



Deliverable I. Overview of the state of play, data gaps and needs

Document	Deliverable I. Overview of the state of play, data gaps and needs		
Date	23/02/2023		
Version	I		
Author(s)	Linda O'Hea, Manon Troucellier, Leonie O'Dowd, David Currie, Hans van Oostenbrugge		
Reviewed by	Susana Rivero Rodríguez	Date	17/03/2023
	Linda O'Hea	Date	20/03/2023
	Joël Vigneau	Date	30/03/2023
Final approval		Date	30/03/2023

FISHN'CO is funded by the European Maritime and Fisheries Fund (EMFF) of the European Commission



Content

Background information.....	I
I. Compilation of information about gaps to reach level of ambition – Thematic Focus Areas	2
TFA 1 – Commercial Fisheries.....	2
1a. Case Study - Small Pelagics in the Baltic.....	2
1b. Case Study - Freezer Trawlers.....	4
TFA 3 - Diadromous species	5
Salmon and Sea trout in the Baltic region	5
Salmon and Sea trout in the NANSEA region.....	7
TFA 4 - Small Scale Coastal Fisheries	9
TFA 6 – Impact of Fishing Activities - Stomach sampling	11
TFA 9 - Research Survey at Sea.....	13
TFA 10 - Biological Data Quality	14
2. Level of Ambition Tables for Thematic Focus Areas (TFAs)	15
TFA 1 – Commercial Fisheries.....	15
1 a. Case Study - Small Pelagics in the Baltic.....	15
1b. Case Study - Freezer Trawlers.....	17
1 c. Case Study - Iberian trawl	19
1d. Case Study - Large Pelagics.....	20
TFA 2 - Marine Recreational Fisheries.....	22
TFA 3 - Diadromous species	24
3 i. Salmon in the Baltic Sea region	24
3 ii. Sea trout in the Baltic Sea region.....	26
3 iii. Salmon in the NANSEA region.....	28
3 iv. Sea trout in the NANSEA region	29
3 v. Eel in the Baltic and NANSEA region.....	30
TFA 4 - Small Scale Coastal Fisheries	31
TFA 5 – Incidental catches of PETS	33
TFA 6 – Additional Data on the Impact of Fishing Activities - Stomach sampling.....	34
TFAs 7& 8 Social and Economic Data	36
TFA 9 Research Survey at Sea.....	38
TFA 10 Biological Data Quality	39
3. WPI – Thematic Focus Area: Biological Data Quality – Final Report.....	40

Biological Data Quality TFA	40
<i>Introduction</i>	40
<i>Objectives</i>	40
<i>Results and Discussion</i>	41
<i>Conclusions and Further Work</i>	60
Draft quality document for Baltic SPF regional pilot	62
Data quality control practices of European fisheries institutes	75
<i>Introduction</i>	75
<i>Objectives</i>	75
<i>Methodology</i>	75
<i>Glossary of terms</i>	76
<i>Response Rate</i>	76
<i>Results and discussion</i>	77
<i>Conclusion</i>	105
<i>Recommendations</i>	106
<i>References</i>	108
Annex 1. Data QC Questionnaire Report – An analysis of the data quality control practices of European fisheries institutes for data checks, editing and imputation	109



Background information

This document has been prepared with the aim of compiling the findings from Fishn'Co WP I – Compiling, identifying and filling information gaps, in one document. The document covers the partial deliverable: D1.1.

The document has been prepared in close cooperation with the regional and pan regional intersessional subgroups of the RCGs NANSEA, Baltic, LP and ECON. In particular, with the close collaboration of Task leaders and the experts in the ten Regional Work Plan Thematic Focus Areas selected as the most relevant for the project, namely: *Commercial Fisheries*, with four case studies (Small pelagics in the Baltic; Freezer trawlers; Iberian trawlers; Large pelagics) and the umbrella group; *Recreational Fisheries*; *Diadromous Species*, salmon and sea trout; *Small Scale Fisheries*; *Incidental Catches of PETS*; *Additional Data on the Impact of Fishing Activities on Marine Biological Resources and Marine Ecosystems*; *Social and Economic Data on Fisheries*; *Social, Economic and Environmental Data on Aquaculture*; *Research Surveys at Sea*; *Biological Data Quality*.

The compilation first identifies the gaps and then presents the map of what exists, what are the best elements and approaches already developed, and what additional information is still necessary to be able to develop Regional Work Plans. With these elements the level of ambition for regional coordination within each Thematic Focus Areas is also identified.

The work developed within WP I has also been integrated into an interactive infographic <https://www.fisheries-rcg.eu/level-of-ambitions/> with the purpose of keeping the viability of the work beyond the lifetime of Fishn'Co, and in final instance strengthen regional coordination.

1. Compilation of information about gaps to reach level of ambition – Thematic Focus Areas

A regional sampling program or any kind of multilateral agreement can be viewed as a process with several steps, ranging from no coordination towards fully coordinated data collection. Data collection on some fisheries and stocks is already partially regionally coordinated with aspects like age reading programs, data uploads and databases already sometimes shared among MS. However, many of these initiatives have in former times not been perceived as integral part of structured progress towards a full regional sampling programme. Such perception has been further confounded by a prevailing idea that data collection from all fisheries can (and would benefit from) undergoing a process that necessarily terminates at the highest level of regionalization.

To identify the current level of coordination and regionalization existing within a given fishery and help set the goal for how coordinated that fishery can/should be in the future, the system of regional coordination steps developed under the RCG were used (Table 1). Not all fisheries can (or need) to be coordinated at the highest level of regional coordination, e.g., if the fishery has very local and national characteristics it probably does not need such coordination even if still gaining by having some parts coordinated. The RCG approach is that regionalization is a process that can have several outcomes, and it is not necessary the final goal to have a full regional coordination for the objectives of improved coordination and regionalization to be fulfilled.



Table 1.- Level of ambition in a regional coordination scale. The levels range from zero (no coordination) to four (joint data collection)

#	Level of ambition
0	No coordination or not relevant
1	Coordinated data reporting
2	Agreed guidelines
3	Common monitoring strategy
4	Joint data collection

The level of ambition is assessed within the Regional Work Plan (RWP) Thematic Focus Areas as introduced in section III.1.a, listed in III.1.d. The RWP Thematic Focus Areas (TFAs) are in line with EU-MAP and the RCG intersessional work programmes.

TFA 1 – Commercial Fisheries

1a. Case Study - Small Pelagics in the Baltic

	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						To analyse the optimal number to be measured / aged
						A regional sampling plan can either cover a stock or a fleet segment. How to ensure coverage of all stocks within a given area.



	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common regional Database						If the sampling plan is covering a fleet segment it is hard to ensure the best coverage of all stocks. It will be important to ensure that all stocks are covered before changing towards a new strategy.
Comparability of results				🎯	🎯	
Harmonisation of data collection/Standardization				🎯	🎯	
Improving knowledge about similarity/difference between countries					🎯	Access to samples: Not all MS have access to the port where the vessels are landing. Better coordination between MS One MS have very small vessels and they do not have freezer capacity on board. An alternative solution needs to be developed.
Data quality and control data				🎯	🎯	Species misreporting, access to control data
End users needs					🎯	
Developing pilot study				🎯	🎯	
Data collection of other variables				🎯	🎯	



Ib. Case Study - Freezer Trawlers

	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method				🎯		Analysis required to identify optimal area/time coverage and sample protocol (sample size, number of samples, number measured, number aged)
Common regional Database				🎯		Harmonisation from national to common protocols may result in different uploading rules.
Comparability of results				🎯		
Harmonisation of data collection/Standardization				🎯		
Improving knowledge about similarity/difference between countries				🎯		A French data submission to the data call has not been received.
Data quality and control data			🎯			
End users needs				🎯		Awaiting feedback from assessment end users. PETS related end user needs to be clarified
Developing pilot study				🎯		Identification and training of candidate observers for the pilot scheme will be undertaken. Updates to sampling schemes and SOPs.
Data collection of other variables				🎯		PETS sampling requires special observer skills, conflicting working times (measuring under deck or observing from bridge).




TFA 3 - Diadromous species

Salmon and Sea trout in the Baltic region

	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						Not applicable
Comparability of results						Identification of practices and methods that don't produce co-dimensional parr density data. Also identification of data that is potentially collected from nontypical rearing habitats (e.g. WFD monitoring). Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise as RWP earliest in medium or long term.
Harmonisation of data collection/Standardization						Mapping of the criteria that is used for selecting the index rivers. For Salmon: index rivers are already designated and criteria for them specified. For Sea trout: ICES WGBAST has recommended to establish one index river per assessment unit. Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise as RWP earliest in medium or long term.
						For Sea trout: Evaluation of catch data in commercial and recreational fisheries. Estimates of retained and released catch of recreational fisheries in marine area and rivers would be needed. This element is linked to the thematic focus area of Recreational fisheries. For Salmon: Evaluation of specifications for unit of effort for different gears in commercial and recreational fisheries. Collection of catch and effort data of commercial fisheries is regulated by EU legislation. Unit of effort, however, may have different specifications in the data MSs supply for the ICES expert groups. In recreational fisheries specification of unit of effort for different gears is needed. Also catches should be reported or estimated separately for retained and released catch. And all this for



	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
						marine area and rivers. This element is linked to the thematic focus area Recreational fisheries. Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise as RWP earliest in medium or long term. More probable to realise in the Baltic than NANSEA region.
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables						Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise as RWP earliest in medium or long term. More probable to realise in the Baltic than NANSEA region.
Other, specific to thematic focus area						Not applicable



Salmon and Sea trout in the NANSEA region

	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						Not applicable
Comparability of results						
Harmonisation of data collection/Standardization						<p>Mapping of the criteria that is used for selecting the index rivers.</p> <p>For Salmon: the index rivers have been selected and based on national competencies and according to what deemed appropriate, affordable and necessary for the management of salmon stocks on national level. Their actual definition and selection within the ICES context is open.</p> <p>For Sea trout: need for sea trout index rivers has not been raised so far.</p> <p>Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise as RWP earliest in medium or long term.</p>
						<p>For Sea trout: Evaluation of catch data in commercial and recreational fisheries. Estimates of retained and released catch of recreational fisheries in marine area and rivers would be needed.</p> <p>This element is linked to the thematic focus area of Recreational fisheries.</p> <p>For Salmon: Evaluation of specifications for unit of effort for different gears in commercial and recreational fisheries. Collection of catch and effort data of commercial fisheries is regulated by EU legislation. Unit of effort, however, may have different specifications in the data MSs supply for the ICES expert groups.</p> <p>In recreational fisheries specification of unit of effort for different gears is needed. Also catches should be reported or estimated separately for retained and released catch. And all this for marine area and rivers. This element is linked to the thematic focus area Recreational fisheries.</p> <p>Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise</p>




	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Improving knowledge about similarity/difference between countries						as RWP earliest in medium or long term. More probable to realise in the Baltic than NANSEA region.
Data quality and control data						Not applicable
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables		🎯				Data workshops by end users are needed for defining the data needs for assessments and for planning the data collection on coordinated basis. This element can be expected to realise as RWP earliest in medium or long term. More probable to realise in the Baltic than NANSEA region.
Other, specific to thematic focus area						Not applicable



TFA 4 - Small Scale Coastal Fisheries

	Current position vs ambition (goal)					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						Some test is needed to check how SSF data fit to these data bases. Coordination between the different data bases is also essential (RDBES and Med & BS)
Comparability of results						Not applicable
Harmonisation of data collection/Standardization						Although this topic has been previously discussed, the implementation of a common procedure has not been reached
Improving knowledge about similarity/difference between countries						There is a lack of knowledge about the degree of similarity/difference between countries in what concerns DCF sampling of biological data from SSF especially regarding: obtained coverage of vessel lengths, strategy/design (is vessel length considered in stratification for sampling or not, is there a separated programme for LSF and for SSF, etc) obtained coverage of species relevant in SSF (but not relevant in LSF)
Data quality and control data						
End users needs						
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						Promote a fishing effort monitoring plan: fishing effort of SSF is less well characterized than LSF (which have mandatory VMS and electronic logbooks).
						Transversal data deficiencies for the SSF analysis: the proposed new Control Regulation could improve these deficiencies, but there is a need to improve in the mid-term.



					<p>Development of a common methodology for analysis of data from real time tracking devices in SSF. Unlike LSF where fishing effort is estimated by mandatory VMS and logbook, in the case of SSF there is a lack of information on the spatio-temporal distribution of fishing effort. Specific approaches for this fleet segment should be implemented.</p>
--	--	--	--	---	---



TFA 6 – Impact of Fishing Activities - Stomach sampling

	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						<p>Two protocols coexist regarding the analysis of preys in stomachs, one based on visual determination of the preys at lab (recommended by WGSAM and FishPI²) and one based on on-board analysis of the stomach volume. One protocol at NANSEA level or one protocol in the IBTS area and another in the Bay of Biscay? 2-3 stomach analysis centers, receiving samples from all countries or each country process the stomach collected during national surveys? Agree on the taxonomic resolution: all preys determined at the lowest taxonomic possible level or commercial fish and invertebrates species at lowest level or fish at lowest level only.</p> <p>Discussion about the pros and cons of the methods before being included in a regionally coordinated work plan</p>
Common regional Database						<p>Collection of stomach content data are time consuming, and good stomach data are scarce. Thus, those data are not usually shared until they are published by their producers. Setting up a relevant embargo time and/or ensuring European funding for the technical staff time (and not basing the work on national money, or on research project money) can allow sharing the data in common database. Contrarily, sharing will depend on good-will...</p>
Comparability of results						<p>Having comparable data on diet is a major goal of the project, but may be hard to reach, as intercalibration works are harder to set up for stomachs than for other aspects (otoliths by ex), as taxonomy of the preys is specific to each area, and as it is hard to based identification of preys on pictures only</p>
Harmonisation of data collection/Standardization						<p>Same as above: an interesting objective that may be hard to reach</p>
Improving knowledge about similarity/difference between countries						<p>Some countries have ongoing stomach sampling programs, that may not directly fit RCG aims, and they will not be eager to modify the protocol and stop their time series</p>
Data quality and control data						<p>Should be considered if each country analyses its own stomach. Not relevant if stomachs are analysed in one centre only</p>



	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
End users needs						WGSAM provided a list of species that is focused on providing estimates of natural mortality by main predators in the North Sea. However, this list may not be comprehensive about trophic interactions in the entire ecosystem.
Developing pilot study						IBTS as a case study in 2022 and next years. Having another case study in other areas could be considered, but would require less coordination, as other regions are surveyed by less countries
Data collection of other variables						Already done for species sampled for biology. To be done for other, but this will represent extra work. The funding for this staff time should be secured, unless it won't be possible to do it in addition with all the work already requested on board.
Other, specific to thematic focus area						Secure funding, notably to fund extra work at sea, and lab work. On the contrary, stomach may potentially not be sampled, or may stay for years frozen before being not analysed and discarded, i.e. being collected for nothing. Define and agree on cost sharing Inclusion on non EU countries



TFA 9 - Research Survey at Sea

	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						Not applicable
Comparability of results						Not applicable
Harmonisation of data collection/Standardization						Not applicable
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						Surveys as listed in Table I of EU MAP: While some surveys already have cost sharing agreements, the new table I needs to be fully reviewed for consensus on surveys selected as candidates for cost sharing
						Agreed reporting templates for survey descriptions: lack of template for describing surveys, no agreed survey descriptions that can be adopted in national and regional work plans
						Agreement of final table structure to capture survey elements of RWP (Table 2.6)



TFA 10 - Biological Data Quality


	Level of ambition					Gaps to reach level of ambition Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						Not applicable
Comparability of results						Not applicable
Harmonisation of data collection/Standardization						Sampling Design Documentation: Template on how to structure a regional sampling design document.
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Standardised method for describing which data checks are being applied by participants in regional sampling programs. Lack of tools available for regional sampling programs. Evaluation of precision for regional sampling programs. Extend existing bias analysis work to the regional level. Not all data is uploaded to international databases (need to have a summary of reasons why). Standardised method for describing how editing and imputing is being applied by participants in regional sampling programs.
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						Not applicable
						Not applicable


2. Level of Ambition Tables for Thematic Focus Areas (TFAs)

As part of Work Package 1 objectives of the FISHNCO project, the current stages of regional coordination were assessed and the level of ambitions for each of the defined thematic focus area for Regional Work Plans collated. These have been represented in the tables that follow. The current status is also available as an interactive infographic. The infographic (deliverable of WVP4) is the result of the compilation, identification and analysis of the status of regional coordination and it is aimed to inform the design of the Regional Work Plan structures.





TFA I – Commercial Fisheries

I a. Case Study - Small Pelagics in the Baltic

The tables below summarises the points made by the ISSG on small pelagics in the Baltic as a contribution to the Fishn’Co project and their level of ambition () for the thematic focus area concerned.

	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
Small Pelagics in the Baltic	0	No coordination or not relevant	3	0	30%
	1	Coordinated data reporting	4	0	
	2	Agreed guidelines	3	0	
	3	Common monitoring strategy	0	7	
	4	Joint data collection	0	3	


*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						It is the intention to have a common protocol defining the minimum amount (kg) per sample, species selection, numbers of ages and length measured, the units used.
						Common protocols on vessel selection, agreement on which part of the fleet to cover (large trawlers) and which part is covered by a national sampling program. Common sampling description (WGCATCH) for all MS to describe before benchmark. Using the same template and the same way to identify the sampling program (template has been developed). Common estimation description (WGCATCH) for all MS to describe before benchmark.
Common regional Database						Data for the case study has been uploaded in the RDBES as a common sampling program. Presently not all the data is uploaded in a common database but only the data from the case study.
Comparability of results						When a common vessel selection protocol and common sampling protocol is adopted, data across MS will be more comparable. Further, the








	Level of ambition					Comments
	0	1	2	3	4	
						ISSG will develop common estimation tools, which will enable comparison of estimates (point estimates and variances) across national strata and against present national estimates.
Harmonisation of data collection/ Standardisation						Annual meeting between those responsible for data collection. Evaluations of the impacts of different sampling designs, sampling protocols and sampling efforts are also ongoing. The last 2 years meetings have been conducted as part of the pilot. However not all MS has participated.
Improving knowledge about similarity/difference between countries						As part of the case study, we have now gathered information on all MS national programs and have started to evaluate how we can align sampling designs and estimation between MS and where it makes sense to keep the national exemptions.
Data quality and control data						Try to ensure a common way to identify mis-reporting. Make control data available for other nations. Common documentation on relevant national checks (RCG / FishCo/ ICES). Agreement on relevant national data checks (based on RDB-ES format)
End users needs						As part of the case study we will conduct analysis on the level of misreporting back in time to be used by the Benchmark process for herring and sprat in 2023. Presently, it has been discussed how to archive reliable information on the misreporting back in time (Scientific observers/ control data / other).
Developing pilot study						A pilot study, where most of the MS participate, has been running for 2 years.
Data collection of other variables						Not applicable
Other, specific to thematic focus area						Systematic age reading workshops. Workshop is already conducted within the ICES system, but not on a regular basis for the sprat and herring in the Baltic Sea.

1b. Case Study - Freezer Trawlers

The tables below summarises the points made by the ISSG as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.

	#	Level of ambition	Counts of Current positions	Counts of goals	Progress v goal*
Case Study - Freezer Trawlers	0	No coordination or not relevant	2	0	
	1	Coordinated data reporting	5	0	
	2	Agreed guidelines	2	1	
	3	Common monitoring strategy	0	8	
	4	Joint data collection	0	0	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).


	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						A comparison of current sampling protocols is currently underway, based on the joint reports currently produced by NED and DEU and a common protocol will be developed. Analysis of historic fleet behaviour and sampling data will be used to design a sampling scheme to both optimise coverage and ensure an appropriate sampling scheme (minimum sample size, number of length measures and number of biological samples) for each species.
Common regional Database						When the common sampling protocol is implemented, data will be uploaded to the RDBES. Currently, national sampling programs upload to the RDB.
Comparability of results						It is aimed to get full comparability of results after the adoption of a common vessel selection protocol and common sampling protocol. Documentation is already in place e.g. NED/DEU joint reports
Harmonisation of data collection/ Standardisation						See common sampling protocol and comparability of results.
Improving knowledge about similarity/difference between countries						A comprehensive analysis of fleet behaviour has been carried out, based on national submissions following a data call. Simulations were conducted to investigate the coverage associated with alternative sampling schemes (random trip, vessel, reference fleet), or national sampling schemes and coordinated. Further information will be provided by the end user assessment scientists.






	Level of ambition					Comments
	0	1	2	3	4	
Data quality and control data			🎯			A compilation of existing national guidelines and operating procedures for observer and self-sampling programmes will be compiled and consolidated.
End users needs				🎯		Stock assessment end user requirements are relatively well defined and will be refined based on feedback from stock coordinators and assessors. End user needs other than assessment groups (e.g. PETS) are required to be taken into account.
Developing pilot study				🎯		A pilot study is to be proposed based on a modification to the existing observer programme. Based on the results, sampling of the complete freezer fleet by a pool of nationally based observers will be considered. The modified observer protocol will focus on the collection of data for target (assessment) species, alongside the current requirements for by-catch monitoring.
Data collection of other variables				🎯		Collection of sensitive by-catch (PETS) is considered by current sampling programs

1 c. Case Study - Iberian trawl

The tables below summarises the points made by the ISSG on Case study Iberian trawl as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.


Commercial Iberian trawl case study	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
	0	No coordination or not relevant			
	1	Coordinated data reporting			
	2	Agreed guidelines			
	3	Common monitoring strategy		1	
	4	Joint data collection		2	


*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						Identify similarities/differences in current sampling protocols of this fishery by institutions/countries (AZTI, IEO, IPMA) and assess if differences can be changed aiming at similar procedures.
Common regional Database						Not applicable
Comparability of results						Not applicable
Harmonisation of data collection/Standardisation						Not applicable
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						Not applicable
Developing pilot study						Define scenarios for sampling design. The selected scenario to be implemented in a pilot study needs to be identified especially taking into account the output from FishPi2 and the sampling protocol. Allocation of sampling effort needs to be defined taking into account the final scenario selected.
Data collection of other variables						Not relevant
Other, specific to thematic focus area						Define aspects for the implementation of the pilot study (timing, costs, additional adjustments); Implement pilot study during one year; Compare results of the pilot study with results of the national sampling plans





1d. Case Study - Large Pelagics

The tables below summarises the points made by the ISSG on LP as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.

	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
Large Pelagics	0	No coordination or not relevant			
	1	Coordinated data reporting			
	2	Agreed guidelines		4	
	3	Common monitoring strategy		1	
	4	Joint data collection			


*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).


	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						<p>A common regional sampling plan onshore samples: Landings samplings, vessels or wells selections, sampling effort allocation, define a common process regarding training/formation. Common process for reporting incidences in the protocol guidelines (and report to higher levels). Question of “faux-poisson”.</p> <p>A common regional sampling plan offshore samples: Observers on boards, sampling effort allocation, define a common process regarding refusal rate, question of integration of EMS, percentage of coverage and proportion of human observers and electronic systems, define a common process regarding training/formation and good practice. Common process for reporting incidences in the protocol guidelines (and report to higher levels).</p> <p>A common protocol for working up biological data samples: Stomach samplings, otoliths, define cooperation in terms of factories (cannery access) and minimum standards. Develop new protocol in relation to new landing location (example Cape Verde).</p>
						Vessel selection
Common regional Database						Not applicable
Comparability of results						Not applicable







	Level of ambition					Comments
	0	1	2	3	4	
Harmonisation of data collection/ Standardisation						A common economic and social data collection: To define more clearly (especially people involved).
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Common Akado software and quality/scripts checking. Question of “faux-poissons” regarding new development.
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						- Target species processes: Tropical tunas treatment. - Bycatch processes: Treatment of bycatch, discard and observers data.

TFA 2 - Marine Recreational Fisheries


The tables below summarises the points made by the ISSG on MRF as a contribution to the fishn'Co project and their level of ambition () for the thematic focus area concerned.

	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
Marine recreational Fisheries	0	No coordination or not relevant	6	6	41%
	1	Coordinated data reporting	3	0	
	2	Agreed guidelines	2	0	
	3	Common monitoring strategy	0	3	
	4	Joint data collection	0	2	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Current position vs ambition (goal)					Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						RDBES +MED & BS and how the MRF data fit into it
Comparability of results						Not applicable
Harmonisation of data collection/ Standardisation						Not applicable
Improving knowledge about similarity/ difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						Although a mandatory list of species to collect data by region exist under the DCF, as a multispecies approach is asked to the different MS, it's important to agree at regional level what potential species to add under the RCGs umbrella based on end users needs
						RCG members expertise on DCF issues together with WGRFS expertise in different technical issues regarding the monitoring of this fishery is essential to improve the regional coordination.
						Incorporation of recreational fisheries data into the assessment WG. The answer to these specific end users needs especially when stocks are shared between different MS needs and important level of coordination.



	Current position vs ambition (goal)					Comments
	0	1	2	3	4	
Developing pilot study						Not relevant
Data collection of other variables						The impact of this fishery should not be considered from a biological impact side only. Other variables are also essential to consider (socioeconomic etc)
Other, specific to thematic focus area						Not applicable



TFA 3 - Diadromous species

The tables below summarises the points made by the ISSG on Diadromous species as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.

3 i. Salmon in the Baltic Sea region

Diadromous species Salmon in the Baltic Sea	#	Level of ambition	Counts of Current positions	Counts of goals	Progress v goal*
	0	No coordination or not relevant	5		
	1	Coordinated data reporting	6		
	2	Agreed guidelines	1	8	
	3	Common monitoring strategy		1	
	4	Joint data collection			

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).


	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						The assessment model of the ICES Baltic salmon and trout assessment working group (WGBAST) takes diverse types of data which is highly challenging to fit under a common sampling protocol.
Common regional Database						Presently catch and effort data of commercial marine fisheries is stored in the InterCatch. In future also catch and effort data on recreational fisheries at sea and all river fisheries should preferably be stored in the regional database (RDBES). Diverse types of other data are used in the assessment too, but these differ a from the regular ICES stock assessments and consequently probably make it unfeasible to comply with RDBES structure. Presently these data are stored in the national databases and compiled by WGBAST. Compiled data sets are stored in ICES SharePoint.
Comparability of results						Harmonise methods and comparability of results for electrofishing survey programs. Parr density data is used by the WGBAST in assessment and electrofishing surveys are included in the NWP of most MS in the region. All MS use standard method in electrofishing but the realised selection of sites is not fully transparent and documented.
Harmonisation of data collection/						Harmonise procedures to designate and run monitoring programs index rivers. Index rivers








	Level of ambition					Comments
	0	1	2	3	4	
Standardisation						are designated but there is room for improved coordination.
			🎯			Commercial catch and effort data are readily available. The coverage and quality of estimates or data on recreational catch and effort in marine and inland waters could be improved. These data are used in the Baltic salmon assessment.
Improving knowledge about similarity/difference between countries			🎯			Differences in data collection is partly described in EG reports but could be improved e.g. regarding selection criteria of electrofishing sites, estimation of river specific potential production capacity, etc.
Data quality and control data			🎯			Fisheries control data would be useful to get available to supplement the other fisheries data.
End users needs				🎯		Data needs of ICES WGBAST are documented in various EG reports.
Developing pilot study			🎯			There are needs for pilot studies on various subjects. None going on presently.
Data collection of other variables			🎯			Other biological sampling like catch sampling for ageing and genotyping where there is room for improved coordination.
Other, specific to thematic focus area						



3 ii. Sea trout in the Baltic Sea region

	#	Level of ambition	Counts of	Counts	Progress v goal*
			Current positions	of goals 	
Diadromous species Sea trout in the Baltic Sea	0	No coordination or not relevant	7		
	1	Coordinated data reporting	4		
	2	Agreed guidelines	1	8	
	3	Common monitoring strategy			
	4	Joint data collection			

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).


	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						The assessment model of the ICES Baltic (WGBAST) takes presently only parr density data. Depending on the model development of the ICES WGTRUTTA there will potentially become a need also for smolt and spawner count data and fisheries data.
Common regional Database						Presently catch data of commercial marine fisheries are stored in the InterCatch. In future also catch data on recreational fisheries at sea and in rivers should preferably be stored in the regional database (RDBES). Parr density data are used in assessment, but is probably unfeasible to comply with RDBES structure. Presently these data are stored in the national databases and compiled by the WGBAST. Compiled data sets are stored at the ICES SharePoint.
Comparability of results						Harmonise methods and comparability of results for electrofishing survey programs. Parr density data is used by the WGBAST in assessment and electrofishing surveys are included in the NWP of most MS in the region. All MS use standard method in electrofishing but the realised selection of sites is not fully transparent and documented.
Harmonisation of data collection/ Standardisation						Harmonise procedures to designate and run monitoring programs index rivers. Index rivers have not been designated despite WGBAST recommendation.
						Commercial catch data are readily available. The coverage and quality of estimates or data on recreational catch and effort in marine and inland waters could be improved. These data are not used in the assessment model but are used as supporting information in formulation of ICES advice.







	Level of ambition					Comments
	0	1	2	3	4	
Improving knowledge about similarity/difference between countries						Differences in data collection is partly described in EG reports but could be improved e.g. regarding selection criteria of electrofishing sites, etc.
Data quality and control data						Fisheries control data would be useful to have available to supplement the other fisheries data.
End users needs						Data needs of ICES WGBAST are documented in various EG reports.
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						




3 iii. Salmon in the NANSEA region

Diadromous species Salmon in the NANSEA region	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
	0	No coordination or not relevant	9		
	1	Coordinated data reporting	1	1	
	2	Agreed guidelines	1	3	
	3	Common monitoring strategy			
	4	Joint data collection			

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						The assessment model of the ICES North-Atlantic salmon assessment working group (WGNAS) takes diverse types of data all of which is highly challenging to fit under a common sampling protocol. Also third countries participate to the WGNAS work.
Common regional Database						ICES WGNAS have had a data call of catch data since 2020 covering all fisheries (commercial, recreational, farmed, ranched, indigenous, subsistence) in all fishing areas (coastal, estuary, river; open sea fisheries don't occur). Data is collected by age/size class of catch. Also estimates of unreported catch are compiled. Presently catch data is stored at the ICES SharePoint.
Comparability of results						Fisheries data
Harmonisation of data collection/ Standardisation						Not applicable
Improving knowledge about similarity/ difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						Data needs are defined by the ICES WGNAS in 2019.
Developing pilot study						Not relevant
Data collection of other variables						New assessment model under development in the WGNAS and consequently the data requirements will likely change in future.
Other, specific to thematic focus area						


3 iv. Sea trout in the NANSEA region

Diadromous species Sea trout in the NANSEA region	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
	0	No coordination or not relevant	9		
	1	Coordinated data reporting			
	2	Agreed guidelines			
	3	Common monitoring strategy			
	4	Joint data collection			



*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						No recognised international end-user for sea trout data in the NANSEA region. ICES WGTRUTTA is developing a general assessment model. No sea trout data collected under DCF so far. Several countries collect data for their national assessment purposes (ISSG Diad has not mapped where and what kinds of data have been collected and how long data series are available). Also third countries participate to the WGTRUTTA work.
Common regional Database						Potential future data needs differ from the regular ICES stock assessments and consequently probably make it unfeasible to comply with RDBES structure.
Comparability of results						Generally some standard methods used in electrofishing by countries in the region.
Harmonisation of data collection/ Standardisation						Not applicable
						Not applicable
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables						Depends on the model development of ICES WGTRUTTA.
Other, specific to thematic focus area						


3 v. Eel in the Baltic and NANSEA region


Diadromous species Eel in the NANSEA region	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
	0	No coordination or not relevant	9		
	1	Coordinated data reporting			
	2	Agreed guidelines	1	1	
	3	Common monitoring strategy			
	4	Joint data collection			

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).







	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						Catch, recruitment and other type of eel data are collected in MS and third countries.
Common regional Database						Presently the eel data are stored in PostgreSQL database hosted with the shiny in EPTB Vilaine (University) server. A lot of effort has been devoted by the WGEEL to get all types of data there. Only recruitment data is used in the present assessment model.
Comparability of results						Not applicable
Harmonisation of data collection/ Standardisation						Not applicable
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Not applicable
End users needs						ICES WKEELDATA has outlined the data requirements in 2021 and how future data calls will fulfil the data needs of WGEEL. Apart from WGEEL also MS are end-user of the data (National eel management plans).
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						

TFA 4 - Small Scale Coastal Fisheries




The tables below summarises the points made by the ISSG on SSF as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.

	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
Small Scale Coastal Fisheries	0	No coordination or not relevant	8	4	25%
	1	Coordinated data reporting	3	0	
	2	Agreed guidelines	2	0	
	3	Common monitoring strategy	0	8	
	4	Joint data collection	0	1	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).


	Current position vs ambition (goal)					Comments
	0	1	2	3	4	
Common sampling protocol/method						Not relevant
Common regional Database						Data base adapted to SSF needs (RDBES + Med & BS)
Comparability of results						Not applicable
Harmonisation of data collection/ Standardisation						Standardization of methodologies for biological data at EU level for the SSF fleet
Improving knowledge about similarity/difference between countries						Make a characterization of the current representativeness of biological data collected within the DCF in each country (in terms of vessel length coverage obtained and species coverage obtained) and identify targets for needed/wanted representativeness
Data quality and control data						Indicators agreed at regional level. Not only for sampling data, also for transversal data collected by the Control Regulation
End users needs						Identify main end users and their needs
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						Promote a fishing effort monitoring plan. It would be desirable to have at least 1/3 of the SSF fleet equipped with real time tracking devices, specifically developed for SSF, to determine spatialized fishing effort.




							<p>Real active vessels vs active registered vessels and low active vessels. In some preliminary analysis, several discrepancies were found. It's also to analyse what happens with the low active vessels, how they are covered etc as this could have in the general analysis made for this fishery.</p>	
								<p>Transversal data deficiencies for the SSF analysis</p>
								









TFA 5 – Incidental catches of PETS


The tables below summarises the points made by the ISSG on PETS as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.


	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
PETS bycatch	0	No coordination or not relevant	0	0	55%
	1	Coordinated data reporting	2	0	
	2	Agreed guidelines	4	0	
	3	Common monitoring strategy	0	6	
	4	Joint data collection	0	0	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).






	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						Improve the protocols for scientific observers sampling PETS bycatch onboard fishing vessels. Need to have an agreement at regional level
Common regional Database						RDBES+ Med & BS regional data bases suited for accommodation of PETSW bycatch data
Comparability of results						Not relevant
Harmonisation of data collection/Standardisation						Standardisation of methodologies: (effort estimates, raising procedures) at regional level
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Indicators agreed at regional level
End users needs						Identification of relevant fisheries concerning PETS bycatch issue it's essential to take decision regarding data collection and coordination level needed.
						Coordinated regional identification of level of effort needed for PETS bycatch data collection
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						

TFA 6 – Additional Data on the Impact of Fishing Activities - Stomach sampling

The tables below summarises the points made by the ISSG on Stomach Sampling as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.

Stomach sampling	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
	0	No coordination or not relevant			
1	Coordinated data reporting				
2	Agreed guidelines				
3	Common monitoring strategy				
4	Joint data collection				


*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).


	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						<p>IBTS 2022 will be use as a case study to test for the implementation of a common protocol regarding stomachs collection in the North Sea in collaboration with WGIBTS. The 5 years rolling scheme regarding species sampled will apply, at least for year 1.</p> <p>Further technical aspects, notably regarding stomach analysis per se, are TORs that will be included in the 2022 ISSG work plan, notably after having more information about financial aspects.</p> <p>It needs to be clarified, whether each country will have to analyse the stomachs in their own laboratories or whether some stomach analysis centers will be established.</p>
Common regional Database						<p>Define a common data format.</p> <p>Ensure that stomach data can be integrated in ICES database</p> <p>Define rules regarding data property (embargo for a defined period after upload?)</p>
Comparability of results						<p>Needed to ensure the aim of the coordinated stomach sampling ie having large spatial and taxonomic resolution for stomach data.</p>
Harmonisation of data collection/Standardisation						
Improving knowledge about similarity/difference between countries						
Data quality and control data						Not applicable









	Level of ambition					Comments
	0	1	2	3	4	
End users needs						Species choice: Identification of potential overlaps for species already included in MSFD programmes Refining the species list recommended by WGSAM to consider species under conservation status (e.g. sharks and rays) Results of the online survey launched in 2021 will allow a better definition of end-users needs, including those not part of the ICES/RCG process.
Developing pilot study						IBTS 2022 will be used as a pilot study for stomach sampling. A pilot study for stomach analysis and data treatment should be planned when more information on funding is available
Data collection of other variables						Biometrical traits (length, mass etc.) are needed to analyze prey data, are recorded for species already included in biological parameters sampling, and should be recorded for species not included in biological monitoring,
Other, specific to thematic focus area						Potential cost issues

TFAs 7& 8 Social and Economic Data


The tables below summarises the points made by the ISSG as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned. The Level of ambition was discussed and agreed during the RCGEcon meeting in 2021. Based on an inventory of the issues of the previous 6 years reports and an action list was made to work on these issues in ISSG-meeting) and progress has been made on a number of them, which is indicated in the table below.

	#	Level of ambition	Counts of Current positions	Counts of goals 	Progress v goal*
Economic data analysis	0	No coordination or not relevant		0	72%
	1	Coordinated data reporting	4	0	
	2	Agreed guidelines	3	7	
	3	Common monitoring strategy		0	
	4	Joint data collection		0	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						For all economic variables standardised methodologies exist and guidelines are available. In some specific cases cross-national cooperation can lead to increased data quality. An identification of cases has been carried out and presented to RCGEcon.
Common regional Database						Aggregate Socio-economic data are now available through the JRC databases. Sharing of detailed economic data is not an ambition.
Comparability of results						For some specific variables (e.g. value of tangible assets and value of intangibles) more testing and implementation of these guidelines needs to be carried out. Work on this is in progress.
Harmonisation of data collection/Standardisation						The handbook provides clear guidelines on the types of survey design and analysis that can be used. This needs to be implemented by MS in their NWP
Improving knowledge about similarity/difference between countries						Not applicable
Data quality and control data						Data validation checks are available at JRC and data are checked during EWG.
End users needs						The current segmentation does not result in optimal homogeneous segments for international comparison and bio-economic modelling. An improved segmentation in which vessels are grouped based on fishing activities



	Level of ambition					Comments
	0	1	2	3	4	
						might be more appropriate and is under development.
Developing pilot study						For social impact analysis, National and community profiles will be necessary to provide information about the social context. First pilots of these profiles have been discussed, but need to be further developed.
Data collection of other variables						Not applicable
Other, specific to thematic focus area						Not applicable



TFA 9 Research Survey at Sea

The tables below summarises the points made by the ISSG on Surveys as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.


	#	Level of ambition	Counts of Current positions	Counts of goals	Progress v goal*
Research Surveys at Sea	0	No coordination or not relevant			25%
	1	Coordinated data reporting			
	2	Agreed guidelines			
	3	Common monitoring strategy	8	1	
	4	Joint data collection		7	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						
Common regional Database						
Comparability of results						
Harmonisation of data collection/Standardisation						
Improving knowledge about similarity/difference between countries						
Data quality and control data						
End users needs						
Developing pilot study						Not Applicable
Data collection of other variables						
Other, specific to thematic focus area						Surveys listed in table I of EU_MAP are selected via the STECF 'Decision Support Tool' (DST) and fulfil criteria which brings them to a minimum of level 3, with several at level 4 displaying joint data collection programmes. Cost sharing agreements are considered as "4+" as they allow financial contributions to redistribute survey effort for MS who have monitoring obligations.
						Level of ambition is to have commonly agreed survey descriptions for surveys in Table I of the EU_MAP. To reduce text in national work plans and ensure consistency with regional reporting




RWP table structure for surveys Table and Text Box 2.6 final agreement for RWPs

TFA 10 Biological Data Quality

The tables below summarises the points made by the ISSG on Data Quality as a contribution to the Fishn'Co project and their level of ambition () for the thematic focus area concerned.

	#	Level of ambition	Counts of Current positions	Counts of goals	Progress v goal*
Biological Data Quality	0	No coordination or not relevant			25%
	1	Coordinated data reporting	3		
	2	Agreed guidelines			
	3	Common monitoring strategy			
	4	Joint data collection		3	

*Progress vs goal calculation is the ratio of the sum of product between the numbers in each column and the level of ambition (0-4).

	Level of ambition					Comments
	0	1	2	3	4	
Common sampling protocol/method						Not applicable
Common regional Database						Not applicable
Comparability of results						Not applicable
Harmonisation of data collection/ Standardisation						Standardised method of describing regional sampling programmes
Improving knowledge about similarity/difference between countries						Standardised method of describing regional sampling programmes
Data quality and control data						Data capture checking documentation, guidance for evaluating data accuracy, data storage documentation, documenting methods of editing and imputing
End users needs						Not applicable
Developing pilot study						Not applicable
Data collection of other variables						Not applicable
Other, specific to thematic focus area						

3. WPI – Thematic Focus Area: Biological Data Quality – Final Report

Biological Data Quality TFA

Introduction

Quality assurance of fisheries data collection is important but difficult to evaluate at a regional level. In the previous DCF workplan templates Table 5A was used by MS to summarise their biological data quality assurance for each of their sampling schemes. In the current DCF workplan templates MS are now required to provide more detailed information on quality assurance by completing an Annex I.1 quality document for each sampling scheme.

The Biological Data Quality TFA aimed to develop common templates and tools that MS can use to complete Annex I.1 in a regional context to improve inter comparability of quality information.

The RCG Data Quality ISSG previously identified a number of gaps in the existing tools and templates and based on these a number of support tasks have been identified including guidance on Sampling Implementation, data capture checks data storage guidance, evaluation of data accuracy (precision and bias) and documenting templates for editing and imputing.

Objectives

The following objectives and tasks were defined for the Biological Data Quality TFA.

1. Produce guidance for Sampling Design
 - 1.1. Collate further examples of sampling design documents from MS not already considered by the ICES PGData and WGQuality groups and the RCG Data Quality ISSG
 - 1.2. Incorporate these further documents into the analysis already performed by the ICES PGData and WGQuality groups and the RCG Data Quality ISSG
 - 1.3. Produce a final template on how to structure a sampling design document.
2. Produce guidance for Sampling Implementation

Handling of "Non-responses & Refusals" will be incorporated in the outputs of Objective 1 so no additional work was required
3. Produce guidance for Data Checks
 - 3.1. Collate national examples of the types of data checks that are implemented
 - 3.2. Categorise these data checks (take into account existing concepts of data quality such as consistency, completeness). Identify any categories of data check that MS are not doing, based on standard data quality concepts.
 - 3.3. Using the categories of data checks identified create a template that MS can use to identify which categories of data check they are implementing and, ideally, point to public code repositories of these checks (if they exist)
4. Produce guidance for Data Storage
 - 4.1. Summarise reasons why MS are not uploading to appropriate international databases
5. Produce guidance for Evaluating data accuracy (precision and bias)
 - 5.1. Identify the different types of estimation that are routinely being performed by MS, and those that would be suitable for regional estimation. Use existing sources of this information such as relevant ICES EG reports (e.g. WGCATCH, WKRDB-EST) and contact national experts as appropriate.
 - 5.2. Using the R language specify the statistical functions required to allow MS to evaluate bias and estimate precision for regional estimation. This should include defining the prerequisites that a MS will need to meet to be able to use the tools (e.g. what types of data the MS must collect, and which data format to use).
6. Produce guidance for Documenting methods of editing and imputing



- 6.1. Collate national examples of the types of editing and imputing that are being performed e.g. identify the techniques and/or libraries that MS are using
- 6.2. Categorise these methods.
- 6.3. Using the categories of methods identified create a template that MS can use to identify which methods of editing and imputation they are implementing and, ideally, point to public code repositories (if they exist)
- 6.4.

Results and Discussion

Objective 1) Produce guidance for Sampling Design

This objective is closely linked to the Data Quality ISSG and involved a collaboration with the Baltic small pelagic fisheries regional pilot study group.

There has been a significant amount of previous work on developing a sampling design document template but that work has concentrated on national sampling programmes. The focus of the current work is to see how a regional programme can be clearly described using the Annex I.1 Biological Data Quality template. The aim is to have a single document which describes the regional programme – this document will need to have input from all countries involved in the sampling. The initial audience for the completed documents will be national institutes, with an aim being to provide it to other interested parties such as ICES benchmark groups in the future.

Since it can be confusing when there are a number of different templates being developed that cover similar concepts it is useful to review how they are related – this is shown in **¡Error! No se encuentra el origen de la referencia..**

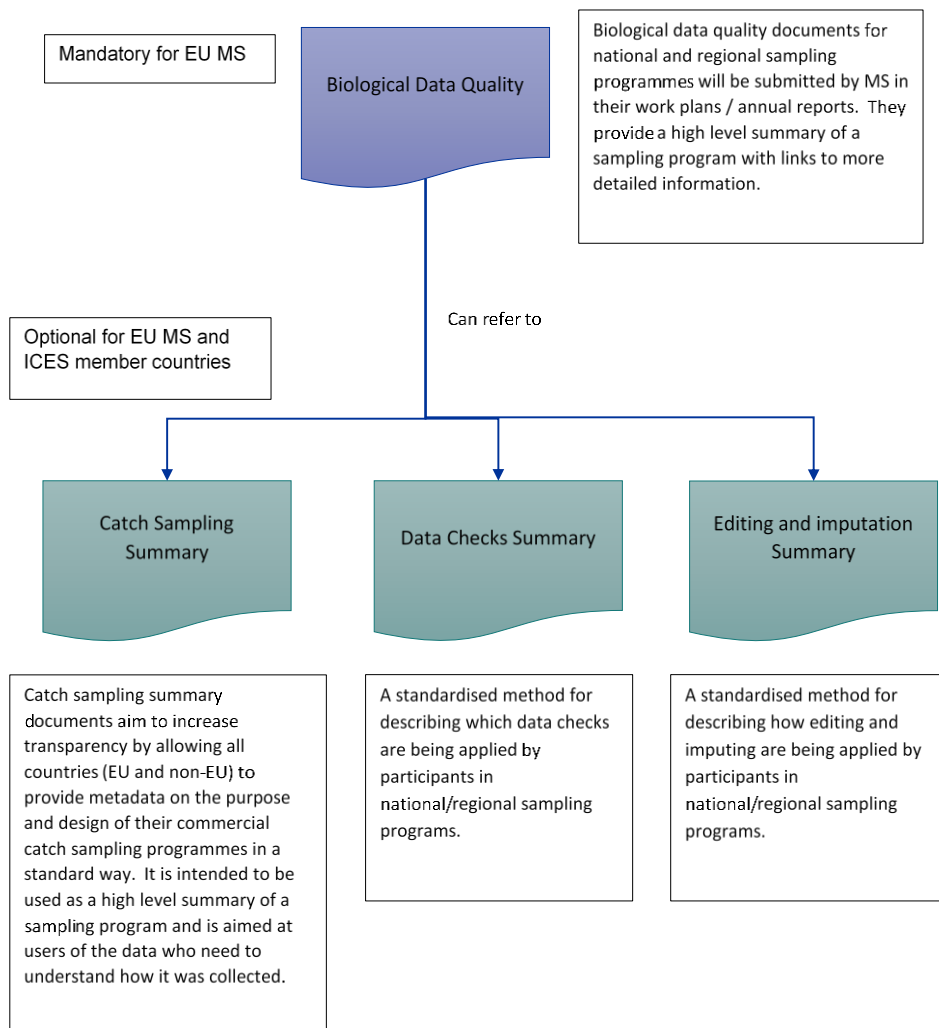


Figure 1. Relationship between templates / documents

The Catch Sampling Summary template has been developed by the ICES PGData and WGQuality groups. The Data Checks and Editing and Imputation Summary templates referred to here have been created as part of the FishNCo project and will be discussed in this section as part of Objectives 3 and 6.

The FishNCo Baltic small pelagic case study was selected as an initial test case to evaluate how the existing Annex I.I biological data quality templates can be used to describe a regional sampling programme. Once agreed it is intended that the other regional case studies can use the completed document as a guide when completing their own documents. This document is presented in the “Draft quality document for Baltic SPF regional pilot” section.

During the creation of this document a few points which will be of common interest to all of the regional case studies were discussed:

1. It was agreed that although there are sections of the national Annex I.I template that do not seem as relevant for regional programmes it was preferable to use the same template rather than try and adapt it for a regional context. When writing the document if there are sections that aren't regionally

- coordinated then a short sentence to explain this should be included e.g. “*Is presently not regionally coordinated*”
2. Some standard text to refer to the 4-point level of regional ambition compared to its current state should be added to the relevant sections e.g.
Regional level of ambition: 3 - “Common monitoring strategy”
Present regional level: 1 - “Coordinated data reporting”
 3. Where there are parts of the population that are not covered by the regional sampling programme but are instead being sampled nationally then the MS, Sampling scheme identifier, and Sampling frame identifier of these national programmes should be specified in the “Description of the population” section. This will ensure a reader can easily find this information.
 4. A summary of regionally coordinated age reading workshops should be included in the “Data Processing” section

Objective 3) Produce guidance for Data Checks and Objective 6) Produce guidance for Documenting methods of editing and imputing

The aims of these Objectives was to produce:

1. A standardised method for describing which data checks are being applied by participants in regional sampling programs.
2. A standardised method for describing how editing and imputing are being applied by participants in regional sampling programs.

The same methodology was used for Objectives 3 and 6:

1. The Data Quality ISSG designed a questionnaire with the aim being to discover what types of data checks, editing, and imputation the institutes cooperating in the RCG and FishNCo project are regularly performing
2. Funding from the FishNCo project was then used to contract an out-sourced resource to:
 - a. Send the questionnaires to relevant people at the institutes
 - b. Collate, categorise, and analyse the questionnaire results
 - c. Using the questionnaire results design template(s) that that MS participating in regional sampling programmes can use to identify which data checks, and which methods of editing and imputation they are implementing
3. The participants in the FishNCo project and the Data Quality then reviewed the templates and suggested any changes required

To give more guidance to institutes when completing the questionnaire, it was decided to limit the scope to processes that are applied to biological sample data from commercial catches that will be used for an analytical stock assessment. The data measured will typically include length-frequency distributions, and common biological parameters such as sex, maturity, age, weight, and length. Data quality processes related to census data (e.g. logbooks, sales notes) are also within the scope of the questionnaire when they are used to produce outputs from the biological data (i.e. when the data is raised).

The main outcomes from the survey are discussed below - the full results and analysis of the survey are presented in the “*Data quality control practices of European fisheries institutes*” section.

Data checks

The primary objective of the questionnaire was to determine if, when and how European fisheries institutes performed data quality control checks, data editing and data imputation. The analysis presented above indicates that most respondents: constrained some values to be physically realistic, used predefined code lists,

performed some form of outlier check, performed some form of spatial data check, performed some form of temporal consistency check, and performed some form of duplication check. These checks were performed regularly as part of the data collection process, as were cross checks with census data and missing values check.

However, whilst most MS performed these checks the point at which checks were performed varied greatly. The reason for performing checks at different points in the process could be attributed to different data capture methods, different time frames for the importing data, or different operating procedures in relation to data collection and checking. At a minimum, institutes should aim to ensure all checks have been performed prior to responding to data calls (at or prior to the point of data extraction). If checks are implemented at a different or additional stage (where checks are being implemented at multiple points), the point, method and type of checks implemented should be documented.

The method for some checks, such as outlier detection and cross checking of spatial data, are similar for many respondents. As many respondents already have a dedicated R script which produces plots which aid in the identification of outliers, it may be possible to produce a standardised R script dedicated to outlier checking and or spatial data plotting, which would be available to all members of the RCG (in turn standardising some checks discussed above). While variety in sampling schemes and data collection practices might limit the effectiveness of such a script, a standardised script containing protocols might prove useful in ensuring checks are in place and use a common method.

Data editing

The consensus for approaches to dealing with errors, inconsistencies and discrepancies was to attempt to correct the sample data where possible, and to exclude the data from outputs where correction is not possible. If data cannot be corrected, institutes should at least aim to document the error prior to deletion. Such a record may help in preventing similar mistakes in future and highlight repeated errors so corrective action(s) can be taken. Such error logging is already used by some MS. The template for logging errors proposed by Wageningen Marine Research may be suitable for logging such errors. If possible, institutes should also log errors even where correction was possible, again to prevent any future errors.

Data Imputation

For dealing with their approach to gaps in Age length keys (ALK's) or weight length keys (WLK's), institutes filled such gaps either by imputing from an average, imputing from a model, imputing from other strata, filling by expert judgement, or leaving the gap. As the course of action often depended on what data from other surveys, strata or sampling schemes was available, a definitive course of action to be taken in the event of an ALK/WLK gap is not appropriate. However, where gaps have been filled, institutes should document which data was imputed and what method was used. If a predicted value from a model was used, details of the model should be recorded. If data is borrowed from other strata or from surveys, the details of the strata or survey should be recorded.

When asked about dealing with gaps in sampling strata, most respondents opted to leave the gap and allow the ICES stock coordinator to decide how to deal with the issue. As this is already a popular course of action, leaving the gaps in the sampling strata and allowing the ICES stock coordinator to deal with them should be the course of action employed by institutes to deal with gap in their sampling strata. Where institutes decide to impute from other strata or surveys, details of what values have been imputed and of the method of imputation should be documented, such that the ICES stock coordinator is aware data has been imputed. This should minimise the chances of already imputed data being further imputed from, increasing data accuracy overall.

Written guidelines

Where asked to list any relevant written guidelines many institutes were not able to provide such guidelines, either because they did not have any or they were not publicly available. As institutes still performed many of these checks without such guidelines, they may be unnecessary, however having Standard Operating Procedures (SOPs) for data quality control recorded in a document would be a useful resource, both at a regional and international level. While such guidelines may contain information sensitive under GDPR, a censored or constrained document could still be appropriate.

Age-readings

While there was some reference to data quality control in relation to otolith readings most respondents did not discuss these practices in their answers. As a result, this report cannot recommend 'best practice' quality control with regards to otolith readings, as it is not supported by the data presented here.

