# Bootstrapping in StoX

StoX is a free and open-source application primarily suited for calculating survey estimates from acoustic-trawl and swept-area surveys. The application is actively used to produce advice for several important European fish stocks, such as herring, mackerel, cod, haddock, sandeel and blue whiting ([Johnsen et al, 2019](https://doi.org/10.1111/2041-210X.13250)).

StoX is designed to support precision estimation, primarily through bootstrapping, where the baseline estimate is re-calculated many times while resampling the data in specific ways. The precision is reported as sample standard deviation and coefficient of variation (sample standard deviation divided by sample mean) of the survey estimates over the bootstrap replicates. Bootstrapping for swept-area and acoustic-trawl survey estimates is described in this document.

## List of expressions and definitions used in StoX

### General definitions in StoX

* **StraumPolygon**: A set of geographical multi-polygons named **Stratum**, inside which average densities are calculated and multiplied by the area of each Stratum to obtain abundance.
* **Survey**: The Strata can be grouped into one or more Surveys, e.g. excluding some strata that were used for experimental purposes.
* **SuperIndividualsData**: The estimated abundance of a Stratum can be distributed to the individual fish used to obtain the estimate, resulting in super-individuals which represent the biotic sampling of e.g. length and age, but are weighted by the abundance.
* **Imputation**: The super-individuals may contain missing values of e.g. weight or age, and the ImputeSuperIndividuals function imputes the missing information. There are two imputation methods available, the RandomSampling method and the Regression method.

The RandomSampling method draws information randomly from individuals of e.g. the same length group, first from the Haul, then from the Stratum if no adequate individuals are found in the Haul, and finally from the entire Survey if no adequate individuals are found in the Stratum.

The Regression method imputes data using a regression model.

### Survey estimation types in StoX

* **Swept-area survey estimate**: Trawl samples are normalized by dividing by tow distance and sweep width to obtain area number density of e.g. fish per length group.
* **Acoustic-trawl survey estimate**: Acoustic backscatter is converted to area density per fish length group by use of species-specific acoustic properties and representative fish length distribution (the AcousticDensity function).

### Definitions for swept-area estimation

* **Station**: A trawl station, represented by time and a geographical position.
* **Haul**: A trawl haul, specified by the gear (type, duration, haul depth, etc.). One Station can include several Hauls.
* **BioticPSU**: The biotic primary sampling unit, which in principle can be collections of Stations. In the current StoX 3.5.0 it is however only possible to have one Station per BioticPSU. BioticPSUs are linked to the Stratum containing the Stations.
* **LengthDistributionData**: The frequency distribution of individual fish lengths of a Haul. For swept-area estimates the LengthDistributionData are normalized by tow distance.
* **MeanLengthDistributionData**: The LengthDistributionData summed over all hauls of a Station and averaged over all Stations of a BioticPSU. If there is only one Haul per Station (and only one Station per BioticPSU, as in the current StoX 3.5.0) the MeanLengthDistributionData is essentially the same as the LengthDistributionData.

### Definitions for swept-area estimation

* **EDSU**: An acoustic log distance or Elementary Distance Sampling Unit, storing the average acoustic backscatter in e.g. 1 nautical mile intervals.
* **AousticPSU**: The acoustic primary sampling units, which are user defined sequences of EDSUs, typically transects. Each AcousticPSU is linked to a Stratum, but can include EDSUs from several Strata if deemed necessary by the user.
* **NASCData**: The acoustic backscatter distributed in depth channels.
* **MeanNASCData**: The NASCData summed over depth channels and averaged over the EDSUs of an AcousticPSU.
* **BioticAssignment**: Each AcousticPSU is assigned a set of Hauls from which length distribution (named **AssignmentLengthDistribution**) is derived and used for calculating density from the MeanNASCData.

## Bootstrapping of swept-area survey estimates

For swept-area survey estimates the bootstrapping is obtained by resampling *with replacement* BioticPSU within each Stratum in the MeanLengthDistributionData. Consequently the BioticPSUs (always Stations in the current StoX 3.5.0) are resampled 0, 1 or more times, and this produces a randomness in the survey estimates that reflects the variation in the input biotic data.

## Bootstrapping of acoustic-trawl survey estimates

For acoustic-trawl survey estimates the bootstrapping consists of the following two resampling processes:

First the AcousticPSUs are resampled *with replacement* within each Stratum in the MeanNASCData. This step incorporates the variation in the acoustic data.

Second, the Hauls assigned to the AcousticPSUs are resampled *with replacement* within each Stratum of the BioticAssignment, to incorporate the variation in the biotic data. In this resampling all Hauls assigned to one or more AcousticPSU in a Stratum are resampled.

There are a few pitfalls with this specific method of resampling assigned Hauls:

### Un-sampled Haul problem

If an AcousticPSU is assigned only a subset of all assigned Hauls of a stratum, there is a positive probability that none of the Hauls assigned to the AcousticPSU are sampled. The probability of this happening increases with smaller subset of assigned Hauls. If none of the assigned Hauls of an AcousticPSU are resampled for a bootstrap replicate, the density for all length groups will be missing (NA) for that AcousticPSU. The NA will then propagate to the abundance of the corresponding stratum, and ultimately to the reports from the bootstrapping (e.g. total abundance over all Strata, averaged over all bootstrap replicates). StoX has the option to remove NAs when reporting from the bootstrapping, but this will treat the NA abundance as 0, which can result in under-estimation!

A similar issue, with the same consequences, occurs if not all assigned Hauls of a stratum contain the target species to be used to calculate density from the acoustic data. In this case there is a positive probability that none of the Hauls with the target species are sampled.

### Failed random sampling imputation problem

Further, an issue specific to imputation with the RandomSampling method occurs when not all assigned Hauls of the Stratum contains the information needed by the imputation. E.g., if one wants to impute age within length groups, and there is only a subset of the Hauls that contain fish with age in specific length group, there is a positive probability that none of these Hauls are sampled, neither in the Stratum nor in the Survey. The imputation of age for fish in the specific length group will then fail, resulting in NA age for that bootstrap run. When reporting the bootstrap data by age, there will be abundance for NA age, thus reducing the abundance for the other present ages.

One solution to this problem is to include an additional imputation using the Regression method, to guarantee that all the desired information gets imputed. The Regression can either be defined with parameters, or estimated from the biotic data using the EstimateBioticRegression function.