

# ICES WKBIOPTIM REPORT 2017

ACOM/SCICOM STEERING GROUP ON INTEGRATED ECOSYSTEM OBSERVATION AND MONITORING

ICES CM 2017/SSGIEOM:32

REF. SCICOM & ACOM

## Report of the Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM)

20–22 June 2017

Lisbon, Portugal



**ICES**  
**CIEM**

International Council for  
the Exploration of the Sea

Conseil International pour  
l'Exploration de la Mer

## **International Council for the Exploration of the Sea Conseil International pour l'Exploration de la Mer**

H. C. Andersens Boulevard 44–46  
DK-1553 Copenhagen V  
Denmark  
Telephone (+45) 33 38 67 00  
Telefax (+45) 33 93 42 15  
[www.ices.dk](http://www.ices.dk)  
[info@ices.dk](mailto:info@ices.dk)

Recommended format for purposes of citation:

ICES. 2017. Report of the Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM), 20-22 June 2017, Lisbon, Portugal. ICES CM 2017/SSGIEOM:32. 150 pp. <https://doi.org/10.17895/ices.pub.8687>

For permission to reproduce material from this publication, please apply to the General Secretary.

The document is a report of an Expert Group under the auspices of the International Council for the Exploration of the Sea and does not necessarily represent the views of the Council.

© 2017 International Council for the Exploration of the Sea

## Contents

---

<b>Executive summary .....</b>	<b>1</b>
<b>1 Introduction .....</b>	<b>2</b>
1.1 Terms of Reference .....	2
1.2 WKBIOPTIM participants and agenda.....	2
1.3 Background to WKBIOPTIM and report outline .....	2
<b>2 Data and script preparation .....</b>	<b>5</b>
2.1 Analyses on the number of individuals collected from biological samples.....	6
2.1.1 Rationale of the analyses.....	6
2.1.2 Development of the scripts.....	6
2.2 Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment.....	8
2.2.1 Rationale of the analyses.....	8
2.2.2 Development of the scripts.....	9
<b>3 Quality Indicators.....</b>	<b>11</b>
<b>4 Sample level analysis.....</b>	<b>14</b>
4.1 Sole in ICES division 7e (sol-echw) 2015 sampling.....	14
4.1.1 Fishery description .....	14
4.1.2 Sampling programmes.....	14
4.1.3 Interpretation of results .....	16
4.2 European hake in Central-Southern Tyrrhenian Sea (GSA 10) .....	20
4.2.1 Stock unit .....	20
4.2.2 Growth and maturity .....	21
4.2.3 Sampling strategy .....	22
4.2.4 Data preparation .....	23
4.2.5 Define sampling strategy and number of simulations .....	23
4.2.6 Outputs.....	24
4.3 African black hakes in CECAF area .....	30
4.3.1 Sampling .....	30
4.3.2 Data.....	30
4.3.3 Output .....	30
4.4 Danish case study with sandeels ( <i>Ammodytes marinus</i> ).....	32
4.4.1 Sampling design.....	32
4.4.2 Sample level - Looking at number of fish sampled per haul .....	34
4.4.3 Results .....	38
4.4.4 Discussion and conclusion .....	39
4.4.5 Future work .....	40

4.5	Adopting Optimization Approach for Commercial Sample to Survey Data .....	41
4.5.1	Data.....	41
4.5.2	Results .....	41
4.6	<i>Pandalus borealis</i> case study .....	46
4.6.1	Fishery .....	46
4.6.2	Sampling .....	46
4.6.3	Methodology .....	47
4.7	Plaice (Germany) case study .....	57
<b>5</b>	<b>Multi-level analysis.....</b>	<b>59</b>
5.1	Sardine in areas 28.8.c and 27.9.a.....	59
5.1.1	Fleet.....	59
5.1.2	Sampling design.....	59
5.1.3	Data.....	59
5.1.4	Exploratory data analysis .....	59
5.1.5	Quality checks .....	61
5.1.6	Raised length composition of the stock .....	63
5.1.7	Scenarios .....	63
5.1.8	Discussion.....	66
5.2	Hake in area 27.9.a.....	67
5.2.1	Fleet.....	67
5.2.2	Data.....	67
5.2.3	Scenarios .....	69
5.3	Sole in areas 28.8ab and 27.4 .....	74
<b>6</b>	<b>ToR d) Identify a road map for the discussion with end-users of optimization perspectives .....</b>	<b>77</b>
<b>7</b>	<b>Conclusions and future work.....</b>	<b>79</b>
<b>8</b>	<b>References .....</b>	<b>81</b>
	<b>Annex 1: List of participants.....</b>	<b>83</b>
	<b>Annex 2: Agenda.....</b>	<b>85</b>
	<b>Annex 3: List of oral presentations.....</b>	<b>86</b>
	<b>Annex 4: Case study documentation.....</b>	<b>87</b>
	<b>Annex 5: Proposal: The second Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM 2).....</b>	<b>143</b>
	<b>Annex 6: Recommendations.....</b>	<b>146</b>

## Executive summary

---

The Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM), chaired by Ana Cláudia Fernandes (Portugal) and Julie Coad Davies (Denmark) was held in Lisbon, Portugal, 20–22 June 2017. Twenty-two participants from 12 countries within the ICES and Mediterranean communities were represented

The workshop focused on practical aspects of optimization of sampling. Prior to the workshop, two sets of R-scripts were developed that used the current exchange format of the Regional Database (RDB) as input. The first set of scripts is useful for cases where there is considerable a priori evidence of oversampling (e.g. several hundreds of measurements per sample over a relatively short size/age-class range). The script implements simulations of several types of sampling strategies (user defined) and produces a set of graphical and numerical outputs that allow the visualization of the consequences of measuring different number of individuals per sample. The second set of scripts can be used to determine the number of fish, hauls and trips that should be taken without significant loss of precision in the final estimates. In both cases scripts were prepared to use “lengths” as the biological parameter to be analysed but can be extended to other biological parameters, like age distribution or discards.

Participants brought their own case-studies and three subgroups were formed: one that tested the first script (sample level), one that tested the second script (multilevel analysis) and one that discussed quality indicators for length/age frequency data. Workshop time was spent introducing participants to the analyses carried out in the scripts, adapting data inputs to different formats (e.g. length frequency vs. length of individual specimens), debugging coding errors, and running simulations of the case-studies. The outputs of the case-studies were analysed during and after the workshop. In what concerns to quality indicators, some possible indicators that can be used for biological parameters were discussed.

WKBIOPTIM identified considerable margin to reduce the sampling effort in some of the case-studies presented without compromising the quality of the data to be used by the end-users. In what concerns both sample-level and multi-level optimization, the R scripts developed can be used to simulate and analyse a range of different sampling scenarios with the outputs being useful in discussions of improvements to national and regional sampling plans. Following the workshop, concrete reduction in the sample sizes collected for some species have been achieved after dialog with data end-users. Such reductions resulted in time savings and facilitated data collection in other stocks. In what concerns quality indicators for other biological parameters and additional quality indicators for length frequency data, a request for advice was sent to WGBIOP and WGCATCH with the aim of including a large array of indicators in future updates of the R-scripts and better adapt them to a wider array of end-users’ needs.

The expansion of the application of the R-scripts to other biological parameters (weights, ages and maturity) is considered of high importance since biological data collection is inherently multivariate and multi-purpose frequently extending far beyond collection of length data. Future developments of the scripts are expected to happen as part of WKBIOPTIM2 which will aim to, among other, include additional biological parameters in the analyses, the integration of additional quality indicators (e.g. effective sample size) and a discussion of the most appropriate balance between them (based on end-users’ needs). It is envisioned that a harmonized toolbox of R-scripts and R-vignettes, possibly encompassed in an R package, will ultimately be produced and aid national labs in the planning of their work.

## 1 Introduction

---

### 1.1 Terms of Reference

The Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM), chaired by Ana Cláudia Fernandes\*, Portugal, and Julie Coad Davies\*, Denmark, met in Lisbon, Portugal, 20–22 June 2017, to:

- a) Discuss indicators of sample quality that can be used in communicating the need and effects of statistical optimization of sampling to end-users (e.g. effective sample size; variability in mean length, age frequency, number of modes in distributions, etc.);
- b) Carry out hands-on work on code for statistical optimization of biological samples based on the CS and CA exchange format of the RDB and sampling strategy used to obtain the data. Code should be general and applicable to samples from different commercial sampling programmes and surveys. Different sampling effort strategies (e.g. fixed number, number dependent on size-span of the sample) and sampling strategies (e.g. simple random, two-stage stratified sampling) should be considered
- c) Test the code developed in a set of case-studies and quantify effects, advantages and disadvantages of different options of statistical optimization at sample level in terms of cost and time-savings involved;
- d) Identify a road map for the discussion with end-users of optimization perspectives

### 1.2 WKBIOPTIM participants and agenda

The list of participants and the agenda for the workshop can be found in Annex 1 and 2 respectively.

### 1.3 Background to WKBIOPTIM and report outline

Several ICES EG's, including e.g. WKPRESISE 2009, PGCCDBS 2012, PGDATA 2015 and WKCOSTBEN 2016, have pointed out that clustering effects in multistage catch sampling programs may lead to effective sample sizes much lower than the number of units sampled. The reason is that, e.g. fish caught during one trip or haul are not a random sample but often have more similar characteristics than the general population of fish they came from. Conjugated with large sample sizes, the latter effect highlights a likely existence of oversampling in the lower stages of many national catch sampling programs (e.g. trips, hauls within trips, fish within hauls), where an excessive number of individuals are possibly sampled that do not accrue much additional information relative to a smaller number. This sampling effort is inefficient at that it does not provide significant additional information on the sample itself nor on the population, i.e. when data are aggregated at higher levels to provide fleet or fishery estimates to end-users. Such inefficiency also has concrete indirect consequences that are not always explicitly considered: when more time/funds than needed are spent sampling one species less time and funds are available to sample other resources and other tasks (e.g. quality checks, data analysis, etc.), i.e. oversampling of one resource frequently reduces the quantity and quality of data available from other resources. Under the new EU-MAUP it is now a requirement that Member States use statistically sound catch sampling programmes when collecting data to meet the needs of various end-users (COM IMPL DEC 2016/1701). Statistically sound sampling allows the determination of the

variance of final estimates and opens way to optimize sampling protocols<sup>1</sup>. It is important that multipurpose sampling schemes such as the EU-MAUP's are optimal in terms of effort and cost spent on sampling for each goal at hand because excessive effort and costs spent on one particular goal are effort and costs not available to meet other. There is an increasing need for efficient resource allocation as new requirements for data collection arise, e.g. determining MSY reference points for previously unassessed stocks, many of which are data limited and the need to record bycatch information during at-sea sampling. In addition, it is likely that new information will be requested in the future on variables for multispecies and Integrated Ecosystem Assessments (IEA's) which will mean that optimising the number of length measurements taken, fish aged and maturity staged will be a priority as these tasks require that significant funds and time are spent by national laboratories.

In 2016 the Workshop on Cost Benefit Analysis of Data Collection in Support of Stock Assessment and Fishery Management (WKCOSTBEN) took some initial steps in examining the issue of costs and benefits of regional data collection programs. Outcomes of WKCOSTBEN were discussed at the Working Group of Commercial Catches (WGCATCH) where it was noted that some national institutes had developed simple statistical tools with the aim of analysing and reducing sample sizes on some clear-cut cases of excessive sampling at sample-level (e.g. situations where several hundreds of individuals from species were being measured from a single haul or trip) (ICES, 2017). Those national methods had not yet been discussed and shared in the wider community and in particular with elements of the ICES Working Group on Biological Parameters (WGBIOP) to whom they would be of interest. A Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM) was so proposed to be co-chaired by Ana Cláudia Fernandes (participant of WGCATCH) and Julie Coad Davies (participant of WGBIOP) so that effective communication between the groups was established (ICES, 2017). As a next step the ICES Planning Group on Data Needs for Assessments and Advice (PGDATA) revised the ToR's of WKBIOPTIM coordinating them with WKCOSTBEN chairs and improving their integration into the ICES cost benefit framework. The need for such a framework is highlighted in the ICES document on "Implementing the ICES strategic plan 2014–2018" (ICES, August 2014). At the PGDATA 2017 meeting ToR's were finalised and preparations were made for the Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM).

PGDATA 2015 (ICES 2016) noted that the utilisation of software routines integrated with the Regional Database (RDB) would allow for increased efficiency in the data compilation process and increase the quality of the evaluation process of fisheries data. Following up on that work, WGCATCH and PGDATA established the development of a set of R scripts based on the widely available RDB format as the main aim of WKBIOPTIM so that national labs could make more effective use of the code developed to quantify the effects of different sampling intensities and sampling designs, and support their discussions on the advantages and disadvantages of different sampling strategies in terms of time and cost savings involved. R (<https://www.r-project.org/>) a freely available language and environment for statistical computing and graphics widely used by fisheries scientists within ICES was selected as the coding language for the scripts of WKBIOPTIM.

---

<sup>1</sup> Optimization: an act, process, or methodology of making something (such as a design, system, or decision) as fully perfect, functional, or effective as possible ([source: https://www.merriam-webster.com/](https://www.merriam-webster.com/))

WKBIOPTIM was from its start envisioned to be a practical workshop with hands-on simulation work using R code for statistical optimization of biological samples from different sampling programmes, and discussions on the outputs from case studies and quality indicators for sample data. Discussions held throughout the workshop should highlight needs and issues to be addressed in follow-on workshops where the end-users would be required to give feedback and input on which further procedures could be examined. By drawing on the expertise of those already involved in WGCATCH and WGBIOP the workshop aimed to bring together participants more familiar with R-coding and sampling design and participants experienced in age and reproduction analysis, with its outcomes being of particular interest to those involved in the annual planning of national sampling protocols, laboratory analysis of biological samples and national work plans in general.



## 2 Data and script preparation

Preparatory work for WKBIOPTIM can be broadly divided into a) the preparation of R- scripts aiming at analysing of the number of individuals that should be collected in individual biological samples, b) the preparation of scripts aiming at multilevel analyses of sampling effort needed to produce core fleet-level estimates for stock assessment, and c) literature review on quality indicators. This work was steered by the chairs of the workshop that coordinated with script developers in several skype meetings.

The conceptual idea around the WKBIOPTIM Toolbox is presented in Figure 2.1 and summarizes much of the work to be developed under WKBIOPTIM. Data in the RDB format is used to perform the analysis and in this first approach some possible scenarios for running simulations for both sample level and national /multilevel are presented. In future, scenarios will be adapted according to input on quality indicators and end-users' needs. Graphical outputs will also be adapted to specific purposes.

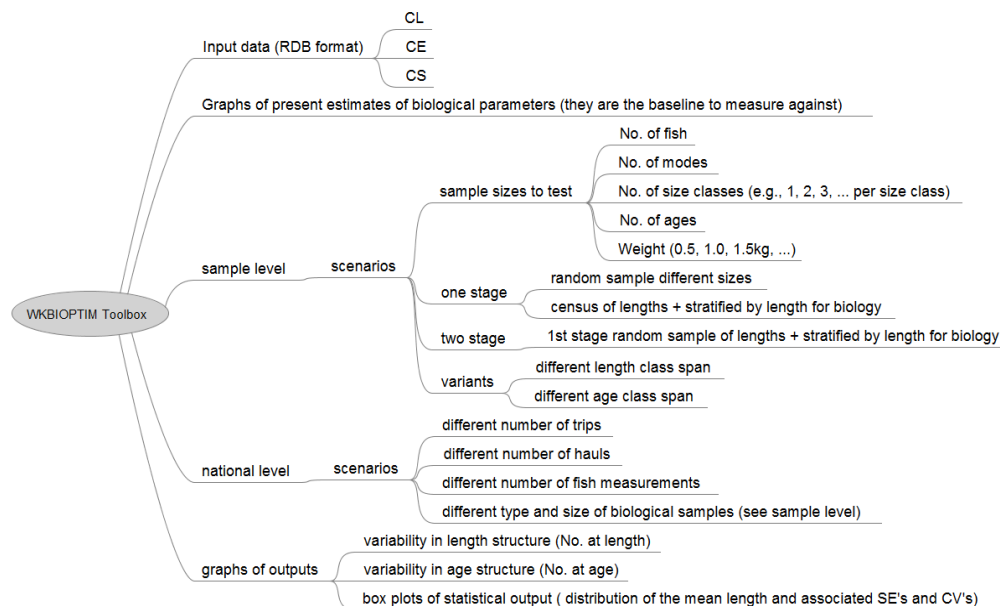


Figure 2.1 - Schematic approach of WKBIOPTIM Toolbox

## 2.1 Analyses on the number of individuals collected from biological samples

### 2.1.1 Rationale of the analyses

It is widely known that collecting an excessive number of samples at lower levels of stratified multistage sampling designs is inefficient at that it costs time and accrues little to the final precision and accuracy of estimates provided to end-users (e.g. WKCOSTBEN). However, end-users and final goals of the sampling programmes may be unknown to the data collectors and frequently change though time with the data collected today being useful in the future for purposes that may require or benefit from precision levels larger than the ones originally in mind (e.g. new research may require finer spatial resolution; new stock assessment models may require precise estimates at spatial and temporal scales smaller than the ones currently used).

The starting point for the development of scripts addressing the number of individuals collected from biological samples was the idea that, in a context of undefined or ever changing end-user needs, well-characterized samples have a better chance at meeting both present and future data needs. Such an idea seemingly ignores the importance of upper levels in the sampling hierarchy which are known to be the ones responsible for most variance (WKCOSTBEN, ICES 2017) *but* does not imply necessarily that too large numbers of individuals are sampled. Rather, one could search for a conservative number of individuals that even if still relatively large for specific present purposes would already represent a progress relative to pre-established sampling targets while cautioning for existing concerns of lab personnel and end-users on future data needs. The main advantage of this approach is that it allows for a conservative reduction in some present sample sizes without a need to specify a priori and simulate all end-user needs. By being conservative such approach is also more easily accepted by end-users and easier to communicate to lab personnel. The latter aspects concur to increase confidence on the discussions of data needs and consequences of the optimization process within national labs, freeing the optimization of more complex fleet-level analysis, and allowing an objective and ready reduction of sampling in cases where excessive number of individuals are quite evidently being collected.

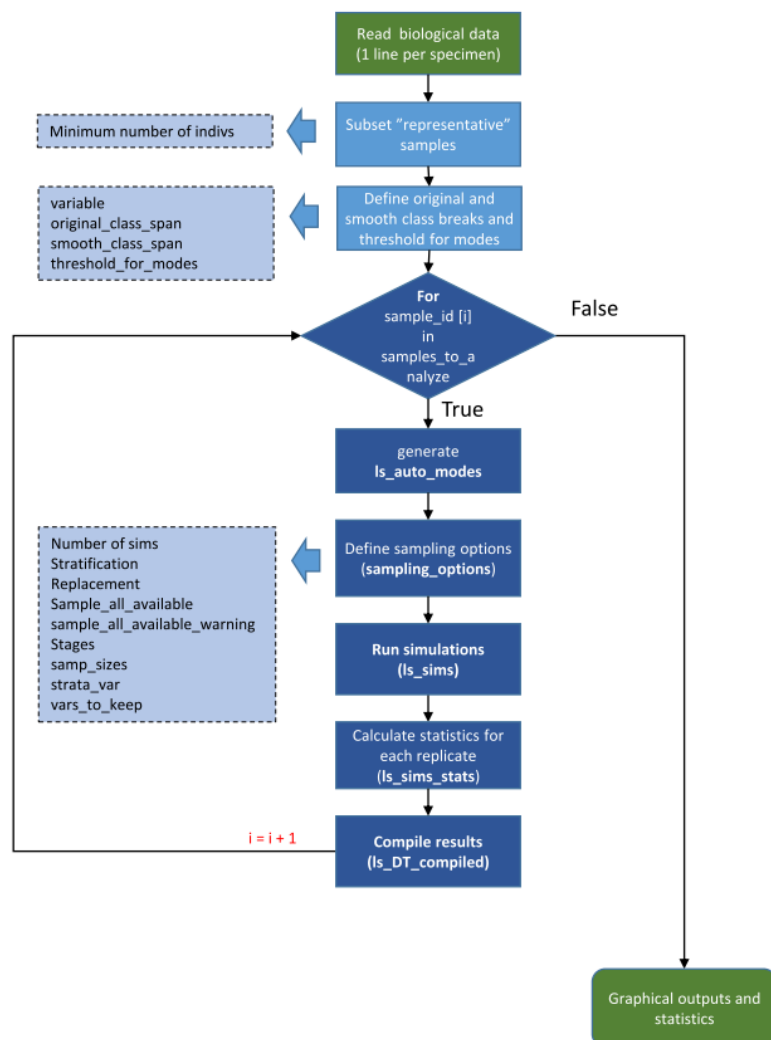
### 2.1.2 Development of the scripts

The development of the R-scripts for analysing the number of individuals that should be taken in each biological sample was undertaken by Nuno Prista (SLU, Sweden) using as a starting point subgroup work carried out during a local Workshop on Sampling Design and Optimization (Lysekil, 31 October–4 November 2016, unpublished). The original scripts used as input an extraction of biological samples (each containing a large number of individuals with length, age, sex and maturity determined) from a local database. The code then simulated a set of sampling designs and sampling intensities in each individual sample and a set of quality indicators for each replicate. Finally, a series of graphs was outputted that corroborated the variation of aggregated quality indices with sample size.

In the preparation of WKBIOPTIM the script was streamlined, annotated and further developed to a) accept RDB CA format as input, b) automatically determine some characteristics and indicators of the original samples (e.g. number of original modes and modes of a smoothed length distribution), c) extended the number options of accommodate a wider range (e.g. stratified sampling). After the workshop a new version of the script was produced with improved annotation, streamlined inputs (adaption to

different column names) and functions and improved speed in the simulations (through the use R-package parallel). The last version of the R script is depicted in Figure 2.1.2.1 and available on the SharePoint accompanied by sample data from the Pandalus case study (Section 4.6) that can be used to exemplify its usage. The following quality indices are included as outputs of the script:

- Mean, standard error of the mean, CV of the mean
- Minimum, Maximum, Median
- Number of size/age classes sampled
- Number of modes, number of modes identical with the modes of original sample
- Number of modes (after smoothing), number of modes identical with the modes of original sample (after smoothing)
- Two sample t-test for equal means
- Two sample Kolmogorov–Smirnov test



**Figure 2.1.2.1 - Flowchart of final version of the script used to analyse the number of individuals that should be collected from each biological sample.**

Contributions from the Training Workshop on Sampling Intensity (MARE/2014/19 Med & BS project, Deliverable 2.5) held in Cyprus on March 2017 following the Recommendation of 2016 RCM Med & BS-LP were also important for the work developed during this workshop. In this Training Workshop methods and scripts developed in the MARE/2014/19 Med & BS project (Deliverable 2.5) were applied and further tested, implementing the *Sampling Design tool* for optimization of sampling intensity (the *Sampling Design tool* was developed to adjust the sampling effort through an analysis of trade-offs between sample size and precision levels, based on COST tools). Tests were conducted on case studies of the Med & BS, using a simplified SDEF data format (named as RCM Med & BS-LP format). The *Sampling Design tool* allows to analyse the evolution of the sampling variability with the number of samples (number of trips), assess whether the current plans over/under-sample the fisheries and propose an “optimal” sampling plan in terms of number of trips for each technical/temporal stratum of the sampling scheme (i.e. by métier and quarter). In the proposed method, the bootstrap approach is used to resampling data from past years taking into account the interannual variability. The CV curves are built highlighting the trade-off between sampling size and precision level, that takes place where the CV curves tend to flatten (as the CV value decrease with the increase of the samples’ number). The reliability of the results can be assessed taking into account the recycling rate in each sampling stratum.

## **2.2 Multi-level analysis of sampling effort needed to produce estimates of catch numbers for stock assessment**

### **2.2.1 Rationale of the analyses**

Data collected in the EU-MAP framework are aimed first to provide stock information for the assessment working groups. National monitoring program and sampling plan have to be tailored to fit this need. Consequently optimization of the sampling plan has to follow the same constraint: to provide the estimates requested by the end-users.

In this framework, analyses were carried out to test how sampling plan changes can affect population estimates. The estimator used in this exercise was the length distribution estimate. Compared to age estimates or discards estimates, this estimate is the most simple to provide to assessment working groups. Moreover measuring length is the archetypal activity in fieldwork related to commercial fisheries and data are widely available in the national databases. Starting from the existing sampling database, simulations were used to modify the sampling plan. The simulation process includes changes in the sampling effort by decreasing the amount of samples or measurements and then computing new length distribution estimates using the modified sampling database. The new estimates were compared with the original length distribution to assess the effect of reducing samples on the estimates.

Various scenarios are used to test modification of the sampling plan and reallocation of the sampling effort. To do so, samples are removed according to the scenario, in a fixed or random way and length distributions are estimated based on the new sample base. The results are compared to the “true” estimates using the original number of samples and comparisons were made (1) visually and (2) using an objective metric. The Earth Mover’s Distance (EMD) was used. In statistics, the (EMD) is a measure of the distance between two probability distributions over a region D (names Wasserstein metric in mathematics) (Rubner, Tomasi, and Guibas 1998). Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over

the region D, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.

### 2.2.2 Development of the scripts

The code used in the simulations was written in R. Data inputs follow the Regional Database (RDB) Exchange Format version 2009 (Jansen *et al.* 2009). It's an older version of the actual format used in the Fishframe database (Jansen *et al.* 2016). Two data objects are used:

- the object CL (Commercial Landings) brings population data related to landing
- the object CS (Commercial Samples) brings samples data related to length

The framework developed for this study integrates the whole processing of fishery data to (1) build population estimates at the stock level (length distribution mainly) and (2) analyse the impact of different sampling strategies in these estimates by simulation. This framework is aimed for reproducibility and transparency, following the recommendation of the reproducible research statement (Gentleman and Lang 2004). Consequently this report is self-consistent: the code used to process and to analyse the fishery data are embedded in the report itself.

Analyses are carried out using the R environment (R Core Team 2012). R is a free software environment for statistical computing and graphics. The reproducibility of the results presented in this report relies on the use of a dialect of the Markdown language called Pandoc for word processing using the Knitr R package. Markdown is a plain text formatting syntax designed so that it can optionally be converted to HTML using a tool by the same name. Pandoc is a Markdown dialect which extends the conversion capability to word processing file (docx, doc and odt) and pdf, among other formats. Pandoc understands a number of markdown syntax extensions, including document metadata (title, author and date), footnotes, tables, figures and references. Knitr is an R package (a set of functions extending the R capabilities). With this package, the R code used to process and analyse the data are included directly in the report. Results are then produced dynamically. This framework has demonstrated the capacity to improve the conduct and the presentation of data analysis in a way that another person can understand and replicate (Baumer *et al.* 2014).

