Detection probability estimates from the 2011 ice-based independent observer survey of bowhead whales near Barrow, Alaska

G. H. Givens[*], S. L. Edmondson[*], J. C. George[†],
B. Tudor[†], R.A. DeLong[‡]R. Suydam[†]

June 9, 2012

Abstract

We estimate visual detection probabilities from the spring 2011 ice-based survey of bowhead whales near Barrow, Alaska. This estimate is intended for use in a future whale abundance estimate and to replace the older detection probability estimates summarized by Zeh and Punt (2005). Standard capture-recapture methods are not directly applicable to the 2011 survey data because of sighting ambiguities, match uncertainty and group size inconsistencies introduced by the data collection protocol. Several bias correction methods are essential to produce accurate estimates. After incorporating these corrections we can apply the general framework of Huggins (1989) using a a weighted data analysis approach. This model estimates the dependence of detection probabilities on covariates associated with the sighting. For our recommended analysis option, the mean estimated detection probability is 0.495, with specific estimates depending strongly on sighting distance and whale group size. Most standard errors are less than 0.03.

1 Introduction

In April and May of 2009, 2010, and 2011, scientists from the North Slope Borough Department of Wildlife Management attempted ice-based counts of bowhead whales (Balaena mysticetus) from the Bering-Chukchi-Beaufort Seas population as the animals migrate northward past Barrow, Alaska. Descriptions of the surveys are given by George et al. (2011; 2012).

These surveys employ a two-perch independent observer protocol. Specifically, two teams of observers stand at fixed survey sites (perches) situated atop pressure ridges near leads and open water. The two perches are sufficiently distant that the teams cannot hear each other or incidentally cue each other about their sightings. All sightings and other data are recorded

[*] Department of Statistics, 1877 Campus Delivery, Colorado State University, Fort Collins, CO 80523 U.S.A. geof@lamar.colostate.edu. This work was supported by the North Slope Borough (Alaska) and the National Oceanic and Atmospheric Administration (through the Alaska Eskimo Whaling Commission).

[†] North Slope Borough, Department of Wildlife Management, Barrow AK, U.S.A.

[‡] DeLongView Enterprises, Box 85044, Fairbanks, AK 99708, U.S.A.

independently at each perch. During some periods (independent observer, or 'IO' periods), both perches were operating. At other times, only one perch, or neither, was used. Typical IO shift periods were 4 hours.

The ultimate purpose of the surveys is to obtain a population abundance estimate. Most previous abundance estimates have been based on ice-based surveys (Braham et al., 1979; Zeh et al., 1986, 1991; Raftery and Zeh, 1998; George et al., 1994; Zeh et al., 2005) The most recent abundance estimate, derived from an aerial photo-identification survey, is 12,631 (95% confidence interval 7,900 to 19,700) by (Koski et al., 2010). The key requirements for estimating abundance from the ice-based survey data are the count of whales seen, the proportion of the population available to be seen as whales migrate past the perches, and an estimate of the detection probability, i.e., the probability of detecting a whale given that it passes the perch within viewing range. Our paper is focused solely on detection probabilities, which may depend on covariates such as visibility conditions and distance of the whale from the perch.

Due to weather and sea ice conditions, the 2009 survey effort failed. Only a few hundred whales were observed–far too few to estimate detection probabilities or abundance. In 2010, the survey was partially successful. There were 399 hours when at least one perch was operating, during which 306 hours of two-perch independent observer effort was maintained. The total numbers of New and Conditional whales (respectively, first or likely first sightings; see Section 2.1 for precise definitions) seen at the primary perch were 1332 and 242, respectively. Of these 1332 New whales, 1216 were seen during 306 hours of IO. Note that many sightings were repeat sightings of the same whale or group made from the same perch or the other one. Despite the partial success of the 2010 survey, it would be difficult to produce a reliable abundance estimate from these data. Ice and weather conditions prevented all survey effort for much of April and May 4-6, and survey effort ended on May 28 before the migration ended. As many as a third to one half of the bowhead population probably passed Barrow during those unmonitored times. However, the 2010 data are ample for detection probability estimation.

The 2011 visual survey was extremely successful. A total of 3379 New and 632 Conditional whales were seen in 859 hours of effort at the primary perch, including 1230 New whales during 180 hours of IO. These counts are within a few whales of the all-time record since surveys began in 1979. Although the total survey effort (859 hours) was much greater in in the previous year due to survey design and weather, the IO effort in 2011 was much less. This change in emphasis from IO and detection probabilities in 2010 to counts in 2011 was planned to optimally use resources and yield the best multi-year dataset for abundance estimation.

This paper describes estimation of detection probabilities from the available survey data. First we describe how the raw survey data are converted into capture/recapture data appropriate for detection probability estimation. After a description of the estimation approach, we present the results and sensitivity analyses. The paper ends with a discussion of our findings and future implications.

2 Data

In Section 3 we describe the statistical model used to estimate detection probabilities (Huggins, 1989). The basis for the analysis is the capture-recapture principle. In a generic case, sighted groups are classified into three categories: those seen by observers in team 1, those seen by team 2, and those seen by both teams. The detection probability (for whale groups) for a team is estimated as the ratio of the number of groups seen by both observers to the total number of animals seen by the other observer team.

However, ordinary capture-recapture methods cannot be directly applied to the bowhead data. Usually, the counts of animal captures and recaptures are known exactly. For example, animals may be marked with bands or matched via photo-identification. In the bowhead case, the best way to count captures and recaptures is less clear because no matching can be done at the moment of sighting or re-sighting (especially at high whale passage rates), and no identifying marks or bands can be placed on the whales. Moreover, even the sightings themselves are subject to confusion since the same whale may be sighted multiple times from the same perch. The sighting and matching methods used in the survey yield very complicated data structures that we call ‘chains’, which include a time- and perch-varying array of covariates, too. In the following subsections we describe the dataset and how it is organized for analysis.

2.1 Linking Sightings

The first stage of our data treatment is to identify captures, in other words whales seen for the first time. One of the most important variables recorded for a sighting is a link code, which indicates whether the observer team believes that it has previously detected the sighted whale(s). The link codes and their meanings are given in Table 1. Each sighting from a single perch is labeled as New (N), Duplicate (R, X, Y, or Z) or Conditional (C), where the latter two categories represent a possible re-sighting of a previously seen whale or group.

These link codes are assigned independently at each perch and pertain only to the sightings from that perch. For example, a New whale seen at one perch might never be seen at the other; or perhaps this New whale is a Y Duplicate within a sequence of previous/future sightings at the other perch, such as N-X-Y-R-Z. Duplicate and Conditional sightings are explicitly perch-specific (as are New sightings) and hence are not recaptures. Due to the presence of Duplicate and Conditional sightings, the total number of sightings reported by a perch exceeds the total number of distinct whale groups sighted.

Connections between related sightings from a single perch are called ‘links’. Links are established by the observer team at the time of sighting using the link codes, and links can only refer backwards in time to previously seen whales. It is required that a New or Conditional whale be identified as the originating sighting for any subsequent sighting coded as a R, X, or Y Duplicate. By definition, C sightings are not linked back to a previous sighting and Z sightings rarely are, but both types may be connected to future sightings via

Link Code	Meaning
N	New whale or group. Observer team is confident that it is seen for the first time.
R	Duplicate. Roll. The sighting is part of a sequence of surface dives or 'roll series' of a previously sighted whale or group. A link is assigned to indicate the associated previous sighting.
X	Duplicate. The observer team is 100% confident that the whale or group can be linked to a specific previous sighting. A link is assigned to indicate the associated previous sighting.
Y	Duplicate. The observer team is about 90% confident that the whale or group can be linked to a specific previous sighting. A link is assigned to indicate the associated previous sighting.
Z	Duplicate. The observer team is quite sure that the whale or group has been previously sighted but the team cannot link it back to a specific previous sighting with 90% confidence. Rarely a link is assigned.
C	Conditional. The observer team cannot determine whether this whale or group is New or a Duplicate of some previous sighting. Links to earlier sightings are forbidden.

