Suggestions for Common Data and Model Structures for Island Closures Analyses

D. S. Butterworth
Marine Resource Assessment and Management Group
Department of Mathematics and Applied Mathematics
University of Cape Town, Rondebosch, 7701

Summary

Following the statement of the recent International Review Panel meeting concerning the desirability of the use of common data sets and model structures for comparative analyses of data from the island closure experiment, suggestions are offered in both regards. In terms of agreed data protocols for the experiment, the datasets held by DEFF need to be updated up to and including 2018, and to include information on the month co-variate, where this is possible. For chick survival, discussion is needed on how to proceed, given that different methods of analysis have been suggested to transform the raw data into the form needed for input to the model used to estimate the island closure effect. For model structure, the default agreed locally and by the Panel in 2016 is suggested. Associated aspects related to the chick survival and especially the chick condition response variable are raised as requiring further discussion.

Background

The Report of the December 2020 International Review Panel on some aspects of the Island Closure Experiment (FISHERIES/2020/DEC/SWG-PEL/REVIEW/07) states that: "The work of the Panel was made more difficult because of the lack of like-vs-like comparisons. It is desirable to use common data sets and common model structures for some of the comparisons."

This matter is indicated to be addressed further at a PWG meeting planned for 28 or 29 January 2021. To facilitate that exercise, this document puts forward suggestions for these common data and model structures for the comparison process to which the Panel referred. In line with the Panel's statement, it must be emphasised that these specifications would not be to prevent individual analysts from exploring the sensitivities of their results to other data and model structure choices as well should they so wish; the purpose is to ensure a common basis for comparison of different estimation methods advanced by those different analysts (e.g. estimators based on annually-aggregated compared to those based on individual data records).

Common Data

Any data upon which management recommendations by any DEFF SWG might be based must be available to all SWG participants for potential analysis, whilst perhaps being subject to certain restrictions as laid out in the SWG Code of Conduct (to which each participant signs acceptance). For the Island Closure Experiment, provision of such information is facilitated through the regular circulation by DEFF of "Penguin data" documents, which detail those data submitted and accepted for such analyses in terms of previously agreed protocols.

The most recent document in this series is FISHERIES/2020/SEP/SWG-PEL/100. This provides the obvious default from which to develop such further suggestions as may be necessary. The focus here is on response variables, for which this document lists eight:

1. Fledging success
2. Chick growth rates
3. Active and potential nests
4. Foraging path length

5. Foraging trip duration
6. Maximum distance foraged
7. Chick condition
8. Chick survival

Note:

- Data for all these variables are available for Robben and Dassen islands; but for Bird and St Croix islands, values are available for variables 4 to 7 only.
- As variable 3 has apparently not been collected after 2016 and has not been considered in recent analyses, it is not considered further here (though note also that the series for variable 2 ends at 2014).
- Tables in FISHERIES/2020/SEP/SWG-PEL/100 list annually aggregated values for these variables; the document advises that individual data are available from Janet Coetzee for all except variables 1 and 3.

Suggestions for issues that need to be addressed

A) Extension to more recent years

The most recent response variable data listed in FISHERIES/2020/SEP/SWG-PEL/100 are for the year 2018.

It is suggested NOT to immediately pursue adding more recent data – this will delay analyses, and any such extra data are unlikely to have much impact on any inter-method comparison exercise whose results are of the greatest current interest.

B) The co-variate month

The Panel report repeatedly indicates the need to include 'month' in analyses, for example: "Include 'month' in the model (where such data are available), as month clearly has an effect on some of the variables being measured."

At present, a month variable is present in the provided data files for response variable 2 (chick growth) and 7 (chick condition) only. For the foraging-related variables, limited information on sampling time, relating only to 'season' is available.

Hence ***it is suggested efforts need to be made to extract and (if available) provide 'month' information on response variables 1, 4, 5 and 6 as soon as possible***; persons should be identified for these tasks. See below re response variable 8 (chick survival).