For example, if the calculus of 1+1 is needed, the code to compute it is written in the report using special hooks, as in this simple example:

```
'''{r test00,warn=FALSE,cache=TRUE,echo=TRUE}
#comment: addition example.
1+1
'''
```

This code is evaluated during the compilation of the report by the Knit R command and it prints the following result:

```
#comment: addition example.
1+1

## [1] 2
```

All the numerical values related to the data, including tables and figures are produced following this procedure. Some hooks are appearing in the report, other were removed to clarify the outputs. Moreover for the case studies analysed here, the original R-Markdown documents are available on the WKBIOPTIM in Annex 4. The document needs the COST R package (Software folder on the WKBIOPTIM SharePoint). To install them:

```
#the path to the library is Local (aka where are the files on your hard disk)
install.packages("/path/to/COSTcore_1.4-0.zip",repos=NULL)
install.packages("/path/to/COSTdbe_1.4-1.zip",repos=NULL)
install.packages("/path/to/COSTeda_1.4.0.zip",repos=NULL)
```

Some other packages are needed and can be installed using CRAN:

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("DATRAS")
install.packages("mapdata")
install.packages("pander")
install.packages('DATRAS',repos='http://www.rforge.net/',type='source')
devtools::install_github("ldbk/earthmovdist")
```

The library needs to be loaded in R. Moreover a set of functions are available in the `wkbioptim_fct.R` files and have to be sourced. Some options related to how the chunks of code are interpreted are fixed (size of the figures...):

```
#R general option:
options(stringsAsFactors = FALSE)
#working directory:
setwd("C:/path/to/the/working/directory")
#chunk option
#knitr option
knitr::opts_chunk$set(cache=TRUE,echo=TRUE, warning=FALSE,
  message=FALSE, fig.height=6,progress=FALSE,verbose=FALSE,
  include=TRUE,dev='png',autodep=FALSE)
#package
library(COSTcore)
library(COSTdbe)
library(COSTeda)
library(dplyr)
library(ggplot2)
library(maps)
library(mapdata)
library(pander)
library(DATRAS)
library(earthmovdist)
#source local file
source("wkbioptim_fct.R")
```

### 3 Quality Indicators

---

**ToR a) Discuss indicators of sample quality that can be used in communicating the need and effects of statistical optimization of sampling to end-users (e.g. effective sample size; variability in mean length, age frequency, number of modes in distributions, etc.);**

There are many steps to be taken when collecting and compiling biological data to be used for species or stock characterization as a first step and ultimately fish stock assessment. It is therefore highly important that quality measures are implemented at each phase of the data collection process to ensure the reliability of the target biological parameters. The Workshop on Methods to Evaluate and Estimate the Accuracy of Fisheries Data used for Assessment (WKACCU, 2008) focused on the detection of bias in a list of key parameters important for stock assessment e.g. species identification, length and age structure and sex-ratio. In 2009 the Workshop on Methods to Evaluate and Estimate the Precision of Fisheries Data used for Assessment (WKPRECISE, 2009) the focus was on the detection of the variability of national level fishery statistics e.g. quantities landed, discards, fishing effort and cpue and biological data in terms of precision. Both workshops give guidelines for best practice on how to examine the accuracy of the data provided to the end-users. When such guidelines are implemented the steps in the data collection process that need improvement can be identified and an indication of the data quality can be achieved. In addition, the need to improve national sampling efforts and to ensure that national sampling schemes are based on statistically sound methods was the focus of a series of three workshops, the Workshop on Practical Implementation of Statistical Sound Catch Sampling Programs (WKPICS, 2011; WKPICS2, 2012; and WKPICS3, 2013).

The focus of WKBIOPTIM has been the sampling effort analysis of biological sampling and the quality of the biological variables under the optimization procedures carried out at the workshop. The optimization procedures discussed prior to and throughout the workshop should only be implemented when national sampling programs are statistically sound and recommendations from abovementioned workshops have been taken into consideration.

Two tasks were identified for this ToR which require feedback from each other and are very dependent on the end-users' needs: 1) Identify possible quality indicators that characterize the main biological properties of the samples (length, age and maturity) and 2) Evaluate the effects of statistical optimization of sampling. Both scripts were prepared using length as the biological variable to be analysed with the aim that they can be extended to other biological parameters.

The amount of time used at the workshop for discussion of this ToR was not as much as intended as time was spent introducing participants to the analyses carried out in the scripts, preparing data inputs originally in different formats (e.g. length frequency vs. length of individual specimens), debugging some coding errors, and running simulations of the case-studies. However, the subgroup working on this ToR carried out a literature search on the very broad topic of quality indicators and suggested that as a starting point several components of data quality need to be considered; precision, accuracy, representativeness, completeness, and comparability (GFCM, 2017). Precision is a measure of the proportion of agreement among replicate measurements; accuracy is the degree to which a recorded measurement varies from a true or expected value; representativeness is the extent to which measurements represents the true value in the population; completeness is the proportion of valid data collected with regards the

total expected and comparability is the extent to which data from different sources can be comparable. When using these components as quality indicators they can be used for classification of the sample, as a measure of comparison with an optimized sample and possibly even of the quality of the yearly national data collected.

We consider:

- 1 ) Completeness: a comparison of the achieved/optimized range to a complete range e.g. length classes;
- 2 ) Comparability: a comparison of the achieved/optimized range to the reference e.g. the full sample the previous year, the full sample;
- 3 ) Precision: a measure of agreement between the achieved/optimized sample and the true sample
- 4 ) Accuracy: a measure of the true value in the sample

We have not included representativeness because this refers to the population. If the indicators of quality (1 and 2) are not satisfied then estimating the statistical indicators (3 and 4) should not be carried out.

Depending on the end-users' needs different statistical tests can be applied to test for 1 and 2 above and this will be included in the future work to be carried out by the group. Included below are some suggested tests for age and length distribution comparisons as a starting point (the first of which is already included in the sample level simulations).

Two independent samples:

- a ) Kolmogorov–Smirnov test: non-parametric (no distribution assumption); compares shapes of two distributions; continuous data
- b ) R: `ks.test(x,y)` or `ksTest(x,y)` in package FSA
- c ) Mann–Whitney U-test (Wilcoxon rank sum test): checks for significant differences in location parameters (medians); non-parametric, based on rank sums, claims equal distributional shapes
- d ) R: `wilcox.test(x,y)`
- e ) Ansari-Bradley test: non-parametric; checks whether the dispersions (scale parameters) are equal (homogeneity); claims that medians are equal
- f ) R: `ansari.test(x,y)`
- g ) Moses test: test on extreme reactions, checks whether extreme values are equally likely in both distributions; non-parametric
- h ) `MosesTest(x,y)` in package

More than two samples:

- a ) Kruskal–Wallis test (location)
- b ) Fligner test (scale)

For the multilevel analyses the Earth Mover's Distance (EMD) was used. This is a measure of the distance between two probability distributions over a region  $D$  (names Wasserstein metric in mathematics) (Rubner, Tomasi, and Guibas 1998). Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region  $D$ , the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.



There was also a discussion among the subgroup around the use of CV as a quality indicator for indices requested by DCF as numbers by species (maturity at length, weight at length, sex-ratio at length and age at length). Here, the relevant questions for WKBIOPTIM are; how are the numbers by species defined and how can the sample optimization procedures be applied while ensuring quality of the requested indices (e.g. effective sample size)? WKBIOPTIM will be requesting input from both WGBIOP and WGCATCH on additional quality indicators for length frequency data and for other biological parameters and it is expected that there will be closer cooperation between the groups in the future on this topic.

## 4 Sample level analysis

---

**ToR b) Carry out hands-on work on code for statistical optimization of biological samples based on the CS and CA exchange format of the RDB and sampling strategy used to obtain the data. Code should be general and applicable to samples from different commercial sampling programmes and surveys. Different sampling effort strategies (e.g. fixed number, number dependent on size-span of the sample) and sampling strategies (e.g. simple random, two-stage stratified sampling) should be considered**

The following case studies were developed both during and after the workshop.

### 4.1 Sole in ICES division 7e (sol-echw) 2015 sampling

#### 4.1.1 Fishery description

The following description relates to the *S. solea* fishery in ICES division 7e (sol-echw) during 2015, as does the data provided to the workshop.

The common sole (*Solea solea*) is a demersal species of flatfish that resides upon sandy or muddy substrata, within which it submerses itself for protection. The maximum length of *S. solea* is around 70cm and its distribution ranges between Eastern Atlantic Ocean and the Mediterranean Sea; though, in winter, the common sole retreats to the southern North Sea due to its comparatively warmer waters. Common sole in the western English Channel (ICES division 7e) is primarily fished by beam and otter trawlers, accounting with 57% and 30% of the landings, respectively. Although is commonly caught as by other gears, such as netters and dredges (ICES, 2016). In 2015, the UK's fleet was responsible for the majority of *S. solea* landings in division 7e, landing 487 tonnes, followed by France with 243 tonnes (ICES, 2016). Discard data indicated that discarding in 2015 was relatively minor for the UK and Belgian fleets (Figures 8.3.3a and 8.3.3b of ICES, 2016). Occasional trips may show some discarding of sole below the minimum landings size.

In general, sole discards are low, total international discards averaged 2.6% of total catch weight in 2015. Discards comprised only 0.8% of UK catch and 0.6% of Belgian catch (ICES, 2016). Discarding from French fleets, however, was higher at 17% of the total catch with demersal trawlers providing the bulk of the discards below the minimum landings size (ICES, 2016).

#### 4.1.2 Sampling programmes

##### Offshore (Observer programme)

England and Wales run an offshore (observer) and onshore ('market') sampling programme to monitor commercial landings and discards of marine fish species. The observer programme is conducted by a random stratified selection of fishing vessels in England and Wales. Under the current sampling procedure, some vessels are excluded from the sampling frame:

- 1) Vessels less than 7m, for health & safety reasons
- 2) Vessels considered unsafe to take observers for reasons other than size.
- 3) Vessels specialising in fishing methods or target species for which a derogation has been granted:
  - a) Clam, oyster and cockle dredgers

- b ) Some pelagic vessels
- c ) Potting vessels
- 4 ) Vessels fishing from foreign ports or outside England and Wales. Vessels subject to bilateral agreements to be sampled in another country, where RCMs consider the métier is effectively sampled by another country.

Non-response rates are recorded. The list of vessels in the sampling frame is stratified by primarily by Coastal Region then by gear group and vessel size LOA (<10m; 10m+). In most regions the nets, trawls and lines are combined into single strata but in some regions where gear specific fisheries are more distinct Nets and Trawl are separated. Beam trawl and Scallop dredge vessels are also kept distinct but are not stratified by region.

### **Onshore (Market sampling)**

The onshore sampling programme targets all shellfish and finfish species landed into ports in England and Wales. Similar to the offshore programme, a regularly updated frame of fishing ports at which all or a defined proportion of the total landings are accessible at auctions, processors or other onshore locations, and from which a stratified random selection of ports and days is made for sampling. Ports excluded from the sampling frame:

- Very small ports, harbours, or other landing sites including beaches, where fish are disposed of locally rather than at auction, and where considerable effort would be required to sample very small amounts.

Ports where access has been denied are included within the frame to capture non-response rates and record the potential biases.

The list of ports in each of the three sampling frames is stratified by:

- 1 ) Quarter,
- 2 ) Region (6 strata) Lists of ports that map closely to ICES divisions, stock boundaries and fleet activities - 1Northeast, 2East, 3Southeast...
- 3 ) Port "size" based on the relative importance of that port within that region
- 4 ) Gear group although only 1 or 2 groups may be specific to the frame.
  - E1 - Demersal trawlers, netters + liners
  - E3 - Pelagic trawlers and seiners
  - E4 - Shellfish pot & trap vessels
  - E5 - Beam trawlers
  - E6 - Scallop dredgers

On the given day at the auction site a sampler will use a unique randomised list of numbers to select the trip to sample and then a similar list to sample the species available. Non-response rates and refusals are recorded. The vessel selection and species selection scheme is necessary in large auctions but at merchants and small markets staffs are often limited to sampling everything that arrives or is available at the time they are there, so the port and day selection process differs between regions.

### Adaptation of tables

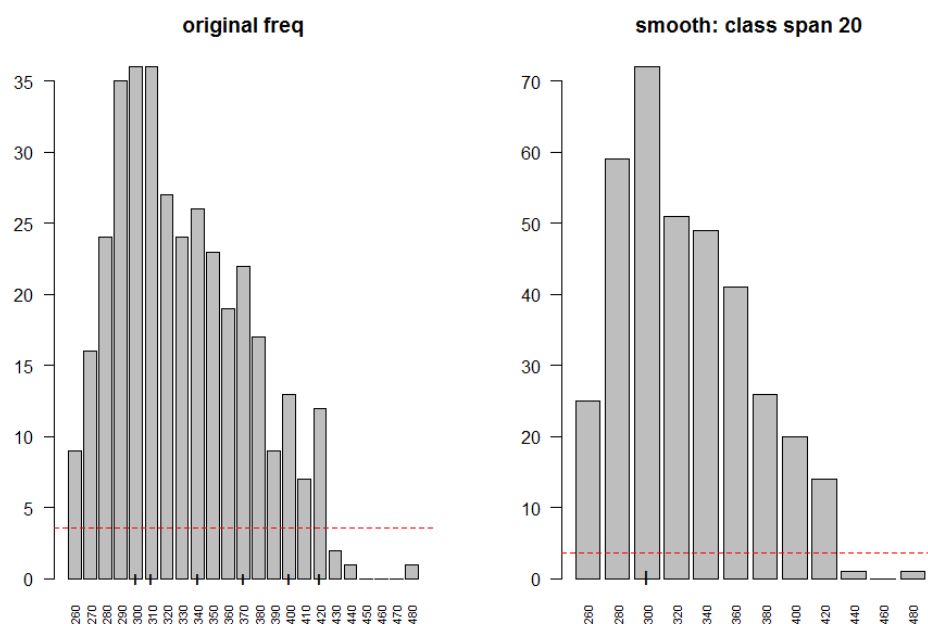
In the UK, only two fish from each size category are aged, therefore, the CA table does not contain all samples collected from a trip. In order to include all samples taken from a trip, data from the HL table was converted into the CA format; furthermore, as weight, maturity and ages are not stored in the HL table, only the impact on length data were examined.

#### 4.1.3 Interpretation of results

##### Original data

All results presented in this section pertain to *S. solea* sampled during trip code 15630 in the UK's market sampling scheme. In total, 359 *S. solea* were sampled during this trip.

The number of modes from the original length frequency data and a smoothed length frequency dataset (created by grouping the lengths into 20mm categories rather than 10) were calculated (Figure 4.1.3.1). Figure 4.1.3.1 shows that 6 modes were present in the original length class, which were: 300, 310, 340, 370, 400 & 420 mm; conversely, only one length frequency mode was found at 300mm when the data were smoothed (Figure 4.1.3.1).



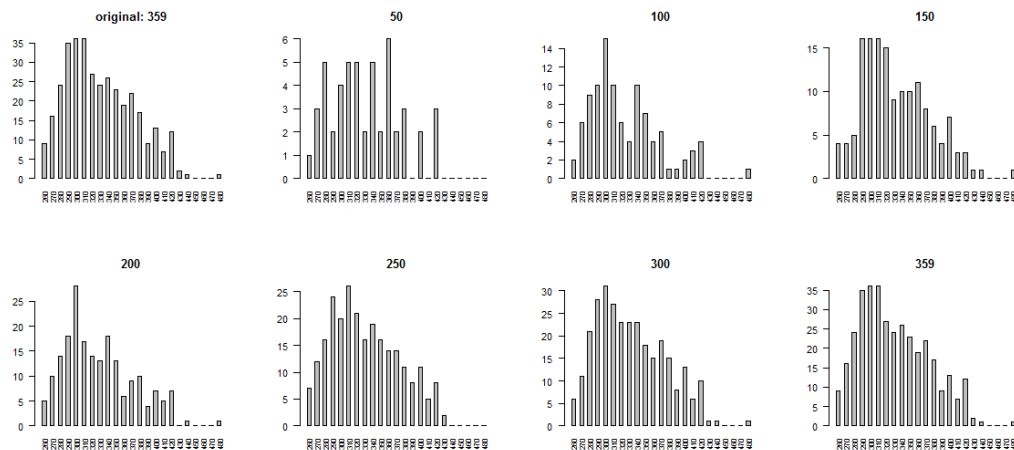
**Figure 4.1.3.1 - The modes present in the original (left) and smoothed (right) length frequency dataset. The smoothed dataset was created by grouping the length frequencies into 20mm categories rather than 10mm. Data taken from sole (*Solea solea*) sampled during trip code 15360 in the UK's market sampling scheme.**

### Reduced sample sizes

#### Length class distributions

Simulations were conducted to determine if reducing the original sample size of 359 to 50, 100, 150, 200, 250 and 300 samples would significantly impact the quality of the data. Visual analysis of the length classes present indicated that length frequencies present in the dataset would not significantly vary if the sample size was reduced to 300

as both the minimum and maximum length classes were present (Figure 4.1.3.2); moreover, while reducing the sample size to 250 produced a similar distribution, the two largest length categories would not have been captured in the sample (Figure 4.1.3.2). Furthermore, reducing the sample size to 200 samples and below significantly reduced the distribution of the data (Figure 4.1.3.2). Consequently, reducing the sample size to either 250 or 300 individuals may be plausible to maintain a similar quality of data.

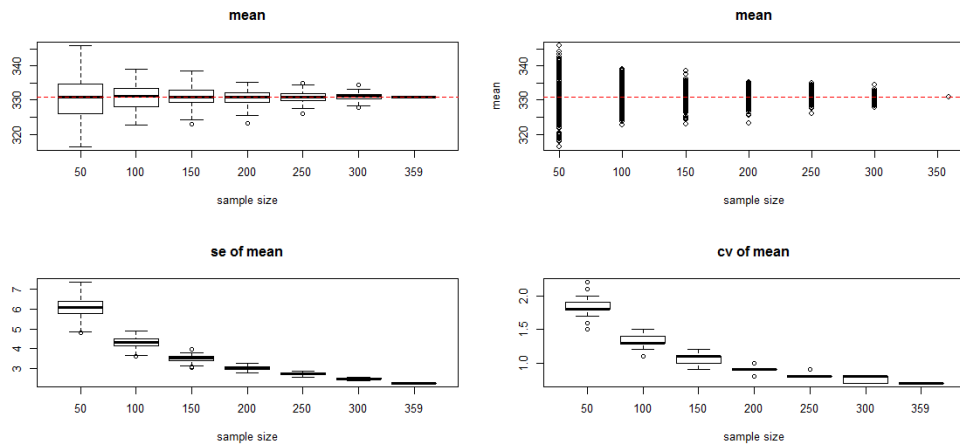


**Figure 4.1.3.2 - The distribution of length classes in the sample if the sample size was reduced.**

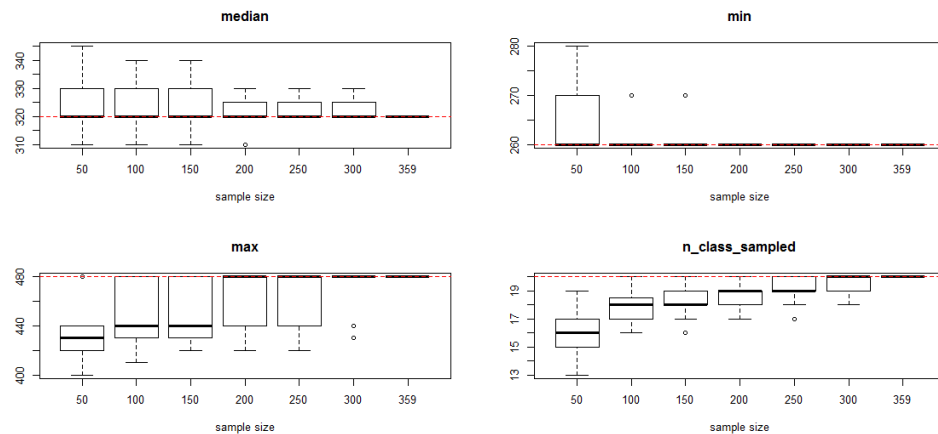
### Bootstrapping

For each reduction in sample size simulated (50, 100, 150, 200, 250 & 300 individuals; 359 individuals were present in the original data), 200 bootstrap samples of different individuals from the original samples were conducted. For each bootstrap simulation, the average length class, standard error (SE), percentage coefficient of variation (CE), median, number of modes (original and smoothed), the number of modes that match the original data (original and smoothed) (Figure 4.1.3.3–Figure 4.1.3.5); furthermore, t and Kolmogorov–Smirnov (KS) tests were conducted to test for statistically significant changes in the distribution of the data (Figure 4.1.3.5). Figure 4.1.3.3 indicated that variation in the mean & spread across the mean length from the original data were not large when the sample size was reduced to either 250 or 300, additionally, the smallest size class was represented in all simulations with 200 and above individuals. Conversely, the median value and number of size classes sampled were not always the same as the original data in any of the simulations, although the variation was minor (Figure 4.1.3.3). Except for two bootstrap simulations, when the sample size was reduced to 300 the maximum size class was always sampled (Figure 4.1.3.4), however, simulations with 250 individuals or less often did not capture the maximum size class present (Figure 4.1.3.4). The number of modes detected in all simulations displayed a high degree of variation, though, when the data were smoothed, the number of modes for all simulations, apart from where the sample size was 50, was always 1 (Figure 4.1.3.5). KS test P values for all bootstrap simulations in all reduced sample size simulations were above 0.05, although, the variation in p values was the lowest in sample size of 200 and above (Figure 4.1.3.5), though, it is worth noting the large number of outliers when the sample size was 200 (Figure 4.1.3.5). On the other hand, t-tests p values were always above or equal to 0.05 in sample sizes of 100 and above (Figure 4.1.3.5); although, a large spread across the mean t-test p value was observed in simulations of 250 and below (Figure 4.1.3.5). From these results, a reduction in sample size from 359 to 300 could be made without any significant risk in the quality of data; indeed, an

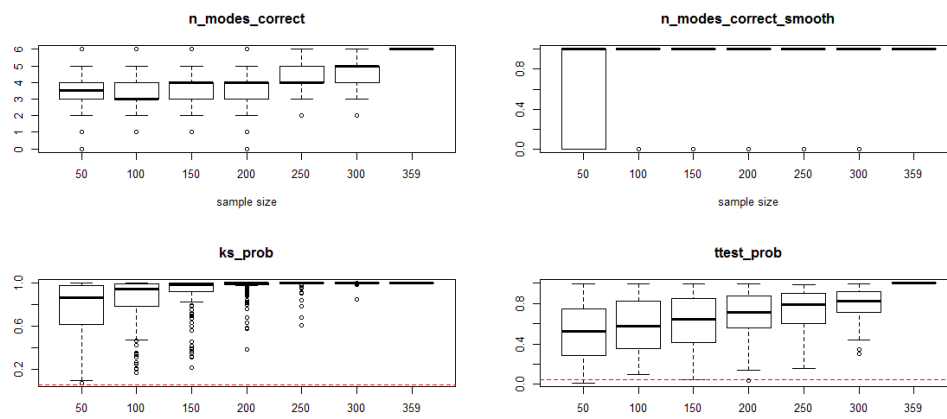
argument could also be made for reducing the sample size to 250, although the risk of reducing the data's quality is much greater.



**Figure 4.1.3.3 - The mean, standard error (SE) and coefficient of variance (CV) of lengths observed in 200 bootstrap samples when the sample size was reduced.**



**Figure 4.1.3.4 - The median, minimum, maximum and number of length classes observed in 200 bootstrap samples when the sample size was reduced.**



**Figure 4.1.3.5 - The number of modes and smoothed modes that were the same as those in the original dataset in addition to the p values of t and Kolmogorov-Smirnov and tests conducted on 200 bootstrap samples when the sample size was reduced.**

## References

ICES, 2016. *ICES WGCSE REPORT*, Copenhagen, Denmark: ICES, ACOM.

## 4.2 European hake in Central–Southern Tyrrhenian Sea (GSA 10)

The presented analyses have been carried out mainly to test if the sampling level scripts (used to investigate oversampling) could be adapted and applied to data from another context such as Mediterranean areas with peculiar fisheries characteristics.

European hake (*M. merluccius*) is one of a key species of the fishing assemblages in the central-southern Tyrrhenian Sea (GSA 10). It is a long living fish mainly exploited by trawlers (catches from trawlers are from a depth range between 60 and 500 m), especially on the continental shelves of the Gulfs (e.g. Gaeta, Salerno, Palermo) but also by artisanal fishers using fixed gears (gillnets, bottom longline). In the Figure 4.2.1 the landings of the species by gear of the last three years (2014–2016) are reported.

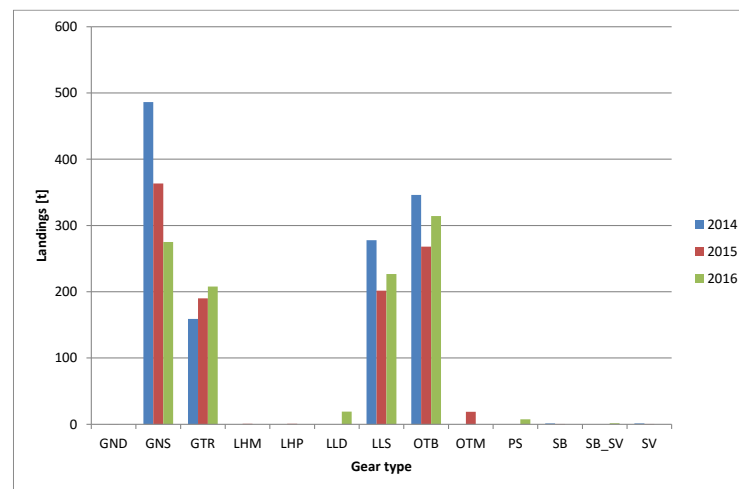


Figure 4.2.1 - *M. merluccius* (European hake) in GSA10 (Italy). Landings [t] by year and gear type

### 4.2.1 Stock unit

For the purpose of this case study the stock of European hake was assumed in the boundaries of the whole GSA 10, even if studies from STOCKMED project (Fiorentino *et al.*, 2014) suggested that the European hake stock in western area is combined on GSAs 9–11, 15–16 and 19. Advices on the status of the stocks have been given on GSA 9–10–11 combined (STECF, 2015).

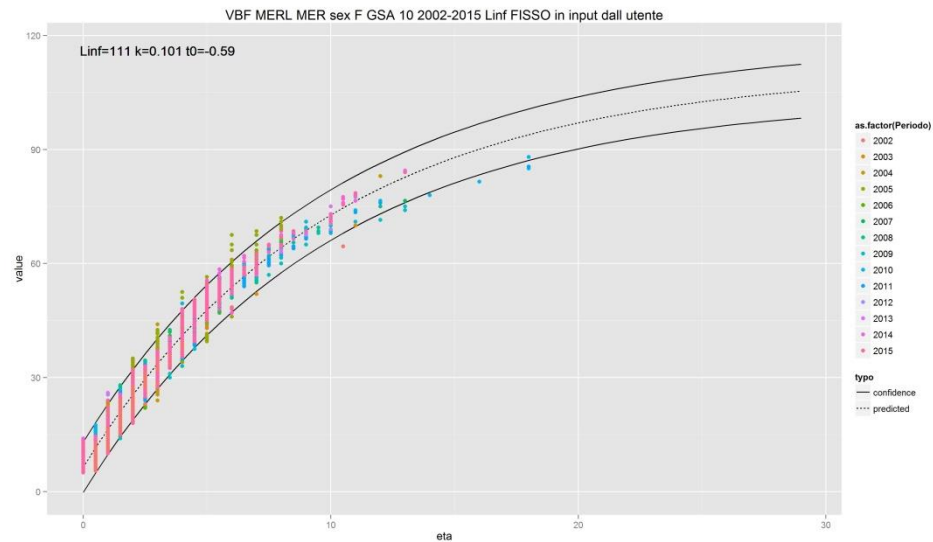
European hake is generally also ranked among species with higher abundance indices in the trawl surveys (e.g. Spedicato *et al.*, 2003). Trawl-survey data have corroborated highest biomass indices on the continental shelf of the GSA 10 (100–200 m; Spedicato *et al.*, 2003), where juveniles (less than 12 cm total length) are mainly concentrated. European hake is considered fully recruited to the bottom at 10 cm TL (from SAMED, 2002). The length structures from trawl surveys are generally dominated by small-medium size individuals, while large size individuals are rare. This pattern might be also due to the different vulnerability of older fish (Abella and Serena, 1998) beside the effect of high exploitation rates. The few large European hake caught during trawl surveys are generally females and inhabit deeper waters. The overall sex ratio (~0.41–0.47) estimated from trawl survey data are slightly skewed towards males. Larger individuals are caught by vessels using bottom longlines.



#### 4.2.2 Growth and maturity

Von-Bertalanffy parameters estimated by sex on age readings from DCF data are shown in Figure 4.2.2.1 (a) and (b).

a )



b )

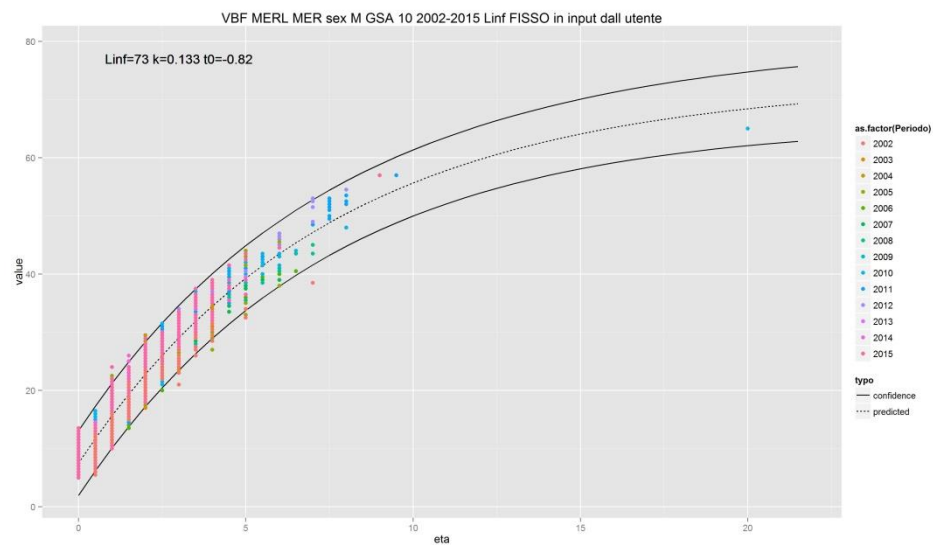
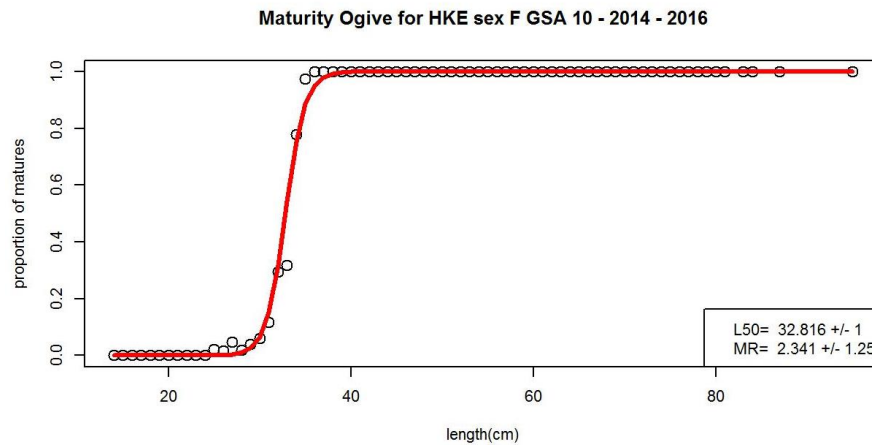


Figure 4.2.2.1 - Females (a) and males (b) VB growth curves of *M. merluccius* in GSA 10.

The observed maximum length of European hake was 95 cm for females and 65 cm for males both registered in the landings. The maturity at length of females is shown in Figure 4.2.2.2 (a) and (b).

a )



b )

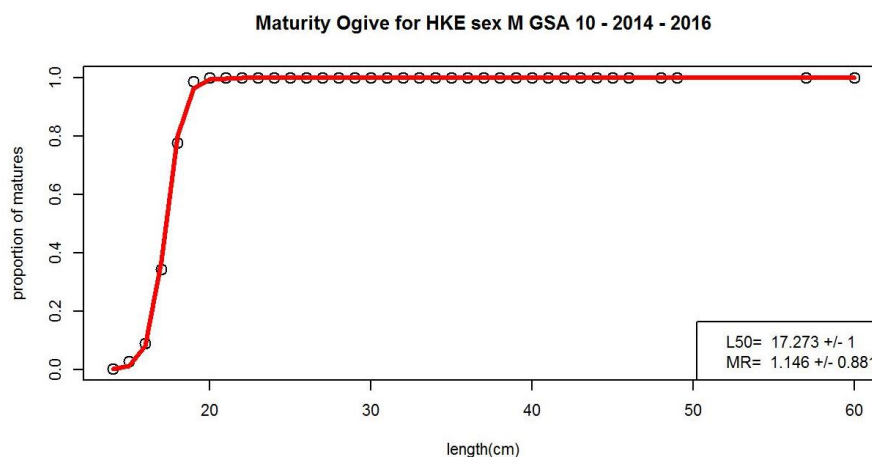


Figure 4.2.2.2 - Maturity ogive of females (a) and males (b) of *M. merluccius* (European hake) in GSA 10.

#### 4.2.3 Sampling strategy

Sampling strategies in GSA 10, is a mix of concurrency-at-sea (sampling directly on board by observers and scientists) and concurrency-at-landing site (sampling directly on landing site). The target population for a given year is the number of fishing trips by métier estimated on the basis of the number achieved in previous three years. Fishing trip is considered equal to the fishing day, considering the fishing habits in the area. The frame population was a subsample of the target population: it is a selection of fishing trips, mainly on spatial and time stratification basis (quarterly) with measurements of the composition of the catch in order to detect seasonal differences in the demographic structure and composition of the landings for the different métier. The sampling is accomplished according to the methods of a two-stage stratified random sampling: the sampling unit belonging to the métier (primary unit) was the fishing trip

(secondary unit). The number of fishing days to be sampled was defined proportionally to the effort (number of days at sea for each métier), and the landings (volume and value) of the previous three years.

#### 4.2.4 Data preparation

The dataset provided for the purposes of the working group contains raw sampling data of *M.merluccius* in GSA 10 for the period 2014–2016. The format used by the distributed scripts for analyses at sample level is the SDEF CA table format as it is quite well adapted to the lowest stage of sampling.

CA table provides information on the métier and quarter for each individual measurement. However 4 columns have been added to the original format for *trip date*, *month*, *quarter* and *métier*. *Year* and *trip code* were redefined to allow the scripts to work with length distributions by quarter and métier merging all the data of the considered period.

#### 4.2.5 Define sampling strategy and number of simulations

The analyses have been performed on data aggregated by gear (level4) grouping the data in three main categories (OTB, NETS and LLS). The aim was to perform the simulations reflecting as much as possible the real sampling stratification by quarter and gear and observe how the length distributions vary with the sample size in each gear. A sampling period of 3 years (2014–2016) (see Table 4.2.5.1 for sample overview) has been taken into consideration in order to integrate the interannual variability and to use bootstrap procedure without replacement to avoid recycling the same individuals when reconstructing length distributions.

Below the sampling options used for the simulations are reported.