Table 1: Link codes for sightings of whales or groups. Every sighting is classified with one of these codes. The three fundamental classifications are New, Duplicate, and Conditional.

those future sightings' backward links. Critically, sightings and links at one perch are totally independent of those from the other perch. If a group was seen by both perches, the links (if any) assigned by each perch are unrelated.

Since time is recorded with each sighting, a sequence of sightings connected by links has a unique link ordering. We call a sequence of linked sightings a 'link chain'. By definition, a link chain must begin with a N or C (or rarely but improperly a Z) and contain no subsequent N or C. Since the perches operate independently, link chains are associated with a single perch. Let an arrow denote a link, with superscripts representing the number of whales reported in the group and subscripts representing perch. It will be implicit that time flows from left to right. Then the link chain $N_1^1 \rightarrow X_1^1 \rightarrow Y_1^2$ represents a link chain created by perch 1, where a new sighting of one whale was linked to an X Duplicate of a single whale, and then to a Y Duplicate where two whales were seen in the group. Note that group size may not be consistent within a chain. The direction of our arrows is, technically, inappropriate because Duplicates are linked backwards in time to previous sightings. However, our notation is more natural to read because it is consistent with time flow. Similarly, we may say that a sighting 'links onward' to another sighting when, more precisely, the future sighting is linked backward in time. Also for simplicity we will omit superscripts or subscripts when they are irrelevant.

We will see later that certain other types of chains may involve loops or other features

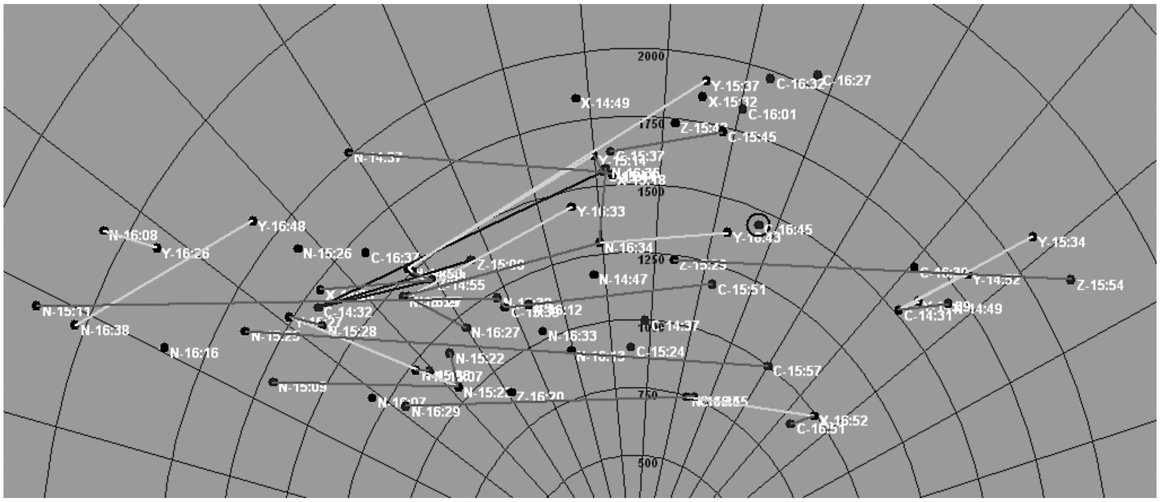

Figure 1: Example of the software display for matching sightings during about two hours on May 8, 2010. Sightings (dots) are labeled by link code and time, and associated by perch according to dot shade. Links and matches are indicated by line segments of different shades. The software displays data but does not attempt any automatic matching.

that preclude the simple notation adopted here. Finally, for brevity later we often adopt the convention that link chains are permitted to be length one, i.e., a single New, Conditional, or Z Duplicate whale, unless the distinction between single- and multi-sighting link chains is important.

2.2 Match Data

2.2.1 Matching sightings

Next we consider the identification of recaptures: whales that were seen from both perches. It is critical to understand that a match connects sightings between perches rather than within a perch. Links connect sightings within a single perch. The observer teams do not communicate, so the process of matching a link chain from one perch to a link chain from the other perch must be done by a third party.

In 2010, matching was attempted both in real time and retrospectively. For the period 1-14 May, observer teams radioed the sighting time, location, swim direction and speed to a command center. The two teams used different radio frequencies. Master matchers in the command center plotted these data using software adapted specifically for this task and tracked sightings approximately in real time, trying to identify between-perch matches. The software was for data display only–matches were determined using human judgment integrating all relevant information. An example of the match display is shown in Figure 1. Sightings are labeled with link code and time. Links and matches are shown with line segments in three different colors. One motivation to attempt real time matching was that

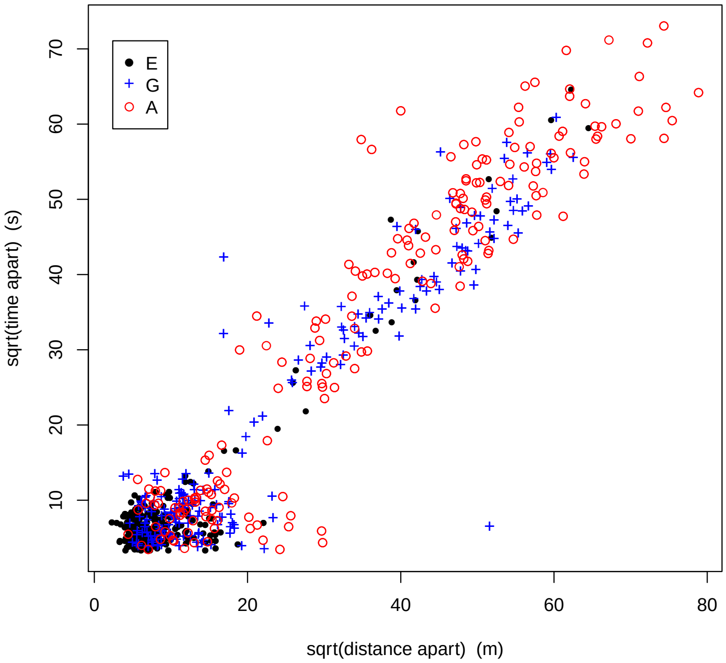

Figure 2: Separation of paired matched sightings in space and time. Each dot represents a match pair. The data have been transformed by taking square roots to better separate the dots. The E/G/A categories are explained in Section 2.2. E matches (black dots) are the most certain, G matches (plus symbols) are intermediate, and A (open circles)matches are the least certain.

the master matchers could radio the perch for additional information if the matchers had a question or needed clarification about a sighting. George et al. (2011) provide much more detail about matching.

At high whale passage rates, it became evident that the real time protocol was not practical. Observers on the perches did not have enough time to record sightings, make links, and also radio back the data to the command center. For the rest of the season, sightings data were collected without real time matching. Several months later, the entire dataset was scrutinized to identify matches. The master matchers used the same software and essentially simulated what would have happened during real time matching, except that the software allowed time to be slowed as much as needed to allow full and careful identification of matches. Match decisions were twice re-assessed and validated through comprehensive reviews later in the year.

In 2011, only post hoc matching was used. Applying the same methods as in 2010, the master matchers produced a matched dataset that they considered to be equivalent to the 2010 match data in terms of data usage and and decision thresholds. The total time spent for the matching effort was about 50 person-weeks weeks for the two surveys.

Figure 2 shows how the sightings comprising a matched pair are separated in space and time, for the 2011 data. Most matched sightings are very close spatially and temporally. The gap in the middle of the plot is a result of whale diving behavior. The sighting pairs in the bottom left portion of the plot are likely to have been sighted by both perches during the same surfacing sequence. The pairs at the top right portion of the plot are likely matches where the sightings were separated by one or more dive periods.

2.2.2 Match chains

Matches explicitly connect sightings, not link chains. However, since every sighting is a member of exactly one perch-specific link chain, a match to any sighting implicitly connects a link chain from one perch to a link chain from the other perch. The entire set of sightings connected by a sequence of matches and links is called a match chain. For 2011, the numbers of match chains, link chains and single sightings were 704, 199, and 3443 respectively.

