

Quality assurance framework

Of the data collected through the DCF Programme

(version 1.0)

Introduction

To assure a high quality of the biological data collected through the DCF Programme, a wide range of validation techniques were implemented. Polish quality assurance framework is a multi-stage process. At first, data entered to the national database are verified in the two-stage validation process supported by a number of completeness, data type and range checks. Export procedures which prepare data sets for external databases (like RDB FishFrame or InterCatch) also perform basic checks.

Additionally, two validation applications were developed, both written in Shiny (R package) and available only via the institute's intranet: Data Quality Check Application and Data Accuracy Check Application.

Quality assurance framework

Database validation

Biological data collected under the DCF are entered into the database using a dedicated web application accessible only in the institute's network. The application provides a two-stage data registration procedure supported by a number of completeness, data type, and range checks. In the first step a user enters the data into the application forms and saves the data in temporary tables for further verification. In the second step a privileged user responsible for the specific species/stock can review, update and check the data. Once the data are approved, they are transferred from temporary tables to the target tables. All the application forms used for data registration are equipped with a drop-down lists of available values. The codes and values in these lists are regularly updated based on the RDB lookup tables and other available sources (EU Fleet Register, WoRMS, Master Data Register, etc.).

Applications for validation

Data Quality Check

The application was designed to enable quality checks of the data collected through the DCF Programme. It has been prepared for biological data from commercial fisheries. Application for quality checks of data from research surveys is under development.

The following visual and quantitative quality analyses of the data stored in the database, are available:

- outliers identification for Weight at Length relationship and Length at Age – a user can inspect the data visually on the scatter plots and mark suspicious points for further checking, or make use of the automatic outliers identification based on the Bonferroni outlier test,
- inconsistency between sample and catch weight,
- biological analyses with missing age – a table with detailed data, as well as a histogram of the number of gaps for all species, are available,
- inconsistency between number of individuals in the length classes and in the biological analyses,

- dates misreporting,
- mean weights of individuals in a sample.

A user can screen the data in the fully interactive mode or download a quality report in HTML/PDF format.

In 2018, the software for data quality checks was successfully deployed as a web application in the Institute's internal network. It is accessible for specialists responsible for species / stocks data analysis..

User manual

See Annex 1

Data Accuracy Check

At present only observer effect analysis is available. A user can display all VMS signals of a chosen vessel and highlight points from trips with observers on-board. The methodology used was based on the ICES WKACCU Report 2008, whereas the example of such analysis applied to the Polish data was performed during RCM Baltic 2016.

The application for data accuracy checks is still under development. Apart from the observer effect analysis which is already available, the following other types of checks were identified and are planned to be further developed and implemented in the near future:

- refusal reasons analysis,
- spatial and temporal coverage of sampling,
- incomplete sampling frame effects,
- random trips vs. expert judgement trips.

Summary

Both applications mentioned above are under constant development. National database and applications are accessible only within internal Institute's network.

Annex 1

Data Quality Check App

User Manual

Rejstry komercyjne

10 000 rekordów danych (ostatni dzień 2019 roku)