```
sampling_options = list (n_sims = 100, stratified = FALSE, replacement = FALSE, sample_all_available = TRUE, sample_all_available_warning = TRUE, stages="one", samp_sizes = c(seq(30,130, by=10), nrow(dfl)), strata_var = "none", vars_to_keep = c("Length_class"))
```

The step size for the definition of the sample size vector has been defined on the basis of the lowest number of individual sampled by quarter for each gear.

**Table 4.2.5.1 - Number of individuals per sample by gear and quarter**

	N. OF INDIVIDUALS			
2014–2016	Q1	Q2	Q3	Q4
NETS	85	162	357	265
LLS_DEF	74	140	285	320
OTB	2161	4176	3949	5184

#### 4.2.6 Outputs

The simulations have been performed using the first version of the scripts (*teste2\_v8\_share.r*) in order to analyse one by one 12 samples obtained as the sum of all individuals measured in a given gear and quarter on the considered period 2014–2016, in order to have an “original sample” to be used for the bootstrap procedure that was representative for each considered sampling strata. The considered “samples” were: OTB\_Q1, OTB\_Q2, OTB\_Q3, OTB\_Q4, NETS\_Q1, NETS\_Q2, NETS\_Q3, NETS\_Q4, LLS\_DEF\_Q1, LLS\_DEF\_Q2, LLS\_DEF\_Q3 and LLS\_DEF\_Q4.

As expected, the “overall” length distributions for all the quarters with the same gear seemed to be quite similar in terms of number of classes and shape. As an example, the results from the 4<sup>th</sup> quarter are reported for each of the three considered gears in terms of:

- mean length and related coefficient of variation;
- median, minimum and maximum lengths and number of sampled classes;
- number of modes considering both the reconstructed distribution and the smoothed one
- original distribution used for the simulations.

The original distributions are assumed to be robust enough and representative of all the possible samples that could be caught by a gear, have been taken as reference for the interpretation of the results. For all the indicators and all the gears it can be observed that, after a certain sample size, the value of the considered indicators tends not to change significantly.

In case of OTB gear (Figure 4.2.6.1) 38 simulations have been run with sample size from 50 to 700 by a step of 20 individuals and equal to 800, 900, 1000, 1500 plus the last simulation that corresponds to the real sample size (5184). It can be observed that the confidence intervals of the mean length in the considered sample sizes (obtained by averaging the mean length of the distributions in the 100 iterations) seem to become smaller when the sample size is higher than about 350 individuals. The plots of CV, number of sampled classes, and median seem to give a quite similar signal, as the curve of CV tends to flatten after about 390 individuals and the number of classes sampled and median seem to be more stable too. The trends of the minimum and maximum sample length class are less stable, as in these classes very few individuals are present so they are not resampled at a high rate.

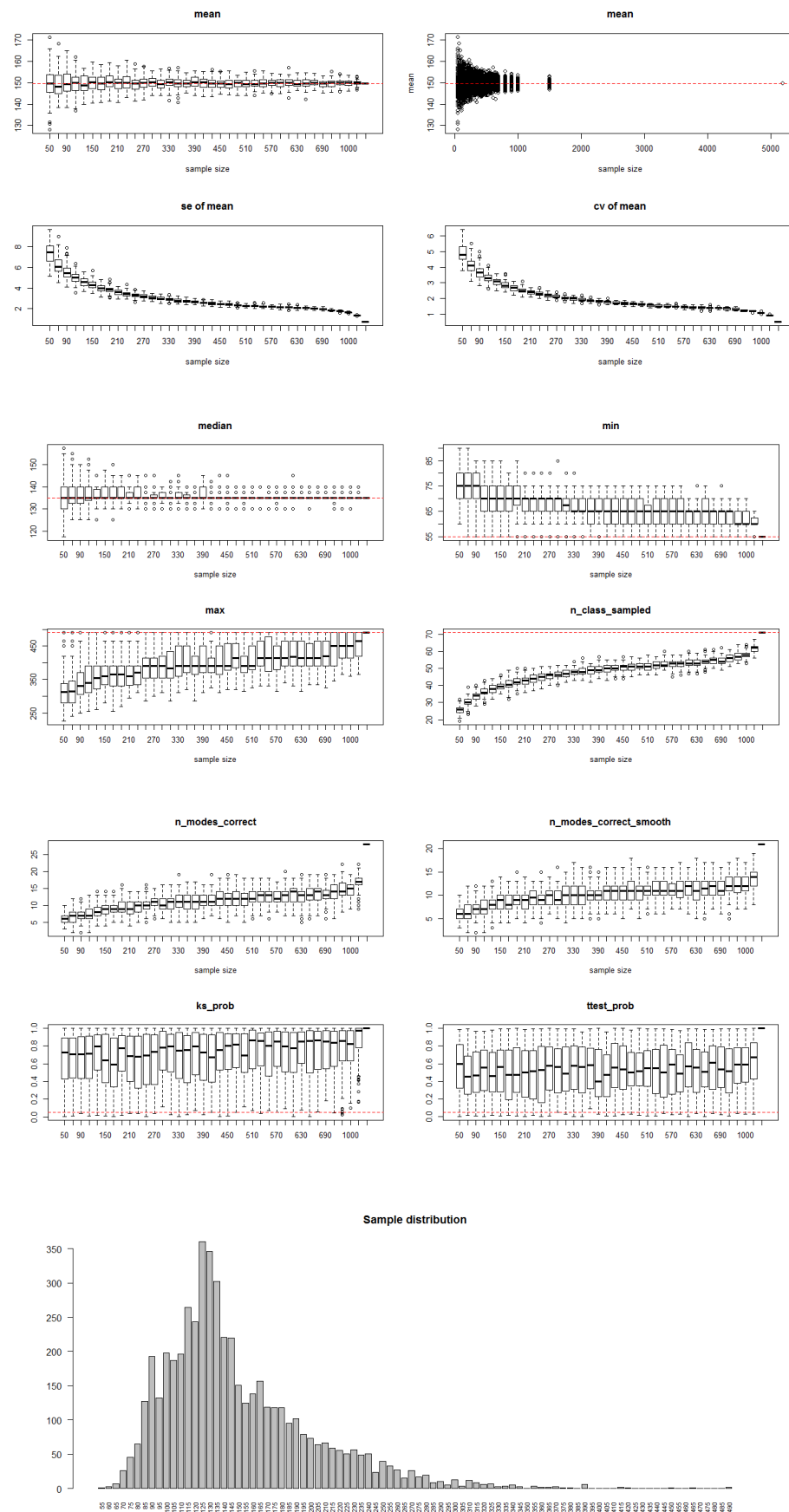


Figure 4.2.6.1 - Results of OTB gear in quarter 4.

For the NETS (Figure 4.2.6.2) 14 simulations have been run with sample sizes ranging from 30 to 150 by a step of 10 individuals and the last simulation equal to the real sample size (265). For LLS\_DEF (Figure 4.2.6.3) 12 simulations have been run with sample size from 30 to 130 by a step of 10 individuals and the last simulation equal to the real sample size (320). In both cases the initial distributions included a small number of individuals so that the maximum sample size could not be very high, given that the resampling was done without replacement. For this reason it is very difficult to identify a sample size that could be considered to be not very different from the original one as the simulation for sample size greater than the size of the reference distribution was not possible.

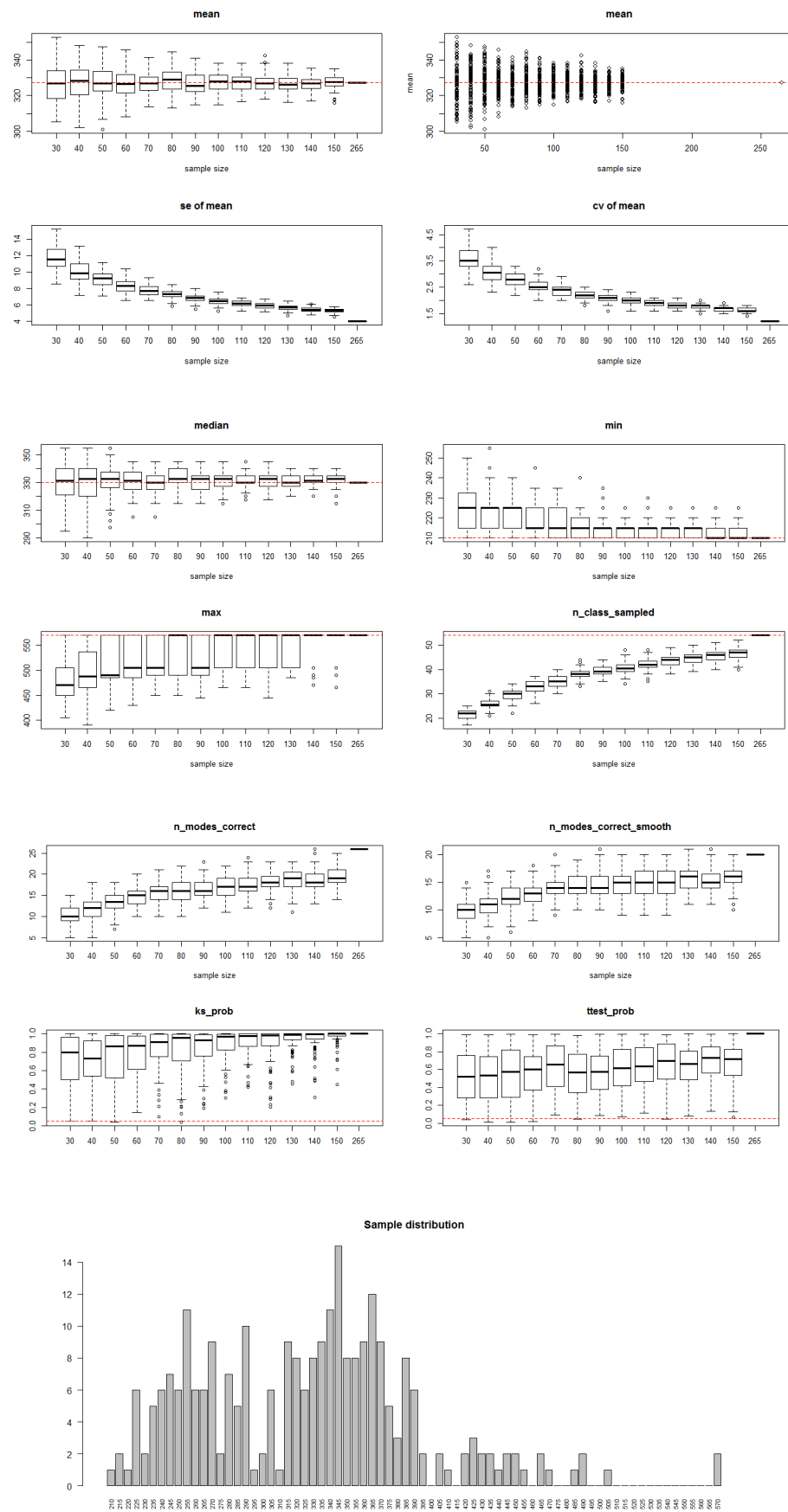


Figure 4.2.6.2 - Results of NETS gear in quarter 4.

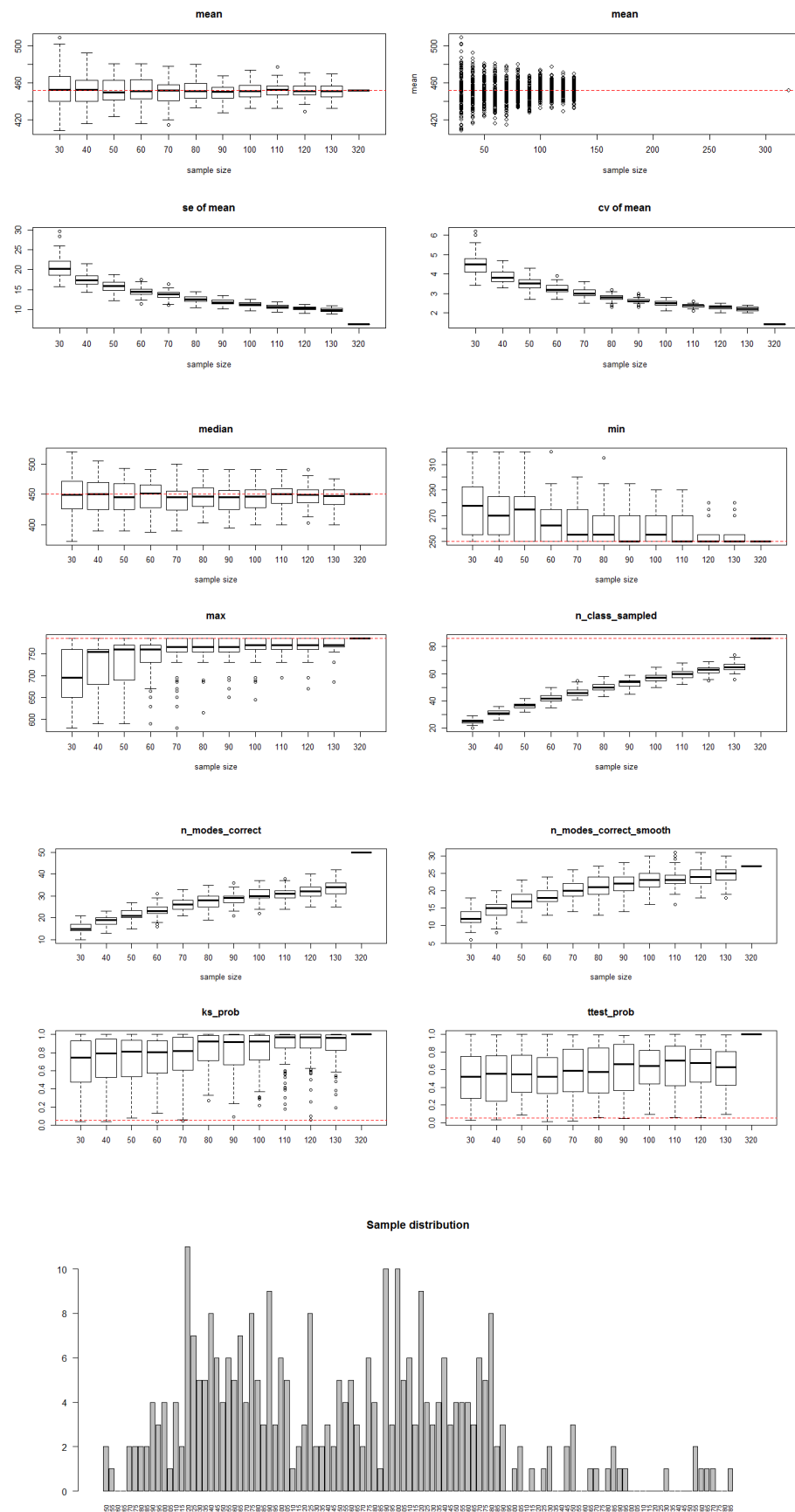


Figure 4.2.6.3 - Results of LLS gear in quarter 4.



## References

- Abella A.J., F. Serena. (1998) - Selettività e vulnerabilità del nasello nella pesca a strascico. *Biol. Mar. Medit.* Vol. 5 (2).
- Fiorentino F., E. Massuti, F. Tinti, S. Somarakis, G. Garofalo, T. Russo, M.T. Facchini, P. Carbonara, K. Kapiris, P. Tugores, R. Cannas, C. Tsigenopoulos, B. Patti, F. Colloca, M. Sbrana, R. Mifsud, V. Valavanis, and M.T. Spedicato, 2014. Stock units: Identification of distinct biological units (stock units) for different fish and shellfish species and among different GFCM-GSA. STOCKMED Deliverable 03: FINAL REPORT. September 2014, 310 p.
- SAMED, 2002 – Stock Assessment in the MEDiterranean. European Commission – DG XIV, Project 99/047 – Final Report.
- Scientific, Technical and Economic Committee for Fisheries (STECF) – Mediterranean assessments part 1 (STECF-15-18). 2015. Publications Office of the European Union, Luxembourg, EUR 27638 EN, JRC 98676, 410 pp.
- Spedicato M.T., Carbonara P., Lembo G., 2003. Valutazione delle risorse demersali dal Fiume Garigliano a Capo Suvero nel triennio 2000–2002. Relazione Finale GRU.N.D. (L.N. 41/82) U.O. 5, COISPA Tecnologia e Ricerca, Bari: 146 pp.

### 4.3 African black hakes in CECAF area

African black hakes, *Merluccius polli* and *M. senegalensis*, are two sympatric species that share distribution areas along different EEZ's in the East Atlantic.

The Spanish demersal trawlers are the main fleet targeting black hake in Northwest African fishing grounds. The French trawlers activity is more stable and this fleet operates mainly in Mauritania, alternating with other fishing grounds (Western Sahara and Senegal) sporadically. Between other practical and political reasons, due to the lack of biological knowledge, these two species are assessed as a unique stock in CECAF by global models. Reported catches are not separated and no age structure is considered.

#### 4.3.1 Sampling

The demersal trawlers activity in the area is monitored by different sources of information: i) vessels spatial and temporal location is registered regularly by satellite (VMS data); ii) efforts and catches are recorded both on-board in the logbooks and in the fish market, where the landings and dates are taken from the sales sheets; iii) length frequencies of landings are obtained by routine measurements (four samples per month) made by samplers, who make length samples of black hake by commercial categories from fresh trawlers in the Cadiz fish market and iv) other high quality and complementary information comes from scientific observers on-board these fleets. Commercial catches, bycatch and discards data are here collected on-board by haul, with accurate geographic information. Also individual size and weight as well as biological data are collected on-board.

The time spent by the observer during a trip is distributed between different tasks covering the data collection of biological information and discard composition of different species. Unlike the other biological sampling of species in ICES area, sampling on board is limited and is the unique source of information to estimate biological parameters. Previous knowledge indicates that the number of samples is adequate and a reduction in the number of individuals measured in each trip possibly is not recommended.

#### 4.3.2 Data

Biological and commercial information is managed through excel files in a similar format than RBD, but without a standardized protocol defined to exchange files for scientific analysis. One of the results of the first WKBIOPTIM is to inform to the CECAF IEO team about the advantages of this tool and to assist in its implementation to organize information in the exchange format RDB.

In this case study, biological information from observers of two trips from commercial vessels in 2016 has been used to run the script at sample level with 282 registers. Variables from observer files have been extracted in a .txt file with the columns:

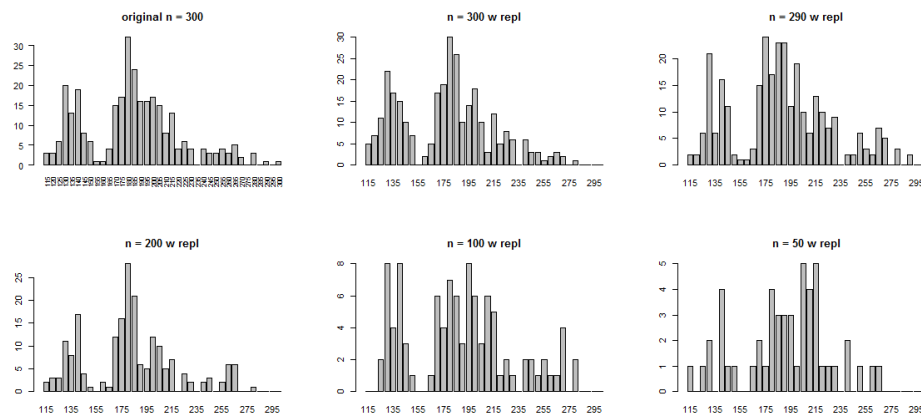
(Date ,Trip, Fishno, Length\_class, Weight, Sex, Maturity\_stage)

where each row corresponds to one fish measured.

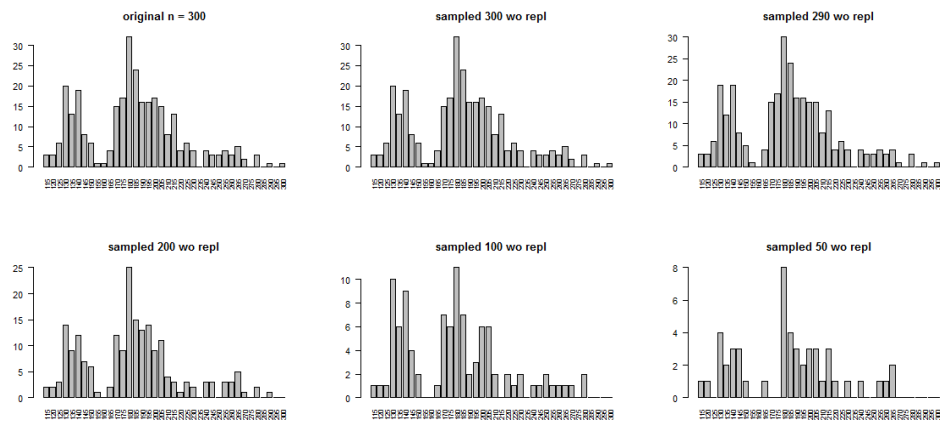
#### 4.3.3 Output

Script "sample\_level\_test\_script\_20170612" was applied to these data to simulate the number of individuals measured to characterize length distribution of black hakes. Some of the plots created for different sample sizes and bar class sizes are shown:

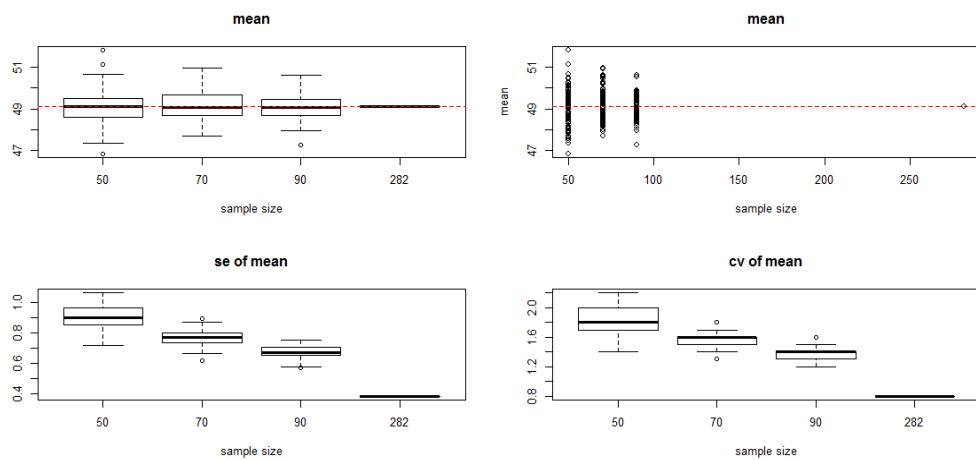
a)

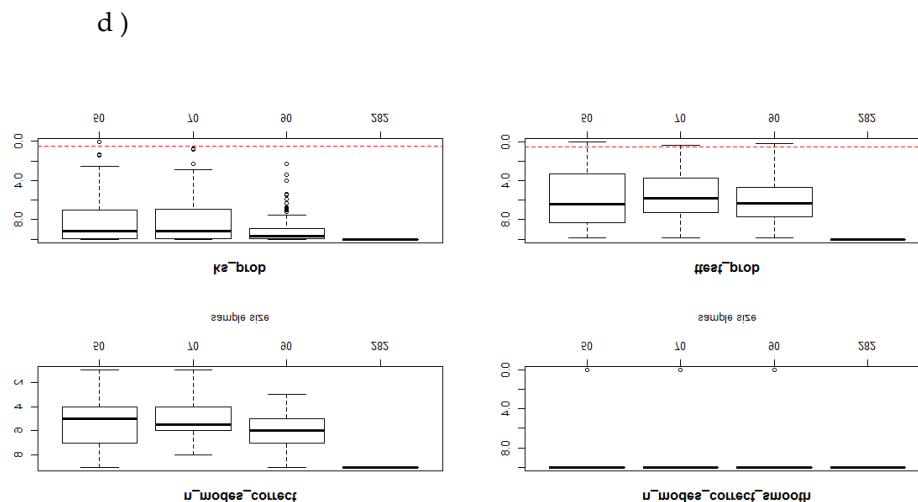


b)



c)





**Figure 4.3.3.1 - Length distributions of black hake for different sample sizes (a and b) and mean length, related SE and CV and by sample size, number of modes considered and statistical output by sample size (c and d)**

Figures 4.3.3.1 a) and b) show length distribution for different sample size of black hakes with and without replacement, respectively. Length patterns and important modes are captured with sample sizes very similar. Low sample sizes distort real values both with and without replacement sampling. Figure 4.3.3.1 c) shows statistics of the distribution of the mean for different sample sizes. Figure 4.3.3.1 d) shows different statistics for the number of coincident modes between the obtained for different sample sizes and the total sample.

Results can't be used to conclude any recommendation about the reduction/increase of samples size, as the data used for this exercise are very limited. Nevertheless, extended files will be used to replicate the example testing other biological variables taking into account additional information from observer data.

#### 4.4 Danish case study with sandeels (*Ammodytes marinus*)

##### 4.4.1 Sampling design

Samples are taken on board the fishing vessels as part of our self-sampling program. A document is provided with instructions on how the samples should be taken depending on what area the vessel is fishing in; 1 sample per day in area 1r and 2r and 1 sample from every haul in areas 4 and 6. A table is provided where vessel's name, skipper and port number should be provided. In addition; gear type used, logbook number, haul start and end position, date and time are requested plus the estimated total weight of sandeel per haul. A random sample of 1–2 kg of sandeel should be taken to ensure that no specific size class is over or under represented. Samples are stored in plastic bags with all the relevant information and frozen as quickly as possible. In principle all Danish fishers having a license for fishing sandeel should self-sample and 1 sample per vessel, day and square was worked up in the lab in 2015.