Match chains may connect a variety of link chains from the two perches, thereby containing multiple matches and multiple links. Note that the matches within a match chain often do not directly connect the earliest sightings within each constituent link chain. Supplementing our previous notation with double arrows to indicate matches, an example match chain might be $N_1^1 \rightarrow X_1^1 \Leftrightarrow N_2^1 \rightarrow Y_2^1 \Leftrightarrow C_1^1$. This represents a link chain from perch 1 (a New whale linked onward to an X duplicate) matched to a link chain from perch 2 (a New whale linked onward to a Y duplicate)–where the match is made between the X duplicate and the New whale at perch 2–and then the Y duplicate is matched to a Conditional sighting at perch 1. The Conditional is not linked to the original link chain at perch 1 even though the master matchers have implicitly asserted that it was an additional sighting of the original chain at perch 1.

The majority of match chains in our dataset are $N_1^1 \Leftrightarrow N_1^1$. With the diversity of cases in the dataset, however, our notation is woefully insufficient. If the first link chain in the previous paragraph was extended so that its X Duplicate was linked onward to an unmatched Y Duplicate at the same perch then the overall match chain would have an additional loose end. If the New whale from perch 1 was matched to the Y Duplicate from perch 2 then the match chain would include a loop. Forks occur when one sighting is matched to several others. In principle a match chain may have quite a few loose ends, loops, and/or forks. Figure 3 shows a match chain of moderately high complexity from the 2010 data.

For the 704 match chains we analyze, the median length is 2, the mean is 2.2, and the longest is 14. To illustrate the potential complexity of chains, consider a whopper from 2010 which has 9 links, 11 matches, 4 forks (one of which is 3-way), 7 loose ends and several loops.

It is impossible for the master matchers to have equal confidence in all match decisions. Thus, to each match they assigned a quality or ‘confidence rating’. Excellent (E) matches were considered to have at least 90% certainty. Good (G) matches were believed to be between 66% and 90% certain. Adequate (A) matches were considered to have between 50% and 66% certainty. In addition to these numerical bounds, the matchers were also given plain language interpretations of the categories: ‘nearly certain’, ‘at least twice as likely as not’, and ‘barely more likely than not’, respectively.

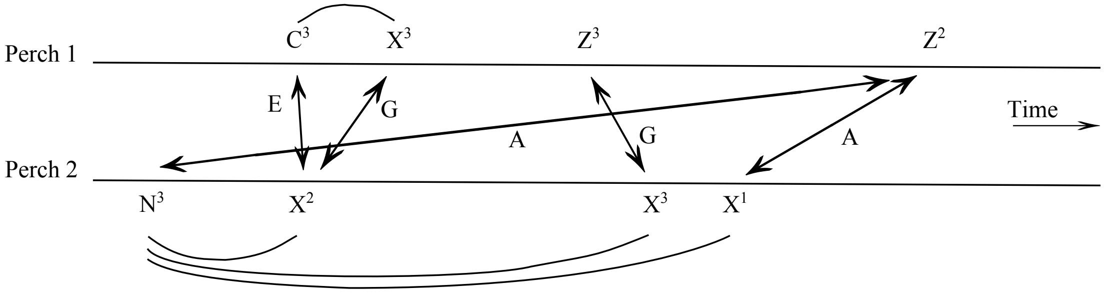

Figure 3: Example match chain structure from 2010. Sightings are ordered in time (flowing from left to right). The vertical axis is irrelevant except to separate the two perches. Arcs represent links and arrows are matches. The sighting notation (N/C/X/Z) is explained in Section 2.1; the match notation (E/G/A) is explained in Section 2.2.2.

2.3 Covariate Data

When a sighting is recorded, a variety of covariate data were collected in addition to the time, sighting location, and link code. Additional covariates such as whale passage rate and chain length (hours) can be derived from the observed information. We list the main covariates in the dataset in Table 2

During exploratory data analysis, we investigated a variety of ways to express certain covariates. The most relevant ones for our analyses are as follows. Distance from the perch is treated as a continuous variable. Visibility is consolidated into three groups: EVG (Excellent or Very Good), GO (Good), and FA (Fair). Although clear, partly cloudy, and overcast weather categories are retained, the remaining categories of heavy fog, heavy rain, heavy snow, light fog, light rain, light snow, and snow/rain mix are consolidated into the category of precipitation. Behavior categories are also consolidated into two groups: easier (breach, spy hop, interaction, tail lobbing, under-ice feeding, and exceptions), and standard (migrating, lingering, resting, trawling, flukes, southbound, unknown, ordinary surfacing, pushing head through hummock, and heard only). Group sizes are consolidated into 1 and > 1, with calves counting towards the total. The rationale for this choice (aside from a good model fit and very few group sizes exceeding two) is that observers felt strongly that the presence of a calf increased detectability at least as much as a second adult whale due to the surface behavior of the cow-calf pair.

Since link and match chains contain several sightings, there is ambiguity about how to assign covariate values to chains. For link chains, the assigned covariate value is taken to be the last value observed within the chain. For match chains, the assigned covariate value is taken to be the value recorded at the first recapture event within the chain. The covariates used in our preferred models vary quite slowly compared to sighting events and are usually assessed consistently by the two perches, so we consider this approach to have at most minor impact compared to other decisions discussed later. For group sizes, the covariate used for model fitting is also taken to be the value observed at the first recapture. However, group sizes introduce an important additional complication in our analysis and are discussed in

Visibility	Unacceptable, Poor, Fair, Good, Very Good, Excellent
Weather	Ten categ. describing cloud cover, precipitation, and fog
Wind Speed and Direction	Miles per hour, degrees clockwise from magnetic north
Ice Coverage	Percent
Lead Condition	Unkn., Closed, Not Contin. (Patchy), Contin., Wide Open
Lead Width	Meters
Swim Direction	North, South, Lingering
Group Size	Number of whales
Calf Presence	Zero or one
Behavior	Fifteen categories, e.g., migrating, trawling, breaching
Number of Observers	Usually 3 or 4 people on a perch
Observer Name	Only for the theodolite operator
Passage Rate	Derived as chains per hour
Chain Length	Derived as total time from first to last sighting
Number of Sightings	Derived as total number of sightings (not whales) in chain

Table 2: Main covariates recorded or derived from sightings data in addition to the time, location, and link code.

detail later; see Section 3.2.

2.4 Data Exclusions

Some data were discarded before estimating detection probabilities:

- Closed or indeterminate lead conditions.
- Distances from perch exceeding 4 km, or missing.
- Visibility rated as poor or unacceptable.
- Single sightings occurring at times when matching would have been impossible. See Section 3.1.1.
- Unlinked X, R, Y, and Z Duplicates. All of these are re-sightings of previously seen groups or single animals. Therefore, these exclusions do not reduce the number of captures from the perch.
- Chains not starting with a New or Conditional sighting.
- A handful of sightings where on-perch calculations of distance (using theodolite data and a hand-held calculator) were wildly inconsistent with post-hoc calculations during matching. In half these cases, an obvious single-digit typographical error was corrected; the other cases were discarded.

All results in this paper pertain to the clean dataset. Note that excluding cases while estimating detection probabilities does not have any more direct effect on abundance estimation. For the latter task, all the sightings can be used.

3 Analytical Methods

3.1 Chain Weighting

Our analysis weights chains to account for differences in the opportunity to match sightings and the confidence in sightings and matches.

3.1.1 Availability Windows

We begin our discussion of weighting by introducing the concept of IO windows. A period of IO effort usually lasted about 4 hours, although longer and shorter times sometimes occurred. We call these periods ‘IO windows’. For a four hour IO window, we define the first and last hours of the period to be the ‘fringe’. The middle two hours are the window ‘core’. For shorter windows, the core will be smaller or even zero. For longer windows, the core will be greater than two hours but the fringes are always one hour.

Based on average swim rates and the experience of observers and matchers, the opportunity to sight a whale that is matched to another sighting is nearly zero before and after one hour from that sighting. Notwithstanding this, the master matchers examined a much longer time frame (2.5 hours ‘lookback’), yet very few sightings were matched beyond one hour apart.

