Appendix B10: Estimation of survey dredge efficiency relative to HabCam.

Tim Miller and Dvora Hart, Northeast Fisheries Science Center, Woods Hole, MA.

Introduction

Using data from a paired-tow calibration experiment, the goal is to estimate the efficiency of the NMFS scallop survey dredge relative to that of the HabCam. The HabCam survey instrument is usually assumed to be 100% efficient so that the absolute efficiency of the survey dredge can be estimated. However, the relative efficiency of the NMFS survey dredge can be estimated without this assumption.

Methods

The data we have to work are for both HabCam and survey dredge at over 140 stations. For the HabCam, we have a number of images of the substrate along a track at each station. For each image, we have the numbers of scallops as well as the estimated area covered by the image. The HabCam captures images continuously along each track, but a thinned subset are used in our analyses. Thinning is intended to make serial correlation of the images within a station negligible. For the dredge, we have the total number of scallops captured at each station as well as an estimate of the swept area.

Statistical models

For these analyses, we consider different probability models for the HabCam and dredge data, but common to all models is our assumption that the expected catch in numbers of the dredge at station i is

$$E(N_{Di} \mid \delta_{Di}, A_{Di}) = q_D \delta_{Di} A_{Di} \qquad (1)$$

and that of the HabCam for photo j at station i is

$$E(N_{Hij} \mid \delta_{Hij}, A_{Hij}) = q_H \delta_{Hij} A_{Hij} \qquad (2)$$

where δ_{Di} and δ_{Hij} are the average density available to the dredge over the entire tow and the average density in the HabCam for image j at station i, and q_D and q_H are the catchabilities for the dredge and HabCam. The respective areas swept by the dredge and in the image j from the HabCam are A_{Di} and A_{Hij} which are assumed known.

The simplest probability model for count data is the Poisson distribution and in gear comparison studies it is common to make use of binomial models which are conditional on the total catch at a given station (e.g., Millar 1992, Lewy et al. 2004). If the density was constant across all of the HabCam images and the dredge, the binomial model would be useful for these data (Appendix B9). However, densities may vary within a station and the numerous HabCam observations at each station allow us to investigate the plausibility of this assumption.

Suppose that each datum for the HabCam and the dredge arises from a Poisson distribution with mean (and variance) given by eqs. 1 and 2, respectively. If we assume the densities for the HabCam photos at station i to be independently and identically distributed as

$$\delta_{Hij} \sim \text{Gamma}\left(\Delta_i, \tau_{1i}\right)$$

where $E(\delta_{Hij}) = \Delta_i$ is the mean density and the variance is $V(\delta_{Hij}) = \Delta_i^2 / \tau_{1i}$, then the catches in each photo arise marginally from a negative binomial distribution with mean and variance

$$E(N_{Hij} \mid A_{Hij}) = q_H \Delta_i A_{Hij} = \mu_{Hi} A_{Hij}$$

$$V(N_{Hij} \mid A_{Hij}) = E(N_{Hij}) + E(N_{Hij})^2 / \tau_{1i}$$

where $\mu_{Hi} = q_H \Delta_i$. As the dispersion parameter τ_{1i} increases, the variability in densities within a station decreases and the observed number in the image approaches the Poisson in distribution.

We can also model variability in densities for the dredge,

$$\delta_{Di} \sim \text{Gamma}\left(\Delta_i, \tau_{2i}\right)$$

so that the marginal distribution of the number caught in the dredge is negative binomial with

$$E(N_{Di}) = q_D \Delta_i A_{Di} = \rho \mu_{Hi} A_{Di}$$

and

$$V(N_{Di}) = E(N_{Di}) + E(N_{Di})^2 / \tau_{2i}.$$

The dispersion parameter for the dredge is distinguishable from that of the HabCam data which allows the variability among the observed average densities in HabCam images to differ from that of the dredge. This model is estimable when there is only a single observation from the dredge at each station because the mean is related to that of the HabCam images by the relative catchability parameter, ρ, which is informed by data from all stations, and because the mean catch per unit area of HabCam images μ_{Hi} is informed by all of the HabCam images at the station. Therefore, the single observation by the dredge can inform the dispersion parameter τ_{2i}. Note that simpler models where $\Delta_i = \Delta$, $\tau_{1i} = \tau_1$, or $\tau_{2i} = \tau_2$ are special cases.