Besides the self-sampling program we receive samples taken by the control onshore – in 2015 we worked up 51 of these samples. The selection of samples is based on the control's protocol and we primarily collect these samples to have 'fishers' independent samples. These samples are not included in this study, since the design of that program is very different from the self-sampling program.

In the lab the sample is sorted and species other than *Ammodytes marinus* are removed, these individuals are measured, counted and weighted. Approximately 150 sandeel (*Ammodytes marinus*) are measured and sorted into semi-centimetre length groups and the remaining fish in the sample are weighed. The number of fish in each length group is counted and all fish in that length group are weighed. 5 fish are taken from each length group and for 3 of these fish otoliths are extracted and age determined, the 2 remaining fish are discarded.

The input data to the model used for the sandeel assessment is catch-at-age (CANUM) stratified per week and ICES statistical rectangle.

### Overview of sampling

In 2015 Denmark sampled 167 trips targeting sandeel and on average 2-3 hauls where sampled per trip – resulting in 445 hauls sampled. The samples may contain species other than *Ammodytes marinus*, but on average 145 *Ammodytes marinus* where sampled per haul – given a total of 64,944 sandeels sampled for length, see Figure 4.4.1.1 for the distribution of number of sandeels sampled per haul.

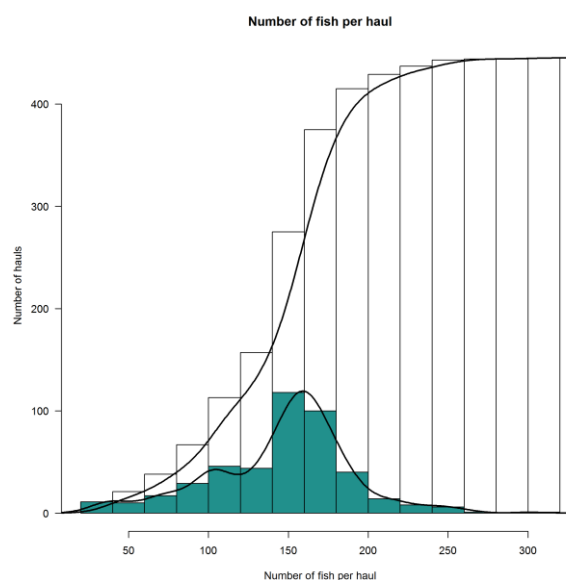


Figure 4.4.1.1 - Histogram showing number of hauls per number of fish sampled per haul (green) and the accumulated distribution (white).

### Problems with present sampling design

Way too many samples – minimise number of trips/ hauls sampled or minimise number of trips/ hauls sample in a combination with number of fish sampled per haul.

In the laboratory it is large numbers of fish sampled for length and then those which are selected for aging which contribute to the work load when working up samples of this species. All fish in the subsample are length measured, divided into length groups and weighed, 5 fish are retained per semi-centre group, 3 of these are dissected, the otoliths extracted and age determined. When this process is repeated over the number of trips and hauls sampled within a short fishing season it becomes a large volume of work. The samples need to be collected from the ports and in addition to the physical working up of the samples in the laboratory there is the paper work which is required when documenting both the vessel, trip and haul data for each sample while double

checking with vms and logbook data that no mistakes have been made aboard the vessel. When considering how resources could be saved in the lab it would be a combination of reducing the number of fish which need to be length measured, weighted and aged combined with a reduction of the number of trips and hauls which are sampled.

#### 4.4.2 Sample level – Looking at number of fish sampled per haul

##### Setup for simulations

Sample id: haul

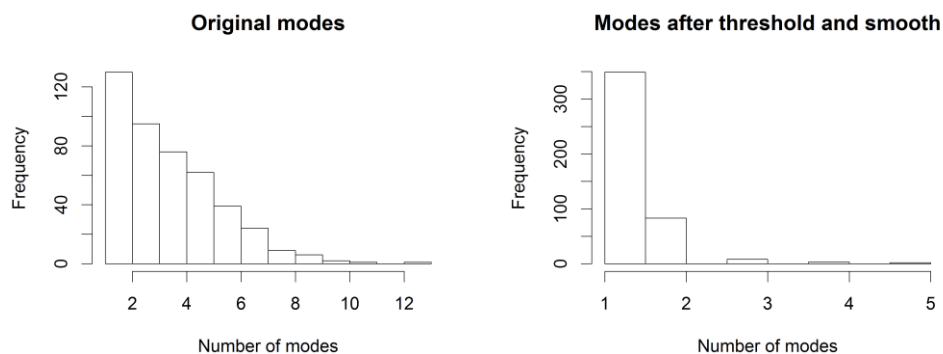
Variable: Length

Original class span: 5 (mm)

Smooth class span: 10 (mm)

Threshold for modes:  $0.025 \times \text{times number of fish in the original sample}$

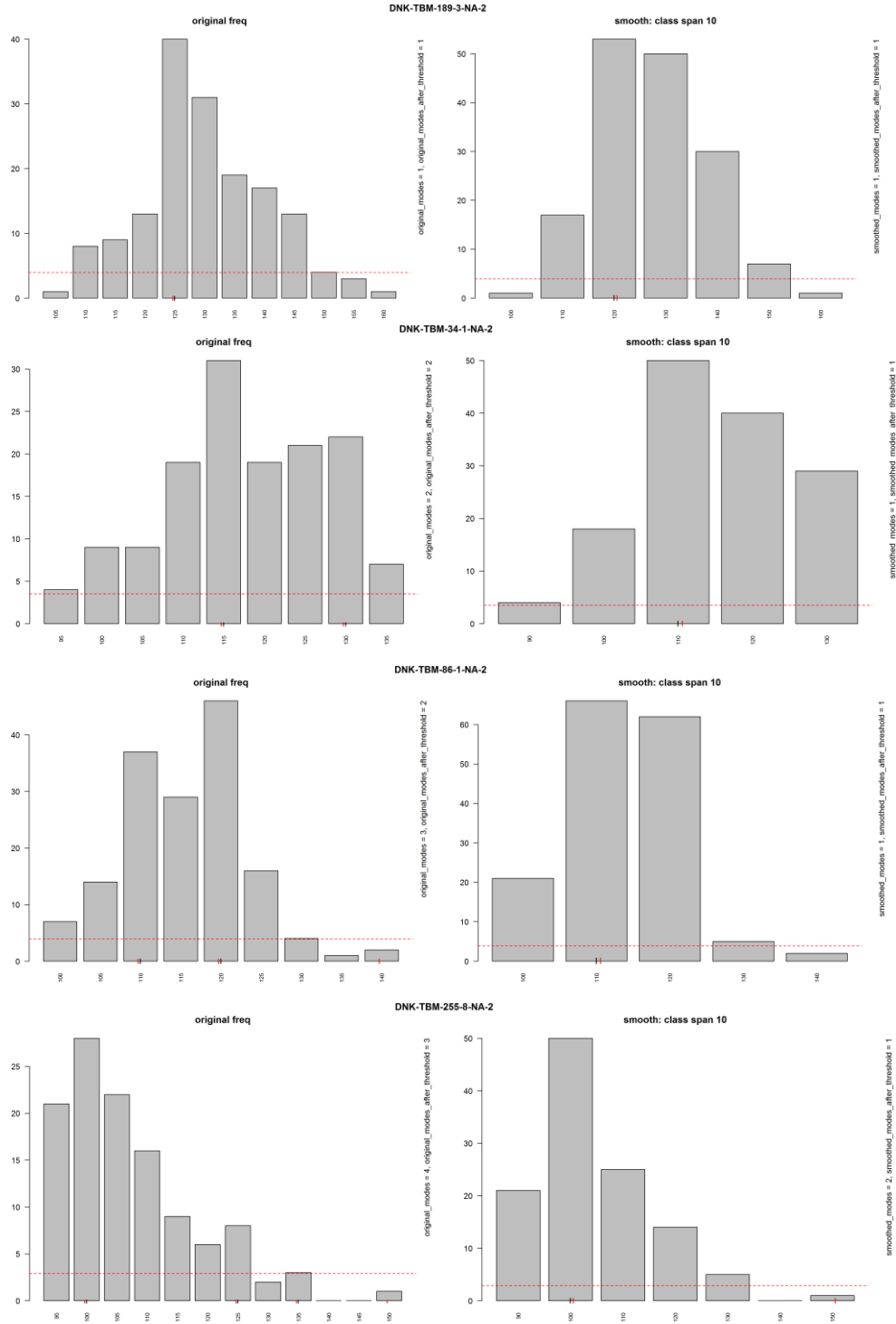
```
sampling_options <- list(n_sims = 500, stratified = FALSE, replacement=TRUE, sample_all_available = TRUE, sample_all_available_warning = TRUE, stages="one", samp_sizes = c(seq(20,300, by=20)), strata_var = "none", vars_to_keep = c("Length_class"))
```

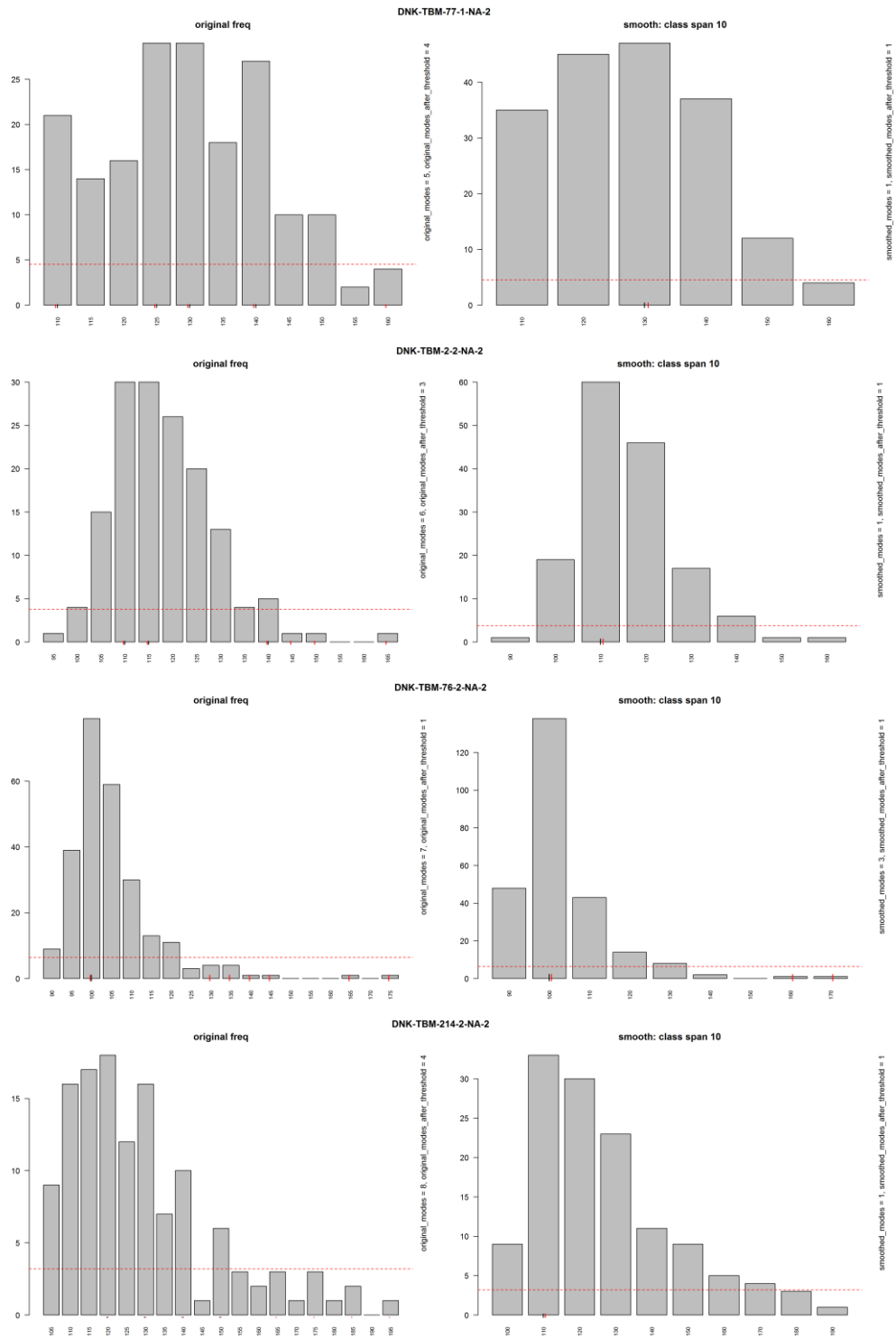


**Figure 4.4.2.1 - Histograms showing number of modes per sample**

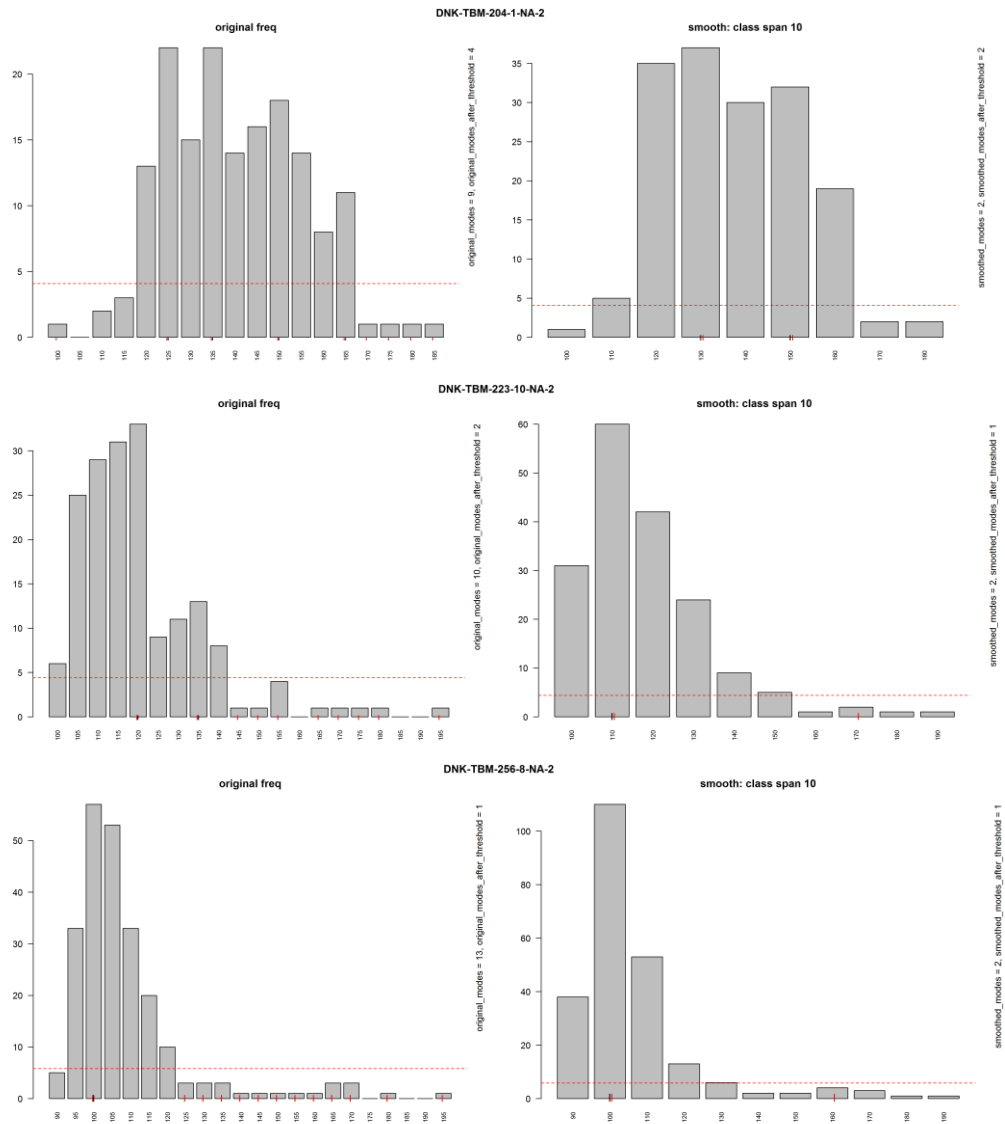
When applying the 'mode'-function to the original data most of the samples have a lot of modes, see Figure 4.4.4.1. A lot of these modes are due to a long tail without any eye catching modes – a relatively small amount of samples show distribution with more than a single mode. Therefore a threshold of  $0.025 \times \text{times number of fish in the original sample}$  was introduced. Further a lot of 'false' modes show up when looking at the distribution with the original class span (1 semi centimetre), therefore a smoother – 1 centimetre - was introduced as well, see Figure 4.4.2.1 for results.

Examples of the results of finding the modes per sample are given in Figure 4.4.2.2.









**Figure 4.4.2.2 - Examples on the results from the mode finding function. 1 random sample selected per number of original modes. Bar plot on the left showing the original modes with (black mark) and without threshold (red mark). Bar plot on the right showing the modes after introducing a smoother.**

#### 4.4.3 Results

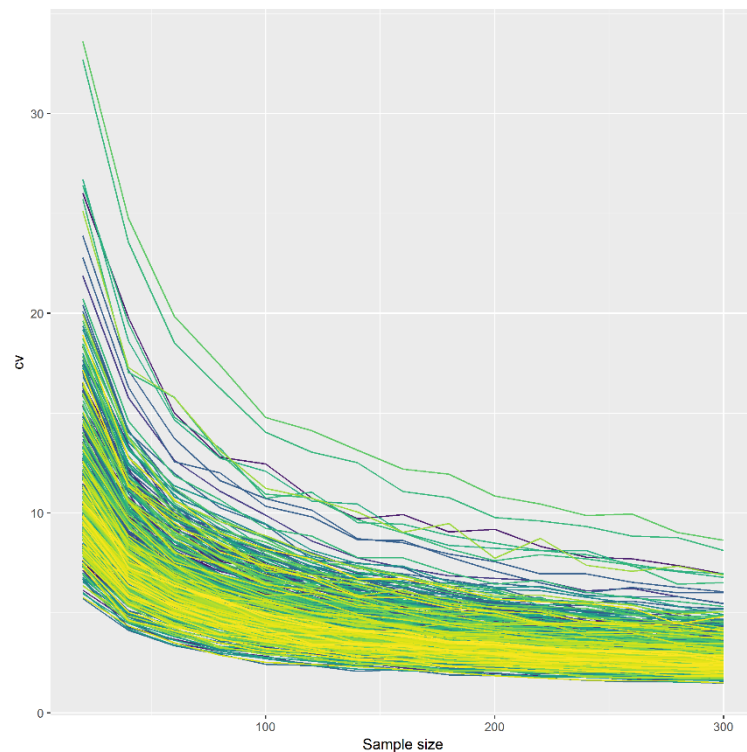


Figure 2.4.3.1 - CV of the mean length per sample size and sample. Each line represents a sample.

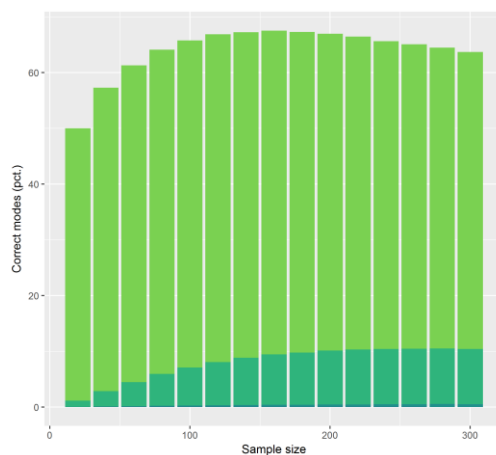


Figure 3.4.3.2 - Number of simulated length distributions with the correct modes (pct.) per sample size and number of modes in the original distribution. All distributions with an added smoother and threshold.

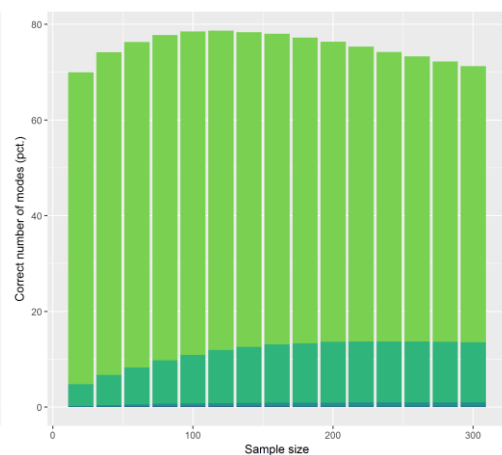


Figure 4.4.3.3 - Number of simulated length distributions with the correct number of modes (pct.) per sample size and number of modes in the original distribution. All distributions with an added smoother and threshold.

**Note:** The percentage of simulations with correct modes and correct number of modes starts to go down at high sample sizes – this is due to the fact that the threshold is scaled accordingly to the number of fish in the original sample – and not the number of fish in the bootstrap sample -> a mode is accepted at a lower threshold for samples with lower samples size than the ones with high samples size.

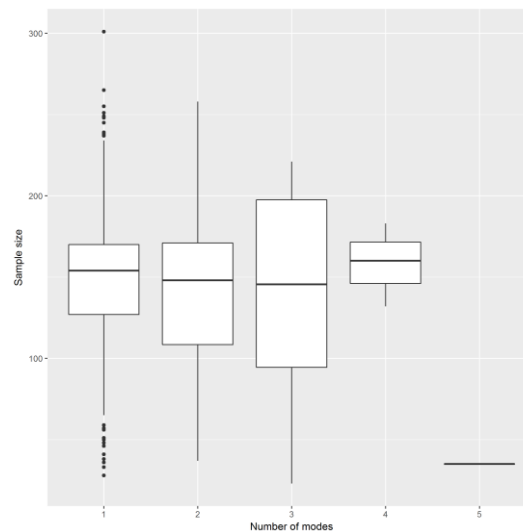


Figure 5.4.3.4 - Relation between number of modes and samples size

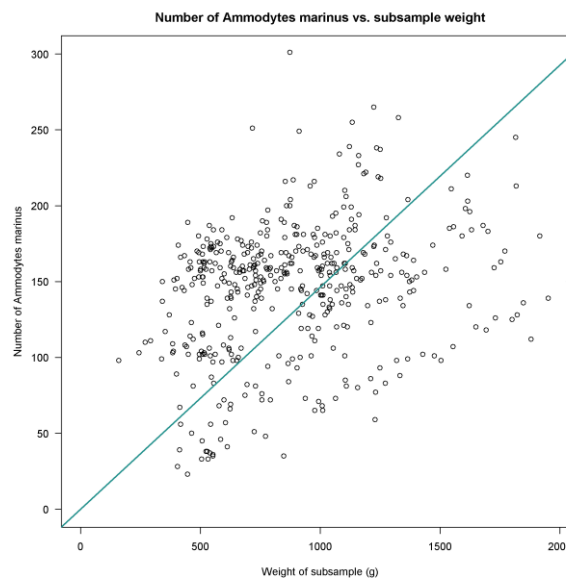


Figure 6.4.3.5 - Number of *Ammodytes marinus* vs. subsample weight. The size of a sample is normally determined by weight and not number. Some samples have a high percentage of other species e.g. *Ammodytes tobianus*, which to some extent is area dependent (not shown).

#### 4.4.4 Discussion and conclusion

In this study we have used 445 samples, so it is a bit tiresome to go through the results sample by sample. The summaries shown hopefully give a valid insight into the underlying samples. We have only looked at a couple of the indicators developed for this WK-CV of mean length and the number of modes. Number of modes seems relevant, since distribution with more than one mode could indicate two distinct year classes in the catches.

If the purpose is to maintain a length distribution as close to the one observed in the original sample, then a sample size of 100 seems to give a good estimate of the mean length and number of modes, see Figures 4.4.3.1 and 7.4.3.3. Number of modes in the length distributions do not seem to be particularly related to number of fish at the samples sizes currently collected, see Figure 8.4.3.4.

#### 4.4.5 Future work

This study looks into maintaining the properties of the length distributions for a single sample when lowering the sampling size. In respect to sandeel - what we really want is to optimize the complete matrix of vessels, trips, hauls and number of length and age measurements and use the estimated length/ age distribution and associated variance per both area (sandeel area) times month and square times week to compare different designs (stratification and number of SU's), the latter resolution being the one used in the model. One way of obtaining this would be to combine the 'multi-level' and 'sample' methods developed at this WK – estimate number at length/ age, compare the variance around the estimates and use the methods developed for single samples to evaluate the estimated distributions. All samples used in this study come from the same known design, so it would be possible to come up with valid estimates. It would also make sense to make a variance-component analysis a la Pennington and Helle (2011) to pinpoint the main sources of variation.

The optimization should take into account; the time spent on the different tasks e.g. picking up samples, measuring and weighing the fish, ageing, quality assurance of the data, entering the data into the database and otolith archiving, since lowering the number of length measurements from e.g. 200 to 100, but keeping the number of trips and hauls sampled constant may not lead to a large reduction in lab time. Further analysis should include simulations which optimise for a combination of both length and age.

#### References

Pennington, M., and Helle, K. 2011. Evaluation of the design and efficiency of the Norwegian self-sampling purse-seine reference fleet.–ICES Journal of Marine Science, 68: 1764–1768.

## 4.5 Adopting Optimization Approach for Commercial Sample to Survey Data

### 4.5.1 Data

The data used to test the proposed sample level algorithm was the North Sea - International Bottom-trawl Survey Quarter 1 in 2005 (NS-IBTS Q1 2005) and a single haul (HaulNo 38) for *Clupea harengus*. The data were obtained from the DATRAS database and was transformed in the RDB format.

The NS-IBT Surveys currently takes place twice a year, in Quarter1 (Jan-Mar) and Quarter3 (typically late Jul–Sept). The aim is to collect data on distribution, relative abundance and biological information on a range of species in ICES Index areas. There are six nations participating in NSIBTSQ1: Norway, Denmark, Germany, Scotland, England and Sweden, and the sampling scheme is based on two nations surveying each statistical rectangle in an ICES area.

### 4.5.2 Results

Table 4.5.2.1 below gives the length classes and the number of measured fish in Haul No 38. Figure 4.5.2.1 shows that there are 18 length classes and 6 modes. Modes are located at classes: 1 (125 mm), 6 (150 mm), 17 (205 mm), 20-21 (220 & 225 mm) and 24 (240 mm).

**Table 4.5.2.1 - Length classes for *Clupea harengus* in Haul No 38 in NSIBTS Q1 2005, and the measured number of fish in each length class.**

LENGTH CLASS (MM)	MEASURED NUMBER
125	1
130	0
135	11
140	21
145	52
150	65
155	55
160	53
165	33
170	15
175	11
180	11
185	6
190	3
195	2
200	0
205	2
210	0
215	0
220	1
225	1
230	0
235	0
240	1
Total	344

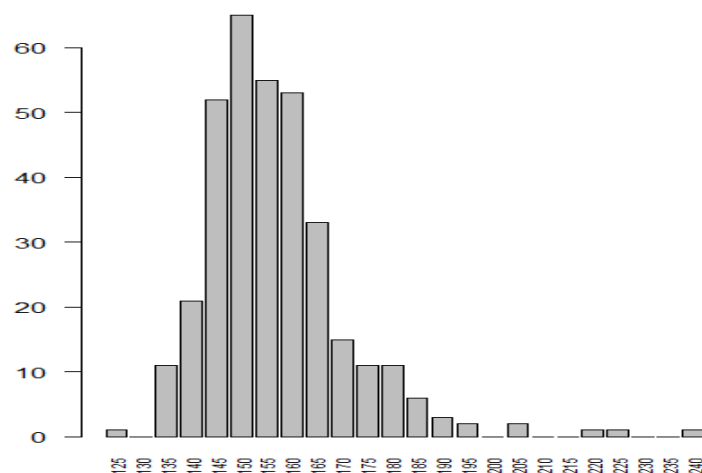


Figure 4.5.2.1 - Histogram of number measured by length class

### Should We Take a Smaller Sample?

The length distributions in the 6 cases showed in Figure 4.5.2.2 seem similar but as the number sampled decreases the modes/peaks change. Note that data are from a single haul/station in year 2005. The results in Table 4.5.2.1 above show that 47% of length classes have  $\leq 6$  records and 29% have 1 record, therefore, sampling with replacement may result in possible selection of the same record, and hence, imprecise estimates i.e. high variance of parameter estimate. In this case, a recommendation of a sample of 300 individuals seems plausible to maintain shape and other relevant statistics.