Note that any unmatched sighting within the IO window core is available for matching during the entire time when it could potentially be matched. Sightings in the fringes of the IO window have less opportunity to be matched. A correction should be applied to prevent downward bias in the number of recaptures and corresponding downward bias in estimated detection probabilities.

Call the 2-hour period centered at a sighting (i.e., ±1 hour) the sighting’s ‘availability window’. Then unmatched chains whose availability window is not wholly contained in an IO window might have been matched if IO effort had been expanded. Thus, the evidence such a sighting provides in favor of a non-match is weaker than an unmatched sighting with complete match availability. Weights are assigned to single sightings to reflect this. For a sighting at one perch, define its ‘match availability weight’ to equal the percent of the sighting’s availability window that overlaps with any IO window(s). The simplest example is a case where one perch operates continuously and sights a whale at the exact instant that the other perch ends IO effort. Then this sighting has weight 0.5 because it can be matched backward to the last hour of the IO period but cannot be matched forward to the subsequent hour when the second perch is inoperative. Counterintuitively, a sighting from one perch that occurred outside an IO window may still have some opportunity to be matched: it may be matched to a sighting within an IO window whose boundary is within 1 hour of the outside sighting. Match chains are assigned availability of 1.

3.1.2 Conditional sightings

Another complexity concerns the treatment of chains (including single sightings) involving Conditional sightings. Zeh et al.(1991) and subsequent analyses have treated a Conditional sighting as half a New sighting when estimating detection probabilities and abundance. We too assign a 'Conditional weight' of 0.5 to any Conditional single.

Conditional sightings also occur in link chains. For example, consider a sequence of sightings like $N_1 \quad C_1 \rightarrow X_1$ where, importantly, the gap indicates that there is no link between the initial New whale and the subsequent Conditional sighting. The absence of a preceding link is compelled by the definition of the term Conditional, and indeed the earlier New whale is merely implicit in the sense that every Conditional whale has at least one New or Duplicate whale before it. Assigning equal odds for a Conditional being New or Duplicate, the above sequence of sightings can be interpreted in two ways. With probability 0.5 the Conditional has not been previously seen so the sequence is equivalent to $N_1 \quad N_1 \rightarrow X_1$. On the other hand, with probability 0.5 the Conditional is actually a re-sighting of some previous whale, in which case the sequence is equivalent to $N_1 \rightarrow X_1 \rightarrow X_1$ for some preceding N_1. Thus a Conditional link chain is also half a sighting.

Treating match chains with leading Conditional sightings at one or both perches is more complex–particularly because the chain must be weighted as a single unit rather than weighting the sightings from each perch separately. Nevertheless, we can apply analogous reasoning to show that we should use Conditional weights of 0.75 if one of the matched sightings is Conditional and 0.5 if both are.

3.1.3 Match uncertainty

Section 2.2 explained the Excellent, Good, and Average ratings the master matchers used to rate their confidence in each match. We will associate probabilities p_E, p_G, and p_A to the three types of matches to indicate the probability that the declared match is correct. In our baseline analysis, we will let $p_E = 1$, $p_G = 0.78$, and $p_A = 0.58$, with the latter two values chosen to be the midpoints of the respective confidence ranges used by the matchers. Sensitivity to these choices is explored in Section 4.2.

These match confidence probabilities pertain to matched sightings, not complete match chains. When one considers that each match connection in a match chain may or may not be correct, it is clear that any observed match chain could be a flawed observation of an underlying truth. Many possible changes to the match calls, in isolation or combination, create alternative configurations of the relevant sighting data, and each configuration can be associated with a corresponding probability weight. Specifically, the other configurations can be enumerated by considering all possible combinations of 'retaining' and 'breaking' declared matches, and the probability of each resulting configuration can be derived using p_E, p_G, and p_A.

$N_1 \Leftrightarrow N_2 \Leftrightarrow C_1$	$p_A p_G$
$N_1 \quad N_2 \Leftrightarrow C_1$	$(1 - p_A)p_G$
$N_1 \Leftrightarrow N_2 \quad C_1$	$p_A(1 - p_G)$
$N_1 \quad N_2 \quad C_1$	$p_A p_G$

Table 3: Possible alternative configurations and probabilities for the match chain $N_1 \overset{A}{\Leftrightarrow} N_2 \overset{G}{\Leftrightarrow} C_1$ where the letters above the match arrows indicate match quality ratings. The original chain is most likely and is listed first.

Consider the match chain $N_1 \overset{A}{\Leftrightarrow} N_2 \overset{G}{\Leftrightarrow} C_1$ where the letters above the match arrows indicate match quality ratings. Table 3 lists the possible alternative configurations for this match chain, and the corresponding probabilities. Note that when $p_E = 1$, as for our recommended analysis, Excellent matches are never broken.

For any chain or alternative configuration, define its 'match confidence weight' to be the probability of that configuration using the probabilities discussed here, such as the example in Table 3. For statistical modeling and analysis, each match chain is replaced by the set of possible configurations, each assigned its corresponding match confidence weight. Finally, we note that the original chain always receives the highest probability since all the match confidence probabilities exceed 0.5.

Replacing match chains with sets of chain configurations weighted by their probabilities may seem to be a bold departure from the raw data in until one can witness the actual matching process. To a master matcher, such an approach seems not only sensible but essential. The master matchers can spend up to 1 hour agonizing over just one potential match decision and debating match confidence ratings as they pore over the data. When they decide to declare a match and rate it as Good, they are relying on that confidence rating and its quantitative description (e.g., 'at least 66% confidence') to express their very real uncertainty about the call.

3.1.4 Summary of weighting

In summary, there are three factors that can affect how a chain is weighted. First, the concept of availability is used to adjust sightings made when there was a reduced recapture opportunity due to the timing of IO and single perch effort. This must increase detection probability estimates. Second, the traditional half-weighting of Conditional sightings is applied to single and link chains, with analogous extensions to match chains involving one or more Conditional sightings in a match. This weighting should be roughly neutral with respect to detection probabilities. Third, match chains are weighted by their match confidence. This has a negative effect on detection probabilities. The overall weight for a chain is defined to be the product of these three weights, with the match confidence weighting implemented using the alternative configuration method in Section 3.1.3.

3.2 Group Size Consistency

Each recorded sighting may be of a single animal or a group. These groups are not whale pods in the conventional sense. Except for cow-calf pairs, migrating bowheads appear to have only weak and probably brief allegiance to any aggregation but their associations are not purely random (Zeh et al., 1993). When the number of whales in a sighted group ('group size') is seen to vary along a chain, this may be attributable to whales joining or leaving a chain, and/or to variation in detection. A poll of observer opinions after the survey finds divided views, with some observers tending to attribute group size inconsistencies to a failure to detect group members, and others tending to attribute it to whales ('transients') joining or leaving groups. Group size inconsistencies are relatively common among match chains. Indeed, 574, 109, 16 and 5 match chains had (maximal) group size inconsistencies of 0, 1, 2, and 3 whales, respectively. This corresponds to an inconsistency rate of 18.5% for match chains. The most common inconsistency is a group size 1 matched to a group size 2.

Along with distance, group size is a dominant factor influencing detection probabilities. The fact that animals may join or leave a chain during the period that the chain is (repeatedly) sighted is of paramount importance because estimation of detection probabilities fundamentally relies on counts of captures and recaptures. When group sizes in a chain vary, these counts are uncertain. Thus, resolving group size inconsistency is a major dilemma that must be addressed in the analysis. It is clear that analysis methods that recognize these potential fleeting group allegiances better reflect the behavioral processes that generate the observed data.

We must also recognize that whales are not individually sighted. For the bowhead survey, capture and recapture events pertain to groups, not individuals. Groups are sighted and recorded on the perches, and groups are plotted and matched during the post-hoc matching phase. All the approaches we will describe below embrace this fact: we will estimate detection probabilities of groups. However, this does not require us to ignore group size inconsistency, nor should we. Nevertheless, the first approach we apply for dealing with group size inconsistency is to ignore it. We label this method 'Inconsistent'. This approach amounts to assuming that there are never transient departures or joinings to a group. Next, we introduce another strategy we call 'deconstruction' to model the possibility of transient group affiliation, With this approach, we may isolate certain possible transients, thereby decomposing a match chain into several parts.