C) Chick survival

The data detailed in FISHERIES/2020/SEP/SWG-PEL/100 include some records where data collection commenced only some period after the chick hatched, so that exposure time is not comparable throughout the data set. Certain recent analyses presented (FISHERIES/2020/NOV/SWG-PEL/117REV) have restricted consideration to records which commence at hatching, as seems desirable to avoid the possible confounding otherwise of exposure time with chick age effects. Continued analyses of these data do, however, first ***require that this revised data set is provided to DEFF in terms of agreed protocols***.

The availability of 'month' information for analyses based on these data could then be achieved by including with each record the date on which the chick concerned hatched for both the dataset where observations start after hatching, as well as the one that is restricted to instances where observations start on the date of hatching.

A further complication is that the data as recorded do not immediately provide survival values; they require prior analysis of some form. Two such approaches have been pursued to date:

i) FISHERIES/2020/JAN/SWG-PEL/06 by OLSPS based on cumulative daily survival to 74 days;
ii) FISHERIES/2020/JUL/SWG-PEL/53rev by Sherley which considers both exponential and log-normal hazard functions (slightly preferring the latter); the results are then transformed to estimates of survival after 74 days.

For the simplest comparable defaults to use to proceed at this time, ***the suggestion is made to work with the survival estimates based on both (i) and (ii), but to limit (ii) to consideration of the exponential hazard function only, and to consider only 'island'', 'year' and 'month'.***

Concerns have been expressed about the possible consequences of colony-specific age-dependence in mortality rates indicated by Kaplan-Meier estimates. This could also have implications for the choice of 74 days as the period for which to compare estimates of cumulative survival. However, it will need time to develop the estimation software required, which needs to allow for incorporation of an alternative plausible functional form as indicated by the Kaplan-Meier results, such as an estimable step-function change in mortality rate after some chick age to be estimated; hence this could be considered only later as a sensitivity.

D) Co-variates

Aside from 'month' as indicated above (as available), and naturally also 'island' and 'year', ***it is suggested that other co-variates not be considered for analyses based on "common" data sets***. However, this is not to preclude analysts from exploring sensitivities which include further co-variates should they wish, provided only that the associated data are made available to DEFF as per existing data protocols.

For certain datasets, for the same colony and year, there can be separate records for two chicks which develop sequentially at the same nest. For the "common" data sets, as is the present situation, it is suggested that such data "pairs" be treated as independent, despite potential correlation. Again, analysts may also pursue sensitivity tests which attempt to take specific account of such non-independence should they wish.

Common Model Structure

Following the 2014 International Panel meeting, proposals were made for a default model structure to use in analysing results from the Island Closure experiment. These were reflected in document FISHERIES/2015/AUG/SWG-PEL/PENG/ALL1 (MARAM/IWS/DEC15/PengD/P1) reported to and discussed at the 2015 Panel meeting. In the light of further computations carried out during the following year, the 2016 Panel meeting (see paragraphs C.1.5.1 and C.1.5.3) recommended estimation using the closure model variant (i.e. with $\lambda_i = 0$) of equation (1) of FISHERIES/2015/AUG/SWG-PEL/PENG/ALL1 (as reproduced in MARAM/IWS/DEC16/Peng Clos/P1arev):

$$-ln\left(F_{y,i,s}\right) = K + \alpha_y + \gamma_s + \lambda_i \frac{C_{y,i}}{\bar{C}_i} + \delta_i X_i + \epsilon_{y,i,s} \quad (1)$$

where

$F_{y,i,s}$	is the penguin response variable for series *s* and island *i* in year *y*,
α_y	is a random year effect reflecting prevailing environmental conditions (assumed to be the same each year, random variation excepted, for both islands in a pair),
γ_s	is an island effect (simply equal to island in most cases, but for fledging success there are two series for Robben Island),
λ_i	is a fishing effect occasioned by the size of the catch taken in the neighbourhood of the island,
$C_{y,i}$	is the anchovy catch taken in year *y* in the 18.148km radius neighbourhood (the area effectively closed on occasion) of island *i*,
$\bar{C}_i$	is the average catch taken over the years for which island *i* was open to the fishery between 2008-2013[1],
δ_i	is a fishing effect occasioned by whether or not the neighbourhood around the island was closed to fishing,
X_i	is a vector with an entry of 1 in years where island *i* is open to the fishery, and 0 where the island is closed, and

