INTERNATIONAL DOLPHIN CONSERVATION PROGRAM

# INTERNATIONAL REVIEW PANEL

## 47TH MEETING

LA JOLLA, CALIFORNIA (USA)
5 JUNE 2009

# DOCUMENT IRP-47-13

# ANALYSIS OF DATA REQUIREMENTS FOR THE ANNUAL COMPARISON OF OBSERVER PROGRAMS AND PROPOSAL FOR THE STANDARDIZATION OF OBSERVER DATA

## 1. BACKGROUND

The responsibility for the at-sea sampling by observers of the fishing activities of vessels over 363 t carrying capacity in the eastern Pacific Ocean (EPO) currently corresponds to eight on-board observer programs, the IATTC international program and the national programs of Colombia, Ecuador, the European Union, Mexico, Nicaragua, Panama, and Venezuela.

These eight programs are responsible for observer training, data processing, and data quality control. As part of efforts to maintain data comparability among observer programs, since 2000 the Secretariat has conducted and presented to the IRP an annual comparison of the summaries of the data from each national observer program with summaries of IATTC data from trips by vessels of the same country. The comparisons included both non-statistical and statistical summary information. Non-statistical summaries were presented for the spatial distribution of sets by set type, the average catch per day, the percentage of sets by set type, and the percentage of dolphin sets with zero mortality. (The first two have not been included since 2004.) Statistical tests of average differences were presented for the percentage of trips involving no sets on dolphins, the average number of days per trip, the average mortality per set, the average number of observer interference infractions per trip, and the average number of procedural infractions per set. Although not all observer programs have been active since 2000, there have been few statistically-significant differences found among national program data and data of the IATTC for the same country. Statistically significant differences were found for one comparison of programs in 2000 for the percentage of trips involving no sets on dolphins and the average mortality per set, and for one comparison of programs in 2004 for the average number of days per trip. In addition, for one comparison of programs, statistically significant differences were found for infraction rates in each year. However, the influential differences in infraction rates occurred at the beginning of the time period, and thus affected subsequent comparison because all available data since 2000 were used annually. The results of the majority of statistical comparisons have not been significantly different (for two-tailed tests at an α-level of 5%). Thus, as measured by these comparisons, the programs have been generally comparable.

Historically, the statistical tests of the comparisons between programs have been based on all available data since 2000. This means that in 2008, comparisons between programs active since 2000 were based on eight years of data (*i.e.*, data from 2000-2007). Not all observer programs have been active since 2000, and thus, for some national programs, fewer years of data might have been used in the statistical analyses. The number of years of data that should be included in the comparisons has not been explicitly addressed. Inclusion of too few years of data can lead to meaningless results, due to inherent interannual variability in the data. Inclusion of all available years of data can lead to results that may unduly reflect the first part of the time period without having meaningful bearing on the present situation. At the 46th meeting of the IRP, the Secretariat was asked to conduct an analysis of the number of years that might reasonably be

used in the statistical tests of the comparison of observer programs.

## 2. SIMULATION

As requested by the 46th meeting of the IRP, a simulation was undertaken to provide guidance on selecting the number of years to be used in the annual comparisons. The simulations were based on average mortality per set because, of the items included in the annual comparison, average mortality per set is the one most directly related to comparability of dolphin mortality. Annual mean differences among programs in average mortality per set of 0.005, 0.01, 0.02, 0.05 and 0.10 animals per set were simulated; these correspond to mean differences in mortality between programs (within a country) of 5, 10, 20, 50 and 100 animals for every thousand sets. The simulation involved generating 1,000 simulated data sets. Each data set was generated by randomly assigning actual observer trips of a country to two hypothetical programs, and then adding a constant amount to the mortality of each dolphin set of the trips of one of the two programs. For each of the 1,000 simulated data sets, a randomization test of the difference between programs in the average mortality per set was conducted using 999 randomizations. This is the same type of test used in the annual comparison of observer programs. (For details of the statistical test used in the annual comparison, see Appendix.) The proportion of the 1,000 tests (one test per data set) that would have rejected the null hypothesis of no difference in average mortality per set between programs was recorded. This simulation was repeated for different numbers of years of data (two to five years), and with data from two different countries. The simulations based on the data of the two countries are illustrative because the variability in the mortality rate data of one country (Country B) is greater than that of the other country (Country A). The years of data used as the basis for the simulation were 2003-2007.