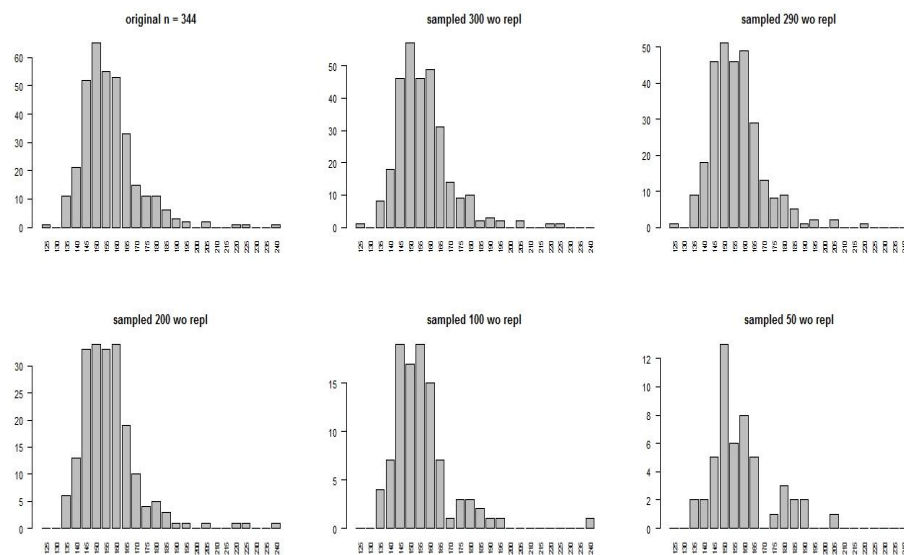
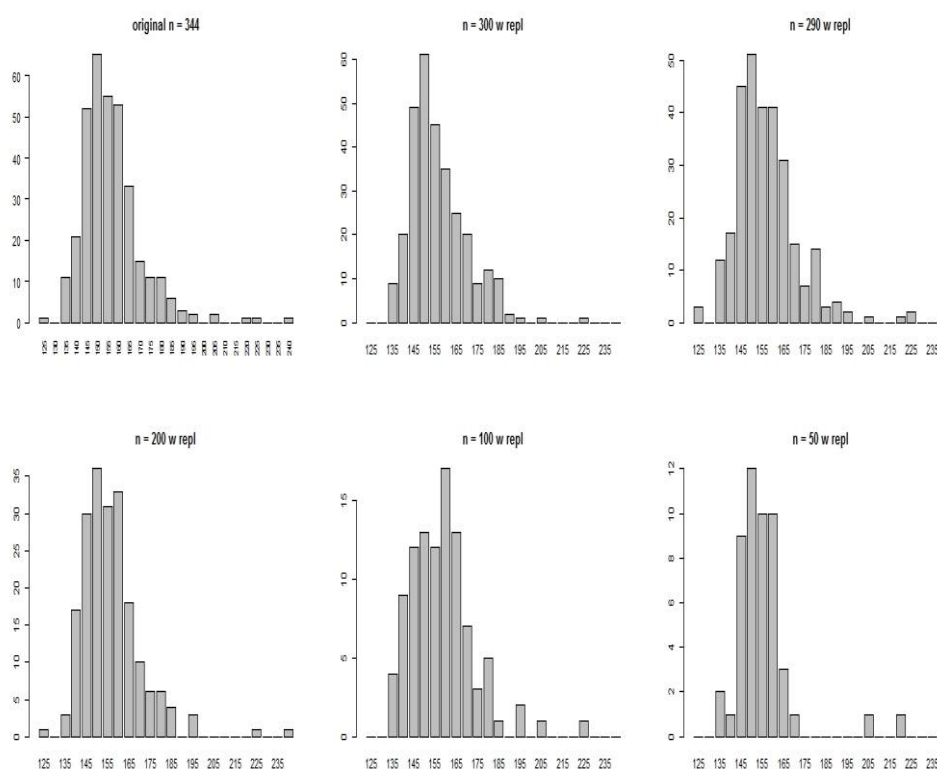


Figure 4.5.2.2 - Random samples of measured individuals of *Clupea harengus* selected without replacement (wo repl)

Similarly, the occurrence of modes differs when sampling with replacement. In this case, the distributions for samples of 300 and 290 seem similar to the original sample of 344 individuals; therefore these can be used in the analysis process.



**Figure 4.5.2.3 - Random samples of measured individuals (*Clupea harengus*) selected with replacement (w repl)**

### Bootstrap resampling

The number of individuals measured was bootstrapped to determine the optimal sample size. Stratification was done with length class as parameter. The number of Bootstrap simulations was set to 100 and possible sample options,  $n$ , of measured individuals were 50; 100; 150; 200; 250 or 300. So, for example, when  $n=50$ , we would have 100 simulations (repeats) of random samples of 50 measured individuals in the randomly selected length classes. Table 4.5.2.2 shows the first 100 simulations for  $n=50$  where most individuals selected were from length class 160 mm and only 1 individual was selected from length classes 135 mm and 140 mm.

**Table 4.5.2.2: Example of randomly selected measured number of individuals for the first of 100 simulations where  $n = 50$ .**

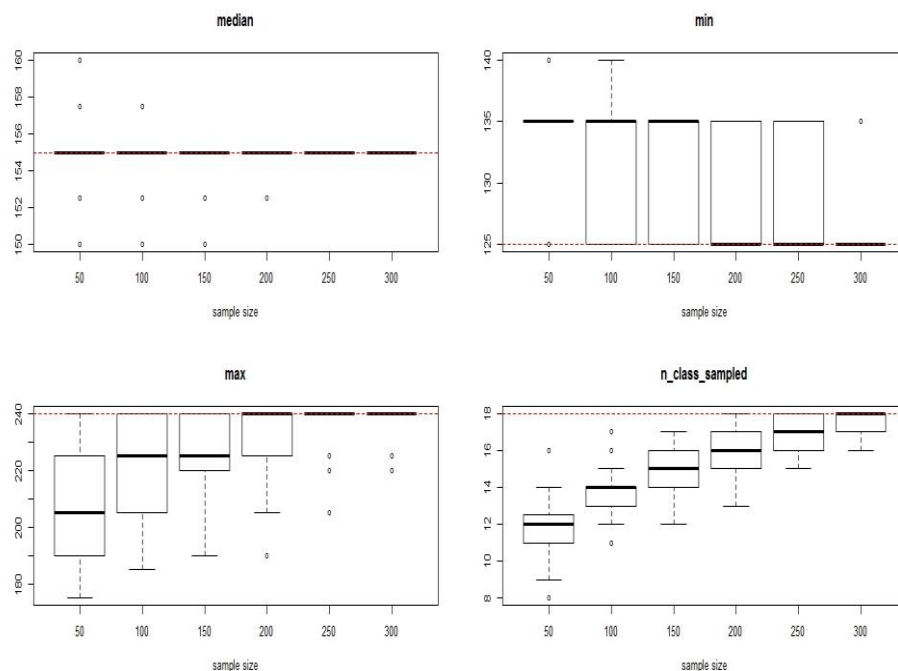
LENGTH CLASS	135	140	145	150	155	160	165	170	175	180	Total
NUMBER MEASURED	1	1	5	11	6	13	4	3	4	2	50

For each of the 100 simulations, the average length class, standard error (SE), coefficient of variation (CV), expressed as a percentage and the median length class, were computed. The overall means for average length classes and average standard error (SE) were also computed. Table 4.5.2.3 gives the mean length class, SE, median and CV for the first 6 simulations when the number of measured individuals was  $n=50$ . It also shows the average of the mean length classes and the average standard error (SE) of the 100 simulations.

**Table 4.5.2.3: The estimated mean length class, standard error (SE), median and coefficient of variation (CV) for the first 6 simulations and the overall means from the 100 simulations.**

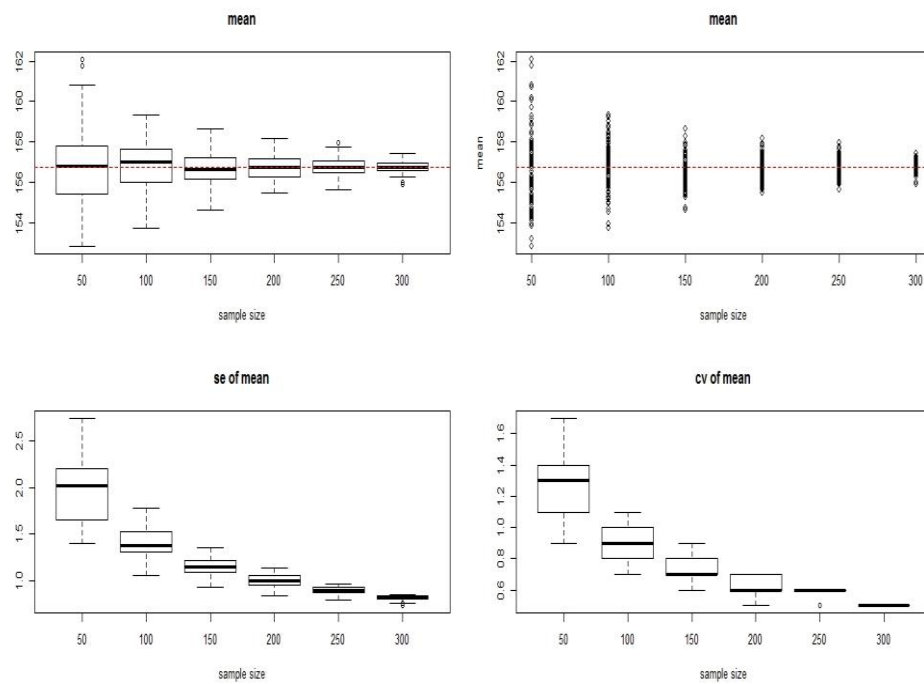
NUMBER	MEAN LENGTH CLASS	SE	MEDIAN	CV (%)
1	157.6	2.21	155	1.4
2	155.4	1.84	155	1.2
3	154.4	1.48	155	1.0
4	155.7	1.66	155	1.1
5	156.1	2.41	155	1.5
6	162.1	2.33	156	1.4
:	:	:		
100				
Average	156.78	1.98		

Figures 4.5.2.4, 4.5.2.5 and 4.5.2.6 are outputs from the simulations. The median, which measures the center of the data in this case, is represented by the line in the box. The whiskers extend from either side of the box and represent the ranges for the bottom 25% and the top 25% of the data values, excluding outliers. But for example, when sample size,  $n=50$ , Figure 4.5.2.4, the first plot in the left panel, we observe some outliers outside the top 25% percent of the data. The outliers would skew the mean value of the length classes upwards in this case. But in this plot, the median length class seems similar for all the sample sizes. In Figure 4.5.2.5, the second plot in the left panel (SE of mean) the results show that the standard error (SE) is quite variable for the different samples, but as expected, it decreases with increasing sample sizes. Similarly, the coefficient of variation of the mean decreases with increasing sample sizes. The CV is a measure of variability in the data, so as sample sizes increase, the variability is expected to decrease, hence, decrease in CV. Figure 4.5.2.6 shows the boxplots on the correct modes identified by simulations and tests performed to ensure quality of simulations. As expected, the results are closest to reality as the sample size increase.

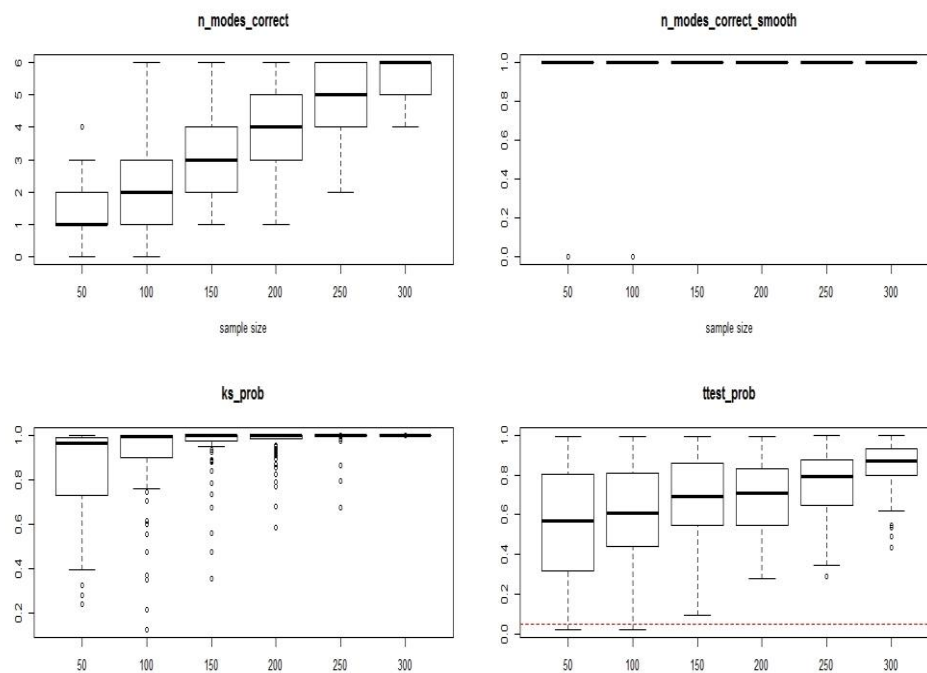


**Figure 4.5.2.4 - Boxplots of the median values, minimum and maximum length classes, and number of classes sampled for the 100 simulations for sample sizes 50; 100; 150; 200; 250 or 300.**





**Figure 4.5.2.5 - Boxplots of the means, standard error (SE) and coefficient of variation (CV) of the 100 simulations for sample sizes 50; 100; 150; 200; 250 or 300.**



**Figure 4.5.2.6: Boxplots of the number of correct modes before and after smoothing and the Kolmogorov-Smirnov probability for the 100 simulations for sample sizes 50; 100; 150; 200; 250 or 300.**

## 4.6 *Pandalus borealis* case study

### 4.6.1 Fishery

In the North Sea and Skagerrak three geographically separated aggregations of the northern shrimp (*Pandalus borealis*) are recognised and assessed as three separate stocks: 1) the Norwegian Deep-Skagerrak stock which is confined to ICES Divisions 4.a east and 3.a, 2) the Fladen Ground stock in ICES Divisions. 4.a west, and 3) the Farn Deep stock in ICES Division. 4.b west (Ulmestrand *et al.*, 2014). The Northern shrimp (*Pandalus borealis*) fishery with bottom trawl in ICES Divisions 3.a and 4.a East is one of the most important Swedish fisheries. In recent years the Sweden has landed ca. 1500 tonnes from a total of 11 000 tonnes landed from this stocks by Sweden, Denmark and Norway. There are about 50 shrimp fishing permits used by vessels specialized in shrimp fishing ( $\geq 10$  t/yr) (Ulmestrand *et al.*, 2014). Larger trawlers normally fish in the eastern and central part of Skagerrak, while smaller trawlers mostly fish in the Swedish coastal zone (Ulmestrand *et al.*, 2014). The shrimp caught by Swedish vessels is size sorted on board and the different fractions sold for different use (public or industry). In each haul, some shrimps may be discarded for one of two reasons: 1) shrimp  $< 15$  mm CL are little marketable and 2) to replace medium-sized, lower-value shrimps with larger and more profitable ones (“highgrading”). Highgrading has been illegal since 2009 in EU waters and since 2016, *Pandalus borealis* is included in the list of EU landing obligation species. As the Swedish fishery has often been constrained by the national quota, discards are a concern so an onboard sampling programme is in place to monitor and estimate their amount so it can be accounted for during stock assessment.

### 4.6.2 Sampling

The Swedish commercial shrimp fishery is sampled on board. 24 fishing trips are observed at-sea each year, 6 per quarter. In each quarter, 3 trips are sampled from each of two métiers. Two observers are present in each trip. The observers sample the catch of all hauls, including shrimp but also fish. With regards to shrimps, the catch is generally conducted through a set of two sieves, with progressively smaller width that are used to separate the shrimps into 3 fractions: the boiled (generally the larger specimens, with higher price, sold whole to the general public), the fresh/raw (generally of intermediate size, with intermediate destined to industry) and the “lus” (generally the smaller sizes, with low price, that may be discarded<sup>2</sup> or sold for transformation). In each haul observers inquire the captain on the amount of boiled, raw and lus he obtained, but due to preservation issues, samples of the boiled and raw fractions are generally only taken from the last haul of the trip; with regards to lus, a sample is taken for species composition from all hauls but full biological analysis, including measurements, are only carried out in the last haul of the trip. Since 2015, in parallel with the sampling of the main 3 fractions, the observers are instructed to take an extra sample from the unsorted catch of each haul (with ca. 2 kg). That sample is brought to the lab for species identification and measurement. This additional sampling is quite burdensome for the observers and has the objective of validating the quantity and length composition estimated from the sum of individual fractions.

---

<sup>2</sup> Presently there is a landing obligation on *Pandalus borealis* so the % of discards has decreased substantially in recent years

The lab time spent processing shrimp samples is considerable (10–12 h spent on a normal shrimp sample). When lab staff was inquired which fraction of the catch was considered less efficient to process, the supplementary catch sample – even if only measured (no other biology is taken) – was identified as the priority for investigations of optimization. If the numbers of shrimps that need to be measured could be reduced without significant impact on for data quality then time would become available for other tasks (e.g. quality control of data from other stocks, data analysis, etc.).

#### 4.6.3 Methodology

##### Data:

Length frequency of individual shrimps from shrimp samples collected by the at-sea sampling programme during 2016 were extracted from the National Database FD2. Between 229 and 555 (mean: 396) shrimps had been measured from 64 samples taken from hauls of 22 trips (1–8 hauls per trip). The data report issued from FD2 was very similar to the RDB CA format with only minor adaptations being required to enter the sample level R-script developed prior to WKBIOPTIM.

##### Simulations:

Script v.11 was used in simulations with the following parameterization:

- Minimum number of individuals per sample: 350 shrimps
- Variable = "Length\_class"
- Original size span = 1
- Smooth size span = 2
- Threshold for mode identification = 1% of length distribution
- Sampling\_options:
  - n\_sims = 500,
  - stratified = FALSE,
  - replacement=FALSE,
  - sample\_all\_available = TRUE,
  - sample\_all\_available\_warning = TRUE,
  - stages = "one",
  - samp\_sizes = c(seq(10,350, by=10),nrow(df1)),
  - strata\_var = "none",
  - vars\_to\_keep = c("Length\_class"))

The above mentioned setup resulted in the consideration for analysis of 49 samples (77% of the samples available) and the carrying out of 500 simulations of random sampling without replacement with sample size 10 to 350 (in 10 mm intervals). Simulation time was ca. 30 min in an HP ZBook 15 G2 equipped with a Processor Intel(R) Core(TM) i7-4710MQ CPU @ 2.50GHz, 2501 Mhz, 4 Core(s), 8 Logical Processor(s) with 6 clusters allocated in the parallel processing stage. For each replicate sample simulated a set of indicators was calculated including: Mean, standard error of the mean, CV of the mean; minimum, maximum, and median; number of size/age classes sampled; number of modes, number of modes identical with the modes of original sample; number of modes (after smoothing), number of modes identical with the modes of original sample (after smoothing); two sample t-test for equal means; two sample Kolmogorov-smirnov test and mean weighed CV of the length distribution (according to Gerritsen and McGrath 2007).