The relative efficiency of the dredge to the HabCam may differ by substrate type. We observe the substrate in each HabCam image, but the dredge track may cover various substrates which are not directly observed. The lack of these observations for the dredge makes estimation of relative efficiency for specific substrates impossible, but because certain substrates are known to be more prevalent in particular strata, we may consider using these broader regions as proxies that can be used as covariates. As such, we defined three regional indicators for the stations in this study depending on the strata where they occur. Sandy bottom is predominant in the Mid-Atlantic region which includes strata 6130, 6140, 6150, 6180, and 6190 and Georges Bank strata 6460, 6470, 6530, 6540, 6550, 6610, 6621, and 6670 whereas rock and gravel substrates are common in Georges Bank strata 6490, 6500, 6510, 6520, 6651, 6652, 6661, 6662, and 6710. We also formed an alternative set of two regional indicators where the two regions with predominantly sandy bottom were combined.

Model fitting

We fit models using programs in AD Model Builder (ADMB 2009). The likelihood function depends on the assumptions about the parameters and distributions and the parameters were estimated in log-space to avoid boundary conditions.

We restricted the data used for model fitting to stations where there was more than 1 scallop observed in the HabCam images because estimating a positive mean catch per image area at the station is impossible when no scallops are observed. We also removed data for stations where there were less than 2 non-zero counts on HabCam images because fitting negative binomial models for these data at each station requires a sufficient number of positive observations to provide estimates of uncertainty. Ultimately, we used data from 140 of the 146 stations in the original data set.

During the analyses, we discovered that fitted models where the negative binomial assumption was made at all stations for the HabCam data converged in the parameter space where the Hessian matrix was not positive definite. Upon inspection, several of the station-specific dispersion parameters were estimated at extremely large values which implied that the data at these stations were better treated with a Poisson model. We fit both negative binomial and Poisson models to the HabCam data at each station and compared the fits by AIC_c (Burnham and Andersen 2002) to determine which stations we could assume were Poisson distributed. These results were corroborated by inspection of the magnitude of the estimated quasi-likelihood dispersion parameters and negative binomial dispersion parameters at each station.

The full set of models that we fit to estimate relative efficiency of the dredge is provided in Table 1. In the first, most basic, set of models (P/P), we assume the Poisson distribution for all of the data for the HabCam and the dredge. In the second set of models (P/NBP), the dredge data are Poisson distributed and the HabCam data from each station arise from either a Poisson or negative binomial distribution depending on the AIC_c values of those models at each station. For the third set of models (NBP/NBP), both the dredge and HabCam data at each station are either Poisson or negative binomial distributed based on the AIC_c values of the model fits to the HabCam data. In the last set of models (NB/NBP), all of the dredge data are negative binomial distributed.

Within each set of models we allow different parameterization assumptions for specific models (Table 1). The marginal scallop density at a given station may either be constant or station-specific. The relative efficiency may either be constant, region-specific (substrate proxy), or station-specific. For models with negative binomial assumptions, dispersion parameters for the HabCam data may either be constant or station-specific.

One last model in the NB/NBP set was fit where the negative binomial dispersion parameter for the dredge was allowed to be station-specific, but similar to the HabCam data, there were stations where the dispersion parameter was estimated extremely high and variance estimation was not possible. We assigned Poisson distributions to stations where the dispersion parameter estimates were greater than 1000.

Results

As one would expect, the use of AIC_c to determine whether the Poisson is preferred by station corresponds well to the magnitude of the estimated quasi-likelihood dispersion parameter for the corresponding stations (Figure 1). When the quasi-likelihood dispersion parameter is equal to one, the variance is equal to the mean which is an implicit assumption for the Poisson model. Because the variance is always greater than the mean for the negative binomial model, the Poisson model which is more parsimonious is expected to have a lower AIC_c value if the quasi-likelihood dispersion parameter is approximately equal or less than one. The AIC_c criterion also corresponds well with magnitude of the estimated negative binomial dispersion parameter when that model is fitted (Figure 2). When the negative binomial dispersion parameter is large the data approach Poisson in distribution.