The results of these simulations are shown in the table below. The 'No difference' column shows the proportion of tests rejecting the null hypothesis when there was in fact no difference between programs. Significant differences would be expected to be found about 5% of the time (based on a two-tailed test at an α-level of 5%; values between 3.6% and 6.4% are acceptable). The simulation results can be used to provide guidance on selecting the number of years of data to use in the annual comparison, under the assumption that the true difference between programs is constant across years. These simulations could be expanded to include, for example, non-constant differences in mean mortality per set across years. This would require guidance from the IRP as to the specific patterns of differences that were of interest.

These results notwithstanding, the Secretariat suggests a transition from the annual comparison exercise carried out since 2000 to instead focusing effort on improving the quality of data collected by all programs and on ensuring standardization of the data collected. Because the data of all programs are used collectively in scientific analyses, it is important that biases are not introduced into the common observer data base as a result of among-program differences in data collection and data editing procedures. For the

| **Years of data used in test** | Proportion of tests rejecting the hypothesis of no difference in mortality per set (two-tailed test at an α-level of 5%) | | | | | |
|---|---|---|---|---|---|---|
| | Simulated difference in the mean mortality per set (animals per set) | | | | | |
| | **No difference** | **0.005** | **0.01** | **0.02** | **0.05** | **0.10** |
| **Country A** | | | | | | |
| 2 | 0.058 | 0.079 | 0.131 | 0.397 | 0.985 | 1.0 |
| 3 | 0.045 | 0.078 | 0.214 | 0.571 | 1.0 | 1.0 |
| 4 | 0.048 | 0.097 | 0.243 | 0.709 | 1.0 | 1.0 |
| 5 | 0.061 | 0.102 | 0.309 | 0.837 | 1.0 | 1.0 |
| **Country B** | | | | | | |
| 2 | 0.060 | 0.049 | 0.057 | 0.111 | 0.382 | 0.919 |
| 3 | 0.052 | 0.065 | 0.067 | 0.164 | 0.625 | 0.999 |
| 4 | 0.055 | 0.060 | 0.092 | 0.196 | 0.793 | 1.0 |
| 5 | 0.045 | 0.060 | 0.092 | 0.247 | 0.889 | 1.0 |

following reasons, the current annual comparison exercise does not provide a framework for ensuring either the standardization of data among programs or that the data collected are necessarily of the highest quality.

1. First, two programs can be statistically similar but both have data of poor quality. Detailed data collection and editing procedures must be in place in order to build a foundation for maintaining the highest possible level of data quality.
2. Second, achieving comparability of data among programs is first and foremost a matter of standardizing observer training and data editing procedures. Thus, the foundation for data comparability should be built through the use of common observer training materials, and common guidelines and procedures for observer debriefing and data editing.
3. Finally, an important goal of any procedures for the standardization of data among observer programs should be to ensure that data of questionable quality are identified and appropriately labeled, so that they may be excluded from future analyses if necessary. The existing annual comparison does not address this goal.

To ensure that data collected by all programs are comparable and of the highest possible quality, it is recommended that the guidelines outlined below, which build on Annex II of the AIDCP, be further developed and adopted by all programs. Some of these were discussed and agreed upon at the 2nd Meeting of the IATTC and National Observer Programs in October 2007, and some, but not all, of them are already being implemented.

## 3. GUIDELINES FOR DATA STANDARDIZATION AND QUALITY CONTROL BY OBSERVER PROGRAMS

Guidelines on the following aspects of data standardization and quality control are presented below.

1. Selection of observers
2. Training of observers
3. Data collection
4. Data editing
5. Data exchange
6. Additional collaboration among programs

### 3.1. Selection of observers

Candidate selection should be a joint activity of all programs.

