Title

Statistical Learning Procedures for Monitoring Regulatory Compliance: An Application to Fisheries Data

Authors
Cleridy E. Lennert-Cody
Richard A. Berk

Publication Date
2011-10-25

Statistical Learning Procedures for Monitoring Regulatory Compliance: An Application to Fisheries Data

Cleridy E. Lennert-Cody

Inter-American Tropical Tuna Commission, USA[1]

Richard A. Berk

University of California, Los Angeles, USA

Summary. As a special case of statistical learning, ensemble methods are well suited for the analysis of opportunistically collected data that involve many weak and sometimes specialized predictors, especially when subject-matter knowledge favors inductive approaches. In this paper, we analyze data on the incidental mortality of dolphins in the purse-seine fishery for tunas in the eastern Pacific Ocean. The goal is to identify those rare purse-seine sets for which incidental mortality would be expected but none was reported. The ensemble method random forests is used to classify sets according to whether mortality was (response = 1) or was not (response = 0) reported. To identify questionable reporting practice, we construct "residuals" as the difference between the categorical response (0, 1) and the proportion of trees in the forest that correctly

[1] *Address for correspondence*: Cleridy E. Lennert-Cody, Inter-American Tropical Tuna Commission, 8604 La Jolla Shores Drive, La Jolla, California, USA 92037-1508.
E-mail: clennert@iattc.org

classify a given set. Two uses of these residuals to identify suspicious data are illustrated. This approach shows promise as a means to identify suspect data gathered for environmental monitoring.

Keywords: Data quality, Ensemble, Environmental monitoring, Fisheries, Random forest

1. Introduction

Reliable data are a prerequisite to effective management of fisheries. Fisheries management involves a number of difficult yet essential tasks including assessing the population status of target and non-target species, and monitoring fishermen's compliance with fishery regulations. The purpose of fishery regulations, such as limits and closures, is to maintain catch at ecologically sustainable levels, but also to minimize the waste of non-target species. Despite good intentions, such management actions can have the unintended consequence of creating an environment in which data provided by onboard fisheries observers may be misreported in order to achieve compliance. While critical to maintaining data integrity, identification of misreported data can be difficult in situations in which it is anticipated to only rarely occur.

In this manuscript, we examine the problem of identifying misreported fisheries data from the purse-seine fishery for tunas in the eastern Pacific Ocean (EPO). We apply the ensemble procedure random forests (Breiman, 2001) to identify instances in which incidental mortality of dolphins was likely as a byproduct of the manner in which tuna were caught. Data from sets for which no mortalities were reported when mortality would be expected are subjected to additional analyses through which suspicious reports and suspect fisheries observers are identified. Our goal is to illustrate an approach to compliance monitoring that may be more generally useful in situations in which violations are anticipated, yet expected to be rare events.

2. The tuna purse-seine fishery

The international purse-seine fishery for tunas in the EPO (Fig. 1) currently produces approximately 20-25% of the world's catch of yellowfin tuna (*Thunnus albacares*) (e.g., IATTC, 2002). Fourteen countries presently participate in this fishery (IATTC, 2004a). The Inter-American Tropical Tuna Commission (IATTC) that was established by an international Convention in 1950 (Bayliff, 2001) is responsible for the conservation and management of this fishery for the member countries. The IATTC operates an international observer program on behalf of the related Agreement on the International Dolphin Conservation Program (AIDCP, 1999). Fisheries observers go to sea aboard the largest size category of fishing vessels in order to collect data on the incidental mortality of dolphins and details of the fishing operations. Additionally, these observers collect data on the local environment, the amounts and species of tuna caught, and the incidental mortalities of non-mammal species, such as sea turtles and a number of fishes (IATTC, 2004b). These data, and other resources, are used by IATTC staff to monitor the status of tuna populations in the EPO, the amounts of incidental mortality of bycatch species, and the compliance of fishermen with catch and bycatch reduction measures. ("Bycatch" refers to any non-target species that is killed incidental to fishing operations.) Enforcement of such measures is the responsibility of the member countries of the IATTC.

To date, dolphins are the only bycatch group involved in this fishery for which mandatory limits exist. Fishermen use the association of tunas with dolphins in the EPO as one means of locating and catching tunas (National Research Council, 1992). Yellowfin tuna (*Thunnus albacares*) is found in association with several species of dolphins, primarily the spotted dolphin (*Stenella attenuata*), the spinner dolphin (*S. longirostris*), and to a lesser extent the common dolphin (*Delphinus delphis*) (Allen, 1985; Hall *et al*., 1999). In order to locate tuna, fishermen may search for signs of dolphins at the sea surface (*e.g*., splashes, animals swimming). To catch the tunas, the fishermen chase and attempt to encircle the dolphins with the purse-seine net. If the fishermen are successful, encircling some percentage of the dolphin herd will also result in capture of tunas. Fishermen then take advantage of the vertical stratification of the tunas and dolphins within the pursed net, with the dolphins being closer to the surface, to release the dolphins before loading the fish aboard the vessel. Incidental mortality of dolphins can occur if the dolphins become entangled in the net prior to release. Yellowfin tuna is also caught with purse-seine nets in the EPO in two other ways: as unassociated schools of fish, and in association with floating objects (*e.g*., flotsam, fish-aggregating devices placed in the water by fishermen) (National Research Council, 1992; Hall, 1998). However, the yellowfin tuna found in association with dolphins tend to be larger on average, and hence more desirable from both economic and ecological perspectives, than those tunas caught as unassociated fish or in association with floating objects (Joseph, 1994; Hall, 1998).

Since the late 1970's, management of marine mammal bycatch associated with this fishery has been the focus of national and international observer programs, legislation, and efforts by conservation organizations (Joseph, 1994; Hall, 1998; Gosliner, 1999). These efforts have resulted in a decrease in the incidental mortality of dolphins from an estimated hundreds of thousand of animals annually in the 1960's to early 1970's, when the first systematic observer sampling program began (Lo and Smith, 1986; Wade, 1995) to less than 5,000 animals per year since 1993 (IATTC, 2004b). Mortality reduction has been approached from several angles. Modifications to fishing gear and adoption of release techniques since the late 1950's (National Research Council, 1992) are responsible for the greatest reduction in mortalities. National and international legislation to establish fleet-specific mortality limits has served to promote reductions in incidental mortalities since the early 1970's (Joseph, 1994; Hall, 1998; Gosliner, 1999). In 1993, an international agreement (Joseph, 1994; Bayliff, 2001) established annual individual vessel limits on dolphin mortality for the first time in this fishery. Following the implementation of these annual vessel limits, the incidental mortality rate has continued to decrease to less than 10% of the 1992 level (IATTC, 2004b).

While individual vessel limits have positive implications for the overall level of incidental mortality associated with this fishery, there is concern that they may also have negative implications for the quality of fisheries observer data. Individual vessel limits have increased each fishermen's personal responsibility for the reduction of dolphin bycatch. As a result, fishermen are likely to make a more concerted effort to release any entangled dolphins alive from the net. However, there is concern that per-vessel limits may have also had the unintended consequence of creating an environment in which fishermen attempt to reduce the reporting of dolphin mortalities by coercing observers to alter their data. The percentage of dolphin sets with misreported mortality data is anticipated to be small, a few percent or less, making these data difficult to identify.

