# MRIP Calibration Workshop White Paper

## Thoughts on Calibration

Jay Breidt and Jean Opsomer

March 9, 2012

Working Paper: Not for citation without permission of the authors

# Thoughts on Calibration

Jay Breidt and Jean Opsomer

March 9, 2012

In the MRIP program, new design and estimation methods have been implemented to replace MRFSS methods. Under the new system, data collection follows an explicit probability sampling design and strict measurement protocols. Estimates computed from the data are fully weighted to reflect the design, and so by construction are unbiased estimates of corresponding population quantities.

Past estimates computed under the original MRFSS did not follow this same approach, so it is not straightforward to weight the data and produce unbiased estimates. Instead, available design information for MRFSS from 2004-2011 was used to create a "retroactive" weighting procedure that accounts at least approximately for the main characteristics of past selection mechanisms. In particular, selection of primary site-days and anglers at selected sites had known design probabilities, and so could be weighted. Common practice in MRFSS was, however, to choose alternate sites when primary sites appeared unproductive. These choices did not follow a rigorous statistical design, so in order to produce weights it was necessary to model the choice procedure using historical data. An additional model was developed to relate interviewer "time on site" to the 24-hour day of angling activity.

To go further back in time, before 2004, requires either an extension of the retroactive weighting described above, or some kind of calibration of old estimates to new estimates. The current document provides a brief description of time series calibration and highlights its major advantages and disadvantages relative to retroactive weighting in the context of MRIP. The decision of which of the two approaches to pursue will require a careful assessment of the available information and resources, as well as a more complete discussion of the costs and benefits of both approaches.

The first of these options, retroactive weighting, requires good quality information on selection methods and confidence that protocols were actually followed in the field. This confidence diminishes as we look further into the past. The retroactive weighting approach also involves significant effort to find/process/or create old design information, to ensure that design assumptions are reasonable, and to ensure weight stability. A significant advantage of the retroactive weighting approach is, however, that the retroactive weights can be applied to all survey variables.

Time series calibration, on the other hand, applies only to a single survey variable at a time. It requires two time series for each variable of interest, one series following the old method and one following the new method. Clearly, a longer overlap period will lead to better data for purposes of building a calibration model. The main advantage of the time series calibration method would be that it is not necessary to find and process old design information, or to model characteristics of the old design if sufficient information was not available. Disadvantages of the time series calibration approach include the significant effort required to develop a valid calibration model, and that the fact this effort would need to be repeated for each survey variable (each species, each state, each mode, etc).

We now briefly describe the steps that would be required to perform time series calibration. We begin by considering the overlap period of both old and new estimates. The old estimates have potential bias due to design-estimation mismatch. The new re-weighted estimates should be nearly unbiased. We therefore "calibrate" to get a consistent time series of estimates over time by modeling the relationship between old and new estimates. Using available old estimates and the modeled relationship, we can then predict what the new results would have been prior to the overlap period.

While various models for the relationship between old and new could be entertained, a basic model for calibration for a particular species in a specific mode could proceed as follows. Let **s**=state and **t**=time point. Suppose that true catch can be written as

**mu(t)+M(s)+m(s,t),**

where

- **mu(t)** is the temporal trend common across states,
- **M(s)** is the state-specific effect that does not change over time
- **m(s,t)** is the temporal trend that is unique to a state.

The new method matches design to estimation and yields unbiased estimates for a given species. Hence, the estimates under the new method can be written as

**New(s,t) = truth + sampling error**

**= mu(t)+M(s)+m(s,t)+e(s,t),**

where the sampling error **e(s,t)** has zero mean. This model for the new estimates can be thought of as the main effect of time, the main effect of state, and the interaction of state and time, plus sampling error.

In contrast with the new method, the old method did not match design to estimation, and so it led to potential bias in estimates. These estimates can be written as

**Old(s,t) = truth + biased sampling error**

**= mu(t)+M(s)+m(s,t)+beta(t)+B(s)+b(s,t)**

where the bias is itself decomposed into three components, specified as follows:

- **beta(t)** is the temporal bias that is common across states
- **B(s)** is the state-specific bias that does not change over time
- **b(s,t)** is the temporal bias and other error that is unique to a state.

We now proceed to estimate the bias components. For states and time periods with both new and old estimates, compute

**Old(s,t)-New(s,t) = beta(t)+B(s)+b(s,t)-e(s,t).**

Imposing standard identifiability restrictions for the model components (and omitting the details here), we can proceed to estimate the bias components as follows:

- **beta(t)** is estimated by averaging over states **s** for each fixed time **t**; call this estimate **Est.beta(t)**
- **B(s)** is estimated by averaging over times t for each state **s**; call this **Est.B(s)**
- Since **e(s,t)** has mean zero, estimate **b(s,t)** by

**Est.b(s,t) = Old(s,t)-New(s,t)-Est.beta(t)-Est.B(s).**

With the model now fitted to the data in the overlap period, we can proceed to calibrate the old estimates, prior to the overlap period. Suppose the overlap period with both **New(s,t)** and **Old(s,t)** runs over periods **t=h+1,…,T.** Consider **Old(s,t)** for **t=1,…,h**, corresponding to estimates using the old method before the overlap period. Calibration proceeds in three steps for **t <= h**:

- Extrapolate **Est.beta(h+1)** down to **Est.beta(t)** using time series and/or regression methods
- Extrapolate **Est.b(s,h+1)** down to **Est.b(s,t) )** using time series and/or regression methods
- Compute the calibrated estimates

**Est.New(s,t)=Old(s,t)-Est.beta(t)-Est.B(s)-Est.b(s,t).**

**Est.New(s,t)** is then the desired calibrated version of **Old(s,t)**.

As already noted, the calibration approach is far from trivial to implement. First, it requires extensive modeling: building time series/regression models for every species in every mode. Second, it requires extensive assumptions: sensible extrapolation back in time assumes that design and measurement methods did not change over time. Third, it requires extensive testing to assess models and assumptions (both statistically and scientifically), wherever possible.

It is possible, in principle, to extend the calibration approach outlined above to include other effects, like wave and its interactions, in the same kind of analysis. For example, the bias effect of wave 6 in New Hampshire might be different from the bias effect of wave 6 in North Carolina, and this systematic difference would need to be reflected in the analysis. Further, it would be possible to incorporate the effects of known design changes, by including indicators (zero-one variables) in the time series or regression models.

The above description of calibration assumes that we are modeling the catch or catch per unit effort for each species separately. An alternative is to apply time series calibration methods to effort estimates. In this setting, either the old or the new estimates could consist of the ratio of the effort from the telephone survey to the sum of the angler-intercept weights from the intercept survey. For the old estimates, these weights would be incorrect in the sense of not matching the design. The methods above could then be applied to the effort estimate series directly, or possibly on some transformed scale, such as logarithmic. Use of such calibrated effort estimates for producing adjusted catch estimates would require further investigation.

To assess the calibration requires that we have available some new-style estimates, with retroactive weighting, that were not used in building the calibration models. Possibilities include setting aside some of the overlap period data for purposes of out-of-sample testing, or building the calibration models for some part of 1998-2003, then later checking these calibrations against re-weighted estimates when they become available. The earliest time period of interest for calibration estimates (prior to 1998) will probably never have re-weighted estimates to check against, because there is not enough design information, and limited quality assurance on the existing data.