Candidates should not be disqualified based on their gender, age, or race.

All candidates must be interviewed.

Candidates must meet the following minimum requirements:

a. Be a university graduate with a degree in biology or related subject (*e.g.*, oceanography, ecology, fisheries biology); recruiting candidates with a degree from a technical discipline closely related to the fishing industry (*e.g.*, fisheries technician, fisheries engineer) is discouraged. Candidates must have completed all credits or the curriculum in their field of study.

b. Be in good physical health, and certified by a physician for at-sea duties.

c. Have completed an at-sea survival course, and provide proof of having passed an at-sea survival test conducted by a government agency. Observers are to pay the cost of the survival course and exam fee, but will be reimbursed for the costs by their observer program upon being hired as an observer.

Whenever possible, candidates must take a psychometric test to assess their psychological fitness

with respect to the job requirements of a fisheries observer.

### 3.2. Observer training

All observer training must be conducted in joint training sessions.

Materials used for observer training must be the identical for all training sessions and for all programs. New materials must be circulated to all programs in draft form for comment.

At the completion of the training course, all candidates must pass a standardized examination in order to be eligible to work as an observer. The standardized examination is to be developed with input from all programs, and updated with new material on an annual basis.

Observer training must include a session dedicated to observer etiquette at sea, and this subject must be covered in the observer manual.

### 3.3. Data collection

All programs must use the same standardized data collection forms, instructions, and manual.

All programs must use the same field guides and species identification keys.

All data forms are to be developed jointly among programs. New materials must be circulated to all programs in draft form for comment.

### 3.4. Data editing

Data editors must, as a minimum, have passed the observer training course, and should preferably have made at least one trip as an observer.

Procedures for the debriefing of observers must be standardized among all programs.

Debriefing of observers must follow the guidelines in a standardized manual (to be developed).

Debriefing of observers must be completed in a timely manner, in accordance with a schedule common to all programs.

All programs must use the same computer programs for editing data.

All programs must edit their data, for both ‘errors’ and ‘warnings,’ in the same manner. Data editing done by all programs must follow the guidelines in a standardized manual (to be developed).

All programs must follow standardized procedures for identification and labeling of questionable data, including species identifications, and catch and bycatch amounts. These procedures must be described in a manual (to be developed).

Program staff must discuss with each observer his/her data mistakes prior to his/her next trip, and provide the observer with a list of his/her latest data errors. The observer must take this error list on his/her next trip for review.

Observers’ performance must be evaluated with standardized criteria, common to all programs.

Each national program and the IATTC program in that country should exchange data editors annually to verify that identical debriefing and editing procedures are being used.

### 3.5. Data exchange

The IATTC and national programs shall:

a. exchange weekly reports of observer departures and arrivals on the respective country’s vessels.

b. exchange preliminary data in a timely manner, in accordance with a schedule agreed upon by all programs.

c. exchange all final edited data in a timely manner, in accordance with a schedule agreed upon by all programs.

### 3.6. Additional collaboration among programs

A periodic meeting (every one to two years) should be held between program staff, including individuals responsible for training observers, for debriefing observers and data editing, and for data bases. The purpose of the meeting is to provide a forum for discussion of program activities, updates/revisions to training materials, editing procedures and computer software, as well as to provide staff with continuing education.

## APPENDIX

### Description of the randomization test used in the comparison of observer programs

To statistically evaluate differences between programs, a randomization test was used to obtain an estimate of the probability that an average annual difference as large as, or larger than, that observed could be due to the chance assignment of trips to programs. The test was performed by randomly assigning trips from the pooled IRP data set for a particular country, by year, to two programs, and then computing the simulated average annual difference in the quantity of interest (e.g., average mortality per set) between programs for the random sample of trips. A total of 4,999 random samples of trips were simulated. The *p*-value for this test was computed as the proportion of simulated average annual differences with an absolute value as large as, or larger, than that actually observed. These calculations represent an approximation to a two-tailed test of the null hypothesis: no difference between programs.