To explain deconstruction, we first define a sub-chain of an original chain to be any chain of connected sightings (links and/or matches) from the original chain with consistent group size that can be obtained by reducing the group sizes at any chosen sightings in the original chain. A chain can be decomposed into a unique set of group size consistent sub-chains by sequentially removing the longest chains of each possible group size, starting with the largest group size and working downward. For example, the sub-chains of a chain having group sizes (2,1,1,2,2) in that order are the chains (1,1,1,1,1), (1,0,0,0,0), and (0,0,0,1,1) where '0' indicates absence of chain membership.

Deconstruction identifies sub-chains of link and match chains that can be reasonably attributed to transients. The following is a deconstruction method based on mode chain group size.

1. Calculate the mode m of group sizes for sightings comprising the chain. (Suppose here that we are beginning with a match chain.) The mode is chosen to be a reasonable estimate of the true group size.

2. Replace all group sizes less than m in the chain with m.

3. Set aside the sub-chain having consistent group size m.

4. Subtract m from the group size of each sighting in the original chain. Remove from the chain any sightings with group size 0.

5. Break all remaining matches in the original chain, if any. Although we break these matches, the remnant group sizes are unchanged.

6. What remains, if anything, will be a collection of isolated (i.e., unlinked) groups and link chains with sightings, possibly with inconsistent sizes.

7. Repeat the above steps separately for each link chain. No repetition is needed for any single group because it already has consistent group size. Continue iterating this process within each sub-chain, and for every sub-chain, until what remains is a set of chains all having internally consistent group sizes, including the original chain with the mode group size.

An elaborate example of deconstruction is presented by the chain having the sequence of group sizes (2,3,1,1,1,3,3,4,4). The longest sub-chain we can extract with consistent group size is (1,1,1,1,1,1,1,1,1), assuming we define the mode to be 1. Subtracting these whales leaves (1,2,0,0,0,2,2,3,3), which must be further deconstructed. The next sub-chain is (0,0,0,0,0,2,2,2,2) since we attack the right hand remnant first. Continuing in this fashion yields the remaining components: (1,1,0,0,0,0,0,0,0), (0,1,0,0,0,0,0,0,0), and (0,0,0,0,0,0,0,1,1). This example is for illustration; most real chains are far simpler.

It is possible that the subtraction step #4 may leave match connections between sightings of non-zero group sizes. The reason that these matches are broken in step #5 is because the original chain represents one recapture event. Failing to break remnant match sub-chains would count the match event more than once.

The mode group size may be a tie between two values. Accordingly, we define two alternative methods: 'Decon↓' rounds modes downward, and 'Decon↑' rounds upward.

Deconstruction both subsumes and creates transients. In step #2, whales are added to match chains to represent true group members that were undetected. In steps #4-6, the extra chains (if any) produced represent whales having transient memberships to the true group recaptured. This approach is a compromise between assuming that every chain represents the largest possible group with no transients and assuming that every chain corresponds to the smallest group size seen with all remaining whales being transients. Finally, recall that group size is a direct observation, not a derived variable. The observers are making an explicit statement about what they see. An appealing aspect of deconstruction is that field observations are taken at face value rather than second-guessing the trained observers after the fact.

The choice between the three methods (Decon↑, Decon↓, and Inconsistent) is a choice about how to count captures and recaptures. Each of the three methods makes a different assumption about transients–from none to many–and each of these assumptions leads to the addition of some number of single chains (or none) to represent transient whales. This is unrelated to how whales are counted for the purpose of making an abundance estimate. It is also unrelated to how group size is assigned to a chain as its covariate value for estimation of detection probabilities. As described in Section 2.3, the group size covariate value is defined to be the size observed at the first recapture event. The same group size covariate value is assigned to any sub-chains of the original chain. This is the most appropriate choice because detection probabilities relate to detection of groups.

3.3 Detection Probability Estimation

We adopt the model of Huggins (1989) for capture-recapture estimation for closed populations. This model assumes that captures at each perch are independent and that the catch history is therefore multinomial for each individual. To form the likelihood, the model conditions on the total number of groups detected. We constrain our analysis to assume that probabilities are equal among the perches. The detection probability for a group is allowed to depend on covariate observations for that group. The effect of covariates is the primary focus of our analysis.

Models are fit using the MARK software (White and Burnham, 1999), using the RMark interface (Laake, 2011). However, all of the analyses described here include consideration of weighted observations. A method for weighted fitting of the Huggins model using MARK/RMark is given in the appendix of our preliminary 2010 survey analysis (Givens et al., 2011). We adopt a revision of that approach here.

Briefly, the ith catch history in the dataset is replicated r_i times, where $r_i = Rw_i$ and w_i is the assigned weight. Weights are rounded and R is chosen so that all r_i are integers. Here we round to tenths and set $R = 10$. Then MARK/RMark is used to estimate parameters using the *unweighted* replicated dataset. The factor R acts as an over-dispersion parameter whose effects can be removed after the fact. It can be shown that

$$\begin{aligned}\widehat{\theta}_w &= \widehat{\theta}_R \\ \mathrm{var}\{\widehat{\theta}_w\} &= R\,\mathrm{var}\{\widehat{\theta}_w\} \\ \Delta AIC_c(w) &= -2\delta_r/R + K_1 - K_2\end{aligned}$$

where $\widehat{\theta}$ represents the maximum likelihood parameter estimates for the weighted (w) and unweighted replicated (R) analyses respectively, and $\Delta AIC_c(w)$ represents the difference in AIC_c values between two weighted models. In these equations, $K_i = 2k_i n/(n - k_i - 1)$ where i indexes two models, k_i denotes the numbers of parameters therein, and δ_R denotes the difference in log-likelihoods from the two models fit to the unweighted, replicated data. The latter quantity can be calculated from the fit to the replicated data.

Model selection is carried out using AIC_c as an objective function (Akaike, 1974; Hurvich and Tsai, 1989). The simplest model within 2.0 units of from the minimum is selected (Burnham and Anderson, 2002). For our data, this was essentially equivalent to choosing the model with the lowest AIC_c. We use a stepwise approach (generally forward selection) considering additive and two-way interactive effects, with backward elimination and tangential explorations conducted when it appears they might be helpful. Model selection was especially focused on the variables generally found to be important during our exploratory data analysis with the 2010 dataset: visibility, distance, lead condition and group size.

Automated model selection methods and model averaging are not used. An alternative model selection criterion like BIC (Schwarz, 1978) could be used to penalize model complexity more severely, but in our case the selected models are so simple and the retained effects so clearly scientifically justifiable that such an option is not pursued. The same reasons eliminate the need for automated model selection methods. There are several reasons not to average models here in the manner of Burnham and Anderson (2002). Most importantly, the model fitting process shows that there is very little model uncertainty in the senses that (i) there were few selection choices presenting ambiguous decrements near 2.0, (ii) the same model was selected for each approach, and (iii) fits to alternative models yielded extremely similar parameter estimates and predictions in most cases. Finally, note that the three methods for addressing group size inconsistency use different datasets of different sizes, thereby yielding incomparable AIC_c values.

4 Results

4.1 Main Findings

All three approaches favor the same model: additive effects for distance and group size. There is no evidence of an interaction. This model will be referred to as the 'simple model' hereafter and it is the one we recommend. We also fit a larger model (the 'complex model') which includes additive effects for visibility, distance, lead condition and group size. These are variables that observers consider likely to affect detection probabilities, and these predictors have often been found important in preliminary research with the 2010 data. Again, no interactions were necessary.

The sample-weighted mean detection probabilities for the two fitted models are shown in Table 4. These employed the Decon$\downarrow$ method, which is the one we recommend to address group size inconsistency. Our recommendation of Decon$\downarrow$ is motivated by our belief that this approach best represents bowhead behavioral patterns and likely observer tendencies.