It is believed that any misreporting of dolphin mortality data should manifest itself as an under-reporting of mortalities. The pressure brought to bear on fishermen by way of the individual vessel limits has been significant: limits have decreased from an allowed 183 dolphins per boat per year in 1993 to around 50 dolphins per boat per year since 1999, approximately one quarter to one third of the 1993 quota. Reaching a limit has financial implications for both fishermen and vessel owners. Vessels that reach or exceed their limits are prohibited from setting on tunas associated with dolphins for the remainder of the year, and any excess mortalities are deducted from subsequent years' limits (AIDCP, 1999, Annex 4, Section III). Moreover, dolphin mortalities can have financial implications even if the limit is not reached. Access to the United States' market requires that no dolphins were intentionally set upon or killed.

3. Fisheries observer data

Data collected by IATTC observers onboard large tuna vessels of the international purse-seine fleet between 1993 and 2002 were used in this analysis. Sampling coverage by IATTC observers over this 10-year period was greater than 65% annually. Only purse-seine sets targeting tunas associated with dolphins (hereafter referred to as "dolphin" sets) were considered. Dolphin sets for which data were not available on all the predictor variables of interest (Table 1) were excluded prior to analysis. In addition, sets for which the observer's estimate of the number of dolphins encircled was zero were excluded because either the observer was not able to obtain a good estimate of the number of animals encircled or dolphin mortality was not an issue because no animals were in the pursed net. (Mortalities that are hypothesized to occur outside the period of observation of the dolphins by the observer (Archer *et al*., 2001) were not considered.) On average, annually 89% of the mortality and 90% of the dolphin sets were retained for analysis. After trimming, data on 52,402 dolphin sets made between 1993 and 2002 were retained. These data were collected by 209 different fisheries observers on 1,976 fishing trips of 201 different fishing captains.

The dolphin mortality data are characterized by many zero-valued observations and a long right tail (Fig. 2; IATTC, 2004b). Annually, fewer than 16% of sets available for analysis had any reported dolphin mortality. In addition, over the 10 year period covered in this analysis, there has been an overall decreasing trend in the average incidental dolphin mortality rate from about 0.52 animals per set in 1993 to about 0.12 animals per set in 2002 (IATTC, 2004b). The large percentage of zero-mortality sets, combined with the fact that occasional sets had reported mortalities of 10s of animals, have made analysis of these data problematic, particularly with more conventional techniques that require specification of a stochastic model (Lennert-Cody *et al.*, 2004). Because it is not only the number of incidental mortalities, but also the occurrence of any incidental mortalities, that may be problematic for fishermen, we regard the problem as a two-group classification problem, in which the response variable takes on a value of 0 for sets with no reported mortality and a value of 1 for sets with reported mortality.

A total of 36 predictors describing environmental conditions, operational problems, fishing operations, use of rescue equipment, and biomass characteristics, were included in this analysis. These predictors are discussed briefly below; a more detailed description of each predictor is presented in Table 1. Previous analyses (Lennert-Cody *et al*., 2004 and references therein) have demonstrated that dolphin mortality can be more likely to occur in the presence of some types of operational problems that make the net difficult to handle and can thus hamper dolphin release efforts. Certain types of fishing operations can make dolphin mortalities more likely to occur because they tend to put the animals into closer proximity with the net for extended periods of time. In addition, large numbers of both dolphins and tunas in the net can make net handling more complex and increase the chance of dolphin mortality. Environmental conditions, such as strong currents, may make dolphin mortalities more likely to occur because they can complicate handling of the net or they may make mortalities difficult to observe, such as rough seas. On the other hand, dolphin rescue procedures can reduce the chance of mortality. Previous analyses of these data (Lennert-Cody *et al*., 2004) have shown that some of the predictors in Table 1 are correlated, and a variety of statistical interactions are to be expected. The product variables required for the latter increase even more the linear dependence among predictors. Moreover, incidental mortality of dolphins has been found to be only weakly correlated with the majority of these predictors. In short, the existing data present substantial challenges.

4. Data analysis using random forests

4.1 *A brief overview of random forests*

Random forests (Breiman, 2001) is an ensemble method that extends the classical idea of classification and regression trees (CART; Breiman *et al*., 1984) by generating predictions based on a large collection of trees ("forest") instead of an individual tree.

For a data set with *n* observations, the general random forest two-group classification algorithm has the following form (Berk, 2005).

1.Take a random sample of size *n* with replacement from the data.

2. Take a random sample without replacement of the predictors.

3. Using the random sample of predictors, construct as usual the first CART partition of the data.

4. Repeat steps 2 and 3 for each subsequent split until the tree is as large as desired. Do not prune.

5. Drop data not selected to be in the sample (*i.e.*, not selected in Step 1) down the tree. (These data are called "out-of-bag" data, or "OOB" data.)

6. Store the class assigned to each of these OOB observations.

7. Repeat Steps 1-6 a large number of times (*e.g.*, 500).

8. Using only the class assigned to each OOB observation, count the number of times over trees that the observation is classified into one category and the number of times over trees it is classified into the other category.

9. Assign each case to a category by a majority vote over the set of trees.

The sampling of the data and the averaging over trees compensates substantially for overfitting. The sampling of predictors as trees are being grown facilitates the construction of very flexible fitting functions in which highly specialized predictors are given the opportunity to contribute (Berk, 2005). Breiman (2001) has shown that random forests has excellent classification and forecasting skill in part because it does not overfit. Experience to date (Breiman, 2001; Berk, 2005) suggests that in general, it performs at least as well as currently available alternatives such as boosted trees (Friedman, 2002).

Note that because the classifications derived from random forests are based on data not used to build the trees, the classifications are true forecasts, and the classification errors are in actuality forecasting errors.

4.2 *Analysis of the data*

Anticipating a certain amount of "data snooping," we randomly divided the data into a training sample of 34,858 observations (two-thirds of the data) and a testing sample of 17,429 observations (one third of the data). The analyses discussed in this section are based on the training sample. The test data set was used to identify unusual observations (see Subsection 4.3 below).

The application of random forests to the dolphin mortality data is complicated by the highly skewed response variable and by the need to take into account the relative costs of false positives and false negatives. The costs for incorrectly classifying sets with dolphin mortality are different from the costs incorrectly classifying sets with no dolphin mortality. And, with different relative costs come different results. For these data, 89% of the sets were not associated with any dolphin mortality ("0" class; "negatives"), while 11% of the sets were ("1" class; "positives"). It follows that it is initially far easier to correctly classify sets for which there was no reported mortality than to correctly classify sets for which mortality was reported. Simply capitalizing on the unbalanced marginal distribution alone, if all sets were classified as having no dolphin mortality, that classification would be correct 89% of the time. But then, all sets with dolphin mortality would be misclassified. Such a result would imply that the cost ratio of false positives (no mortality classified as mortality) to false negatives (mortality classified as no mortality) was infinite.

With the highly unbalanced marginal distribution of the response variable, and an assumption of equal costs for false positives and false negatives, random forests (available in R as *randomForests*) produced unsatisfactory results. Most sets in which dolphin were reported killed were misclassified, and the balance of false positives to false negatives implied a cost ratio of 11 to 1 (Table 2A). For our application, these implied costs were undesirable.