That the negative binomial assumption is better for many stations is also reflected in lower AIC_c values (over all stations) for fitted models that allow it (Table 2). The models where the Poisson distribution is assumed for both the dredge and HabCam observations at all stations had the poorest fits based on AIC_c. The lowest AIC_c value for any P/P model was approximately 10,000 units greater than the best fits among other classes of models that we considered.

Fits for two of the models converged but the Hessian was not positive definite and variance estimation was not possible (NBP/NBP M_5 and NB/NBP M_5). These models were among the best fits with regard to AIC_c, but a model with the Poisson assumption for the dredge data and negative binomial or Poisson assumptions for the HabCam data provided the same maximized log-likelihood with fewer parameters and a positive-definite hessian matrix (P/NBP M_5). Although P/NBP M_5 provided the best fit, it is parameterized with station-specific relative efficiencies which cannot be used to infer the efficiency of the dredge in previous years.

The model with the lowest AIC_c that can be used to infer efficiency of the dredge throughout the time series is NB/NBP M_6 which allowed different relative efficiencies for the regions predominant in gravel and sand. The estimated relative efficiency of the dredge is 0.462 (0.006 SE) in the sandy regions and 0.401 (0.011 SE) in the gravel regions.

Discussion

We found that among the fitted models the best fit was provided by allowing the calibration factors to be station-specific. This was not practical for the uses here in the scallop assessment, but these results imply that there is substantial heterogeneity in the relative efficiency of the dredge. A better model would allow a further hierarchy to describe the variation in the relative efficiency, which is an important avenue of analyses in the future.

Of the applicable models that we fit, the best model allowed different relative efficiencies for the regions with predominantly sandy and gravel substrates. The higher relative efficiency of the dredge in the sandy region is expected because the dredge is intended to operate optimally in finer substrates rather than coarse substrates such as gravel and rock.

Finally, it should be noted that these analyses were carried out with swept areas for the dredge based on nominal tow path estimates. Work carried out concurrent to this study suggests that the true tow path is about 4-10% more than those used here. An additional adjustment to our estimates of survey dredge sampling efficiency may be required in some applications.

References

ADMB Project. 2009. AD Model Builder: automatic differentiation model builder. Developed by David Fournier and freely available at admb-project.org.

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer-Verlag, New York.

Lewy, P., Nielsen, J. R., and Hovgård, H. 2004. Survey gear calibration independent of spatial fish distribution. Can. J. Fish. Aquat. Sci. 61: 636–647.

Millar, R. B. 1992. Estimating the size-selectivity of fishing gear by conditioning on the total catch. J. Stat. Assoc. Am. 87: 962-968.