Recommendations

1. When data quality control checks are implemented, institutes should ensure that the type of check, timing of the check (both the point during the data collection process and the date), and a brief description of the check are documented.
2. Where checks are performed at multiple points during the data collection process, institutes should ensure that datasets / samples are marked such that users are aware what checks have been already performed or where data has been edited or imputed.
3. Where the method of check is broadly similar among institutes attempts should be made to produce a standardised SOP, ideally at a WG level, detailing the method used to perform the checks.
4. Where the method of check is broadly similar among institutes attempts should be made to produce an R script to conduct these checks which is available to all users.
5. Where errors, inconsistencies or discrepancies are found in the data, information about the cause of the error and course of action taken to rectify it should be recorded. Records will allow users to identify common sources of error in data collection process.
6. Where institutes are imputing data from a predicted average/model/survey or from other strata to fill gaps in ALK's or WLK's, institutes should clearly document *what* data has been imputed, *where* the data was imputed from and *when* the data was imputed. As imputation may be performed at multiple points or by different users, it is essential that all users, from local to working group level, are aware what data is 'real' data and what data has been predicted or imputed.
7. Where gaps are found in sampling strata, a standardised course of action should be decided on at WG level. Based on the analysis conducted in this report, the most suitable course of action is to leave the gaps and allow the ICES stock coordinator to decide on how best to deal with them.
8. Further research should be conducted to collect information on data checks, editing and imputation with regards to age-reading among institutes.

Based on the survey a template has been created ("BioDataQualityTFA_Data_checks_template") that allows MS to:

- i. identify which categories of data check they are implementing
- ii. identify which methods of editing and imputation they are implementing

The intention is that MS could complete this template for particular sampling programmes and then publically share it e.g. the Marine Institute, Ireland has done this and examples are available at <https://www.dcmapp-ireland.ie/documents/methodologies> Once the completed template is publically available it can then be referenced from documents such as Annex I.I Biological Quality Documents.

During discussions it was identified that some MS thought that the questions related to imputation should be removed from the template – MS are free to adapt the template for their own use and can either ignore or remove the imputation questions if they wish to. This discussion should also be considered if any future work is done to modify this template. The recommendations and template can be further discussed and refined at MS institutes and within the Regional Coordination Groups. These are seen as good routes to make people aware of this work, and to further improve the outputs.

It was thought that it would be useful for regional sampling pilot programmes to try completing the template and this could be a first step towards harmonising data checks in a regional context. It is also useful to note that if a regional programme identifies checks that should be applied to the entire regional dataset (rather than national portions of that programme) it could be possible to implement those checks in the ICES Regional Database & Estimation System (RDBES).

A template has also been proposed that would allow MS to document any changes that are made to their data (“BioDataQualityTFA_Error_Log_template”) – this template is only intended for internal tracking of changes by MS and it is not proposed that the results should be made public.

Objective 4) Produce guidance for Data Storage

This work was completed by Data Quality ISSG and the output is in part III of the RCG NA NS&EA RCG Baltic 2021 report <https://datacollection.jrc.ec.europa.eu/docs/rcg> . For convenience a summary is included here:

Type of detailed data	Reason for not uploading to international database	Comment
Eel	No database with common access exists	<p>WGEEL collect and store some types of data from member states for the use of the group. Data collected by WGEEL included landings, recruitment, yellow eel standing stock, silver eel time series, and recreational catches.</p> <p>There have been discussions about storing diadromous data in the RDBES but these are at an early state.</p>
Salmon	No database with common access exists	<p>Some aggregated salmon data (i.e. landings, BMS landings and number of fish damaged by seals) from recent years has been uploaded to InterCatch.</p> <p>WGBAST collect and store some types of data from member states for the use of the group. This includes biological sampling, number of fish from stockings, fish stocking magnitude, recreational catches, electrofishing data, fish ladder data, and smolt trapping results.</p> <p>There have been discussions about storing diadromous data in the RDBES but these are at an early state.</p>

Type of detailed data	Reason for not uploading to international database	Comment
Mediterranean	No database exists	An initiative to solve this is ongoing since an EU funded project to develop a regional database for the Mediterranean & Black Sea region has begun.
Freshwater	No database exists	Unknown if any international initiatives are ongoing or planned.
Southern waters and other regions	No database exists	Unknown if any international initiatives are ongoing or planned.
Recreational	No database exists	Aggregated data could end up in the new RDBES (if it is found possible and appropriate), but detailed data may often consist of questionnaires and are not currently planned to end up in a common international database.
National crustaceans, cephalopods, shellfish surveys	No database exists	Unknown if any international initiatives are ongoing or planned.

It can be seen that in general the reason why detailed data has not been uploaded to an international database is that a suitable database does not exist. It should be seen as a positive trend that where an international database does exist MS are generally submitting the relevant data to it.

For future work-plans / annual reports MS are advised to make a comment on why datasets are not in an international database, if that is the case.

Objective 5) Produce guidance for evaluating data accuracy (precision and bias)

The aims of this Objective are:

1. Create tools to allow the evaluation of precision for regional sampling programs (by extending previous work done in the “Background document for response to special request regarding precision and bias based on RDBES format”)
2. Extend existing bias analysis work to regional sampling programmes.

The work on this Objective was planned by the Data Quality ISSG made use of an out-sourced resource funded by the FishNCo project. Before describing the work done it is useful to first define what is meant by data accuracy in this context:

“Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure.

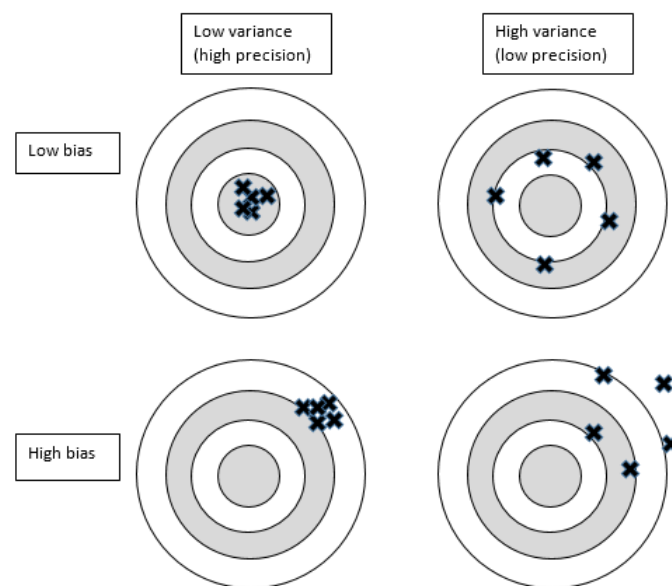
...

The concept of accuracy relates a numerical estimate to its true value according to an agreed definition. The closer the estimate is to its true value, the more accurate it is. The difference between the estimate and the true value is called the error of the estimate and error is thus a technical term to represent the degree of lack of accuracy. The error has a random component (variance) as well as a systematic component (bias). It is sometimes better to speak of uncertainty than error, when the term error risks to be confused with a mistake committed, which is a very different matter.”

p98, European Statistical System (ESS) handbook for quality and metadata reports — 2020 edition

In the context of this project the concept of data accuracy is explicitly linked with the terms “precision” and “bias”. In this case precision can be considered to be inversely related to variance i.e. a higher variance in the random component of the uncertainty means a lower precision.

An informal example which is often given to illustrate the difference between variance and bias is that of trying to shoot arrows at a target - ideally we’d like all our arrows to be in the centre. The diagram below illustrates how the arrows might hit the target in different variance and bias scenarios



Clearly a desired situation is to have both low variance (high precision) and low bias in our estimates although this may not always be possible in practice.

It should be noted that there can be a number of different types of bias occurring at different points in the data collection and advice production cycle – in this report we only consider bias that may occur as a result of sampling, not other biases such as those that may be present in particular estimators, or stock assessment models.

Precision / variance analysis

Regarding the precision analysis work the following deliverable was defined:



- An R implementation of an appropriate statistical algorithm for calculating the variance of point estimates from a multi-country, multi-stage, hierarchical commercial fisheries sampling program. It accepts input data in the ICES Regional Database & Estimation System (RDBES) data format <https://github.com/ices-tools-dev/RDBES>. The data will be in the ICES Regional Database & Estimation System (RDBES) data format <https://github.com/ices-tools-dev/RDBES> – example data will be provided.

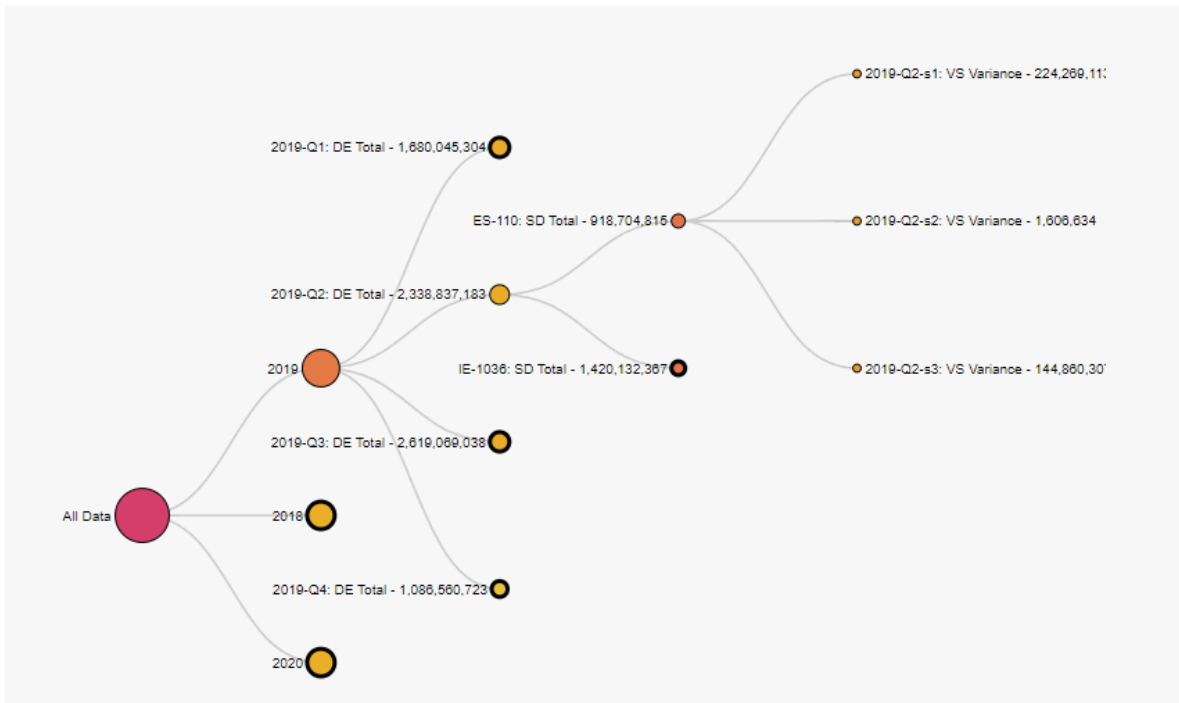
The work on evaluating data precision proceeded in close collaboration with the ICES Working Group on Estimation with the RDBES data model (WGRDBES-EST). WGRDBES-EST steered the choice of algorithm and members of the group were available for discussions. It is expected that the code delivered will be incorporated into a future R package on this topic. The deliverable was informed by the work already done on this topic in the Second Workshop on Estimation with the RDBES data model <https://doi.org/10.17895/ices.pub.7915>

WGRDBES-EST are developing a package (“RDBEScore”) to support design-based estimation using the RDBES – as part of this work functions to estimate totals/means using a generalised Horvitz-Thompson estimator and estimate variance using the Sen-Yates-Grundy formulation have been written. For the FishNCo project an RMarkdown script has been written to display the estimates and variance values in an interactive way. The latest version of the script can be found at <https://github.com/ices-tools-dev/RDBEScore/tree/main/FishNCo>

In this section examples of the graphical outputs from this script are presented – they use test data which has been generated for a fictitious regional sampling program incorporating Ireland, France, and Spain, centred on the island of Ireland.

Tree Diagram for Variance of Vessel Selection Strata

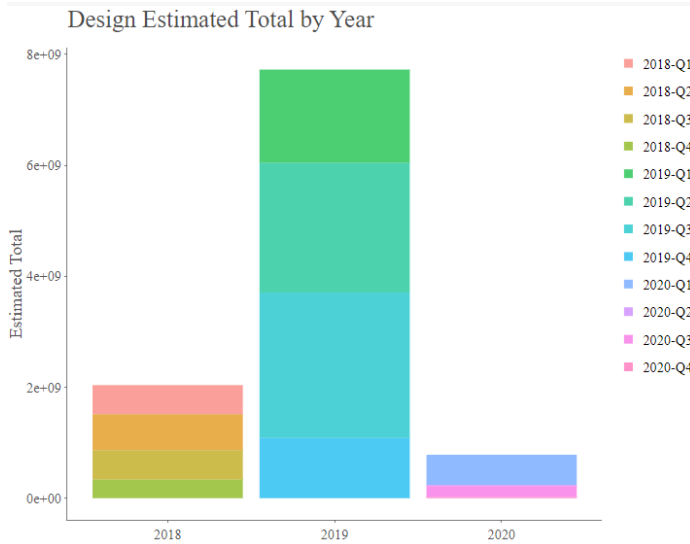
Plot showing the variance for the vessel selection strata. The tree displays the parent strata, Design and Sampling Details. In the interactive version the user can click on a node to expand its child strata.



Data Visualisations

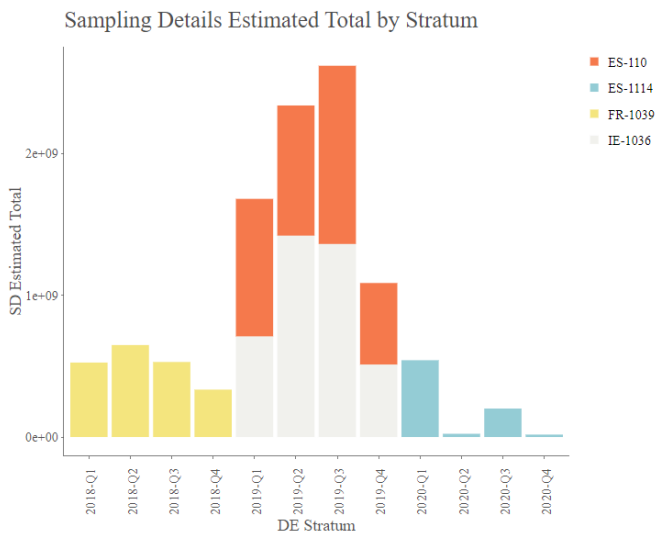
Design (DE)

Graph showing the Estimated Total per Year, coloured by Design Stratum Name.



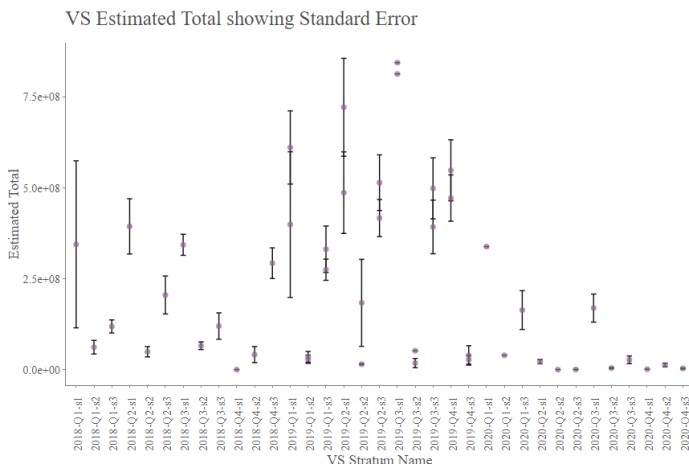
Sampling Details (SD)

Graph showing the Estimated Total by Stratum, coloured by Sampling Details Stratum Name.



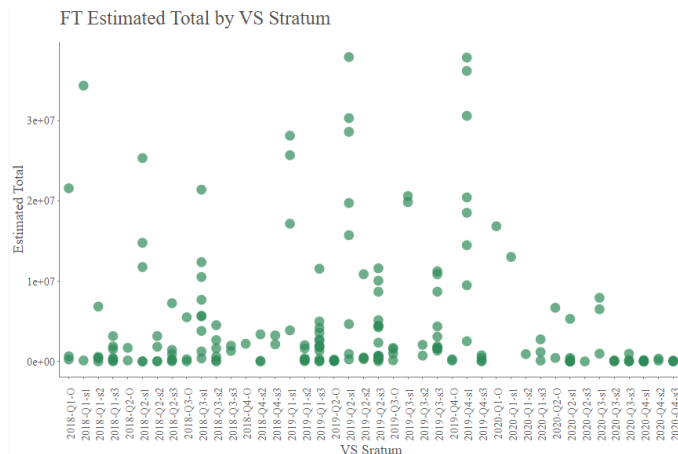
Vessel Section (VS)

The graph shows the Vessel Selection estimated total per Vessel Selection stratum name. The error bars show the standard error of the values with smaller bars depicting higher confidence levels. For the error bars with Standard Error value of NA, the confidence level is unknown.



Fishing Trip (FT)

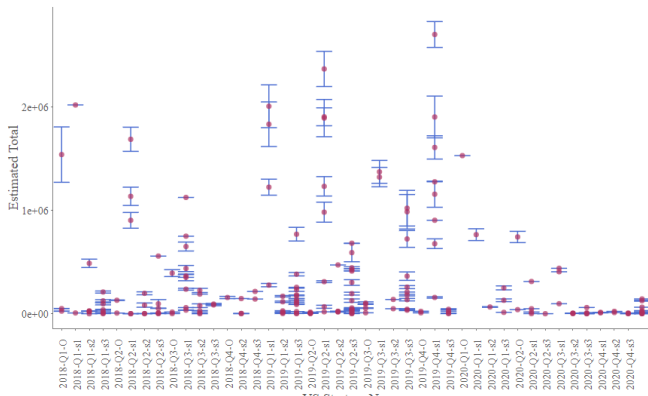
This graph shows the Fishing Trip estimated totals by Vessel Selection stratum name.



Fishing Operation (FO)

The graph shows the Fishing Operation estimated total per Vessel Selection stratum name. The error bars show the standard error of the values with smaller bars depicting higher confidence levels. For the error bars with Standard Error value of NA, the confidence level is unknown.

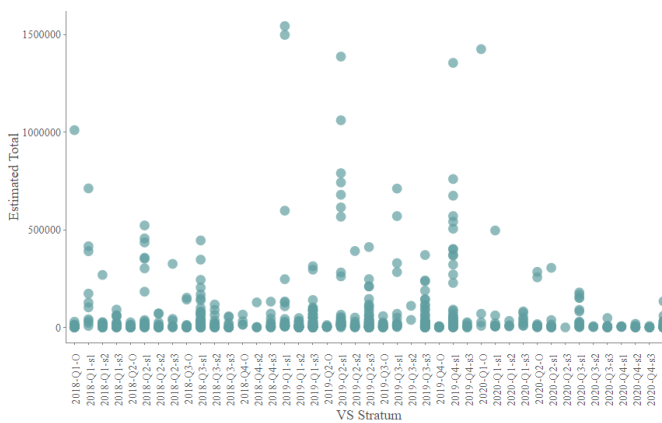
FO Estimated Total showing Standard Error



Species Selection (SS)

This graph shows the Species Selection estimated totals by Vessel Selection stratum name.

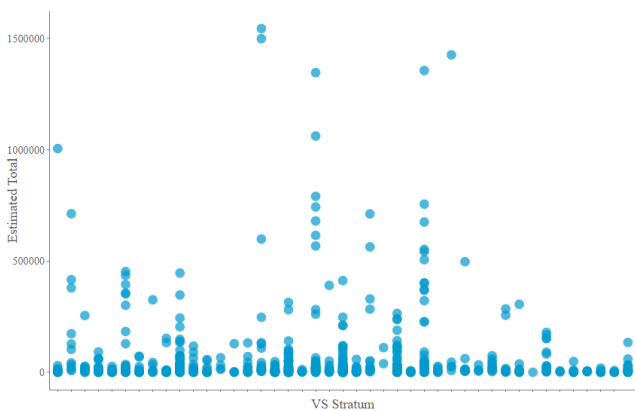
SS Estimated Total by VS Stratum



Sample (SA)

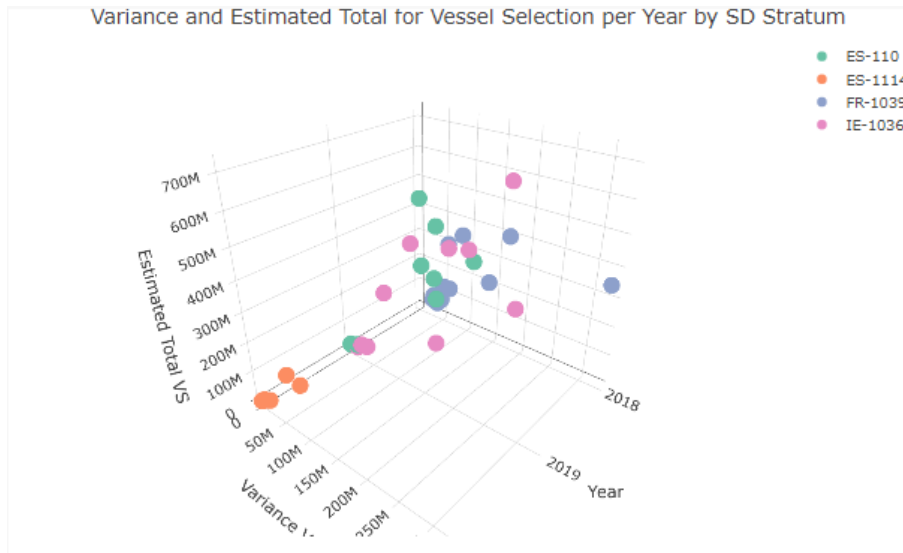
This graph shows the Sample estimated totals by Vessel Selection stratum name.

SA Estimated Total by VS Stratum



3D Plot

3D graph showing the variance and estimated total per Year, coloured by Sampling Details stratum name.



Bias analysis

Regarding the bias analysis work the following deliverable was defined:

- An Rmarkdown script/s that illustrates potential biases in commercial sampling data – including tabular and graphical components. It accepts input data in the ICES Regional Database & Estimation System (RDBES) data format <https://github.com/ices-tools-dev/RDBES>. The deliverable was informed by the work already done on this topic in the “EU request on providing output on evaluating data accuracy (precision and bias) for design-based estimation at a national level” <https://doi.org/10.17895/ices.advice.7641> and other relevant work.

The Data Quality ISSG identified that the most important bias topic is to compare the sampling programme data to the commercial fishing effort and landings data to illustrate its coverage. (However, it should be remembered that any discrepancy between the sampling and fishing effort coverage do not lead to a bias when the sampling is done randomly following a well-designed protocol.) Stock assessors are often interested in the stability of a time series and how the latest year’s data correlates with that time series so information about how the coverage has varied over time was included.

There were 5 main types of sampling coverage considered:

- Country coverage
 - In a regional sampling programmes multiple countries are likely to be sampling data. The national source of the sampling data can be compared to national fishing activity. The relevant variables include sampling country, vessel flag country.
- Spatial coverage



- Compare the spatial information of the samples with the spatial information of the overall fishing activity. The relevant variables considered include: RCG region, ICES division, ICES statistical rectangle, GFCM rectangle, landing ports.
- Temporal coverage
 - Compare the temporal information of when the samples were taken with the temporal information of the overall fishing activity. The relevant variables include: Year, quarter, month, season.
- Technical coverage
 - Compare the technical information for the sampled data with that of the overall fishing activity. The relevant variables include: commercial size category, gear, mesh, level 6/5 metiers, national metiers, by-catch mitigation devices.
- Landings/effort coverage
 - Compare the sample data to the species that were landed, or the fishing effort. The relevant variables include: sampled weight, landed weight, fishing effort.

R functions were developed and the latest version of the code, along with RMarkdown scripts to illustrate their use, can be found at <https://github.com/ices-tools-dev/RDBEScore/tree/main/FishNCo>

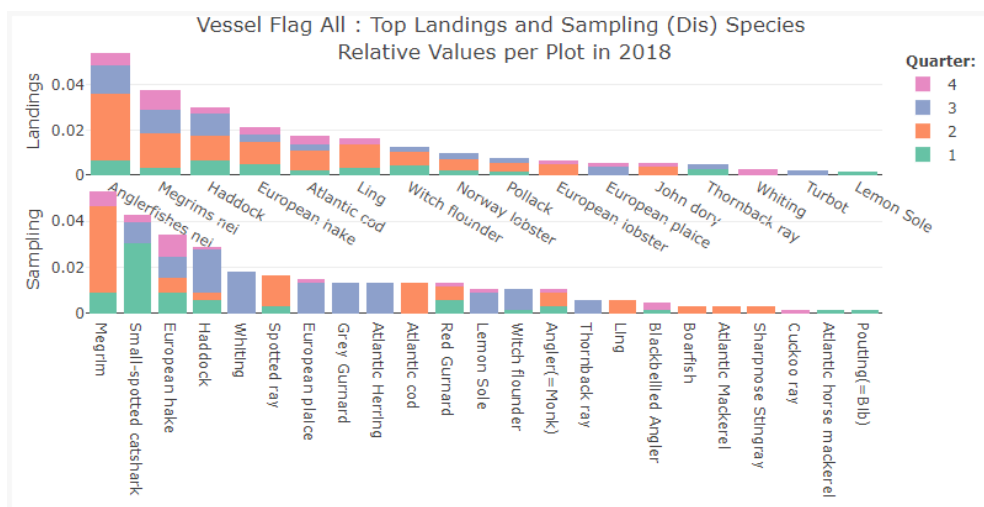
In this section examples of the graphical outputs from these functions are presented – they use test data which has been generated for a fictitious regional sampling program incorporating Ireland, France, and Spain, centred on the island of Ireland.

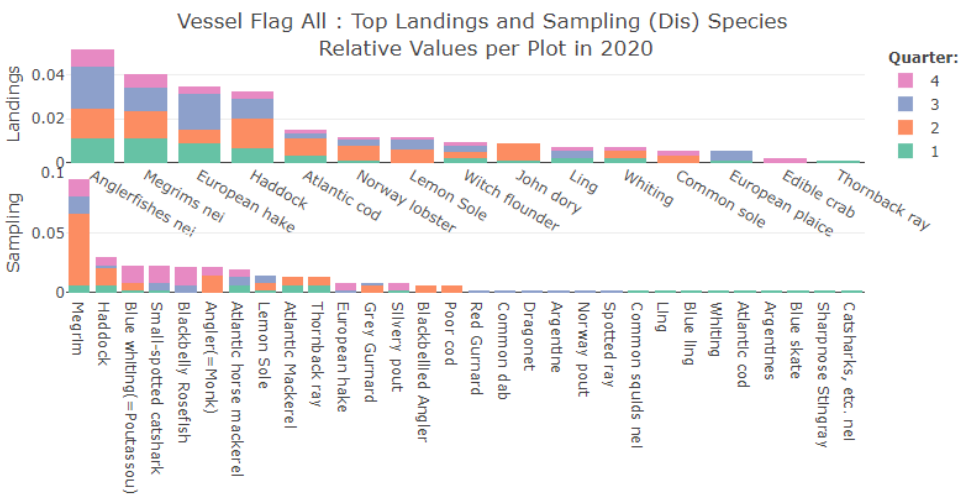
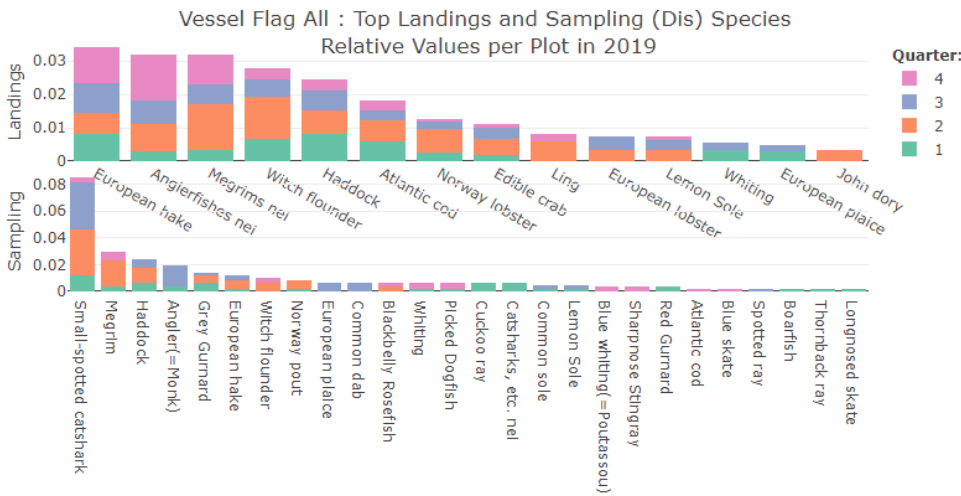
Landings Species Plot - All years

Sampling Catch Category: Discards

This output shows the top species discards for each year in the test data - 2018, 2019 & 2020. The bars represent relative species count values per year. The bars are coloured per quarter.

biasLandings(dataToPlot = testData, var="species", CatchCat = "Dis")



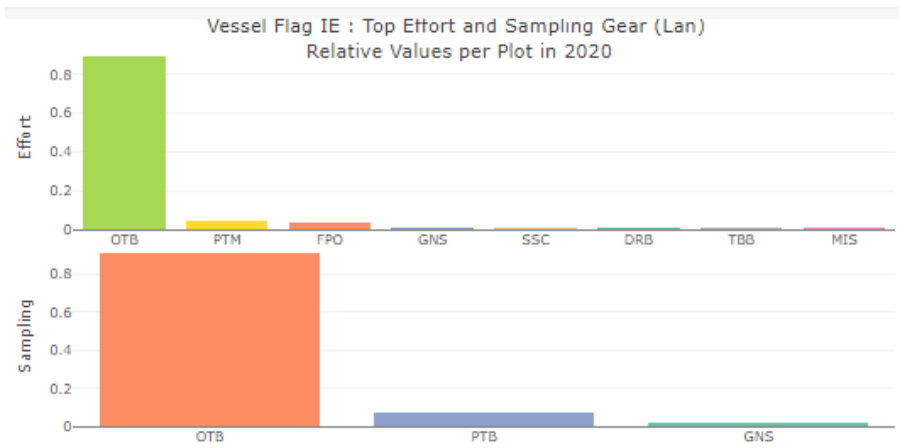


Effort Gear Plot for 2020, Irish Vessels

Sampling Catch Category: Landings

The following output shows the top gear used in effort and sampling landings for Irish vessels in 2020. The bars represent relative gear count values per plot.

biasEffort(dataToPlot = testData, year= 2020, Vessel_flag= "IE", var="gear", CatchCat = "Lan")

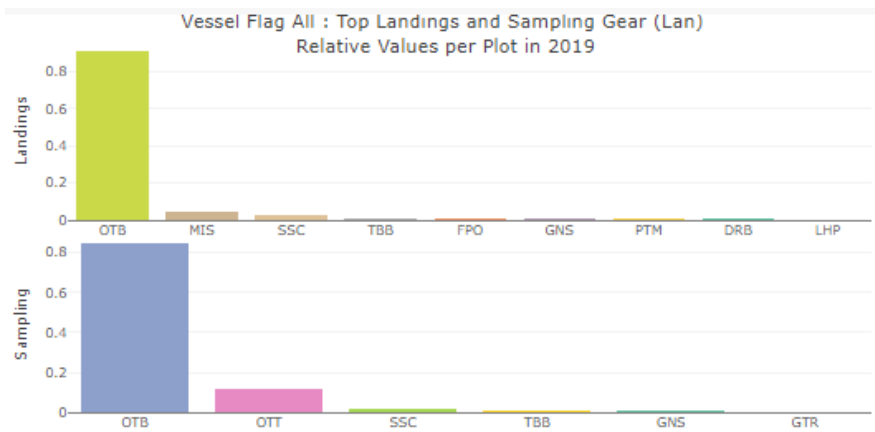


Landings Gear plot for 2018, All Vessels

Sampling Catch Category: Landings

The following output shows the top gear used in landings and sampling landings for all flag vessels in 2019. The bars represent relative gear count values per plot.

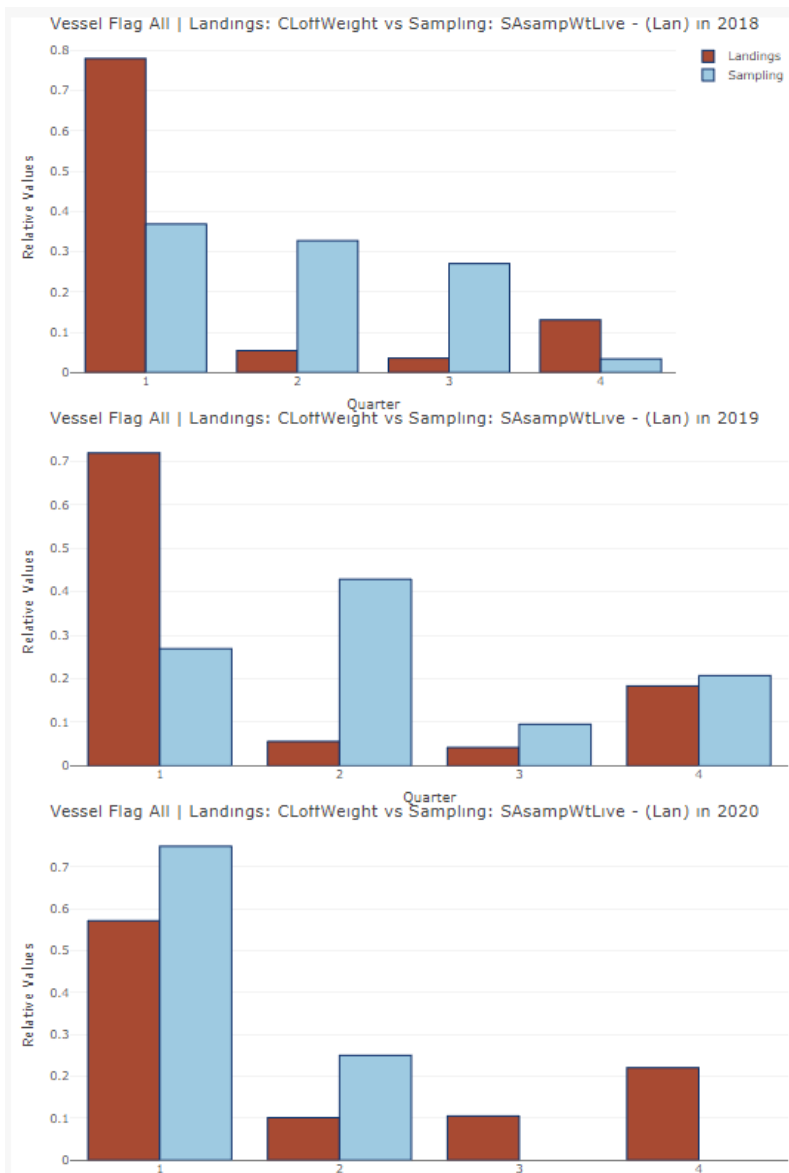
biasLandings(dataToPlot = testData, year=2019,var="gear", CatchCat = "Lan")



Landings Temporal plot for all years - Plotting Landings official Weight vs Sampling Live Weight

Sampling Catch Category: Landings The following output shows the comparison of the landings variable 'CloffWeight' vs the sampling variable 'SAsampWtLive' for sampling landings. Each plot represents data for one year, and the bars show data per quarter. The values are relative per graph.

biasLandings(dataToPlot = testData, CommercialVariable="CloffWeight", SamplingVariable = "SAsampWtLive", CatchCat = "Lan")

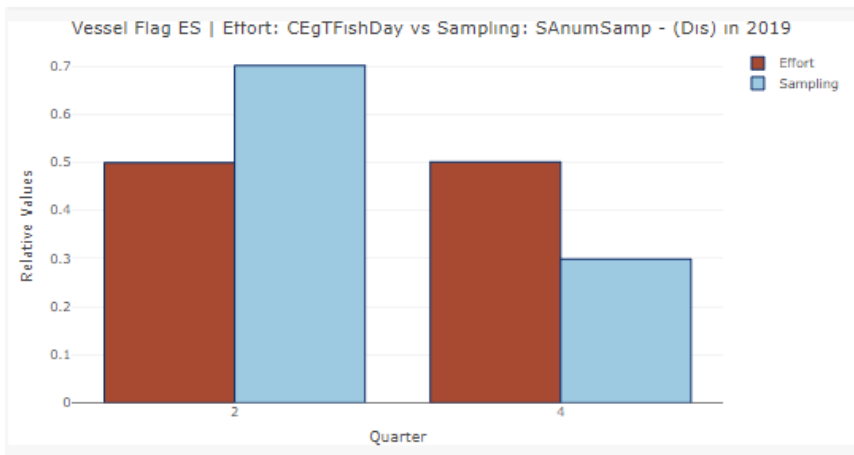


Effort Temporal Plot for 2019, Quarters 1 & 2, Spanish Vessels

Sampling Catch Category: Discards

This output shows the comparison of the effort variable 'CEgTFishDay' vs the sampling variable 'SAnumSamp' for sampling discards in Spanish vessels. The graph shows data for 2019, and the bars show data for quarters 2 and 4. The values are relative per variable for 2019 Q2 & Q4.

```
biasEffort(dataToPlot = testData, year=2019,quarter=c(2,4), Vessel_flag = "ES", CommercialVariable="CEgTFishDay", SamplingVariable = "SAnumSamp", CatchCat = "Dis")
```

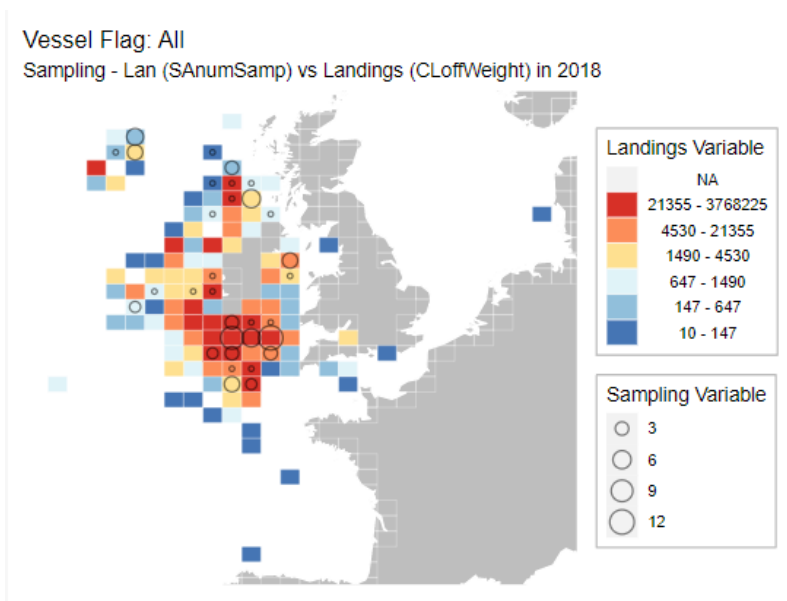


Landings Spatial Plot (Type points) for 2018

Sampling Catch Category: Landings

The below output shows the spatial location of the landings variable 'CLOffWeight' and sampling variable 'SAnumSamp' for sampling landings per statistical rectangle. The rectangles are coloured by the landings variable with red representing a higher value and blue representing a lower value. The circles show the sampling variable with higher values representing a larger circle. This output shows data for 2018.

```
biasLandings(dataToPlot = testData, year=2018,var="Statrec",CommercialVariable="CLOffWeight",SamplingVariable="SAnumSamp", CatchCat = "Lan", SpatialPlot = "Points" )
```

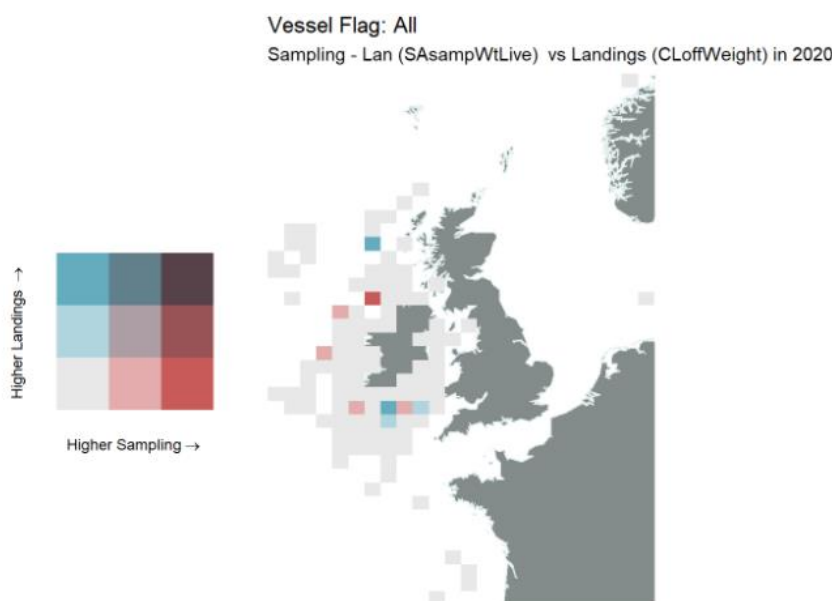


Landings Spatial Plot (Type bivariate) for 2020

Sampling Catch Category: Landings

This bivariate plot shows the spatial location of the landings variable 'CLOffWeight' and sampling variable 'SAsampWtLive' for sampling landings per statistical rectangle. The areas in blue represent higher values for the landings variable. The areas in red represent higher values for the sampling variable. Beige colours show low values for both landings and sampling. Purple shows high values for the landings variable and high values for the sampling variable. This output shows data for 2020.

```
biasLandings(dataToPlot = testData,
year=2020,var="Statrec",CommercialVariable="CLOffWeight",SamplingVariable="SAsampWtLive",
CatchCat = "Lan", SpatialPlot = "Bivariate")
```



Conclusions and Further Work

Objective 1) Produce guidance for Sampling Design

The draft biological data quality document presented here should be server as a basis for the document that is ultimately submitted as part of a regional work programme by the MS participating in the Baltic small pelagic fishery regional programme. Other regional sampling programmes or pilots can also use it as a guide when completing their own documents.

Objective 3) Produce guidance for Data Checks and Objective 6) Produce guidance for Documenting methods of editing and imputing

MS should consider the recommendations presented here and be encouraged to use the templates to support their future national and regional biological quality documents.

Objective 4) Produce guidance for Data Storage

MS have shown a strong willingness to submit data to international databases when they exist. In the cases of the database gap identified then MS are recommended to cooperate with international organisations such as ICES to fill them.

**Objective 5) Produce guidance for evaluating data accuracy (precision and bias)**

WGRDBES-EST will continue to develop the graphical tools produced during this project with the aim being to develop them into a standardised R package that can be used to visualise and explore RDBES data – this will complement the “RDBEScore” package that is already under development by that group.



Draft quality document for Baltic SPF regional pilot

MS: DNK, EST, FIN, LAT, LIT, POL, GER, SWE
Region: Baltic region
Sampling scheme identifier: Baltic SPF regional
Sampling scheme type: Commercial fishing trip
Observation type: Not coordinated
Time period of validity: 2023-2024
<p>Short description:</p> <p>This is a regional sampling program to collect length and age samples from the mixed sprat and herring fishery conducted by commercial vessels operating in ICES Subareas 27.3 using self-sampling, observer sampling or sampling on shore. The aim is to estimate length-composition, catch in numbers by age, and mean weight of fish by age, caught by commercial trawlers by quarter and subdivision.</p> <p>The sampling program is still a trial to test what and how much it is possible to standardize regional sampling and therefore in most countries run in parallel with national sampling programs covering the same fleet / stocks</p> <p>At the moment the sample selection method varies between countries, mainly due to practicalities, but the countries have agreed on standardized protocols for sub-sampling of biological parameters.</p>
Description of the population
<p>Population targeted:</p> <p>Pelagic trawlers participating in the herring and sprat fisheries of Subareas 27.3 – the sampling area is the Baltic sea from Kattegat to northern Baltic: 27.3.a-d.20-29+32.</p> <p>All herring and sprat commercially caught in the Baltic Sea for which estimates of length or age composition is required</p>
<p>Population sampled:</p> <p>The scheme samples fishing trips from the most important Baltic trawlers participating in consumption and industrial small-pelagic fisheries for herring and sprat.</p> <p>In principle all herring stocks and the one sprat stock in the Baltic can be sampled in this sampling program, however, in reality not all MS fleets are covering all the areas, as is indicated in figure 1.</p>

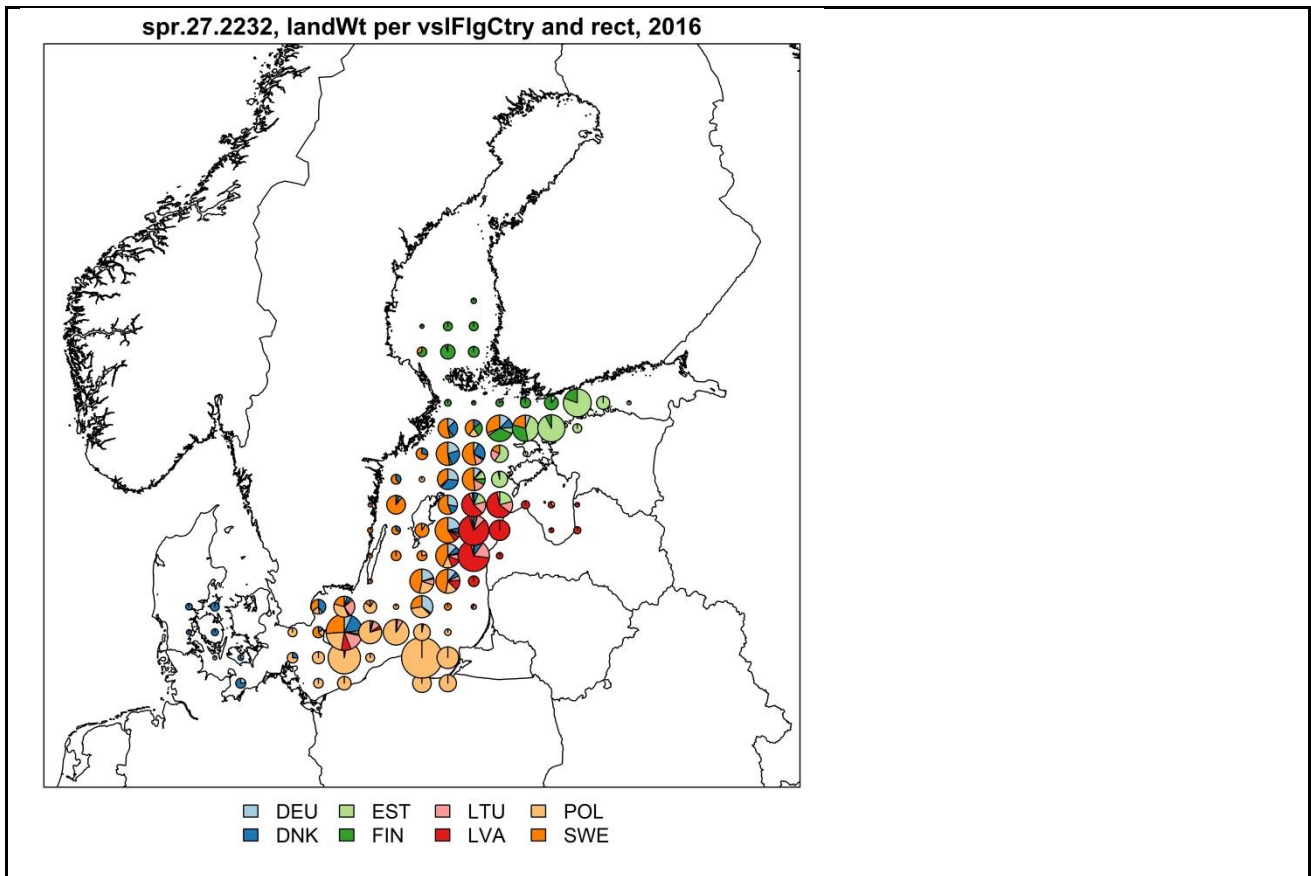


Figure 2 Catch of sprat in the Baltic in 2016 by MS

Stocks covered by MS participating in the Baltic SPF regional program:

Stock	MS
her.27.20-24	DK/SE
her.27.25-2932	DK/FI/EE/LT/LV/PL/SE
her.27.28	LV/EE
her.27.3031	FI/SE
spr.27.22-32	DK/PL/SE/FI/EE/LT/LV

With some national adaptations, the vessel included in 2021 were larger trawlers fishing sprat and herring in the Baltic:

Country	Number of vessels included in the sampling frame
DK	8
SE	15
PL	30
FI	17
LT	13 (5 landing in LT)
EE	24
LV	40
GE	17



In general (with some national adaptations), all vessels below 25 meters, gillnetters landing herring or vessels with a very mixed fishery are **not** covered in this regional program but are instead targeted in national On-Shore sampling programs. This includes gillnetters and smaller trawlers.

The following table gives the identifiers the present national sampling programmes – details can be found in the relevant national workplan <https://datacollection.jrc.ec.europa.eu/wp-np-ar>

MS	Sampling scheme identifier	Sampling frame identifier
DEU	OF Self-Sampling	Baltic herring active 2224
DEU	OF Self-Sampling	Baltic sprat
DNK	Baltic small pelagic RSP	Sprat
EST	OnShoreCommercialPelagic	OSF PEL
EST	OnShoreCommercialPelagicGOR	GOR PEL
FIN	On shore sampling program targeting pelagic trawl fishery of herring and sprat	OTM_SPF
LTU	SO-SEA-COM-SS	BS-TR
LTU	SO-SHORE-COM-SS	BS-TR
LVA	GOR PEL-I (SciObsAtSea)	GOR PEL-I
LVA	GOR PEL-I (SelfAtSea)	GOR PEL-I
LVA	OSF PEL-I (SciObsAtSea)	OSF PEL-I
LVA	OSF PEL-I (SelfAtSea)	OSF PEL-I
POL	Baltic small pelagic RSP	Pelagics_RSP
SWE	CommSelfAtSea - Selected species/stocks	Active SmallPelagics HER, SPR - 27.3.a-d.20-29, 27.4
SWE	CommSelfAtSea - Selected species/stocks	Active SmallPelagics HER, SPR - 27.3.d.24-29

For information the figures below compare herring and sprat landings from 2018 that would be considered in-frame and out-of-frame.

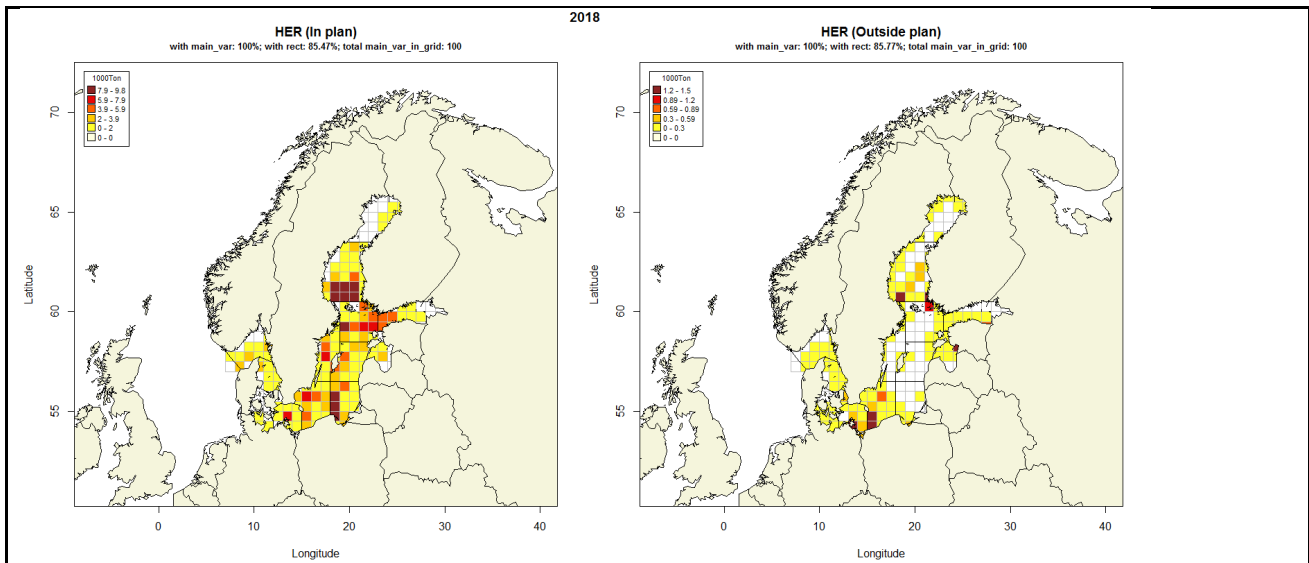


Figure 3 Herring landings inside and outside the regional sampling plan by ICES square based on 2018 data.

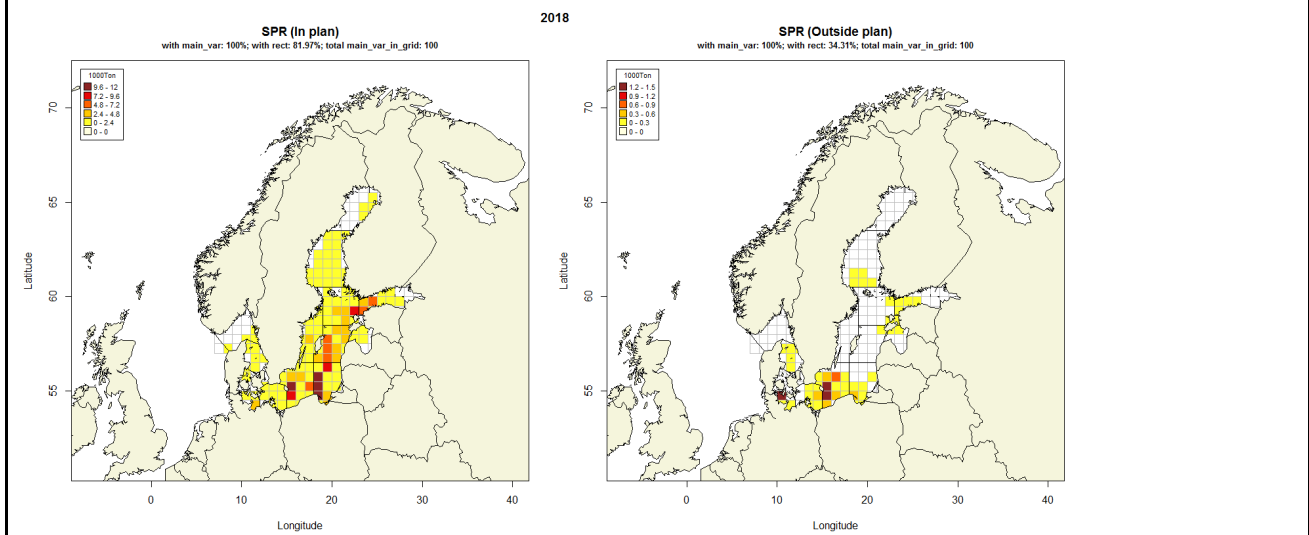


Figure 4 Sprat landings inside and outside the regional sampling plan by ICES square based on 2018 data.