The relative cost of false positives to false negatives ultimately must be set by fisheries managers, not an easy task in this situation because outcomes are vague and therefore difficult to value. For example, stocks of spotted and spinner dolphins most affected by the fishery are depleted relative to pre-fishery levels (Wade, 1994). However, the long-term health of these populations remains a matter of much debate (NMFS, 2002). As a result, the question of how much effect any under-reporting of dolphin mortality could have is unresolved, and therefore, it is difficult to put an ecological cost on under-reported mortalities. This is particularly true because alternative purse-seine fishing modes are not without their own serious bycatch concerns, albeit with respect to other species (Hall, 1998). For a variety of legal and procedural reasons, human costs (*e.g.*, mistakenly accusing fisheries observers of misreporting data) are equally difficult to assess. Nonetheless, given that dolphin mortalities are possibly under-reported, it is clearly desirable to place added importance on properly classifying a dolphin set that had reported mortality. This suggests minimally setting the relative cost of false negatives as being twice that of false positives. It may be argued that the relative costs should be higher.

To illustrate our approach, we set the relative cost of false negatives as twice that of false positives. In order to implement these relative costs, we exploited an option in the software that allows one to alter the resampling probabilities used to construct the sample data (Step 1 of the random forest algorithm). In effect, two strata were constructed, one for the sets with reported dolphin mortality and one for sets with no reported dolphin mortality. The former were sampled with a probability of .37, and the latter were sampled with a probability of .63. The less common event of dolphin mortality was oversampled to achieve in the end the 2 to 1 cost ratio of false negatives to false positives. The resulting random forest classifier based on 5,000 trees incorrectly classified sets that had reported dolphin mortality 44% of the time and incorrectly classified sets that had no reported dolphin mortality 11% of the time (Table 2B).

In random forests, a measure of the importance of each predictor is the decline in forecasting skill that occurs when values of that predictor are randomly shuffled, and hence its relationship with the outcome is scrambled (Berk, 2005). As anticipated, the average decline in prediction accuracy showed that a few of the predictors were far more important than others (Fig. 3). For example, randomly shuffling the “duration of backdown” or “net canopies” reduced forecasting accuracy by about 4% (*e.g.*, from a misclassification rate of 44% to 48%). When “dolphin species” and “number of dolphins” were each shuffled, forecasting accuracy declined by about 3%. These relatively small effects are anticipated when there are many predictors that are at least moderately correlated. The forecasting skill lost when a given predictor is shuffled will be in part replaced by the forecasting skill of the other predictors with which it is correlated.

From other output (not shown), it is clear that these effects are consistent with our understanding about what puts dolphin at risk within the net. Longer backdown times extend the period during which dolphins are in close proximity to the net where they risk entanglement, and may lead to the formation of net canopies which can trap dolphins below the water's surface. Species such as the spinner dolphin and common dolphin are more likely to exhibit rapid swimming behavior within the net, increasing chances of entanglement, whereas the spotted dolphin is more likely to wait passively to be released (Pryor and Shallenberger, 1991; Schramm, 1997). The greater the number of dolphins encircled, the greater the chances of mortality because of the sheer magnitude of the biomass in the net, and because large groups require more time to be released from the net, increasing the duration of backdown and the chance that net canopies will form. The agreement of the direction of these effects with knowledge of fishing operations gives the random forest results subject-matter credibility.

Given our concern about misreporting, there are grounds for questioning these random forest results. Errors in the response variable might distort the findings. However, there was good reason *a priori* to believe that misreporting, while critical for the integrity of the monitoring, was relatively rare. First, IATTC staff review each observer's data with the observer immediately upon his return from sea. During this review, attempts are made to identify suspect data through discussions with the observer about the details of each purse-seine set. All data are subsequently checked with computer programs to identify numerical inconsistencies, and all inconsistencies are reviewed with the observer. A final data review is then conducted by a different group of IATTC staff (without the observer present). We believe it unlikely that there would be widespread misreporting that went unnoticed during all three stages of the data review process. Second, the most useful predictors, as well as the direction of predictor effects, are consistent with results of analyses conducted on data prior to initiation of the individual vessel limits in 1993 (see references in Lennert-Cody *et al*., 2004). Finally, based on rough estimates of the number of sets that might be affected by misreporting, we would expect only a few percent or less of dolphin sets to have misreported mortality data. While a small amount of misreporting creates major data analysis challenges, the random forest results are not likely to be materially affected.

4.3 Identifying unusual observations

Because available information suggests that misreporting of incidental mortality of dolphins is most likely to occur as an under-reporting of mortality, we focus on identifying unusual *negative* "residuals" derived from the random forest classifier applied to the test data set. The observed value of the response is the class (*i.e*., 0 = no reported mortality; 1 = reported mortality). For the i^{th} set, we can define a "residual," r_i, to be the difference between the observed response (*i.e*., 0 or 1) and p_i, which is the proportion of trees that correctly classify set i. The value of p_i represents the stability of the classification over random samples of the data and random samples of predictors. A larger value of p_i implies that the classification is robust.

The residuals have a bimodal distribution (Fig. 4). Values of r_i between -0.5 and 0.5 correspond to sets that would have been considered correctly classified according to a majority vote rule. Large positive values of r_i are for those sets for which mortality was observed, but p_i was considerably less than 1.0. For these residuals, there was in fact reported dolphin mortality but the set cannot be classified in a stable manner. Large negative values of r_i are for those sets for which no mortality was reported, but p_i was considerably greater than 0. Here, there is a strong tension between what was reported and the classification; the observer reported no mortality but random forests confidently claims otherwise.

How large do the negative residuals have to be before they should be considered unusual? Under the assumption that the distribution of positive residuals is not compromised by misreporting, the percentiles of the right-hand tail of the positive residuals can be used to define "unusual" for the left-hand tail of the negative residuals. To remove any differences in scale between the positive and negative residuals, we first standardize by dividing the negative residuals by their mean and standard deviation, and do the same for the positive residuals (Fig. 5). Using the ".05 level" as a threshold, for example, the 95^{th} percentile of the standardized positive residuals corresponds to a value of 1.78. Then, the value of –1.78 is applied to the negative residuals. If the positive and negative residuals have the same distribution, 5% of the negative residuals would have a value of -1.78 or less. In fact, the negative residuals are a bit more skewed because the region to the right of –1.78 contains 6.1% of the negative residuals. This is not a large difference, but under the assumption that systematic under-reporting of dolphin mortality is relatively rare, a large difference is not likely.

Use of the uncompromised tail of the distribution of residuals to define the percentiles for the compromised tail will be most appropriate when individual observations are the focus of the data screening. However, in our example, the oversight issue is less a matter of suspect data on individual sets and more about suspect observers who may have provided suspect data on several sets. Our observations are nested within 209 fisheries observers, and because some observers may be more inclined to misreport than others, this nesting needs to be taken into account. If observers all went to sea on the same number of vessel trips per year, and observed the same number of sets, the number of unusually negative residual per observer could be an indicator of the degree to which individual observers may be misreporting data. However, observers do not necessarily make the same number of trips per year, and the number of dolphin sets per trip is variable. Thus, it is necessary to consider the number of unusually large negative residuals within observers, taking the number of sets per observer into consideration.

To this end, we employ the binomial distribution within observers, for those observers with at least one unusually negative residual. For each such observer, we compute the probability, q, of obtaining r large negative residuals (*i.e.*, "successes"), in n sets (*i.e.*, "trials"), using an assumed binomial probability of .055 per trial. This binomial probability is the proportion of all residuals that were unusually negative, computed as the number of residuals less than or equal to -1.78 divided by the total number of residuals. If all observers were equally likely to have negative residuals falling within the region we have defined as suspect, the probability of having such residuals would be .055. Large values of r, relative to n, could suggest a systematic tendency to under-report dolphin mortality. We focus on those observers whose overall proportion of unusually negative residuals (*i.e.*, r/n) exceeds the assumed binomial probability of 0.055.