[1] The purpose of the $\bar{C}_i$ factor is to scale the catch values in Equation (1) and consequently lead to more readily interpreted λ estimate values, which become more comparable to estimated δ values. The year range 2008-2013 is as was used previously (e.g. MARAM/IWS/DEC14/Peng/B12rev, FISHERIES/2017/DEC/SWG-PEL/38), and has therefore also been used here to maintain comparability of the new with the earlier λ estimates.

$\epsilon_{y,i,s}$ is an error term.

Note that the model is set up in log-space for the response variable, implying an error term for the observed variable value itself with a constant CV as a default. The response variables themselves (the *F*'s) were defined as follows, so that a negative value of δ is to be expected if closure is beneficial to the penguins. The *F* values thus relate to the aggregated values (r) given in the Tables of FISHERIES/2020/SEP/SWG-PEL/100 as follows:

1	Fledging success	$lnF = \ln(r)$
2	Chick growth rates	$lnF = \ln(r)$
4	Foraging path length	$lnF = -\ln(r)$
5	Foraging trip duration	$lnF = -\ln(r)$
6	Maximum distance foraged	$lnF = -\ln(r)$
7	Chick condition	$lnF = \ln(r)$
8	Chick survival	$lnF = \ln(r)$

It is suggested then that the following equation provide the common model structure*:*

$$-ln(F_{y,i,s}) = K + \alpha_y + \gamma_s + \delta_i X_i + \epsilon_{y,i,s} \quad (2)$$

This structure is intended to apply to both annually-aggregated and individual-based approaches, but for simplicity is shown here as it would apply for the former. For annual-aggregated data, the *F* variable would be the year factor estimated from some GLM/GLMM applied to individual data which is standardised to take month into account by including it as a co-variate (minimally as a colony-specific fixed effect). For individual-based approaches, the equation would include subscripts for both month and the individual datum.

Given, however, that variables 7 (chick condition) and 8 (chick survival) are recent additions to the set of variables considered which were accepted for analysis some years ago, and have perhaps not been as fully discussed in the context of how best to relate the measured *r* to the modelled *F* response variable, some further discussion and perhaps modifications in that regard are perhaps warranted.

<u>Chick condition</u>

As defined, the condition factor for a particular chick can be negative. While the use of annually aggregated means is no problem for the structure of equation (2) above, as these means are all positive, a difficulty arises with application to individual data given that logs are to be taken. Certainly therefore, some transformation needs to be introduced.

The simplest possibility might seem to be:

$$F = r - \min(r) + 1 \quad (3)$$

where although the subtraction of min(*r*) ensures positivity, the further addition of 1 is included to avoid taking logs of values in the range [0 ; 1] where major non-linearity might be introduced. Fig. 1 compares the histograms of these individual data (aggregated over year and closure status) for each of the four islands in the experiment before and after this transformation. These plots indicate that the addition of 1 is possibly unnecessary, as the proportion of the data for which $r - \min(r)$ is less than 1 is well below 1% for all of the four islands, so that the "in principle" concern raised about non-linearity probably hardly applies in practice.

However and importantly, using this transformation serves to emphasise a different problem. Fig. 2 compares estimates of the closure effect parameter δ for the aggregated model for each island across alternative transformations. Although the signs of the estimates do not change (except for the St Croix δ estimate which changes from negative to positive), their magnitude does (substantially), rendering comparison with the Panel's suggested threshold value for δ of -0.1 problematic.