Stratification:

Primary sampling unit are vessel, vessel*trip, weeks or vessel*month, depending on the MS (see details under WGCATCH sampling template). (Add link to WGCATCH sampling template.)

The program is stratified into national lists of vessels. The stratification aims to achieve good spatial coverage over the broad geographical range of the fisheries as well as adequate number of samples and representation of fishing for human consumption and industrial uses. Detailed information on strata by MS can be found the “WGCATCH sampling template”.

Presently there is no consensus on effort allocation. However, based on the 2018 data the table below gives an overview of how many samples by MS could be inside and outside the plan.

	Initial (2018)	In-frame	Out-of-frame
DEU	41	7	34
DNK*	74	0	74
EST	106	8	98
FIN	86	2	84
LTU	8	0	8
LVA	91	7	84
POL	36	12	24
SWE	147	1	146
TOTAL	589	37	552

Figure Numbers of samples in 2018 by MS. 552 samples were used in the allocations

*Danish samples includes landings by other flag countries

Sampling design and protocols

Regional level of ambition: 3 - “Common monitoring strategy”

Present regional level: 1 - “Coordinated data reporting”

Sampling design description:

- Probabilistic sampling design – varies by MS.
- Active trawlers targeting the sprat/ herring fishery.
- The sampling frame is stratified into national vessel lists
- Minimum sampling size (3-5 kg)
- Minimum number of fish per sample for biological analysis (50/ species)
- Vessels outside the regional program are covered by national program

See the WGCATCH sampling template for a more detailed description. (Add link to WGCATCH sampling template.)

Biological sampling protocols:

- A 5 kg random sample is provided from a trip with information on the given haul the sample has been taken from.
- All 5 kg is sorted into species (mainly herring and sprat but other species can be present).
- Random sample of approximately 50 individuals by species is selected for length, weight and age analysis. In some countries, the selection is conducted by measuring the weight of 10 individuals and add fish until the weight of the 10 individuals x 5 has been reach. The length is measured in scm.
- The same individuals as were selected for length are selected for weight measurement. The weight is measured in g. (non-stratified)
- The same individuals as were selected for length are selected for age measurement (non-stratified)



- It is not mandatory in the regional sampling program to collect other biological parameters however, some MS are collecting information on sex, maturity, stomach fullness, parasites and genetics.

Is the sampling design compliant with the 4S principle?

Yes, although this varies by MS

Regional coordination:

Strata of small pelagic sub-scheme that targeting the herring and sprat fisheries with active gears in the Central Baltic: Y

Link to sampling design documentation:

(Add link to WGCATCH sampling template.)

Some additional information:

Danish sampling program was before 2020 an ad hoc sampling program where control agency sampled vessels based on a quota system to cover the main part of the landings. As the main part of the Danish landings in the Baltic are conducted in a few but very large trips this was not the optimal ways of sampling. Since 2020 Denmark has sampled the small pelagic in the Baltic according to the new regional design. This indicates that all larger trawlers > 25 meters are included if they have more than 95% sprat/herring landings. These vessels are all asked to take 1 sample per trip. Further, an additional at land sampling program has been sat in place covering all vessel length. Not all sampling sites are cooperating and refusal rates on landing sites are therefor included. Further species misreporting has occurred back in time, mainly with over reporting of herring and underreporting of sprat. This has been partly compensated for in the data delivery for stock assessment as Denmark for some years used corrected data based on control samples used by month and area on the fleet. It has however not been done systematically back in time. In April 2020 a new and very detailed control system has been emplaced for all industrial landings in Denmark with a very large sampling intensity conducted on every landing, this has improved the quality of the data.

Latvia sampling program. Each year the Fisheries department of the Latvian Ministry of Agriculture prepares the list of vessels and companies that have the fishing permit in the Baltic Sea and the Gulf of Riga. The vessel list consists of information on vessel name, fish species and fishing subdivisions. The vessel list is sorted by fishing type and subdivision to create three segments:

- Pelagic fishery in the Central Baltic (34 vessels in 2021);
- Pelagic fishery in the Gulf of Riga (22 vessels in 2021);
- Demersal fishery (31 vessels in 2021).

Each vessel can be included in one or several segments. Not all vessels that have fishing rights participate in the actual fishery. In the pelagic fishery, six biological samples are collected each month – three samples from the pelagic fishery in the Central Baltic and three samples from the pelagic fishery in the Gulf of Riga. For each segment, fishing vessels are randomly selected from the initial vessel list using Simple Random Sampling Without Replacement (SRSWOR). After the vessel selection, it is checked whether the vessel is active and participates in the fishery of interest. If the vessel is active (according to electronic logbooks), a call is made to the company owner or other contact person to arrange the biological sample or observer

participation for the next trip. If the vessel doesn't participate in the fishery of interest or doesn't fish for other reasons, the next vessel is selected according to the same principles. In case when the random selection of vessels shows the vessel that was already selected in a given quarter, this vessel is ignored and the procedure is repeated. The vessel selection process is documented to ensure the traceability of the process.

The Swedish sampling program was before 2020 a sampling program that relied on quota sample to obtain samples from each subdivision, quarter and fishery type (consumption, industrial) from control and market sources. Given the lack of scientific control over the sampling and uncertainty in the raising totals (possible bias in species position of fleet level totals; alongside possible bias in totals considered as consumption and industrial), bias and precision of final estimates have remained largely non-investigated. Since 2020 Sweden has sampled the small pelagic in the Baltic according to the new regional design, that now is based on probabilistic vessel and trip selection and self-sampling. The <2020 sampling design remained in place but is only used as a last-resort back-up to secure data if refusals threaten data collection itself. The move towards the regional design is expected to significantly improve the quality of the data but its emphasis on the larger industrial vessels now requires special consideration of some smaller vessels fishing for consumption.

Estonia sampling. Is an ad hoc sampling program which aims to collect samples from all active trawlers from each subdivision during active fishing period. During the pilot program in 2020 and 2021 probabilistic sampling scheme was tried (probabilistic selection of vessel), however due to the nuance rich fisheries behavior it was difficult to guarantee that all subdivisions were covered with enough samples. The difficulty laid in the fact that it was hard to predict which vessels were going to fish in which area/stock, especially as subdivision 28.1 (Gulf of Riga) comprises of a separate herring stock. Same vessels can fish both in open sea or in Gulf of Riga, and the fishing location is determined by many variables. Within the framework of regional sampling Estonia will continue to find solutions on how to move to probabilistic vessel selection.

German sampling program. The declining number of vessels in the German pelagic fishing fleets and more automated catch handling processes onboard led to a switch from observer trips to self-sampling in the last few years. Fishermen are providing mixed catch samples following an agreed sampling protocol onboard. Germany is collecting around 20-25 catch samples per year from the relevant fleets, where one sample contains around 50kg of fish. Neither the vessels nor the sampling time however are chosen randomly. Sprat samples are provided by 1-2 trawler, herring is provided by less than 10 trawler that are usually pair-trawling in the main herring distribution areas, thus missing smaller herring populations and fishing areas. Sampling times are fixed to two times per week, but extra samples might be added opportunistically.

Polish sampling program. In 2017 Poland implemented a new sampling design plan, moving gradually from metier based and purely opportunistic sampling towards the plan based on statistics. The sampling scheme for the Baltic Sea region was based on the main types of fisheries exploiting fish stocks subject to sampling requirements, with the use of a combination of at-sea and on-shore schemes, e.g. "Demersal at sea and on shore", "Pelagic at sea and on shore", etc. After three years, in 2020 Poland improved the design and the following approach was applied to a new sampling plan. The stratification has been specified based



on vessels' length category now. To define the sampling intensity per each stratum per quarter, half of the total annual number of samples was distributed proportionally to the quarterly distribution of landings. The second half of the total number of samples was distributed proportionally to the total number of trips. Moreover, Poland has carried out an additional sampling of small pelagics, according to the methodology agreed by the regional subgroup.

Lithuania sampling program. Selection procedure: direct contact with vessel owner to discuss possibility of accepting of observer. 0 (zero) landings in Lithuania, so only sampling at sea possible. Embarking and disembarking of observer in the ports out of Lithuania, therefore logistics (observers travelling) was main limitation for conducting the sampling. Due to travel restrictions in 2020 none of the vessel was selected for sampling. Number of vessels fishing for small pelagic is very small (in 2021 only 13 and only 5 of them have made landings in Lithuania). It makes sampling probability very unequal. Most sprat is landed in Demark, so samples were collected by Danish observers according to the agreement. Since 2021 this agreement started to be replaced by coordinated actions in the framework of this pilot study.

Only landings of herring and sprat for human consumption are made in Lithuania. These fishes are caught by trawls with mesh size more than 32 mm. However, majority of sprat and significant part of herring are landing for industrial purposes out of Lithuania. These fishes are caught by trawls with mesh size 16 -20 mm. Due to it, data on length distribution collected from landings in Lithuania may be different from average total.

Target population is midwater trawlers targeting spart and/or herring. The sampling scheme for herring caught by small scale coastal fleet is running in parallel.

Finnish sampling program. Finnish sampling is based on on-shore sampling program targeting pelagic trawl fishery of herring and sprat. The stocks for sampling are Central Baltic Herring (SD 25-29, 32), Bothnian Sea Herring (SD 30) and Bothnian Bay Herring (SD 31) – the latter two have always belonged to same management unit and to same assessment unit since 2017 as well as the Baltic Sprat stock. Biological data are collected mostly from sampling of commercial trawl fisheries (OTM_SPF and PTM_SPF). Sampling of Herring (and sprat) is based on length stratified sub-sampling scheme, where target number of specimen for biological data is 1/ 0.5 cm length-class/sampled trip (the number of specimens is increased for maturity sampling in spring before spawning time). The herring stock-related biological data (i.e. age-length relation) is used also with the trap-net length distributions – and vice versa.

Finland has started the statistically sound sampling scheme (4S) from the trawl fisheries targeting herring and sprat, where it has been in force from the beginning of year 2019. The selection of PSU for herring (and sprat) is to do random sampling from a draw list, where probability of a fishing unit to be selected for sampling in certain SD and quarter is weighted by its previous years' combined catch of herring and sprat in the same SD and Q. During each quarter the sampling personnel go through the draw list in free order, recording all relevant info (sampling, refusal, out of area, etc.) of the interaction into our sampling database SUOMU, which also has the lottery function needed in the process. Additional lottery draw of PSU's will be done to reach the sampling target if there is a deficit.

Risks and mitigations for the regional sampling program





Different local issues have been presented from different MS. For Lithuania landing sites are often abroad and not easily accessible for observers, this has given some challenges in respect to receive the samples. Further it has not been possible to ask the fisherman to bring the sample back to the home harbour.

In Finland the self-sampling was not possible due to the storing issues onboard the vessels which cause the sample quality to be very poor. Therefore, the Finnish sampling program has been slightly changed to have a similar selection procedure but the sample is taken from the unsorted landings on shore. In Estonia the self-sampling is also not possible due to storing issues onboard the vessels and harbors. In addition, some vessel frequently use abroad landings sites from where it's a challenge to receive a sample.

In Sweden a reduction in sampling of catches for consumption was observed when the regional program was implemented. This reduction was partially related to the sampling frame being dominated by large vessels that fish essentially for industrial purposes. Improved stratification will be implemented in 2022 to reduce this aspect and improve coverage of smaller vessels that remain in the target area and fish for consumption.

A brief summary of the existing time-series:

Time period	Description Denmark
1994 - 2019	Ad Hoc Sampling (NPAH)
2020 – present	Simple Random Sampling Without Replacement (SRSWOR)
	Description Estonia
- present	Ad Hoc Sampling (NPAH)
	Description Latvia
-2016	Ad Hoc Sampling (NPAH)
2017-present	Simple Random Sampling Without Replacement (SRSWOR)
	Description Finland
1974-1997	Simple random sampling on ad hoc basis
1998-2019	Length-stratified random(quota-) sampling on ad hoc basis
2019-2020	Length-stratified random(quota-) sampling on probabilistic basis
2021-present	Simple random sampling on probabilistic basis
	Description Germany
1992 - present	Non-Probabilistic Judgement Sampling (NPJS)
	Description Lithuania
2004-2016	Ad Hoc Sampling (NPAH)
2017-present	Simple Random Sampling With Replacement (SRSWR)*
	Description Poland
2004-2016	Ad Hoc Sampling (NPAH)
2017-present	Simple Random Sampling Without Replacement (SRSWOR)
Time period	Description Sweden
-2019	Ad Hoc Sampling (NPAH)
2020 – present	Simple Random Sampling Without Replacement (SRSWOR)

Further information

More information on this regional sampling program can be found in the 2021 and 2022 RCG reports:

RCG NANSEA RCG Baltic 2022. Regional Coordination Group North Atlantic, North Sea & Eastern Arctic and Regional Coordination Group Baltic. 2022. Part I Report, 101 pgs. Part II Decisions and

Recommendations, 13 pgs. Part III, Intersessional Subgroup (ISSG) 2021-2022 Reports, 159 pgs. (<https://datacollection.jrc.ec.europa.eu/docs/rcg>)

RCG NA NS&EA RCG Baltic 2021. Regional Coordination Group North Atlantic, North Sea & Eastern Arctic and Regional Coordination Group Baltic. 2021. Part I Report, 78 pgs. Part II Decisions and Recommendations, 16 pgs. Part III, Intersessional Subgroup (ISSG) 2020-2021 Reports, 350 pgs. (<https://datacollection.jrc.ec.europa.eu/docs/rcg>)

Compliance with international recommendations:

Yes

Link to sampling protocol documentation:

Online documentation accessible to public will be prepared during 2022-2024.

Some additional information:

RCG NANSEA RCG Baltic 2022. Regional Coordination Group North Atlantic, North Sea & Eastern Arctic and Regional Coordination Group Baltic. 2022. Part I Report, 101 pgs. Part II Decisions and Recommendations, 13 pgs. Part III, Intersessional Subgroup (ISSG) 2021-2022 Reports, 159 pgs. (<https://datacollection.jrc.ec.europa.eu/docs/rcg>)

RCG NA NS&EA RCG Baltic 2021. Regional Coordination Group North Atlantic, North Sea & Eastern Arctic and Regional Coordination Group Baltic. 2021. Part I Report, 78 pgs. Part II Decisions and Recommendations, 16 pgs. Part III, Intersessional Subgroup (ISSG) 2020-2021 Reports, 350 pgs. (<https://datacollection.jrc.ec.europa.eu/docs/rcg>)

Compliance with international recommendations:

Yes

Sampling implementation

Regional level of ambition: 3 - “Common monitoring strategy”

Present regional level: 1 - “Coordinated data reporting”

Recording of refusal rate:

Yes

Refuses and non-responses are recorded. However, as this program is based on self-sampling it is not always straightforward to record if a given sample was collected on the selected trip or from another trip/ haul. Different MS are receiving different refusal rates.

Member state	Vessels in the frame	Refusal rate
DK	8	38%
SE	15	
PL	30	
FI	17	0% (12% couldn't be reached)
LT	5 (landing in LT)	0% only on-shore sampling

EE	24	
LV	40	0%
GE	17	50%

Monitoring of sampling progress within the sampling year:

Routine follow-up meetings are organized between MS are organized minimum 2 a year. At this meeting both the sampling protocols, are reading workshop, species misreporting etc. are discussed.

Data capture

Regional level of ambition: 1 - *“Coordinated data reporting”*

Present regional level: 0 - *“No coordination or not relevant”*

Means of data capture:

Is presently not regionally coordinated

Data capture documentation:

Is presently not regionally coordinated

Quality checks documentation:

Is presently not coordinated, however is planned to be part of the coordination. The BioDataQualityTFA could be used as a common documentation.

Regular international age reading workshops are held but presently no other international data checks are conducted.

72

Data storage

Regional level of ambition: 4 - *“Joint data collection”*

Present regional level: 2 - *“Agreed guidelines”*

National database:

Database name	Location (e.g. host institute)	Format (database / spreadsheet)	Years of data stored
Fiskeline	DTU Aqua	database	1990-present
Fiskdata 2	SLU Aqua	database	
NPZDR	NMFRI (MIR)	database	2004-present
DMAR-01	Thünen-OF	database	2002-present
BIODATA	BIOR	database	2003-present
SUOMU	LUKE	database	2009-present



	EMI-UT	database									
<p>International database:</p> <p>Small pelagic scheme targeting the herring and sprat fisheries: RDB/RDBES at ICES uploaded as common name “Baltic SPF regional” to the RDB-ES</p> <table border="1"> <thead> <tr> <th>Database name</th> <th>Location (e.g. host institute)</th> <th>Format (database / spreadsheet)</th> <th>Years of data stored</th> </tr> </thead> <tbody> <tr> <td>RDBES</td> <td>ICES</td> <td>database</td> <td>2021-present</td> </tr> </tbody> </table> <p>Quality checks and data validation documentation:</p> <p>Common documentation and agreement on relevant national data checks based on RDBES format. (RCG/ FishnCo/ ICES) will be developed</p>				Database name	Location (e.g. host institute)	Format (database / spreadsheet)	Years of data stored	RDBES	ICES	database	2021-present
Database name	Location (e.g. host institute)	Format (database / spreadsheet)	Years of data stored								
RDBES	ICES	database	2021-present								
<p>Sample storage</p> <p>Regional level of ambition: 0 - “No coordination or not relevant”</p> <p>Present regional level: 0 - “No coordination or not relevant”</p> <p>Storage description:</p> <p>Is presently not regionally coordinated</p> <p>Sample analysis:</p> <p>Is presently not regionally coordinated</p> <p>Additional information:</p>											
<p>Data processing</p> <p>Regional level of ambition: 4 - “Joint data collection”</p> <p>Present regional level: 1 - “Coordinated data reporting”</p> <p>Evaluation of data accuracy (bias and precision):</p> <p>Scripts will be developed based on the RDBES data format that make use of common functions being developed by groups such as the ICES WGRDBES-EST.</p>											



Age reading comparison. It has been agreed to quality ensure the age reading on a regional level regular and as a minimum before benchmarks. Dates for last regional age reading exercise via SmartDots indicted in the table per stock

Stock	year	MS
her.27.20-24	2018	Reported in WGBIOP 2018, Annex 3, p 46-47
her.27.25-2932	2022	DK, POL, SWE, GER, LV, LT, EE & FIN
her.27.28	2015	WGBIOP 2017 Report, Annex 5, p 75
her.27.3031	2019	SWE, FIN
spr.27.22-32	2022	DK, POL, SWE, GER, LV, LT, EE

Editing and imputation methods:

A design-based estimator is under development. Documentation will be available in RDBES scripts and outputs when that system is in production.

Quality document associated to a dataset:

Documentation will be available in RDBES scripts and outputs when that system is in production.

Link to estimation documentation;

Documentation on estimation will be made available using the WGCATCH common estimation template https://github.com/ices-eg/wg_WGCATCH/blob/master/templates/WGCATCH_estimation_template.xlsx

Validation of the final dataset:

Final validation takes place when data is compiled at ICES stock coordination level.

Data quality control practices of European fisheries institutes

An analysis of the data quality control practices of European fisheries institutes for data checks, editing and imputation

Introduction

The aim of this survey was to collect information on the data checking, editing and imputation practices of 18 different fisheries institutes across the EU. Under the 'Biological Data Quality' thematic working area of the FishNCo project, the collection of this data will aid in the strengthening of EU fisheries data collection by developing Regional Work Plans for the EU Regional Coordination Groups (RCG). In addition to the collation and analysis presented in this report, the data collected in these questionnaires will also aid in the production of a data quality process template. This template will allow members of RCG's to record, efficiently and concisely, any data checking, editing or imputation process they implement in the future.

The questionnaire itself is composed of 5 sections. Sections 1 (not published) and 2 (Respondent information) collected information about the respondents, their respective roles their institutes. Section 3 (Data checks) collects information on if, when and how data checks are performed during the data collection process. Section 4 (Data editing) collects on any how inconsistencies, errors or discrepancies are dealt with during the data collection process. Section 5 (Imputation) collects information on how gaps in Age length Keys (ALK's), Weight length Keys (WLK's) and sampling strata are addressed during the data collection process.

Objectives

The objectives of this report are as follows:

1. Collect, collate, and categorise data on data checks, editing and imputation performed by EU fisheries Institutes during the collection of fisheries data.
2. Summarise and analyse the collected data to determine if, when and how such checks are performed by EU fisheries Institutes.
3. Present the collected data and analysis in report which clearly and concisely communicates the observed results.
4. Use the summary and analysis conducted to create a data quality control checks, editing and imputation template to be used in the collection of fisheries data by EU fisheries Institutes.

Methodology

A questionnaire composed of 5 sections was composed. Respondents were asked to respond using free text answers and to include diagrams, images, and written guidelines where relevant and possible. These questionnaires were distributed on the 25/05/2021. After responses were received, all responses were collated in a spreadsheet, with each columns representing a question and each row the response from a specific institute. A response cut-off date of 22/05/2021 was set and responses received following this date are not included in the analysis.

A duplicate matrix was then created, and inductive categorisation was used to categorise responses. For questions 3.2, 3.4,3.5,3.6,3.7,3.8 and 3.9, answers were broken down into three sections 1) Whether the check was performed, 2) At what point in the data capture process was the check performed and 3) How the check was performed. For all other questions (Section 2, Q3.2,3.10,3.11, Section 4 and Section 5), respondents answers focused on describing the check, editing or imputation process address in the question.

The results were then plotted using R version 4.04 (R Core Team, 2021) and the 'ggplot2' package (Wickham, 2016). For each question, a plot showing the frequencies of each categorised answer, a prose analysis of the responses with supporting quotes, and a table showing how each respondent was categorised was presented.



The findings of the analysis and recommendations were then summarised in the conclusion. These findings were then incorporated into a data quality control template, which will allow users to record the time, type and method of data quality control checks they implemented in future. Each check was categorised based on data properties presented by West (2011).

Caveat: While every effort has been made to ensure as much detail of respondent's answers was captured, categorisation of textual data necessitates some reduction in data resolution. Full, uncategorised responses are available in the relevant appendices, and users are encouraged to refer to these for greater detail and clarity where required.

Glossary of terms

Data collection method

EDC: Electronic data capture, usually by means of an electronic measuring board (in the case of fish) or callipers (in the case of *Nephrops*).

Point of data collection

Ad-Hoc: Checks are only performed as necessary during the data collection process, but not on a regular basis or at a defined point in the process.

Data capture: The recording of data, either manually on paper or by means of electronic data capture.

Data entry: The inputting of paper transcribed data to a temporary digital workbook such as an excel sheet/Microsoft access database.

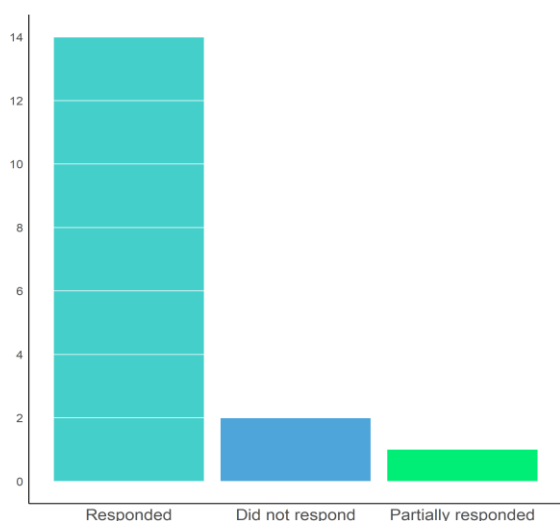
Data import: The transfer of data from temporary digital workbooks/databases to the primary database. After being imported to the primary database the data should be ready for extraction.

Data extraction: The withdrawal of data from the primary workbook, usually in response to data calls for RCG's, WG's etc.

R scripts: Can refer to R markdown documents (.Rmd) or Simple R files (.R)

Response Rate

Of the 18 institutes asked to complete the survey, 15 responded within the timeframe, one responded late and two did not respond. The names, acronyms, and response status of all those contacted are detailed in table 1. Of the three institutes who did not complete the questionnaire, two were unable to be contacted, while one replied when contacted but stated they would not be able to complete the questionnaire in the allotted timeframe. As a result, the response of IPMA (Portugal) has been added as an appendix to the report, but their answers have not been included in the analysis.



s surveyed for this report.

his report.

	Acronym	Response status
	DRP-RAA	Responded
ndbouw- en	ILVO	Responded
larine Institute	FEAS-MI	Responded
	AZTI	Responded
Institute of Food Safety, Animal Health and Environment	BIOR	Responded
Instituto Español de Oceanografía	IEO	Responded
Luonnonvarakeskus	LUKE	Responded
National Marine Fisheries Research Institute	NMFRI	Responded
Stichting Wageningen Research	WMR	Responded
Swedish University of Agriculture and Sciences	SLU	Responded
Technical University of Denmark	DTU	Responded
Thuenen Institute	THN	Responded
Klaipeda University	KU	Responded
University of Tartu Estonian Marine Institute	EMI	Responded

Institut de Recherche pour le Développement	IRD	Did not respond
Institut Français de Recherche pour l'Exploitation de la Mer	IFREMER	Did not respond
Instituto Português do Mar e da Atmosfera	IPMA	Responded (not analysed)

Results and discussion

Section 2 – Institute information

The questions in section 2 were aimed at gathering basic information about the respondents. Respondents were asked 1) What countries they worked in, 2) What lab or institute they worked in, 3) Whether their lab or institute had any relevant accreditations or certifications and 4) What data they thought about when completing sections 3,4 and 5.

As the questionnaire covered a range of topics in the data collection process, most responses required the input of personnel in various roles e.g. Data manager, Database administrator, Sampling co-ordinators, Onboard observers. Where institutes stated clearly which answers had been offered by different personnel,



their responses were separated into two different responses, indicated by the Institute abbreviation followed by a or b (e.g IEO(a)). Institutes whose responses were separated in this way were: Instituto Español de Oceanografía, Stichting Wageningen Research, Swedish University of Agriculture and Sciences and Technical University of Denmark.

Q2.1 Which country do you work in?

A map showing the country of origin (q2.1) and response status of all those contacted can be seen in figure 2. As a response was received from IPMA (Portugal) following the response deadline, the response was included in the appendices, but was not included in the analysis. Hence, Portugal was categorised as 'Partially responded'.

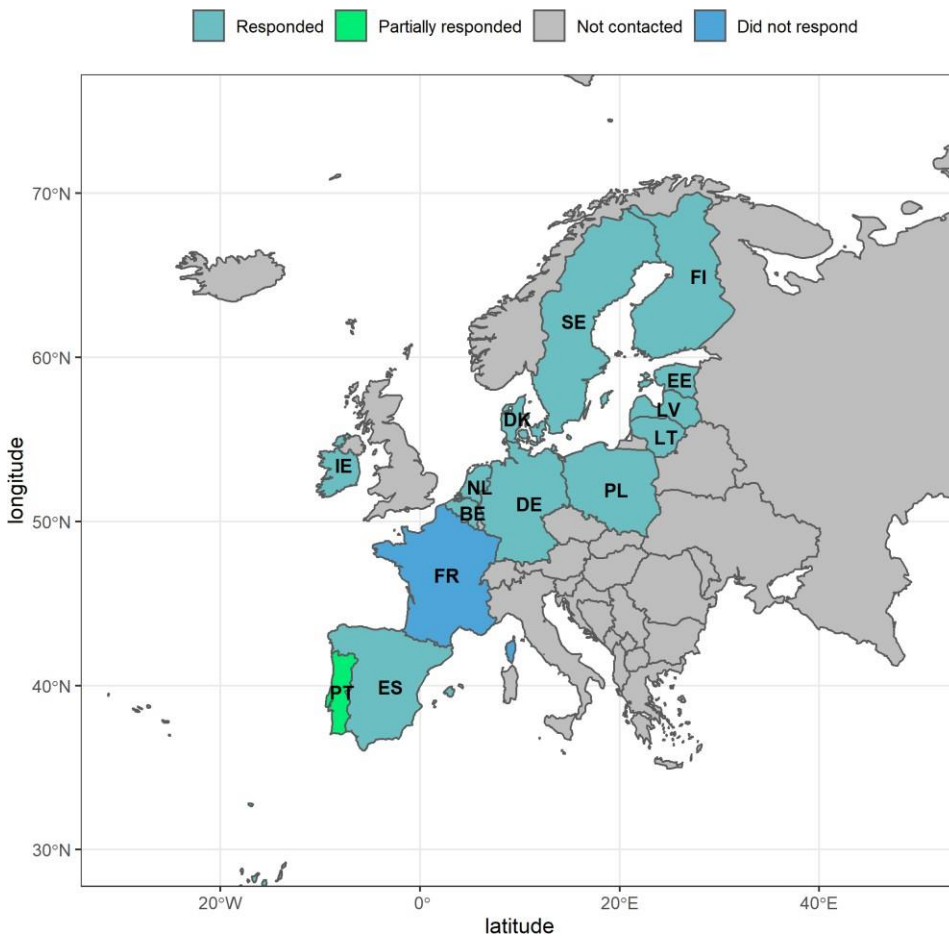


Figure 2: Map showing response status by country (indicated with 2 letter alpha codes) of institutes surveyed for this report

Table 2: Response status, 2 letter alpha code and country name of all institutes surveyed for this report.

Country		Institute
BE	Belgium	ILVO
DE	Germany	THN
DK	Denmark	DTU(a)

Country		Institute
		DTU(b)
EE	Estonia	EMI
ES	Spain	IEO(a)
		IEO(b)
		AZTI
FI	Finland	LUKE
FR	France	IFREMER
		IRD
IE	Ireland	FEAS -MI
LT	Lithuania	KU
LV	Latvia	BIOR
NL	Netherlands	WMR(a)
		WMR(b)
PL	Poland	NMFRI
PT	Portugal	IPMA
	Portugal – Autonomous Region of the Azores (RAA).	DRP-RAA
SE	Sweden	SLU(a)
		SLU(b)

Q2.2 Which institute or laboratory do you work in?

Q2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

The name of the lab or institute in which respondents worked can be found in table 3 in addition to any relevant certifications or accreditations the lab or institute holds. Only five respondents listed relevant certifications, four of which were ISO accreditations. The only other accreditation listed was IODE accreditation.

Table 3: Full name and relevant accreditations of all respondents.

Institute (Short)	Institute/Lab	Relevant certifications
AZTI	AZTI	No

Institute (Short)	Institute/Lab	Relevant certifications
BIOR	Institute of Food Safety, Animal Health and Environment "BIOR", Fish resources re-search department, Marine laboratory.	No
DRP-RAA	Regional Directorate for Fisheries in the Azores (DRP/RAA).	No
DTU(a)	DTU Aqua	No
DTU(b)	DTU Aqua	No
EMI	Estonian Marine Institute, University of Tartu	No
FEAS -MI	Marine Institute, Fisheries Advisory & Ecosystems Services	IODE accreditation
IEO(a)	Instituto Español de Oceanografía (IEO).	No
IEO(b)	Centro Nacional INSTITUTO ESPAÑOL DE OCEANOGRAFÍA (IEO, CSIC).	No
ILVO	ILVO Marine research (Flanders research institute for agriculture, fisheries and food.)	ISO 17025
LUKE	Natural resources institute Finland, Luke	No
NMFRI	National Marine Fisheries Research Institute in Gdynia, Poland	No
SLU(a)	Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences	No
SLU(b)	Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences	No
THN	Thünen Institute of Sea Fisheries	NA
KU	Marine Research Institute of Klaipeda University	ISO 14001, ISO 45001, ISO 9001
WMR(a)	Wageningen Marine Research.	ISO 9001
WMR(b)	Wageningen Marine Research.	ISO 9001

Q2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

When asked about which data they considered while completing this survey, respondent answers ranged from general ("Fish stock rather", KU) to extremely specific (see IEO(b), table 3). Due to broad nature of the answers, categorisation was not used, and full respondents' full answers can be found in table 4.

Table 4: Sampling schemes or stocks which respondent considered while completing this questionnaire.

Institute Q2.4 (Data considered when completing the survey)	
AZTI	Data from our sampling schemes (at the market and on board) and official data corresponding to ICES areas
BIOR	Data from Baltic Sea demersal trawlers
DRP-RAA	All relevant stocks and sampling schemes are monitored from commercial fisheries in ICES Division 10a2 (Azorean fleet).



DTU(a)	a) Estimated amount of discard for different ICES assessment WG's
DTU(b)	b) Estimated age distribution of landings of commercial stocks for different ICES assessment WG's, where the sampling is stratified per commercial size categories
EMI	Stock assessment-related data for Baltic herring (Central Baltic Herring and the Gulf of Riga herring stocks), and the Balticsprat in Sd. 22-32
FEAS -MI	<p>The questions are answered for the Demersal Catch Sampling At-Sea programme, which follows the flow of data collected during an at-sea sampling programme from collection to analysis to reporting.</p> <p>The Demersal Catch Sampling At-Sea programme is comprised of demersal at-sea and Nephrops at-sea sampling. The Nephrops at-sea sampling has similar but slightly different protocols to the demersal at-sea. Landings data from at-sea sampling is uploaded to the Stockman database.</p>
IEO(a)	Data from our length sampling programme, both market and on-board, in the ICES area under the DCF/EUMAP. Tuna fish-eries excluded.
IEO(b)	<p>The biological variables data (Fisheries independent data) on the stocks for the ICES Area are carried out according to 2 differentiated sampling designs, depending on the biological characteristics of each species:</p> <p>- Small pelagic species: the sample/subsample is selected by a Simple Random Sampling (SRS). The sample is entirely bio-logically analyzed (various biological variables are collected on each sampled fish until the expected number of samples is reached).</p> <p>Engraulis encrasicolus (ane.27.8),Micromesistius poutassou (whb.27.1-9No14),Sardina pilchardus (pil.27.8c9a),Scomberscombrus (mac.27.nea),Scomber colias 8, 9, Trachurus trachurus (hom.27.2a4a5b6a7a-ce-k8),Trachurus trachurus (hom.27.9a),Engraulis encrasicolus (ane.27.9a),Sardina pilchardus (9as),Scomber scombrus (9as)</p> <p>- Demersal and benthic species: the sample is stratified by length classes. A Simple Random Sampling (SRS) is applied for the selection of the samples in each length stratum. A fixed number of specimens from each length class is biologically sampled and various biological variables are collected on each individual. The sample attempts to represent the full length range of the catch, so the least abundant length classes are preferably selected for sampling.</p> <p>Lepidorhombus boscii (ldb.27.8c9a),Lepidorhombus whiffiagonisboscii (meg.27.7b-k8abd),Lepidorhombus whiffiagonis- boscii (meg.27.8c9a),Lophius budegassa (ank.27.78abd),Lophius budegassa (ank.27.8c9a),Lophius piscatorius (mon.27.78abd),Lophius piscatorius (mon.27.8c9a),Conger conger (all areas),Helicolenus dactylopterus (all areas),Merluc- cius merluccius (hke.27.3a46-8abd),Merluccius merluccius (hke.27.8c9a),Molva molva all areas (lin.27.3a4a6-9No14),Phy- cis blennoides all areas (gfb.27.nea),Trisopterus spp all areas (T. luscus)</p> <p>The samples of the following species usually come from surveys although could be occasionally sampled from commercial</p>

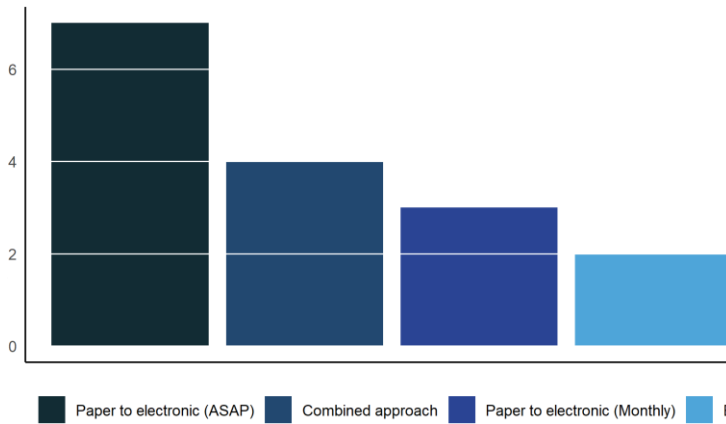
Section 3 – Data Checks

Section 3 of this questionnaire asked respondents about what data checks they implemented during the data collection process, when they performed these checks, and how they performed these checks. In addition, it asked respondents about their data collection methods and about any relevant guidelines or written processes they had with regards to data checks.

Q3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)



When is the data input into an electronic recording system?



or to inputting it into an electronic recording system immediately or survey (“captured on paper and then inputted into an electronic recording system”). Four respondents employed a combined approach using paper before inputting the data electronically as soon as possible.

Where a combined approach was used, EDC was often employed when sampling *Nephrops norvegicus* and paper transcription for other samples (“The only electronically device used in our commercial sampling is a calliper used for measuring the carapace length (mm) of *Nephrops* and shrimps. Everything else is captured on paper and entered in our national database as soon as possible”, DTU(a)). Paper transcription with monthly digitisation of data was employed by IEO(a,b) and DRP-RAA. Finally, some institutes (ILVO, SLU) use EDC exclusively for data collection (“seagoing observers register sample data at sea directly in the database using a custom developed Smartfish application. The application is run on a rugged tablet coupled to an electronic measuring board.”, ILVO).

Categorised answers can be found in table 5, while full answers for each country can be found in the relevant appendix. Where countries employed a combined approach, the primary method and secondary method are listed into table 5.

Table 5: Categorised answers of all respondents to Q3.1

Institute	Method	Primary	Secondary
AZTI	Paper to electronic (Monthly)	NA	NA
DRP-RAA	Combined approach	Paper to electronic (Monthly)	EDC
EMI	Paper to electronic (Monthly)	NA	NA
FEAS -MI	Combined approach	EDC	Paper to electronic (ASAP)
IEO(a)	Paper to electronic (Monthly)	NA	NA
IEO(b)	Combined approach	Paper to electronic (Monthly)	EDC
IFSAHE	Paper to electronic (ASAP)	NA	NA
ILVO	EDC	NA	NA
LUKE	Paper to electronic (ASAP)	NA	NA
NMFRI	Paper to electronic (ASAP)	NA	NA
SLU(a)	Paper to electronic (ASAP)	NA	NA
SLU(b)	EDC	NA	NA
THN	Paper to electronic (ASAP)	NA	NA
KU	Paper to electronic (ASAP)	NA	NA
WMR(a)	Combined approach	EDC	Paper to electronic (Annually)
WRM(b)	Paper to electronic (ASAP)	NA	NA

Q 3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).



Do you constrain the values of properties in your data recording system to be physically realistic?

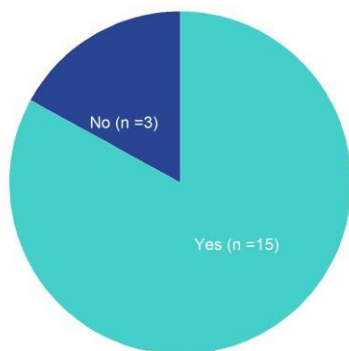
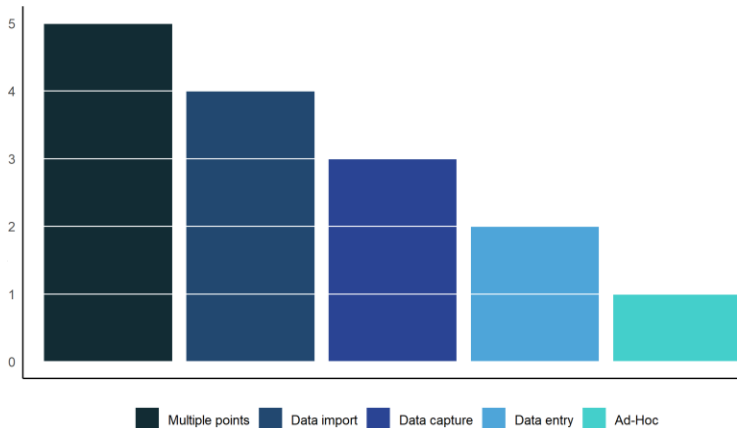


Figure 4: Frequency of responses to Q3.2 – were checks performed?

When asked whether values of properties were constrained in their data recording system, the majority of respondents (n = 15) answered yes. Only three respondents (AZTI, IEO (b), LUKE) answered no. However, while values were not constrained, two of those who answered no (IEO(b), AZTI) did check the data prior to data extraction (“No for most of the stocks, however data are checked just after data extraction.”, IEO (b)).

At what point do you perform these checks?



Form the check?

Of those who answered yes (n = 5) carried usually both at the point of data capture and outlier search, plotting boxplots or histograms, and data extraction”, THN.). Four respondents carry any database (“Data is checked against common out of range errors at the step of entering into the database.”, NMFRI). Three performed the check at the point of data capture (“There is a constrain for extreme values on age, length and weight by species in the data recording system (during data capture).”, WMR(a), and two (DTU(a), DTU(b)). at the point of data entry. Only one institute constrained values to be physically realistic on an Ad-Hoc basis (EMI).

Table 6: Categorized responses of all institute to Q3.2 – if they perform the check and when they perform the check.

Institute	Check performed	Point of check
AZTI	No	NA
BIOR	Yes	Multiple points
DRP-RAA	Yes	Multiple points
DTU(a)	Yes	Data entry
DTU(b)	Yes	Data entry
EMI	Yes	Ad-Hoc
FEAS-MI	Yes	Multiple points
IEO(a)	Yes	Data import
IEO(b)	No	NA
ILVO	Yes	Data capture

LUKE	No	NA
NMFRI	Yes	Data import
SLU(a)	Yes	Data import
SLU(b)	Yes	Multiple points
THN	Yes	Multiple points
KU	Yes	Data import
WMR(a)	Yes	Data capture
WMR(b)	Yes	Data capture

When asked to describe the type of constraints they had in place, 12 respondents constrained values to be within a reasonable range. This could apply to fish length (“measurements must between 3.01mm to 99.99mm”, FEAS-MI.), weights (“.. individual weight between 1 – 50000 grams, etc”, KU), or non-biological variables (“Some of the numeric fields in our national database has constrains, so only realistic values can be entered e.g. wind direction”, DTU (a)). Three respondents constrained their data entry such that the user could only choose from pre-defined lists, limiting the entry of incorrect or unrealistic values (“The data file contains predefined values that can be assigned to the following biological parameters: sex and maturity. At the top of the datasheet 10 rectangles are located. For each rectangle excel macro is assigned. We are using a 6-scale maturity scale. Sex is defined as numbers, 1 is male and 2 is female. In the rectangles all combinations of sex and maturity are predefined..”, BIOR). Three respondents had physically realistic constraints in place with regards to catch and sample weights, usually checking that sample weight was not greater than catch weight (“Sample weights are checked by comparing the length frequency of the sample and sample weight cannot be larger than the total weight”, SLU(b)). Finally, three respondents had input restrictions on their database, where users were prevented or warned by the data entry software when erroneous or missing values were present (“A general species-specific length-weight key check is applied for every weight registration (sample and individual weight). A notification is displayed for an abnormal weight. The user can reject the notification or choose to change the initially registered weight..”, ILVO. , “Our Commercial Port Sampling Application (Stockman) contains data validation ensuring required fields have been entered i.e. Sampling Place, Landing Port Sampler...”, FEAS-MI.)

A summary of the constraints in place by respondents can be seen in table 7. Full details can be found in the relevant appendices.

Table 7: Method of constraints used by all respondents in Q3.2

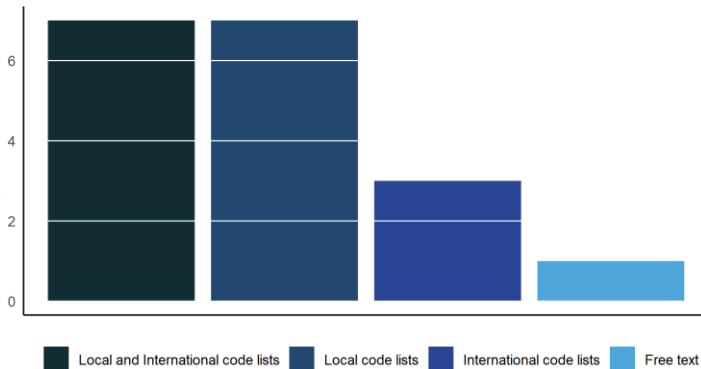
Institute	Reasonable range	Pre-defined lists	Catch and Sample weights	Input restrictions
BIOR	X	X		
DRP-RAA	X		X	
DTU(a)	X			
DTU(b)	X			
EMI	X			
FEAS-MI	X		X	X
IEO(a)	X			
ILVO	X	X		
NMFRI	X			
SLU(a)	X		X	X
SLU(b)	X		X	X
THN				
KU	X	X		



Institute	Reasonable range	Pre-defined lists	Catch and Sample weights	Input restrictions
WMR(a)				
WMR(b)	X			

Q 3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Do you use defined code lists for storing categorical information electronically?



de list to store categorical data during the data <clusively international code lists. FAO codes (“International 3-letter code (FAO code) aries.”, BIOR).

al information on these lists was recorded in the questionnaire, with respondents usually only stating that they used local code lists (“Yes, local code lists.”, WR(b)).

Six respondents used a combination of local and international code lists (“local/working and ICES codes”, KU., “Nearly all of the codes lists are local, but the most relevant ones, species, area etc., have a field with International codes”, DTU(a). International lists were again drawn from either ICES or FAO codes, however two respondents (FEAS-MI, AZTI) also use codes from the world register of marine species (WoRMS)

Only one respondent did not use code lists as the primary means of recording categorical data, although code lists were used for some information (“This depends on categorical information, e.g. areas, gear and metier are defined as in ICES vocabularies. Otherwise mostly free text.”, EMI).

Categorised responses by Institute can be seen in Table 8.

Table 8: summarised responses of all respondents to Q3.3

Institute	Q3.3
AZTI	Local and International code lists
BIOR	International code lists
DRP-RAA	Local code lists
DTU	International code lists
DTU	Local and International code lists
EMI	Free text
FEAS -MI	Local and International code lists
IEO	Local code lists
IEO	Local code lists
ILVO	International code lists
LUKE	Local and International code lists
NMFRI	Local code lists
SLU	Local code lists
SLU	Local code lists

THN	Local and International code lists
KU	Local and International code lists
WR	Local and International code lists
WR	Local code lists

Q3.4 Do you perform any outlier checks on your data? If yes, please explain:

Q3.4.1 Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

All respondents (n = 18) stated that they did perform outlier checks on their data. In terms of properties checks, all respondents checked biological properties for outliers, including length-weights (“Yes. Analysis and detection of outliers for biological parameters, their weight-length relationships and ranges.”, IEO(b)), length-age (“biological parameters i.e. length-weight, length-age.”, LUKE) and maturity (“Number of individuals length, Age range, Length range, Sex ratio, Maturity stage”, WMR(b)).

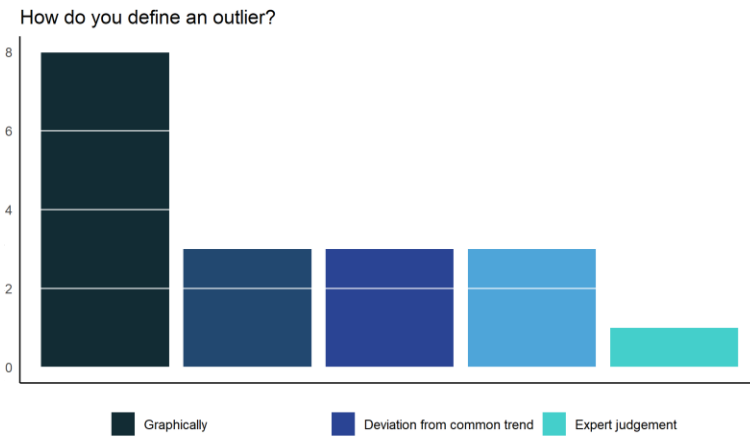
Other properties commonly checked for outliers included discard weights per haul (“Discards weights per haul and species compared to an estimated weight based on the length distribution of the sample (Routine)”, DTU(b)) and catch and sample weight (“Unexpected sample weights; High raising factors; Missing raising factors; Negative discards (discard weight larger than total catch weight); Sample weight larger than total discards”, FEAS-MI.). Some respondents also checked census data (“We do check length distributions, landings, etc...”, IEO(a)), discard rates, spatial data (“Positions have been visualised on a map, haul duration has been checked using Microsoft Power Bi”, ILVO) and Haul or trip information (“Excessive tow length or fishing speed; Zero tow length; Impossible or unexpected shoot or haul positions; Short tow duration; Negative tow duration...”, FEAS-MI). Most respondents checked a combination of these properties, as can be seen table 9.

Table 9: Properties checked for outliers by respondents.

Institute	Biological parameters	Discard weights per haul	Catch and sample weights	Census data	Discard rates	Spatial data	Haul data
AZTI	X		X				
BIOR	X						
DRP-RAA	X						
DTU(a)	X	X	X				
DTU(b)	X	X	X				
EMI	X						
FEAS -MI	X	X	X	X	X	X	X
IEO(a)	X						
IEO(b)	X						
ILVO	X		X			X	X
LUKE	X						
NMFRI	X		X				
SLU(a)	X	X	X				
SLU(b)	X						
THN	X	X			X		

KU	X					
WMR(a)	X			X		X
WMR(b)	X			X		X

Q3.4.2 How do you define an outlier?



define outliers.

so graphically. Of these eight, some specified ships and boxplots between biological variables I(a).), while others did not (“Visual, extreme variation”, DTU(a)). Boxplots, histograms, and graphs used.

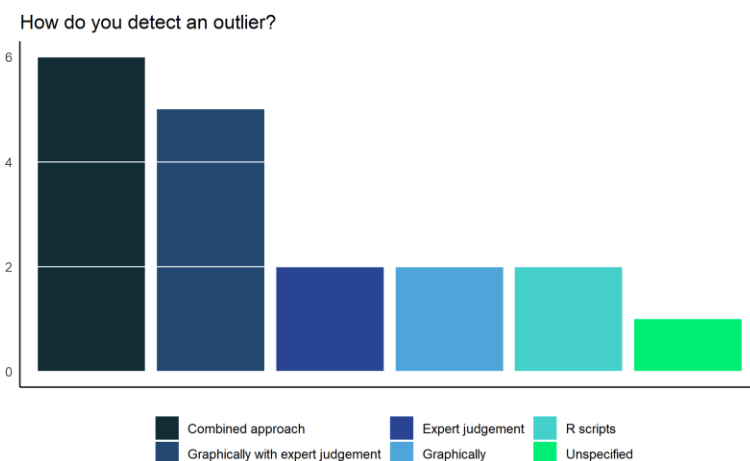
Three respondents defined outliers through comparison with historical data (“Outliers are defined by comparison to historical data. Points that fall outside 95% of historical data points are considered to be outliers.”, FEAS-MI, “comparison of discards rates over the years”, THN).

Three defined outliers as observations deviating from a common trend, however they did not specify if this trend was observed visually, numerically or through expert judgement (“Value far apart from other values or values that are frequently the result of an error”, IEO(a), “An observation is considered an outlier when it deviates significantly from a commontrend of observations in the same group.”, NMFRI).

Three respondents defined outliers numerically, using either Fultons coefficient (“After entering the weight that does not match the settings (“Fulton’s coefficient is >2 or less than 0.5), cell is coloured in red and additional data checking is performed.”, BIOR), Cookes distance (“For length and landings we use Cook distance to detect outliers”, IEO(b)) or residuals following modelling (“Exp of residual is less than 0.5 or more than 2”, KU).

One respondent defined an outlier based on expert judgement, however no further information was offered (“According to expert experience.”, EMI).

Q3.4.3 How do you check for outliers? (e.g. graphically using expert judgement, R scripts)



used to detect outliers.

respondents utilised a combined approach, using a for outliers. The combined methods of these



Where respondents had a single approach for detecting outliers, graphical detection with expert judgement was the most common method (n = 5) “Graphically using expert judgment, creating common graphs such as scatter plots, histograms, box plots in R with ggplot2 package”, IEO(a)). To ensure potential outliers were in fact outliers and not extreme values, expert judgement was considered essential (“Identification of outliers can be done visually on the available plots and tables... Expert judgement is important in the outliers identification process because in some cases an outlier is connected with natural reasons, e.g. diseases, parasites, poor condition.”, NMFRI).

Two respondents detected outliers graphically, and while expert judgement may have played a role, this was not stated in the answers. Two respondents used R scripts to detect outliers, though no additional information on the script itself was offered (“scripts mostly”, THN). Finally, one respondent did not state how they conducted their outlier check, just that it was conducted (“internal calculations to Toughbook”, SLUB(b)).

Table 10: Primary and secondary methods used to check for outliers by respondents who employed a combined approach to question 3.4.3

Institute	Primary	Secondary
BIOR	Graphically	Excel
KU	Graphically	R scripts
WMR(a)	Expert judgement	R scripts
FEAS-MI	Graphically	R script
IEO(a)	Graphically with expert judgement	R scripts
ILVO	Graphically with expert judgement	R scripts

Q3.4.4 At what points are the checks performed? (e.g. at data capture, during data extraction, ad- hoc).