Because q is related to n, and because we have a range of values of n, from several sets to several hundred sets per observer, we group the values of q by intervals of the values of n in order to compare q among observers. With the exception of the first interval, interval boundaries were selected to accommodate the sparseness of the values of n, while not allowing too large of a range of probabilities of the mean count within each interval. Intervals grossly represent the range of n divided into quintiles. With the exception of the first interval, the range of probabilities of obtaining the mean count within each interval varied by 0.10 or less. (Dividing the first interval into two sub-intervals ($n \leq 11$ and $n > 11$) gave similar results.)

We note that the residuals we are using (like true residuals) build in a bit of dependence, and in any case, there is no convincing way to determine if, within observers, the reports are independent. However, the q need not be taken literally. The values of q can be used as a descriptive measure with which to identify a relatively small number of observers who fall some distance away from the rest.

Comparing the q among observers illustrates the point that the data of some observers are, relatively speaking, more suspect than those of other observers (Fig. 6). Those observers with the smallest values of q within each interval of n would be considered the most suspicious. Thus, further analysis might start by focusing on those individuals with values of q falling within the lower whiskers of each group. In this example, there are no true outliers of q in the classical sense. However, we note that this is not necessarily a problem. A lack of true outliers would be consistent with varying degrees of misreporting.

We do not know for certain that large values of r, relative to n, indicate misreporting; however, the limited ancillary data available support its use for screening. Over the last decade, for a small group of observers ("problem" group), misreporting has been either identified through outright confession or strongly suggested by way of reports and evidence provided by other observers and by fishermen. A quantile-quantile plot of the proportion of unusually negative residuals, including zero-valued proportions, by observer for the general pool observers that went to sea between 1993 and 2002, and this problem group (Fig. 7), shows that the distribution of the problem group is generally shifted toward larger values. This supports the interpretation that larger numbers of unusually negative residuals, relative to total residuals, may signal problem observers.

5. Concluding comments

In this manuscript we have presented an application of random forests to the problem of identifying suspect data in the two-group classification problem. Methods for identifying anomalous data were proposed, based on "residuals" computed from the observed class (0 or 1) and the proportion of trees in the random forest that correctly classified the observation. This method will be applicable to any data screening problem that can be cast in a presence/absence context (*e.g.*, occurrence of values above or below a certain threshold).

Development of techniques of this type is particularly important for use in applications in which obtaining control data is logistically infeasible and/or prohibitively expensive. Environmental monitoring, and fisheries applications in particular, are examples. In the case of fisheries, management often involves the use of fishery-dependent data that may be provided directly by the fishermen in the form of logbooks. Under-reporting of bycatch can occur because of misreporting, or because during peak fishing times, accurate documentation of bycatch has a low priority for fishermen (Walsh *et al*., 2002). Data quality becomes a concern when these data are used to assess population status, and compliance with regulations such as closures and limits. The costs of obtaining more reliable data using at-sea observers can be high and dependent on vessel size; smaller vessels may have no room to accommodate observers, particularly on multi-day fishing trips. Moreover, as our example suggests, fisheries observer data are not without their problems. Electronic monitoring is being used in limited cases, but while typically less expensive than at-sea observers, can still be costly, and has limited scope.

Acknowledgements
We thank Robin Allen, William Bayliff, Richard Deriso, Martín Hall and Michael Scott for reviewing this manuscript, and Nickolas Vogel for data base assistance.

References

AIDCP (1999) Agreement on the International Dolphin Conservation Program. [Available from: http://www.iattc.org].

Allen, R.L. (1985) Dolphins and the purse-seine fishery for yellowfin tuna. In: *Marine Mammals and Fisheries*, Beddington, J.R., Beverton, R.J.H. and Lavigne, D.M. (eds). pp. 236-252. London: George Allen & Unwin.

Archer, F., Gerrodette, T., Dizon, A., Abella, K., and Southern, S. (2001) Unobserved kill of nursing dolphin calves in a tuna purse-seine fishery. *Marine Mammal Science* **17**, 540-554.

Bayliff, W.H. (2001) Organization, functions and achievements of the Inter-American Tropical Tuna Commission. Special Report 13. Inter-American Tropical Tuna Commission, La Jolla, California, U.S.A. 92037. [Available from: http://www.iattc.org].

Berk, R.A. (2005) An introduction to ensemble methods for data analysis. *Sociological Methods and Research*, forthcoming.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984) *Classification and Regression Trees*. Chapman and Hall.

Breiman, L. (2001) Random forests. *Machine Learning Journal* **45**, 5-32.

Friedman, J.H. (2002) Stochastic gradient boosting. *Computational Statistics and Data Analysis* **38**, 367-378.

Gosliner, M.L. (1999) The tuna-dolphin controversy. In: *Conservation and management of marine mammals*, Twiss, J.R. Jr. and Reeves, R.R. (eds). Smithsonian. pp. 120-155.

Hall, M.A. (1998) An ecological view of the tuna-dolphin problem: impacts and trade-offs. *Reviews in Fish Biology and Fisheries* **8**, 1-34.

Hall, M.A., García, M., Lennert-Cody, C., Arenas, P. and Miller, F. (1999) The association of tunas with floating objects and dolphins in the eastern Pacific Ocean: A review of the current purse-seine fishery. pp. 87-194. In: Special Report 11. Inter-American Tropical Tuna Commission, La Jolla, California, U.S.A. 92037.

IATTC (2002) Annual Report of the Inter-American Tropical Tuna Commission, 2001. Inter-American Tropical Tuna Commission, La Jolla, California., U.S.A. 92037. [Available from: http://www.iattc.org].

IATTC (2004a) Fishery Status Report No. 2. Inter-American Tropical Tuna Commission, La Jolla, California., U.S.A. 92037. [Available from: http://www.iattc.org].

IATTC (2004b) Annual Report of the Inter-American Tropical Tuna Commission, 2003. Inter-American Tropical Tuna Commission, La Jolla, California., U.S.A. 92037. [Available from: http://www.iattc.org].

Joseph, J. (1994) The tuna-dolphin controversy in the eastern Pacific Ocean: biological, economic and political impacts. *Ocean Development and International Law* **25**, 1-30.

Lennert-Cody, C.E., Minami, M., and Hall, M.A. (2004) Incidental mortality of dolphins in the eastern Pacific Ocean purse-seine fishery: correlates and their spatial association. *Journal of Cetacean Research and Management* **6**, 151-163.

Lo, N.C.H. and Smith, T.D. (1986) Incidental mortality of dolphins in the eastern tropical Pacific, 1959-72. *Fishery Bulletin* **84**, 27-34.

National Research Council. (1992) *Dolphins and the tuna industry*. National Academy of Sciences, Washington D.C.

NMFS (2002) Report of the Scientific Research Program Under the International Dolphin Conservation Program Act. Southwest Fisheries Science Center, National Marine Fisheries Service. 17 September 2002. [Available from: http://swfsc.nmfs.noaa.gov/IDCPA/TunaDol_rep/].

Pryor, K. and Shallenberger, I.K. (1991) Social structure in spotted dolphins (*Stenella attenuata*) in the tuna purse seine fishery in the eastern tropical Pacific. Pp. 161-196. In; K. Pyror and K.N. Norris (eds.) *Dolphin Societies: Discoveries and Puzzles*. Berkeley: University of California Press.