This points to a more fundamental problem for this and some others of the response variables being considered. Indeed, it is closely related to a point which the Panel has emphasised in the past: the importance of being able to relate a change in the value of a response variable to a change in the magnitude of the trend in penguin abundance. For chick survival, a value of δ of about -0.1 corresponds to a change in this trend of about 1% pa (see Appendix C of MARAM/IWS/DEC16/Peng Clos/P1a). For most of the response variables available, however, the relationship is unknown, and interpretation of the estimate for δ relies on the assumption of proportionality of the relative change in the response variable and the penguin population increase rate. That is already a stretch for response variables which are necessarily positive, such as foraging duration. However, it becomes more problematic when the variable for an individual record (for a penguin) can have a value which may also be negative – if the mean of these values is close to zero, a closure effect value can then appear very large when expressed as a proportion of that mean, which may be misleading. Note that this problem applies as much to analyses of individual as annual-aggregated data.

Therefore, the issue here is not foremost simply to seek some simple mathematical function which ensures the positivity of this variable after transformation, but rather the question of how does the value of the response variable relate to quantities that determine penguin demographics? Consequently, ***it is suggested that*** rather than focus immediately on the choice of a transformation, ***discussion should first consider the most likely form of the relationship between the value the chick condition response variable and the penguin population growth rate.***

Chick survival

A concern here is that one might expect variability to be likely to increase with chick mortality (expressed as an effectively annual rate $M = -lnS$) rather than independent of lnS as conventional regressions based on equation (2) assume, e.g. the error term in that equation might be more likely to have a constant CV rather than a constant variance. However, taking logs of equation (2) to address this would distort the assumption of additivity of the year, island and closure effects in this space, leading to difficulties in interpretation of the results in terms of how the value of δ impacts the penguin population growth rate.

This relates to the Panel's more general advice to check that residuals are homoscedastic in case differential weighting of data is needed in the maximisation of the likelihood in the model fitting process. The best way to handle this for this variable would seem to be to continue with the current assumption of a constant variance, and then to check whether the actual residuals that result from the regression fits show any trend with the predicted values of M. If they do, that could be addressed by then introducing a weighting term in the minimisation used to achieve the regression, where that term effectively scales the residuals in a manner that removes that trend with M. For an individual data analysis approach, the discrete nature of the observations confounds relating the comments above to residuals *per se*, but nevertheless the Panel's advice requires checking that the variance structure of the default distributional assumption used for the model is consistent with its fit to the data, and modifying that distributional assumption appropriately should that turn out not to be the case

Concluding remark

The content of this document has been developed to fall in line with perceived general agreement that further analyses of the island closure experiment data should be pursued swiftly, but obviously in line with the Panel's comments, specifically as regards use of common datasets and model structures to aid comparisons. More could be suggested as regards further data searches and interrogation, but such extensions would need to be weighed against the further time which they would require and the consequent delay in obtaining finalising analyses.

Acknowledgements

Andrea Ross-Gillespie is thanked for assistance with the computations and preparation of the Figures.