Figure 9: Frequency of responses to question 3.4.4

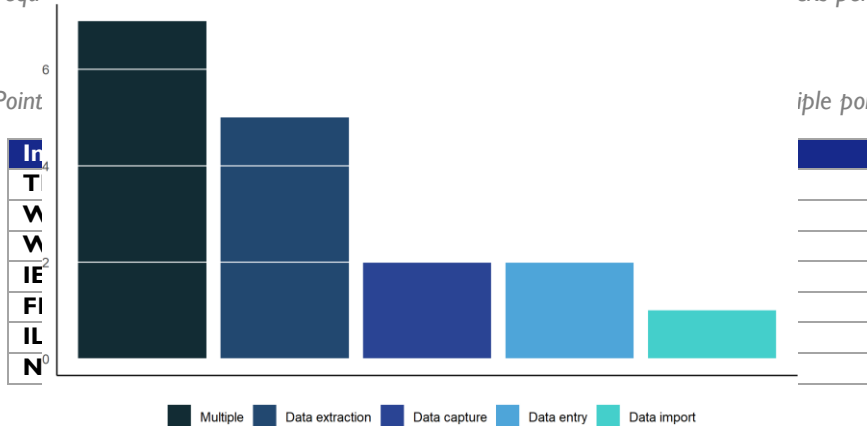


Table 11: Point of check performed

Institute	Point of check
BIOR	Data entry
DRP-RAA	Data extraction
DTU(a)	Data extraction
DTU(b)	Data extraction
IEO(a)	Data extraction
ILVO	Data extraction
WMR(a)	Data extraction
FEAS-MI	Data extraction
KU	Data extraction
NMFRI	Data extraction
SLUB(b)	Data extraction
THN	Data extraction

Table 12: Summarised responses of all respondents to question 3.4.

Institute	Outlier definition	Outlier detection	Point of check
AZTI	Graphically	Graphically with expert judgement	Data extraction
BIOR	Graphically	Combined approach	Data entry
DRP-RAA	Graphically	Graphically	Data extraction
DTU(a)	Graphically	Expert judgement	Data extraction
DTU(b)	Graphically	Expert judgement	Data extraction

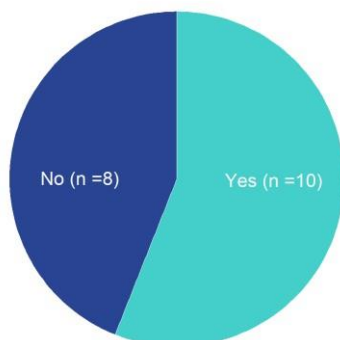


Institute	Outlier definition	Outlier detection	Point of check
EMI	Expert judgement	Graphically	Data capture
FEAS -MI	Comparison with historical data	Combined approach	Multiple
IEO(a)	Numerically	Combined approach	Multiple
IEO(b)	Deviation from common trend	Graphically with expert judgement	Data import
ILVO	Comparison with historical data	Combined approach	Multiple
LUKE	Deviation from common trend	Graphically with expert judgement	NA
NMFRI	Deviation from common trend	Graphically with expert judgement	Multiple
SLU(a)	Graphically	R scripts	Data extraction
SLU(b)	Numerically	Unspecified	Data capture
THN	Comparison with historical data	R scripts	Multiple
KU	Numerically	Combined approach	Data entry
WMR(a)	Graphically	Combined approach	Multiple
WMR(b)	Graphically	Graphically with expert judgement	Multiple

Q 3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

Do you perform any cross checks of sample data with census data?

Q3.5 – do you perform cross checks with census data?



10 respondents stated that they did perform cross checks of sample data with census data, 10 respondents stated that they did not.

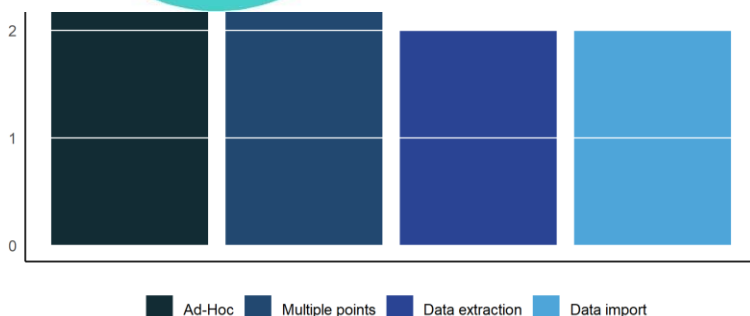


Figure 11: Frequency of categorised responses to Q3.5 – At what point do you perform these checks?

When asked at what point during the data collection process, they performed a cross check between census and sample data, three respondents stated that checks were only performed on an Ad-Hoc basis (“Not as a

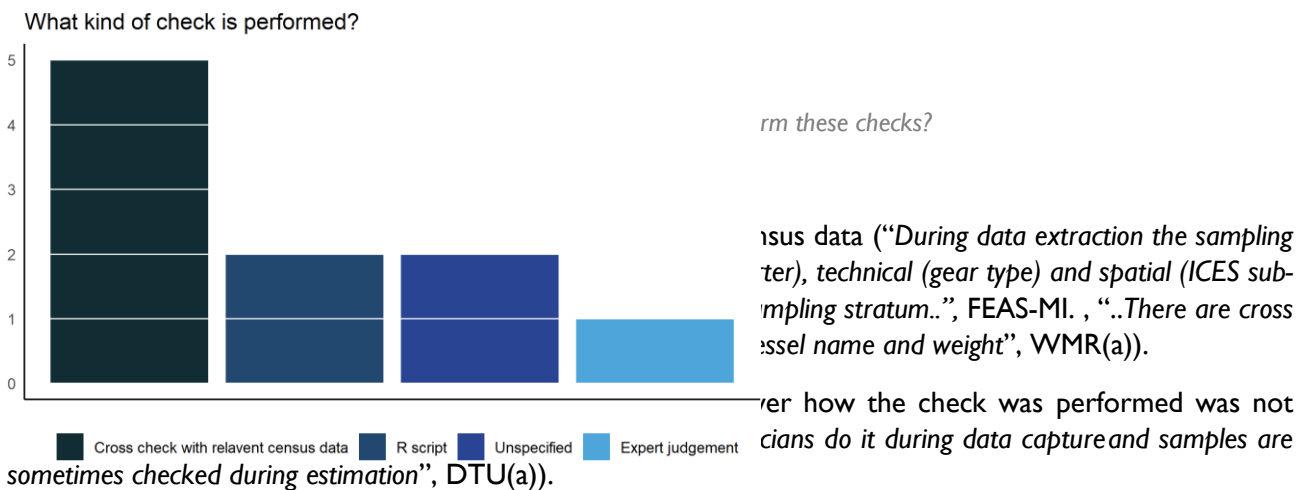
routine. On a more ad-hoc basis, some technicians do it during data capture and samples are sometimes checked during estimation”, DTU(a)).

Three respondents performed the checks at two or more points in the data collection process, (“..During data capture and extraction, at-market and at-sea sampling are cross-checked with salesnotes and logbooks...”, DRF-RAA).

Two respondents performed the checks at the point of data extraction, usually prior to answering data calls (“It is done during the data quality process before answering data calls.”, AZTI).

Two respondents performed the check at data import, when data was being imported into the primary database (“Data on fishing effort and landings for the sampled trip are imported into IMPORT workbook after all these data are recorded into national fisheries data information system

... Simple R script extracts relevant data based on logbook number and landing data.”, KU)



Two respondents incorporated the checks into an R script, which automatically cross-checked census and sample data (“The pairing/crosschecking process between the sampled trips and the official data consists in crossing both sources through an R script in order to assign to each sampled trip the corresponding fishing trip of the NVDP (metierized database of official data)”, IEO(a)).

A single respondent (THN), cross checked census and sample data by way of expert judgement, however no further information on the process was offered (“Not on regular basis and only based on expert judgement”, THN).

Q3.6 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).



Ten respondents did conduct some form of missing value checks during the data collection process. Six respondents did not conduct missing value checks, while missing value checks were not relevant to the data in question for two respondents (WMR(a), EMI).

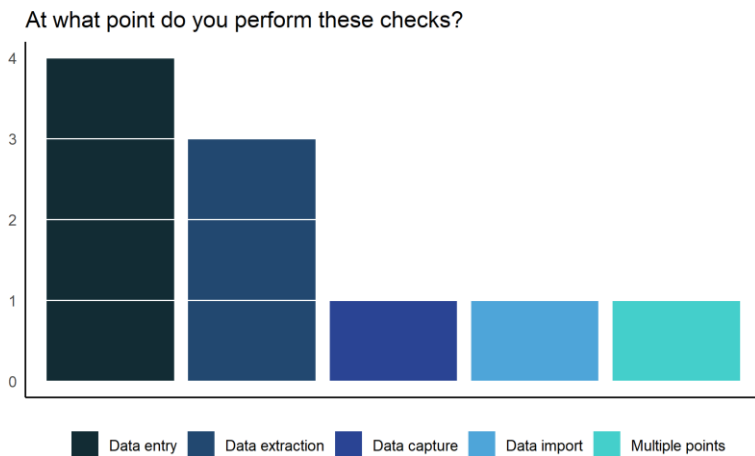
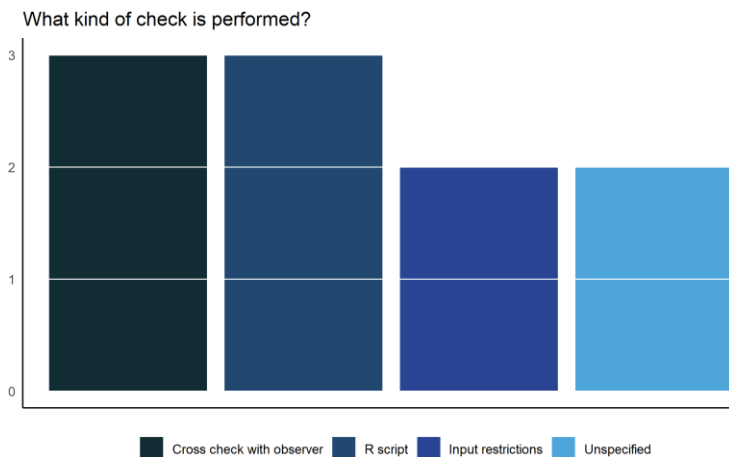


Figure 14: Frequency of categorised responses to Q3.6 – At what point do you perform missing value checks?

Of the ten respondents who did perform the checks, four performed the checks at the point of data entry. Three respondents conducted the check at the data extraction phases prior to answering data calls. One respondent performed the check at the point of data capture, one at the point of data import, and one performed at the checks at multiple points during the data collection process.



What kind of check is performed?

...ts cross checked data with original observer ... error or a true zero (“In cases of mismatch, ... cated that both discards and landings have been

... check for missing values in fish length weights ... in R using the command “table(Dataset\$weight, ... the R script created to detect some missing values:

missing individual weight, missing sex.”, KU).

For two respondents, their data entry software employed restrictions which ensured all required fields were filled (“The data entry software ensures that all mandatory information is registered. For biological parameters, the shiny application designed for data quality control, allows to list all records where age information has not yet been registered.”, NMFRI, “Our data recording system (SIRENO) doesn’t allow the introduction of missing values/zeros for length variable.”, IEO(b), preventing missing values being input with the data.

Two respondents did not specify how they conducted the check, with their answers, only noting if and when the check was performed.

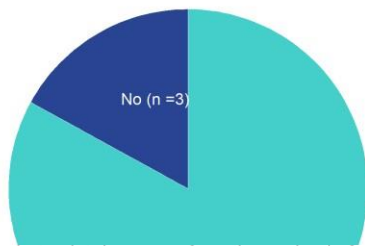


Table 13: Summary of categorised responses to for all respondents to Q3.6

Institute	Checks	Point	Method
AZTI	No	NA	NA
BIOR	Yes	Data entry	Cross check with observer
DRP-RAA	No	NA	NA
DTU(a)	Yes	Data extraction	Cross check with observer
DTU(b)	No	NA	NA
EMI	Not relevant	NA	NA
FEAS -MI	No	NA	NA
IEO(a)	Yes	Data import	Input restrictions
IEO(b)	Yes	Data extraction	Unspecified
ILVO	Yes	Multiple points	R script
LUKE	Yes	Data extraction	Unspecified
NMFRI	Yes	Data entry	Unspecified
SLU(a)	Yes	Data entry	R script
SLU(b)	Yes	Data capture	Cross check with observer
THN	No	Data capture	Unspecified
KU	Yes	Data entry	R script
WMR(a)	Not relevant	NA	NA
WMR(b)	No	NA	NA

Q 3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Do you perform any spatial data checks?



Responses to Q3.7– do you perform spatial data checks?

Of the 18 respondents who performed any spatial data checks, 15 respondents answered that they did. Three did not perform such check, accepting spatial information as is (“No spatial checks yet. Logbook information.”, KU).

At what point do you perform these checks?

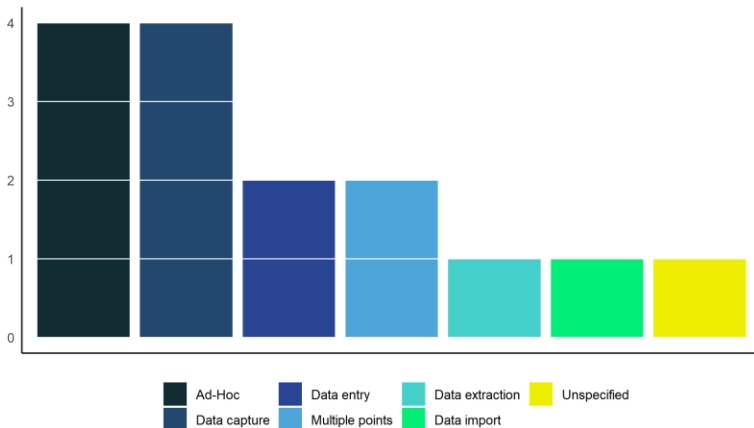
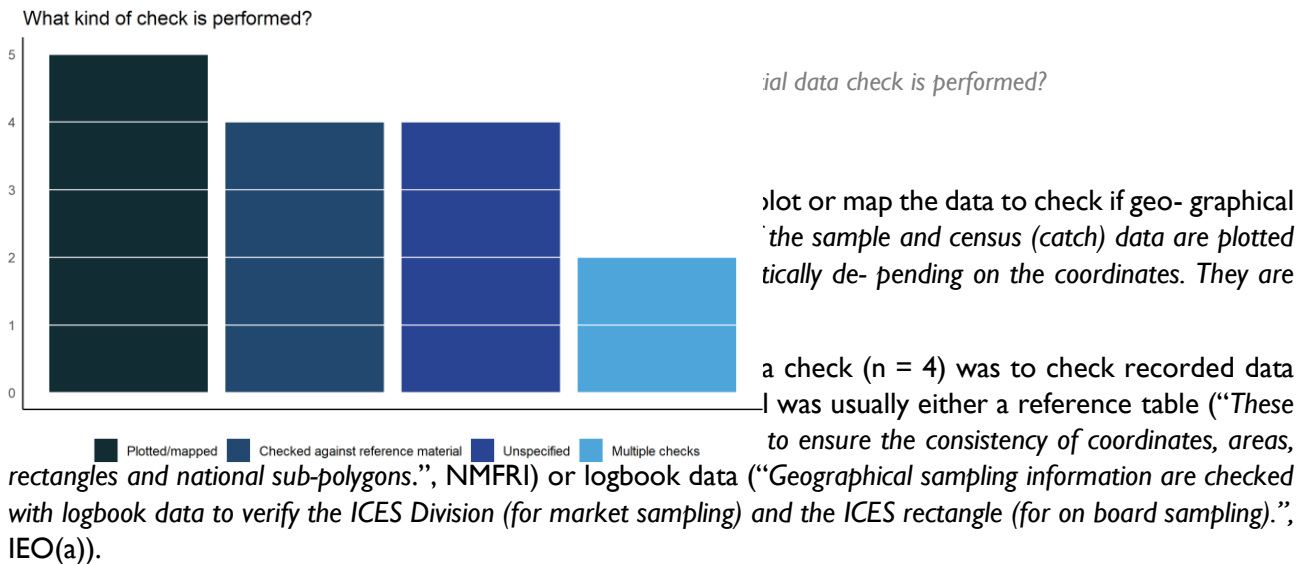


Figure 17: Frequency of categorised responses to Q3.7 – At what point do you performed spatial data checks?



Four Respondents performed the spatial data checks on an Ad-Hoc basis. Four respondents conducted the check at the point of data capture. Two respondents performed the check at data entry, and two performed the check at multiple points during the data collection process. One respondent performed the check during data extraction, one at the point of data import and one did not specify when they performed this check.



Four respondents did not specify how they conducted the check, instead only stating if and when the check was performed during the data collection process (“Not many. Some during the estimation.”, SLU(a)).

Two respondents employed a combined approach (FEAS-MI, DRP-RAA) creating both plots of the data and either checking against reference material (“...These are corrected either visually by plotting positions on a map (Fig. 10) or by reference to original data sheets.”, FEAS-MI) or checking species presence absence in that area (“At the time of data extraction, the spatial distribution is visualized, and wrong coordinates are corrected (which usually occurs due to data entry errors - transposition error). Ad-hoc crossing of areas with the presence/absence of species is also carried out, but not systematically.”, DRP-RAA).

Categorised answers of all respondents can be seen in Table 14.

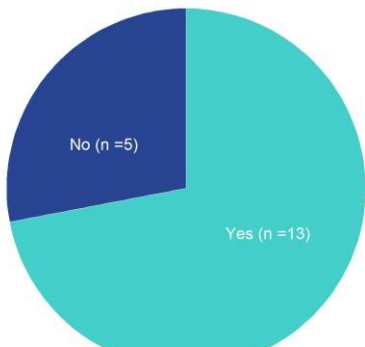
Table 14: Categorised responses for all respondents to question 3.7

Institute	Checks	Point	Method
AZTI	Yes	Data capture	Plotted/mapped
BIOR	Yes	Ad-Hoc	Unspecified
DRP-RAA	Yes	Data entry	Multiple checks (Plotted and mapped, Species Presence/Absence)
DTU(a)	Yes	Ad-Hoc	Plotted/mapped
DTU(b)	Yes	Ad-Hoc	Unspecified
EMI	Yes	Data capture	Checked against reference material
FEAS -MI	Yes	Unspecified	Multiple checks (Plotted and mapped, Checked against reference material)
IEO(a)	Yes	Data import	Checked against reference material
IEO(b)	No	NA	NA
ILVO	Yes	Data extraction	Plotted/mapped
LUKE	No	NA	NA
NMFRI	Yes	Data entry	Checked against reference material
SLU(a)	Yes	Ad-Hoc	Unspecified
SLU(b)	Yes	Data capture	Checked against reference material
THN	Yes	Data capture	Unspecified

KU	No	NA	NA
WMR(a)	Yes	Multiple points	Plotted/mapped
WMR(b)	Yes	Multiple points	Plotted/mapped

Q 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

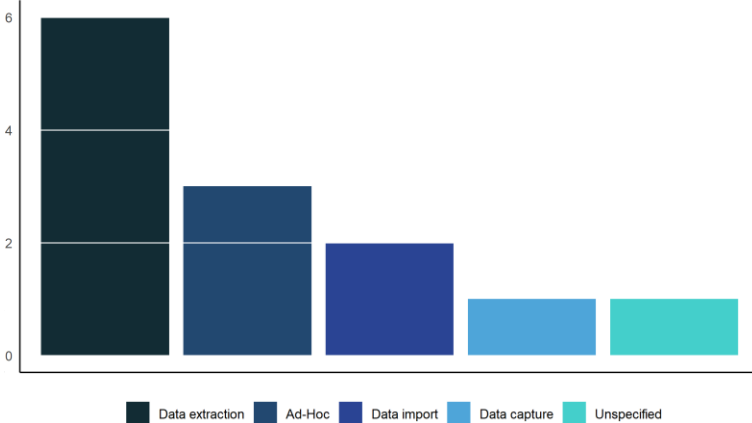
Do you perform any temporal data checks?



Responses to Q3.8 – Do you perform any temporal data checks?

When asked if they performed any temporal data checks, 13 respondents stated that they did perform this and 5 respondents stated that they did not perform any temporal data checks.

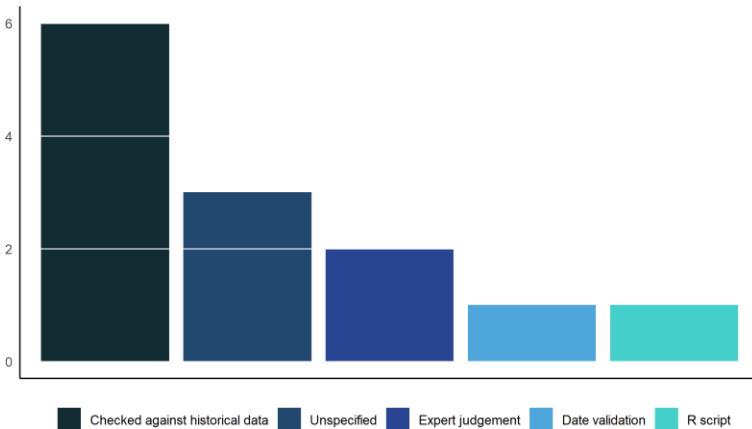
At what point do you perform these checks?



When do you perform temporal data checks?

When asked when they performed temporal data checks, 6 respondents carried out the checks at the point of data extraction, 3 respondents carried out the checks at the point of ad-hoc, 2 respondents carried out the checks at the point of data import, 1 respondent carried out the checks at the point of data capture, and one respondent did not specify.

What kind of check is performed?



What kind of temporal data check is performed?

When asked what kind of check was performed, 6 respondents checked the data against historical data, 3 respondents did not specify, 2 respondents relied on expert judgement, 1 respondent used date validation, and 1 respondent used an R script.

When asked when they performed the check, just if and when the check was performed (“Yes. During the estimation.”, SLU(b)).

Two respondents relied on expert judgement to cross check temporal data (“Expert judgement used to quality check certain parameters is therefore built over the years.”, ILVO).

One respondent validated trip dates by cross checking sample data with known trip information (“The check consists in ensuring that the sample date is within or close to the trip dates, depending on the type of fishery.”, NMFRI).

One respondent conducted the check through use of an R script which generated summary statistics for a variety of parameters and checked them against values from previous years and quarters (“Simple R script for description of summary data statistics by species, year, quarter and metier...”, KU).

Categorised answers of all respondents can be seen in table 15.

Table 15: Categorised responses for all respondents to Q3.8

Institute	Checks	Point	Method
AZTI	No	NA	NA
BIOR	Yes	Data extraction	Checked against historical data
DRP-RAA	Yes	Ad-Hoc	Unspecified
DTU(a)	Yes	Unspecified	Checked against historical data
DTU(b)	No	NA	NA
EMI	No	NA	NA
FEAS -MI	Yes	Data extraction	Checked against historical data
IEO(a)	Yes	Data import	Expert judgement
IEO(b)	Yes	Ad-Hoc	Unspecified
ILVO	Yes	Data import	Expert judgement
LUKE	No	NA	NA
NMFRI	Yes	Data extraction	Date validation
SLU(a)	Yes	Data extraction	Unspecified
SLU(b)	Yes	Data capture	Checked against historical data
THN	No	NA	NA
KU	Yes	Ad-Hoc	R script
WMR(a)	Yes	Data extraction	Checked against historical data
WMR(b)	Yes	Data extraction	Checked against historical data

Q 3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

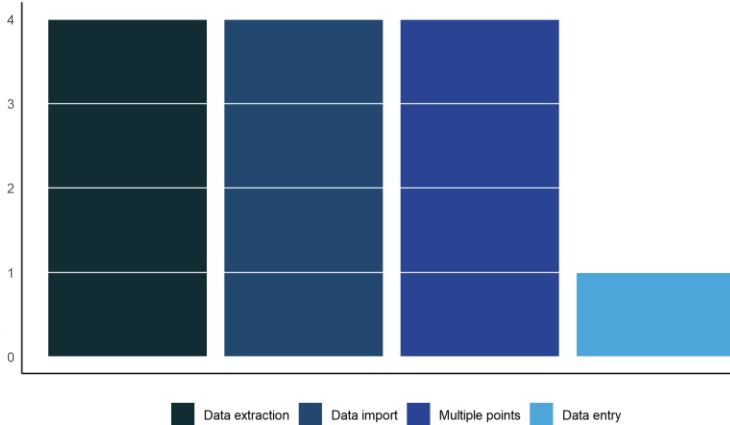
Do you perform any duplication checks?



Figure 22: Frequency of categorised responses to Q3.9 – Do you perform any duplication checks?

When asked whether they conducted any duplication checks during the data collection process, 13 respondents stated that they did. Five respondents stated that they did not perform any duplication checks.

At what point do you perform these checks?



Frequency of categorised responses to Q3.9 – At what point do you perform these checks?

Four respondents carried out the duplication checks at multiple points during data extraction (“Yes, duplications are checked for the database (for things that cannot be checked when delivering data to ICES.”, SLUB(b), “During data import and extraction the number of rows in the original data set is checked against the number of rows of the same data set when the distinct values are filtered out.”, WMR(a)). One respondent carried out the duplication check at the point of entering the data into the primary database (“The database constraints prevent from entering duplicates in some data entry steps”, NMFRI).

What kind of check is performed?

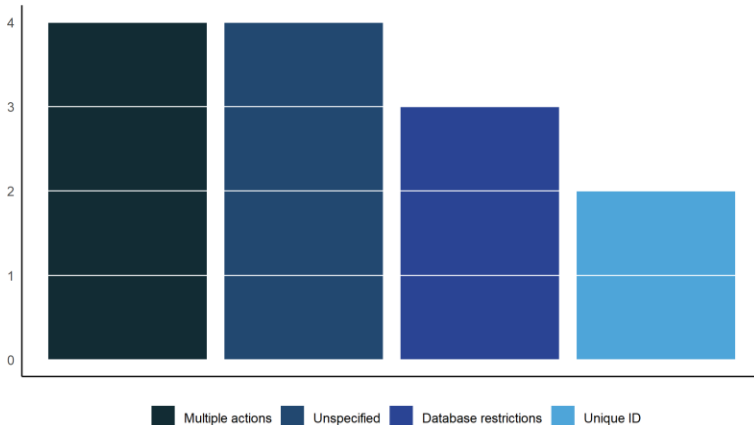


Figure 24: Frequency of categorised responses to Q3.8 – What kind of duplication check is performed?

When asked how they performed the duplication checks, four respondents did not specify how they performed the check, just if and when the check was performed (“We have some duplication checks for sampling data. We do it during the data quality process before answering data calls”, AZTI).

Three respondents had constraints or restrictions on their database which prevented the entry of duplicate records (“SIRENO database or icrOS system doesn’t allow the introduction of duplicates data.”, IEO(b), “Yes, the Smartfish application does not allow users to create duplicated samples during the data capture process. Similar process is valid when working with the age reading tool Smartdots”, ILVO).

Two respondents utilised unique IDs for each sample, where the same ID cannot be used twice. Unique sample IDs were generated either through primary and foreign keys (“All tables in the national database related with primary and foreign keys, which reveal the duplications”, THN) or through unique combinations of haul, biological and date information collected (“Yes, duplications are checked for at several occasions, when importing data from the field, ad hoc in the database... Things that are compared are eg. but not only: • The combination any vessel and fromdate time must be unique. • The combination fish number and catch id must be unique.”, SLU(b)).

Four used a combined approach from preventing duplicate entries. Three of these used unique ID’s and parallel tables (“During data import and extraction the number of rows in the original data set is checked against the number of rows of the same data set when the distinct values are filtered out. Furthermore, each sample is assigned to a unique sample ID. A unique sample ID can’t be entered in the database twice”, WR(b)), and one used database restrictions and parallel tables (“The database constraints prevent from entering duplicates in some data entry steps. Checksums are available at the level of entering biological data. Moreover, a relation with a parallel system for PSU selection, enables to identify potential duplicates.”, NMFRI). Details of combined approaches can be found in table 16.

Categorised answers of all respondents can be seen in table 16.

Table 16: Categorised responses for all respondents to question 3.9.

Institute	Checks	Point	Method
AZTI	Yes	Data extraction	Unspecified
BIOR	No	NA	NA
DRP-RAA	Yes	Data extraction	Database restrictions
DTU(a)	No	NA	NA
DTU(b)	No	NA	NA
EMI	No	NA	NA
FEAS -MI	Yes	Data import	Multiple points (Unique ID, Parallel table)
IEO(a)	Yes	Data import	Unspecified
IEO(b)	Yes	Data import	Database restrictions
ILVO	Yes	Multiple points	Database restrictions
LUKE	No	NA	NA
NMFRI	Yes	Data entry	Multiple points (Database restrictions, Parallel table)
SLU(a)	Yes	Data extraction	Unspecified
SLU(b)	Yes	Multiple points	Unique ID
THN	Yes	Data extraction	Unique ID
KU	Yes	Data import	Unspecified
WMR(a)	Yes	Multiple points	Multiple points (Unique ID, Parallel table)
WMR(b)	Yes	Multiple points	Multiple points (Unique ID, Parallel table)

3.10 Please let us know about any other relevant data checks which have not already been described in your answers

When asked about any other relevant data checks they performed, ten respondents stated did not have any

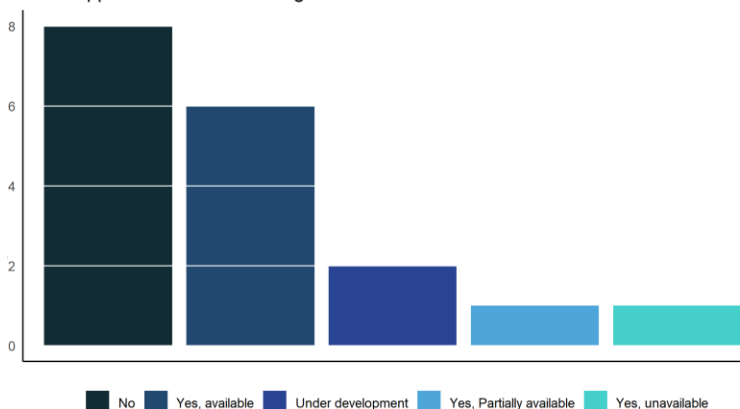
other relevant checks or they left the question blank. For the eight respondents who answered the question, categorisation was not appropriate so table 17 below shows their full responses in addition to links to data where possible. Associated images for answers can be found in the relevant appendices.

Table 17: Full responses for respondents who gave details of any additional data checks they performed in the data collection process.

Institute	q3.10	Links
AZTI	We check census data for errors in species identification, for these species which are clearly wrong because they cannot be present in our waters. We check metier & area combination.	
BIOR	<p>As I mentioned above, I am working in the sea alone. Biological data with the otoliths are collected and returned in special paper books. For each individual fish such information is collected, length, full weight, sex, maturity and otoliths.</p> <p>Otoliths are wrapped in page similar to an envelope. At this example is cod with length 47 cm, weight 1,03 kg, female with maturity stage 5.</p> <p>After data input in Excel file, the age reader receives paper books with otoliths and file with the entered data. During the otolith preparation for age reading additional data quality check is performed, if necessary, corrections are made.</p>	
DTU(a)	Ad-a) Different relevant checks are done as a routine on the at-sea observer trips per trip and quarter, see attached pdf's	
DTU(b)	Ad-a) Different relevant checks are done as a routine on the at-sea observer trips per trip and quarter, see attached pdf's	
EMI	Since our data is uploaded to ICES RDB, the RDB data checking system performs many checks.	
FEAS -MI	<p>F:\Logbooks_Current_report – for some checks on the logbook data that is used to raise the sample data to the population level Length/Frequency plots are generated during data entry. This plot updates automatically within Nemesys as commercial data is electronically captured at sea.</p> <p>Figure Yes. An example of one of the sections in the Nephrops Measuring System (Nemesys) Data Validation Reports and similar length frequency/plots have been added into our commercial port sampling data entry application (Stockman)</p> <p>QC Weights added into Nemesys -described above Voice Report Validation tool for validating entered commercial discards data. Data is entered through paper sheets into our Commercial Discards Database, and the entered is validated through a Voice Reporting Application.</p>	
IEO(a)	http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework	http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework

Q3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

Do you have written processes or guidelines which define your approach to data checking?



written process or guidelines relevant to your

guidelines for their data quality control checks, their data checking. Six respondents did use and in table 18. Two respondents do not have had such guidelines but due to GDPR sensitive on request. Finally, one respondent had such the intellectual property of their institute.

Institute	q3.11	Link
AZTI	Yes, unavailable	NA
BIOR	No	NA
DRP-RAA	Under development	NA
DTU(a)	No	NA
DTU(b)	No	NA
EMI	Yes, available	https://www.envir.ee/sites/default/files/andmetootluse_juhend.pdf
FEAS - MI	Yes, partially available	Censored version available upon request.
IEO(a)	Yes, available	http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework
IEO(b)	No	NA
ILVO	Yes, available	Available upon request
LUKE	No	NA
NMFRI	Yes, available	tinyurl.com/dpadesdd
SLU(a)	No	NA
SLU(b)	No	NA
THN	Under development	NA
KU	No	NA
WMR(a)	Yes, available	Image provided - see appendix
WMR(b)	Yes, available	Image provided - see appendix

Section 4 – Data editing

Section 4 asked respondents about data editing and to outline their procedure for dealing with any errors, inconsistencies or discrepancies found in their data.



Q4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them?(e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

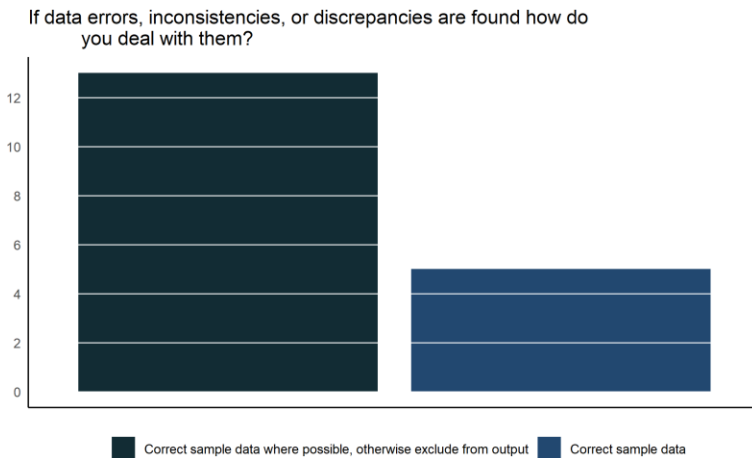


Figure 26: Frequency of categorised responses to Q4.1 – How do you deal with any errors, inconsistencies or discrepancies found in your data?

When asked how they dealt with errors, inconsistencies or discrepancies found in the data, there appears to be a broad consensus among respondents, with all respondents answering that they attempted to correct the error in the sample if possible.

13 respondents stated when errors, inconsistencies or discrepancies are found, they attempted to correct the data where possible, and if data could not be corrected it was excluded from outputs “If a data point is identified as an outlier, first it is examined if it’s a wrong entry and if not, it is transmitted to the laboratory technicians to check if the value is an actual observation or a mistake. If the technician points it out as a mistake the data is removed from the database and consequently excluded from any output.”, WRM(b). Five respondents stated that they corrected the sample data where possible, however they did not state how they dealt with data that could not be corrected (“Sample data will be corrected when possible before data supply”, THN). Overall, data correction was generally carried out by referring to the original data collection sheets (“... must be reviewed by the supervisors, usually implying review of the original sampling sheets.”, IEO(a)).

Categorised answers of all respondents can be seen in table 19.

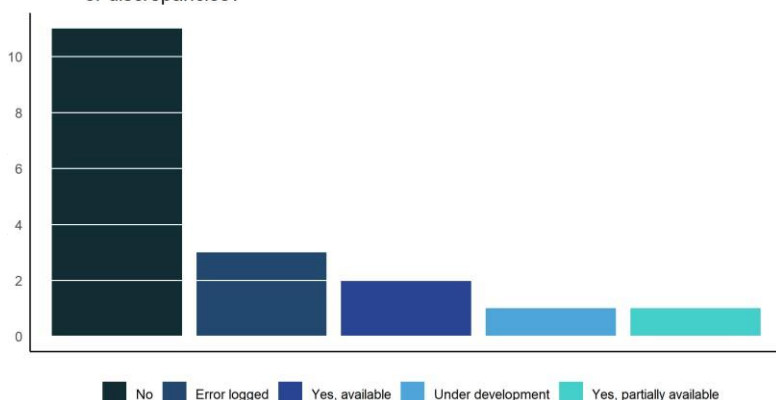
Table 19: Categorised responses for all respondents to Q4.1

Institute	q4.1
AZTI	Correct sample data where possible, otherwise exclude from output
BIOR	Correct sample data where possible, otherwise exclude from output
DRP-RAA	Correct sample data
DTU(a)	Correct sample data where possible, otherwise exclude from output
DTU(b)	Correct sample data where possible, otherwise exclude from output
EMI	Correct sample data where possible, otherwise exclude from output
FEAS -MI	Correct sample data where possible, otherwise exclude from output

IEO(a)	Correct sample data
IEO(b)	Correct sample data where possible, otherwise exclude from output
ILVO	Correct sample data where possible, otherwise exclude from output
LUKE	Correct sample data where possible, otherwise exclude from output
NMFRI	Correct sample data
SLU(a)	Correct sample data where possible, otherwise exclude from output
SLU(b)	Correct sample data where possible, otherwise exclude from output
THN	Correct sample data
KU	Correct sample data
WMR(a)	Correct sample data where possible, otherwise exclude from output
WMR(b)	Correct sample data where possible, otherwise exclude from output

Q 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies?



delines for dealing with any errors, inconsistencies

idelines for dealing with such errors, I
s. Three respondents outlined the process
event similar errors in the future (“The data
l documents during the data checking process
owing mandatory fields need to be field in the

Yes/No), Who”, WMR(a)). Two respondents were able to provide the guidelines or documentation which defined their approach to dealing with such errors. One respondent stated that they are currently developing such guidelines, and one respondent was able to provide only some of their guidelines, as others contained sensitive information unavailable for publication.

Table 20: Categorised responses for all respondents to Q4.1 – Do you have any guidelines for dealing with errors, inconsistencies, and discrepancies in your data? Where respondents provided a link, the link has also been given in the table.

Institute	q4.2	link
AZTI	No	NA
BIOR	No	NA
DRP-RAA	No	NA
DTU(a)	No	NA
DTU(b)	No	NA
EMI	Yes, available	https://www.envir.ee/sites/default/files/andmetootluse_juhend.pdf
FEAS -MI	Yes, partially available	https://wwz.ifremer.fr/cost/
IEO(a)	No	NA
IEO(b)	No	NA
ILVO	Yes, available	See data extraction protocol for ICES combined data call
LUKE	No	NA

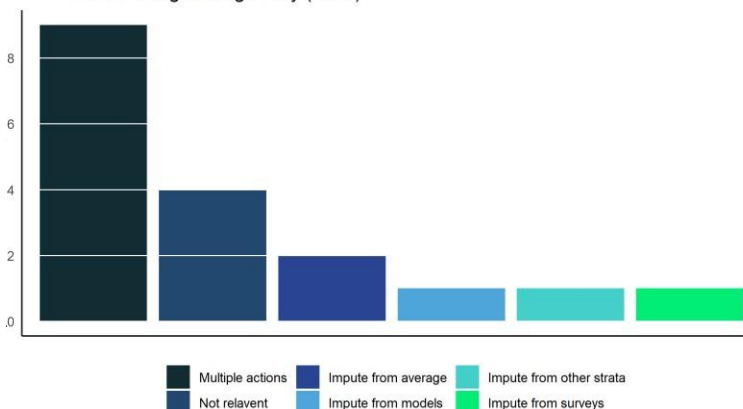
Institute	q4.2	link
NMFRI	No	NA
SLU(a)	No	NA
SLU(b)	No	NA
THN	Under development	NA
KU	Error logged	NA
WMR(a)	Error logged	NA
WMR(b)	Error logged	NA

Section 5 – Data imputation

Section 5 asked respondents about their approach to dealing with any gaps in their data. Specifically, respondents were asked about gaps in age length keys (ALK's) , weight length keys (WLK's) and sampling strata.

Q5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)?



with any gaps in your ALK's or WLK's?

Weight length keys, two respondents imputed from averages”, LUKE, “In cases of gaps in ALK or

binomial logistic model (“To deal with gaps in quarterly sampled, age-length keys (ALK) are modelled using a binomial logistic model (Gerritsen et al., 2006)”, ILVO).

One respondent imputed values from other strata where available (“For age data the ALK are merged across technical strata but there still might be gaps. To make things efficient, an assumption that the differences in the ALK between areas are minor enough to be ignored, so age data from all areas are combined into one but the quarterly stratification is kept.”, FEAS-MI).

One respondent dealt with gaps by imputing a value from fisheries independent surveys (“Impute missing values from surveys, if possible.”, SLU(b)).

Most respondents (n = 9) employed a combination of the above actions. For example, some imputed values from survey data, before filling any further gaps based on expert judgement (“age length key (ALK) of the commercial sampling is completed with the age-length survey data and the missing values are completed by an age expert judgement.”, IEO(b)). Others attempted to impute values from averages, followed by surveys followed by models (“Missing values are imputed first from averages, then from surveys, then from models.”, WMR(a)).

For three respondents, ALK's and WLK's were not relevant to their data.

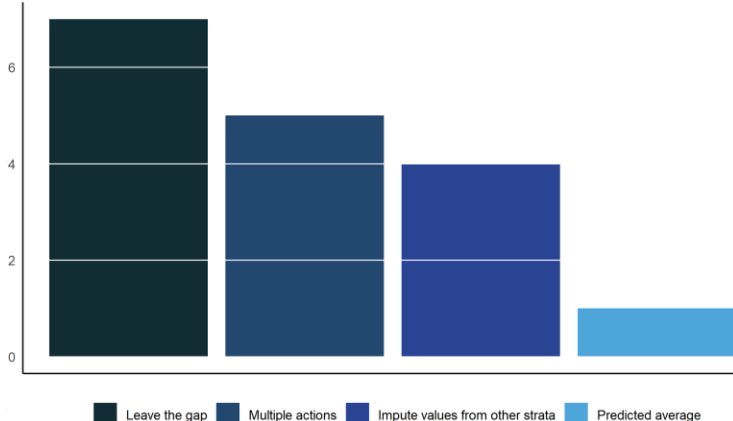
Table 21: Categorised responses to Q5.1 – dealing with gaps in ALK's and WLK's. Where respondents employed a combined approach, all their responses are listed.



Institute	q5.1_1	Action 1	Action 2	Action 3
AZTI	Multiple actions	Impute from average	Impute from other strata	NA
BIOR	Multiple actions	Impute from average	Fill by expert judgement	NA
DRP-RAA	Not relevant	NA	NA	NA
DTU(a)	Multiple actions	Impute from average	Impute from models	NA
DTU(b)	Multiple actions	Impute from average	Impute from models	NA
EMI	Not relevant	NA	NA	NA
FEAS - MI	Impute from other strata	NA	NA	NA
IEO(a)	Not relevant	NA	NA	NA
IEO(b)	Multiple actions	Impute from other strata	Fill by expert judgement	Leave the gaps
ILVO	Impute from models	NA	NA	NA
LUKE	Impute from average	NA	NA	NA
NMFRI	Impute from average	NA	NA	NA
SLU(a)	Not relevant	NA	NA	NA
SLU(b)	Impute from surveys	NA	NA	NA
THN	Multiple actions	Impute from other strata	Impute from surveys	NA
KU	Multiple actions	Impute from average	Impute from models	Impute from surveys
WMR(a)	Multiple actions	NA	NA	NA
WMR(b)	Multiple actions	Impute from average	Impute from surveys	Impute from models

Q5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

How do you deal with any gaps in your sampling strata?



How do you deal with any gaps in your sampling strata?

7 respondents opted to leave the gaps in the data to deal with them (“Since the implementation of the stock coordinator after the integration of all data at the national level but at Stock Data Coordination level. This is recorded in national database.”, NMFRI).

1 respondent in the sampling strata. These included leaving the gaps in the major stratum that has insufficient samples then the sample data can either be deleted for that stratum or it can be submitted with a warning. It is preferable to let the ICES stock coordinator deal with gaps. For species that are reported by length and for which there is no biological sampling (i.e. weights-at-length) the length-weight parameters will need to be supplied to estimate the sample weights... an Age-Length Key then becomes a Length-Length key, which is a convoluted way of raising the data has the functionality of merging strata etc.”, FEAS-MI), and imputing from survey data followed by filling gaps based on expert judgment (“For small pelagic stocks, age length key (ALK) of the commercial sampling is completed with the age-length survey data and the missing values are completed by an age expert judgement.”, IEO(b)).



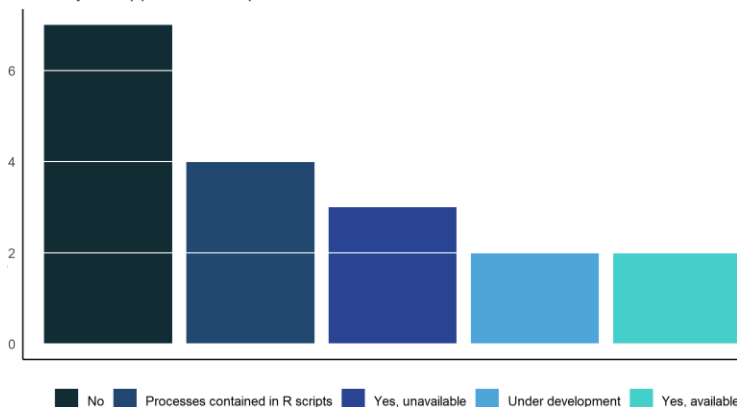
Four respondents imputed values from other strata to fill gaps (“Strata, commercial size categories, do not match the ones in InterCatch, so missing values are imputed from other strata.”, DTUB(b), “Usually, impute missing values from other strata.”, DRP-RAA).

Table 22: Categorised responses to Q5.2– How do you deal with any gaps in your sampling strata?

Institute	q5.2_1	q5.2_2	q5.2_3
AZTI	Multiple actions	Leave the gap	Impute values from other strata
BIOR	Leave the gap	NA	NA
DRP-RAA	Impute values from other strata	NA	NA
DTU(a)	Leave the gap	NA	NA
DTU(b)	Impute values from other strata	NA	NA
EMI	NA	NA	NA
FEAS - MI	Multiple actions	Leave the gap	Impute values from other strata
IEO(a)	Impute values from other strata	NA	NA
IEO(b)	Multiple actions	Impute values from survey data	Expert judgement
ILVO	Leave the gap	NA	NA
LUKE	Multiple actions	Leave the gap	Impute values from other strata
NMFRI	Leave the gap	NA	NA
SLU(a)	Impute values from other strata	NA	NA
SLU(b)	Multiple actions	Impute values from other strata	Leave the gap
THN	Leave the gap	NA	NA
KU	Predicted average	NA	NA
WMR(a)	Leave the gap	NA	NA
WMR(b)	Leave the gap	NA	NA

Q5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g.structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it

Do you have written processes or guidelines which define your approach to imputation?



written guidelines for dealing with any gaps in your

s or guidelines which defined their approach to guidelines. Four respondents included written There are two R markdown documents for data usually. Training is also given to data submitters on s documented in scripts. Its most important steps at.”, SLU(a)).

Table 23: Categorised responses for all respondents to Q5.3– Do you have any written guidelines for dealing with any gaps in your sampling strata?

Institute	q5.3
AZTI	Under development
BIOR	No
DRP-RAA	No
DTU(a)	Processes contained in R scripts
DTU(b)	Processes contained in R scripts
EMI	No
FEAS -MI	Processes contained in R scripts
IEO(a)	Yes, unavailable
IEO(b)	No
ILVO	Yes, available
LUKE	No
NMFRI	No
SLU(a)	Processes contained in R scripts
SLU(b)	No
THN	Under development
KU	Yes, available
WMR(a)	Yes, unavailable
WMR(b)	Yes, unavailable

Conclusion

Data checks

The primary objective of this questionnaire was to determine if, when and how European fisheries institutes performed data quality control checks, data editing and data imputation. The analysis presented above indicates that most respondents: constrained some values to be physically realistic (Q3.2), used predefined code lists (Q3.3), performed some form of outlier check (Q3.4), performed some form of spatial data check (Q3.7), performed some form of temporal consistency check (Q3.8), performed some form of duplication check (Q3.9). Checks were performed regularly as part of the data collection process were cross checks with census data (Q3.5) and missing values check (Q3.6). However, whilst most checks were performed, the point at which checks were performed varied greatly. The reason for performing check at different points in the process could be attributed to different data capture methods, different time frames for the importing data or different operating procedures in relation to data collection and checking. At a minimum, institutes should aim to ensure all checks have been performed prior to responding to data calls (at or prior to the point of data extraction). If checks are implemented at a different or additional stage (where checks are being implemented at multiple points), the point, method and type of checks implemented should be documented.

The method for some checks, such as outlier detection and cross checking of spatial data, are similar for many respondents. As many respondents already have a dedicated R script which produces plots which aid in the identification of outliers, it may be possible to produce a standardised R script dedicated to outlier checking and or spatial data plotting, which would be available to all members of the RCG (in turn standardising some/multiple checks discussed above). While variety in sampling schemes and data collection practices might limit the effectiveness of such a script, a standardised script containing protocols might prove useful in ensuring checks are in place and are of a common method.

Data editing

The consensus for approaches to dealing with errors, inconsistencies and discrepancies was to attempt to correct the sample data where possible, and to exclude the data from outputs where correction is not possible. If data cannot be corrected, institutes should at least aim to document the error prior to deletion. Such a record may help in preventing similar mistakes in future and highlight repeated errors so corrective action(s) can be taken. Such error logging is already in place by WMR(a,b) and KU. The template for logging errors proposed by WMR may be suitable for logging such errors (“*SampleID, Species, DateChecked, ErrorDescription, ActionsTaken* (e.g. *excluded, corrected*), *Reason, DateProcessed, Re-imported (Yes/No), Who*”, WMR(a,b)). If possible, institutes should also log errors even where correction was possible, again to prevent any future errors.

Data Imputation

For dealing with their approach to gaps in Age length keys (ALK's) or weight length keys (WLK's), institutes filled such gaps either by imputing from an average, imputing from a model, imputing from other strata, filling by expert judgement, or leaving the gap. As the course of action often depended on what data from other surveys, strata or sampling schemes was available, a definitive course of action to be taken in the event of an ALK/WLK gap is not appropriate. However, where gaps have been filled, institutes should document which data was imputed and what method was used. If a predicted value from a model was used, details of the model should be recorded. If data is borrowed from other strata or from surveys, the details of the strata or survey should be recorded.

When asked about dealing with gaps in sampling strata, most respondents opted to leave the gap and allow the ICES stock coordinator to decide how to deal with the issue. As this is already a popular course of action, leaving the gaps in the sampling strata and allowing the ICES stock coordinator to deal with them should be the course of action employed by institutes to deal with gap in their sampling strata. Where institutes decide to impute from other strata or surveys, details of what values have been imputed and of the method of imputation should be documented, such that the ICES stock coordinator is aware data has been imputed. This should minimise the chances of already imputed data being imputed from, increasing data accuracy overall.

Written guidelines

Where asked to list any written guidelines relevant to sections three, four and five, many institutes were not able to provide such guidelines, either because they did not have any or they were not publicly available. As institutes still performed many of these checks without such guidelines, they may be unnecessary, however having SOP's for data quality control recorded in a document would be a useful resource, both at a regional and international level. While such guidelines may contain information sensitive under GDPR, a censored or constrained document could still be appropriate.

Age - readings

While there was some reference to data quality control in relation to otolith readings (FEAS-MI, ILVO), most respondents did not discuss these practices in their answers. As a result, this report cannot recommend 'best practice' quality control with regards to otolith readings, as it is not supported by the data presented here.

Recommendations

Based on the analysis conducted in this report, the following recommendations are proposed for data quality control practices.



1. When data quality control checks (such as those discussed in section 3) are implemented, institutes should ensure that the type of check, timing of the check (both the point during the data collection process and the date), and a brief description of the check are documented.
2. Where checks are performed at multiple points during the data collection process, institutes should ensure that datasets / samples are marked such that users are aware what checks have been already performed or where data has been edited or imputed.
3. Where the method of check is broadly similar among institutes (e.g. Q3.4 - outlier detection, Q3.8 - spatial data checking etc), attempts should be made to produce a standardised SOP, ideally at a WG level, detailing the method used to perform the checks.
4. Where the method of check is broadly similar among institutes (e.g. Q3.4 - outlier detection, Q3.8 - spatial data checking etc), attempts should be made to produce an R script to conduct these checks which is available to all users.
5. Where errors, inconsistencies or discrepancies are found in the data, information about the cause of the error and course of action taken to rectify it should be recorded. Records will allow users to identify common sources of error in data collection process.
6. Where institutes are imputing data from a predicted average/model/survey or from other strata to fill gaps in ALK's or WLK's, institutes should clearly document *what* data has been imputed, *where* the data was imputed from and *when* the data was imputed. As imputation may be performed at multiple points or by different users, it is essential that all users, from local to working group level, are aware what data is 'real' data and what data has been predicted or imputed.
7. Where gaps are found in sampling strata, a standardised course of action should be decided on at WG level. Based on the analysis conducted in this report, the most suitable course of action is to leave the gaps and allow the ICES stock coordinator to decide on how best to deal with them.
8. Further research should be conducted to collect information on data checks, editing and imputation with regards to age-reading among institutes.



References

West, M., 2011. *Developing high quality data models*. Burlington, MA: Morgan Kaufmann, pp.4 - 6.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.



Annex I. Data QC Questionnaire Report – An analysis of the data quality control practices of European fisheries institutes for data checks, editing and imputation



FISHN'CO Data QC Questionnaire Report

An analysis of the data quality control practices of European fisheries institutes for data checks, editing and imputation.

Michael Kinneen



Author:

This report was written by Michael Kinneen (OSMS) on behalf of the Marine Institute, Ireland.