Schramm, Y. (1997) Activity level of offshore spotted (*Stenella attenuata*) and eastern spinner dolphins (*S. longirostris*), during purse seining in the eastern Pacific Ocean. Master of Science Thesis, Universidad Autónoma de Baja California, Ensenada, Baja California, México.

Wade, P. R. (1994) Abundance and population dynamics of two eastern Pacific dolphins, *Stenella attenuata* and *Stenella longirostris orientalis*. Ph.D. dissertation, University of California, San Diego.

Wade, P.R. (1995) Revised estimates of incidental kill of dolphins (Delphinidae) by the purse-seine tuna fishery in the eastern tropical Pacific, 1959-1972. *Fishery Bulletin* **93**, 345-354.

Walsh, W.A., Kleiber, P., McCracken, M. (2002) Comparison of logbook reports of incidental blue shark catch rates by Hawaii-based longline vessels to fishery observer data by application of a generalized additive model. *Fisheries Research* **58**, 79-94.

Table 1. Predictors used in the analysis of mortality presence/absence data. For categorical variables, the number of levels in shown in parentheses. More details can be found in Lennert-Cody *et al*. (2004) and references therein.

Predictor	Type	Description
Environmental variability		
Latitude	Continuous	Decimal degrees; including latitude^2 and latitude^3.
Longitude	Continuous	Decimal degrees; including longitude^2 and longitude^3.
Season	Categorical (3)	January-April; May-August; September-December.
Year	Categorical (10)	1993-2002.
Strong currents	Categorical (2)	Presence/absence.
Sea state	Continuous	Beaufort scale.
Fishing area	Continuous	Historical fishing activity (number of dolphin sets, 1959-1992, by 5º square area).
Operational problems		
Gear malfunctions	Categorical (3)	No gear malfunctions; minor malfunctions; major malfunctions. Major malfunctions directly affect the ability of captain and crew to conduct dolphin rescue and release procedures.
Net collapses	Categorical (3)	Absence; presence of unintentional net collapses; presence of intentional net collapses. Net collapses occur when opposite sides of the purse-seine come together.
Net canopies	Categorical (2)	Presence/absence of net canopies. Net canopies are billows of webbing that form along the sides of the net, typically at the surface.
Dolphin safety panel	Categorical (2)	Presence/absence of adequate coverage of the backdown channel.
Captain skill	Categorical (2)	Limited; more extensive. Determined from the captain's previous fishing activity (average number of dolphin sets per year for the previous three years): <30 dolphin sets per year; ≥ 30 dolphin sets per year.
Noisemakers	Categorical (2)	Presence/absence of noisemakers used to direct the herd of dolphins.
Multiple sets	Categorical (2)	Presence/absence of repeat sets made on this herd of dolphins.
Quota utilization	Categorical (4)	Percentage of the vessels annual mortality quota utilized to date: not applicable; 0-5%; 5-25%; >25%.
Fishing operations		
Start time of the set	Continuous	Local time of the release of the net skiff.
Night sets	Categorical (2)	Presence/absence adequate day light to finish the set.
Duration (decimal hours) of:		
Approach	Continuous	Time between initial sighting and release of the first speed boat.
Chase	Continuous	Time between the release of the first speed boat and release of the net skiff.

Encirclement	Continuous	Time between the release of the net skiff and the point at which the bottom of the net. is pursed and the rings are above the water (“rings up”).
Net retrieval	Continuous	Pre-backdown net retrieval. Time between “rings up” and the start of the “backdown”.
Backdown	Continuous	The procedure used to release dolphins from the net.
Rescue equipment use during backdown		
Speedboats	Categorical (3)	None; dolphin rescue by speedboat crew; dolphin rescue by speedboat crew with mask.
Rafts	Categorical (3)	None; dolphin rescue with raft; dolphin rescue with raft and mask.
Swimmers, divers	Categorical (4)	None; dolphin rescue by swimmers without mask; dolphin rescue by swimmers with mask; dolphin rescue by S.C.U.B.A. divers.
After-backdown rescue	Categorical (5)	After conducting the backdown procedure to release dolphins: none; dolphin rescue by swimmers without mask; dolphin rescue by swimmers with mask; dolphin rescue by S.C.U.B.A. divers; other rescue.
Biomass characteristics		
Number of dolphins	Continuous	Number of animals in the purse-seine net once the net has been pursed.
Dolphin species	Categorical (4)	Species of dolphin encircled: spotted dolphins, spotted and spinner dolphins or spinner dolphins alone, common dolphins; other dolphin species involved (possibly including spotted, spinner and common dolphins).
Tons of tuna:		
Yellowfin tuna; Other tunas	Continuous	Tons of tuna, yellowfin and other tunas (metric tons), in the purse-seine net once the net has been pursed.
Herd cohesion:		
At chase; at encirclement	Categorical (3)	Cohesiveness of the herd at the time of chase and at the time of encirclement: all animals in one group; animals in several groups; animals in many groups.

Table 2. Misclassification rates (“confusion tables”) for the random forest with unaltered priors (A), and the random forest with altered priors (B).

(A)

	Predicted		
Reported	*0*	*1*	*Misclassification rate*
0	30,700	345	0.01
1	3,190	783	0.81

(B)

	Predicted		
Reported	*0*	*1*	*Misclassification rate*
0	27,544	3,501	0.11
1	1,733	2,195	0.44

Figures

Figure 1. Numbers of purse-seine sets on tunas associated with dolphins, by 1º square area, 1993-2002.

Figure 2. Frequency distribution of incidental dolphin mortality per set (mps) for sets with reported mortalities, 1993 and 2002.

Figure 3. Changes in random forest prediction accuracy associated with the 20 most important of 36 predictors for class "1" (presence of dolphin mortality). (Variables are defined in Table 1.)

Figure 4. Frequency distribution of residuals, r_i, from random forest classification.

Figure 5. Frequency distributions of standardized values of r_i.

Figure 6. Boxplots of values of observer-specific probabilities (q), grouped according to values of the total number of residuals per observer (n).

Figure 7. Quantile-quantile plot of the proportion of unusually negative residuals, by observer, for the general observer pool and the problem group. The solid line indicates the one-to-one line.

Figure 1.