Table 4 also shows the sample-weighted mean the would be obtained when the parameter estimates from the Decon$\downarrow$ simple model are applied to the non-deconstructed data. This is what would be done when forming an abundance estimate. The result differs very little.

Although useful as overall summaries, the sample-weighted means cannot reveal whether

Quantity	Estimate	Std. Error	Confidence Interval
Sample Mean, Simple Model	0.495		
. . .to orig. data	0.492		
Sample Mean, Complex Model	0.494		
Fit for single whale, 3000 m	0.377	0.023	(0.332, 0.423)
Fit for single whale, 2000 m	0.475	0.018	(0.440, 0.510)
Fit for group, 1000 m	0.645	0.032	(0.583, 0.709)

Table 4: Detection probability estimates using the Decon↓ method for the simple and complex models from analysis of the 2011 survey data. The top portion of the table shows sample-weighted means of detection probability estimates from each model, and the mean that would be obtained if the parameter estimates from our preferred simple model were applied to the non-deconstructed dataset. Predictions for single whales and groups at several distances are also shown in the bottom half of the table. These estimates are also based on the simple model and *Decon* ↓.

the methods identify any differing covariate effects. Consequently, we also show in Table 4 the predicted detection probabilities for groups at three distances and two group size categories (group size = 1 and > 1). Since large whale groups near the perch are the easiest to detect, the estimated detection probabilities for groups > 1 at 1000 m is largest. Single whales are more difficult to detect, especially at greater distances, so the corresponding estimate at 3000 m is smallest. Group sizes of 1 predominate in the dataset and 2000 m is very close to the dataset average, so the detection probability estimate shown for single whales at 2000 m is essentially the central value. The standard error for this case is also smallest because of the greater frequency and centrality of such cases.

General detection probability estimates can be expressed via the following equations. Let d and s denote the distance and group size of a sighting, respectively. Then the detection probability p can be calculated as

$$p = \frac{\exp\{\beta_0 + \beta_1 d + \beta_2 I(s = 1)\}}{1 + \exp\{\beta_0 + \beta_1 d + \beta_2 I(s = 1)\}} \tag{1}$$

where $I(s = 1) = 1$ for group size = 1 a single whale and 0 for group size > 1. For our recommended model (simple) and group size approach (Decon↓), the parameter estimates are $\widehat{\beta}_0 = 1.0139$, $\widehat{\beta}_1 = -4.0251 \times 10^{-4}$, and $\widehat{\beta}_2 = -0.3104$.

Our results confirm that increasing distance from the perch has a very strong negative effect on detection probabilities, which (for singles) are as high as 67% near the perch and 29% at 4000 m. The results also indicate that, compared to a single, the odds of detecting a group are increased by a multiplicative factor of $\exp\{.3014\} = 1.35$.

Method	Simple Model	Complex Model
Decon↓	0.495	0.492
Inconsistent	0.509	0.509
Decon↑	0.506	0.504

Table 5: Sample-weighted mean estimated detection probabilities for the three methods of addressing (or ignoring) group size inconsistency. Results are shown for both fitted models (simple and complex).

4.2 Other Results

Our analysis also finds that detection probabilities vary significantly between observers. Although it is common to find observer effects in studies like ours, there is one surprising aspect here. On the perch, search success is a team effort, yet it is only the operator of the theodolite whose identity is recorded for the sighting. The theodolite operator is not the only–nor the primary–person discovering whales, particularly considering the narrow field of view of the device. Moreover, team memberships were continuously remixed so that any particular theodolite operator used the device with many different combinations of teammates. For these reasons, it is difficult to explain the observer effects. Despite an extensive analysis not reported here, we have found no correlation between observer effects and the level of observer experience (e.g., total hours of effort during the survey) or mean whale passage rate during the observers effort.

Our current plan for the future abundance estimate is to ignore observer effects. Such unmodeled extra heterogeneity will tend to cause a downward bias in abundance estimates using standard capture-recapture abundance models (Carothers, 1973, 1979; Otis et al., 1978; Seber, 1982; Pollock et al., 1990; Hwang and Chao, 1995; Pledger and Efford, 1998; Pledger and Phillpot, 2008). Bias is undesirable, but if it is unavoidable then downward bias in abundance is preferable to upward for the sake of conservative management. Since the observer heterogeneity is substantial in our case, any resultant bias can be viewed as a safety guard against the detection probability bias corrections discussed in this paper. In other words, if one does not accept our bias corrections but cannot supply replacements, they may be considered to be 'canceled out' by uncontrolled observer heterogeneity.

We use sensitivity analyses to investigate some of the choices made in our analysis. First, we can ask whether the methods for addressing group size inconsistencies strongly influence the results. Table 6 shows that the three methods produce quite similar sample-weighted mean detection probabilities. This is not the whole story, however. The estimated detection probability for groups > 1 obtained from the Decon↓ approach is roughly 0.07 and 0.10 lower than the ones obtained from the Decon↑ and Inconsistent methods, respectively. Although this may seem like a large difference, consider that less than 20% of the data have group sizes exceeding 1. The net effect of, say, a 0.085 decrease in group effect is a sample-average detection probability reduction of only $0.20 \times 0.085 \approx 0.017$, which is consistent with our

Quantity	Estimate	Std. Error	Confidence Interval
Single, 3000 m			
$(p_E, p_G, p_A) = (0.9, 0.66, 0.5)$	0.323	0.023	(0.280, 0.370)
$(p_E, p_G, p_A) = (1.0, 0.78, 0.58)$	0.377	0.023	(0.327, 0.418)
$(p_E, p_G, p_A) = (1.0, 0.89, 0.65)$	0.391	0.024	(0.346, 0.438)
Single, 2000 m			
$(p_E, p_G, p_A) = (0.9, 0.66, 0.5)$	0.416	0.018	(0.380, 0.452)
$(p_E, p_G, p_A) = (1.0, 0.78, 0.58)$	0.474	0.018	(0.439, 0.509)
$(p_E, p_G, p_A) = (1.0, 0.89, 0.65)$	0.494	0.018	(0.459, 0.529)
Group, 1000 m			
$(p_E, p_G, p_A) = (0.9, 0.66, 0.5)$	0.654	0.034	(0.585, 0.717)
$(p_E, p_G, p_A) = (1.0, 0.78, 0.58)$	0.717	0.031	(0.654, 0.773)
$(p_E, p_G, p_A) = (1.0, 0.89, 0.65)$	0.734	0.029	(0.675, 0.790)

Table 6: Results of sensitivity analysis using high, medium, and low match confidence weights with the Decon↑ method for addressing group size inconsistency. Although Decon↑ is not our recommended approach, it is used here because it produces intermediate results with the medium weights in Table 5.

reported findings.

The second sensitivity analysis investigates the choice for (p_E, p_G, p_A). Recall that the master matchers were instructed to adopt $(0.90, 0.66, 0.50)$ as their decision thresholds. In our trials, we evaluate high, medium and low values. The medium values are those used in the main analyses, namely $(1.00, 0.78, 0.58)$ where the latter two values are the midpoints of the decision boundaries. The high set is $(1.00, 0.89, 0.65)$, and the low set is $(0.90, 0.65, 0.50)$. Table 6 shows estimates obtained using each of these choices. Although we recommend the Decon↓ approach, results in Table 6 employ the Decon↑ approach since it is the one that yields the more central marginal mean. We find that varying (p_E, p_G, p_A) operates in the expected direction: higher matching confidence probabilities yield higher detection probability estimates. However, the magnitude of the changes in the estimates is not extraordinary, especially considering that our high and low scenarios are the absolute extremes of what could be chosen. In other words, observers can't be more than 100% certain, nor would they declare a match that they believe is more likely not one. Overall, our results clearly support our medium choice as a reasonable one.

5 Discussion

It is worthwhile reviewing the big picture surrounding this work. The 2010 survey was not successful for the purpose of estimating abundance, but it was very successful as an independent observer experiment. The 2011 survey was nearly unprecedentedly successful for obtaining sighting counts, and also produced much additional IO data. Together, the IO data from these two surveys clearly present the best opportunity in the history of bowhead surveys to estimate ice-based detection probabilities.