*Selection of an appropriate sample size:*

Preliminary analysis indicated that the mean-weighted CV (MWCV) was a relatively robust indicator for the description of the quality of length frequency in *Pandalus borealis*. Accordingly the variation of MWCV with sample size was used to select an efficient sample size for the samples of *Pandalus borealis*. The results obtained with MWCV were further confirmed by visually inspecting the variation of the other statistics with sample size.

**Results**

The 49 samples selected for analyses contained between 12 and 19 size classes and included samples with varied sizes and from different quarters. The original MWCV of the samples varied between 13.2% (in a sample of  $n = 513$  individuals) to 18.7% (in a sample of  $n = 350$  individuals) with a mean of 15.9%.

Graphical outputs obtained from the script indicated the expected fast reduction in MWCV with sample size with increased precision indicators. Summary plots of observations obtained from 3 illustrative samples are presented in Figure 4.6.3.1–4.6.3.4.

Table 4.6.3.1 summarizes the % of simulated replicates with maximum MWCV <30%, 25% and 20%, respectively at different sample sizes. It shows that based on the current reference set of 49 samples, when 230 shrimps are measured (i.e. ca. 1.5 kg of shrimp) the mean-weighted CV is very unlikely to ever be >25%. More detailed analyses of replicates with this sample size confirm that the resulting length frequencies are nearly indistinguishable from the original length frequencies of the samples maintaining important aspects like the allowing for characteristics like estimation of the mean or median, minimum, maximum, and a correct identification of the main modes of the distribution with only minimal error (Figures 4.6.3.5–5.6.3.7).

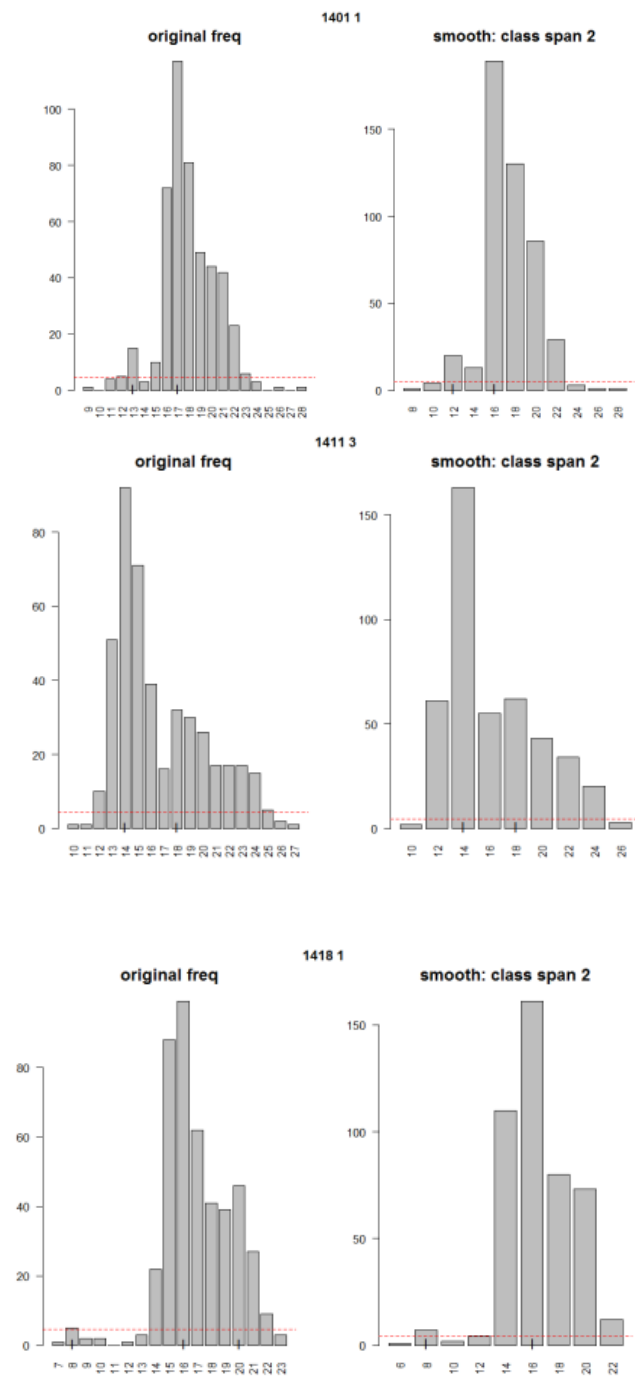


Figure 4.6.3.1 - Demonstrative samples from the *Pandalus* fishery. Sample 1401 1: January, 477 individuals, 9–28 mm in length; Sample 1411 3: May, 433 individuals, 10–27 mm in length; Sample 1418 1: September, 450 individuals, 7–23 mm in length.

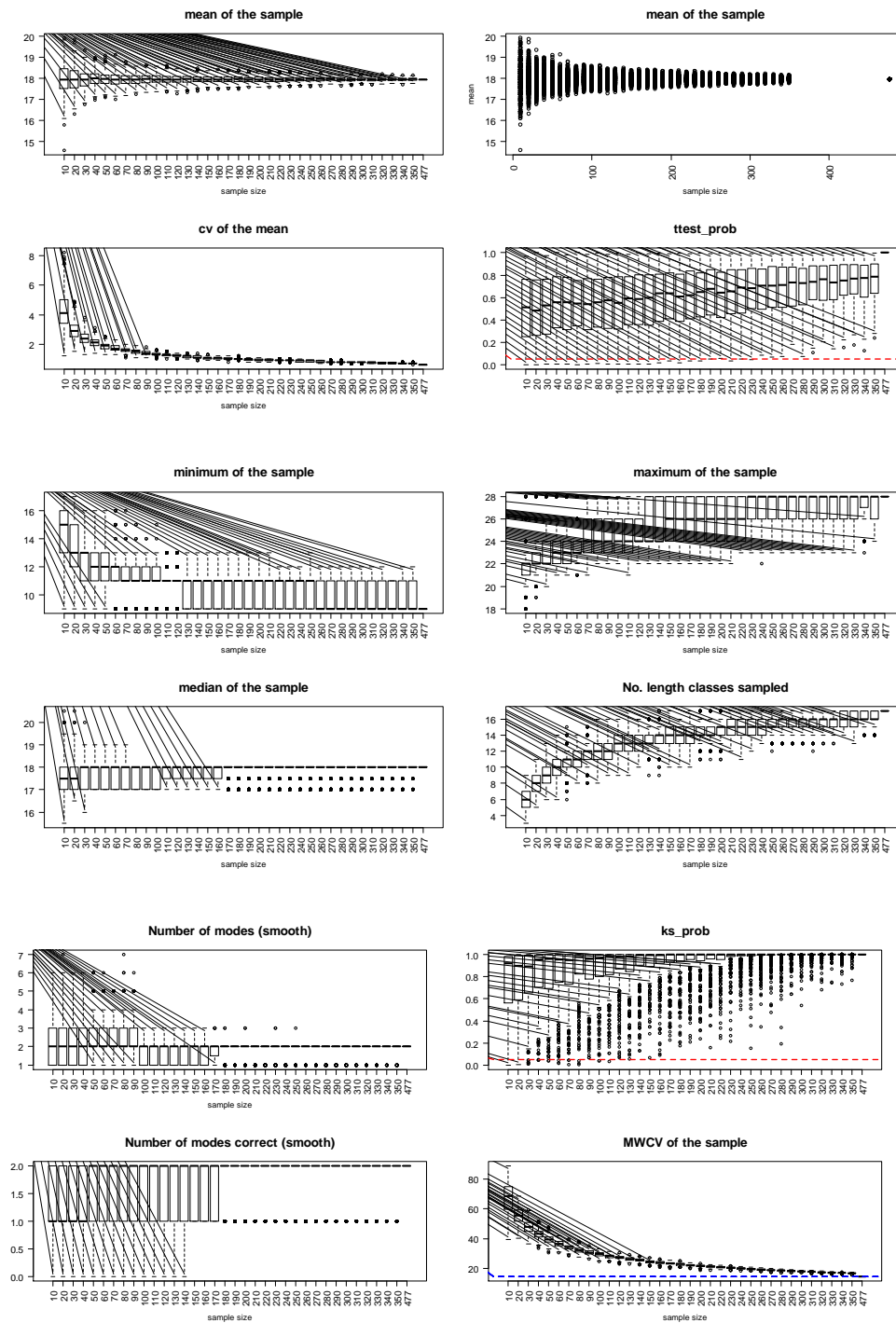


Figure 4.6.3.2 - Example of graphical outputs obtained for sample 1401 1 of shrimp (sample details in Fig 4.6.3.1)

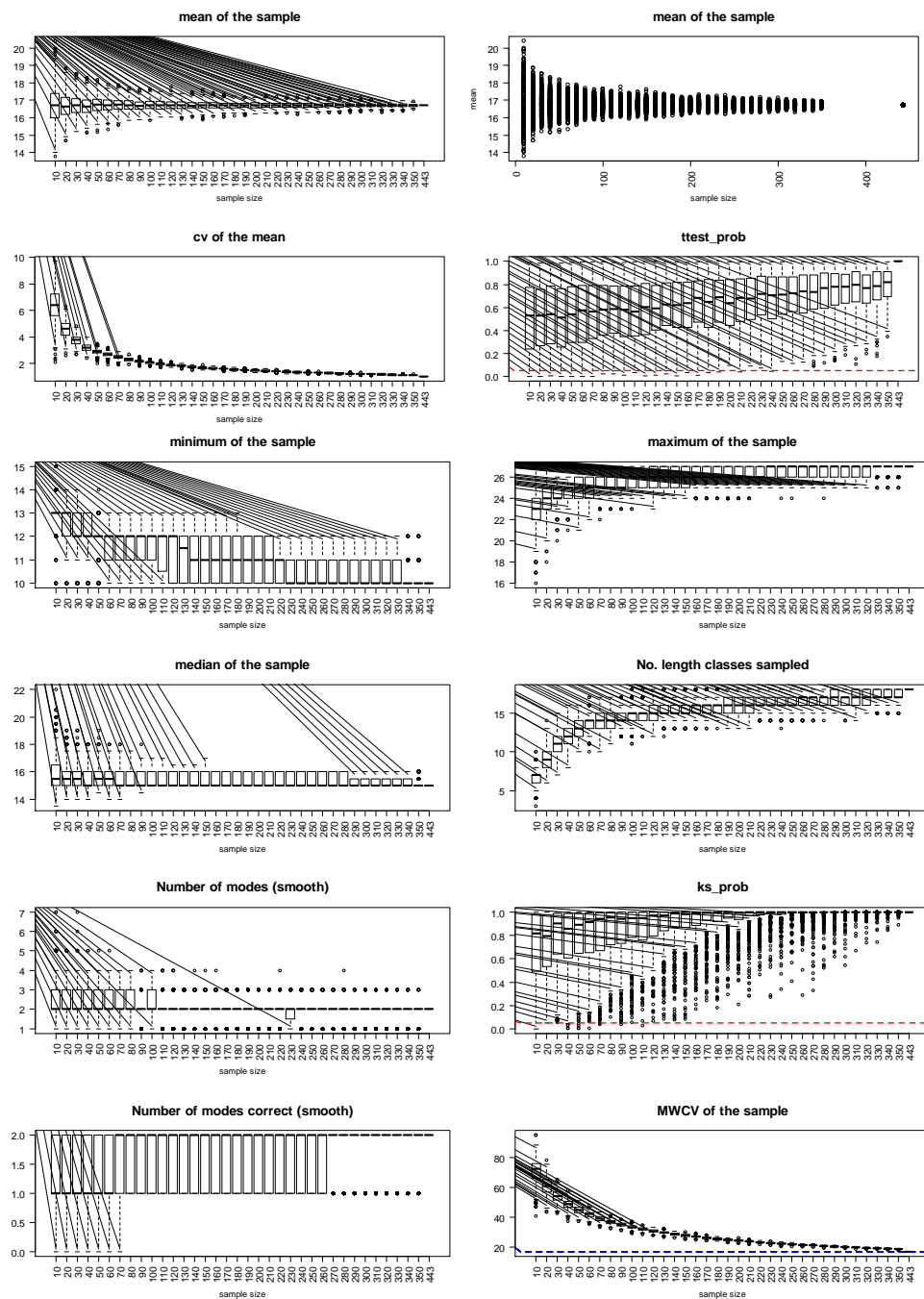


Figure 4.6.3.3 - Example of graphical outputs obtained for sample 1411 3 of shrimp (sample details in Fig 4.6.3.1)

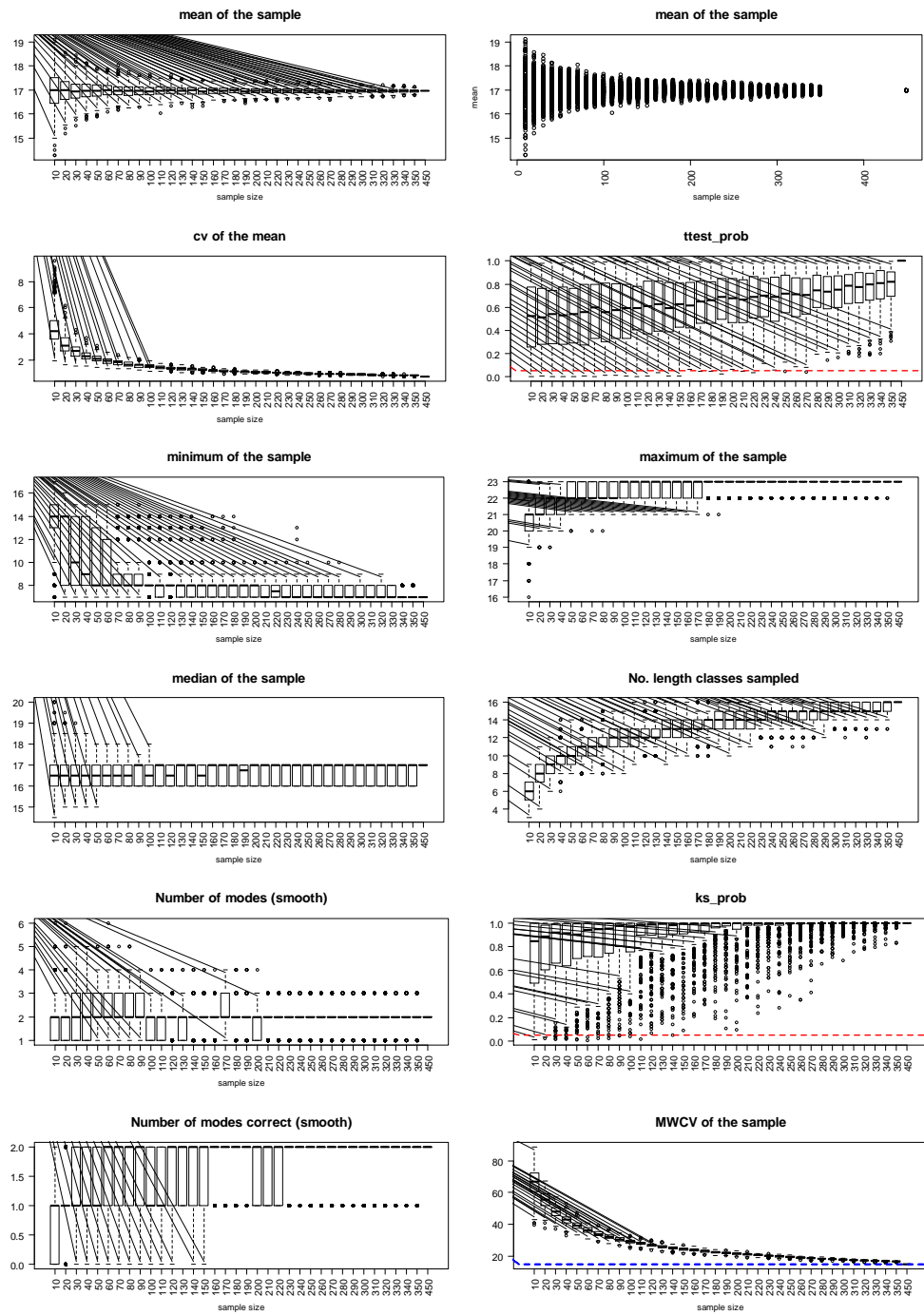


Figure 4.6.3.4 - Example of graphical outputs obtained for sample 1418 1 of shrimp (sample details in Fig 4.6.3.1)



Table 4.6.3.1 - Percentage of samples with three levels of maximum MWCV per sample size.

Sample size	Aprox Weight (g)	Max MWCV < 30%	Max MWCV < 25%	Max MWCV < 20%
100	629	0	0	0
110	691	6	0	0
120	754	22	0	0
130	817	53	0	0
140	880	80	0	0
150	943	92	2	0
160	1006	98	10	0
170	1069	<b>100</b>	24	0
180	1132	100	45	0
190	1195	100	69	0
200	1258	100	80	0
210	1321	100	92	0
220	1383	100	95	0
230	1446	<b>100</b>	<b>100</b>	<b>2</b>
240	1509	100	100	8
250	1572	100	100	14
350	2201	<i>100</i>	<i>100</i>	<b>100</b>

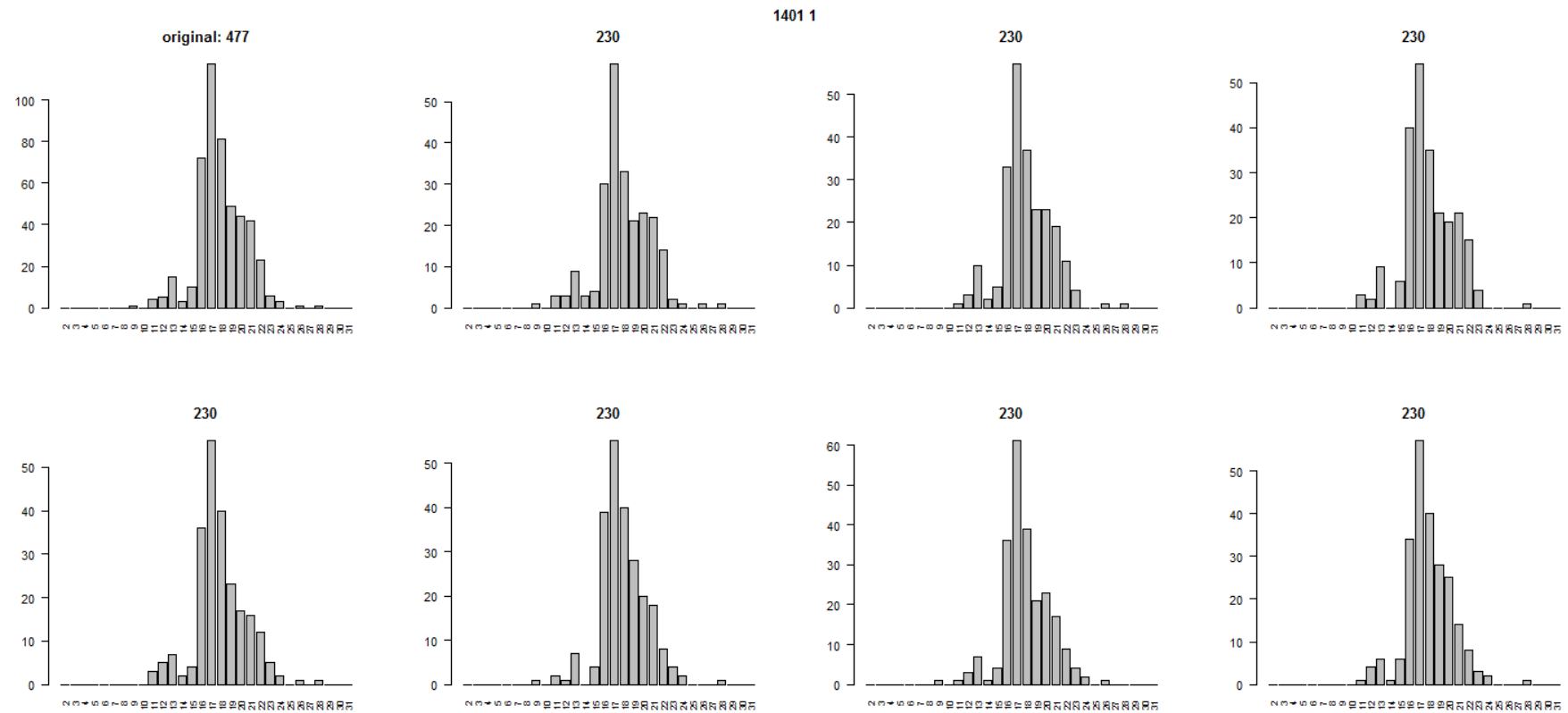


Figure 4.6.3.5 - Demonstrative example of 7 replicates of 230 shrimps from sample 1401 1 (sample details in Fig 4.6.3.1)

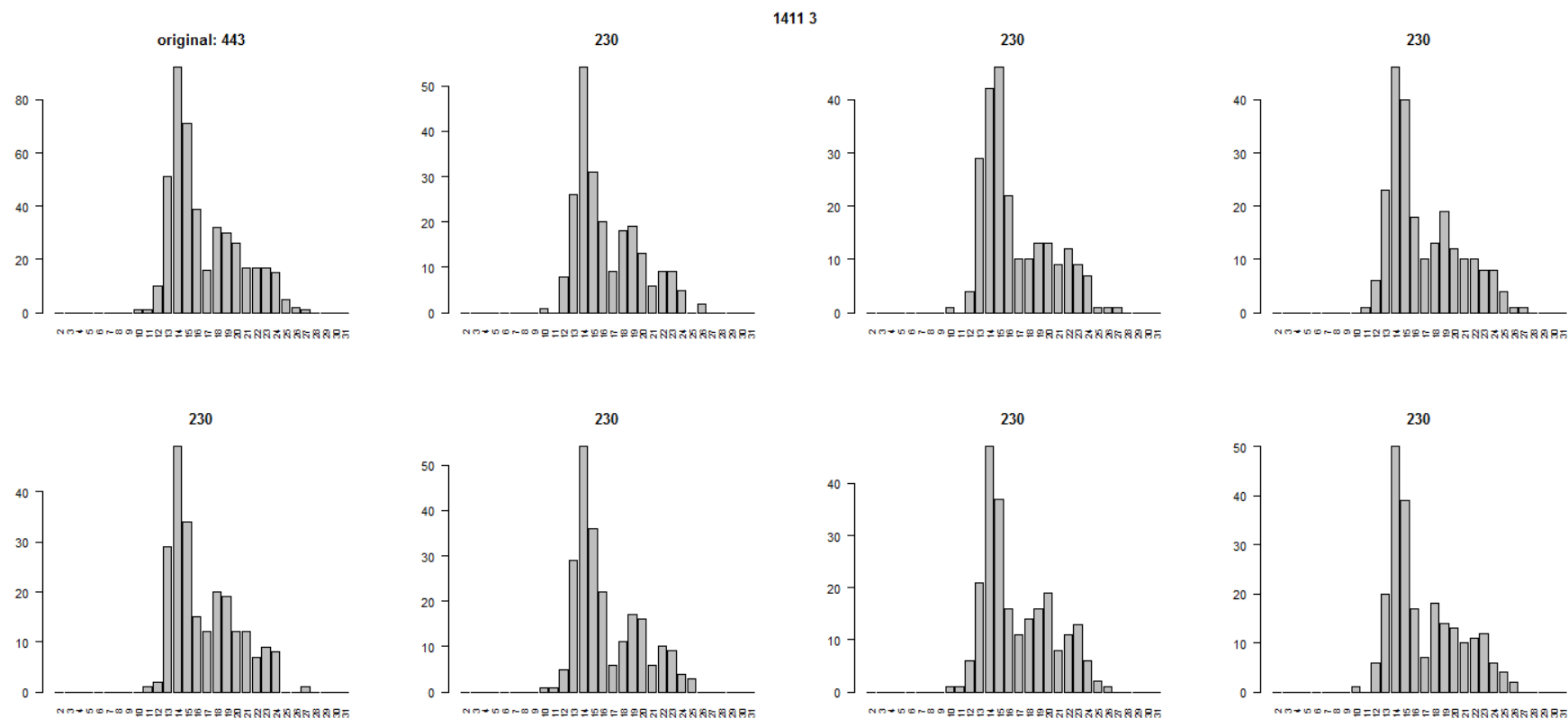


Figure 4.6.3.6 - Demonstrative example of 7 replicates of 230 shrimps from sample 1411 3 (sample details in Fig 4.6.3.1)

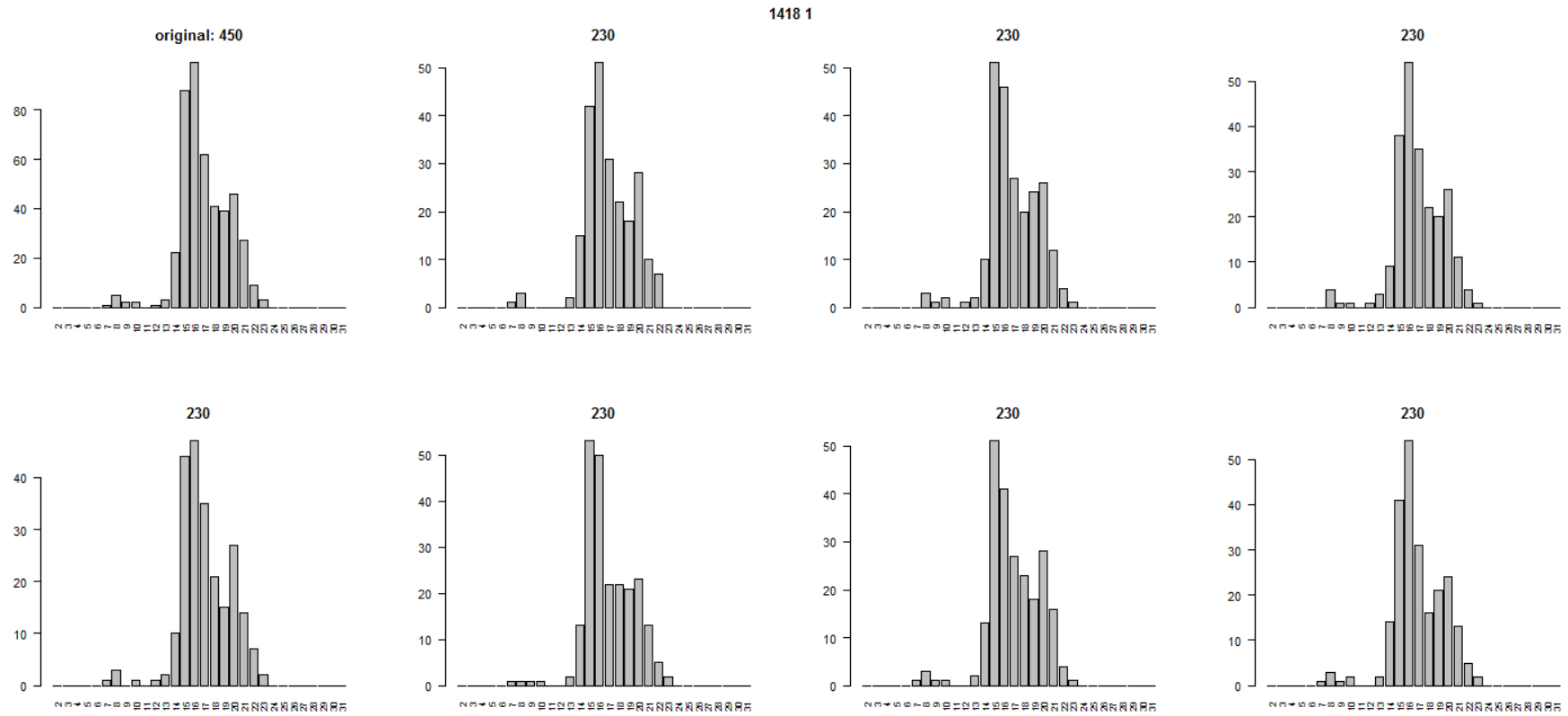


Figure 4.6.3.7 - Demonstrative example of 7 replicates of 230 shrimps from sample 1418 1 (sample details in Fig 4.6.3.1)

## Conclusion

Catch samples with 1.5 kg of *Pandalus borealis* appear to be a reasonable sampling target for observers collecting shrimp samples from hauls of the Swedish shrimp fishery. Such value has highly conservative in that it allows for every haul to be well characterized, a situation that is hardly required for many common end-uses of these data but that could be useful for future spatial analyses. If it is adopted, the new sampling goal represents a ca. 25% reduction in number of shrimps measured while still producing good enough length distributions for nearly all conceivable end-uses (both present and future). Such time and cost reductions are invaluable in that they can be applied to data collection of other resources. It is therefore recommended that observers aim to collect such samples having only the care to ensure that they bring to the lab a slightly higher weight (e.g. 2 kg) when other shrimp seasons are very frequent in the catch sample.

## Reference list

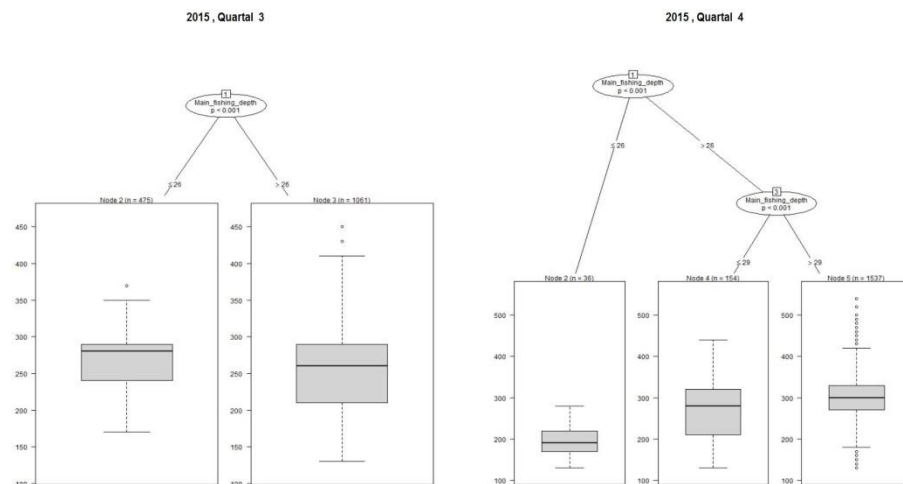
Gerritsen H. D. and D. McGrath (2007). Precision estimates and suggested sample sizes for length–frequency data. *Fishery Bulletin* 106:116-120.

Ulmestrand M, S. Munch-Petersen, G. Søvik and O. Eigaard (2014), The Northern shrimp (*Pandalus borealis*) Stock in Skagerrak and the Norwegian Deep (ICES Divisions IIIa and IVa East). Serial No. N6370. NAFO SCR Doc. 14/065

## 4.7 Plaice (Germany) case study

The case study analysed plaice (*Pleuronectes platessa*) and took a closer look at at-sea-sampling (which is only one sampling type implemented in Germany) in the area 4. The analysis was conducted by means of the proposed R-code. It has shown that, in general, sampling effort can be reduced compared to the existing one.

In future study other approaches to reveal the amount of length/age measurements which provide the best compromise between total sampling effort and estimation precision can be explored. One example is the “change point detection method”, see i.e. Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). Another promising area of future research might be an investigation of factors can affect length distribution. R package party offers a reliable tool for the identifying the suite of characteristics/factors that are important for the case study species. So, the example presented demonstrates that the fishing depth seems to be the most important variable explaining plaice length distribution, for the quarters 3 and 4 in 2015. The tree is split off on the values of one factor (independent variable) at a time such that the overall variance of the mean length (dependent variable) is minimized at each split. Terminal nodes indicate the mean length value of plaice relative condition assigned to the node. The nodes were detected from the original sample for métier TBB\_DEF\_70-99\_0\_0.



**Figure 4.7.1 – Tree split illustrating the important variables explaining plaice length distribution for quarters 3 and 4 in 2015**

## References

Zeileis A., Leisch F., Hornik K., Kleiber C. (2002) strucchange: An R package for testing for structural change in linear regression models. Journal of Statistical Software, 7 (2) (2002), pp. 1–38

## 5 Multi-level analysis

**ToR c) Test the code developed in a set of case-studies and quantify effects, advantages and disadvantages of different options of statistical optimization at sample level in terms of cost and time-savings involved**

The analyses were performed on 3 species combining data from different countries or from one single country:

- sardine in areas 27.8.c and 27.9.c : analyses using Spanish data,
- hake in area 27.9.a: analyses using Portuguese data.
- sole in areas 28.8.a,b and 27.4: analyses using the Belgium and French data,

### 5.1 Sardine in areas 28.8.c and 27.9.a

*This is a presentation of main points and outcomes for this case study. A full version of the case study with all data, analysis and figures can be found in Annex 4.*

The Southern stock of sardine (*Sardina pilchardus* in divisions 8.c and 9.a, stock code: sar-soth) is exploited by Spain and Portugal. This analysis focuses on the Spanish sampling data. In 2015, total Spanish catches were 6986 t. using the commercial sampling data for the year 2015, this section presents a framework where length distribution estimates are built using different subsamples of the original database. The way these subsamples are built are called 'scenario'.

This study, using numerical simulation, provides some indication on how sampling effort could be reallocated, keeping the same quality of information for ONE species and ONE country. The simulations were performed without taking into account the information related to other stocks (concurrent sampling) and to other parameters (e.g. discards), and without any consideration to practical issues related to the organization of the sampling plan.

#### 5.1.1 Fleet

Sardine is mainly captured by purse-seiners. The Spanish fleet targets anchovy (*Engraulis encrasicolus*), mackerel (*Scomber scombrus*), horse-mackerel (*Trachurus trachurus*) and sardine (*Sardina pilchardus*) with some characteristic seasonal patterns for some of the species (anchovy in spring; mackerel in winter). The gear must have a minimum mesh size of 14 mm.

#### 5.1.2 Sampling design

Overall Spanish sampling design in 2015 was based on métiers following the ranking system according to the DCF. For purse-seiners, sampling frame was constructed with a list of ports based on effort and landings where days and trips were selected. Frame was stratified by quarter with a systematic monthly allocation of effort. Full concurrent sampling (all species) was in place.

#### 5.1.3 Data

National data are loaded. Some corrections are made to homogenise the data, due to some discrepancies in variable type. (See Annex 4 for specific programming details)

#### 5.1.4 Exploratory data analysis

In this section, a short exploratory analysis of the data are provided.

## Global figures

Landings and sampling effort are computed by country and ICES Subdivision.

Table 5.1.4.1 - Sardine fishery for Spain in 2015

COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	FISH MEASURED
ESP	8c	1933	35	3225
ESP	9a	5053	80	5510

## Métier

Landings and sampling effort by métier are summarized. The number of métier is reduced using the ICES rules on main métiers: here métiers fishing less than 90% of the cumulated catches are coded as MIS\_MIS\_0\_0\_0.

Table 5.1.4.2 - Landings and sampling effort by main métier, country and subarea in 2016. MIS\_MIS métiers are generated using the ICES rules.

MÉTIER	COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	HAULS SAMPLED	FISH MEASURED
PS_SPF-0_0_0	ESP	8c	1916	35	35	3225
MIS_MIS-0_0_0	ESP	8c	16.63	0	0	0
PS_SPF-0_0_0	ESP	9a	4902	72	72	5023
MIS_MIS-0_0_0	ESP	9a	151.4	8	8	487

## Landings

Landings are mapped by ICES rectangle, quarter and country.

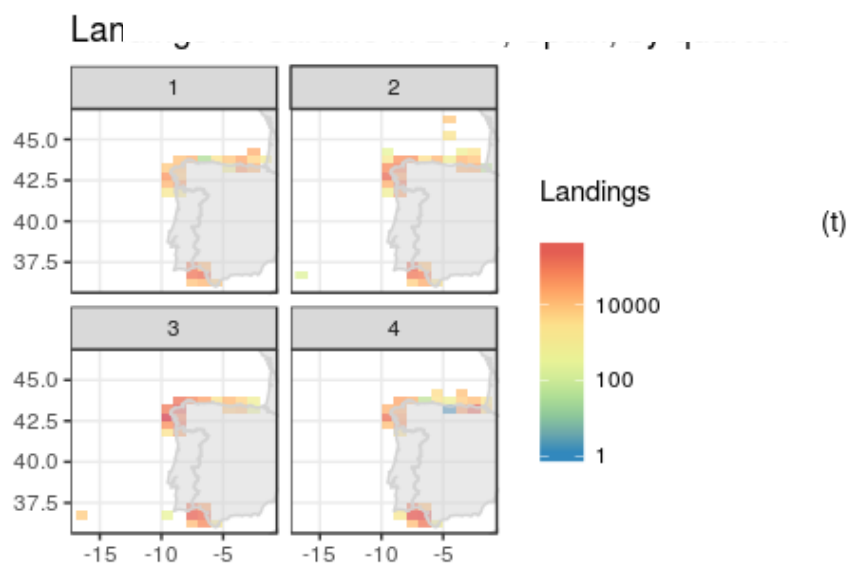


Figure 5.1.4.1 – Spatial distribution of landings of sardine in 2015, by quarter