Regional Coordination Group
Baltic



Regional Coordination Group
on Economic Issues



Regional Coordination Group
Large Pelagics



Regional Coordination Group
North Atlantic
North Sea & Eastern Arctic



Contents

Introduction	4
Objectives	5
Methodology	6
Glossary of terms.....	7
Response Rate.....	8
Questionnaire results.....	10
Q2.1 Which country do you work in?	11
Q2.2 Which institute or laboratory do you work in?	12
Q2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions?)	12
Q2.4 Which data have you thought about when answering these questions?	13
Section 3 - Data checks	
Q 3.1 When is the data entered into an electronic recording system such as a database?	17
Q 3.2 Do you constrain the values of properties in your data recording system to be physically realistic?.....	19
3.3 Do you use defined code lists for storing categorical information electronically?.....	22
Q3.4 Do you perform any outlier checks on your data? If yes, please explain:	24
Q 3.5 Do you perform any cross checks of sample data with census data?	28
Q 3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas).....	32
Q 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years).....	35
Q 3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice).	38
3.10 Please let us know about any other relevant data checks which have not already been described in your answers.....	41
Q3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.	43
Section 4 - Data editing	
Q4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them?	46
Q 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies?	48
Section 5 - Data imputation	
Q5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)?	51
Q5.2 How do you deal with any gaps in your sampling strata?	53
Q5.3 Do you have written processes or guidelines which define your approach to imputation?	55
Conclusion	57
Recommendations	59
References.....	60
Appendices	61
AZTI – Fundacio AZTI (Spain)	61
BIOR - Institute of Food Safety, Animal Health and Environment, Fish resources research department, Marine laboratory (Latvia)	64
DRP/RAA - Regional Directorate for Fisheries in the Azores.....	71





DTU(a,b) – Denmark technical University Aqua	74
EMI - Estonian Marine Institute, University of Tartu	79
FEAS-MI – Fisheries Ecosystem Advisory Services, Marine Institute (Ireland)	82
IEO(a) - Instituto Español de Oceanografía (Spain)	95
IEO(B) - INSTITUTO ESPAÑOL DE OCEANOGRAFÍA ,Centro Nacional (Spain)	101
ILVO - Marine research (Flanders research institute for agriculture, fisheries, and food.)(Belgium)	112
IPMA – Instituto Instituto Português do Mar e da Atmosfera (Portugal)	120
LUKE - Natural resources institute Finland	124
NMFRI - National Marine Fisheries Research Institute (Poland)	128
SLU(a) - Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences	132
SLU(b) - Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences	136
THN - Thünen Institute of Sea Fisheries (Germany)	140
KU - Marine Research Institute of Klaipeda University (Lithuania)	143
WMR(a,b) - Wageningen Marine Research (Netherlands)	149





Introduction

The aim of this survey was to collect information on the data checking, editing and imputation practices of 18 different fisheries institutes across the EU. Under the 'Biological Data Quality' thematic working area of the FishNCo project, the collection of this data will aid in the strengthening of EU fisheries data collection by developing Regional Work Plans for the EU Regional Coordination Groups (RCG). In addition to the collation and analysis presented in this report, the data collected in these questionnaires will also aid in the production of a data quality process template. This template will allow members of RCG's to record, efficiently and concisely, any data checking, editing or imputation process they implement in the future.

The questionnaire itself is composed of 5 sections. Sections 1 (not published) and 2 (Respondent information) collected information about the respondents, their respective roles their institutes. Section 3 (Data checks) collects information on if, when and how data checks are performed during the data collection process. Section 4 (Data editing) collects on any how inconsistencies, errors or discrepancies are dealt with during the data collection process. Section 5 (Imputation) collects information on how gaps in Age length Keys (ALK's), Weight length Keys (WLK's) and sampling strata are addressed during the data collection process.





Objectives

The objectives of this report are as follows.

1. Collect, collate, and categorise data on data checks, editing and imputation performed by EU fisheries Institutes during the collection of fisheries data.
2. Summarise and analyse the collected data to determine if, when and how such checks are performed by EU fisheries Institutes.
3. Present the collected data and analysis in report which clearly and concisely communicates the observed results.
4. Use the summary and analysis conducted to create a data quality control checks, editing and imputation template to be used in the collection of fisheries data by EU fisheries Institutes.





Methodology

A questionnaire composed of 5 sections was composed. Respondents were asked to respond using free text answers and to include diagrams, images, and written guidelines where relevant and possible. These questionnaires were distributed on the 25/05/2021. After responses were received, all responses were collated in a spreadsheet, with each columns representing a question and each row the response from a specific institute. A response cut-off date of 22/05/2021 was set and responses received following this date are not included in the analysis.

A duplicate matrix was then created, and inductive categorisation was used to categorise responses. For questions 3.2, 3.4,3.5,3.6,3.7,3.8 and 3.9, answers were broken down into three sections 1) Whether the check was performed, 2) At what point in the data capture process was the check performed and 3) How the check was performed. For all other questions (Section 2, Q3.2,3.10,3.11,Section 4 and Section 5), respondents answers focused on describing the check, editing or imputation process address in the question.

The results were then plotted using R version 4.04 (R Core Team, 2021) and the 'ggplot2' package (Wickham, 2016). For each question, a plot showing the frequencies of each categorised answer, a prose analysis of the responses with supporting quotes, and a table showing how each respondent was categorised was presented.

The findings of the analysis and recommendations were then summarised in the conclusion. These findings were then incorporated into a data quality control template, which will allow users to record the time, type and method of data quality control checks they implemented in future. Each check was categorised based on data properties presented by West (2011).

Caveat: While every effort has been made to ensure as much detail of respondent's answers was captured, categorisation of textual data necessitates some reduction in data resolution. Full, uncategorised responses are available in the relevant appendices, and users are encouraged to refer to these for greater detail and clarity where required.





Glossary of terms

Data collection method

EDC : Electronic data capture, usually by means of an electronic measuring board (in the case of fish) or callipers (in the case of *Nephrops*).

Point of data collection

Ad-Hoc : Checks are only performed as necessary during the data collection process, but not on a regular basis or at a defined point in the process.

Data capture : The recording of data, either manually on paper or by means of electronic data capture.

Data entry : The inputting of paper transcribed data to a temporary digital workbook such as an excel sheet/Microsoft access database.

Data import : The transfer of data from temporary digital workbooks/databases to the primary database. After being imported to the primary database the data should be ready for extraction.

Data extraction : The withdrawal of data from the primary workbook, usually in response to data calls for RCG's, WG's etc.

R scripts: Can refer to R markdown documents (.Rmd) or Simple R files (.R)





Response Rate

Of the 18 institutes asked to complete the survey, 15 responded within the timeframe, one responded late and two did not respond. The names, acronyms, and response status of all those contacted are detailed in table 1. Of the three institutes who did not complete the questionnaire, two were unable to be contacted, while one replied when contacted but stated they would not be able to complete the questionnaire in the allotted timeframe. As a result, the response of IPMA (Portugal) has been added as an appendix to the report, but their answers have not been included in the analysis.

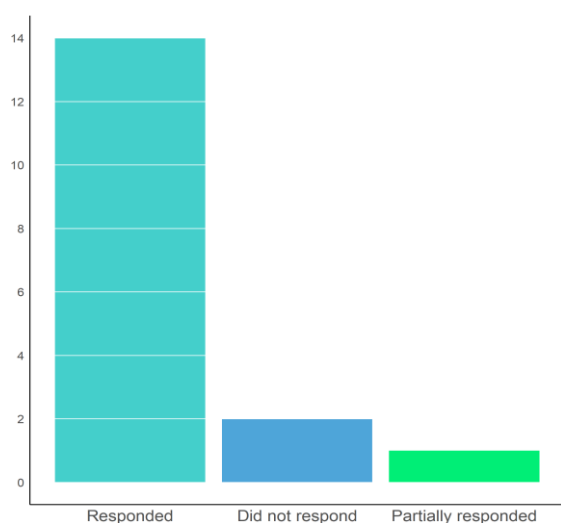


Figure 1: Frequency of response status of all respondents surveyed for this report.

Table 1: Response status for all institutes surveyed for this report.

Institute Name	Acronym	Response status
Azores Regional Directorate of Fisheries	DRP-RAA	Responded
Eigen Vermogen van het Instituut voor Landbouw- en visserijonderzoek	ILVO	Responded
Fisheries Ecosystem, Advisory Services - Marine Institute	FEAS-MI	Responded
Fundación AZTI	AZTI	Responded
Institute of Food Safety, Animal Health and Environment	BIOR	Responded
Instituto Español de Oceanografía	IEO	Responded
Luonnonvarakeskus	LUKE	Responded
National Marine Fisheries Research Institute	NMFRI	Responded
Stichting Wageningen Research	WMR	Responded
Swedish University of Agriculture and Sciences	SLU	Responded
Technical University of Denmark	DTU	Responded
Thuenen Institute	THN	Responded
Klaipeda University	KU	Responded
University of Tartu Estonian Marine Institute	EMI	Responded





Institut de Recherche pour le Développement	IRD	Did not respond
Institut Français de Recherche pour l'Exploitation de la Mer	IFREMER	Did not respond
Instituto Português do Mar e da Atmosfera	IPMA	Responded (not analysed)





Questionnaire results

Section 2 - Institute information

The questions in section 2 were aimed at gathering basic information about the respondents. Respondents were asked 1) What countries they worked in, 2) What lab or institute they worked in, 3) Whether their lab or institute had any relevant accreditations or certifications and 4) What data they thought about when completing sections 3, 4 and 5.

As the questionnaire covered a range of topics in the data collection process, most responses required the input of personnel in various roles e.g. Data manager, Database administrator, Sampling co-ordinators, Onboard observers. Where institutes stated clearly which answers had been offered by different personnel, their responses were separated into two different responses, indicated by the Institute abbreviation followed by a or b (e.g. IEO(a)). Institutes whose responses were separated in this way were: Instituto Español de Oceanografía, Stichting Wageningen Research, Swedish University of Agriculture and Sciences and Technical University of Denmark.





Q2.1 Which country do you work in?

A map showing the country of origin (q2.1) and response status of all those contacted can be seen in figure 2. As a response was received from IPMA (Portugal) following the response deadline, the response was included in the appendices, but was not included in the analysis. Hence, Portugal was categorised as ‘Partially responded’.

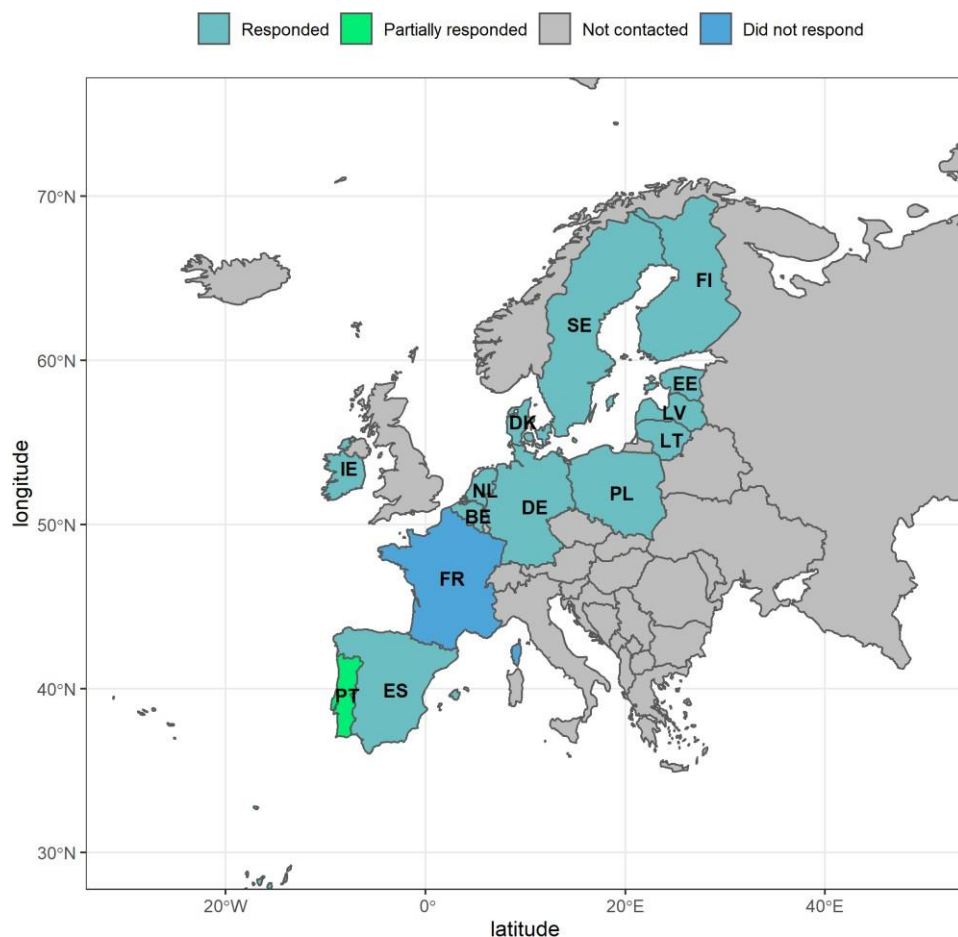


Figure 2: Map showing response status by country (indicated with 2 letter alpha codes) of institutes surveyed for this report.

Table 2: Response status, 2 letter alpha code and country name of all institutes surveyed for this report.

Country		Institute
BE	Belgium	ILVO
DE	Germany	THN
DK	Denmark	DTU(a)
		DTU(b)
EE	Estonia	EMI
ES	Spain	IEO(a)
		IEO(b)
		AZTI
FI	Finland	LUKE
FR	France	IFREMER
		IRD
IE	Ireland	FEAS -MI
LT	Lithuania	KU
LV	Latvia	BIOR
NL	Netherlands	WMR(a)
		WMR(b)
PL	Poland	NMFRI
PT	Portugal	IPMA
	Portugal – Autonomous Region of the Azores (RAA).	DRP-RAA
SE	Sweden	SLU(a)
		SLU(b)



Q2.2 Which institute or laboratory do you work in?

Q2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

The name of the lab or institute in which respondents worked can be found in table 3 in addition to any relevant certifications or accreditations the lab or institute holds. Only five respondents listed relevant certifications, four of which were ISO accreditations. The only other accreditation listed was IODE accreditation.

Table 3 : Full name and relevant accreditations of all respondents.

Institute(Short)	Institute/Lab	Relevant certifications
AZTI	AZTI	No
BIOR	Institute of Food Safety, Animal Health and Environment "BIOR", Fish resources research department, Marine laboratory.	No
DRP-RAA	Regional Directorate for Fisheries in the Azores (DRP/RAA).	No
DTU(a)	DTU Aqua	No
DTU(b)	DTU Aqua	No
EMI	Estonian Marine Institute, University of Tartu	No
FEAS -MI	Marine Institute, Fisheries Advisory & Ecosystems Services	IODE accreditation
IEO(a)	Instituto Español de Oceanografía (IEO).	No
IEO(b)	Centro Nacional INSTITUTO ESPAÑOL DE OCEANOGRAFÍA (IEO, CSIC).	No
ILVO	ILVO Marine research (Flanders research institute for agriculture, fisheries and food.)	ISO 17025
LUKE	Natural resources institute Finland, Luke	No
NMFRI	National Marine Fisheries Research Institute in Gdynia, Poland	No
SLU(a)	Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences	No
SLU(b)	Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences	No
THN	Thünen Institute of Sea Fisheries	NA
KU	Marine Research Institute of Klaipeda University	ISO 14001, ISO 45001, ISO 9001
WMR(a)	Wageningen Marine Research.	ISO 9001
WMR(b)	Wageningen Marine Research.	ISO 9001





Q2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

When asked about which data they considered while completing this survey, respondent answers ranged from general (“Fish stock rather”, KU) to extremely specific (see IEO(b), table 3). Due to broad nature of the answers, categorisation was not used, and full respondents’ full answers can be found in table 4.

Table 4: Sampling schemes or stocks which respondent considered while completing this questionnaire.

Institute	Q2.4 (Data considered when completing the survey)
AZTI	Data from our sampling schemes (at the market and on board) and official data corresponding to ICES areas
BIOR	Data from Baltic Sea demersal trawlers
DRP-RAA	All relevant stocks and sampling schemes are monitored from commercial fisheries in ICES Division 10a2 (Azorean fleet).
DTU(a)	a) Estimated amount of discard for different ICES assessment WG’s
DTU(b)	b) Estimated age distribution of landings of commercial stocks for different ICES assessment WG’s, where the sampling is stratified per commercial size categories
EMI	Stock assessment-related data for Baltic herring (Central Baltic Herring and the Gulf of Riga herring stocks), and the Baltic sprat in Sd. 22-32
FEAS -MI	The questions are answered for the Demersal Catch Sampling At-Sea programme, which follows the flow of data collected during an at-sea sampling programme from collection to analysis to reporting. The Demersal Catch Sampling At-Sea programme is comprised of demersal at-sea and Nephrops at-sea sampling. The Nephrops at-sea sampling has similar but slightly different protocols to the demersal at-sea. Landings data from at-sea sampling is uploaded to the Stockman database.
IEO(a)	Data from our length sampling programme, both market and on-board, in the ICES area under the DCF/EUMAP. Tuna fisheries excluded.





IEO(b)	<p>The biological variables data (Fisheries independent data) on the stocks for the ICES Área are carried out according to 2 differentiated sampling designs, depending on the biological characteristics of each species:</p> <ul style="list-style-type: none"> - Small pelagic species: the sample/subsample is selected by a Simple Random Sampling (SRS). The sample is entirely biologically analyzed (various biological variables are collected on each sampled fish until the expected number of samples is reached). <p>Engraulis encrasicolus (ane.27.8), Micromesistius poutassou (whb.27.1-9No14), Sardina pilchardus (pil.27.8c9a), Scomber scombrus (mac.27.nea), Scomber colias 8, 9, Trachurus trachurus (hom.27.2a4a5b6a7a-ce-k8), Trachurus trachurus (hom.27.9a), Engraulis encrasicolus (ane.27.9a), Sardina pilchardus (9as), Scomber scombrus (9as)</p> <ul style="list-style-type: none"> - Demersal and benthic species: the sample is stratified by length classes. A Simple Random Sampling (SRS) is applied for the selection of the samples in each length stratum. A fixed number of specimens from each length class is biologically sampled and various biological variables are collected on each individual. The sample attempts to represent the full length range of the catch, so the least abundant length classes are preferably selected for sampling. <p>Lepidorhombus boscii (ldb.27.8c9a), Lepidorhombus whiffiagonisboscii (meg.27.7b-k8abd), Lepidorhombus whiffiagonisboscii (meg.27.8c9a), Lophius budegassa (ank.27.78abd), Lophius budegassa (ank.27.8c9a), Lophius piscatorius (mon.27.78abd), Lophius piscatorius (mon.27.8c9a), Conger conger (all areas), Helicolenus dactylopterus (all areas), Merluccius merluccius (hke.27.3a46-8abd), Merluccius merluccius (hke.27.8c9a), Molva molva all areas (lin.27.3a4a6-9No14), Phycis blennoides all areas (gfb.27.nea), Trisopterus spp all areas (T. luscus)</p> <p>The samples of the following species usually come from surveys although could be occasionally sampled from commercial</p>
---------------	---





	landings: Zeus faber all areas, Mullus surmuletus all areas, Loligo vulgaris 8c, 9a, Pagellus bogaraveo (sbr.27.9), Parapenaeus longirostris 9a, Sepia officinalis all areas
ILVO	All biological sample data from commercial sampling at sea trips that are used for analytical stock assessments and hereto linked census data (logbooks and sales notes).
LUKE	Salmon catch samples from coastal fyke-net fishery in ICES SD22-32 in the Baltic Sea, self-sampling by selected fishers and catch samples from commercial HER and SPR fishery
NMFRI	Data collected in all sampling schemes.
SLU(a)	Market sampling of cod landings in the west coast of Sweden
SLU(b)	Pot fishery for Norwegian lobster. (Length, weight, sex, maturity in females and diseases.)
THN	Raised biological commercial data of the German commercial fleet (except Baltic), by-catch data
KU	Fish stock rather
WMR(a)	KMG: All landing data, all landing sampling schemes
WMR(b)	HO: commercial data collected on board.





Section 3 - Data Checks

Section 3 of this questionnaire asked respondents about what data checks they implemented during the data collection process, when they performed these checks, and how they performed these checks. In addition, it asked respondents about their data collection methods and about any relevant guidelines or written processes they had with regards to data checks.





Q 3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

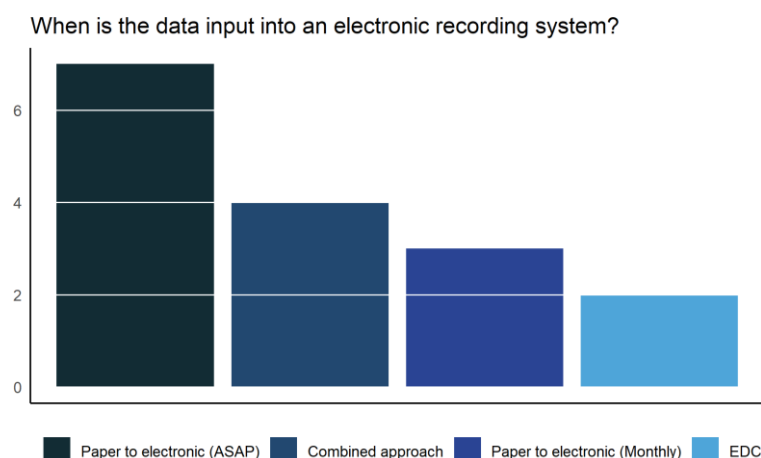


Figure 1: Frequency of categorised responses to question 3.1

The majority of respondents (n = 7) recorded data on paper prior to inputting it into an electronic recording system as soon as possible, usually following the sampling activity or survey (*“captured on paper and then transcribed as soon as possible after each sampling activity”*, THN). Four respondents employed a combined approach, where both electronic data capture (EDC) and recording on paper before inputting the data electronically as soon as possible. Where a combined approach was used, EDC was often employed when sampling *Nephrops norvegicus* and paper transcription for other samples (*“The only electronically device used in our commercial sampling is a calliper used for measuring the carapace length (mm) of Nephrops and shrimps. Everything else is captured on paper and entered in our national database as soon as possible”*, DTU(a)). Paper transcription with monthly digitisation of data was employed by IEO(a,b) and DRP-RAA. Finally, some institutes (ILVO, SLU) use EDC exclusively for data collection (*“seagoing observers register sample data at sea directly in the database using a custom developed Smartfish application. The application is run on a rugged tablet coupled to an electronic measuring board.....”*, ILVO).

Categorised answers can be found in table 1, while full answers for each country can be found in the relevant appendix. Where countries employed a combined approach, the primary method and secondary method are listed into table 1.



Table 1: Categorized answers of all respondents to Q3.1

Institute	Method	Primary	Secondary
AZTI	Paper to electronic (Monthly)	NA	NA
DRP-RAA	Combined approach	Paper to electronic (Monthly)	EDC
EMI	Paper to electronic (Monthly)	NA	NA
FEAS -MI	Combined approach	EDC	Paper to electronic (ASAP)
IEO(a)	Paper to electronic (Monthly)	NA	NA
IEO(b)	Combined approach	Paper to electronic (Monthly)	EDC
IFSAHE	Paper to electronic (ASAP)	NA	NA
ILVO	EDC	NA	NA
LUKE	Paper to electronic (ASAP)	NA	NA
NMFRI	Paper to electronic (ASAP)	NA	NA
SLU(a)	Paper to electronic (ASAP)	NA	NA
SLU(b)	EDC	NA	NA
THN	Paper to electronic (ASAP)	NA	NA
KU	Paper to electronic (ASAP)	NA	NA
WMR(a)	Combined approach	EDC	Paper to electronic (Annually)
WRM(b)	Paper to electronic (ASAP)	NA	NA





Q 3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Do you constrain the values of properties in your data recording system to be physically realistic?

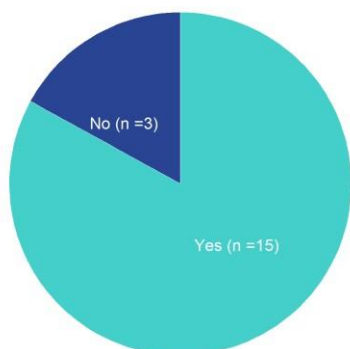


Figure 1: Frequency of responses to Q3.2 – were checks performed?

When asked whether values of properties were constrained in their data recording system, the majority of respondents (n = 15) answered yes. Only three respondents (AZTI, IEO (b), LUKE) answered no. However, while values were not constrained, two of those who answered no (IEO(b), AZTI) did check the data prior to data extraction (“No for most of the stocks, however data are checked just after data extraction.”, IEO (b)).

At what point do you perform these checks?

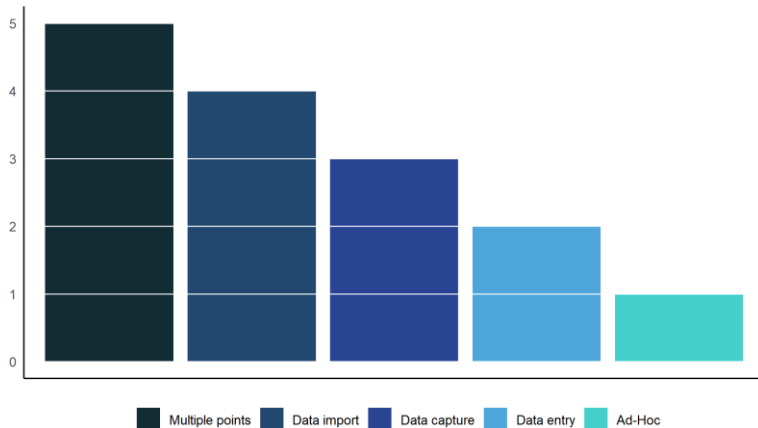


Figure 2: Frequency of categorised responses to Q3.2 – When did they perform the check?

When asked when these checks were performed, the majority of those who answered yes (n = 5) carried out checks at multiple points during the data collection process, usually both at the point of data capture and during data extraction (“Checking is in place e.g. by automatic outlier search, plotting boxplots or histograms, comparison with length-weight relationships etc., during data input and data extraction”, THN.). Four respondents implemented the check prior to importing the data into the primary database (“Data is checked against common out of range errors at the step of entering into the database.”, NMFRI). Three performed the check at the point of data capture (“There is a constrain for extreme values on age, length and weight by species in the data recording





system (during data capture).”, WMR(a), and two (DTU(a), DTU(b)). at the point of data entry. Only one institute constrained values to be physically realistic on an Ad-Hoc basis (EMI).

Table 1: Categorised responses of all institute to Q3.2 – if they perform the check and when they perform the check.

Institute	Check performed	Point of check
AZTI	No	NA
BIOR	Yes	Multiple points
DRP-RAA	Yes	Multiple points
DTU(a)	Yes	Data entry
DTU(b)	Yes	Data entry
EMI	Yes	Ad-Hoc
FEAS-MI	Yes	Multiple points
IEO(a)	Yes	Data import
IEO(b)	No	NA
ILVO	Yes	Data capture
LUKE	No	NA
NMFRI	Yes	Data import
SLU(a)	Yes	Data import
SLU(b)	Yes	Multiple points
THN	Yes	Multiple points
KU	Yes	Data import
WMR(a)	Yes	Data capture
WMR(b)	Yes	Data capture

When asked to describe the type of constraints they had in place, 12 respondents constrained values to be within a reasonable range. This could apply to fish length (“measurements must between 3.01mm to 99.99mm”, FEAS-MI.), weights (“.. individual weight between 1 – 50000 grams, etc”, KU), or non-biological variables (“Some of the numeric fields in our national database has constrains, so only realistic values can be entered e.g. wind direction”, DTU (a)). Three respondents constrained their data entry such that the user could only choose from pre-defined lists, limiting the entry of incorrect or unrealistic values (“The data file contains predefined values that can be assigned to the following biological parameters: sex and maturity. At the top of the datasheet 10 rectangles are located. For each rectangle excel macro is assigned. We are using a 6-scale maturity scale. Sex is defined as numbers, 1 is male and 2 is female. In the rectangles all combinations of sex and maturity are predefined..”, BIOR). Three respondents had physically realistic constraints in place with regards to catch and sample weights, usually checking that sample weight was not greater than catch weight (“Sample weights are checked by comparing the length frequency of the sample and sample weight cannot be larger than the total weight”, SLU(b)). Finally, three respondents had input restrictions on their database, where users were prevented or warned by the data entry software when erroneous or missing values were present (“A general species-specific length-weight key check is applied for every weight registration (sample and individual weight). A notification is displayed for an abnormal weight. The user can reject the notification or choose to change the initially registered weight..”, ILVO. , “Our Commercial Port Sampling Application (Stockman) contains data validation ensuring required fields have been entered i.e. Sampling Place, Landing Port Sampler...”, FEAS-MI.)

A summary of the constraints in place by respondents can be seen in table 2. Full details can be found in the relevant appendices.





Table 2: Method of constraints used by all respondents in Q3.2

Institute	Reasonable range	Pre-defined lists	Catch and Sample weights	Input restrictions
BIOR	X	X		
DRP-RAA	X		X	
DTU(a)	X			
DTU(b)	X			
EMI	X			
FEAS-MI	X		X	X
IEO(a)	X			
ILVO	X	X		
NMFRI	X			
SLU(a)	X		X	X
SLU(b)	X		X	X
THN				
KU	X	X		
WMR(a)				
WMR(b)	X			





Q 3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

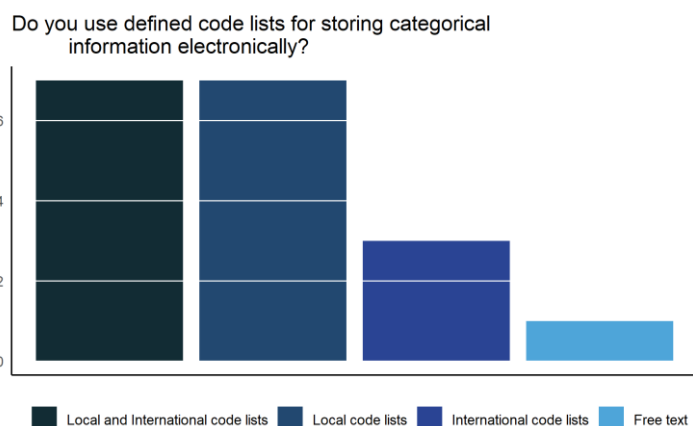


Figure 1: Frequency of summarised responses to question Q3.3

Almost all respondents (n = 15), used some form of code list to store categorical data during the data collection process. Two respondents (BIOR, ILVO) used exclusively international code lists. International lists were usually a combination of ICES and FAO codes (*“International 3-letter code (FAO code) list for fish species, international code lists such as ICES vocabularies.”*, BIOR).

Seven respondents employed local code lists. Little additional information on these lists was recorded in the questionnaire, with respondents usually only stating that they used local code lists (*“Yes, local code lists.”*, WR(b)).

Six respondents used a combination of local and international code lists (*“local/working and ICES codes”*, KU., *“Nearly all of the codes lists are local, but the most relevant ones, species, area etc., have a field with International codes”*, DTU(a)). International lists were again drawn from either ICES or FAO codes, however two respondents (FEAS-MI, AZTI) also use codes from the world register of marine species (WoRMS)

Only one respondent did not use code lists as the primary means of recording categorical data, although code lists were used for some information (*“This depends on categorical information, e.g. areas, gear and metier are defined as in ICES vocabularies. Otherwise mostly free text.”*, EMI).

Categorised responses by Institute can be seen in table 1.



Table 1: summarised responses of all respondents to Q3.3

Institute	Q3.3
AZTI	Local and International code lists
BIOR	International code lists
DRP-RAA	Local code lists
DTU	International code lists
DTU	Local and International code lists
EMI	Free text
FEAS -MI	Local and International code lists
IEO	Local code lists
IEO	Local code lists
ILVO	International code lists
LUKE	Local and International code lists
NMFRI	Local code lists
SLU	Local code lists
SLU	Local code lists
THN	Local and International code lists
KU	Local and International code lists
WR	Local and International code lists
WR	Local code lists





Q3.4 Do you perform any outlier checks on your data? If yes, please explain:

3.4.1 Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

All respondents (n = 18) stated that they did perform outlier checks on their data. In terms of properties checks, all respondents checked biological properties for outliers, including length-weights (“Yes. Analysis and detection of outliers for biological parameters, their weight–length relationships and ranges.”, IEO(b)), length-age (“biological parameters i.e. length-weight, length-age.”, LUKE) and maturity (“Number of individuals length, Age range, Length range, Sex ratio ,Maturity stage” , WMR(b)).

Other properties commonly checked for outliers included discard weights per haul (“Discards weights per haul and species compared to an estimated weight based on the length distribution of the sample (Routine)”, DTU(b)) and catch and sample weight (“Unexpected sample weights; High raising factors; Missing raising factors; Negative discards (discard weight larger than total catch weight); Sample weight larger than total discards”, FEAS-MI.). Some respondents also checked census data (“We do check length distributions, landings, etc...”, IEO(a)), discard rates, spatial data (“Positions have been visualised on a map, haul duration has been checked using Microsoft Power Bi,”, ILVO) and Haul or trip information (“Excessive tow length or fishing speed; Zero tow length; Impossible or unexpected shoot or haul positions; Short tow duration; Negative tow duration...”, FEAS-MI). Most respondents checked a combination of these properties, as can be seen table 1.

Table 1: Properties checked for outliers by respondents.

Institute	Biological parameters	Discard weights per haul	Catch and sample weights	Census data	Discard rates	Spatial data	Haul data
AZTI	X		X				
BIOR	X						
DRP-RAA	X						
DTU(a)	X	X	X				
DTU(b)	X	X	X				
EMI	X						
FEAS -MI	X	X	X	X	X	X	X
IEO(a)	X						
IEO(b)	X						
ILVO	X		X			X	X
LUKE	X						
NMFRI	X		X				
SLU(a)	X	X	X				
SLU(b)	X						
THN	X	X			X		
KU	X						
WMR(a)	X			X		X	
WMR(b)	X			X		X	





3.4.2 How do you define an outlier?

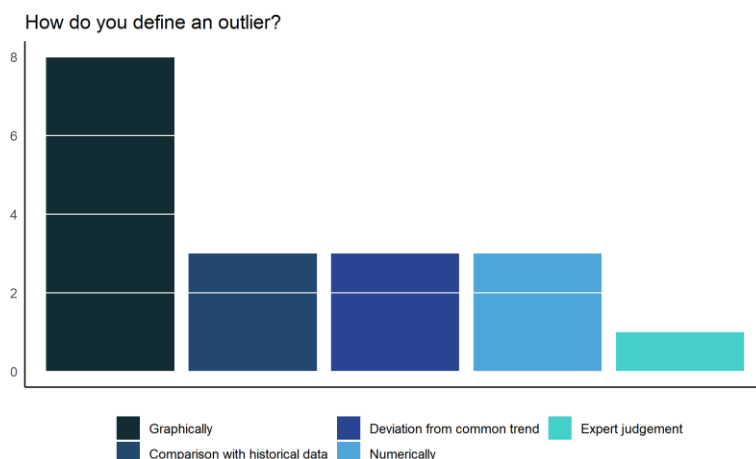


Figure 1: Frequency of summarised responses to Q3.4.2 methods used to define outliers.

When asked how they defined an outlier, eight respondents did so graphically. Of these eight, some specified the type of plot used (*“atypical values in several types of relationships and boxplots between biological variables (length, weight, age,...); unusual biological variables collected”*, SLU(a).), while others did not (*“Visual, extreme percentage. Never found a good approximation with standard deviation”*, DTU(a)). Boxplots, histograms, and length-weight scatterplots were among the common types of graphs used.

Three respondents defined outliers through comparison with historical data (*“Outliers are defined by comparison to historical data. Points that fall outside 95% of historical data points are considered to be outliers.”*, FEAS-MI, *“comparison of discards rates over the years”*, THN).

Three defined outliers as observations deviating from a common trend, however they did not specify if this trend was observed visually, numerically or through expert judgement (*“Value far apart from other values or values that are frequently the result of an error”*, IEO(a), *“An observation is considered an outlier when it deviates significantly from a common trend of observations in the same group.”*, NMFRI).

Three respondents defined outliers numerically, using either Fultons coefficient (*“After entering the weight that does not match the settings (‘Fulton’s coefficient is >2 or less than 0.5), cell is coloured in red and additional data checking is performed.”*, BIOR), Cookes distance (*“For length and landings we use Cook distance to detect outliers”*, IEO(b)) or residuals following modelling (*“Exp of residual is less than 0.5 or more than 2”*, KU).

One respondent defined an outlier based on expert judgement, however no further information was offered (*“According to expert experience.”*, EMI)



3.4.3 How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

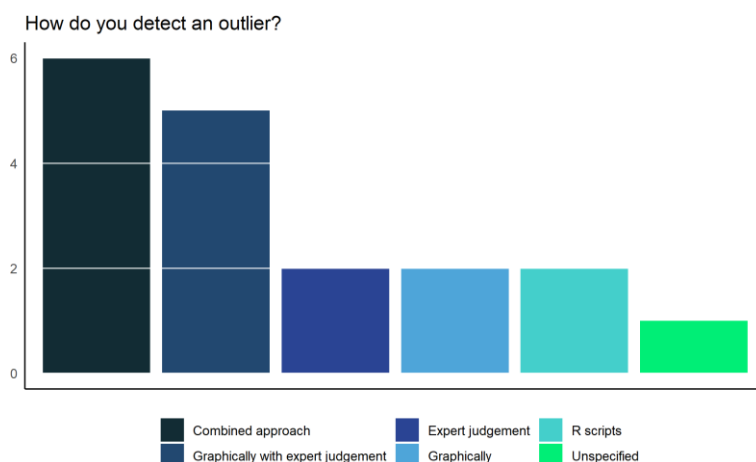


Figure 2: Frequency of summarised responses to question Q3.4.3 – Method used to detect outliers.

When asked how the outlier check was conducted, six respondents utilised a combined approach, using a combination graphical, expert judgement and R scripts to check for outliers. The combined methods of these five respondents can be seen in table 2.

Where respondents had a single approach for detecting outliers, graphical detection with expert judgement was the most common method (n = 5) “Graphically using expert judgment, creating common graphs such as scatter plots, histograms, box plots in R with ggplot2 package”, IEO(a)). To ensure potential outliers were in fact outliers and not extreme values, expert judgement was considered essential (“Identification of outliers can be done visually on the available plots and tables.... Expert judgement is important in the outliers identification process because in some cases an outlier is connected with natural reasons, e.g. diseases, parasites, poor condition.”, NMFRI).

Two respondents detected outliers graphically, and while expert judgement may have played a role, this was not stated in the answers. Two respondents used R scripts to detect outliers, though no additional information on the script itself was offered (“scripts mostly”, THN). Finally, one respondent did not state how they conducted their outlier check, just that it was conducted (“internal calculations to Toughbook”, SLUB(b)).

Table 2: Primary and secondary methods used to check for outliers by respondents who employed a combined approach to question 3.4.3

Institute	Primary	Secondary
BIOR	Graphically	Excel
KU	Graphically	R scripts
WMR(a)	Expert judgement	R scripts
FEAS-MI	Graphically	R script
IEO(a)	Graphically with expert judgement	R scripts
ILVO	Graphically with expert judgement	R scripts



3.4.4 At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

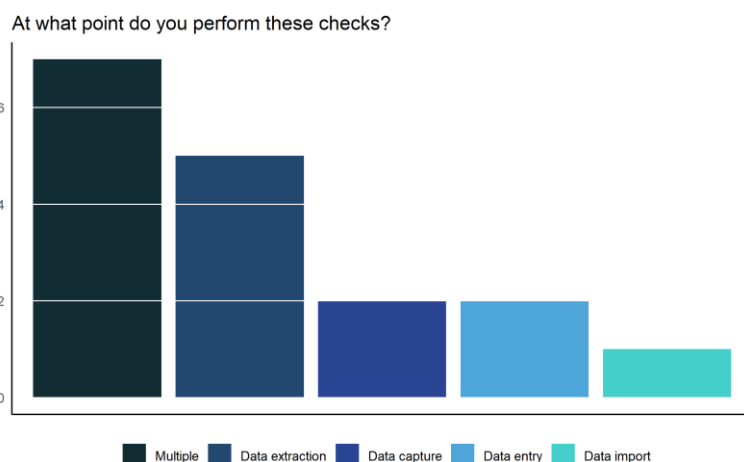


Figure 3: Frequency of summarised responses to question Q3.4.4 – When were outlier checks performed.

Table 3: Points where outlier checks were performed by respondents who answered 'multiple points' to question 3.4.4

Institute	Point 1	Point 2	Point 3
THN	Data entry	Data extraction	NA
WMR(a)	Data capture	Data import	Data extraction
WMR(b)	Data capture	Data import	Data extraction
IEO(a)	Data import	Data extraction	NA
FEAS -MI	Data entry	Data import	Data extraction
ILVO	Data entry	Data extraction	NA
NMFRI	Data entry	Data extraction	NA

Table 4: Summarised responses of all respondents to question 3.4.

Institute	Outlier definition	Outlier detection	Point of check
AZTI	Graphically	Graphically with expert judgement	Data extraction
BIOR	Graphically	Combined approach	Data entry
DRP-RAA	Graphically	Graphically	Data extraction
DTU(a)	Graphically	Expert judgement	Data extraction
DTU(b)	Graphically	Expert judgement	Data extraction
EMI	Expert judgement	Graphically	Data capture
FEAS -MI	Comparison with historical data	Combined approach	Multiple
IEO(a)	Numerically	Combined approach	Multiple
IEO(b)	Deviation from common trend	Graphically with expert judgement	Data import
ILVO	Comparison with historical data	Combined approach	Multiple
LUKE	Deviation from common trend	Graphically with expert judgement	NA
NMFRI	Deviation from common trend	Graphically with expert judgement	Multiple
SLU(a)	Graphically	R scripts	Data extraction
SLU(b)	Numerically	Unspecified	Data capture
THN	Comparison with historical data	R scripts	Multiple
KU	Numerically	Combined approach	Data entry
WMR(a)	Graphically	Combined approach	Multiple
WMR(b)	Graphically	Graphically with expert judgement	Multiple



Q 3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

Do you perform any cross checks of sample data with census data?

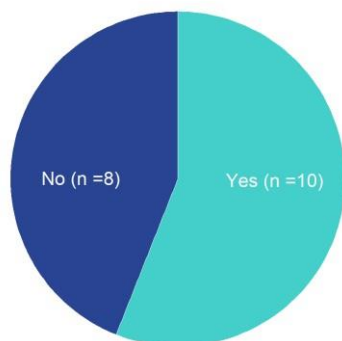


Figure 1: Frequency of categorised responses to Q3.5 – do you perform cross checks with census data?

When asked whether they cross checked sample data with census data, 10 respondents stated that they did while eight did not.

At what point do you perform these checks?

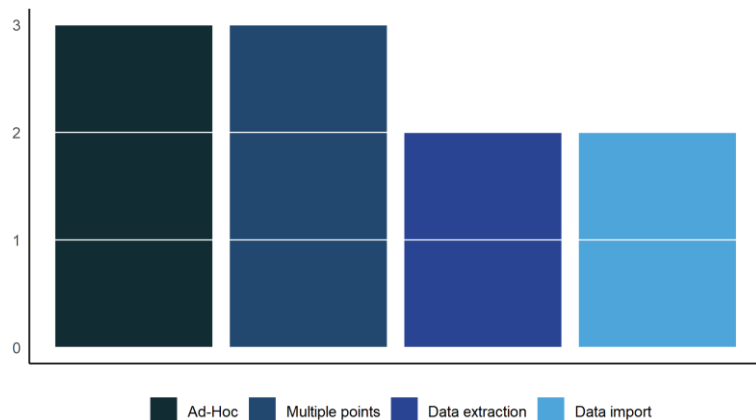


Figure 2: Frequency of categorised responses to Q3.5 – At what point do you perform these checks?

When asked at what point during the data collection process, they performed a cross check between census and sample data, three respondents stated that checks were only performed on an Ad-Hoc basis (*“Not as a routine. On a more ad-hoc basis, some technicians do it during data capture and samples are sometimes checked during estimation”*, DTU(a)).

Three respondents performed the checks at two or more points in the data collection process, (*“..During data capture and extraction, at-market and at-sea sampling are cross-checked with sales notes and logbooks...”*, DRF-RAA).

Two respondents performed the checks at the point of data extraction, usually prior to answering data calls (*“It is done during the data quality process before answering data calls.”*, AZTI).





Two respondents performed the check at data import, when data was being imported into the primary database (*"Data on fishing effort and landings for the sampled trip are imported into IMPORT workbook after all these data are recorded into national fisheries data information system ... Simple R script extracts relevant data based on logbook number and landing data."*, KU)

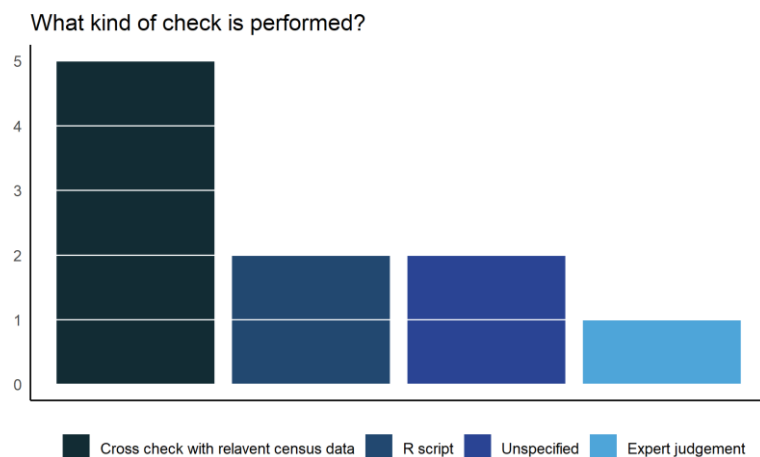


Figure 3: Frequency of categorised responses to Q3.5 – How do you perform these checks?

Five respondents compared sampling data with the relevant census data (*"During data extraction the sampling levels are checked against commercial landings using temporal (quarter), technical (gear type) and spatial (ices sub-division) variables to check if there are sufficient samples for each sampling stratum."*, FEAS-MI. , *"..There are cross checks between the sample and the trip in respect to area, metier, vessel name and weight"*, WMR(a)).

Two respondents stated that checks were performed, however how the check was performed was not specified (*"Not as a routine. On a more ad-hoc basis, some technicians do it during data capture and samples are sometimes checked during estimation"*, DTU(a)).

Two respondents incorporated the checks into an R script, which automatically cross-checked census and sample data (*"The pairing/crosschecking process between the sampled trips and the official data consists in crossing both sources through an R script in order to assign to each sampled trip the corresponding fishing trip of the NVDP (metierized database of official data)"*, IEO(a)).

A single respondent (THN), cross checked census and sample data by way of expert judgement, however no further information on the process was offered (*"Not on regular basis and only based on expert judgement"*, THN).



Q3.6 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

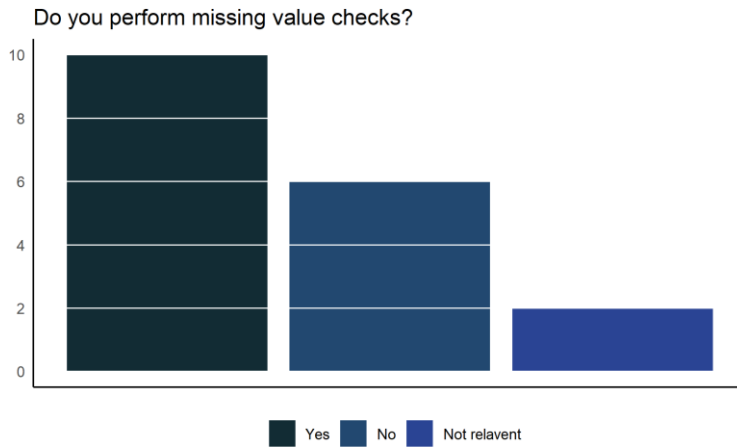


Figure 1: Frequency of categorised responses to Q3.6 – do you perform missing value checks?

Ten respondents did conduct some form of missing value checks during the data collection process. Six respondents did not conduct missing value checks, while missing value checks were not relevant to the data in question for two respondents (WMR(a), EMI).

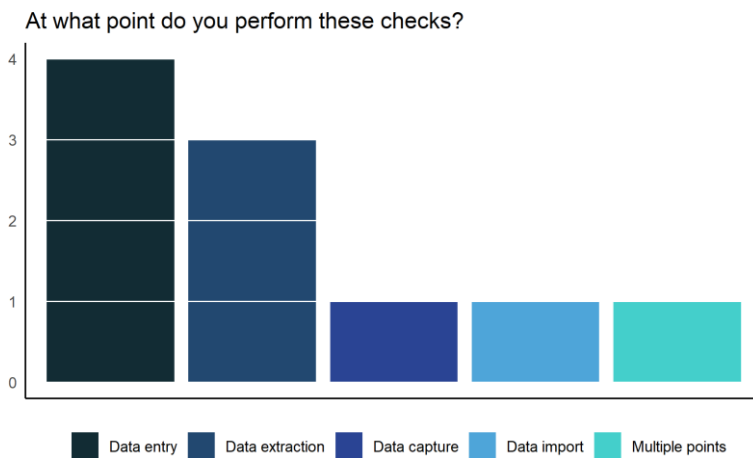


Figure 2: Frequency of categorised responses to Q3.6 – At what point do you perform missing value checks?

Of the ten respondents who did perform the checks, four performed the checks at the point of data entry. Three respondents conducted the check at the data extraction phases prior to answering data calls. One respondent performed the check at the point of data capture, one at the point of data import, and one performed at the checks at multiple points during the data collection process.

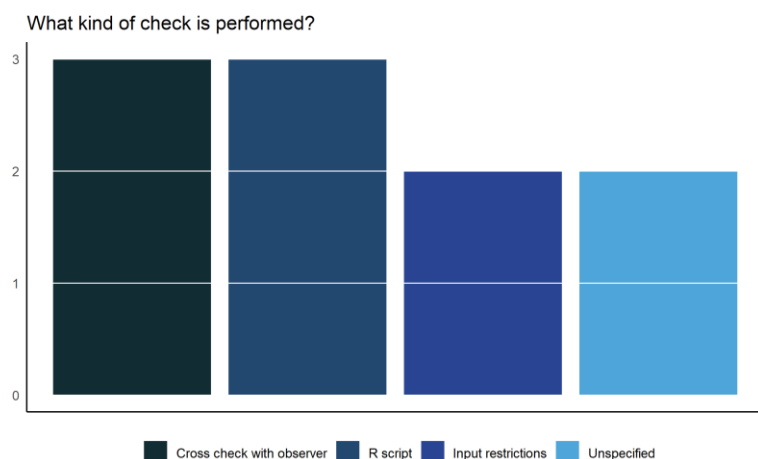


Figure 3: Frequency of categorised responses to Q3.6 – What kind of missing value check is performed?

When asked how they performed the check, three respondents cross checked data with original observer record sheets, to determine whether the missing value was an error or a true zero (*“In cases of mismatch, then the observers are asked to check. The same is true, if it is indicated that both discards and landings have been work up, but no recording of discard is found..”*, DTU(a)).

Three respondents conducted the check by using an R script to check for missing values in fish length weights (*“During data extraction, length and weight ranges are investigated in R using the command “table(Dataset\$weight, use.NA=“always”)”.*”, ILVO), weight and sex (*“For Baltic Sea simple R script created to detect some missing values: missing individual weight, missing sex.”*, KU).

For two respondents, their data entry software employed restrictions which ensured all required fields were filled (*“The data entry software ensures that all mandatory information is registered. For biological parameters, the shiny application designed for data quality control, allows to list all records where age information has not yet been registered.”*, NMFRI, *“Our data recording system (SIRENO) doesn’t allow the introduction of missing values/zeros for length variable.”*, IEO(b), preventing missing values being input with the data.

Two respondents did not specify how they conducted the check, with their answers, only noting if and when the check was performed.

Table 1: Summary of categorised responses to for all respondents to Q3.6

Institute	Checks	Point	Method
AZTI	No	NA	NA
BIOR	Yes	Data entry	Cross check with observer
DRP-RAA	No	NA	NA
DTU(a)	Yes	Data extraction	Cross check with observer
DTU(b)	No	NA	NA
EMI	Not relevant	NA	NA
FEAS -MI	No	NA	NA
IEO(a)	Yes	Data import	Input restrictions
IEO(b)	Yes	Data extraction	Unspecified
ILVO	Yes	Multiple points	R script
LUKE	Yes	Data extraction	Unspecified
NMFRI	Yes	Data entry	Unspecified
SLU(a)	Yes	Data entry	R script
SLU(b)	Yes	Data capture	Cross check with observer
THN	No	Data capture	Unspecified
KU	Yes	Data entry	R script
WMR(a)	Not relevant	NA	NA
WMR(b)	No	NA	NA



Q 3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Do you perform any spatial data checks?

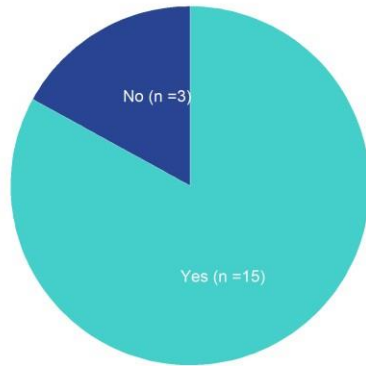


Figure 1: Frequency of categorised responses to Q3.7– do you perform spatial data checks?

When asked whether they conducted any spatial data checks, 15 respondents answered that they did. Three respondents did not perform any such check, accepting spatial information as is (“No spatial checks yet. Logbook records accepted as reliable spatial information.”, KU).

At what point do you perform these checks?

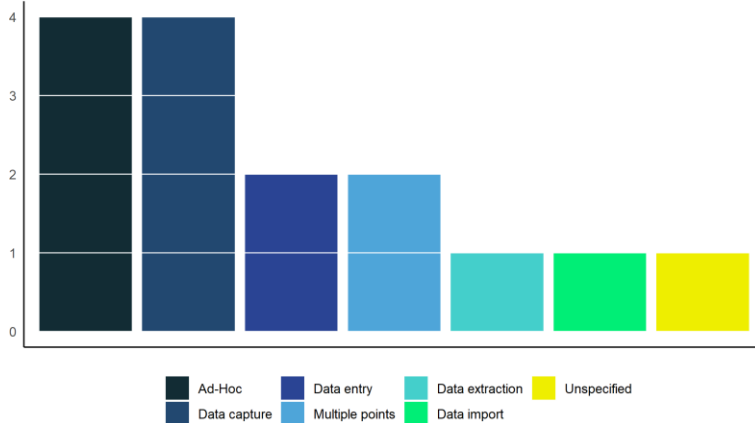


Figure 2: Frequency of categorised responses to Q3.7 – At what point do you performed spatial data checks?