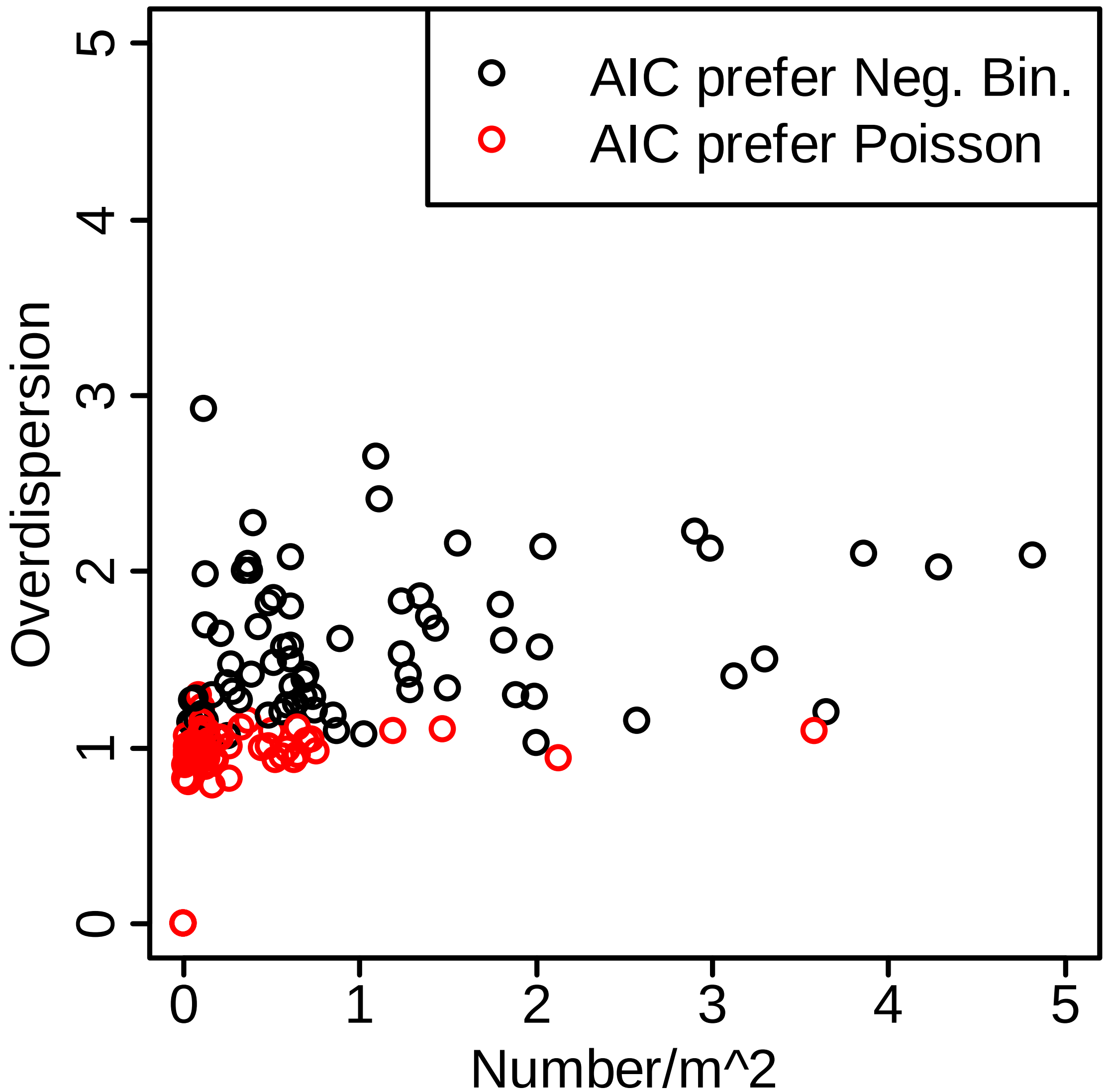
Appendix B10-Table 1. Models fitted to the HabCam and dredge data.

Model	Description	Parameters
P/P M_0	Dredge data and HabCam data are Poisson distributed. Density is constant, relative catchability is constant.	ρ and μ_H
P/P M_1	Dredge data and HabCam data are Poisson distributed. Density is station-specific, relative catchability is constant.	ρ and μ_{Hi}
P/P M_2	Dredge data and HabCam data are Poisson distributed. Density is station-specific, relative catchability is region-specific (Gravel/Sand).	ρ_s and μ_{Hi}
P/P M_3	Dredge data and HabCam data are Poisson distributed. Density is station-specific, relative catchability is region-specific (GB Gravel/GB Sand/MA Sand).	ρ_r and μ_{Hi}
P/P M_4	Dredge data and HabCam data are Poisson distributed. Density is station-specific, relative catchability is station-specific.	ρ_i and μ_{Hi}
P/NBP M_0	Dredge data are Poisson distributed and HabCam data are either Poisson or negative binomial distributed. Density is constant, relative catchability is constant, dispersion is constant	ρ, μ_H, τ_1
P/NBP M_1	Dredge data are Poisson distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is constant, dispersion is constant.	ρ, μ_{Hi}, and τ_1
P/NBP M_2	Dredge data are Poisson distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is constant, dispersion is station-specific.	ρ, μ_{Hi}, and τ_{1i}
P/NBP M_3	Dredge data are Poisson distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is region-specific (Gravel/Sand), dispersion is station-specific.	ρ_s, μ_{Hi}, and τ_{1i}
P/NBP M_4	Dredge data are Poisson distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is region-specific (GB Gravel/GB Sand/MA Sand), dispersion is station-specific.	ρ_r, μ_{Hi}, and τ_{1i}
P/NBP M_5	Dredge data are Poisson distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is station-specific, dispersion is station-specific.	ρ_i, μ_{Hi}, and τ_{1i}
NBP/NBP M_0	Dredge data and HabCam data at each station are either Poisson or negative binomial distributed. Density is constant, relative catchability is constant, dispersion parameters are constant.	ρ, μ_H, τ_1, and τ_2
NBP/NBP M_1	Dredge data and HabCam data at each station are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is constant, dispersion parameters are constant.	ρ, μ_{Hi}, τ_1, and τ_2
NBP/NBP M_2	Dredge data and HabCam data at each station are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is constant, HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ, μ_{Hi}, τ_{1i}, and τ_2
NBP/NBP M_3	Dredge data and HabCam data at each station are either Poisson or negative binomial distributed with a common dispersion parameter. Density is station-specific, relative catchability is region-specific (Gravel/Sand), HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ_s, μ_{Hi}, τ_{1i}, and τ_2
NBP/NBP M_4	Dredge data and HabCam data at each station are either Poisson or negative binomial distributed with a common dispersion parameter. Density is station-specific, relative catchability is region-specific (GB Gravel/GB Sand/MA Sand), HabCam dispersion is station-specific,	ρ_r, μ_{Hi}, τ_{1i}, and τ_2