## Stratification

The stratification defines the spatial, temporal and technical support on which population estimates are computed. The stratification follows the ICES data call policy: quarter, ICES area and métier level 6.

timeStrata	spaceStrata	techStrata
quarter	area	foCatEu6

## Validation and consolidation

According to the stratification, the data are validated: samples are tested using quality checks.

### 5.1.5 Quality checks

#### Sample weights

Subsamples weights are checked using the generic sardine weight length relationship. If the samples weights differ from the theoretical weight by 50% then this sample is removed.

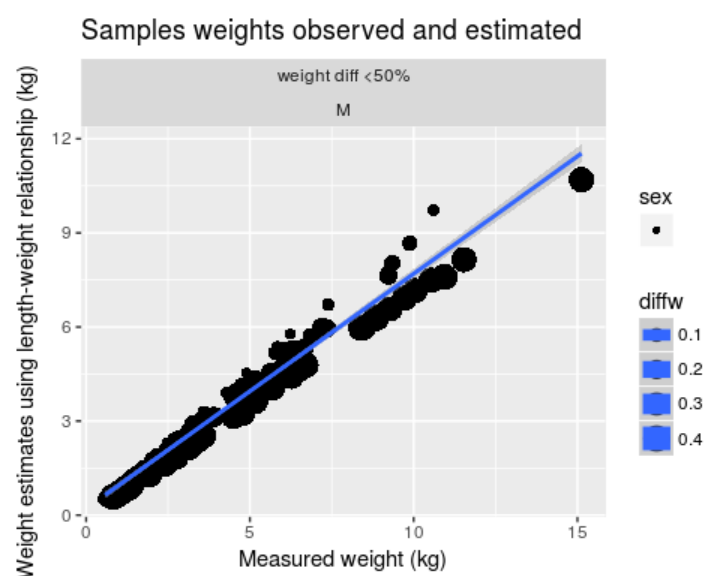


Figure 5.1.5.1 - Length weight relationship quality check

#### Sample outliers

Sample outliers are tested using the delta approach following Vigneau and Mahevas (2007). This index helps to detect sample outliers and sample heterogeneity when catch-at-length is estimated using the ratio estimator. In this report, in order to ensure the repeatability of our results, a fixed threshold is applied to detect outliers. In a strata, all the samples with delta values outside the quantile at 2.5%% and 97.5% are flagged as outliers. The delta value outliers' detection for samples are computed but not used in this report:

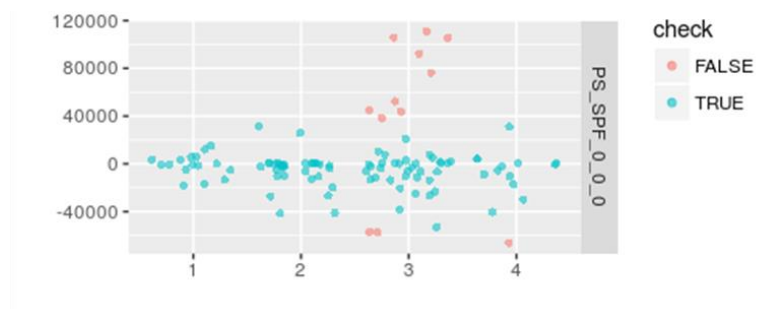


Figure 5.1.5.2 – Detection of outliers using the Delta measure (COST package)

### Sample coverage

The sample coverage is checked according to the stratification: the percentage in numbers of samples by strata is presented against the percentage of landings in each modality (red line).

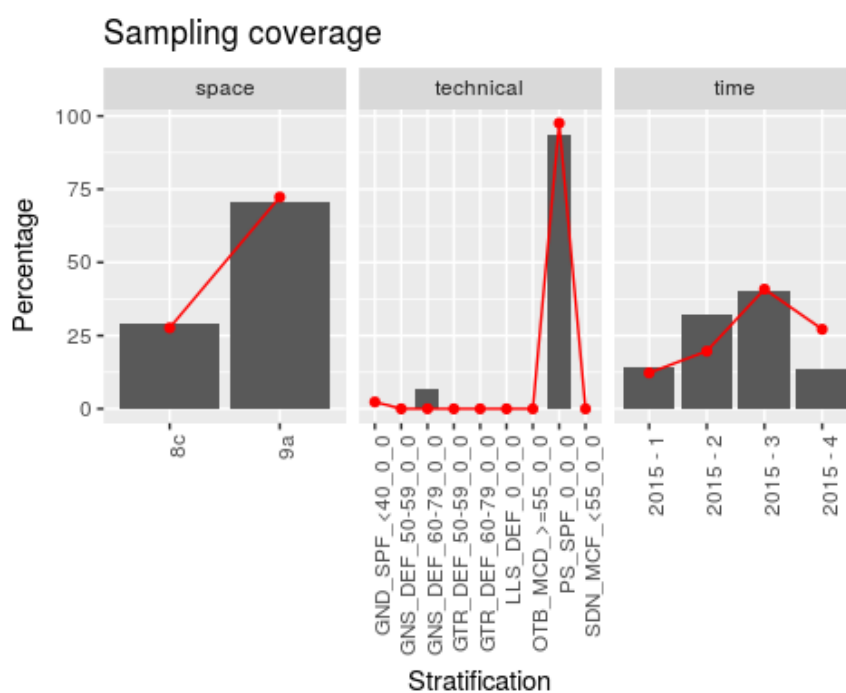


Figure 5.1.5.3 – Sample coverage in relation to landings

### 5.1.6 Raised length composition of the stock

The length composition is estimated using ratio-estimator based on landings. To summarize our graphical output, length distributions are combined by ICES Subarea and quarter.

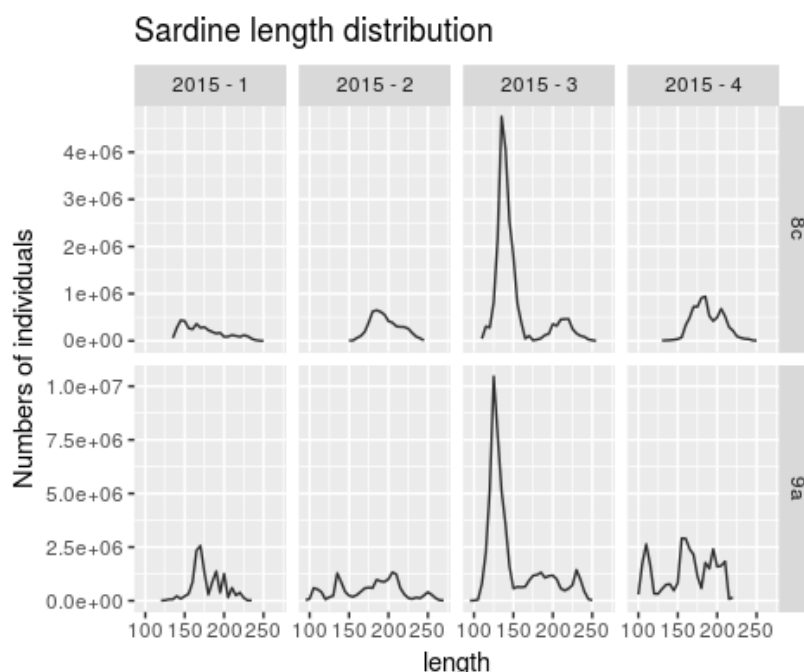


Figure 5.1.6.1 - Raised length composition for sardine.

### 5.1.7 Scenarios

From the first analyses, the length distribution of the sardine in the areas 27.8.c and 27.9.a are estimated using the dataset collected by Spain in 2015. These length distributions are considered the "true" estimates of the catches. This section is now dedicated to the estimation of the same parameters, with changes in the sampling plan. These changes are scenarios testing some modification of the sampling plan and reallocation of the sampling effort. To do so, samples are removed according to the scenario, in a fixed or random way and length distribution are estimated based on the new samples base. The results are compared to the "true" estimates using the original number of samples. To compare the 2 distribution, the Earth Mover's Distance (EMD) is used. In statistics, the (EMD) is a measure of the distance between two probability distributions over a region D (names Wasserstein metric in mathematics) (Rubner, Tomasi, and Guibas 1998). Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region D, the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.

The 2 scenario tested in this study are:

- Scenario 1: decrease the number of sampled trips.
- Scenario 2: decrease the number of fish sampled by trip.

### Scenario 1

In this scenario, the sampling effort was decreased, from 100% to 10% of the number of original trips sampled. Trips are sampled randomly with replacement in 30 simulations.

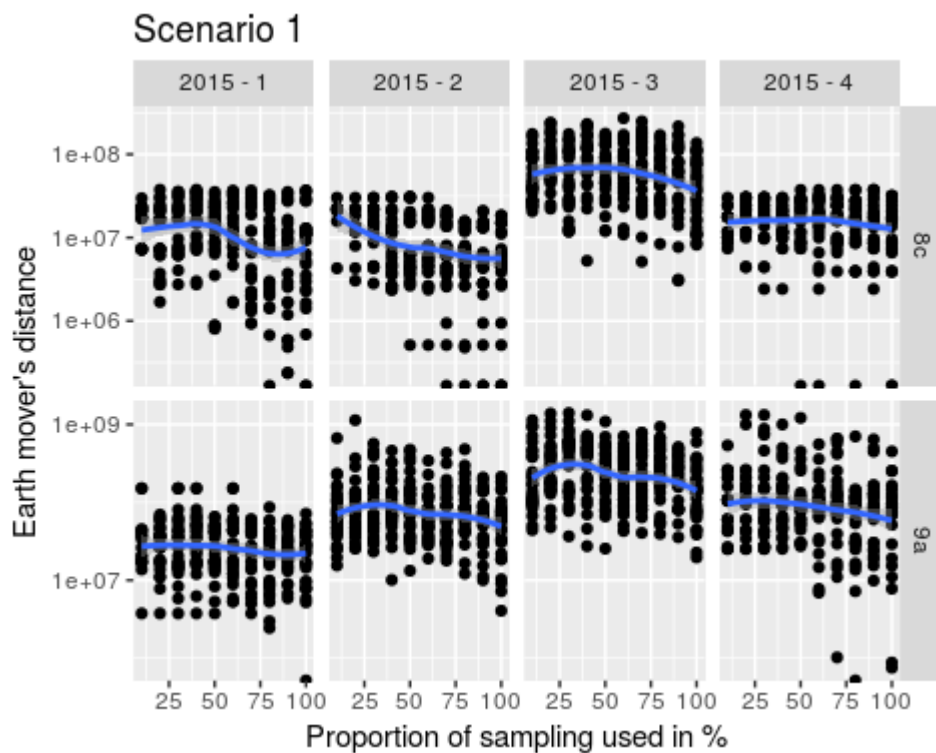


Figure 5.1.7.1 The earth mover's distance (EMD) results

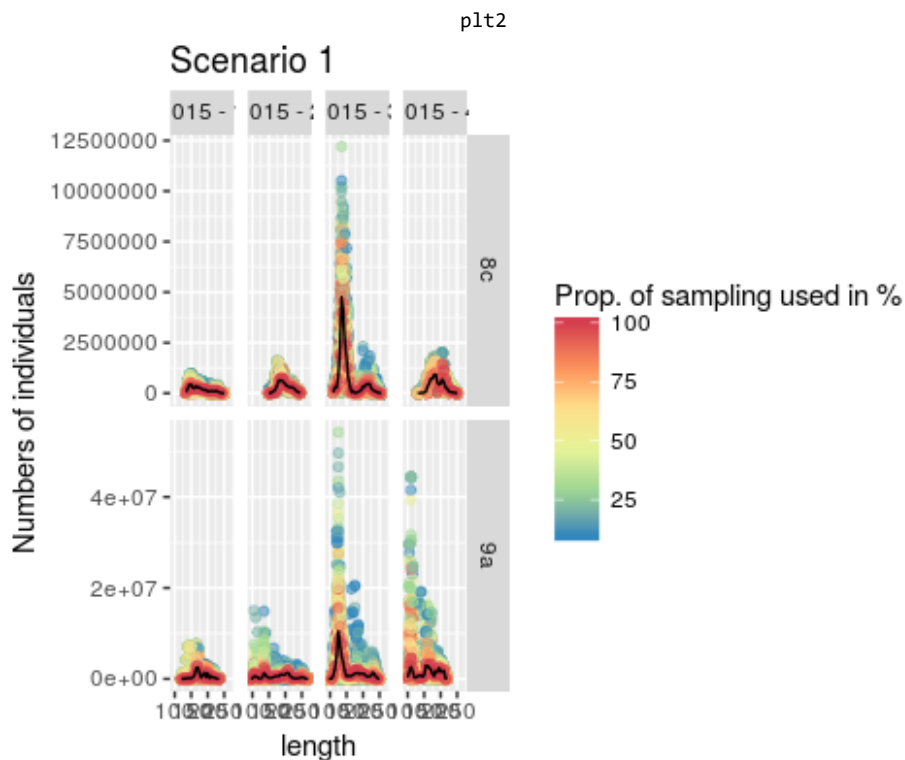


Figure 5.1.7.2 - The estimated length distributions

The earth mover's distance (EMD) (Figure 5.1.7.1) decreases logically with the increase in number of trips used to estimate the length distribution (the blue line is a LOESS smoother) for some strata (quarter 2 in 8c for example). For other strata the EMD curve is flat: decreasing the number of trips sampled did not change the shape of the length distribution. For the quarter 1 in 8c, the sampling effort could be reduced to 50% without loss of information. This is confirmed by the visual inspection of the length distribution (Figure 5.1.7.2). For the other strata, the length distribution computed with 50% to 100% of the trips (red to orange dots) convey the same amount of information than the original length distribution (black line) for the subarea 27.8c.

## Scenario 2

In this scenario, the impact of the number of fish sampled by haul inside a trip is investigated. For each haul, 10–100 fish are taken randomly with replacement from the original measurements, and length distributions are computed using the new samples. 30 simulations were performed using this scenario.

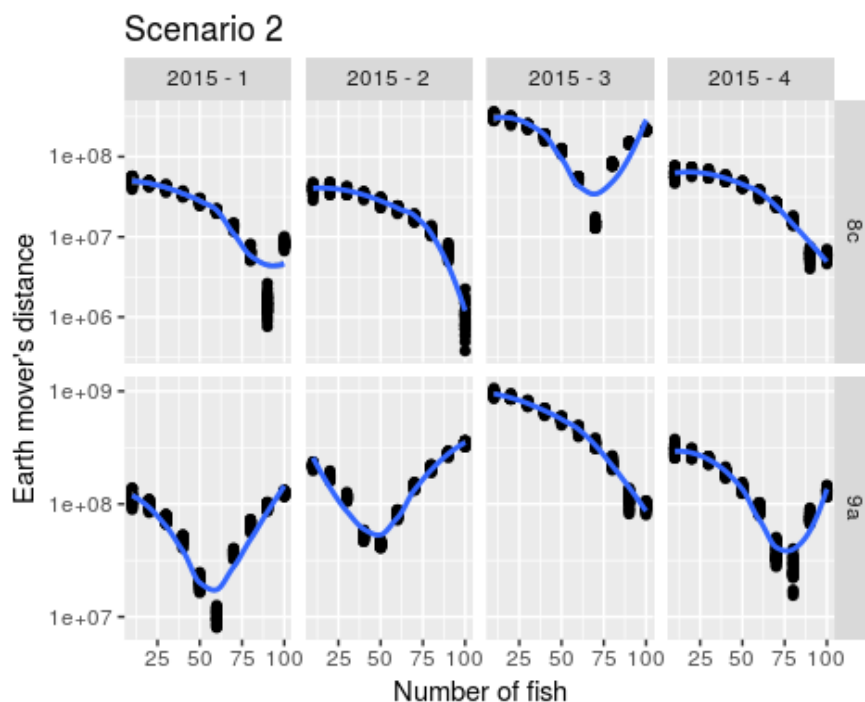


Figure 5.1.7.3 The earth mover's distance (EMD)

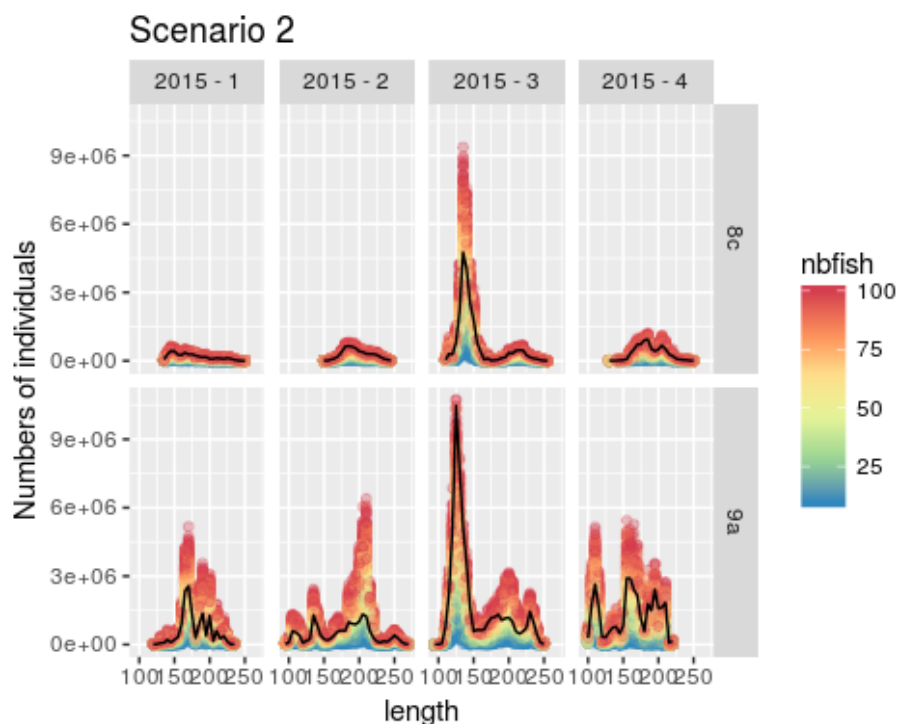


Figure 5.1.7.4 The estimated length distributions

The earth mover's distance (EMD) reaches local minima, in some strata (Figure 5.1.7.3), and this should indicate an optimal number of fish to be measured. It is not clear what variables are determining these minima. While sampling protocols are the same in all the area, samplers and port procedures (access to the fish, sale, etc.) differs. The simulation using replacement explain why in the area 9c, around the local minima, an increase or a decrease in number of fish modify the length distribution. Some latitude at + or - 10 fish is possible. Graphically (Figure 5.1.7.4) the estimated length distributions are similar to the original one when the number of fish is high.

### 5.1.8 Discussion

Optimization analyses prove to be a valuable tool to explore sampling design and implementation. Detection of sample outliers with the delta index is already included in the Spanish procedure; implementing a common fixed threshold would be a step forwards a general quality improvement. No modifications are needed in relation to the sample coverage of this stock, which shows a correct connection between sampling effort and landings in all the three stratifications. Nevertheless, the scenario analyses for the sampling effort show that a decrease in number of trips could be performed without a substantial impact in the information provided. A better understanding of this EMD analyses can be used for adjacent years thus ensuring a clear indication for change. As a second factor for consideration, this analyses need to be done considering the other species covered under the sampling of this fleet. With the current concurrent sampling scheme the reduction of trips have to ensure similar information for all the sampled stocks.

### References

Rubner, Y., C. Tomasi, and L. J. Guibas. 1998. "A Metric for Distributions with Applications to Image Databases." *IEEE International Conference on Computer Vision*, 59–66.

Vigneau, Joel, and Stephanie Mahevas. 2007. "Detecting Sampling Outliers and Sampling Heterogeneity When Catch-at-Length Is Estimated Using the Ratio Estimator." *ICES Journal of Marine Science: Journal Du Conseil* 64 (5): 1028–32

## 5.2 Hake in area 27.9.a

*This is a presentation of main points and outcomes for this case study. A full version of the case study with all data, analysis and figures can be found in Annex 4.*

Southern hake stock comprises the Atlantic coast of Iberian Peninsula corresponding to ICES division 8.c and 9.a. Hake in these divisions is caught in the trawl and artisanal mixed fisheries together with other fish species and crustaceans. In 2014, total landings were 2443 t. using the commercial at-market sampling data for the year 2014, this document presents a framework where length distribution estimates for landings are built using different subsamples of the original database. The way how these subsamples are built are called 'scenario'.

Numerical simulations used in this case study provide some possible indication on how sampling effort could be reallocated, keeping the same quality of information for ONE species and ONE country. The simulations were performed without taking into account the information related to other stock (concurrent sampling) and to other parameters (discards) and without any consideration to practical issues related to the organization of the sampling plan.

### 5.2.1 Fleet

The Portuguese fleets targeting hake operate mainly in division 9.a. Trawl fleet comprises two distinct components – the trawl targeting demersal fish (OTB\_DEF) that operates along the entire Portuguese coast and the trawl fleet targeting crustaceans (OTB\_CRU) that operates mainly in the southwest and south coasts. Artisanal fisheries include gillnetters, longlines and a MIS\_MIS métier group of vessels is also present. This is related to the fact that a major part of Portuguese polyvalent fleet can operate with more than one gear in the same trip making the métier level 6 definition difficult to allocate for each trip. During the last year IPMA and DGRM (Portuguese administration) are joining efforts to come up with common criteria for the classification of trips from multigear vessels.

### 5.2.2 Data

National data are loaded and combined.

**Table 5.2.2.1 - Hake fishery for Portugal in 2014**

COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	FISH MEASURED
PRT	9.a	2443	522	26166

### Métier

Landings and sampling effort by métier are summarized in Table 5.2.2.2. The number of métiers is reduced: only métier catching hake and sampled are kept (with at least 7 trips sampled). The other métiers are labelled "Other".

Table 5.2.2.2 - Landings and sampling effort by main métiers, country and subarea in 2014.

MÉTIER	COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	FISH MEASURED
MISMIS-0_0_0	PRT	9.a	899.1	202	9308
OTBDEF- >=55_0_0	PRT	9.a	657.9	134	9795
GTRDEF- >=100_0_0	PRT	9.a	326.4	78	1892
Other	PRT	9.a	197.8	8	112
LLSDEF-0_0_0	PRT	9.a	72.68	8	616
GNSDEF-60- 79_0_0	PRT	9.a	10.68	31	1311
OTBCRU- >=55_0_0	PRT	9.a	6.116	62	3132

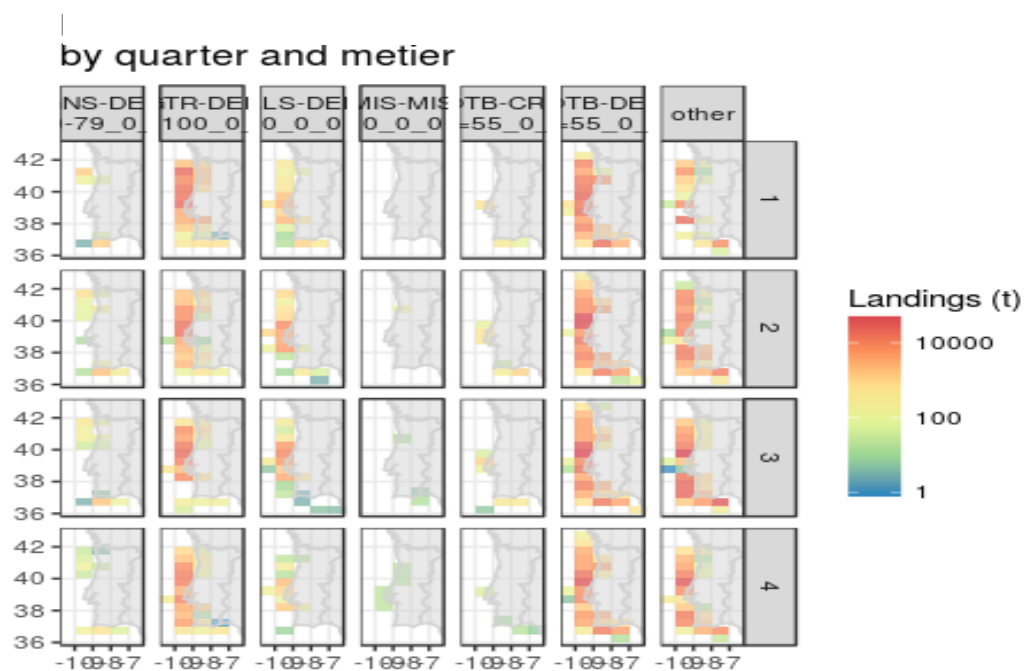


Figure 5.2.2.1 – Spatial distribution of hake by quarter and metier for 2014.

### Sample coverage

The sample coverage is checked according to the stratification and the percentage in numbers of samples by strata is presented against the percentage of landings (Figure 5.2.2.2).



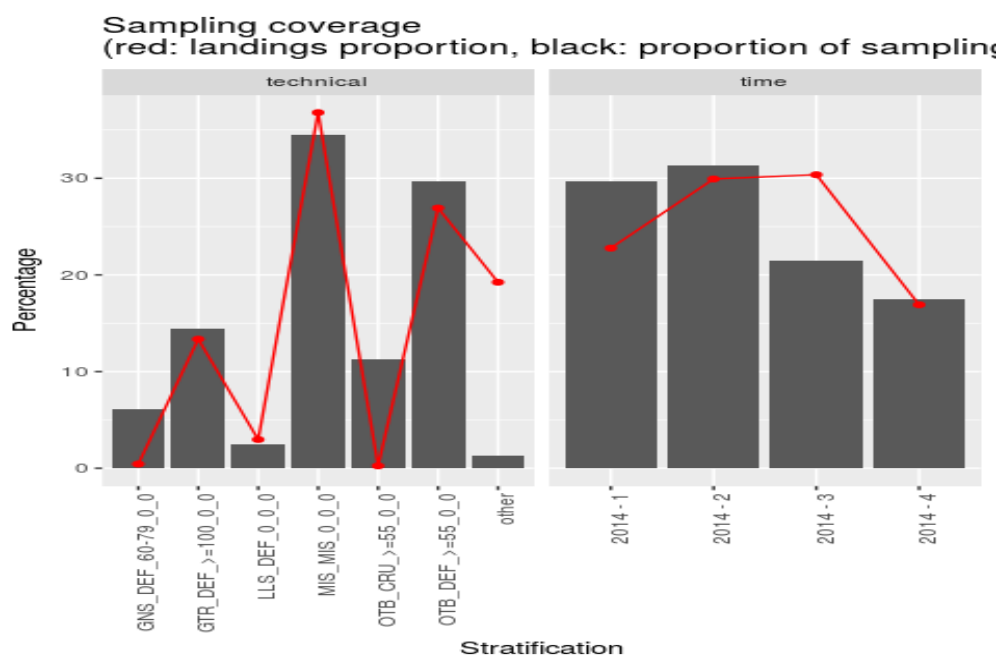


Figure 5.2.2.2 – Sample coverage of per métier and quarter. Red line represents the landing proportion and the black bars the sample coverage of the landings, for hake in 2014.

### Statistical analysis

The framework developed for this study integrates the whole processing of fishery data to (1) build population estimates at the stock level (length distribution mainly) and (2) to analyse the impact of different sampling strategies in these estimates by simulation.