Four Respondents performed the spatial data checks on an Ad-Hoc basis. Four respondents conducted the check at the point of data capture. Two respondents performed the check at data entry, and two performed the check at multiple points during the data collection process. One respondent performed the check during data extraction, one at the point of data import and one did not specify when they performed this check.



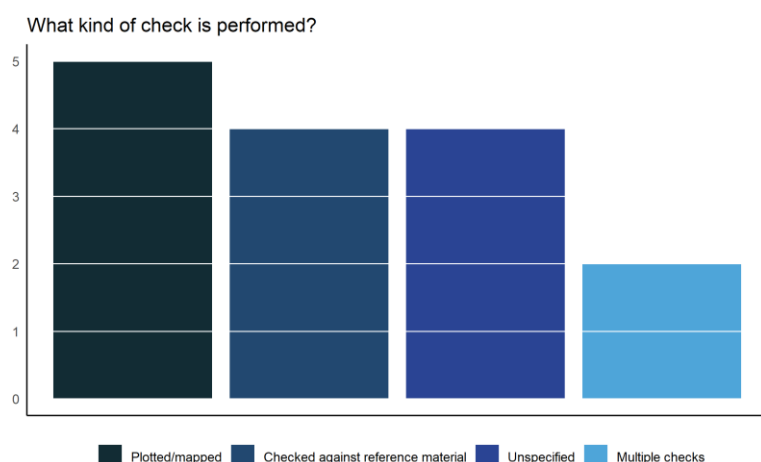


Figure 3: Frequency of categorised responses to Q3.7 – What kind of spatial data check is performed?

The most common spatial data check (n = 5) conducted was to plot or map the data to check if geographical coordinates were realistic and accurate (*“Yes, the coordinates of the sample and census (catch) data are plotted in a map.”*, WMR(a), *“Area and rectangles are calculated automatically depending on the coordinates. They are also plotted in a map to detect clearly wrong positions.”*, AZTI).

Following mapping, the next most common form of spatial data check (n = 4) was to check recorded data against reference material to identify errors. Reference material was usually either a reference table (*“These checks are carried out using a set of reference tables which enable to ensure the consistency of coordinates, areas, rectangles and national sub-polygons.”*, NMFRI) or logbook data (*“Geographical sampling information are checked with logbook data to verify the ICES Division (for market sampling) and the ICES rectangle (for on board sampling).”*, IEO(a)).

Four respondents did not specify how they conducted the check, instead only stating if and when the check was performed during the data collection process (*“Not many. Some during the estimation.”*, SLU(a)).

Two respondents employed a combined approach (FEAS-MI, DRP-RAA) creating both plots of the data and either checking against reference material (*“...These are corrected either visually by plotting positions on a map (Fig. 10) or by reference to original data sheets.”*, FEAS-MI) or checking species presence absence in that area (*“At the time of data extraction, the spatial distribution is visualized, and wrong coordinates are corrected (which usually occurs due to data entry errors - transposition error). Ad-hoc crossing of areas with the presence/absence of species is also carried out, but not systematically.”*, DRP-RAA).

Categorised answers of all respondents can be seen in table 1.





Table 1: Categorised responses for all respondents to question 3.7

Institute	Checks	Point	Method
AZTI	Yes	Data capture	Plotted/mapped
BIOR	Yes	Ad-Hoc	Unspecified
DRP-RAA	Yes	Data entry	Multiple checks (Plotted and mapped, Species Presence/Absence)
DTU(a)	Yes	Ad-Hoc	Plotted/mapped
DTU(b)	Yes	Ad-Hoc	Unspecified
EMI	Yes	Data capture	Checked against reference material
FEAS -MI	Yes	Unspecified	Multiple checks (Plotted and mapped, Checked against reference material)
IEO(a)	Yes	Data import	Checked against reference material
IEO(b)	No	NA	NA
ILVO	Yes	Data extraction	Plotted/mapped
LUKE	No	NA	NA
NMFRI	Yes	Data entry	Checked against reference material
SLU(a)	Yes	Ad-Hoc	Unspecified
SLU(b)	Yes	Data capture	Checked against reference material
THN	Yes	Data capture	Unspecified
KU	No	NA	NA
WMR(a)	Yes	Multiple points	Plotted/mapped
WMR(b)	Yes	Multiple points	Plotted/mapped





Q 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Do you perform any temporal data checks?

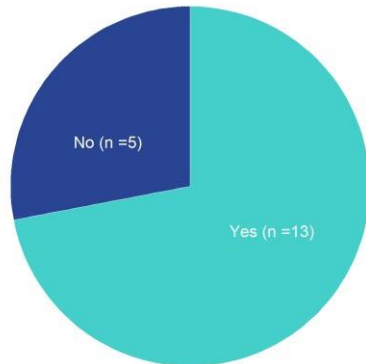


Figure 1: Frequency of categorised responses to Q3.8 – Do you perform any temporal data checks?

When asked whether performed any temporal data checks, 13 respondents stated that they did perform this check, while five respondents stated that they did not perform any temporal data checks.

At what point do you perform these checks?

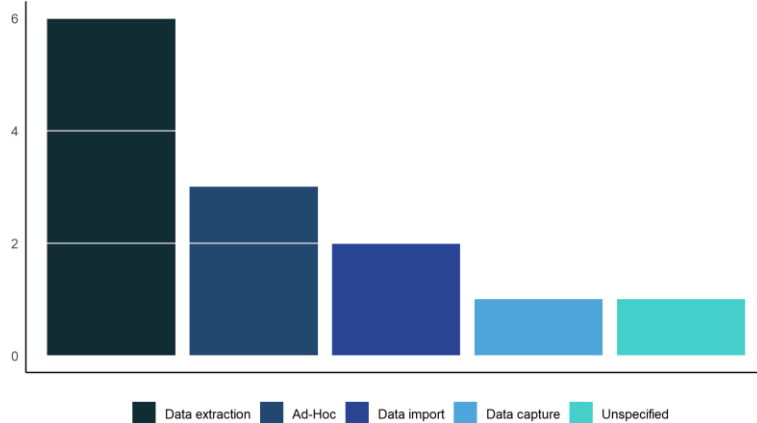


Figure 2: Frequency of categorised responses to Q3.8 – At what point do you perform temporal data checks?

Six respondents performed the check at the point of data extraction, prior to answering data calls. Three respondents performed the check only on an Ad-Hoc basis, one of which stated that temporal checks only occurred in response to other studies (“...checks (quarters or years) are usually carried out as part of other studies, not as part of the sampling process itself.”, IEO(b)). Two respondents carried out the checks at the point of data import. One respondent carried out the check during data capture, and one respondent did not specify when they carried out the check.

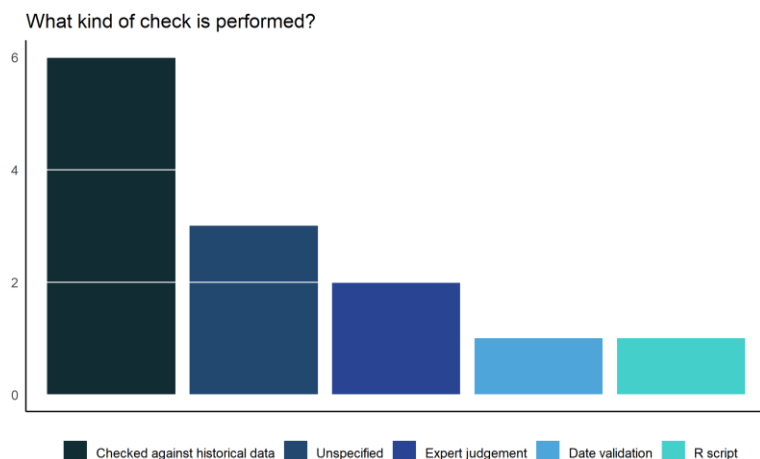


Figure 3: Frequency of categorised responses to Q3.8 – What kind of temporal data check is performed?

When asked how they performed the check, six respondents did so by checking the data against historical data, usually comparing data from previous years or quarters (*“Comparison of data by years, quarters is performed in the annual report of the institute”*, BIOR, *“Cumulative length frequency distributions for each stock metier are compared across quarters to check if merging of temporal strata is sensible. During data extraction sampling levels are checked against commercial landings by quarter to ensure that there are sufficient samples in each temporal stratum”*, FEAS-MI).

Three respondents did not specify how they performed the check, just if and when the check was performed (*“Yes. During the estimation.”*, SLU(b)).

Two respondents relied on expert judgement to cross check temporal data (*“Expert judgement used to quality check certain parameters is therefore built over the years.”*, ILVO).

One respondent validated trip dates by cross checking sample data with known trip information (*“The check consists in ensuring that the sample date is within or close to the trip dates, depending on the type of fishery.”*, NMFRI).

One respondent conducted the check through use of an R script which generated summary statistics for a variety of parameters and checked them against values from previous years and quarters (*“Simple R script for description of summary data statistics by species, year, quarter and metier...”*, KU).

Categorised answers of all respondents can be seen in table 1.



Table 1 : Categorised responses for all respondents to Q3.8

Institute	Checks	Point	Method
AZTI	No	NA	NA
BIOR	Yes	Data extraction	Checked against historical data
DRP-RAA	Yes	Ad-Hoc	Unspecified
DTU(a)	Yes	Unspecified	Checked against historical data
DTU(b)	No	NA	NA
EMI	No	NA	NA
FEAS -MI	Yes	Data extraction	Checked against historical data
IEO(a)	Yes	Data import	Expert judgement
IEO(b)	Yes	Ad-Hoc	Unspecified
ILVO	Yes	Data import	Expert judgement
LUKE	No	NA	NA
NMFRI	Yes	Data extraction	Date validation
SLU(a)	Yes	Data extraction	Unspecified
SLU(b)	Yes	Data capture	Checked against historical data
THN	No	NA	NA
KU	Yes	Ad-Hoc	R script
WMR(a)	Yes	Data extraction	Checked against historical data
WMR(b)	Yes	Data extraction	Checked against historical data





Q 3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Do you perform any duplication checks?

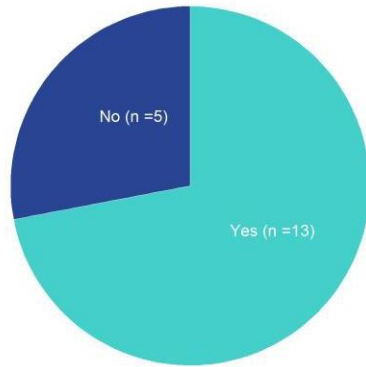


Figure 1: Frequency of categorised responses to Q3.9 – Do you perform any duplication checks?

When asked whether they conducted any duplication checks during the data collection process, 13 respondents stated that they did. Five respondents stated that they did not perform any duplication checks.

At what point do you perform these checks?

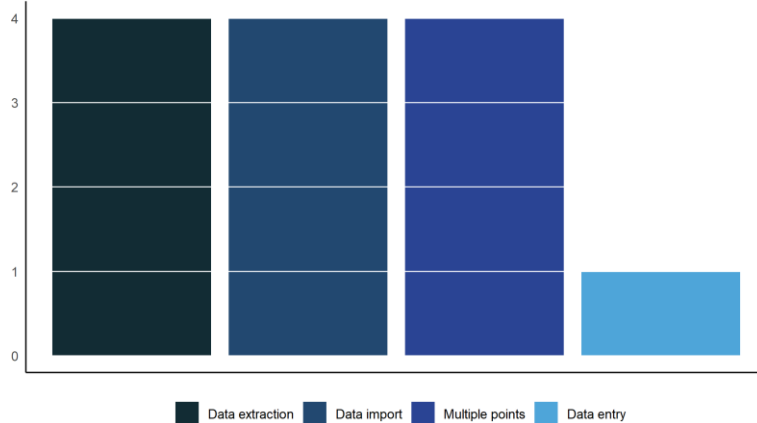


Figure 2: Frequency of categorised responses to Q3.9 – At what point do you perform data duplication checks?

Four respondents carried out the check at the point of data extraction. Four respondents carried out the check at the point of data import. Four respondents carried out duplication checks at multiple points during the data collection process, usually at the data import and data extraction (“Yes, duplications are checked for at several occasions, when importing data from the field, ad hoc in the database (for things that cannot be checked when registration or import of electronic data occurs) and when delivering data to ICES.”, SLUB(b), “During data import and extraction the number of rows in the original data set is checked against the number of rows of the same data set when the distinct values are filtered out..”, WMR(a)). One respondent carried out the duplication check at the point of entering the data into the primary database (“The database constraints prevent from entering duplicates in some data entry steps”, NMFRI).



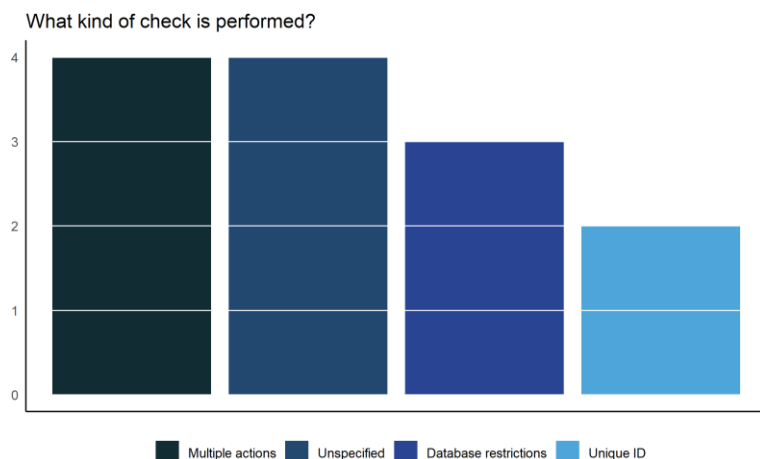


Figure 3: Frequency of categorised responses to Q3.8 – What kind of duplication check is performed?

When asked how they performed the duplication checks, four respondents did not specify how they performed the check, just if and when the check was performed (*“We have some duplication checks for sampling data. We do it during the data quality process before answering data calls”, AZTI*).

Three respondents had constraints or restrictions on their database which prevented the entry of duplicate records (*“SIRENO database or icrOS system doesn’t allow the introduction of duplicated data.”, IEO(b), “Yes, the Smartfish application does not allow users to create duplicated samples during the data capture process. Similar process is valid when working with the age reading tool Smartdots”, ILVO*).

Two respondents utilised unique IDs for each sample, where the same ID cannot be used twice. Unique sample IDs were generated either through primary and foreign keys (*“All tables in the national database related with primary and foreign keys, which reveal the duplications”, THN*) or through unique combinations of haul, biological and date information collected (*“Yes, duplications are checked for at several occasions, when importing data from the field, ad hoc in the database... Things that are compared are eg. but not only: • The combination any vessel and fromdatetime must be unique. • The combination fish number and catch id must be unique.”, SLU(b))*).

Four used a combined approach from preventing duplicate entries. Three of these used unique ID’s and parallel tables (*“During data import and extraction the number of rows in the original data set is checked against the number of rows of the same data set when the distinct values are filtered out.. Furthermore, each sample is assigned to a unique sample ID. A unique sample ID can’t be entered in the database twice”, WR(b)*), and one used database restrictions and parallel tables (*“The database constraints prevent from entering duplicates in some data entry steps. Checksums are available at the level of entering biological data. Moreover, a relation with a parallel system for PSU selection, enables to identify potential duplicates.”, NMFRI*). Details of combined approaches can be found in table 1.

Categorised answers of all respondents can be seen in table 1.



Table 1: Categorised responses for all respondents to question 3.9.

Institute	Checks	Point	Method
AZTI	Yes	Data extraction	Unspecified
BIOR	No	NA	NA
DRP-RAA	Yes	Data extraction	Database restrictions
DTU(a)	No	NA	NA
DTU(b)	No	NA	NA
EMI	No	NA	NA
FEAS -MI	Yes	Data import	Multiple points (Unique ID, Parallel table)
IEO(a)	Yes	Data import	Unspecified
IEO(b)	Yes	Data import	Database restrictions
ILVO	Yes	Multiple points	Database restrictions
LUKE	No	NA	NA
NMFRI	Yes	Data entry	Multiple points (Database restrictions, Parallel table)
SLU(a)	Yes	Data extraction	Unspecified
SLU(b)	Yes	Multiple points	Unique ID
THN	Yes	Data extraction	Unique ID
KU	Yes	Data import	Unspecified
WMR(a)	Yes	Multiple points	Multiple points (Unique ID, Parallel table)
WMR(b)	Yes	Multiple points	Multiple points (Unique ID, Parallel table)





3.10 Please let us know about any other relevant data checks which have not already been described in your answers

When asked about any other relevant data checks they performed, ten respondents stated did not have any other relevant checks or they left the question blank. For the eight respondents who answered the question, categorisation was not appropriate so table 1 below shows their full responses in addition to links to data where possible. Associated images for answers can be found in the relevant appendices.

Table 1: Full responses for respondents who gave details of any additional data checks they performed in the data collection process.

Institute	q3.10	Links
AZTI	We check census data for errors in species identification, for these species which are clearly wrong because they cannot be present in our waters. We check metier & area combination.	
BIOR	As I mentioned above, I am working in the sea alone. Biological data with the otoliths are collected and returned in special paper books. For each individual fish such information is collected, length, full weight, sex, maturity and otoliths. Otoliths are wrapped in page similar to an envelope. At this example is cod with length 47 cm, weight 1,03 kg, female with maturity stage 5. After data input in Excel file, the age reader receives paper books with otoliths and file with the entered data. During the otolith preparation for age reading additional data quality check is performed, if necessary, corrections are made.	
DTU(a)	Ad-a) Different relevant checks are done as a routine on the at-sea observer trips per trip and quarter, see attached pdf's	
DTU(b)	Ad-a) Different relevant checks are done as a routine on the at-sea observer trips per trip and quarter, see attached pdf's	
EMI	Since our data is uploaded to ICES RDB, the RDB data checking system performs many checks.	
FEAS -MI	F:\Logbooks_Current_report – for some checks on the logbook data that is used to raise the sample data to the population level Length/Frequency plots are generated during data entry. This plot updates automatically within Nemesys as commercial data is electronically captured at sea. Figure Yes. An example of one of the sections in the Nephrops Measuring System (Nemesys) Data Validation Reports and similar length frequency/plots have been added into our commercial port sampling data entry application (Stockman) QC Weights added into Nemesys -described above Voice Report Validation tool for validating entered commercial discards data. Data is entered through paper sheets into our Commercial Discards	





	Database, and the entered is validated through a Voice Reporting Application.	
IEO(a)	http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework	http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework





Q3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

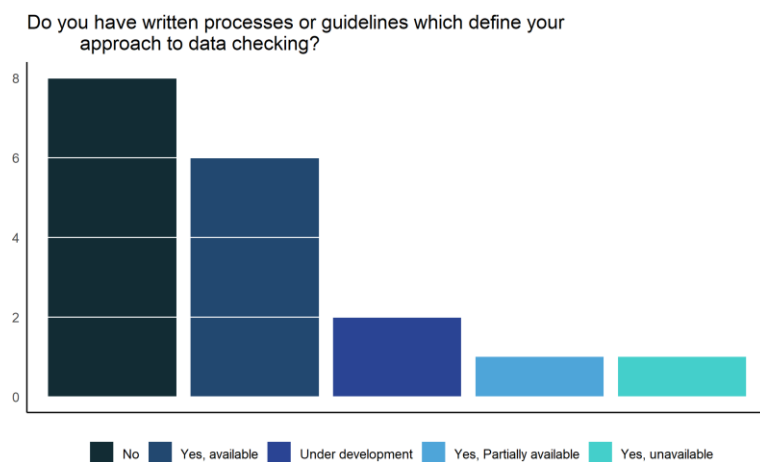


Figure 1: Frequency of categorised responses to Q3.10 – Do you have any written process or guidelines relevant to your approach to data checks?

When asked whether they used any written processes or guidelines for their data quality control checks, eight respondents did not use any such written guidelines for their data checking. Six respondents did use guidelines which were made available, links for which can be found in table 1. Two respondents do not have but are currently developing such guidelines. One respondent had such guidelines but due to GDPR sensitive information, stated they could only provide a censored version on request. Finally, one respondent had such documentation but did not want to provide it as they consider it the intellectual property of their institute.





Table 1: Categorised responses for all respondents to Q3.11

Institute	q3.11	Link
AZTI	Yes, unavailable	NA
BIOR	No	NA
DRP-RAA	Under development	NA
DTU(a)	No	NA
DTU(b)	No	NA
EMI	Yes, available	https://www.envir.ee/sites/default/files/andmetootluse_juhend.pdf
FEAS - MI	Yes, partially available	Censored version available upon request.
IEO(a)	Yes, available	http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework
IEO(b)	No	NA
ILVO	Yes, available	Available upon request
LUKE	No	NA
NMFRI	Yes, available	tinyurl.com/dpadesdd
SLU(a)	No	NA
SLU(b)	No	NA
THN	Under development	NA
KU	No	NA
WMR(a)	Yes, available	Image provided - see appendix
WMR(b)	Yes, available	Image provided - see appendix





Section 4 - Data editing

Section 4 asked respondents about data editing and to outline their procedure for dealing with any errors, inconsistencies or discrepancies found in their data.





Q4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

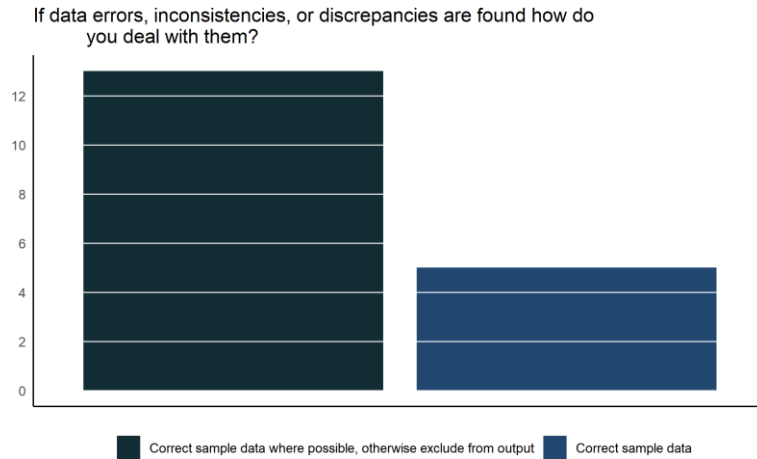


Figure 1: Frequency of categorised responses to Q4.1 – How do you deal with any errors, inconsistencies or discrepancies found in your data?

When asked how they dealt with errors, inconsistencies or discrepancies found in the data, there appears to be a broad consensus among respondents, with all respondents answering that they attempted to correct the error in the sample if possible.

13 respondents stated when errors, inconsistencies or discrepancies are found, they attempted to correct the data where possible, and if data could not be corrected it was excluded from outputs “*If a data point is identified as an outlier, first it is examined if it’s a wrong entry and if not, it is transmitted to the laboratory technicians to check if the value is an actual observation or a mistake. If the technician points it out as a mistake the data is removed from the database and consequently excluded from any output.*”, WRM(b). Five respondents stated that they corrected the sample data where possible, however they did not state how they dealt with data that could not be corrected (“*Sample date will be corrected when possible before data supply*”, THN). Overall, data correction was generally carried out by referring to the original data collection sheets (“*... must be reviewed by the supervisors, usually implying review of the original sampling sheets.*”, IEO(a)).

Categorised answers of all respondents can be seen in table 1.



Table 1: Categorised responses for all respondents to Q4.1

Institute	q4.1
AZTI	Correct sample data where possible, otherwise exclude from output
BIOR	Correct sample data where possible, otherwise exclude from output
DRP-RAA	Correct sample data
DTU(a)	Correct sample data where possible, otherwise exclude from output
DTU(b)	Correct sample data where possible, otherwise exclude from output
EMI	Correct sample data where possible, otherwise exclude from output
FEAS -MI	Correct sample data where possible, otherwise exclude from output
IEO(a)	Correct sample data
IEO(b)	Correct sample data where possible, otherwise exclude from output
ILVO	Correct sample data where possible, otherwise exclude from output
LUKE	Correct sample data where possible, otherwise exclude from output
NMFRI	Correct sample data
SLU(a)	Correct sample data where possible, otherwise exclude from output
SLU(b)	Correct sample data where possible, otherwise exclude from output
THN	Correct sample data
KU	Correct sample data
WMR(a)	Correct sample data where possible, otherwise exclude from output
WMR(b)	Correct sample data where possible, otherwise exclude from output





Q 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

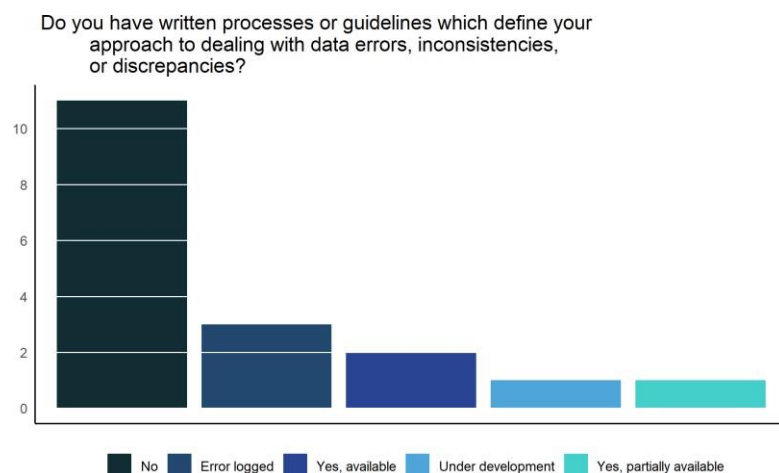


Figure 1: Frequency of categorised responses to Q4.2 – do you have any guidelines for dealing with any errors, inconsistencies or discrepancies found in your data?

When asked whether they had any written processes or guidelines for dealing with such errors, 11 respondents answered that they did not have any such guidelines. Three respondents outlined the process they use to record such errors when they arise, attempting to prevent similar errors in the future (*“The data errors, inconsistencies and/or discrepancies are recorded in dedicated documents during the data checking process annually. For example, if an error is found in the sample data the following mandatory fields need to be field in the documentation template SampleID, Species, DateChecked, ErrorDescription, ActionsTaken (e.g. excluded, corrected), Reason, DateProcessed, Re-imported (Yes/No), Who”*, WMR(a)). Two respondents were able to provide the guidelines or documentation which defined their approach to dealing with such errors. One respondent stated that they are currently developing such guidelines, and one respondent was able to provide only some of their guidelines, as others contained sensitive information unavailable for publication.



Table 1: Categorized responses for all respondents to Q4.1 – Do you have any guidelines for dealing with errors, inconsistencies, and discrepancies in your data? Where respondents provided a link, the link has also been given in the table.

Institute	q4.2	link
AZTI	No	NA
BIOR	No	NA
DRP-RAA	No	NA
DTU(a)	No	NA
DTU(b)	No	NA
EMI	Yes, available	https://www.envir.ee/sites/default/files/andmetootluse_juhend.pdf
FEAS -MI	Yes, partially available	https://wwz.ifremer.fr/cost/
IEO(a)	No	NA
IEO(b)	No	NA
ILVO	Yes, available	See data extraction protocol for ICES combined data call
LUKE	No	NA
NMFRI	No	NA
SLU(a)	No	NA
SLU(b)	No	NA
THN	Under development	NA
KU	Error logged	NA
WMR(a)	Error logged	NA
WMR(b)	Error logged	NA





Section 5 - Data imputation

Section 5 asked respondents about their approach to dealing with any gaps in their data. Specifically, respondents were asked about gaps in age length keys (ALK's) , weight length keys (WLK's) and sampling strata.





Q5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

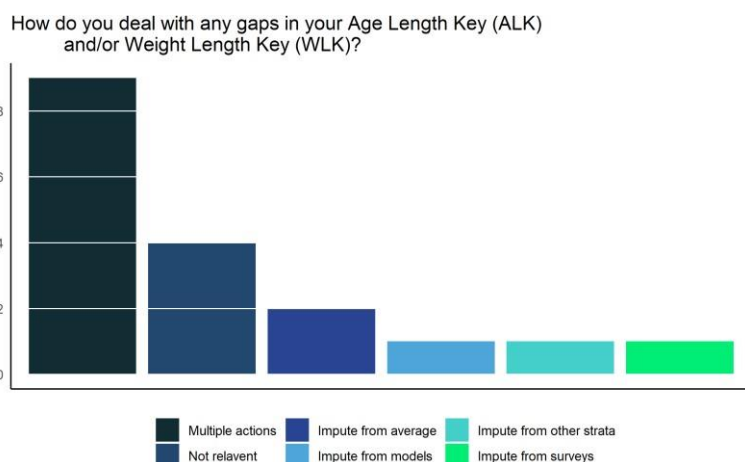


Figure 1: Frequency of categorised responses to Q5.1– How do you deal with any gaps in your ALK’s or WLK’s?

When asked how they dealt with any gaps in Age length keys or Weight length keys, two respondents imputed a value from an average (“*HER and SPR: impute missing values from averages*”, LUKE, “*In cases of gaps in ALK or WLK, average values are used if available.*”, NMFRI).

One respondent imputed values to fill the gaps from a multinomial logistic model (“*To deal with gaps in ALKs and to assure good estimates for length categories which are poorly sampled, age-length keys (ALK) are modelled based on the observed ALKs using a multinomial logistic regression model (Gerritsen et al., 2006)*”, ILVO).

One respondent imputed values from other strata where available (“*For age data the ALK are merged across technical strata but there still might be gaps. To make things efficient, an assumption that the differences in the ALK between areas are minor enough to be ignored, so age data from all areas are combined into one but the quarterly stratification is kept.*”, FEAS-MI).

One respondent dealt with gaps by imputing a value from fisheries independent surveys (“*Impute missing values from surveys, if possible.*”, SLU(b)).

Most respondents (n = 9) employed a combination of the above actions. For example, some imputed values from survey data, before filling any further gaps based on expert judgement (“*age length key (ALK) of the commercial sampling is completed with the age-length survey data and the missing values are completed by an age expert judgement.*”, IEO(b)). Others attempted to impute values from averages, followed by surveys followed by models (“*Missing values are imputed first from averages, then from surveys, then from models.*”, WMR(a)).

For three respondents, ALK’s and WLK’s were not relevant to their data.



Table 1: Categorised responses to Q5.1 – dealing with gaps in ALK's and WLK's. Where respondents employed a combined approach, all their responses are listed.

Institute	q5.1_1	Action 1	Action 2	Action 3
AZTI	Multiple actions	Impute from average	Impute from other strata	NA
BIOR	Multiple actions	Impute from average	Fill by expert judgement	NA
DRP-RAA	Not relevant	NA	NA	NA
DTU(a)	Multiple actions	Impute from average	Impute from models	NA
DTU(b)	Multiple actions	Impute from average	Impute from models	NA
EMI	Not relevant	NA	NA	NA
FEAS -MI	Impute from other strata	NA	NA	NA
IEO(a)	Not relevant	NA	NA	NA
IEO(b)	Multiple actions	Impute from other strata	Fill by expert judgement	Leave the gaps
ILVO	Impute from models	NA	NA	NA
LUKE	Impute from average	NA	NA	NA
NMFRI	Impute from average	NA	NA	NA
SLU(a)	Not relevant	NA	NA	NA
SLU(b)	Impute from surveys	NA	NA	NA
THN	Multiple actions	Impute from other strata	Impute from surveys	NA
KU	Multiple actions	Impute from average	Impute from models	Impute from surveys
WMR(a)	Multiple actions	NA	NA	NA
WMR(b)	Multiple actions	Impute from average	Impute from surveys	Impute from models





Q5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

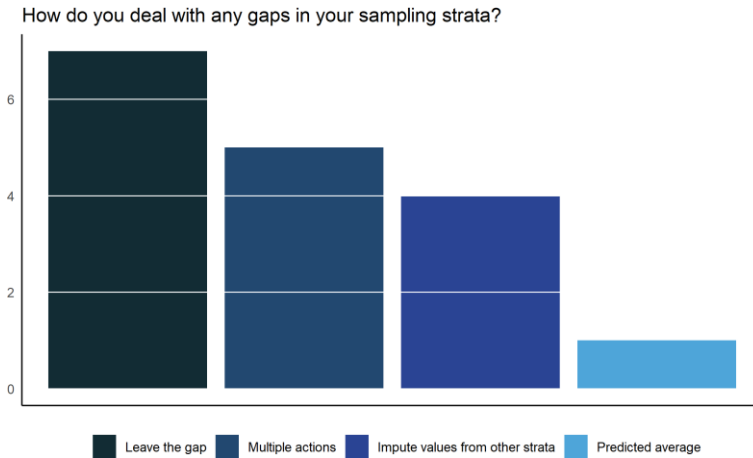


Figure 1: Frequency of categorised responses to Q5.2– How do you deal with any gaps in your sampling strata?

When asked how they dealt with gaps in sampling strata, seven respondents opted to leave the gaps in the data, allowing the ICES stock coordinator to decide how best to deal with them (“*Since the implementation of InterCatch (IC), we do not apply imputations, as it can be done by the stock coordinator after the integration of all international data...*”, IEO(a), “*Imputation is not performed at national level but at Stock Data Coordination level. Data are provided to end user "as-is" (as collected, validated and recorded in national database).*”, NMFRI).

Five respondents performed multiple actions to deal with gaps in the sampling strata. These included leaving the gap followed by imputing from other strata (“*If there is a major stratum that has insufficient samples then the sample data can either be deleted for that stratum or it can be submitted with a warning. It is preferable to let the ICES stock coordinator deal with gaps .For species that are reported by length and for which there is no biological sampling (i.e. weights-at-length) the length-weight parameters will need to be supplied to estimate the sample weights... an Age-Length Key then becomes a Length-Length key, which is a convoluted way of raising the data has the functionality of merging strata etc.*”, FEAS-MI), and imputing from survey data followed by filling gaps based on expert judgment (“*For small pelagic stocks, age length key (ALK) of the commercial sampling is completed with the age-length survey data and the missing values are completed by an age expert judgement.*”, IEO(b)).

Four respondents imputed values from other strata to fill gaps (“*Strata, commercial size categories, do not match the ones in InterCatch, so missing values are imputed from other strata.*”, DTUB(b), “*Usually, impute missing values from other strata.*”, DRP-RAA).





Table 1: Categorised responses to Q5.2– How do you deal with any gaps in your sampling strata?

Institute	q5.2_1	q5.2_2	q5.2_3
AZTI	Multiple actions	Leave the gap	Impute values from other strata
BIOR	Leave the gap	NA	NA
DRP-RAA	Impute values from other strata	NA	NA
DTU(a)	Leave the gap	NA	NA
DTU(b)	Impute values from other strata	NA	NA
EMI	NA	NA	NA
FEAS -MI	Multiple actions	Leave the gap	Impute values from other strata
IEO(a)	Impute values from other strata	NA	NA
IEO(b)	Multiple actions	Impute values from survey data	Expert judgement
ILVO	Leave the gap	NA	NA
LUKE	Multiple actions	Leave the gap	Impute values from other strata
NMFRI	Leave the gap	NA	NA
SLU(a)	Impute values from other strata	NA	NA
SLU(b)	Multiple actions	Impute values from other strata	Leave the gap
THN	Leave the gap	NA	NA
KU	Predicted average	NA	NA
WMR(a)	Leave the gap	NA	NA
WMR(b)	Leave the gap	NA	NA





Q5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it

Do you have written processes or guidelines which define your approach to imputation?

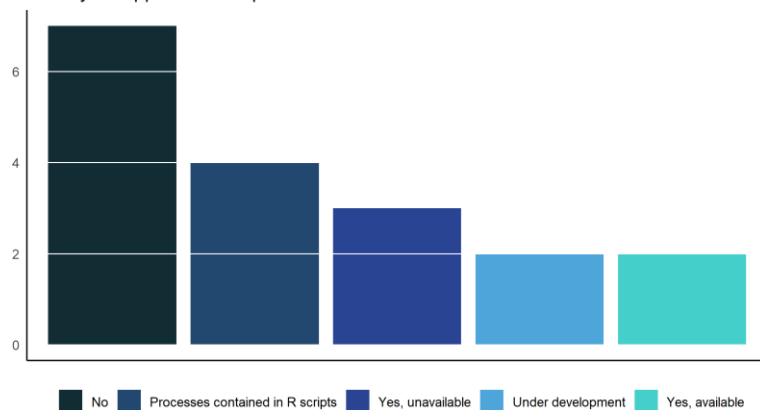


Figure 1: Frequency of categorised responses to Q5.3– Do you have any written guidelines for dealing with any gaps in your sampling strata?

Respondent were asked whether they had any written processes or guidelines which defined their approach to imputation. Seven respondents stated they did not have any such guidelines. Four respondents included written guidelines in the R script which they used for imputation (“*There are two R markdown documents for data submitters to follow, based on COST functions. These are updated annually. Training is also given to data submitters on these documents prior to data extraction.*”, FEAS-MI, “*Imputation is documented in scripts. Its most important steps are also documented as notes to stock coordinator in InterCatch format.*”, SLU(a)).

Table 1: Categorised responses for all respondents to Q5.3– Do you have any written guidelines for dealing with any gaps in your sampling strata?

Institute	q5.3
AZTI	Under development
BIOR	No
DRP-RAA	No
DTU(a)	Processes contained in R scripts
DTU(b)	Processes contained in R scripts
EMI	No
FEAS -MI	Processes contained in R scripts
IEO(a)	Yes, unavailable
IEO(b)	No
ILVO	Yes, available
LUKE	No
NMFRI	No
SLU(a)	Processes contained in R scripts
SLU(b)	No
THN	Under development
KU	Yes, available
WMR(a)	Yes, unavailable
WMR(b)	Yes, unavailable







Conclusion

Data checks

The primary objective of this questionnaire was to determine if, when and how European fisheries institutes performed data quality control checks, data editing and data imputation. The analysis presented above indicates that most respondents: constrained some values to be physically realistic (Q3.2), used predefined code lists (Q3.3), performed some form of outlier check (Q3.4), performed some form of spatial data check (Q3.7), performed some form of temporal consistency check (Q3.8), performed some form of duplication check (Q3.9). Checks were performed regularly as part of the data collection process were cross checks with census data (Q3.5) and missing values check (Q3.6). However, whilst most checks were performed, the point at which checks were performed varied greatly. The reason for performing check at different points in the process could be attributed to different data capture methods, different time frames for the importing data or different operating procedures in relation to data collection and checking. At a minimum, institutes should aim to ensure all checks have been performed prior to responding to data calls (at or prior to the point of data extraction). If checks are implemented at a different or additional stage (where checks are being implemented at multiple points), the point, method and type of checks implemented should be documented.

The method for some checks, such as outlier detection and cross checking of spatial data, are similar for many respondents. As many respondents already have a dedicated R script which produces plots which aid in the identification of outliers, it may be possible to produce an standardised R script dedicated to outlier checking and or spatial data plotting, which would be available to all members of the RCG (in turn standardising some/multiple checks discussed above). While variety in sampling schemes and data collection practices might limit the effectiveness of such a script, a standardised script containing protocols might prove useful in ensuring checks are in place and are of a common method.

Data editing

The consensus for approaches to dealing with errors, inconsistencies and discrepancies was to attempt to correct the sample data where possible, and to exclude the data from outputs where correction is not possible. If data cannot be corrected, institutes should at least aim to document the error prior to deletion. Such a record may help in preventing similar mistakes in future and highlight repeated errors so corrective action(s) can be taken. Such error logging is already in place by WMR(a,b) and KU. The template for logging errors proposed by WMR may be suitable for logging such errors ("*SampleID, Species, DateChecked, ErrorDescription, ActionsTaken (e.g. excluded, corrected) ,Reason, DateProcessed ,Re-imported (Yes/No), Who*", WMR(a,b)). If possible, institutes should also log errors even where correction was possible, again to prevent any future errors.

Data Imputation

For dealing with their approach to gaps in Age length keys (ALK's) or weight length keys (WLK's), institutes filled such gaps either by imputing from an average, imputing from a model, imputing from other strata, filling by expert judgement, or leaving the gap. As the course of action often depended on what data from other surveys, strata or sampling schemes was available, a definitive course of action to be taken in the event of an ALK/WLK gap is not appropriate. However, where gaps have been filled, institutes should document which data was imputed and what method was used. If a





predicted value from a model was used, details of the model should be recorded. If data is borrowed from other strata or from surveys, the details of the strata or survey should be recorded.

When asked about dealing with gaps in sampling strata, most respondents opted to leave the gap and allow the ICES stock coordinator to decide how to deal with the issue. As this is already a popular course of action, leaving the gaps in the sampling strata and allowing the ICES stock coordinator to deal with them should be the course of action employed by institutes to deal with gap in their sampling strata. Where institutes decide to impute from other strata or surveys, details of what values have been imputed and of the method of imputation should be documented, such that the ICES stock coordinator is aware data has been imputed. This should minimise the chances of already imputed data being imputed from, increasing data accuracy overall.

Written guidelines

When asked to list any written guidelines relevant to sections three, four and five, many institutes were not able to provide such guidelines, either because they did not have any or they were not publicly available. As institutes still performed many of these checks without such guidelines, they may be unnecessary, however having SOP's for data quality control recorded in a document would be a useful resource, both at a regional and international level. While such guidelines may contain information sensitive under GDPR, a censored or constrained document could still be appropriate.

Age - readings

While there was some reference to data quality control in relation to otolith readings (FEAS-MI, ILVO), most respondents did not discuss these practices in their answers. As a result, this report cannot recommend 'best practice' quality control with regards to otolith readings, as it is not supported by the data presented here.





Recommendations

Based on the analysis conducted in this report, the following recommendations are proposed for data quality control practices.

1. When data quality control checks (such as those discussed in section 3) are implemented, institutes should ensure that the type of check, timing of the check (both the point during the data collection process and the date), and a brief description of the check are documented.
2. Where checks are performed at multiple points during the data collection process, institutes should ensure that datasets / samples are marked such that users are aware what checks have been already performed or where data has been edited or imputed.
3. Where the method of check is broadly similar among institutes (e.g. Q3.4 - outlier detection, Q3.8 - spatial data checking etc), attempts should be made to produce a standardised SOP, ideally at a WG level, detailing the method used to perform the checks.
4. Where the method of check is broadly similar among institutes (e.g. Q3.4 - outlier detection, Q3.8 - spatial data checking etc), attempts should be made to produce an R script to conduct these checks which is available to all users.
5. Where errors, inconsistencies or discrepancies are found in the data, information about the cause of the error and course of action taken to rectify it should be recorded. Records will allow users to identify common sources of error in data collection process.
6. Where institutes are imputing data from a predicted average/model/survey or from other strata to fill gaps in ALK's or WLK's, institutes should clearly document *what* data has been imputed, *where* the data was imputed from and *when* the data was imputed. As imputation may be performed at multiple points or by different users, it is essential that all users, from local to working group level, are aware what data is 'real' data and what data has been predicted or imputed.
7. Where gaps are found in sampling strata, a standardised course of action should be decided on at WG level. Based on the analysis conducted in this report, the most suitable course of action is to leave the gaps and allow the ICES stock coordinator to decide on how best to deal with them.
8. Further research should be conducted to collect information on data checks, editing and imputation with regards to age-reading among institutes.





References

West, M., 2011. *Developing high quality data models*. Burlington, MA: Morgan Kaufmann, pp.4 - 6.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.





Appendices

AZTI – Fundacio AZTI (Spain)

Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

- 1.1. What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

- 2.1. Which country do you work in?

Spain

- 2.2. Which institute or laboratory do you work in?

AZTI

- 2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No

- 2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Data from our sampling schemes (at the market and on board) and official data corresponding to ICES areas





3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Sampling data is captured on paper and transcribed to the Data base in about 1 month time

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

We don't have these constrains in place, we check for outliers during the data quality process before answering data calls. Values checked for outliers/non realistic values are sampling data for length, biological parameters, sampled weight, spatial position, duration of the haul and dates

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes, our data base has its own code lists, based on national and international code list (Spain fleet register, ASFIS, WORMS, LOCODE, list of metiers accepted) . We update the list manually, every time we have data which is not included.

3.4. Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

Biological parameters, catch and sample weight.

- How do you define an outlier?

For length data we use the graphs developed in FishPi. For the rest of biological parameters the outlier is the observation out of the interquartile range.

For catch and sample weights we use boxplots and visual identification

- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

Graphically with boxplots and using expert judgement.

At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

During the data quality process before answering data calls

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

We do some cross checks with sampling and census data for the trawl fleet. The checks consist in comparing the landing weight observed by the sampler with the official weight. It is done during the data quality process before





answering data calls. If there is an inconsistency between the sample and census data, we double-check that it is not a mistake. If it is correct, we use sampling data to calculate the non-reported landings in InterCatch
We do some general cross checks with sampling and census data for all fleets, for species assignation. This allow us to correct species which are wrongly identified in the census data, and estimate the share of species which are landed together (monkfishes, megrims, etc)

- 3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

No

- 3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

On board sampling data: Area and rectangles are calculated automatically depending on the coordinates. They are also plotted in a map to detect clearly wrong positions.

- 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

No

- 3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

We have some duplication checks for sampling data. We do it during the data quality process before answering data calls

- 3.10. Please let us know about any other relevant data checks which have not already been described in your answers

We check census data for errors in species identification, for these species which are clearly wrong because they cannot be present in our waters. We check metier & area combination.

- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

We have an internal protocol detailing all the data check, but it is not ready to be shared. We are working on that.

4. Editing





- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

We check the sample data with the register in paper, and correct the sample data in the DB. If we cannot correct it, we usually remove the data or replace with average values/expert judgement.

- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

We do not have written processes or guidelines which define our approach to dealing with data errors, inconsistencies, or discrepancies.

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

ALK: We usually fill it in by hand in case there is a small gap between unsampled sizes with respect to the observed size range. However, in case the number of samples is very small, we try to complement the ALK with other sources: sometimes from the campaigns, other times from adjacent regions or adjacent periods.

Translated with www.DeepL.com/Translator (free version)

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

For the length distribution of landings, we leave the gaps or impute missing values from other strata, depending on the instructions given by the stock coordinator at the institute.

For the length distribution of discards, we impute missing values from other strata.

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

We have some internal documents describing this (txt files stored together with data call files). But they are not ready to be shared. We plan to compile all instructions in a single protocol.





Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

- 1.1. What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

- 2.1. Which country do you work in?

Latvia

- 2.2. Which institute or laboratory do you work in?

Institute of Food Safety, Animal Health and Environment "BIOR", Fish resources research department, Marine laboratory.

- 2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No, our institute has not any accreditations or certifications relevant to these questions.

- 2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Data from Baltic Sea demersal trawlers

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

- 3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

During the trip information about collected material on haul level is captured in paper format. After returning from the trip as soon as possible collected information is entered into electronic format (Excel flat databases). After that data are imported to the Access database. Later data is prepared in national database Biodata format and imported



to it. There is no specific deadline for importing data into Biodata information system, it depends on my time. Usually information from several trips is imported at the same time.

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

During the data entry length, weight and age data are checked and constrained for minimal and maximal values. Excel data validation tool is used. The data file contains predefined values that can be assigned to the following biological parameters: sex and maturity. At the top of the datasheet 10 rectangles are located. For each rectangle excel macro is assigned.



We are using a 6-scale maturity scale. Sex is defined as numbers, 1 is male and 2 is female. In the rectangles all combinations of sex and maturity are predefined. For example, "14" means male with maturity stage number 4. For the entering sex and maturity data, the cell for sex is selected. After clicking with a mouse on the rectangle "14", in the cell for sex information about gender is entered, in the cell for maturity information about maturity stage is entered, end then macro select next fish cell for sex.

Sex	Maturity
1	4
2	5

Additionally, during the data import or direct input into the national database Biodata, additional data quality is checked. Database programmers can give full information about restrictions for data input.

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

International 3-letter code (FAO code) list for fish species, international code lists such as ICES vocabularies.

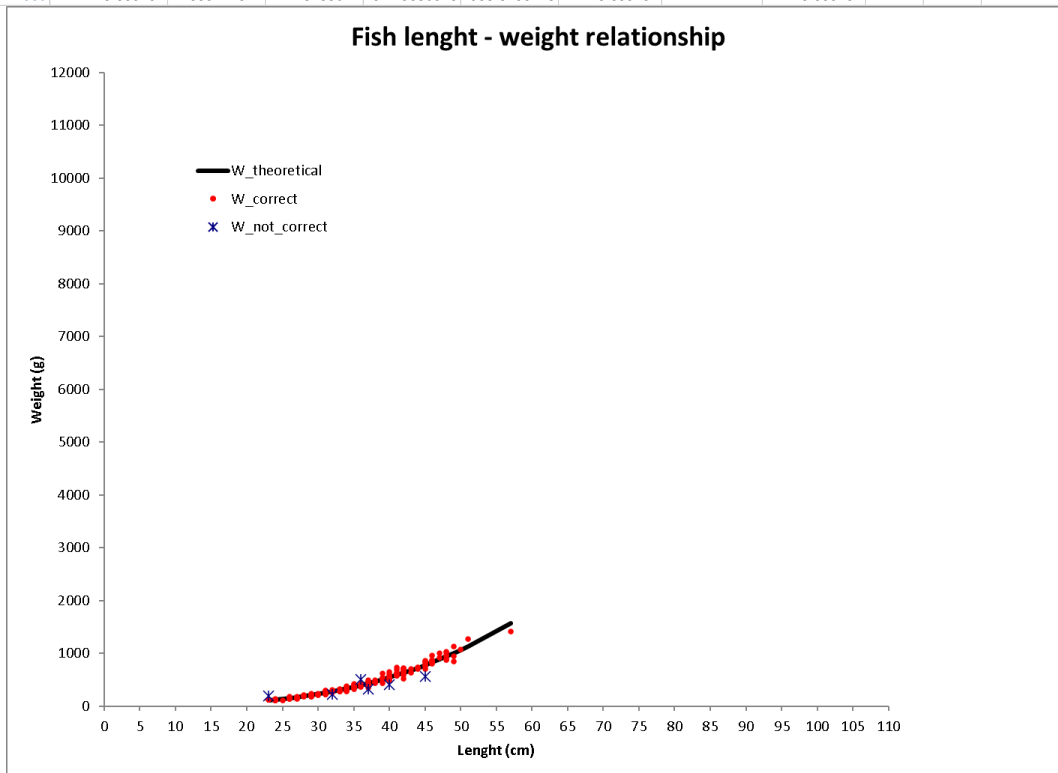
3.4. Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)
- How do you define an outlier?
- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)
- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).



Once the data for the respective trip has been entered, the length & weight relationship are analysed. A linear regression model is developed by determining the trend line and forecast confidence intervals. Those records that are outside confidence intervals are marked as erroneous and re-checked and if necessary, corrected. Excel macro is used for this checking. As a result, we obtain a graph with visual info and a table with problematic fish weights. In the table we receive information about haul number, fish number and problematic fish weight.

Zv al			Svars	W_theoretical	Log_L	Log_W	W_min	W_max	W_dif	W_dif_+	W_dif_-	W_1	W_2	W_correct	W_not_correct
2	43	23	190	105.7473809	1.361727836	2.278753601	84.76743346	131.9198673	84.25261907	84.25261907					190
2	13	32	220	282.0691381	1.505149978	2.342422681	226.1075091	351.8812753	-62.06913808		-62.06913808				220
2	85	36	505	400.2415232	1.556302501	2.703291378	320.834865	499.3013365	104.7584768	104.7584768					505
2	14	37	320	434.1831583	1.568201724	2.505149978	348.0425866	541.6435294	-114.1831583		-114.1831583				320
2	123	40	410	547.3446442	1.602059991	2.612783857	438.7531899	682.8124933	-137.3446442		-137.3446442				410
2	87	45	560	776.653751	1.653212514	2.748188027	622.5680918	968.8756248	-216.653751		-216.653751				560



In the last years Fulton's coefficient is used to check the length-weight relationship during the data entry. Excel conditional formatting option is used to check data quality. After entering the weight that does not match the settings (Fulton's coefficient is >2 or less than 0.5), cell is coloured in red and additional data checking is performed.

Length (L)	cm	Svars(g)
33	33	310
31	31	3000
34	34	320
32	32	310
34	34	300
32	32	300
33	33	330





- 3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

No special data checking with census data. Usually, data obtained from observed trip contains more detailed information. Observer task is to obtain real fishing information not to check census data quality.

- 3.6. Do you perform any missing values checks? (e.g. missing values vs. "true zeros"). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

If we find that something is missing in the collected biological data, together with the age determiner, a decision is made what to do with this information about the fish, leave or discard.

- 3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

No special data check for this type of information. ICES rectangles and Areas are entered with excel user-defined formulas, that uses coordinates. For data quality checking, the FishFrame database screening tool is used. If something is wrong, coordinates are checked.

Additionally, the national database Biodata calculate area, rectangle and national rectangle from the coordinates.

- 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Comparison of data by years, quarters is performed in the annual report of the institute. No such information checking during data capture.

- 3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

No specialized check for data duplications. A person who collects data makes data input also. Information about input data amount is checked with excel database formulas. This information is used for each survey report. During the cruise report preparation, information about collected material is checked again.

- 3.10. Please let us know about any other relevant data checks which have not already been described in your answers.





As I mentioned above, I am working in the sea alone. Biological data with the otoliths are collected and returned in special paper books. For each individual fish such information is collected, length, full weight, sex, maturity and otoliths. Otoliths are wrapped in page similar to an envelope. At this example is cod with length 47 cm, weight 1,03 kg, female with maturity stage 5.



After data input in Excel file, the age reader receives paper books with otoliths and file with the entered data. During the otolith preparation for age reading additional data quality check is performed, if necessary, corrections are made.

- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No. During the years and experience we try to make data checking better. No special document about it.

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

If data errors are found, original data and outputs are corrected, necessary data into databases are corrected or updated.

- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No.





5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

Any gaps in ALK and WLK are filed together with the age reader. If it is possible, averages between 2 existing data are used. If it is necessary, based on age reader experience gap is filled. The chance of gaps is very low. During the trip biological data are collected for the needs of ALK. For example, for the cod before direct cod fishery ban, otoliths during the trip were collected as a minimum 30 fishes from each 5-centimetre group for each subdivision.

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Data are used for calculations in strata from which data are collected. No data borrowing.

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

No.