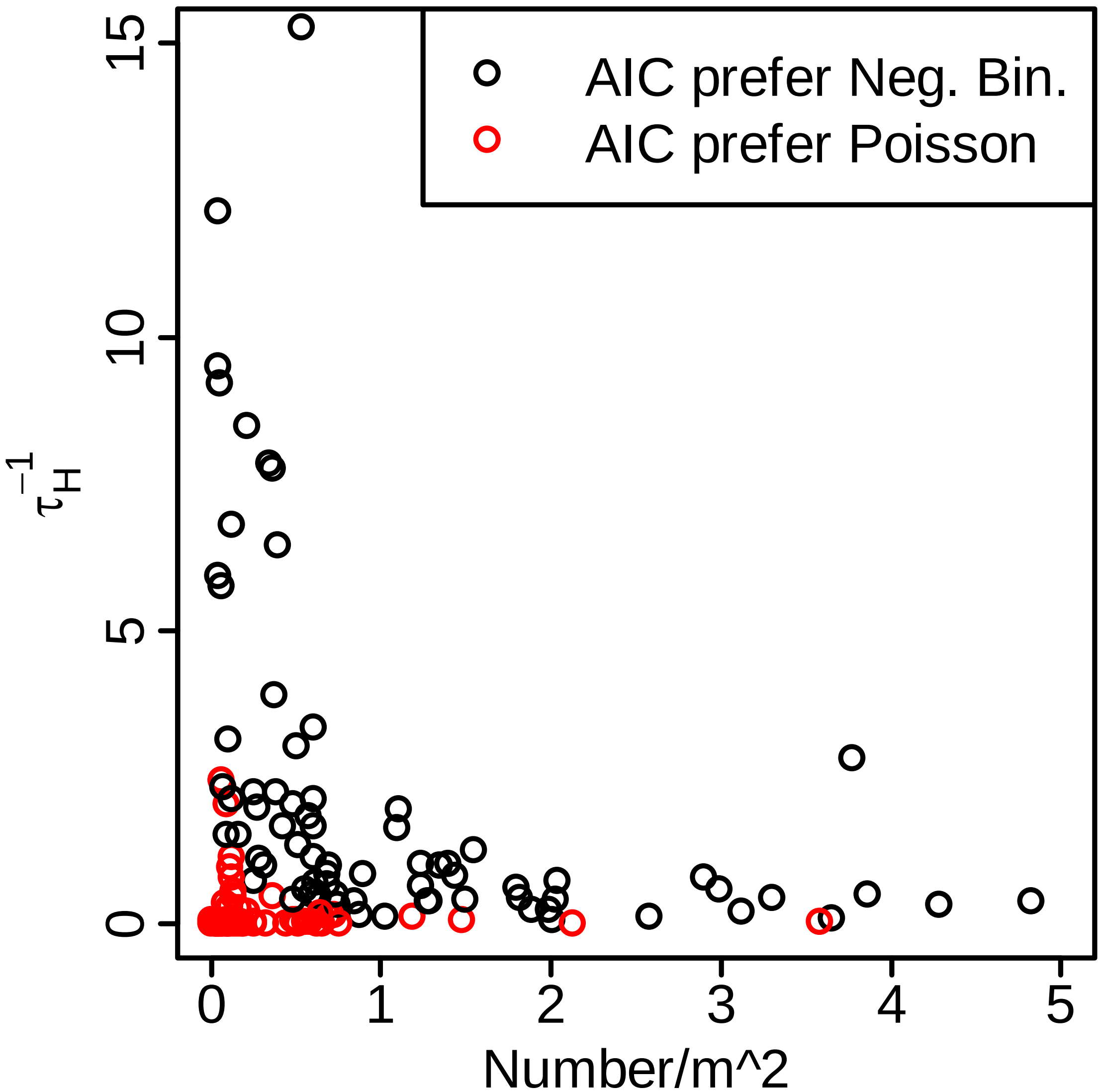
	dredge dispersion parameter is constant.	
NBP/NBP M_5	Dredge data and HabCam data at each station are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is station-specific, HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ_i, μ_{Hi}, τ_{1i}, and τ_2
NB/NBP M_0	Dredge data are negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is constant, relative catchability is constant, dispersion parameters are constant	ρ, μ_H, τ_1, and τ_2
NB/NBP M_1	Dredge data are negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is constant, dispersion parameters are constant.	ρ, μ_{Hi}, τ_1, and τ_2
NB/NBP M_2	Dredge data are negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is constant, HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ, μ_{Hi}, τ_{1i}, and τ_2
NB/NBP M_3	Dredge data are negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is region-specific (Gravel/Sand), HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ_s, μ_{Hi}, τ_{1i}, and τ_2
NB/NBP M_4	Dredge data are negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is region-specific (GB Gravel/GB Sand/MA Sand), HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ_r, μ_{Hi}, τ_{1i}, and τ_2
NB/NBP M_5	Dredge data are negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is station-specific, HabCam dispersion is station-specific, dredge dispersion parameter is constant.	ρ_i, μ_{Hi}, τ_{1i}, and τ_2
NB/NBP M_6	Dredge data are either Poisson or negative binomial distributed and HabCam data are either Poisson or negative binomial distributed. Density is station-specific, relative catchability is region-specific (Gravel/Sand), HabCam dispersion is station-specific, dredge dispersion parameter is station-specific.	ρ_s, μ_{Hi}, τ_{1i}, and τ_{2i}

