same as the Animal risk. Animal risk and Human risk happen very early, starting from duration 0. On the other hand, Inundation risk and Natural Hatching risk happen significantly later; there is no Inundation risk or Natural Hatching risk before duration 30 days. However, as shown in section 3.3, after duration 40 days or so, the Inundation and Natural Hatching risk increase sharply, overtaking Animal risks by duration 55 days. One overall conclusion is: once a nest is created, it has higher chance to be disturbed by animals than humans at the early stage (before duration 30 days), except at High Human Impact beaches. After duration 40 days, the Inundation and Natural Hatching risks increase sharply, surpassing Animal risk by duration 55 days, with Natural Hatching becoming increasingly more likely as duration increases. Of the major predictor class variables examined (River, Year, Human Impact status), year had very little effect, river had some effect on Inundation risk (with Itenez being riskier than Paragua), while higher level of Human Impact status seriously increases the risk of Human poaching. The nestdate covariate had a small positive effect in increasing the risk of Inundation and Natural Hatching.

REFEREENCES

1. Cantor A . SAS survival analysis techniques for medical research. Cary, NC, USA .2003.

2. Chen YQ, Wang MC (2000). Estimating a treatment effect with the accelerated hazard models. Controlled clinical trials 21:369-380.

3. Collett D. Modeling Survival Data in Medical Research. Boca Raton: Chapman & Hall\CRC. 2003.

4. Cox DR (1972). Regression models and life tables. Journal of the Royal Statistical society Series B 34:187-220.

5. Kaplan E.L, Meier P (1958). Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53, 457-481.

6. Klein J, Moeschberger M .Survival analysis: techniques for censored and truncated data. New York : Springer, 2003.

7. Lipman A (2008). Dissertation, Chapter 4, 8-10.

8. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D and Uerbach AD (2004). A note on competing risks in survival data analysis. British Journal of Cancer 91:1229 – 1235.