**DRP/RAA - Regional Directorate for Fisheries in the Azores****Annotation:**

Data Collection Framework in the Azores has only recently been of the responsibility of the Regional Government of the Azores (since the end of 2018). The entire process of preparing databases, guidelines, and written procedures had to start anew, which, associated with the pandemic situation of recent years, means that these procedures are still far behind. It is, however, expected that by the end of 2021, this situation will be largely resolved. For that reason, there are no written processes, and the quote “Not applicable” was used in some answers.

Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

- 1.1. What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

- 2.1. Which country do you work in?

Portugal – Autonomous Region of the Azores (RAA).

- 2.2. Which institute or laboratory do you work in?

Regional Directorate for Fisheries in the Azores (DRP/RAA).

- 2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No.

- 2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

All relevant stocks and sampling schemes are monitored from commercial fisheries in ICES Division 10a2 (Azorean fleet).

3. Data checks



When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Biological data is registered on paper forms and, depending on the data source, transcribed on a quarter or monthly basis. Additionally, length-frequency data from the Fishmetrics system uses images captured from fish boxes to obtain length composition, which are available in a cloud.

Lotaçor, S.A. (the local state-owned company that provides the public service of organising the first sale of fish) electronically record landings (sales notes) daily.

Electronic fishing logbooks from fishing vessels of 12 meters' length overall or more are electronically transmitted on a daily basis. Vessels between 10 and 12 meters use traditional paper logbooks, which are monthly, entered into an electronic recording system. All vessels under 10 meters are exempted.

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes. Length/weight range, crossing dates, catch and sample weights. All checks are performed at data capture, during data validation, before and after electronic recording, and during extraction.

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes. Lists of valid local values are available for several data such as fleet ID and segmentation, fishing ports, fishing gears, métiers, catch fraction, species names, measured length. ICES vocabularies are used for data such as gear type, the unit of effort, and stock code.

3.4. Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)
- How do you define an outlier?
- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)
- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

Yes. Biological variables are checked for linear relationships between pairs of data (e.g. total vs. furcal length, total vs. dressed weight). Outlier analyses are checked graphically during data extraction and are defined as values outside the range of quartiles.

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

Yes. During data capture and extraction, at-market and at-sea sampling are cross-checked with sales notes and logbooks for trip duration or duplication, species misidentification, and landed weights. The inconsistencies usually





are related to misreporting (either on species identification or differences in catchweight determination – differences between landings and catch). These situations are corrected and reported in parallel: official data (from census) and “real” data (from samples).

- 3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Usually, no. When it occurs, only ad-hoc checks are performed.

- 3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

At the time of data extraction, the spatial distribution is visualized, and wrong coordinates are corrected (which usually occurs due to data entry errors - transposition error). Ad-hoc crossing of areas with the presence/absence of species is also carried out, but not systematically.

- 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

It can occur but not systematic.

- 3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes. At data extraction, a verification for duplication of samples is performed.

- 3.10. Please let us know about any other relevant data checks which have not already been described in your answers.

Not applicable.

- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

Documentation on Quality checks for data capture, processes, evaluation accuracy, and data processing evaluation will be available during 2021.

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

Data are replaced with correct values once validation applies.





- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No.

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

Not applicable.

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Usually, impute missing values from other strata.

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

No.

DTU(a,b) – Denmark technical University Aqua

Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1 .About you (answers will not be published)





1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1. Which country do you work in?

Denmark

2.2. Which institute or laboratory do you work in?

DTU Aqua

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No

2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

- Estimated amount of discard for different ICES assessment WG's
- Estimated age distribution of landings of commercial stocks for different ICES assessment WG's, where the sampling is stratified per commercial size categories

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Ad a-b) The only electronically device used in our commercial sampling is a calliper used for measuring the carapace length (mm) of Nephrops and shrimps. Everything else is captured on paper and entered in our national database as soon as possible

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc)."





Ad a-b) Database – entry - numeric: Partly. Some of the numeric fields in our national database has constrains, so only realistic values can be entered e.g. wind direction, but most of the numeric fields is only constrained by the length of the field in the database, which often is set unrealistically high e.g. mesh size is numeric(5,1). We have implemented a set-up, so it is possible to set realistic values for age, length and weight per species, but the set-up has only been used in a short period, since the technicians was tired of all the warnings.

Ad a-b) Database – entry - character: see point 3.3

Ad a-b) Elsewhere: No

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Ad a-b) All categorical information have defined code lists in our national database, except skipper contact details. All forms has a free text field for remarks. Nearly all of the codes lists are local, but the most relevant ones, species, area etc., have a field with International codes

3.4. Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)
 - Age and weight per length (individual measurements) (Routine)
 - Discards weights per haul and species compared to an estimated weight based on the length distribution of the sample (Routine)
 - Estimated mean weight per length / Age vs. total weight (SOP check on results). If extreme, this is often due to an outlier (Routine)

- How do you define an outlier?

5.3..1. Visual, extreme percentage. Never found a good approximation with standard deviation

- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

5.3..1. Expert judgement

- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

5.3..1. Data extraction or estimation

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

Ad a-b) Not as a routine. On a more ad-hoc basis, some technicians do it during data capture and samples are sometimes checked during estimation

3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).





Ad a) Yes, our national database has fields that indicate if everything in a hauls has been worked up. In cases of mismatch, then the observers are asked to check. The same is true, if it is indicated that both discards and landings have been work up, but no recording of discard is found. The checks are performed during data extraction.

Ad b) No

3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Ad a-b) Rectangles are constrained within areas in our national database

Ad a) Coordinates are checked ad-hoc doing data entry (mapping function in our national database) and as a routine with maps in R-markdown reports just after data entry, see example in attached pdf's. Ad-hoc checks during estimation if samples are causing troubles

Ad b) Some technicians do it ad-hoc and ad-hoc checks during estimation if samples are causing troubles

3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Ad a) Estimates of discarded amount are compared with the last 5 years as a routine.

Ad b) Not as a routine

3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Ad a-b) No

3.10. Please let us know about any other relevant data checks which have not already been described in your answers

Ad-a) Different relevant checks are done as a routine on the at-sea observer trips per trip and quarter, see attached pdf's

3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No and lot of the checks are included in the scripts we use for extraction or estimation

4. Editing

4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)





Overall, if inconsistencies are found then data are checked e.g. against original papers, re-reading of ages, census data. If no error is found, then the value is accepted. In rare case, when an outlier is spotted just before submission of data and it has a strong influence on the result, then the sample is left out and checked afterwards.

- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

ALK's: Impute by either average or model

WLK's: Impute by model

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Ad a) Strata match the ones in InterCatch, so the gaps are left blank

Ad b) Strata, commercial size categories, do not match the ones in InterCatch, so missing values are imputed from other strata

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

No, only rule of thumbs in the head of the estimator for the none-modelled imputations, but everything done in the past is documented in SAS or R scripts



**EMI - Estonian Marine Institute, University of Tartu****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1 About you (answers will not be published)

- 1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2 About your work-place

- 2.1 Which country do you work in? **Estonia**

- 2.2 Which institute or laboratory do you work in? **Estonian Marine Institute, University of Tartu**

- 2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No

- 2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Stock assessment-related data for Baltic herring (Central Baltic Herring and the Gulf of Riga herring stocks), and the Baltic sprat in Sd. 22-32

3 Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

- 3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

At first data is captured on paper and then transcribed to local database (usually monthly).

- 3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).





Checking is ad hoc- uploaded data is checked against the reasonable value ranges (relying on the expert experience).

3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

This depends on categorical information, e.g. areas, gear and métier are defined as in ICES vocabularies. Otherwise mostly free text.

3.4 Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

Biological parameters: mean weights at age, mean length at age, total length range, length-weight relationship.

- How do you define an outlier? **According to expert experience.**
- How do you check for outliers? (e.g. graphically using expert judgement, R scripts) **Graphically**
- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc). **Data capture**

3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

No cross checks of sample data with census data is made.

3.6 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **NA**

3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Yes, at data capture, Cross-consistency between rectangle and area codes are checked during data capture**

3.8 Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **No**

3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **No**





3.10 Please let us know about any other relevant data checks which have not already been described in your answers.

Since our data is uploaded to ICES RDB, the RDB data checking system performs many checks.

3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

Yes, https://www.envir.ee/sites/default/files/andmetootluse_juhend.pdf

4. Editing

4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

If errors are found, they are corrected or the respective data line is deleted if correction is not possible. All possible errors are corrected prior to uploading to InterCatch.

4.2 Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

https://www.envir.ee/sites/default/files/andmetootluse_juhend.pdf

5 Imputation

If you have different imputation processes for different end-users please make these clear in your answers

5.1 How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

NA (no ALKs used in herring/sprat sampling data)

5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

NA

5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

No written guidelines.





FEAS-MI – Fisheries Ecosystem Advisory Services, Marine Institute (Ireland)

Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.2 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1. Which country do you work in?

Ireland

2.2. Which institute or laboratory do you work in?

Marine Institute, Fisheries Advisory & Ecosystems Services

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation).

In February 2019 the Marine Institute received international **IODE accreditation** of its Data Management Quality Management Framework (DM-QMF) by the UNESCO International Oceanographic Commissions (IOC) International Oceanographic Data and Information Exchange programme (IODE).

The overall aim on the DM-QMF is to support continual improvement of the quality of the data, products and services delivered by the Marine Institute through assuring the quality of the processes and procedures used in the generation of the data and products.

2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

The questions are answered for the **Demersal Catch Sampling At-Sea programme**, which follows the flow of data collected during an at-sea sampling programme from collection to analysis to reporting.



The **Demersal Catch Sampling At-Sea programme** is comprised of demersal at-sea and *Nephrops* at-sea sampling. The *Nephrops* at-sea sampling has similar but slightly different protocols to the demersal at-sea. Landings data from at-sea sampling is uploaded to the Stockman database.

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Nephrops

The **Nemesys** Application is used at sea for electronically measuring *Nephrops* Commercial Samples at sea. Measurements (Fig. 1) and weights of individual *Nephrops* are captured electronically through the use of wireless callipers and Bluetooth weighing scales



Figure 1. Measuring Nephrops using a callipers

Data is recorded electronically at sea into a local database (stored on the individual tablet) before being uploaded by the Sampler into a central SQL Server Database when back on dry land. Between Marine Institute Surveys and Commercial Sampling an average of 100,000 *Nephrops* measurements have electronically been captured annually since 2002.

Demersal

Demersal data is captured on paper and then either (a) manually inputted to the database via a remote desktop or (b) the trip data is manually entered to Excel sheets on a remote desktop. An R script is then run by the database administrator to format the data and automatically upload to the database.

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

The **Nemesys** Application uses data validation while individual *Nephrops* are electronically weighed and measured. Validation checks include: unknown gear type not allowed, unknown vessels not allowed, measurements must be between 3.01mm to 99.99mm (this check is made while individual prawn are actually be measured at sea). Quantity Checks (QC) are also made while the weight measurements are undertaken.

The QC Weight Validation checks the Catch Males, Total Catch Females (all catch female weights are summed), Discards Males and Total Discards Females (all discards female weights are summed) while each *Nephrops* Sample is being electronically captured at sea. A formula is applied to the all lengths captured to each of the four



categories and the Functional Unit to calculate the predicated weight. The calculated predicated weight is then compared and checked to ensure its within 20% of the actual weights that were entered. The QC Weights parameters are uploaded as part of the Sample Metadata.

Our Commercial Port Sampling Application (Stockman) contains data validation ensuring required fields have been entered i.e. Sampling Place, Landing Port Sampler, Recorder, Inputted By and Sampling Event Date, Vessel, ICES Division, Gear, Species, Sample Quantity, Total Quantity and Fish Lengths must be 0 and 999. Length/Frequency, and Length plots are generated as data is being entered.

Demersal biological data (length, weight, sex if relevant) is not constrained as outliers may be valid samples. Trip and catch data for at sea sampling programme are not constrained during data entry but are reviewed during the quality control procedures to ensure that information is realistic (see question 3.4 below for details).

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Dropdowns (based upon managed reference tables) are used throughout data entry within **Nemesys** and **Stockman**: **Nemesys** uses dropdowns for the following: Sample Type, Vessel, Date (calendar control, Landing Port, Sampled By, Functional Unit, Fishing Grounds (linked to the Fishing Grounds Dropdown) , Gear and category

Stockman uses dropdowns to record: Sampling Place, Landing Port Sampler, Recorder, Inputted By and Sampling Event Date, Vessel, ICES Division, Gear, Species.

Demersal database for at sea sampling programme uses dropdowns to record the following.

'Cruise' information: year, port code, cruise code, survey type, departure date, completion date, boat name, departure port, discharge port, personnel, input by, validated by, reported by and discard sample collected.

'Haul' information: gear, success code, ICES Division, fishing ground, wind direction, wind force, sea state, swell direction, sea swell and ground type.

'Sample' information: Presentation (round/gutted), sample type (catch/discards/landings), grade (small/medium/large/ungraded) and whether measurements were taken as a sample or all fish were measured. It should be noted that the size of fish in discard sample is cross referenced against the relevant minimum landing size during data entry.

3.4. Do you perform any outlier checks on your data? If yes, please explain:

3.4.1. Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

Database consistency: Duplicate trips; Trips without hauls; Hauls with missing trips; Duplicate hauls; Hauls with success code 1 (successful haul with samples) or 5 (non-random samples) and no samples; Hauls with success code 1 or 5 and no catch or landings; Samples with missing hauls; Landings with missing hauls; Duplicate samples; Sample headers without samples; Species that do not exist in the species table; Missing success code (foul haul or valid haul); Measured landings that do not exist in the bulk catch table.





Raising factors: Unexpected sample weights; High raising factors; Missing raising factors; Negative discards (discard weight larger than total catch weight); Sample weight larger than total discards; High proportion of discards; Low catch rate or landings rate; High catch rate or landings rate; Weight of measured discard fish larger than sample weight; Unexpected proportions of non-fish discards.

Tow data: Excessive tow length or fishing speed; Zero tow length; Impossible or unexpected shoot or haul positions; Short tow duration; Negative tow duration; Missing tow duration; Long tow duration; Tow shot before previous tow was hauled; Tow year does not match year in cruise code; Tow dates outside cruise dates.

Length data: Any fish that are larger than the 99th percentile * 1.5 or smaller than 1st percentile * 0.5, are identified as outliers.

3.4.2. How do you define an outlier?

Outliers are defined by comparison to historical data. Points that fall outside 95% of historical data points are considered to be outliers.

3.4.3. How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

The following plots are generated using R scripts as part of a quality control report and are reviewed for presence of outliers (see Fig. 2):

- 1) Sample weight is plotted against raising factor
- 2) Proportion of discards is plotted against raising factor
- 3) Landings rate (kg/h) is plotted against catch rate (kg/h)
- 4) Haul duration (h) is plotted against calculated fishing speed (nm/h)



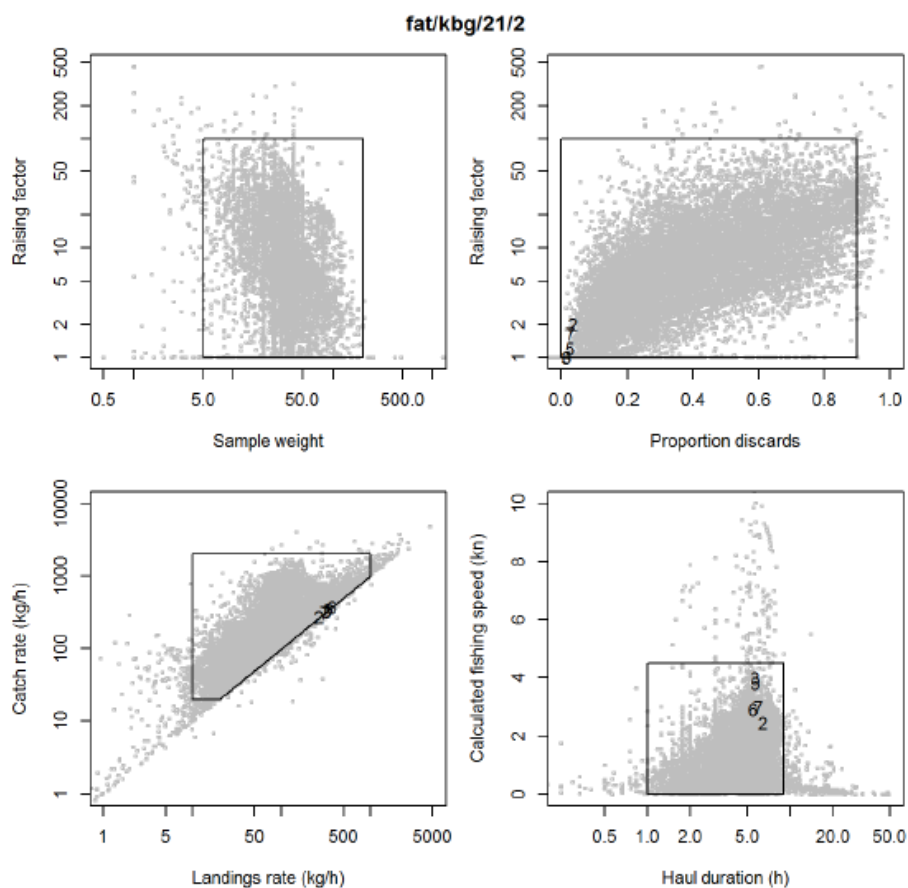


Figure 2. (a) Plot of sample weight against raising factor, (b) Plot of proportion of discards against raising factor (c) Plot of landings rate (kg/h) against catch rate and (d) Plot of haul duration h, against calculated fishing speed nm/h

3.4.4. At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

Quality control checks are primarily performed after the data has been entered into database, although basic quality control also occurs during data entry. The quality control report is generated using R scripts for each trip and reviewed accordingly. Original datasheets are digitally scanned as PDF files and are available for reference if needed during quality control process.

Further quality control checks are performed during data extraction on length to weight relationships (Fig. 3), age length keys (Fig. 4) and length frequency distributions (Fig. 5). During each of these quality control steps individual samples can be selected and checked for errors. Delta plots are also produced (expected weight of each sample divided by its actual weight) and outliers on the delta plots are defined as length distributions that are different from the expected distribution, although this does not mean that they are necessarily incorrect (Fig. 6).



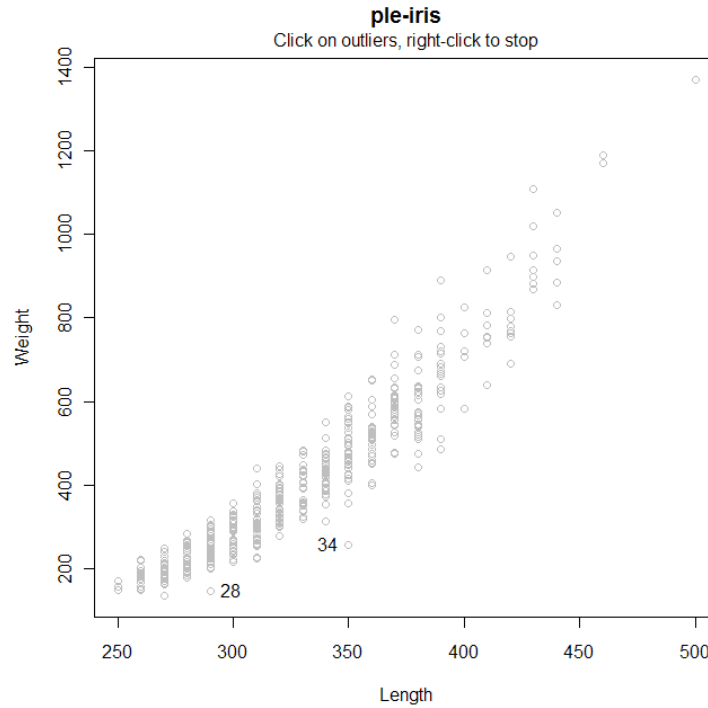


Figure 3. Example of a quality control check examining a length-weight relationship

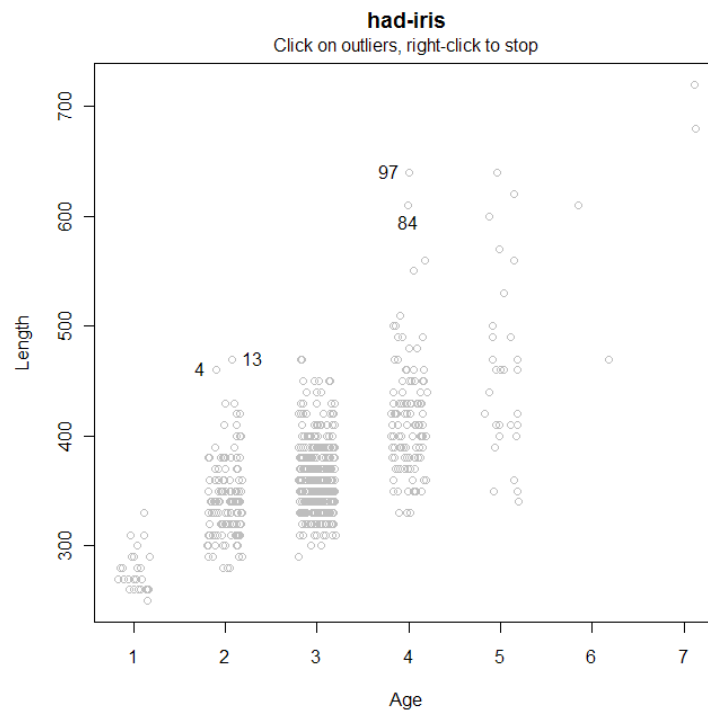


Figure 4. Example of a quality control check showing an age-length relationship

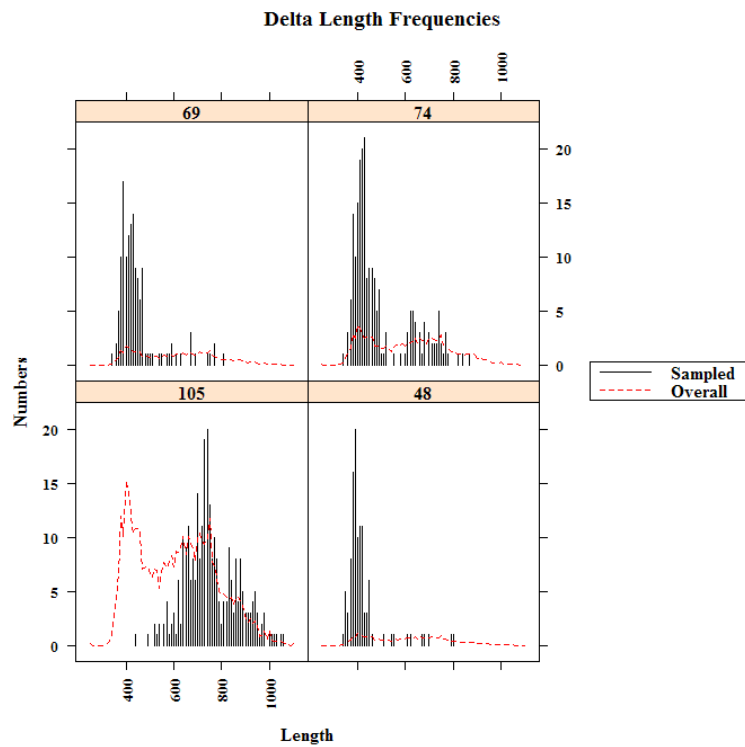


Figure 5. Example of a length frequency distribution from landings data

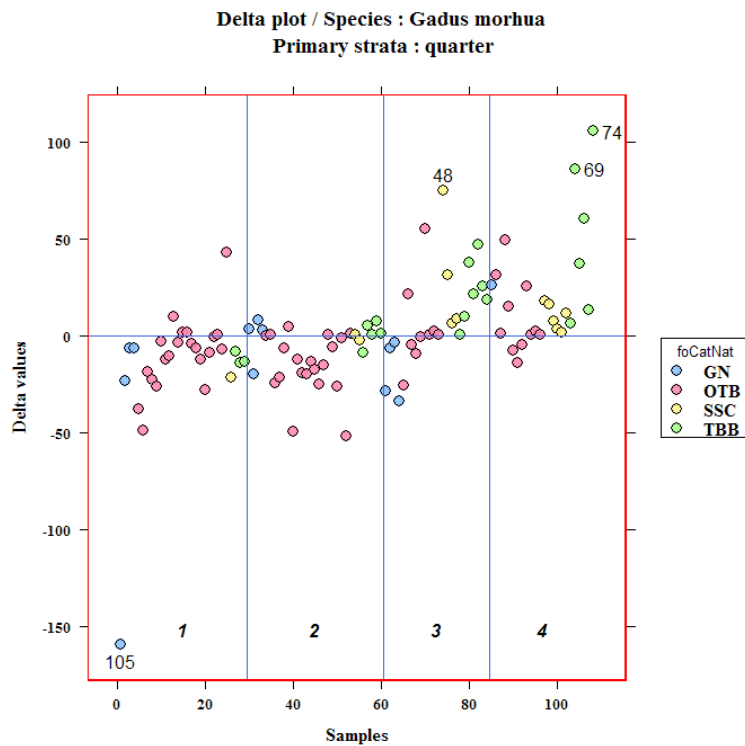


Figure 6. Example of a delta plot from landings data

After data extraction the contribution of each sample to overall landings and discards estimates are plotted for each metier so that any over-representative trips or hauls can be identified and checked for errors (Fig. 7 & 8).

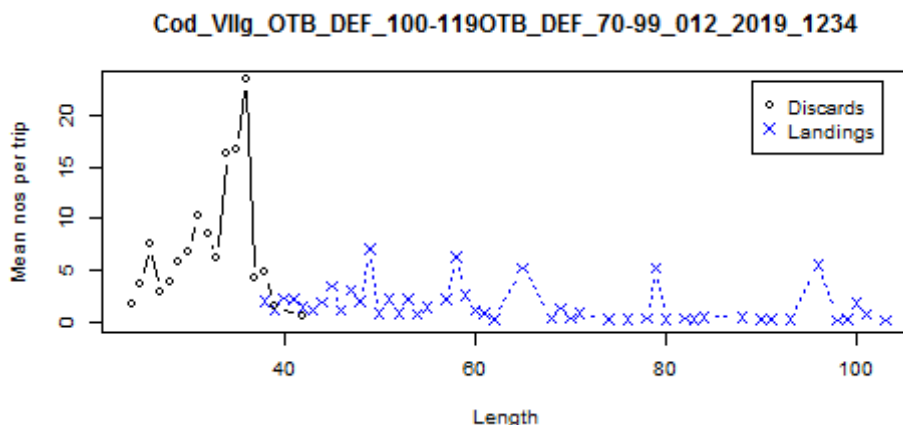


Figure 7. An example of at-sea sampling data, showing discard and landings mean numbers at length for a metier

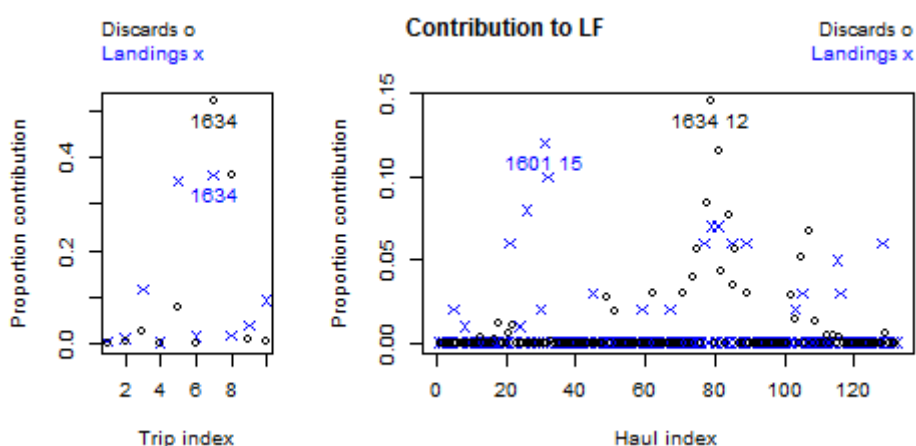


Figure 8. Examples of at-sea sampling data, showing proportions of discards and landings per trip and per haul for the metier in Fig.7

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

During data extraction the sampling levels are checked against commercial landings using temporal (quarter), technical (gear type) and spatial (ices sub-division) variables to check if there are sufficient samples for each sampling stratum (Fig. 9). If there are insufficient samples then strata may have to be merged before data consolidation. Commercial species composition and logbook discards are not used in data extraction process.

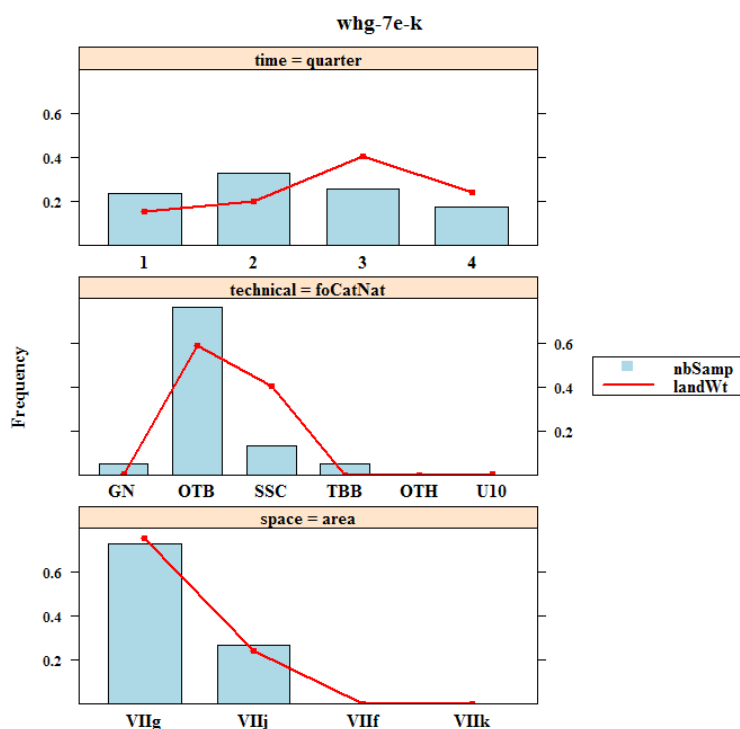


Figure 9. An example of sampling levels checked against commercial landings per temporal, technical and spatial variables

3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

At present no spatial modelling of sample data is performed in order to fill in missing values during data extraction. However, this is something that is being considered by the Marine Institute in order to maximise the statistical utility of the sampling program.

In some cases, missing data is treated as a true zero (e.g. a species that is not present in a simple of discards) in other cases it is just missing data (e.g. a species was landed but not sampled)

3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

The quality control report checks for excessive tow lengths or fishing speeds, zero tow lengths and impossible or unexpected shoot or haul positions. These are corrected either visually by plotting positions on a map (Fig. 10) or by reference to original data sheets.

During data extraction the cumulative length frequency distributions for each stock are compared across ICES sub divisions to check if merging of spatial strata is sensible. Sampling levels are checked against commercial landings by ICES sub divisions to ensure that there are sufficient samples in each spatial stratum (see 3.5 above).



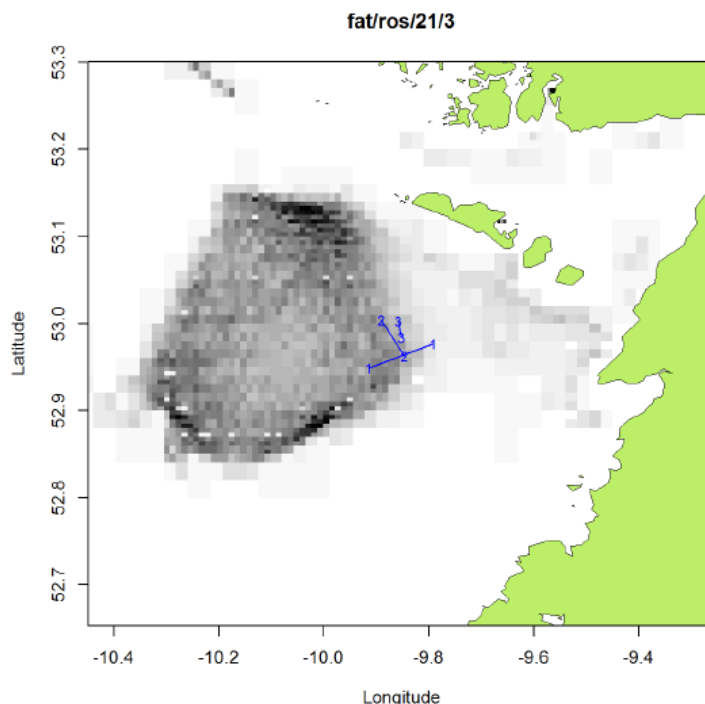


Figure 10. An example of spatial data checks

3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Cumulative length frequency distributions for each stock metier are compared across quarters to check if merging of temporal strata is sensible. During data extraction sampling levels are checked against commercial landings by quarter to ensure that there are sufficient samples in each temporal stratum (see 3.5 above).

Due to the disruption caused to sampling activities by the Covid-19 pandemic, the sampling levels for 2020 were compared to previous years to identify stocks that had reduced information available.

3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Reference lists are stored within the databases through controlled tables. Duplicate checks and updates are made by a Database Administrator when new entries are added to ensure duplicates are not introduced into the reference tables.

The Samples Number are generated by the co-ordinators of the sampler programs prior to data entry.

Both Stockman / Commercial Demersal Discards Database use a unique cruise code.

The Cruise Code is automatically generated within Stockman and is unique.

The Discards Database contains duplication checks on the Cruise Code field and manual entry of the Cruise Code is required.

Both Nemesys and Stockman produce length/frequency plots as data is entered, which aid in highlighting duplicates within data entry.

3.10. Please let us know about any other relevant data checks which have not already been described in your answers

F:\Logbooks_Current_report – for some checks on the logbook data that is used to raise the sample data to the population level

Length/Frequency plots are generated during data entry. This plot updates automatically within Nemesys as commercial data is electronically captured at sea.

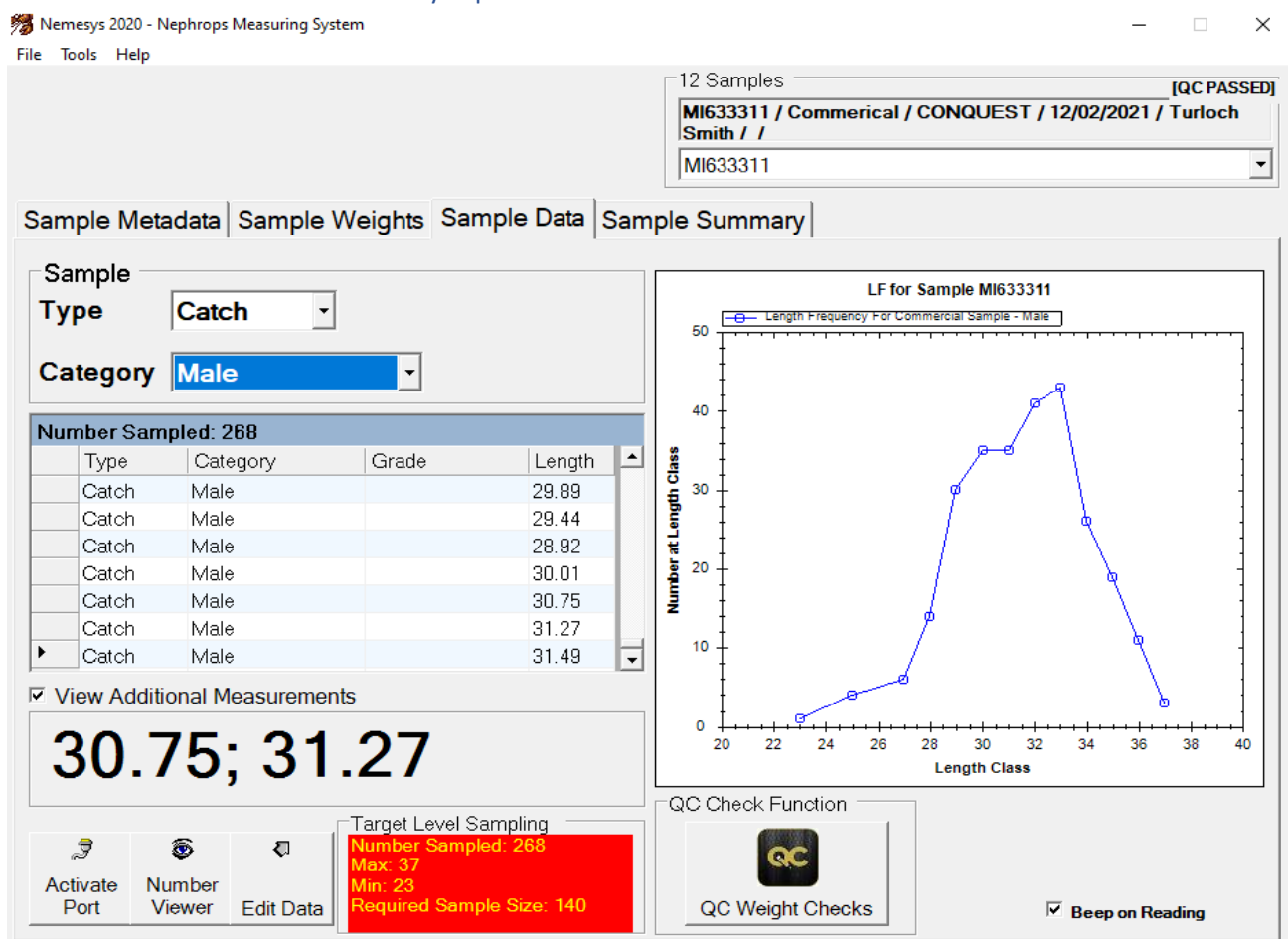


Figure 11. An example of one of the sections in the Nephrops Measuring System (Nemesys)

Data Validation Reports and similar length frequency/plots have been added into our commercial port sampling data entry application (Stockman)

QC Weights added into Nemesys -described above

Voice Report Validation tool for validating entered commercial discards data. Data is entered through paper sheets into our Commercial Discards Database, and the entered is validated through a Voice Reporting Application.

3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.



This answers to this questionnaire have generally been taken from the Marine Institute's "catch sampling quality control report" and the "COST data extraction for ICES WG" document. Both of these documents are produced using R scripts incorporating SQL database queries. These documents may contain sensitive data under GDPR although censored versions can be provided.

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

During the initial quality control review data error, inconsistencies or discrepancies are ideally resolved by referring to original data sheets (input errors), data entry personnel or communication with the sampler. If these issues cannot be resolved the success code for the problematic haul can be changed to either unsuccessful or unsampled. During data extraction age, length and weight samples can be edited or removed from the analysis.

- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

The Marine Institute's "Quality Control Report" was produced in house by Hans Gerritsen in 2013. Unfortunately these reports contain sensitive information regarding vessel names and fishing locations and cannot be shared. However, R Scripts and SQL queries used to generate reports can be provided on request.

Data extraction quality control checks were based on COST (Common Open Source Tools) methodologies adapted by Hans Gerritsen. Documentation and software packages are available at <https://wwz.ifremer.fr/cost/>.

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)





For age data the ALK are merged across technical strata but there still might be gaps. To make things efficient, an assumption that the differences in the ALK between areas are minor enough to be ignored, so age data from all areas are combined into one but the quarterly stratification is kept.

An R function is used to highlight the gaps in the ALK and another R function fills those gaps.

The function to fill the gaps is not used for length-only stocks - instead the fudged age data can be replicated across quarters (or areas)

5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

For LW data, the aim is to end up with sufficient samples in each (time/space/technical) stratum. For strata with minor landings it may be acceptable to have only one or two samples, for major strata, the aim is for at least 5 samples. If there are zero samples then the landings tonnage only is submitted and the ICES stock coordinator will have to deal with the gap in the sampling data. If there is a major stratum that has insufficient samples then the sample data can either be deleted for that stratum or it can be submitted with a warning. It is preferable to let the ICES stock coordinator deal with gaps.

For species that are reported by length and for which there is no biological sampling (i.e. weights-at-length) the length-weight parameters will need to be supplied to estimate the sample weights. Dummy data is created for the cs@ca slot so that almost exactly the same procedure as the ALK species can be followed. Therefore, instead of age data, the ages in the cs@ca slot are populated with lengths, and an Age-Length Key then becomes a Length-Length key, which is a convoluted way of raising the data has the functionality of merging strata etc.

5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

There are two R markdown documents for data submitters to follow, based on COST functions. These are updated annually. Training is also given to data submitters on these documents prior to data extraction.

Cost Data Extraction.Rmd

Discard Data Extraction.Rmd



**IEO(a) - Instituto Español de Oceanografía (Spain)****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1 . About you (answers will not be published)

1.2 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1. Which country do you work in?

Spain.

2.2. Which institute or laboratory do you work in?

Instituto Español de Oceanografía (IEO).

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No.

2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Data from our length sampling programme, both market and on-board, in the ICES area under the DCF/EUMAP. Tuna fisheries excluded.

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative





3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

For the IEO, there is a subcontracted company who engage the samplers for the on-shore and at-sea sampling. They are also in charge for the electronic recording of the sampling information. IEO assumes the responsibility of the correct dumping into IEO's database (SIRENO) of the sampling information typed by the contracted company in a way that ensures the information saved complies with the criteria, format and protocol agreed by the IEO:

- Verification of the periodic dumping of samples in SIRENO.
- Standardization of the information placed under the criteria established by the sampling team.
- Errors correction.
- Review and maintenance of the integrity of the stored data.

In addition, during the process, SAP proceeds to identify recurring errors during typing. These errors are communicated to the outsourced company for the correct typing in the following months. The process allows the analysis of the problems in the reception of the data and its communication to the subcontractor company to avoid the repetition of the same problems.

The development of the process has led to the development of a framework to guarantee the quality of the information stored. The original initial controls (more focused on the format, accommodation in terms of master tables, etc.), are being expanded with elements such as the detection of outliers (e.g., landings weights), or taxonomic control (e.g., improbable identifications).

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

There is a wide and continuously increasing number of checkings that can be made to the fisheries data to ensure data stored and, therefore, any subsequent use, have a proper quality.

Checkings have been separated into 7 different groups, referring to the main purpose of the check.

This work was originally based in the GFCM Data Collection Reference Framework (GFCM, 2017), and was adapted to our own needs and experience. From these 7 groups, there is one related to data physically realistic where we check the value of variables must be reasonable and probable, example of these checks are the detection of doubtful species for the area/gear and outliers in length range of species. They are all checked in the phase of dumping into IEO's database.

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes, one of the referred 7 groups of checks is related to the conformity with masters. IEO database limit the values of most variables to existing masters in the database. Examples of these checks are vessel code in official Spanish census (CFPO), port, gear, metier, species, etc.

3.4. Do you perform any outlier checks on your data? If yes, please explain:





5.3.1. Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

Biological variables (maturity, etc) are carried out by other team.
We do check length distributions, landings, etc.

5.3.2. How do you define an outlier?

Depends on the variable. For length and landings we use Cook distance to detect outliers. A process is then followed to clarify if the value is considered a mistake or an outlier.

5.3.3. How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

The process is carried out in several phases that involve different checks and treatments of the information received. Basically the process consists in 4 steps carried out immediately after the information is taken and some other analysis carried out annually before the sampling year is closed to start the processes to give estimates to WGs:

1) Pre-dump revision

Once the data of the contracted company has been received, an R script is used to homogenize the information with respect to the existing one in SIRENO (IEO's fisheries data base) and to detect errors.

In addition, the files are exported to the appropriate format required for direct dumping into SIRENO. This process is carried out by the data integrity manager.

2) Dump

It consists of pouring information from the previous phase into SIRENO. It is carried out by SIRENO's computer service (mainly because some information can only be incorporated properly within the database, i.e. SOP weights).

3) Post-dump revision

From the SIRENO output reports, an R script is used to detect errors and warnings.

- Errors : there is an objective mistake and must be fixed.
- Warnings : there could be a mistake or an evidence of a possible error.

4) Post-dump corrections

Errors and warnings from the "Post-dump revision" must be reviewed by the supervisor of the geographical area, as it can't be done directly by the data integrity manager nor automatically. It usually entangles revision of originally paper sampling sheets and/or communication with the sampler.

Annual analysis before closing the sampling year

A set of analysis done after all information is uploaded and checked.

Refers to statistical and graphical analysis for the length distribution and landings by metier and species.

Distribution, mode, temporal variation, outlier detection based on Cook distance, etc.

5.3.4. At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).





Please, see above.

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

1) Pairing/crosschecking process trips

IEO receives official fisheries data (logbooks, sales notes and the operational fishing fleet census) during the first months of the following year from the Ministry. IEO is in charge of the hierarchical classification into metiers (Decision of the Commission of November 6, 2008), which constitute the fundamental units of aggregation of fishery information.

SAP team builds the NVDP (metierized database of official data) that our team uses to review the design of the sampling plan, monitor fisheries dynamics, process sampling information, etc.

The pairing/crosschecking process between the sampled trips and the official data consists in crossing both sources through an R script in order to assign to each sampled trip the corresponding fishing trip of the NVDP.

Pairing is done for both, the on-shore sampled trips (RIM) and at-sea sampled trips (OAB); including contacted OAB trips which were rejected.

This procedure is set after the annual sampling review (see "QAF EMSAP Data Integrity, Data quality" deliverable) with the aim of:

- Assign the ID logbook of the NVDP to the sampled trips.
- Contrast and consolidate the information of the sampled trips.
- For RIM trips:
 - Record catches profile and catches by species.
 - Confirm the fleet activity.
 - Georeference the sampled trip: fishing Division and ICES rectangle.
 - Obtain specific trip variables not collected by the samplers (e.g. fishing days, deep, etc).
 - Obtain the sale location (sales notes cross checking), the location where landings are accessible for sampling (e.g. relevant to assess the coverage done to Spanish fleet landing abroad)
- For OAB trips, same information is collected. In this case, since most part of variables are collected on-board by the observers, information from logbooks are specially relevant for those trips where observer couldn't get on-board (refusals, etc).

Pairing methodology

This process is done through R after preparation of both datasets. Match is mainly done based in:

- NVDP: Landing Date/Fishing date and vessel ID code.
- Sampling: Sampling Date and vessel ID code.

Matches are reviewed based on supervisors knowledge of specific port or fleets dynamics.





2) Comparative analysis of LPUEs

Comparative analysis of DPUEs between Logbooks (DP), sale notes (NV) and sampling data allows the identification of problems in:

- Selection of trips.
- Metiers coverage.
- Concurrence of sampling (e.g. not access to certain species)
- Accessibility problems in ports to certain species, categories or part of the catches.

The landings comparison report is a routine task within the team to perform after the annual closure of samplings. The results will be examined at the sampling team meeting so that the team can investigate the problems of representativeness and concurrence in sampling.

No changes of the observed (sampled) is done. No changes in the official registration (e.g. logbooks) can be done either.

3.6. Do you perform any missing values checks? (e.g. missing values vs. "true zeros"). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes. Most part of variables in the data base are checked to avoid missing values during the pre-dump and post-dump phases already explained.

3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Geographical sampling information are checked with logbook data to verify the ICES Division (for market sampling) and the ICES rectangle (for on board sampling).

3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Graphical output for landings by month and species.

3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes. Checks to detect duplicated trips in different port, with different gear or different metier; and checks to detect duplicated categories and duplicated sexes in the same category are applied.





- 3.10. Please let us know about any other relevant data checks which have not already been described in your answers

See below.

- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

<http://www.proyectosap.es/index.php/documentacion-publica/category/323-quality-assurance-framework>

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

Yes, checking for discrepancies between data in records dumped to the database/s and data in the original records registered by the samplers/observers on the market and on-board.

The discrepancy checking is the verification that the information in the database corresponds to the information collected in the sampling statements.

Part of this process is taken during the quality checks review where errors and warnings identified by the algorithm ("Post-dump revision") must be reviewed by the supervisors, usually implying review of the original sampling sheets

- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it

As explained before ("Post-dump revision") the algorithm distinguish errors and warnings for the

- Errors : there is an objective mistake and must be fixed.
- Warnings : there could be a mistake or an evidence of a possible error.

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers



Regional Coordination Group
Baltic



Regional Coordination Group
on Economic Issues

100



Regional Coordination Group
Large Pelagics



Regional Coordination Group
North Atlantic
North Sea & Eastern Arctic



- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

Our team doesn't deal with Age Length Key data, please see the document provided by the IEO team working in biological data.

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Since the implementation of InterCatch (IC), we do not apply imputations, as it can be done by the stock coordinator after the integration of all international data. However, it is must continued to be applied for the transmission of mixed species data (species of the same Family with joint TAC). To do this, we apply the ratio of species of the same metier-quarter from the previous year.

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

The imputation of the mixed species ratio (explained in the previous point) was detailed to ICES when requested.

IEO(B) - INSTITUTO ESPAÑOL DE OCEANOGRAFÍA ,Centro Nacional (Spain)

Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

- 1.1. What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION





2. About your work-place

2.1. Which country do you work in?

Spain

2.2. Which institute or laboratory do you work in?

Centro Nacional INSTITUTO ESPAÑOL DE OCEANOGRAFÍA (IEO, CSIC)

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No.





2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

The biological variables data (Fisheries independent data) on the stocks for the ICES Area are carried out according to 2 differentiated sampling designs, depending on the biological characteristics of each species:

- **Small pelagic species:** the sample/subsample is selected by a Simple Random Sampling (SRS). The sample is entirely biologically analyzed (various biological variables are collected on each sampled fish until the expected number of samples is reached).

- *Engraulis encrasicolus* (ane.27.8)
- *Micromesistius poutassou* (whb.27.1-91214)
- *Sardina pilchardus* (pil.27.8c9a)
- *Scomber scombrus* (mac.27.nea)
- *Scomber colias* 8, 9
- *Trachurus trachurus* (hom.27.2a4a5b6a7a-ce-k8)
- *Trachurus trachurus* (hom.27.9a)
- *Engraulis encrasicolus* (ane.27.9a)

- *Sardina pilchardus* (9as)
- *Scomber scombrus* (9as)

- **Demersal and benthic species:** the sample is stratified by length classes. A Simple Random Sampling (SRS) is applied for the selection of the samples in each length stratum. A fixed number of specimens from each length class is





biologically sampled and various biological variables are collected on each individual. The sample attempts to represent the full length range of the catch, so the least abundant length classes are preferably selected for sampling.

- *Lepidorhombus boscii* (ldb.27.8c9a)
- *Lepidorhombus whiffiagonisboscii* (meg.27.7b-k8abd)
- *Lepidorhombus whiffiagonisboscii* (meg.27.8c9a)
- *Lophius budegassa* (ank.27.78abd)
- *Lophius budegassa* (ank.27.8c9a)
- *Lophius piscatorius* (mon.27.78abd)
- *Lophius piscatorius* (mon.27.8c9a)
- *Conger conger* (all areas)
- *Helicolenus dactylopterus* (all areas)
- *Merluccius merluccius* (hke.27.3a46-8abd)
- *Merluccius merluccius* (hke.27.8c9a)
- *Molva molva* all areas (lin.27.3a4a6-91214)
- *Phycis blennoides* all areas (gfb.27.nea)
- *Trisopterus spp* all areas (*T. luscus*)

The samples of the following species usually come from surveys although could be occasionally sampled from commercial landings:

- *Zeus faber* all areas





- *Mullus surmuletus* all areas
- *Loligo vulgaris* 8c, 9a
- *Pagellus bogaraveo* (sbr.27.9)
- *Parapenaeus longirostris* 9a
- *Sepia officinalis* all areas

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

- 3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

For most of the stocks, data from samplings are captured on paper and transcribed to the IEO SIRENO database as soon as possible.

Anchovy data from the Gulf of Cádiz (9a_S) is captured electronically with a tailored software/hardware system (icrOS) and data are subsequently uploaded to the IEO SIRENO database. This is going to be extended to *S. pilchardus* and *S. colias* from the Gulf of Cádiz in the near future.

- 3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

No for most of the stocks, however data are checked just after data extraction. We check the range of all parameters, lengths-size relationships and codes.





For *E. encrasicolus* (and the other small pelagic fishes) from 9a-s, when using the icrOS system, the numerical values are not constrained (now under development), however, with weight data some sort of constraining is performed providing the data is electronically captured from the scales. Regarding to categorical information, the system constrains the data input to the possible values of the key. No values outside the key can be input. However, wrong values can be input by the user and there is no check for that.

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes, we use a local code list, defined for the SIRENO database.

3.4. Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

Yes. Analysis and detection of outliers for biological parameters, their weight–length relationships and ranges.

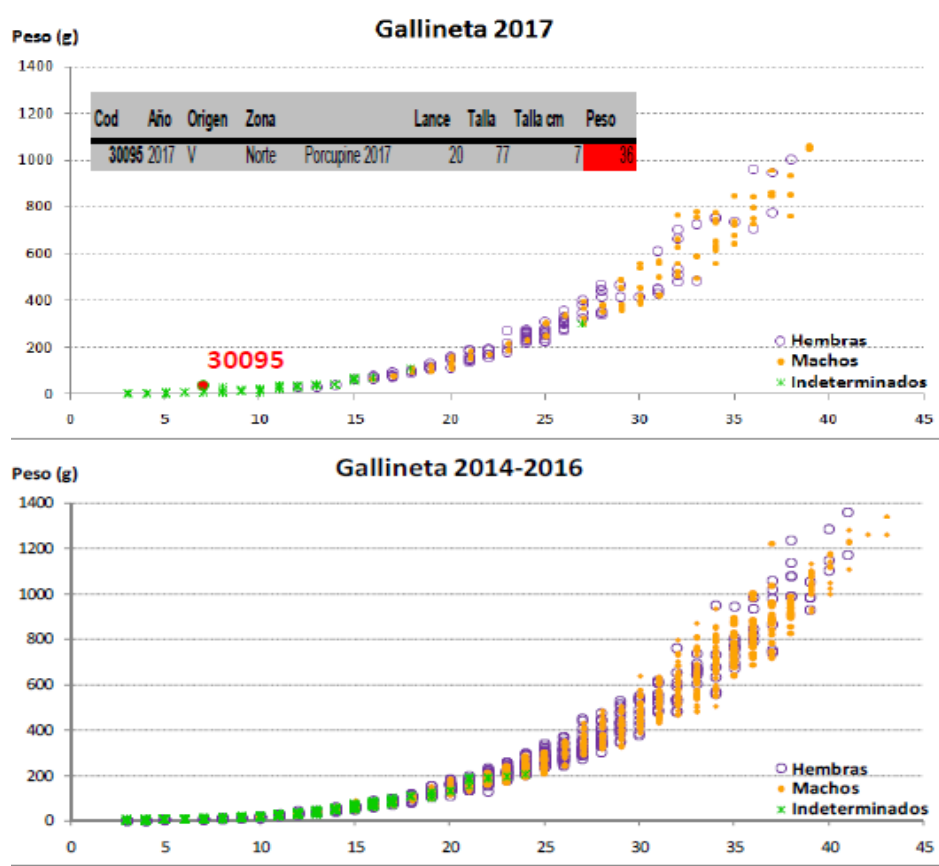
- How do you define an outlier?

