North Pacific Fisheries Commission

NPFC-2021-SSC PS07-WP20

Outline of the protocol for catch at size (CAS) data for Russian stick-held dip net fishery

Vladimir Kulik, Dmitriy Antonenko, Aleksei Baitaliuk, and Oleg Katugin

Pacific branch of the Russian Federal Research Institute of Fisheries and Oceanography (TINRO)

1. Introduction

NPFC-PS members need to submit the catch at size (CAS) data that is necessary for developing age-/size- structured stock assessment model. This report describes the outline of protocol to make the CAS data for Russian stick-held dip net fishery.

2. Data set preparation

Original length frequencies are available in TINRO from Russian scientific observers who collected that information on board motherships and fish factories, along with data on exact location and date of catch, as well as catch weight, if possible, in metric tons (mt), and vessel name from the logbooks of those vessels. Each sample included fork-length measurements of at least 100 individuals of fish with 1 cm precision for length measurement. The average weight of fish in a sample was calculated with 0.1 g precision. There were 4,257 samples collected (Fig. 1) which included more than 739 thousand of fish measured.

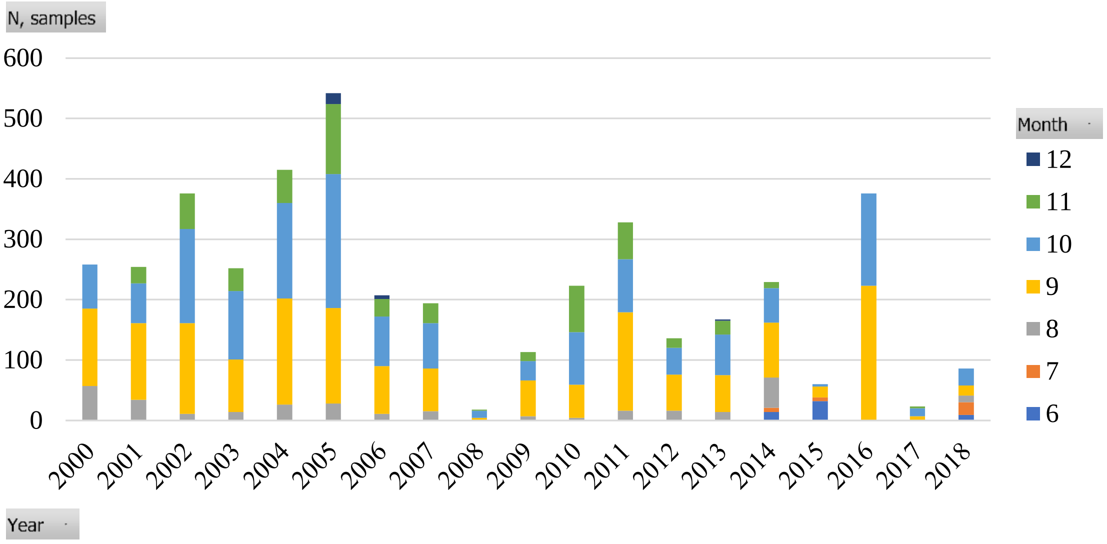

Figure 1 – number of samples of saury fork length measurements by Year and Month collected by Russian scientific observers

Interquartile range was from 100 to 223 fish in a sample.

Unfortunately, 1,606 out of 4,257 sample records did not include average weight of fish. Therefore, we imputed average weight for those records using Year, Month, and coordinates of catch as covariates in machine learning methods. We tested Random Forest or RF (Liaw, Wiener, 2002) and Stochastic Gradient Boosting or GBM (Greenwell et al., 2020) on 25% out-of-bag samples in caret package (Kuhn, 2021) for R programming language (R Core Team, 2021). Optimization of hyperparameters was conducted by Cross-Validation with 10 fold, repeated 5 times seeking for the lowest RMSE following "one standard error" rule of Breiman et al. (1984). The winner was RF (mtry=2) with RMSE = 15.5 ($r^2 = 0.64$, MAE = 10.8), which had significantly (Bonferroni $p < 0.001$) higher coefficient of determination and significantly (Bonferroni $p < 0.001$) lower error metrics than GBM had. The highest variable importance was found for the Year. Longitude contributed around 1.5% comparing to the Year, Month contribution was around 0.65% while latitude contribution was close to 0%.

The same approach was taken for imputation of missing catch volume records (849 out of 4,257). Again, the winner was RF, (mtry=2) with RMSE = 35.4 ($r^2 = 0.22$, MAE = 22.7). The highest variable importance was found for the Year. Longitude contributed around 24.2% comparing to the Year, Month contribution was around 10.7% while latitude contribution was close to 0%.

We imputed average weight of fish (g) or catch volume (mt) for those cells where they were not available using RF. Thus, we have got a dataset where for each sample out of 4,257 rows the exact location, date, length frequencies by 1 cm interval and exact or estimated catch volume and average weight of fish were known, but 23 records were dropped due to unknown source (vessel name) or other unrecoverable losses of information. Therefore 4,234 rows were kept in the final dataset.

3. Procedure for estimation of CAS

The length frequency in each fishery area in each period was assumed to be homogeneous. Then, we estimated the CAS data using the following procedure.

1) Total number of fish (Nc) in each sample was calculated as

N_c=1e+06C/W,

where N_c – total fish number (individuals) in catch, C – catch weight (mt), W – average weight of fish in a sample (g).

2) Estimation of fish quantity from length frequency ($N_{c,j,i}$) in catch for each sample

$N_{c,j,i} = N_j N_c / N_s$,

where N_j – number of fish (individuals) in a sample at 1 cm length intervals, N_s – total number of fish measured in each sample.

3) Estimation of fish quantity by length frequency ($N_{c,Y,M,R,j}$) in catch for each period and region

$$N_{c,Y,M,R,j} = \sum_{i=1}^{n} N_{c,j,i}$$

where $N_{c,Y,M,R,j}$ – number of fish in a given Year (*Y*), Month (*M*), Region (*R* for 1x1 degree cell) and 1 cm length interval (*j*) as a sum from all samples (n) extrapolated to their catch.

4. Results

The result of calculation for the time period from 2000 to 2018 has been shared among the NPFC Members through the collaboration facility (https://collaboration.npfc.int/comment/249). After 2018, there were no scientific observations on length frequency data.

5. References

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

Brandon Greenwell, Bradley Boehmke, Jay Cunningham and GBM Developers (2020). gbm: Generalized Boosted Regression Models. R package version 2.1.8. https://CRAN.R-project.org/package=gbm

Max Kuhn (2021). caret: Classification and Regression Training. R package version 6.0-88. https://CRAN.R-project.org/package=caret

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Breiman, Friedman, Olshen, and Stone. (1984) Classification and Regression Trees. Wadsworth.