Appendix B10-Table 2. Number of parameters, maximized log-likelihood value and AIC_c for each fitted model. Log-likelihood and AIC_c values are in parentheses for models without invertible hessian matrices.

Model	No. Parameters	Log-Likelihood	AIC_c
P/P M_0	2	-278,850.0	557,704.0
P/P M_1	141	-72,019.1	144,320.8
P/P M_2	142	-71,581.8	143,448.2
P/P M_3	143	-71,578.9	143,444.4
P/P M_4	280	-62,693.0	125,948.5
P/NBP M_0	3	-250,341.0	500,688.0
P/NBP M_1	142	-63,288.6	126,861.8
P/NBP M_2	242	-60,667.5	121,820.8
P/NBP M_3	243	-60,511.4	121,510.7
P/NBP M_4	244	-60,503.0	121,495.9
P/NBP M_5	381	-57,444.1	115,654.8
NBP/NBP M_0	4	-94,524.7	189,057.4
NBP/NBP M_1	143	-58,743.3	117,773.2
NBP/NBP M_2	243	-57,932.3	116,352.5
NBP/NBP M_3	244	-57,924.1	116,338.1
NBP/NBP M_4	245	-57,918.5	116,328.9
NBP/NBP M_5	382	(-57,444.1)	(115,656.8)
NB/NBP M_0	4	-78,974.0	157,956.0
NB/NBP M_1	143	-58,706.5	117,699.6
NB/NBP M_2	243	-57,895.1	116,278.1
NB/NBP M_3	244	-57,893.8	116,277.5
NB/NBP M_4	245	-57,893.7	116,279.3
NB/NBP M_5	382	(-57,444.1)	(115,656.8)
NB/NBP M_6	315	-57,730.5	116,094.1



Appendix B10-Figure 1. Estimated overdispersion and mean observed number/m^2 from fitted quasi-likelihood model for HabCam count data at each station with log link. Red points indicate that the Poisson model was preferred based on AIC$_c$.



Appendix B10-Figure 2. Estimated (inverse) negative binomial dispersion parameter and mean observed number/m ^{2} for HabCam count data at each station. Red points indicate that the Poisson model was preferred based on AIC $_c$.