Value far apart from other values or values that are frequently the result of an error (writing, measurement, etc)

- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

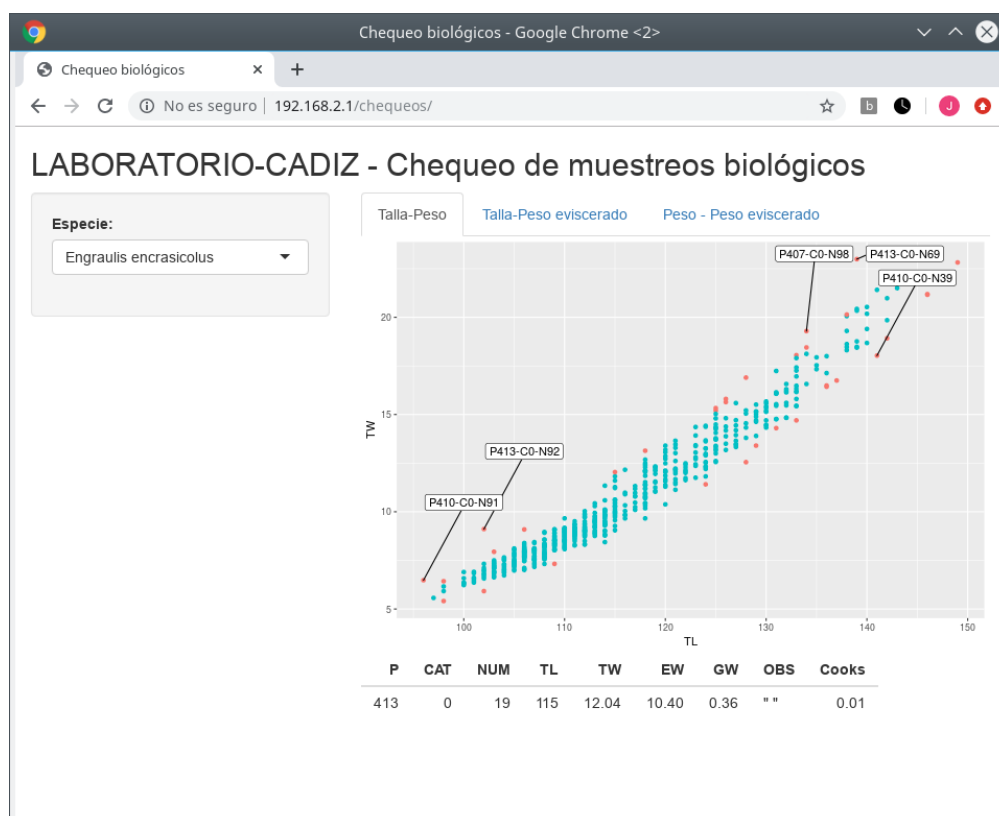
Graphically using expert judgment, creating common graphs such as scatter plots, histograms, box plots in R with (ggplot2 package).





For *E. encrasicolus* (and the other small pelagic fishes) from 9a-s, a Shiny application is used after sampling is complete.





- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

Checks are usually carried out at the end of the sampling and also by analyzing certain relationships between parameters

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

No.

3.6. Do you perform any missing values checks? (e.g. missing values vs. "true zeros"). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Checks are performed during data extraction. Our data recording system (SIRENO) doesn't allow the introduction of missing values/zeros for length variable.

In the case of using IcrOS system (small pelagic species from 9a-S Gulf of Cadiz), the IcrOS system doesn't allow the occurrence of missing values/zeros in selected variables like total length or total weight or any other mandatory variable. For other numerical variables, no check is performed.

Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).





No verification of spatial data is performed

- 3.7. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Temporal consistency data checks (quarters or years) are usually carried out as part of other studies, not as part of the sampling process itself.

- 3.8. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

SIRENO database or icrOS system doesn't allow the introduction of duplicates data

- 3.9. Please let us know about any other relevant data checks which have not already been described in your answers

- 3.10. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No, there is no written process for data checking

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

Depending on the error it could be tackled correcting the sample data (like some typing errors), while others are excluded from output/calculations or marked as outliers/errors.





- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

For small pelagic stocks, age length key (ALK) of the commercial sampling is completed with the age-length survey data and the missing values are completed by an age expert judgement. In addition, In the case of maturity of anchovy from the Gulf of Cádiz, for maturity ogives we impute missing maturity percentages from historical data.

For ALKs of benthic stocks (megrim and monkfish species), if there are values in any length of the length distributions to which the ALK will be applied, the gaps of those lengths in the ALKs are covered by the percentage from a wide time-series data, but also taking into account the strength of the cohorts in the ALK analysed, all by an age expert judgement.

For demersal stocks, age data are often scarce and it is difficult to construct a reliable annual ALK, so the gaps are not filled in.

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

It depends on several factos.

For small pelagic species, the most common procedure would be imputing information (LFD, ALK) from the adjacent (time-) strata (i.e. quarter) is imputed to missing values although under expert judgement.





For ALKs of benthic stocks (megrim and monkfish species), the gaps in the ALKs are covered by the percentage from a wide time-series data, but also taking into account the strength of the cohorts in the ALK analysed, all by an age expert judgement.

For demersal stocks, the gaps are not filled in.

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

No



**ILVO - Marine research (Flanders research institute for agriculture, fisheries, and food.)(Belgium)****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.2 Which country do you work in? **Belgium**

2.3 Which institute or laboratory do you work in? **ILVO Marine research (Flanders research institute for agriculture, fisheries and food.)**

2.4 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation) **Yes, the age reading lab is ISO 17025 certified.**

2.5 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s). **All biological sample data from commercial sampling at sea trips that are used for analytical stock assessments and hereto linked census data (logbooks and sales notes).**

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.2 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly).

- **ILVO Marine seagoing observers register sample data at sea directly in the database using a custom developed Smartfish application. The application is run on a rugged tablet coupled to an electronic measuring board. The same system is used as well for fish sampled at the fish auction (in case of self-sampling events) and for fish sampled in the fish lab for biological parameters such as age, sex, maturity, individual weight and length. Otoliths, fishscales and spines are processed using the Smartdots application, which is directly linked to the database. Sample and total weights and metadata such as haul- and shoot positions and net configurations are**





transcribed by the sea-going observer from paper to the Smartfish application upon return from sampling at sea.

- The sales and logbook data are provided by Department “Landbouw & Visserij”. ILVO Marine performs a number of quality checks on these data:
 - Temporal consistency in the numbers over the years (e.g. landings in weight and value, effort, ...)
 - Missing links between logbook and sales records
 - Validity of the dates
 - Validity of ICES statistical rectangle versus ICES areas
 - Validity of the gear type

3.3 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

- Data capture quality checks performed within the Smartfish application are the following:
 - Most data is entered through normalized vocabularies (vessel, stations, species, sample categorization, ...) in drop-down menus.
 - All length data is registered with a custom developed measuring board. The length data are shown in a graph during recording. There are no length constraints applied. However, visual inspection of the graphs allows detection of unrealistic values.
 - A general species-specific length-weight key check is applied for every weight registration (sample and individual weight). A notification is displayed for an abnormal weight. The user can reject the notification or choose to change the initially registered weight. Detected deviations are always logged and can be traced.
- Data capture quality checks performed within the custom developed Smartdots application (software for age reading using images) are the following:
 - All otoliths are read by two persons for the determination of the age. No age constraints are applied in the SmartDots application.
 - Reference otoliths collections are used for guaranteeing continuous good quality of the age readings.
 - Participations in otolith exchange exercises and workshops is done on a regular basis.

3.4 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

- Yes, all data is normalized. We use international code lists such as ICES vocabulary, FAO species codes, EU Fleet register and EU fisheries codes.

3.5 Do you perform any outlier checks on your data? If yes, please explain:





Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates). How do you define an outlier? How do you check for outliers? (e.g. graphically using expert judgement, R scripts). At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

- **Outliers are defined as unusually high or low values using expert judgement based on data from previous years to define 'high' or 'low'. Outliers are also defined in graphs as points that do not follow the apparent trend in the graph.**
- **Data management quality checks are performed in the Smartfish application, Microsoft Power Bi and R.**
 - **When all data is entered in the Smartfish application, the trip is labelled with the status 'raw'. Depending on his role, a user can change the status from 'raw' to 'validated' and from 'validated' to 'consolidated'. Once the trip status is set to 'consolidated' all data become read-only. More details about the validation performed at a status change can be found in SmartFish Required - validation - defaults.pdf (see attached with this document)**
 - **The sea-going observer can change the status from 'raw' to 'validated' when:**
 - **Mandatory fields are checked in the Smartfish application,**
 - **Positions have been visualised on a map, haul duration has been checked using Microsoft Power Bi,**
 - **Length frequency distributions and length-weight relationships are checked using visual inspection of graphs in Microsoft Power Bi.**
 - **A scientist can change the status from 'validated' to 'consolidated' when:**
 - **Haul duration and positions are double checked in Power Bi,**
 - **Length frequency distributions and length-weight relationships are double checked in Power Bi,**
 - **Catch per unit of effort (CPUE) per species per haul is calculated and inspected in a graph in Power Bi for abnormally high or low values,**
 - **Quality checks on age data are performed by plotting age-length relationships in graphs using an R script.**
- **Data extraction quality checks are performed in R for the purpose of the ICES combined fisheries data call in February-April of each year. These quality checks are performed on a stock level.**
 - **Data are raised on a stock level, when a number of thresholds are met:**
 - **Discard quantity is provided when:**
 - **At least 2 trips and 65 hauls are sampled**
 - **OR at least 2 trips and ≥ 70 kg landings are sampled**
 - **OR at least 2 trips and ≥ 20 kg discards are sampled**
 - **Discard length distributions are provided when:**
 - **At least 2 trips and 65 hauls are sampled and ≥ 30 discard length measurements are available**





- **OR at least 2 trips and ≥ 70 kg landings are sampled and ≥ 30 discard length measurements are available**
- **OR at least 2 trips and ≥ 20 kg discards sampled weight and ≥ 30 discard length measurements are available**
 - **Landings length distributions are provided when:**
 - **At least 2 trips and 65 hauls are sampled**
 - **OR at least 2 trips and ≥ 70 kg landings are sampled**
 - **OR at least 2 trips and ≥ 20 kg discard weights are sampled and 100 length measurements are available.**
- **Haul duration is verified. Shoot and haul positions are plotted to identify mistakes during imputation. Abnormalities are checked with a seagoing observer and corrected.**
- **Sample weights should not be larger than total weights. Mistakes are checked with the paper documents of that specific trip and with a seagoing observer.**
- **Length ranges and weight ranges are checked per fate category using expert judgement. Abnormalities in sample weights are checked using an LWK (see further). Outliers in length data are checked with a seagoing observer, verified using an LWK (when it is a large outlier) and removed from the database when it appears to be a true mistake in data capture.**
- **Boxplots are made to identify outliers in total weights per fate category. Abnormalities are verified with the paper documents of that specific trip and checked with a seagoing observer.**
- **Delta plots, as implemented in the COST package, are used to identify outliers in the length distributions per haul. Outliers are checked using an LWK. The calculated sample weight is then compared to the recorded sample weight. When a difference of more than 30% is found, the recorded sample weight is modified to the estimated weight in consultation with a seagoing observer.**
- **Discard rates are only calculated if a number of thresholds are met. Outliers are therefore rare.**
- **Outliers in the census data (logbooks and sales data) are detected as described under 3.1.**

3.6 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

- **Thorough cross checks are not done on a regular basis. However, there are two situations where these cross-checks have been performed:**
 - **For the purpose of the latest ICES combined fisheries data call, we verified the amount of BMS landings registered on fleet level and compared it to the sample data. Inconsistencies were found, because observers are not able to distinguish between BMS and discards on a haul level. BMS data were therefore not uploaded to the ICES InterCatch platform.**





- For the purpose of 2 recent sole benchmarks, auction landings data were checked with landings data registered by the seagoing observers, which highlighted some differences. Further investigation led to the correction of the Belgian TBB_DEF_70-99 sole landings data in 27.7d and 27.7fg (more information: ICES WKFLATCSNS 2020; ICES WKNSEA 2021).

3.7 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

- Yes, when changing the status from ‘raw’ to ‘validated’, the Smartfish application checks whether required fields are completed (see answer under 3.4). During data extraction, length and weight ranges are investigated in R using the command “table(Dataset\$weight, use.NA=”always”)”.
- Missing values in the census data (logbooks and discards) are detected as described under 3.1.

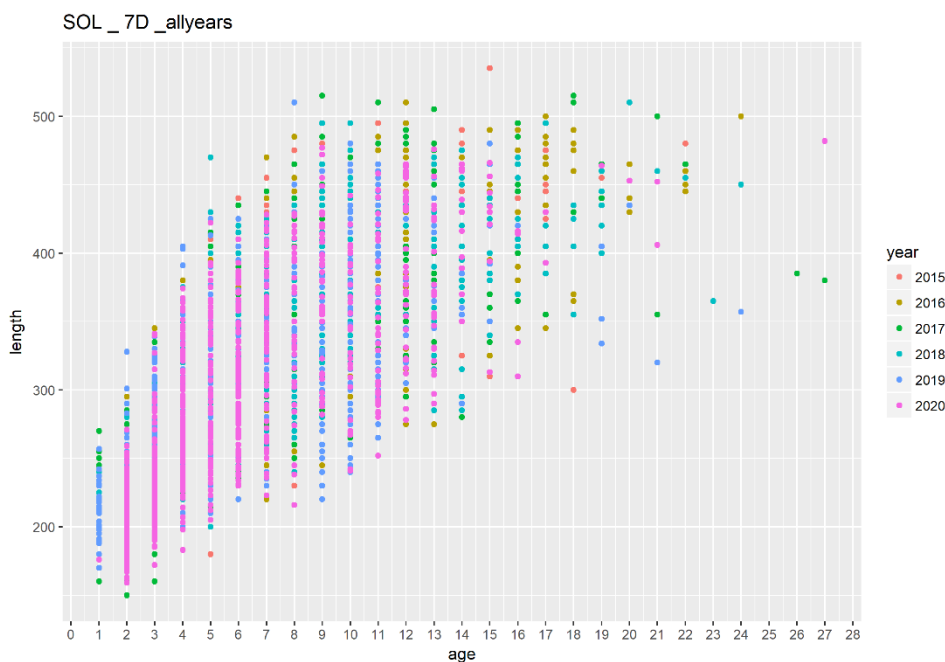
3.8 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

- Yes, registered haul and shoot positions are visualised in Smartfish, PowerBi and R. These checks are performed during data management and data extraction quality control (see answer under 3.4). The Smartfish application allows the automatic allocation of ICES statistical rectangles and ICES areas/divisions to the inserted shoot and haul coordinates.

3.9 Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

- Yes, when checking age-length keys, different years are plotted to check for temporal consistency (see figure below). Additionally, the same group of scientists performs the data extraction for the purpose of the ICES combined fisheries data call. Expert judgement used to quality check certain parameters is therefore built over the years.





- **Temporal consistency data checks are also performed on the logbook and sales data (as described under 3.1):**
 - Landings in weight and value per ICES area, per gear
 - Effort in fishing hours and fishing days

3.10 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

- **Yes, the Smartfish application does not allow users to create duplicated samples during the data capture process. Similar process is valid when working with the age reading tool Smartdots.**
- **Duplicates in the logbook and sales data are checked by the Department “Landbouw & Visserij” prior to sending the data to ILVO Marine.**

3.11 Please let us know about any other relevant data checks which have not already been described in your answers.

/

3.12 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

- **Yes, documents are available describing quality checks of measurements of biological parameters: length, weight, age and maturity. As age reading is done under the ISO Norm 17025, the whole process of the reading itself is also described and documented.**
- **An internal protocol (in Dutch) describing how to perform the quality checks using Microsoft Power Bi (see answer under 3.4), is available. It is used to check recent sampling data (current year) as well as older sampling data (e.g. when time series for specific stocks are requested for benchmark meetings)**





- A protocol is available concerning data extraction for the ICES combined fisheries data call. Additionally, when data is raised for a certain stock, a template needs to be completed and saved on a shared drive to document which outliers were checked/modified.
- A protocol describing the data quality checks on logbook and sales data is also available.

4. Editing

4.2 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

- **If possible, data errors are corrected. If not possible to correct, data are excluded from any output.**
- **Outliers in sample weights are checked using LWKs and adjusted when deviating more than 30% from the estimated value in consultation with a sea-going observers.**

4.3 Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

- **Yes, the data extraction protocol for the ICES combined fisheries data call describes these editing guidelines.**

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

5.2 How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

- **To deal with gaps in ALKs and to assure good estimates for length categories which are poorly sampled, age-length keys (ALK) are modelled based on the observed ALKs using a multinomial logistic regression model (Gerritsen et al., 2006). Note that at least 35 individuals with length-age data should be available. If not, the ALK is not considered of high quality and age distributions are not uploaded to InterCatch (the exception is Data Limited Stocks, for which ICES request all available data).**
- **Parameters defining length-weight relationships (a and b in the equation: $Weight \sim a \times Length^b$) are only used when length-weight data from at least 30 individuals are available. If this threshold is not met, a and b parameters are borrowed from other countries or length distributions are not provided depending on the stock.**

5.3 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)





- **Belgium is obliged to sample the TBB_DEF_70-99 métier in the North Sea and Western Waters and the TBB_DEF_>120 métier in the North Sea. Filling gaps across métiers is not done. When the PSU per quarter is too low for a certain stock according to our national thresholds (see 3.4), data are provided by year.**

5.4 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

- **Yes, the data extraction protocol for the ICES combined fisheries data call documents the imputation approaches (in Dutch). For the purpose of the WKNSEA 2021 benchmark for cod 27.47d20, a working document was provided specifying the thresholds and imputation approaches used for this data call (in English), which is similar to the general protocol for the ICES combined fisheries data call.**



**IPMA – Instituto Instituto Português do Mar e da Atmosfera (Portugal)****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

- 1.1. What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

- 2.1. Which country do you work in?

Portugal.

- 2.2. Which institute or laboratory do you work in?

IPMA.

- 2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No.

- 2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

DCF onshore and onboard sampling programmes.

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

- 3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Data is captured on paper and then entered in the database as soon as possible / daily, typically by the same person that captured the data.





3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes, there are restrictions.

-All text/alphanumeric/categorical fields have limited code lists during data entry (e.g. port, vessel, metier, commercial species, size category, sampled species, maturation scale and state, etc...). No free text fields, except for supplementary observation fields.

-Numeric variables do not have code lists but for some/many of these variables, values are:

. constrained during data entry stage:

(e.g.

in onboard sampling: geographical coordinates – degree limited to values possible, minutes limited to 0-60, seconds limited to 0-60;

in onboard sampling: date – end dates/times limited to after start dates/times;

etc).

. checked during quality control stage (i.e. after data entry into the database and prior to data extraction for use):

(e.g.

in onshore sampling: if length is not between plausible values for the species (i.e. from literature and from sampling data from previous years), value is rechecked in the paper record;

in biological sampling: if length-weight, or length-gutted weight, or weight-gutted weight relationship has outliers, values are rechecked in the paper record;

in onboard sampling: if haul duration is not between plausible values for that metier (i.e. from sampling data from previous years), value is rechecked in the paper record;

etc).

When errors are found in sampling data (during data quality control, extraction or analysis stages), sampling data is corrected in the database only to reflect what was recorded on paper, except for unequivocal measurement unit errors that can also be corrected.

3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes, local code lists (described in 3.2). Local code lists based on international code lists.

3.4. Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)
- How do you define an outlier?
- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)
- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

In biological sampling data, outliers in length-weight, or length-gutted weight, or weight-gutted weight relationships are identified during data quality control stage (i.e. after data entry into the database and prior to data extraction for use), using R scripts. These outliers are identified based on confidence interval around the regression model used.





- 3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

Checks between sample data and census data are done for the following cases:

-For onboard sampling – Landed commercial species name and weight per haul are recorded by the observer, but during data entry stage this data is checked with census data (sales notes) though the latter is at trip level and not at haul level; Data is adjusted during data entry if needed.

-For onshore sampling – In each sampled landing event/trip, all combinations of commercial species*commercial size category are sampled (with the list of combinations obtained from pre-sales notes). In the national sampling database, for each sampled landing event/trip the list of combinations of commercial species* commercial size category is recorded based on census data (sales notes). If any combination is missing from the comparison between the pre-sales notes and sales notes, then during data entry stage it is recorded as landed but not sold.

- 3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Missing value checks are done during data entry stage and during data quality control stage (i.e. after data entry into the database and prior to data extraction for use).

They are numerous and will not be listed here.

- 3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes. For onboard sampling data during quality control stage, for some métiers (e.g. bottom otter trawl) maps are produced to check for hauls in unexpected locations (e.g. haul coordinates on land).

- 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Temporal consistency checks in sampling data, during data entry stage and data quality control stage (i.e. after data entry into the database and prior to data extraction for use) are only done for the following cases:

Number and list of species sampled (annual), number of lengths sampled per species (annual), number of trips sampled per métier onshore and onboard (annual and temporal).

Additional checks are done during data analysis stage.

- 3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Duplication checks are done during data entry stage and data quality control stage (i.e. after data entry into the database and prior to data extraction for use):

Duplication of trips is checked during data entry and data quality control.

Duplication of landing event/trip, commercial species*commercial size category*fraction, haul (in onboard sampling), box, sampled species, individual is checked during data entry stage.





3.10. Please let us know about any other relevant data checks which have not already been described in your answers

-

3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

-

4. Editing

4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

When errors are found in sampling data (during data entry, quality control, extraction or analysis stages), sampling data is only corrected in the database to reflect what was recorded on paper, except for unequivocal measurement unit errors that can also be corrected. Otherwise, other errors are not corrected in the database, and they are not excluded during data extraction, but data can be excluded during the data analysis stage prior to data submission.

4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

-

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

Imputing missing values from averages/surveys (depends on the species).

5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Imputing sampling data from other strata/year.

5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

-



**LUKE - Natural resources institute Finland****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1 Which country do you work in? **Finland**

2.2 Which institute or laboratory do you work in? **Natural resources institute Finland, Luke**

2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

No

2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Salmon catch samples from coastal fyke-net fishery in ICES SD22-32 in the Baltic Sea, self-sampling by selected fishers and catch samples from commercial HER and SPR fishery

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative





3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

SAL: In the first place individual sample information (place, date, weight, length, sex, adipose fin, etc) is written by hand on the scale sample envelopes by fishers. Data is recorded to the database in connection of age reading. Logical error checking is carried out (weight vs. length, size vs. age etc.)

HER and SPR: Data is captured electronically (e.g. the condition factor and length-and weight limits for certain species are automatically checked during measuring to eliminate errors). The herring age readings are done from sliced and stained otoliths, which is considered most reliable age-reading method for northern Baltic herring.

3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

SAL: Checking of recording errors: weight 500-30000g, length 40-140 cm,
HER and SPR: see point 3.1

3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

For species the FAO 3-alpha coding is used, for gear ISSCFG, 2016 used. For statistical rectangle a national coding that is transformable to e.g. to ICES rectangles. Area and sub-division follow ICES coding.

3.4 Do you perform any outlier checks on your data? If yes, please explain:

5.3.5. Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

biological parameters i.e. length-weight, length-age.

5.3.6. How do you define an outlier?

HER and SPR: by the degree of deviation from the average.

5.3.7. How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

graphically using expert judgement

5.3.8. At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this? **No**





- 3.6 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). HER and SPR: **Yes, during data extraction.**
- 3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **No**
- 3.8 Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **No checks between quarters.**
- 3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **No**
- 3.10 Please let us know about any other relevant data checks which have not already been described in your answers **None**
- 3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it. **No**

4. Editing

- 4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?) **HER and SPR: If data errors, inconsistencies or discrepancies cannot be traced down, the data will not be used. If age is missing from a certainly 0- or 1-age class individual (very small at the end or beginning of a year), it will be filled in as 0 or 1.**
- 4.2 Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it. **No**

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers





- 5.1 How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys) **SAL: no imputation; HER and SPR: impute missing values from averages**
- 5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata) **SAL: no imputation; HER and SPR: impute missing values from other strata**
- 5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it. **No**



**NMFRI - National Marine Fisheries Research Institute (Poland)****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2 About your work-place

2.1 Which country do you work in?

Poland

2.2 Which institute or laboratory do you work in?

National Marine Fisheries Research Institute in Gdynia, Poland

2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

None

2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Data collected in all sampling schemes.

3 Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

The example graphical outputs of the data quality assurance software can be found in a document entitled "Data quality check description". A link to the document is provided in point 3.11 of this questionnaire.





3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Data is captured on paper and transcribed to a centralised database system through a dedicated web application as soon as possible. Data is entered to the database in a two-stage process. Newly entered data are attributed with a status indicating that they are waiting for approval. Then, another person verifies the data and approves it.

3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Data is checked against common out of range errors at the step of entering into the database. This is implemented by using predefined limits for some properties, e.g. minimum and maximum length for a given species.

3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Categorical data are stored in the database using code lists which are regularly updated. Local code lists are used but consistency with ICES vocabularies is ensured.

3.4 Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

Outliers checks are implemented in a shiny application which is accessible from the Institute's internal network. Types of checks which are available: catch weight vs. sample weight, number of fish measured vs. number of fish aged by length class, mean weight of fish in a sample, age-length plots, length-weight plots, histograms of age groups and length classes.

- How do you define an outlier?

An observation is considered an outlier when it deviates significantly from a common trend of observations in the same group.

- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

Identification of outliers can be done visually on the available plots and tables. In case of age-length and length-weight relations there is a possibility to run automatic outliers identification which is based on Bonferroni test. Expert judgement is important in the outliers identification process because in some cases an outlier is connected with natural reasons, e.g. diseases, parasites, poor condition.

- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

Outliers checks are available in the web application immediately after the data is entered into the database. Data is checked regularly, but with a higher intensity before extracting it for sending to external databases (e.g. RDB, InterCatch, JRC).

3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?





Cross checks of sample data with census data is performed ad-hoc and concerns mainly species composition. The results from the comparison are used for data corrections if needed, before sending the data to external databases. Another ad-hoc check concerns fishing location information and is performed in case of suspicious entries.

3.6 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

The data entry software ensures that all mandatory information is registered. For biological parameters, the shiny application designed for data quality control, allows to list all records where age information has not yet been registered.

The protocol for collecting biological information at sea specifies the type of sampling as concurrent. Therefore, if information on a specific species or catch category is not registered, it is considered a “true zero”.

3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Spatial data checks are performed at the step of entering data to the database. These checks are carried out using a set of reference tables which enable to ensure the consistency of coordinates, areas, rectangles and national sub-polygons.

3.8 Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

The ad-hoc date validation is performed mainly before data extraction. The check consist in ensuring that the sample date is within or close to the trip dates, depending on the type of fishery.

3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

The database constraints prevent from entering duplicates in some data entry steps. Checksums are available at the level of entering biological data. Moreover, a relation with a parallel system for PSU selection, enables to identify potential duplicates.

3.10 Please let us know about any other relevant data checks which have not already been described in your answers.

None.

3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

The document describing data checking process can be found on the Polish DCF website:
https://dcf.mir.gdynia.pl/?page_id=367 The name of the document is "Data quality check description".





4 Editing

- 4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

If data errors are found, they are corrected in the raw data registered in the database.

- 4.2 Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No data editing process documentation is available.

5 Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1 How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

In cases of gaps in ALK or WLK, average values are used if available.

- 5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Imputation is not performed at national level but at Stock Data Coordination level. Data are provided to end user "as-is" (as collected, validated and recorded in national database).

- 5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

No data imputation process documentation is available.



**SLU(a) - Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

2.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1. Which country do you work in? [Sweden](#)

2.2. Which institute or laboratory do you work in? [Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences](#)

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation) [No](#)

2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s). [Market sampling of cod landings in the west coast of Sweden](#)

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly) [Data is captured on paper forms and recorded manually into the databased](#)





- 3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Yes. The following checks are some of those done when data is saved into the database: is there are data in the haul [only 1 haul considered since it is landings]?, is there length data for each individual?, is there a total weight for the catch? is the sample weight is not larger than the total weight? are all mandatory fields filled in? is there is a length frequency when specimen exist?**
- 3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies) **Yes, mostly local code lists**
- 3.4. Do you perform any outlier checks on your data? If yes, please explain:
- 3.4.1. Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates) **catch and sample weights. biological parameters, discards weights per haul, catch and sample weights, trip info (e.g., days at sea)**
- 3.4.2. How do you define an outlier? **Total weight < sampled weight; theoretical weight of sample (as obtained via length-weight relationship) very different from registered sampled weight; atypical catch fraction values (several types of box-plot analysis, e.g., by gear, etc); atypical lengths (several types of box-plot analysis, e.g., by fraction, by size category, by gear, etc); atypical values in several types of relationships and boxplots between biological variables (length, weight, age,...); unusual biological variables collected;**
- 3.4.3. How do you check for outliers? (e.g. graphically using expert judgement, R scripts) **R-scripts**
- 3.4.4. At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc). **During data extraction, prior to estimation**
- 3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this? **No.**
- 3.6. Do you perform any missing values checks? (e.g. missing values vs. "true zeros"). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Yes. We have r-checks on, e.g., missing and duplicated trips, missing samples, missing total landing and sample weight values, missing lengths in length frequencies, missing length frequencies, missing specimens, missing biological data (various types of biological data)**
- 3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Not many. Some during the estimation.**





- 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Yes. During the estimation.**
- 3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Yes. During data extraction, prior to estimation**
- 3.10. Please let us know about any other relevant data checks which have not already been described in your answers
- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it. **Checks are documented in the r-script...**

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?). **Using weight and lengths as an example (can be generalized). When a clear error which value we can deduce with confidence we correct in the database (e.g., a wrong digit). If the error is clear and we cannot deduce its value with certainty, we delete it. If the error is not clear, we keep the data as is in the database and leave it up to the estimator/end-user to make a decision on inclusion/non-inclusion in each particular analysis**
- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it. **No. But we keep notes in scripts of what we do so consistency in handling the situations is kept between years.**

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys) **We do not use ALK or WLK in this programme – fish are taken at random from size categories, no length stratification involved.**
- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata) **In general there are few gaps. When so, we input maintaining the imputation strategy documented and reasonably consistent throughout the years.**





- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it. [Imputation is documented in scripts. Its most important steps are also documented as notes to stock coordinator in InterCatch format.](#)



**SLU(b) - Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1. What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1. Which country do you work in? [Sweden](#)

2.2. Which institute or laboratory do you work in? [Institute of Marine Research, Department of Aquatic resources, Swedish University of Agricultural Sciences](#)

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation) [No](#)

2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s). [Pot fishery for Norwegian lobster. \(Length, weight, sex, maturity in females and diseases.\)](#)

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly). [Data is captured in an electronic protocol on a tough book in the field. The electronic protocol is developed by the institute and is designed for different sampling types of on-board sampling \(and surveys\). The user chooses a sampling type of the current trip and is then steered through a defined workflow, with flexibility due to differences in work schedule on different fishing vessels. Measurements as length can be made with an electronic calliper, connected by Bluetooth or USB-wire to the tough book. All data from the trip is transferred to the main database when coming back to the institute. For safety, a copy of the data is made on an USB-stick regularly during the trip.](#)

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). [When entering data into the electronic protocol length and weight are checked towards historical/known length-weight relationships +/- 30% for the measured species. When outliers are detected observers](#)





get a question so they check if they are correct. If yes, the outlier value can be stored anyway. Sample weights are checked by comparing the length frequency of the sample and sample weight cannot be larger than the total weight. When the data is entered into the database, additional checks are made. These checks are also made for data inserted manually in the database, so there are some overlaps. (Checks if: there are data in each haul, there is length data for each individual, there is a total weight for the catch, the sample weight is not larger than the total weight, all mandatory fields are filled in, the number of hauls is not larger than the number of sampled stations, there is a length frequency if there is specimen data.)

- 3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies). [The person performing data capture choose e.g. latin names and gear types in predefined lists in the electronic protocol. Only comments can be made in free text.](#)
- 3.4. Do you perform any outlier checks on your data? If yes, please explain:
- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates) [At sea, outliers of the data are identified using length- weight relationships \(in the case of individual specimens\).](#)
 - How do you define an outlier? [+/- 30% weight than theoretical weight for each given length](#)
 - How do you check for outliers? (e.g. graphically using expert judgement, R scripts) [internal calculations to toughbook](#)
 - At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc). Data capture [data capture at sea](#)
- 3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this? No.
- 3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). [At sea, the protocol controls that all values needed for a sample type is there, per individual when measuring individuals, per haul, when verifying the haul when ending the haul registration and per trip when verifying the trip, before entering the harbour. This because no values should be forgotten before leaving the boat.](#)
- 3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). [The position is not yet checked in the electronic protocol, but as long as you have wifi/satellite contact the position can be captured automatically at the sight by pushing a button. If not, it needs to be entered manually. You can also choose an ICES rectangle. The rectangle or position defines which sampling target you have \(you define this beforehand\) and the number of sampled specimens per species is restricted due to this. \(When collection a number of otoliths per length class in fish, the protocol jumps to the individual sampling page automatically as long as you still have individuals to sample. When a length class is full/ready the protocol stops jumping to the sampling page and stays on the length measurement page. To facilitate the correct number of sampled otoliths.\)](#)
- 3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). [Length-weight relationships checked at sea are made of recent measured values. \(In species where the seasonal variation is large and where there is enough data, there will be seasonal relationships to compare with in the future.\)](#)





- 3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). **Yes, duplications are checked for at several occasions, when importing data from the field, ad hoc in the database (for things that cannot be checked when registration or import of electronic data occurs) and when delivering data to ICES. Things that are compared are eg. but not only:**
- The combination any vessel and fromdatetime must be unique.
 - The combination fish number and catch id must be unique.
 - The combination length group and catch id must be unique.
 - The combination species, processing, preservation, size must be unique.
 - The combination station, species and sub sample must be unique.
- 3.10. Please let us know about any other relevant data checks which have not already been described in your answers
- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it. No

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?) **If possible, sample data is corrected and data outputs are updated. If sample data cannot be corrected (often the case since it is often impossible to correct it after the sampling is done), the data is excluded from the estimation and outputs.**
- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it. **No written processes or guidelines.**

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys) **Impute missing values from surveys, if possible.**
- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata). **If surrounding quarters are well sampled, missing values are usually imputed from the nearest quarter (Q1-Q2 or Q3-Q4), or two quarters are pooled in the estimation. If sampling is poor in nearby quarters as well, the gap is left and no estimate is submitted.**





5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it. [No written guidelines.](#)





THN - Thünen Institute of Sea Fisheries (Germany)

Questions

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1. Which country do you work in?

Germany

2.2. Which institute or laboratory do you work in?

Thünen Institute of Sea Fisheries

2.3. Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

NA

2.4. Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Raised biological commercial data of the German commercial fleet (except Baltic), by-catch data

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1. When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

captured on paper and then transcribed as soon as possible after each sampling activity, electronic recording system under development

3.2. Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Checking is in place e.g. by automatic outlier search, plotting boxplots or histograms, comparison with length-weight relationships etc., during data input and data extraction.





3.3. Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes, local code lists and international code lists such as ICES vocabularies

3.4. Do you perform any outlier checks on your data? If yes, please explain:

yes

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)
biological parameters, discards weights per haul and rates
- How do you define an outlier?
deviation from statically average over a certain threshold depending on parameter, boxplots, comparison of discards rates over the years
- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)
scripts mostly
- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).
during data input and data extraction

3.5. Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

Not on regular basis and only based on expert judgement

3.6. Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes, at data capture

3.7. Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes, at data capture

3.8. Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Not known

3.9. Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Yes, during data extraction. All tables in the national database related with primary and foreign keys, which reveal the duplications

3.10. Please let us know about any other relevant data checks which have not already been described in your answers

Histograms/density plots for some species; temporal/spatial coverage of sampling trips





- 3.11. Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

Work in progress

4. Editing

- 4.1. If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

Sample data will be corrected when possible before data supply

- 4.2. Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.

Work in progress

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

- 5.1. How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

Impute missing values from aggregated data (e.g. missing ALKs are replaced by yearly ALKs) or from survey data

- 5.2. How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

On national basis leaving the gaps

- 5.3. Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

Work in progress



**KU - Marine Research Institute of Klaipeda University (Lithuania)****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1 Which country do you work in?

Lithuania

2.2 Which institute or laboratory do you work in?

Marine Research Institute of Klaipeda University

2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

ISO 14001:2015; ISO 45001:2018; ISO 9001:2015

2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

Fish stock rather

3. Data checks

When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

The data collected on field are recorded on a paper note. If some measurements are carried out in laboratory data usually entered into excel worksheet. As soon as all data are collected, they are entered into IMPORTING excel workbook.





IMPORTING workbook contains sheets with information about measurements, trips, stations, official catches during sampled trip. Then after few data quality checks, data are imported into local ACCESS database. As backup CSV files are saved.

3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Partially. Some validation rules are set in IMPORT workbook: for Baltic Sea length diapason 5 – 10000 millimetres, individual weight between 1 – 50000 grams, etc. Values for sex, maturity, age from predefined lists are allowed only.

3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Mixed: local/working and ICES codes

3.4 Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

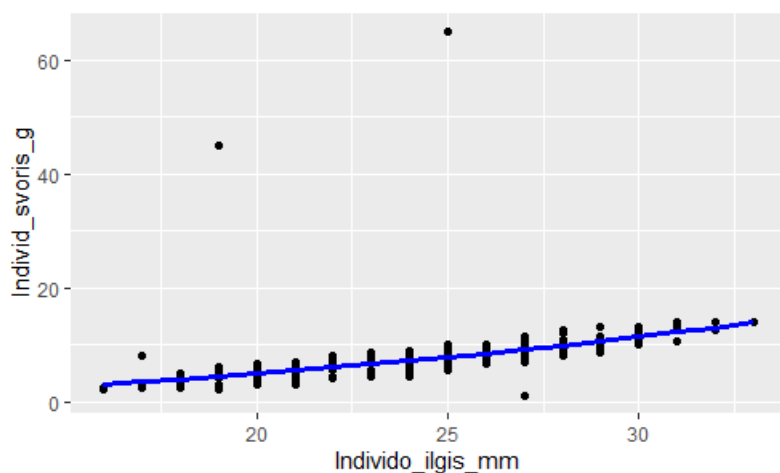
Recorded biological parameters: length - weight. This check is for identification of possible typing errors rather than for finding of outliers.

- How do you define an outlier?

Exp of residual is less than 0.5 or more than 2 (“rex” in the table below).

- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

R script: linear regression of log length and weight, graph, and table with outliers (see example for Pandalus borealis below).



	Individuo_ilgis_mm	Individ_svoris_g	pred	resid	rex	svpred
1	27	1.1	2.208689	-2.1133786	0.1208290	9.103771
2	17	8.0	1.243524	0.8359173	2.3069293	3.467813
3	25	65.0	2.048126	2.1262612	8.3834640	7.753358
4	19	2.0	1.475573	-0.7824254	0.4572955	4.373539
5	19	45.0	1.475573	2.3310899	10.2891495	4.373539







- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

After entered into IMPORT workbook but before importing into local Access database.

3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

During the sampling at sea researcher makes only raw estimations of catch, he has no real possibility to access the catches better than crew. Data on fishing effort and landings for the sampled trip are imported into IMPORT workbook after all these data are recorded into national fisheries data information system administrated by Fisheries service under Ministry of Agriculture. Simple R script extracts relevant data based on logbook number and landing data.

3.6 Do you perform any missing values checks? (e.g. missing values vs. “true zeros”). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

*Partially: for example, in the workbook for *Sebastes marinus* the crosscheck between length measurement and individual weight and collection of otoliths is performing during data entering. In a case when no one individual weight or otolith for measured length group is not recorded the coloured indicator “not enough” appears in the “checks” sheet. Observer is obligated to weight and take otolith for the relevant group during next sampling. For Baltic Sea simple R script created to detect some missing values: missing individual weight, missing sex.*

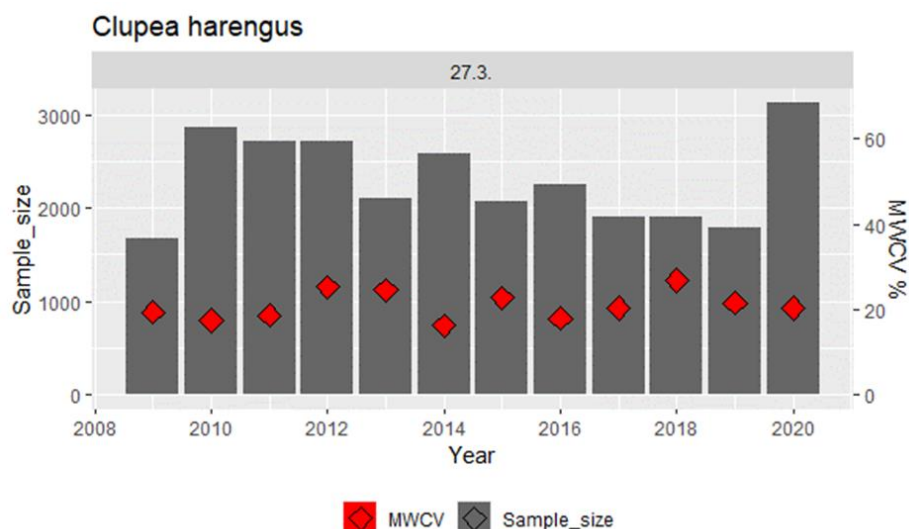
3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

No spatial checks yet. Logbook records accepted as reliable spatial information.

3.8 Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Simple R script for description of summary data statistics by species, year, quarter and metier. The output of this summary is information about sample size, mean value, median, quantiles, length diapason, mean weighted variation coefficient, etc. We do not consider this procedure as check, but it is useful when our sampling plans or sampling effort are discussed (example of one of outputs below).





3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

Simple R script created to indicate duplicated records in the IMPORT workbook

3.10 Please let us know about any other relevant data checks which have not already been described in your answers

3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

No written guidelines for data checking. Some basic checks to be performed by observer during sampling are described in the sampling protocols. Working list of Rscripts dealing with data is prepared.

4. Editing

4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

If data errors are detected during check of data entered into IMPORT workbook (see answer 3.4), additional check of primary (paper) records, contact with researcher carried out the exact sampling and then editing manually.

4.2 Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.





No formal guidelines for dealing with data errors. If some errors appears repeatedly, working discussion is imitated to find out the reasons of mistakes and possible ways to avoid it in the future.

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

5.1 How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

Some procedures to avoid gaps are integrated into primary data recording workbooks (see answer 3.6) missing data are collected during next sampling effort.

If some information is missing for the samples taken in Lithuania targeted vendor sampling is taken.

If some information (e.g.: age) is missing from Baltic Sea commercial fisheries sampling gaps are fulfilled with data collected during surveys within same area during the same quarter.

If not possible to get any data gaps are fulfilled with predicted average.

5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

Lithuanian fleet is small, number of strata is small (e.g. distant fleet one strata per sampling scheme, one or two vessels per sampling scheme), so no share of data between strata. For the Baltic sea there are some stratification based on gear type within sampling scheme. Therefore, no real time sampling strata checks are performed. If some gaps appear during to preparing of data to load into ICES data bases or to deliver under requests of data calls predicted average values within same strata are used.

5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

Principles and protocols are available in website Duomenų rinkimo programa – Jūros tyrimų institutas (ku.lt). R scripts are shared only internally within laboratory. We are not experts in R so public sharing of our scripts we consider as useless.



**WMR(a,b) - Wageningen Marine Research (Netherlands)****Questions**

We'd like you to answer the following questions – please provide URLs to publically available resources where they are relevant to the answers.

1. About you (answers will not be published)

1.1 What are the roles of the people completing this questionnaire? (Please use broad terms rather than specific job titles)

REMOVED PRIOR TO PUBLICATION

2. About your work-place

2.1 Which country do you work in?

Netherlands

2.2 Which institute or laboratory do you work in?

Wageningen Marine Research

2.3 Has your institute achieved any accreditations or certifications which are relevant to these questions? If so, please list them. (e.g. ISO 9001:2015, CoreTrustSeal, IODE accreditation)

ISO 9001:2015

2.4 Which data have you thought about when answering these questions? E.g. it might be all data from a named sampling scheme, or data collected from a named stock(s).

A: All landing data, all landing sampling schemes

B: commercial data collected on board

3. Data checks



When answering these questions please provide examples of graphical outputs or scripts if they would be informative

3.1 When is the data entered into an electronic recording system such as a database? (e.g. it is captured electronically, it is captured on paper and then transcribed as soon as possible, it is entered monthly)

Data Type	Data Format	Time of Import
Sale slips (A)	Electronically	Annually
Logbooks (A)	Electronically	Annually
Biological sample data (Landings) (A)	Electronically/Paper	Annually
Biological sample data (Bycatch) (B)	Paper and inserted electronically as soon as possible	Quarterly

3.2 Do you constrain the values of properties in your data recording system to be physically realistic? (e.g. lengths can only in a plausible range). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

A and B: There is a constrain for extreme values on age, length and weight by species in the data recording system (during data capture).

3.3 Do you use defined code lists for storing categorical information electronically? (e.g. No, free text; Yes, local code lists; Yes, international code lists such as ICES vocabularies)

Yes, local code lists.

3.4 Do you perform any outlier checks on your data? If yes, please explain:

- Which properties do you check? (e.g. biological parameters, discards weights per haul, catch and sample weights, census data, discard rates)

A and B:

- Individual weights
- Sample weights
- Size class weights (market sampling)
- Number of individuals length measured
- Number of individuals age measured
- Age range
- Length range
- Sex ratio
- Maturity stage
- Age-Length Matrix





- Spatial position
- Logbook data

- How do you define an outlier?

A, B: Outliers are data points with a significant distance from the majority of other observations. The outliers are visually identified with box-plots using the interquartile range criterion ($Q1 - 1,5IQR$ or $Q3 + 1,5IQR$), histograms and expert knowledge. If a data point is identified as an outlier, first it is examined if it's a wrong entry and if not, it is transmitted to the laboratory technicians to check if the value is an actual observation or a mistake.

- How do you check for outliers? (e.g. graphically using expert judgement, R scripts)

A: R scripts and expert judgement.

B: Graphically using expert judgement.

- At what points are the checks performed? (e.g. at data capture, during data extraction, ad-hoc).

A and B: During data capture, data import and data extraction.

3.5 Do you perform any cross checks of sample data with census data? (e.g. species composition, landing weights, unwanted catch weights). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc). If there is an inconsistency between the sample and census data how do you handle this?

B: No

A: Yes. There are cross checks between the sample and the trip in respect to area, metier, vessel name and weight. For the auction sampling, the size class sample weight is also checked against the reported size class weight from the auctions.

3.6 Do you perform any missing values checks? (e.g. missing values vs. "true zeros"). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

A: Not relevant for the landings.

B: Not done for bycatch

3.7 Do you perform any spatial data checks? (e.g. coordinates, rectangles, areas). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

A and B: Yes, the coordinates of the sample and census (catch) data are plotted in a map.





3.8 Do you perform any temporal consistency data checks? (e.g. checking the variation of data with quarters/years). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

A and B:

Variation in raised data is checked with previous years after data extraction

3.9 Do you perform any duplication checks? (e.g. checking that the same sample is not entered into a database twice). If yes, please describe the checks and at what points they are performed (e.g. at data capture, during data extraction, ad-hoc).

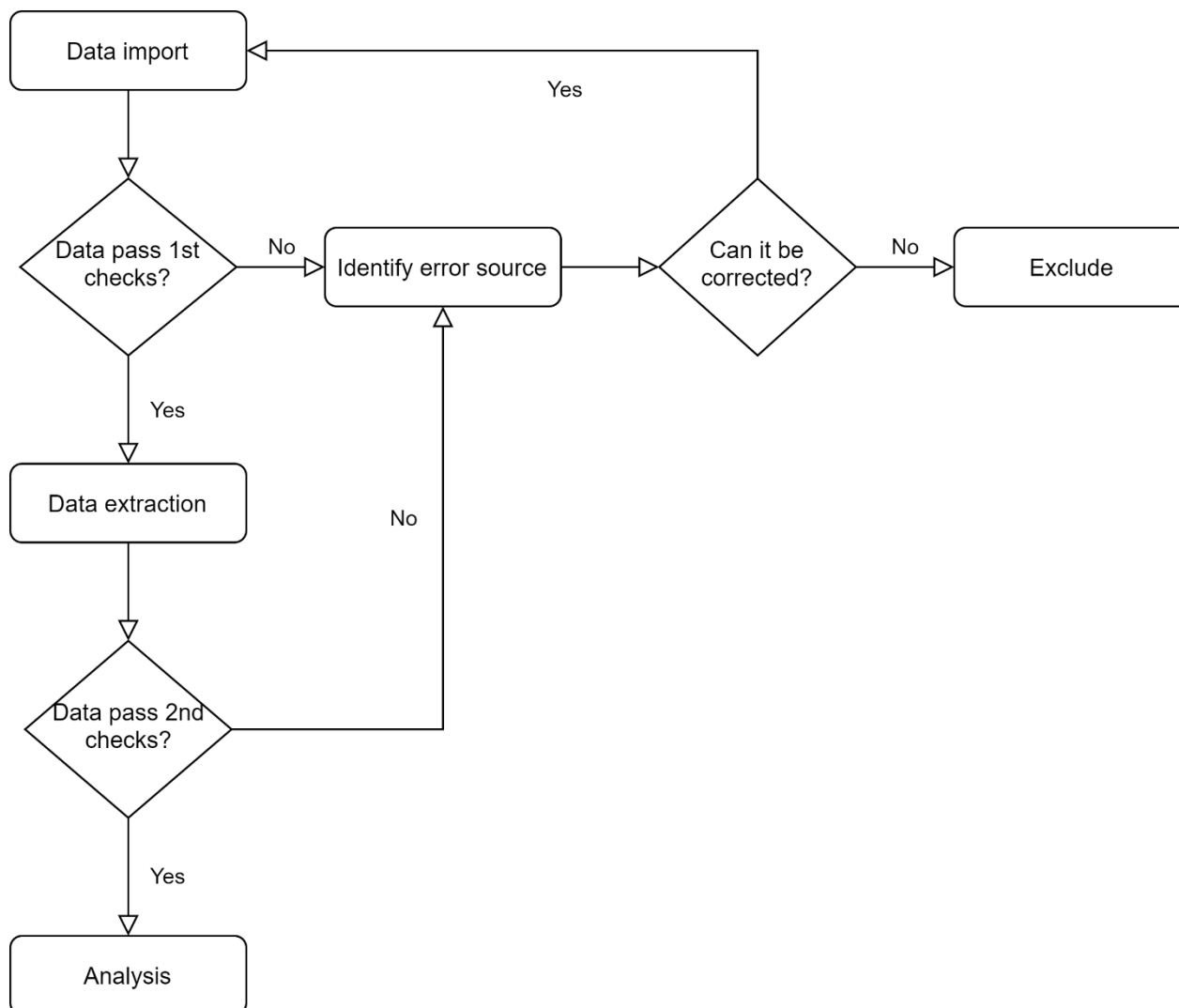
B and A: During data import and extraction the number of rows in the original data set is checked against the number of rows of the same data set when the distinct values are filtered out.. Furthermore, each sample is assigned to a unique sample ID. A unique sample ID can't be entered in the database twice.

3.10 Please let us know about any other relevant data checks which have not already been described in your answers

3.11 Do you have written processes or guidelines which define your approach to data checking? If so and you are allowed to share it, please provide a copy of the document or a link to it.

The diagram below describes the data checking procedure of the raw data starting at the import to the beginning of the analysis. This workflow applies to both census and sample data.





4. Editing

4.1 If data errors, inconsistencies, or discrepancies are found how do you deal with them? (e.g. do you correct the sample data, exclude the data from any outputs, replace with average values, correct data outputs such as InterCatch files?)

A: Depending on the type of error either it is corrected or excluded. If an error is not found until after the data analysis then the steps in the data checking process described above restarts. The InterCatch files are the final output of the analysis and they are not corrected manually.

B: If a data point is identified as an outlier, first it is examined if it's a wrong entry and if not, it is transmitted to the laboratory technicians to check if the value is an actual observation or a mistake. If the technician points it out as a mistake the data is removed from the database and consequently excluded from any output.

4.2 Do you have written processes or guidelines which define your approach to dealing with data errors, inconsistencies, or discrepancies? If so and you are allowed to share it, please provide a copy of the document or a link to it.





The data errors, inconsistencies and/or discrepancies are recorded in dedicated documents during the data checking process annually. For example, if an error is found in the sample data the following mandatory fields need to be field in the documentation template:

- SampleID
- Species
- DateChecked
- ErrorDescription
- ActionsTaken (e.g. excluded, corrected)
- Reason
- DateProcessed
- Re-imported (Yes/No)
- Who

5. Imputation

If you have different imputation processes for different end-users please make these clear in your answers

5.1 How do you deal with any gaps in your Age Length Key (ALK) and/or Weight Length Key (WLK)? (e.g. leave the gaps, impute missing values from averages/models/surveys)

A and B: mMissing values are imputed first from averages, then from surveys, then from models.

5.2 How do you deal with any gaps in your sampling strata? (e.g. leave the gaps, impute missing values from other strata)

A and B: Leave the gaps.

5.3 Do you have written processes or guidelines which define your approach to imputation? (note that a written process could be in the form of a document or scripts e.g. structured R markdown scripts or similar). If so and you are allowed to share it, please provide a copy of the document or a link to it.

The approach to imputation is included in the scripts that are used for the data analysis and raising procedures. These scripts are considered intellectual property of the insitute and can not be shared at this stage.