WYBIERZ WYBRANE ROKI

Rok: 2018

Kod gminy: 000

WYBIERZ

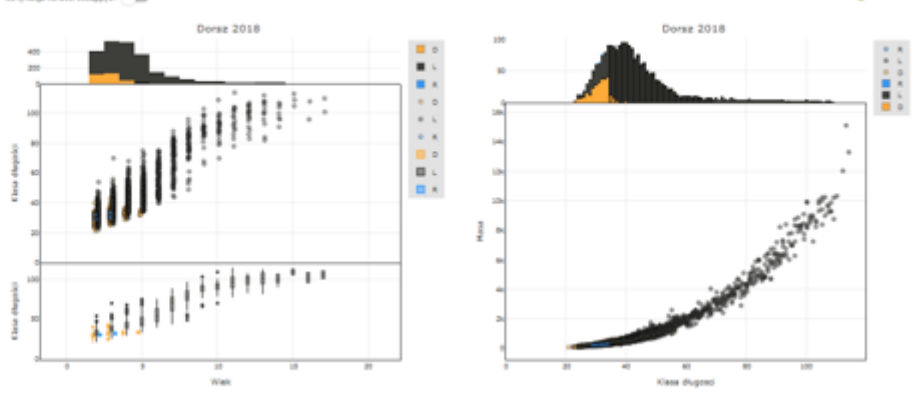
TABELA WYNIKI WARTOŚCI OBLICZANE

Okres: DZIŃ, MIESIĄC, KWARTAŁ, LATO

Analiza zależności wieku, długości i masy

Okres	Data Początek	Data Koniec	Okres analizy	Wiek analizy	Waga analizy	Wzrost analizy	Wzrost analizy	Wzrost analizy	Wzrost analizy	Wzrost analizy	Wzrost analizy	Wzrost analizy
1	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
2	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
3	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
4	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
5	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
6	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
7	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
8	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
9	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
10	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
11	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000
12	2018-01	2018-01	1	0	000	000	000	000	000	000	000	000

Wykresy: HISTORIA WARTOŚCI OBLICZANYCH



Technical Requirements

Recommended browsers are Google Chrome and Mozilla Firefox. Using Internet Explorer, might cause unavailability of some elements of the application.

Interactive mode

Sidebar panel

How do you want to analyse the data?
To start, choose the way you want to analyse the data.

Analysis by year	Choose this option if you want to analyse all the data from the given year.
Analysis by fishing trip	Choose this option if you want to analyse all the data from the given fishing trip.

Year
Choose a year to analyse the data from. Available years: 2004 - 2019



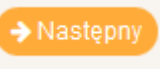




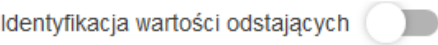
Trip number



This option will be available if you choose “Analysis of the fishing trip”. You shall enter a trip id here (usually four-number code). In case of inappropriate id, you will be informed about it.

Species

Choose a species you are interested in. If you don't define any, as a result you will get the summary for all species. If you choose a species, then you will get the summary only for the chosen one, and in addition in the Plots tab, there will be accessible age-length and weight-length plots. Moreover, you will be able to investigate outliers.

Buttons

Button	Description
	Help button. Includes a short description of possible actions in the given section.
	Search and edit. Press to start searching for the data in the database, or to edit the already set parameters. Yellow colour – active button, grey colour – inactive button. Depending on the data size, it may take a while after clicking the Search button, to get the result.
	Next. Button to go to the next species. You can either use this or select species manually from the drop down list. Having chosen the last species from the list and pressing the button, no species will be chosen.
	Buttons to copy and download a table in a chosen format. Bear in mind, while copying or downloading, if the table has more than one page, only the page that is currently visible, will be taken into account. Because of that, if you are interested in the whole table, before pressing any of these buttons, expand the number of rows to display by clicking 'All'.
	Deselect. If there are any selected rows in the table (they are also highlighted on the plot), you can cancel it by pressing this button.
	Select all. Press to select all the rows from the table. It is useful, when you have filtered out the table, e.g. by choosing only one vessel, and you want to highlight the filtered rows on a plot. So instead of choosing each row separately, you can use this button.
	Highlight outliers. Press to run the algorithm finding outliers for weight-length relationship. Having pressed the button, you will see some values highlighted on the plot – these will be the potential outliers. They should however not be treated as wrong values, but rather as an indicator of which data you should look into more deeply or consult with an expert.
	Outliers identification. Choose this if you want to start the process of identifying outliers on the plots. When the option is active, by clicking a point on the plot, you will automatically add it to the list of potential outliers. Afterwards you will be able to download the list, and check the values. If a chosen point is successfully added to the

	list, there will appear yellow message in the right bottom corner. If you choose the same point twice, a message will appear in the right bottom corner, saying that the point has already been chosen.
	Delete table. If your table of outliers contains some data, you can delete the whole table. What is important, you are only deleting the table from your view. You are not making any modifications in the database.
	Delete the selected rows. If your table of outliers contains some data, you can delete some rows by selecting them, and pressing the button.

Tabs

Tables

In this tab there are all the tables with errors.

Table	Error description
Sample weight vs catch weight	The sample weight is bigger than catch weight
Numbers at age and numbers at length	Number at length is smaller than numbers at age
Inconsistencies in the dates	Sample time does not fall between the start and end time of the trip
Numbers at length vs number of fish in a sample	Sum of numbers at length does not equal number of fish in a sample
Missing age	Set of analysis with missing age
Mean weight	Mean sample weight per species.

Plots





In this tab there are plots and tables with age-length and weight-length relationships.

Outliers

In this tab there is a table with values, that have been marked as potential outliers.