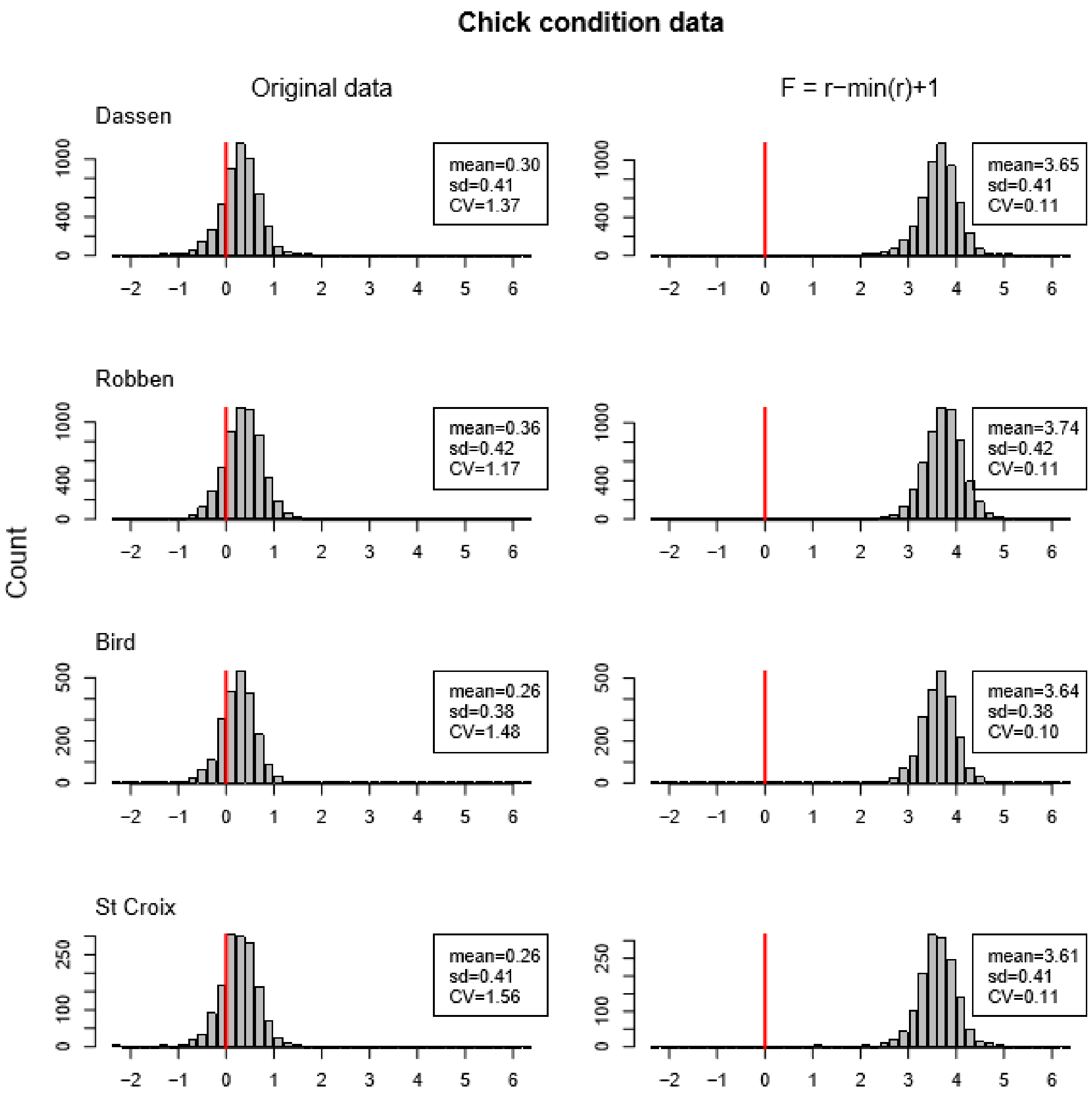

Figure 1: Histograms of the chick condition data (combined over years and irrespective of closure status) are shown for each island. The left column shows the original data, while the right column shows the data after they have been transformed using equation (3). The mean, standard deviation and CV for each distribution is indicated in the legend. The vertical red line marks 0.