Our ultimate goal for detection probability estimation is to combine the survey data from both years. Unfortunately, we were unable to complete this analysis before the meeting. It may be tempting to discount the need for including the 2010 data. However, this would fail to appreciate how the surveys were planned and how we can pool information from both years statistically.

Early during the 2010 survey it became apparent that sufficient coverage of the migration would be unlikely, so effort was overwhelmingly dedicated to ensuring that we obtained a large amount of high quality independent observer data. For the 2011 survey, the critical need became obtaining equally good count data. Thus in 2011 we prioritized continuous effort from one perch instead of more limited effort from both perches.

This strategy is reflected in Figure 4 which shows the periods of single-perch and IO effort during the survey season in each year. Heavy lines indicate IO periods, with the boundaries indicated with an ‘x’. Light lines indicate periods where only one perch was operating. In 2010, virtually all effort was for IO, and in 2011 coverage with at least one perch was nearly continuous throughout the portion of the season when the largest numbers of whales pass Barrow.

The statistical merits of a combined analysis are as follows. With respect to the marginal annual detection probabilities, the data from only a single perch effectively contribute to the point estimate for that perch. If this were the whole story, we could use detection probability estimates from only 2011 data, along with the 2011 count data, to estimate whale abundance. However, the combined data from both years can jointly be informative about the effects of covariates like distance on detection. By using both datasets, more precise estimation of such effects is possible (assuming there is no interaction between the covariate and year). Indeed, exploratory analysis of the combined data indicates that this approach may be promising. These more precise detection probability estimates may then be applied to the 2011 count data to estimate abundance. Also, compare the option of using the combined 2010-2011 estimates to the status quo. The alternative would be to apply much less recent detection probabilities to the 2011 counts. Clearly the modern data will offer a major improvement.

With respect to statistical methodology, perhaps the greatest focus in this paper has been on bias correction. The challenge presented by the bowhead survey is that the animals, environment, and survey protocol force us to work with data carrying obvious biases. In this paper, we have chosen to confront these biases directly and quantitatively. To balance the possibilities of partial group detection and transient group membership, we introduce the Deconstruction methods. To correct for missing opportunity for matches, we weight cases based on match availability. Finally, to reflect the master matchers’ doubt when declaring matches, we consider weighted possibilities.

In our view, these three sources of bias are so clearly present and possibly substantial that it would be a mistake to fail to account for them. In practice, we find that the Deconstruction adjustments have limited effect, while the other two corrections have greater impact, but

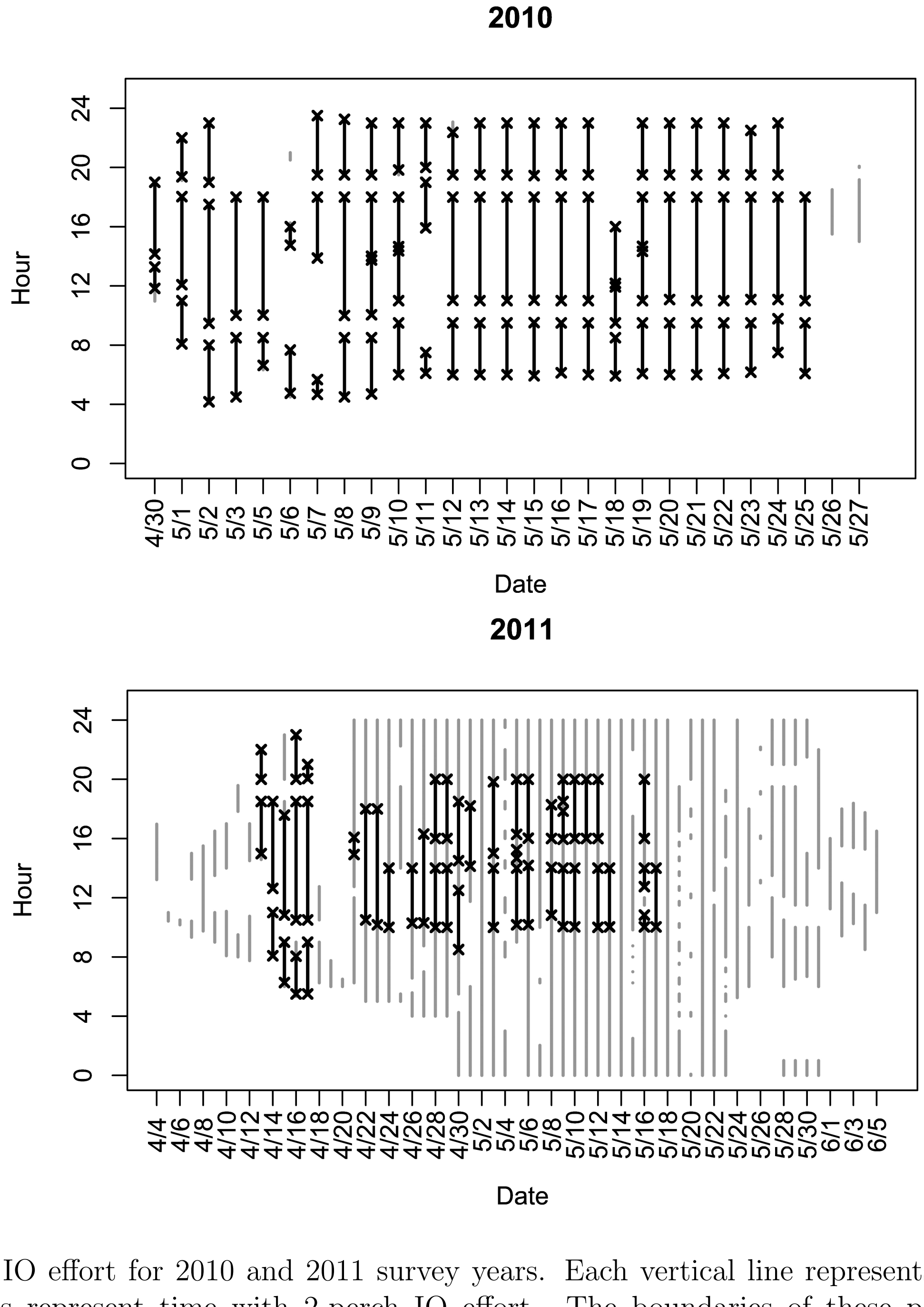

Figure 4: IO effort for 2010 and 2011 survey years. Each vertical line represents one day. Black lines represent time with 2-perch IO effort. The boundaries of these periods are indicated with an ‘x’. Gray lines indicate periods when only one perch was used.

in opposing directions. Overall, the net effect of our corrections is to reduce detection probability estimates compared to a naive uncorrected analysis of the raw data.

Previous estimates of detection probabilities include those of Zeh and Punt (2005). Some of their detection probability estimates are 0.72±0.06 (EVG visibility, ≤2 km), 0.60±0.08 (FA visibility, ≤2 km), 0.40±0.11 (GO visibility,2 km), and 0.33±0.12 (FA visibility, >2 km). Broadly speaking, we estimate lower detection probabilities on average, and the decline of detection probabilities with increasing distance is slightly weaker than theirs. Our standard errors are considerably smaller. At the 2011 meeting of the Scientific Committee of the IWC, Givens et al. (2011) presented preliminary estimates of detection probabilities using only the 2010 survey data. Those estimates did not employ bias corrections as used here, so those results should now be considered unreliable.

One can only speculate about the reasons for the difference between 2011 and the historical estimates. First, the distribution of group sizes appears to have changed from the earliest surveys. For group sizes of 1, 2, and 3, the observed frequencies are 84%, 13%, and 3%, respectively averaging over 2010-2011, but 72%, 23%, and 5% respectively for the combined years of 1975-1978, 1987, and 1988 (Zeh et al., 1993). Annual variation in weather conditions and environmental conditions may be important. Of course observers have also changed over the years, although there is no indication that the distribution of observer abilities is much different now compared to past years. Although scientists have tried to maintain protocol and equipment constant throughout the history of these surveys, there have been inevitable changes and improvements. Finally, there have been marked changes in environmental conditions, particularly ice, over the more than 30 years since bowhead surveys were first undertaken. Changes in whale swim behavior associated with changes in ice conditions may affect the ease of detection.