Tables

All the tables

It is possible to copy or download each table in a given format using the buttons:    

One can also search through the whole table using the field **Search:**

Notice that letter size matters. It is also possible to search through a given column with the field

Greyish field mean that there is only one value in a column. So there is no sense in searching through it. It is possible, in some columns, to use regular expressions (e.g. if you want to filter out all the records with Catch Category D or R, you can type `[D|R]` in the appropriate column. Using the arrows next to the column names, you can easily sort the column.

To increase the number of rows to display modify .

To go to another page of the table use

Age-length and weight-length tables

The interaction between table and plots is possible. If you select a row in the table, the appropriate value will be highlighted on the plot. If you want to highlight all filtered rows, either select row by

row, or use the button

If you want to cancel a row selection, click it once again and if you want to deselect all the rows use

Weight-length table

Additionally for weight-length table, you can use the built in algorithm of outliers detection. To do so,

use

Bear in mind, that it's only algorithm highlighting potential outliers, and you should always verify these values.

Plots

All the plots

Each plot is an interactive plot. In the right upper corner there is panel with various options like download as png, zoom in, zoom out, cut, move, etc. It is also possible to hide some elements by clicking on them in the legend part. Second click will display an element again. A double click will hide everything except from the clicked value.

Age-Length and weight-length plots

To make it easier to analyse the plot, the values have been slightly jittered to avoid overlapping of points. Except for that all the landings value are slightly moved right, all discard are slightly moved left to enable distinction between these categories.

To display some more information about a point, hover over it.

Above the plot there is a button (i.e. "Outliers **identification**").

It is inactive by default. To activate it, press the button. By doing that it will be possible to add points to the list of potential outliers, by clicking them on the plots. If a chosen point is successfully added to the list, a yellow message will appear in the right bottom corner. If you choose the same point twice, a message will appear in the right bottom corner, saying that the point has already been chosen.

There are also histograms and boxplots added to the both types of the plots.

Report Generator

The application gives also the possibility to generate a report with all the errors and plots. Having defined all the needed parameters you only have to wait a while and the *html* report should be ready.

Set parameters

Report type

Select what kind of a report you want to generate

Report for the year	Choose this option if you want to analyse all the data from the given year.
Report for the species	Choose this option if you want to analyse all the data for a given species.
Report for the fishing trip	Choose this option if you want to analyse all the data from the given fishing trip.

Species

This option will be available if you have chosen “Report for the species”.

Year

Choose a year to analyse the data from. Available years: 2004 - 2019

Trip number

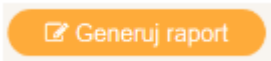
This option will be available if you have chosen “Report for the fishing trip”. You shall enter a trip id here (usually four-number code). In case of inappropriate id, you will be informed about it.

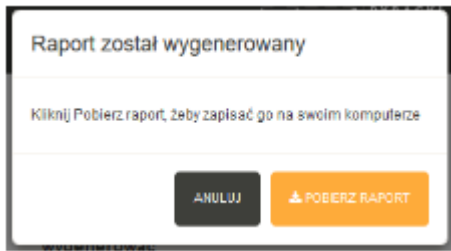
Analysis types

Table	Error description
Sample weight vs catch weight	The sample weight is bigger than catch weight
Numbers at age and numbers at length	Number at length is smaller than numbers at age
Inconsistencies in the dates	Sample time does not fall between the start and end time of the trip
Numbers at length vs number of fish in a sample	Sum of numbers at length does not equal number of fish in a sample
Missing age	Set of analysis with missing age
Mean weight	Mean sample weight per species.
Age – length	Age – length relationship
Weight – length	Weight – length relationship

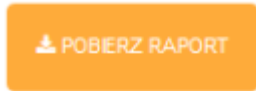
Generate report

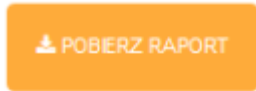
A rectangular button with a light orange background and a dark orange border. It contains a small icon of a document with a checkmark and the text "Generuj raport" in a sans-serif font.

To generate a report press the button . Until all parameters needed are defined, the button will stay inactive (grey). Having pressed the button, there should appear a progress bar. If the report is generated successfully, the window should appear:



with two buttons: to download it or cancel.



By choosing  (i.e. "Download") a second progress bar should appear. The download time depends on the size of the data. Do not close the browser until the download is completed.