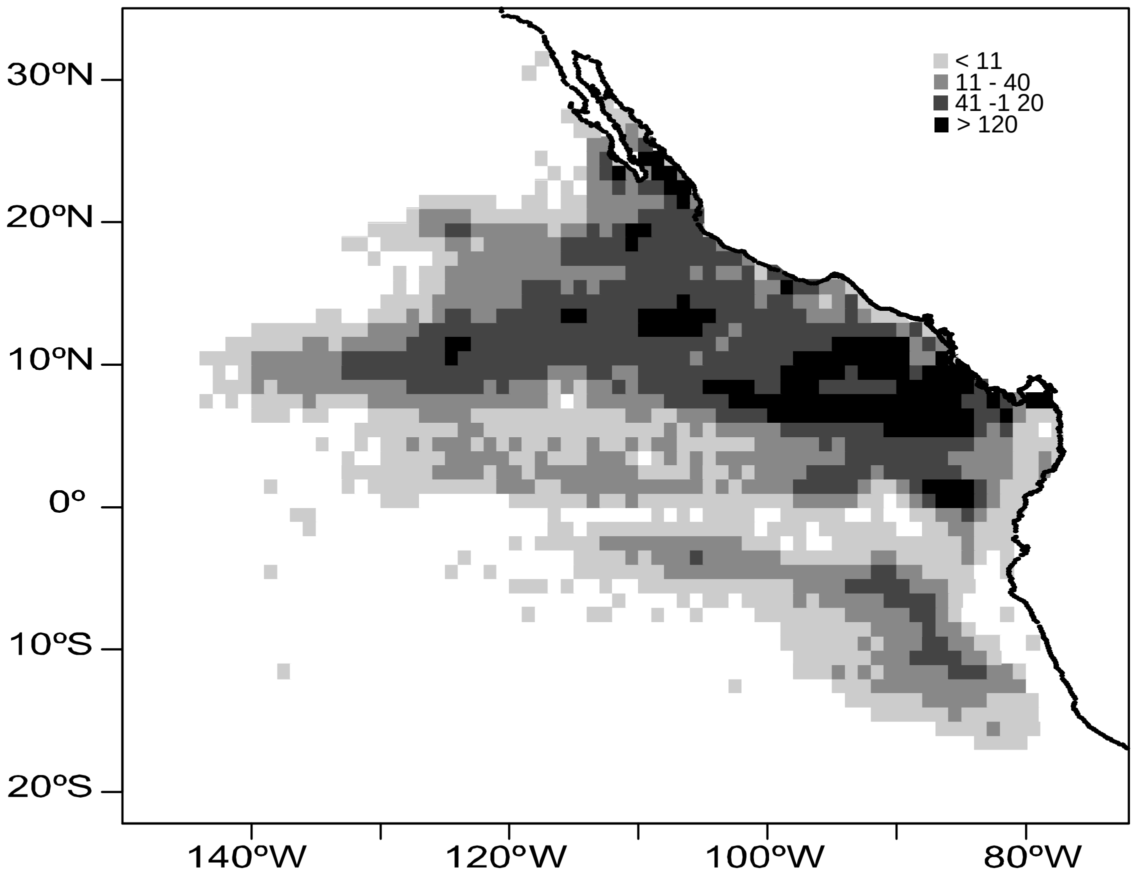

< 11
11 - 40
41 -1 20
> 120
30ºN
20ºN
10ºN
0º
10ºS
20ºS
140ºW
120ºW
100ºW
80ºW

Figure 2.

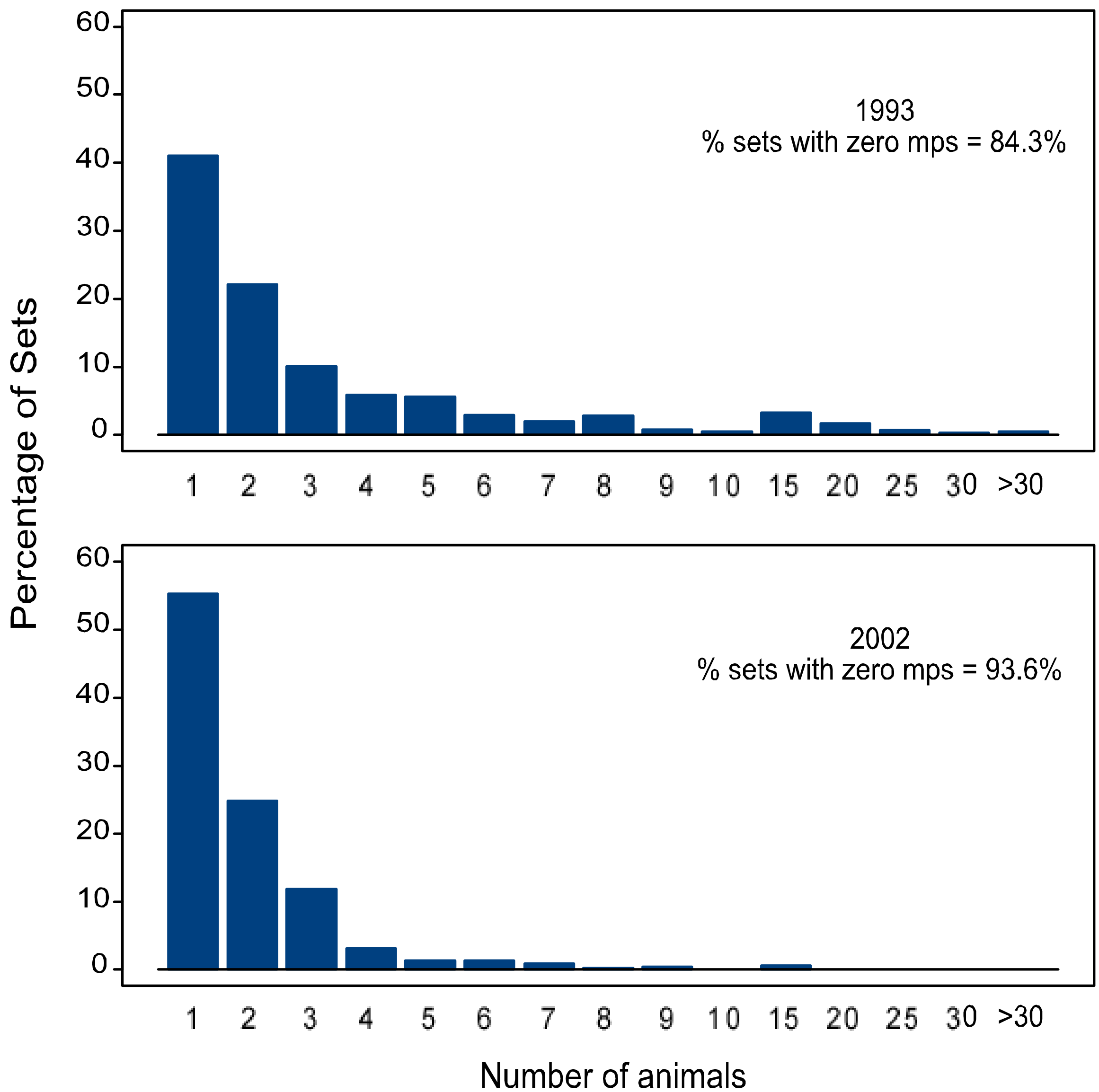

Percentage of Sets
1993
% sets with zero mps = 84.3%
2002
% sets with zero mps = 93.6%
0
10
20
30
40
50
60
1
2
3
4
5
6
7
8
9
10
15
20
25
30
>30
Number of animals

Figure 3.