The 2010 and 2011 surveys are the first ones using an independent observer design rather than the removal method. This change in methodology might contribute to the differences in detection probabilities in the following manner. The past (removal) method was extremely difficult for observers to carry out effectively at high whale passage rates because the north perch was required to receive detailed data radioed from the south perch and attempt to re-detect those whales while also simultaneously maintaining its own separate sighting and data recording effort. One can imagine that, under busy conditions, the demands of receiving and confirming the south perch's sighting data diminished the north team's ability to detect New whales. This phenomenon would lead to an increased number of recaptures at the expense of new sightings, thereby pushing detection probability estimates upward. The IO survey design does not suffer from this problem and therefore might be expected to produce lower detection probability estimates.

The most tempting speculative exercise is to make a rough guess of abundance. We will indulge that temptation here, but only if the reader accepts that our guess is highly speculative. We begin with the subset of the 2011 dataset used to estimate detection probabilities and then estimate corresponding abundance using a Horvitz-Thompson-like approach (1952). Then we apply the following scaling factors: (i) the ratio of chains used for detection probability estimation and consequently in the raw Horvitz-Thompson estimate to the total chains in the dataset; (ii) the ratio of the number of hours of IO availability (i.e., our IO windows plus one hour on each side) to the total number of hours between the days of first and last IO survey effort (13 April – 17 May); (iii) the historical proportion of bowhead groups passing within the 4 km viewing range of the perches; (iv) down-weighting Conditionals by 0.5; (v) the mean group size. The result is 14,700. Note that this does not adjust for whale passage rate within the period of IO effort, nor for the proportion of the season before IO began (4 April – 12 April) and after it ended (17 May – 5 June); see Figure 4. The latter adjustment could raise this guess substantially. This guess is not inconsistent with the estimate from the last major ice-based effort in 2001 which yielded a population abundance of 10,470 and an estimated annual rate of increase of 0.034 Zeh et al. (2005), which extrapolates to 14,200. Now please forget you ever saw that wild guess and wait until next year!

Our goal in presenting these results is to provide readers with an introduction to the types of challenges inherent in independent observer data analysis for Bering-Chukchi-Beaufort Seas bowhead ice-based surveys. The methods discussed here and the corresponding detection probability estimates will be used to produce a bowhead population abundance estimate from the ice-based survey data within the coming year. We welcome suggestions and queries about this work since our goal is to produce an abundance estimate that is as accurate and defensible as possible, and as conservative as is reasonable given the complexities of the survey and data.

6 Acknowledgments

This work was supported by the North Slope Borough (Alaska) and the National Oceanic and Atmospheric Administration (through the Alaska Eskimo Whaling Commission). We thank the many observers, matchers, and other colleagues who conducted this challenging survey with amazing tenacity and professionalism. Jeff Laake is thanked for advice on data analysis and using RMark, and Dave Rugh is thanked for his thoughts on various aspects of this analysis. We also acknowledge the cooperation of the Alaska Eskimo Whaling Commission (AEWC) and thank BP Alaska for providing additional funding for field logistics. We also thank the whale hunters of Barrow who supported our studies and allowed us to conduct survey operations near their camps on the sea ice. Finally, Judy Zeh is thanked for her advice and wisdom about virtually every aspect of the bowhead surveys over 30 years.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716–723.

Braham, H., Krogman, B., Leatherwood, S., Marquette, W., Rugh, D., Tillman, M., Johnson, J., and Carroll, G. (1979). Preliminary report of the 1978 spring bowhead whale research program results. *Rep. Int. Whal. Commn.*, 29:291–306.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach, 2nd Edition.* Springer-Verlag, New York.

Carothers, A. (1973). The effects of unequal catchability on Jolly-Seber estimates. *Biometrics*, 29:79–100.

Carothers, A. (1979). Quantifying unequal catchability and its effect on survival estimates in an actual population. *J. Animal Ecology*, 48:863–869.

George, J., Givens, G., Herreman, J., DeLong, R., Tudor, B., Suydam, R., and Kendell, L. (2011). Report of the 2010 bowhead whale survey at Barrow with emphasis on methods for matching bowhead whale sightings from paired independent observations. Paper SC/63/BRG3 presented to the IWC Scientific Committee, June 2011.

George, J., Herreman, J., Givens, G., Suydam, R., Mocklin, J., Clark, C., Tudor, B., and DeLong, R. (2012). Brief overview of the 2010 and 2011 bowhead whale abundance surveys near Point Barrow, Alaska. Paper SC/64/AWMP7 presented to the IWC Scientific Committee, June 2012.

George, J. C., Zeh, J., Suydam, R., and Clark, C. (1994). Abundance and population trend (1978-2001) of the western Arctic bowhead whales surveyed near Barrow, Alaska. *Marine Mammal Science*, 20:755–773.

Givens, G. H., Edmondson, S., George, J., Tudor, B., DeLong, R., and Suydam, R. (2011). Estimation of detection probabilities from the 2010 ice-based independent observer survey of bowhead whales near Barrow, Alaska. Paper SC/63/BRG1 presented to the IWC Scientific Committee, June 2011.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76:133–140.

Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika*, 76:297–307.

Hwang, W.-D. and Chao, A. (1995). Quantifying the effects of unequal catchabilities on Jolly-Seber estimates via sample coverage. *Biometrics*, 51:128–141.

Koski, W., Zeh, J., Mocklin, J., Davis, A., Rugh, D., George, J., and Suydam, R. (2010). Abundance of Bering-Chukchi-Beaufort bowhead whales (*balaena mysticetus*) in 2004 estimated from photo-identification data. *Journal of Cetacean Research and Management*, 11:89–99.

Laake, J. (2011). RMark package for R. http://www.phidot.org/software/mark/rmark.

Otis, D., Burnham, K., White, G., and Anderson, D. (1978). *Statistical inference from capture data on closed animal populations*. *Wildlife Monographs*, No. 62, 135pp.

Pledger, S. and Efford, M. (1998). Correction of bias due to heterogeneous capture probability in capture-recapture studies of open populations. *Biometrics*, 54:888–898.

Pledger, S. and Phillpot, P. (2008). Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal*, 50:1022–1034.

Pollock, K., Nichols, J., Brownie, C., and Hines, J. (1990). *Statistical Inference for Capture-Recapture Experiments*. *Wildlife Monographs*, No. 107, 97pp.

Raftery, A. E. and Zeh, J. E. (1998). Estimating bowhead whale population size and rate of increase from the 1993 census. *Journal of the American Statistical Association*, 93:451–463.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Seber, G. (1982). *The Estimation of Animal Abundance and Related Parameters, 2nd Edition*. Griffin, London, UK.

White, G. and Burnham, K. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study*, 46, Suppl.:120–138.

Zeh, J., , and Punt, A. (2005). Updated 1978-2001 abundance estimates and their correlations for the Bering-Chukchi-Beaufort Seas stock of bowhead whales. *Journal of Cetacean Research and Management*, 7:169:175.

Zeh, J., Clark, C., George, J., Withrow, D., Carroll, G., , and Koski, W. (1993). Current population size and dynamics. In Burns, J., Montague, J., and Cowles, C., editors, *The Bowhead Whale*, pages 409–489. Special Publication No. 2 of the Society for Marine Mammaology, Lawrence, KS.

Zeh, J., George, J., Raftery, A., and Carroll, G. (1991). Rate of increase, 1978-1988, of bowhead whales, *balaena mysticetus*, estimated from ice-based census data. *Marine Mammal Science*, 7:105:122.

Zeh, J., Ko, D., Krogman, B., and Sonntag, R. (1986). A multinomial model for estimating the size of a whale population from incomplete census data. *Biometrics*, 42:1–14.