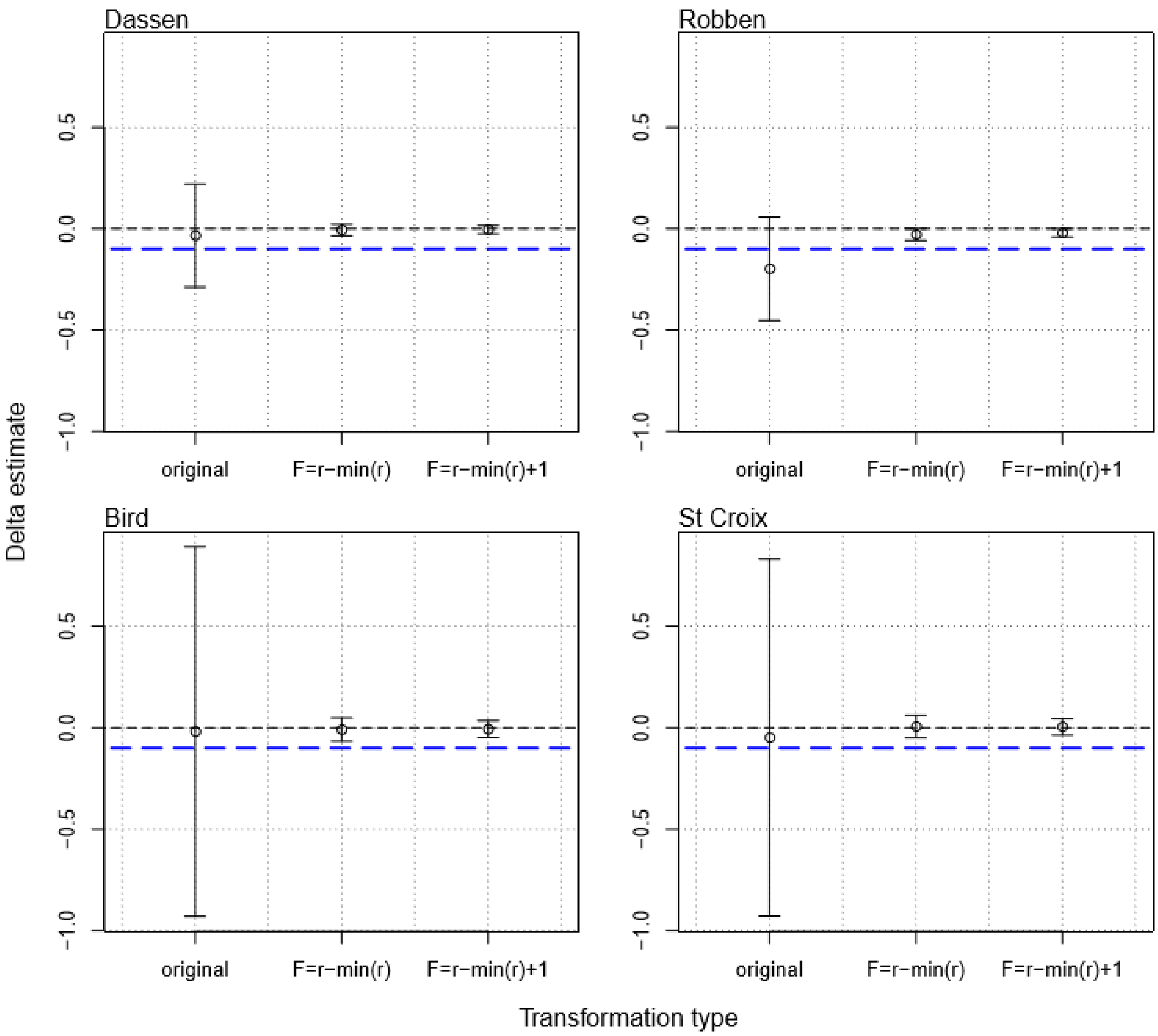

Figure 2: Zeh plots showing the estimates of the δ (closure) effect when Equation (2) is applied to three data sets: (a) the annual averages of the original data (r), (b) the annual averages of the transformation of Equation (3) $F = r - \min(r)$ and (c) the annual averages of a second transformation $F = r - \min(r) + 1$. The point estimate is indicated by the open circle and the error bars show the rough 95% confidence interval given by twice the standard error of the estimate. The dashed blue line shows the Panel's suggested threshold for δ of -0.1.