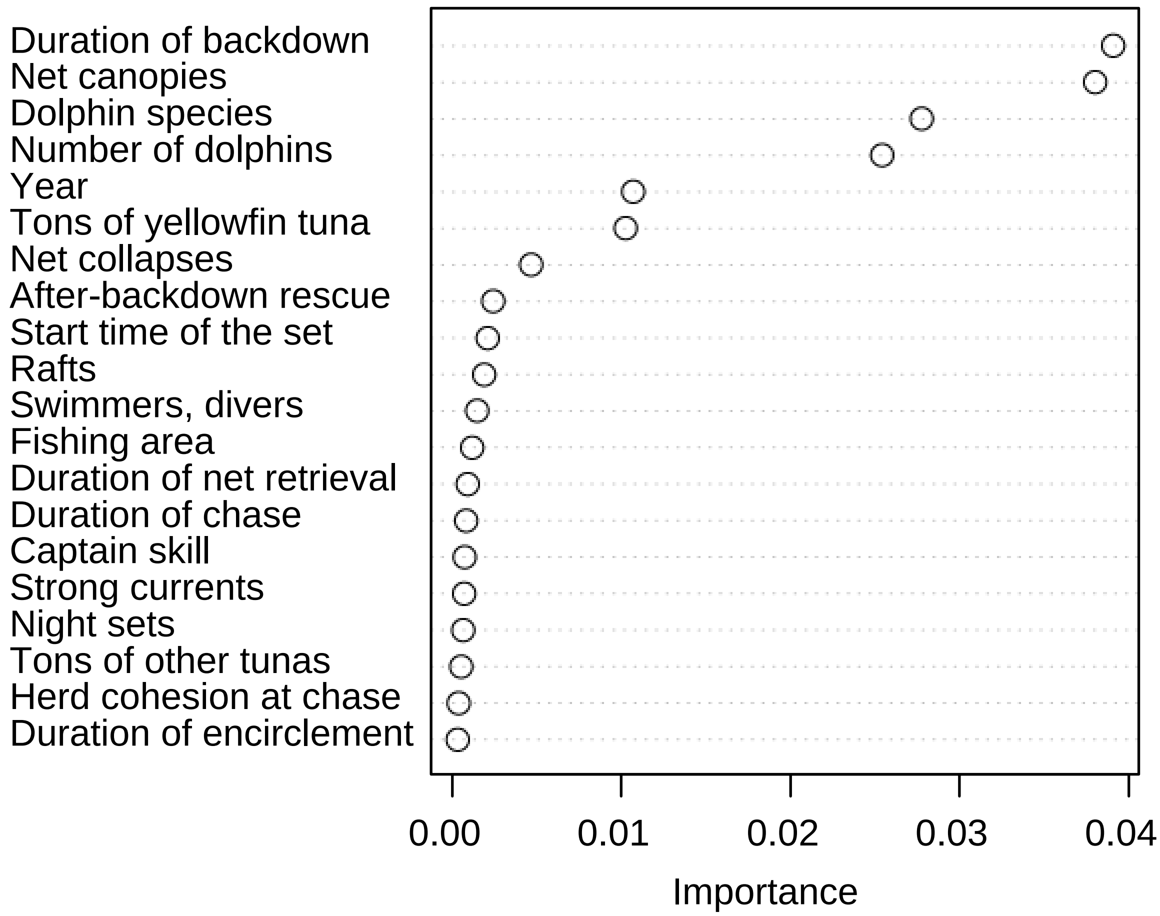

Duration of backdown
Net canopies
Dolphin species
Number of dolphins
Year
Tons of yellowfin tuna
Net collapses
After-backdown rescue
Start time of the set
Rafts
Swimmers, divers
Fishing area
Duration of net retrieval
Duration of chase
Captain skill
Strong currents
Night sets
Tons of other tunas
Herd cohesion at chase
Duration of encirclement
0.00
0.01
0.02
0.03
0.04
Importance

Figure 4.

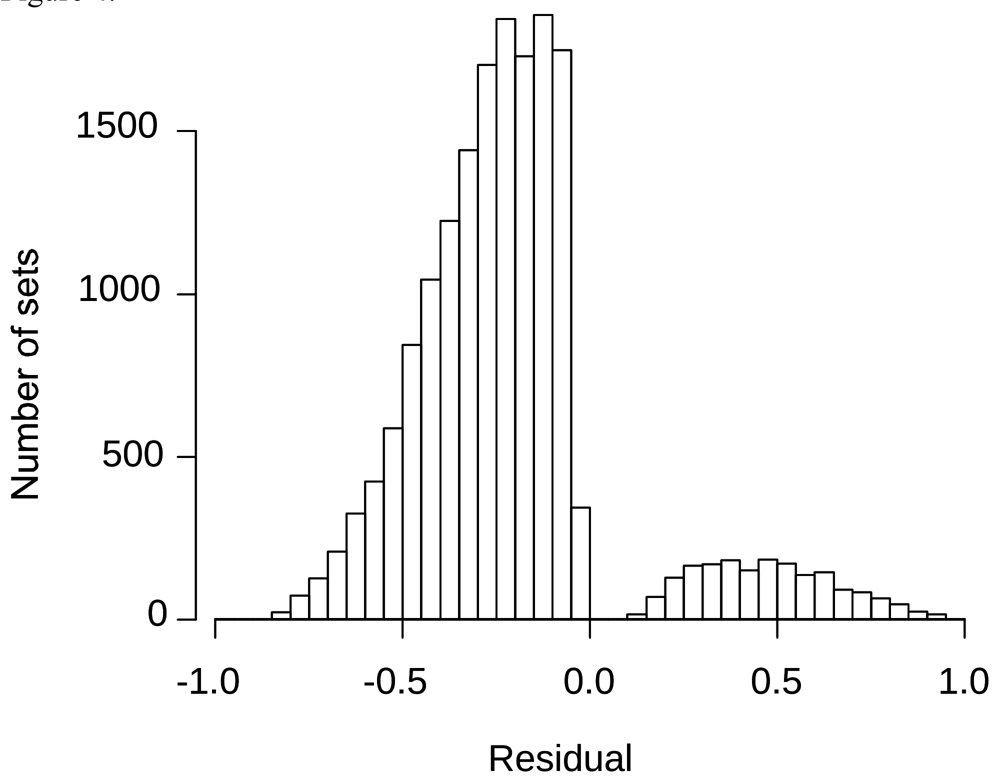
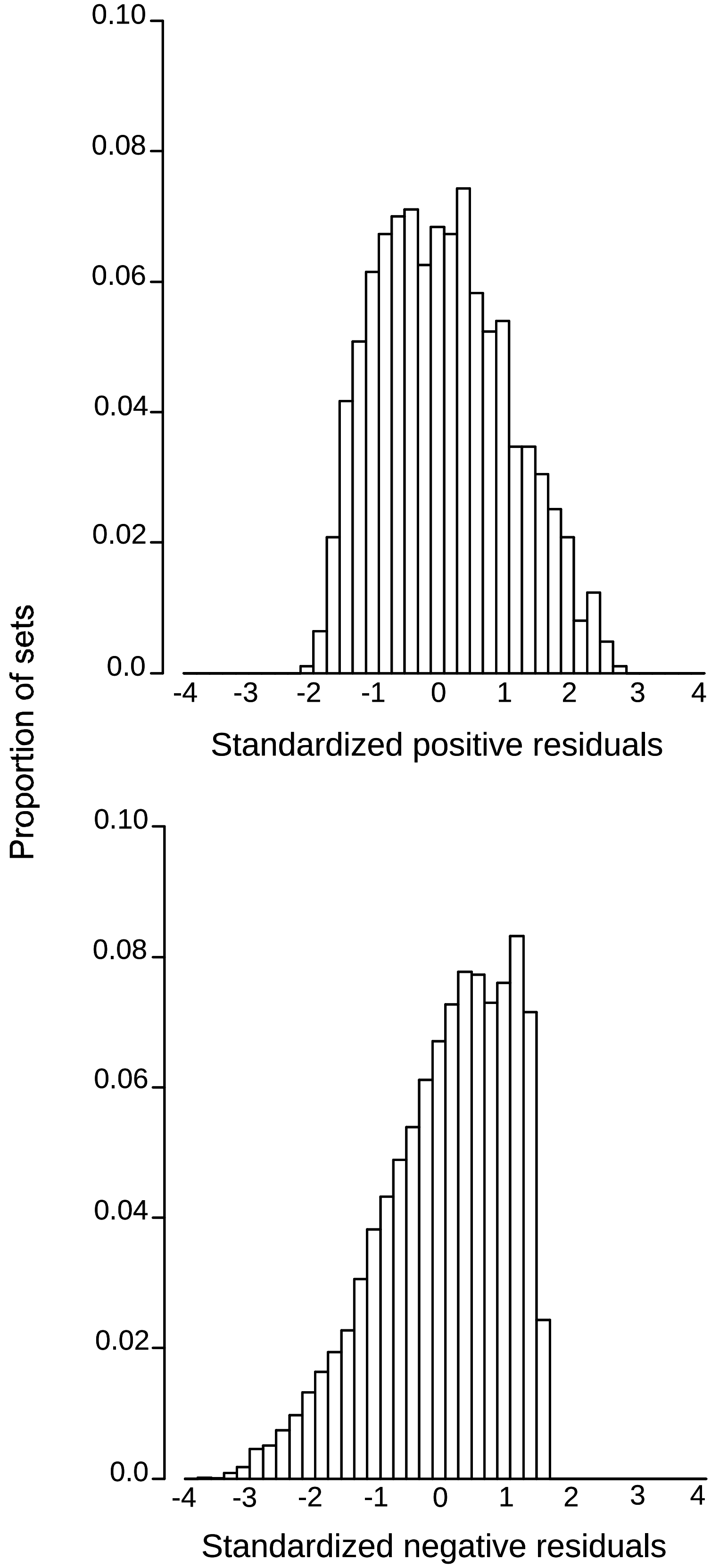
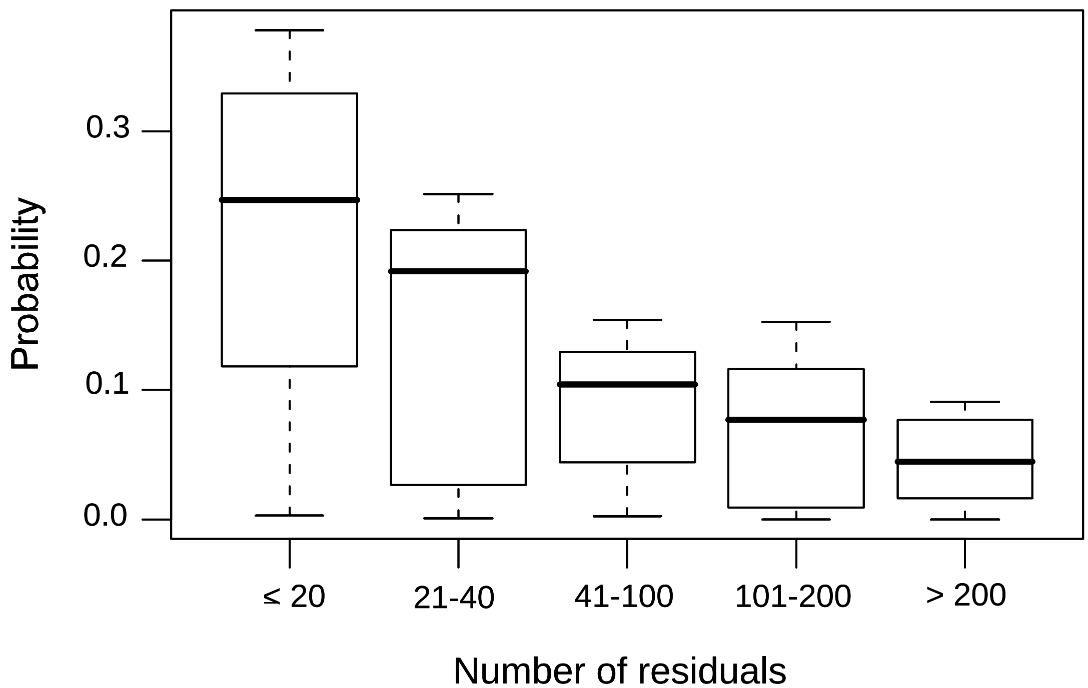


Figure 5.



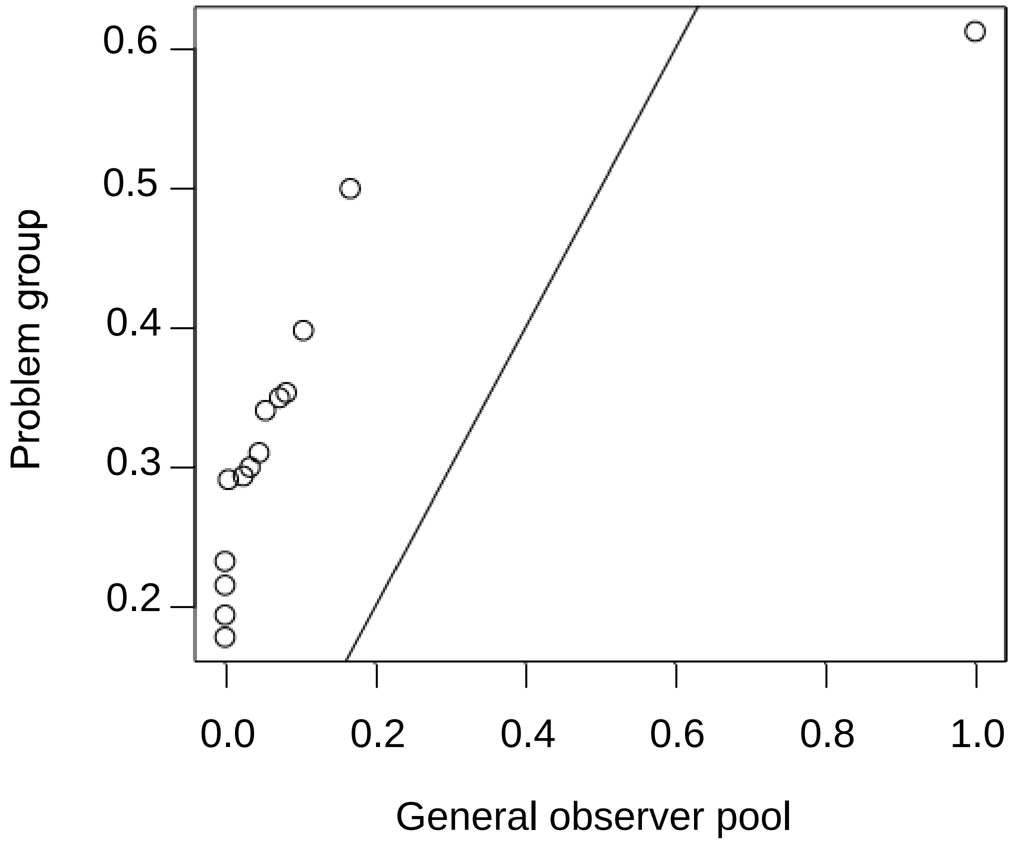
0.10
0.08
0.06
0.04
0.02
0.0
-4
-3
-2
-1
0
1
2
3
4
Standardized positive residuals
Proportion of sets
0.10
0.08
0.06
0.04
0.02
0.0
-4
-3
-2
-1
0
1
2
3
4
Standardized negative residuals

Figure 6.



0.3
0.2
0.1
0.0
Probability
≤ 20
21-40
41-100
101-200
> 200
Number of residuals

Figure 7.


0.6
0.5
0.4
0.3
0.2
Problem group
0.0
0.2
0.4
0.6
0.8
1.0
General observer pool