### 5.2.3 Scenarios

From the first analyses, the hake length distribution in the area 27.9.a is estimated using the dataset collected by Portugal in 2014 (Figure 5.2.3.1). These length distributions are considered the "true" estimates of the information on landings by national sampling program. This section is now dedicated to the estimation of the same parameters, with changes in the sampling plan. These changes are scenarios testing some modification of the sampling plan. To do so, samples are removed according to the scenario, in a fixed or random way and length distribution are estimated based on the new samples base. The results are compared to the "true" estimates using the original number of samples. To compare the two distributions, the Earth Mover's Distance (EMD) is used (see Section 5.1.7).

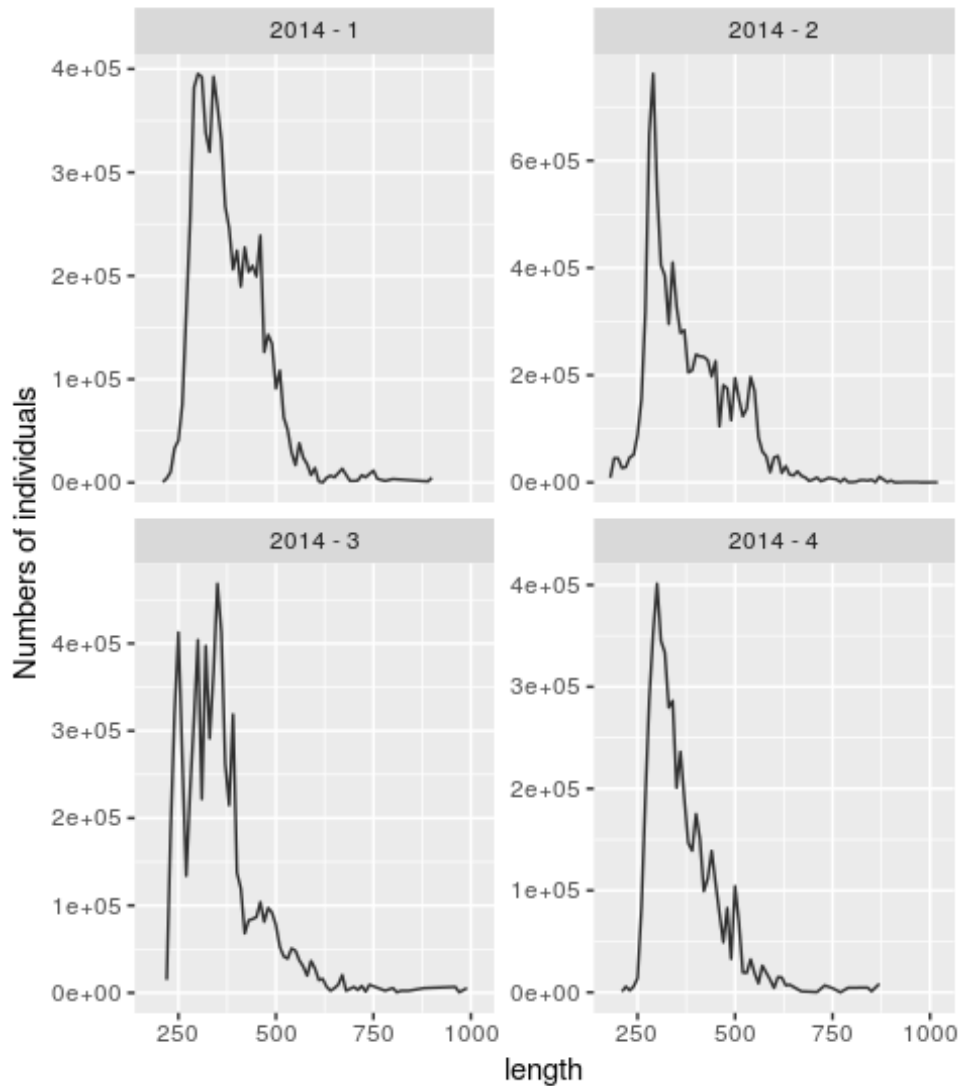


Figure 5.2.3.1 Hake length distribution by quarter, combining all métiers.

The 2 scenario tested in this report are:

- Scenario 1: decrease the number of sampled trips.
- Scenario 2: decrease the number of fish sampled by trip.

### Scenario 1

In this scenario, the sampling effort was decreased, from 100–10% of the number of original trip sampled. Trips are sampled randomly with replacement in 30 simulations.

The earth mover's distance (EMD) (Figure 5.2.3.2) decrease logically with the increase of number of trips used to estimates the length distribution (the blue line is a LOESS smoother) for some strata (quarter 1 and 4). For the third quarter the EMD curve is rather flat: decreasing the number of trips sampled did not change the shape of the length distribution. For the quarter 2 a local minimum is reached at 80% meaning a reduction of the sampling effort is possible without any loss of information.

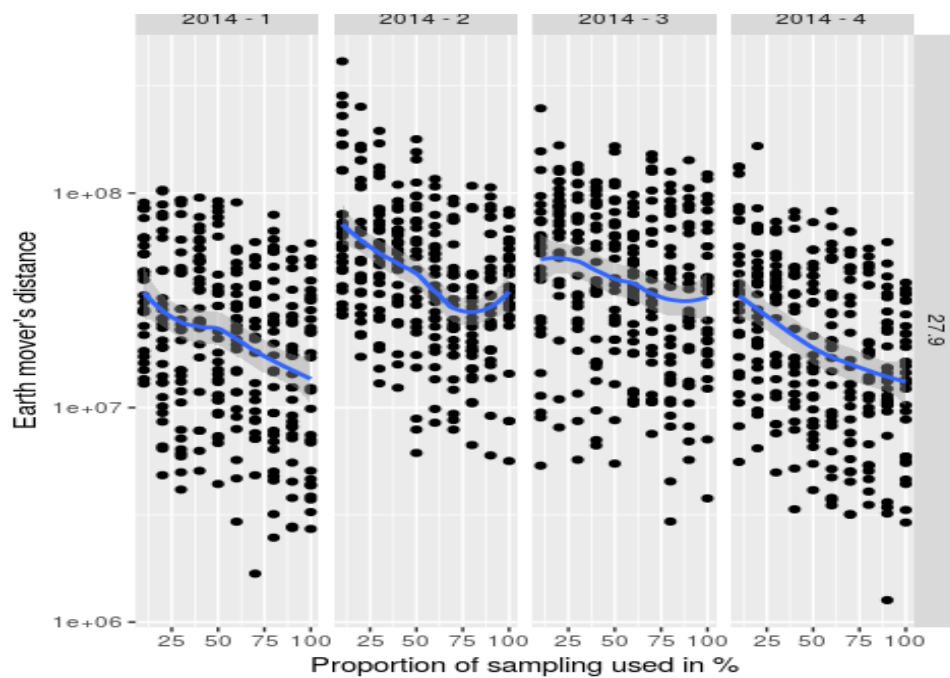


Figure 5.2.3.2 - Scenario 1: distances (y-axis) between the original length distribution and the length distribution computed with less samples (x-axis) by quarter (dots: distance for each replication, blue line: polynomial fit)

This is also confirmed by the visual inspection of the length distribution in Figure 5.2.3.3. For the other strata, the length distribution computed with 75–100% of the trips (red to dark orange dots) convey the same amount of information than the original length distribution (black line).

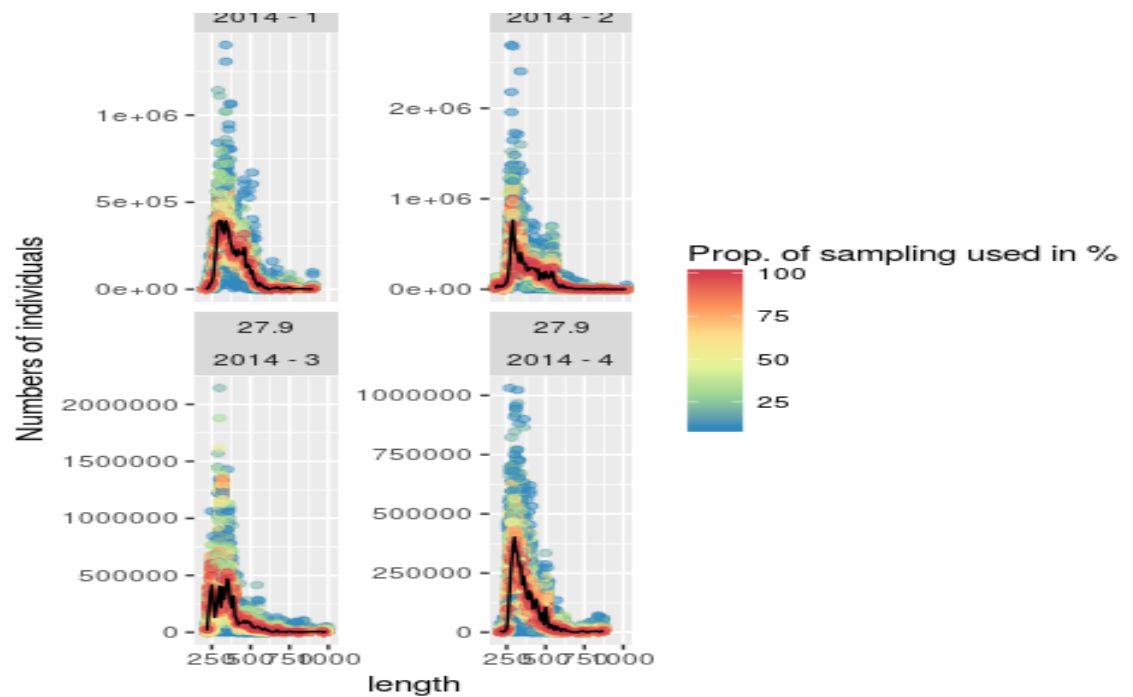


Figure 5.2.3.3 – Scenario 1: Original length distribution (black line) and the simulated ones using lower sampling intensities in terms of number of trips (coloured dots).

### Scenario 2

In this scenario, the impact of the number of fish sampled by haul inside a trip is investigated. For each haul, 10–100 fish are taken randomly with replacement from the original measurements, and length distributions are computed using the new samples. 30 simulations were performed using this scenario.

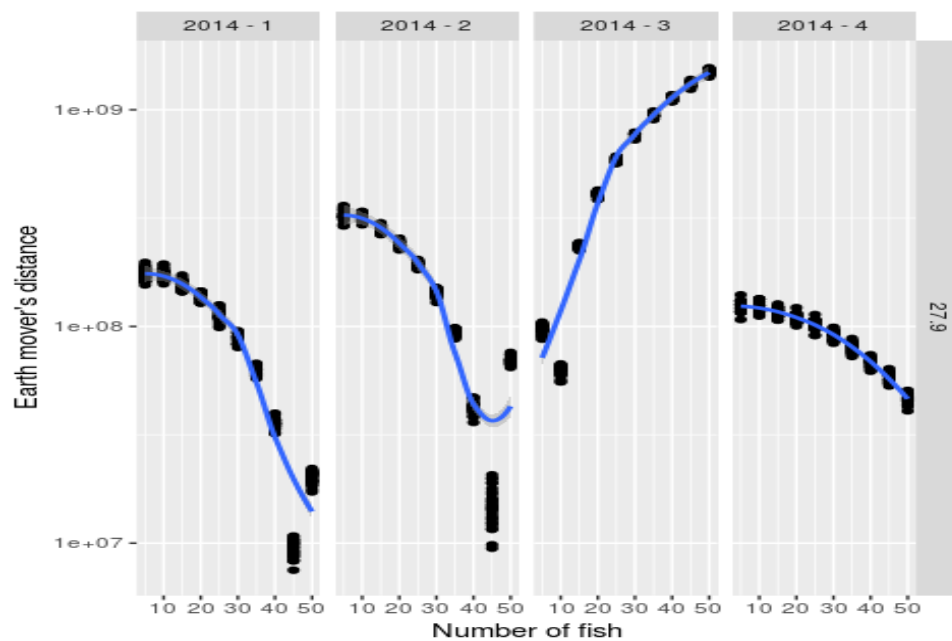


Figure 5.2.3.4 - Scenario 2: distances (y-axis) between the original length distribution and the length distribution computed with fewer measurements (x-axis) by quarter (dots: distance for each replication, blue line: polynomial fit).

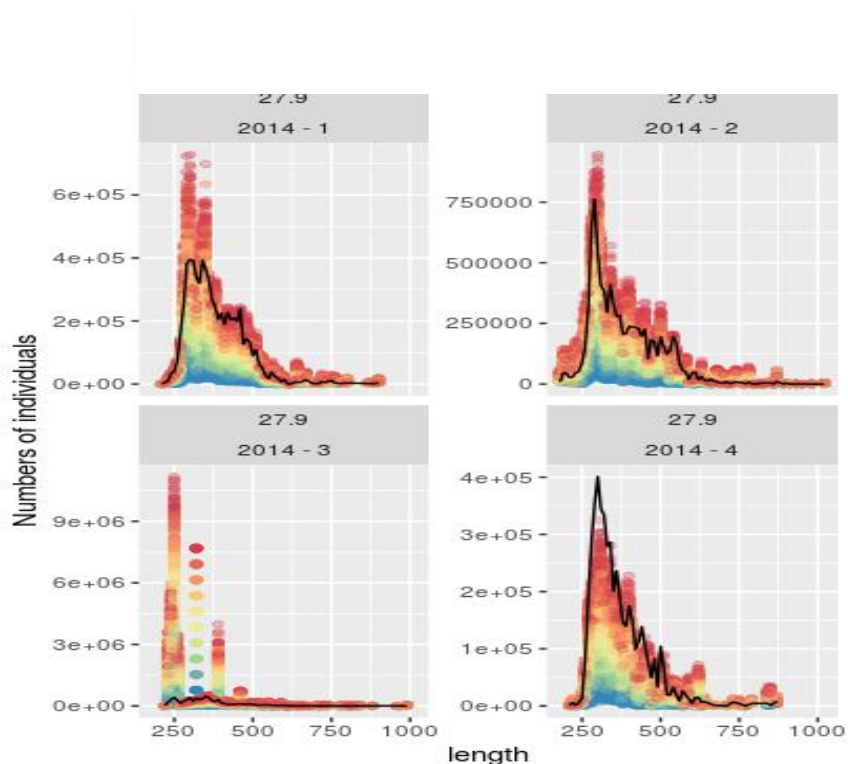


Figure 5.2.3.5 – Scenario 2: Original length distribution (black line) and the simulated ones using lower sampling intensities in terms of number of fish measured (coloured dots).

Visual inspection of Figures 5.2.3.4 and 5.2.3.5 shows that the earth mover's distance (EMD) increase sharply with the number of fish measured for the quarter 1 and 2. In this case, decreasing the amount of fish by sample, changes the length distribution. For the quarter 3 the increase is smoother, and the number of fish can be diminished. For the quarter 3, the trend is the opposite, due to original number of fish available (less than 50). Some length classes are not well represented in the original samples, as seen in the graphical representation of the simulated distribution. Bootstrapping the fish measurements brings noise in the final length distribution.

### Discussion

This procedure seems to be a good tool to be used or adapted to when analyses of the sampling effort is the aim. It also includes data pre-screening, validation and consolidation procedures (see Annex 4). Scenarios results show that considering the sampling effort in number of trips there is some space for reduction without much loss of information. In what concerns to the second scenario, it seems that the reduction of the number of fish measured must be taken with more caution. Results show that some differences in the length distribution may compromise the final estimates using this procedure. Taking this into account, maybe for this case the analysis should also consider species spatial distribution and size categories as levels of the analysis. Indications exist that these variables may influence the optimization procedure for this species (Silva *et al.*, 2017). Also effective sample size procedures in terms of number of trips, must take in consideration the kind of sampling strategies in use (concurrent vs. species focus), not to compromise the information collected for other species/stocks targeted by that same fleets.

### References

- ICES. 2011. Report of the Working Group on the Assessment of Southern Shelf stocks of Hake, Monk and Megrin (WGHMM), 5–11 May 2011, ICES Headquarters, Copenhagen. ICES CM 2011/ACOM:11.625 pp
- Silva C, Azevedo M, Chaves C, Coelho R, Costa AMC, Dinis D, Dorés S, Fernandes AC, Gonçalves P, Lino PG, Mendes H, Moura T, Nunes C, Oroszlányová M, Pinto D, Silva MC, 2017. Report of the Workshop on Sampling Effort for Biological Parameters (WKSEBP), IPMA, Lisbon 18–20 April. Relat. Cient. Téc. do IPMA (<http://ipma.pt>), nº 17, 55 p + 4 Anexos.

## 5.3 Sole in areas 28.8ab and 27.4

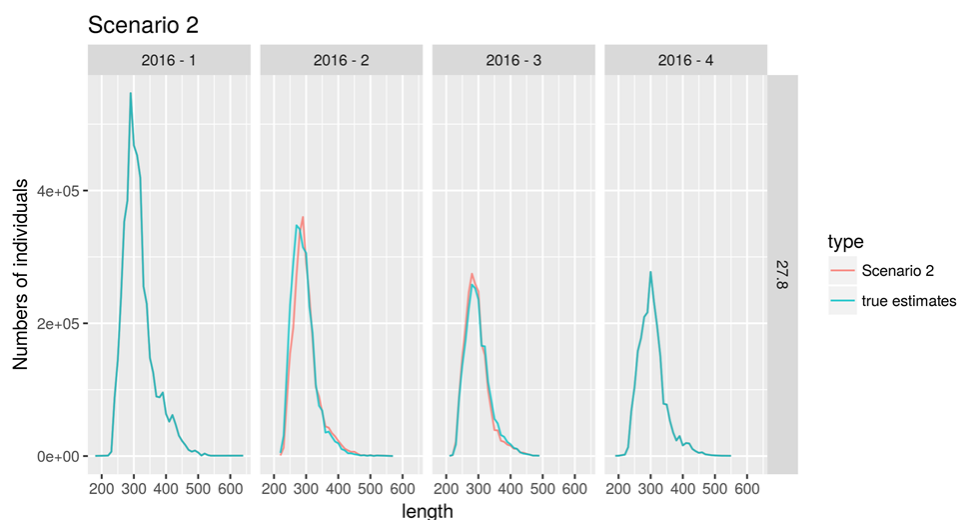
*This is a presentation of main points and outcomes for this case study. A full version of the case study with all data, analysis and figures can be found in Annex 4.*

The stock sol.27.8ab is mainly exploited by France and Belgium. In 2016, total catches were 3346 t (91% from France and 9% from Belgium) according the ICES Advice 2016. Using the commercial sampling data for the year 2016 for the countries involved in the exploitation of this stock, simulations were used to provide some indication on how sampling effort could be reallocated without loss of information regarding the length distribution at the multi-level level. After some preliminary analyses, the overall quality of the data were assessed (outlier detection...) and a common stratification set up, 6 scenarios were tested:

- Scenario 1: remove a semester of French market samples in the Bay of Biscay.
- Scenario 2: remove the Belgian samples in the Bay of Biscay.
- Scenario 3: remove the French samples in North Sea.

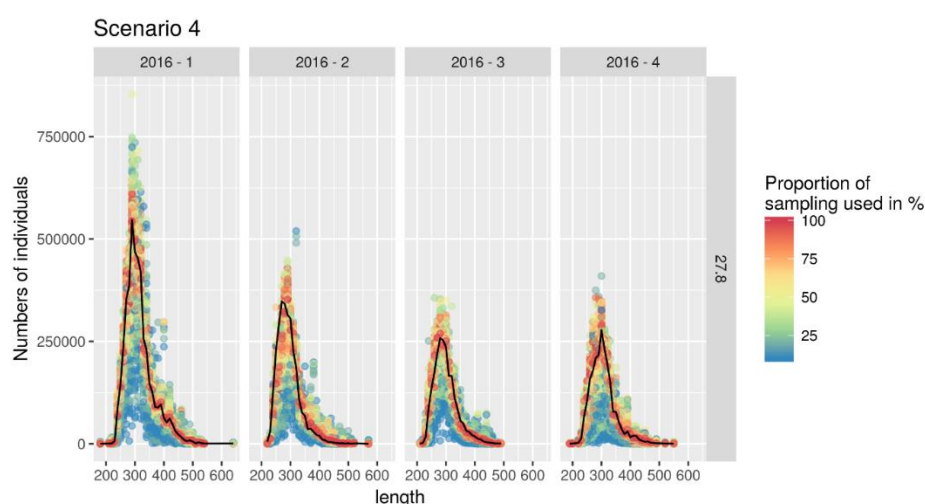
- Scenario 4: decrease French sampling effort in Bay of Biscay.
- Scenario 5: decrease the number of sampled hauls in the sample.
- Scenario 6: decrease the number of fish sampled by hauls in the sample.

Detailed results of the six scenarios are presented in the Annex 4. Main results show that in some case, some sample reallocation can be done at the regional level without loss of information regarding the estimates of the length distribution. For example, in scenario 2, the length distribution estimates are almost the same for the area where Belgian samples were removed (Figure 5.3.1).



**Figure 5.3.1 - Length distribution estimates by quarter and ICES division, estimates using the complete set of data (blue) vs. estimates where Belgian samples were removed in the division 27.8 (red).**

In the same area, the overall sampling intensity can be reduced in term of number of samples in this area up to almost 75 %, maintaining the length distribution estimates as informative as the original one (Figure 5.3.2).



**Figure 5.3.2 Length distribution estimates by quarter in division 27.8. In black is presented the estimates based on the complete set of data. The coloured dots are the estimates built using a reduced proportion of samples (from red - 100% to blue 10%). 30 simulations were performed for each level of sample reduction.**

These results underline that there is room for improvement of the sampling plans if the regional level is considered. But other scenarios highlight some high dependence on some national samples to build length population estimates for other area (scenarios 1 and 3). At the national level, optimization is generally in place as shown by these simulations. These results are preliminary and further investigation are needed to improve the simulation and the bootstrap procedure, taking into account the differences in the stratification of the sampling program (e.g. at seavs.onshore, or national differences). The idea behind this study was to provide to end-users an analytical framework to optimize sampling plans based on existing data, in a transparent way, code is embedded in this report and can be checked and data format is the fishframe (RDB) format used in the RCG's.



## 6 ToR d) Identify a road map for the discussion with end-users of optimization perspectives

---

The R-scripts developed and tests during WKBIOPTIM allow national sampling design teams and laboratory staff to start considering and carrying out their own analyses of sampling effort of national case-studies.

The rationale behind sample size analyses at biological-sample level aims deliberately at obtaining results conservative enough so that most common end-uses of data should still be properly accounted for even if final fleet-level estimates are not explicitly outputted. The rationale behind multi-level analyses of sampling effort aiming at optimizing fleet-level estimates explicitly considers one type of end-users (in the present case stock assessment) and being more focused is able to attain more significant reductions not only on the number of individuals measured but also other components of the sampling programmes (number of hauls, number of trips).

The decision between which approach to use depends on the circumstances. In cases where one end-user is clearly dominant in importance and stable in time, multi-level analyses targeting the main levels of the sampling hierarchy of programmes should be the most suitable as they are fit for specific purposes and attain more significant reductions in time and costs that can be applied elsewhere. Precautions should be taken if these analyses are carried out at a national level and subsequently pooled with data from other countries for estimation purposes as the CV of the international estimate will depend of the relative contributions from the different countries. In cases where end-users are less clearly defined or prone to changes in time, or the need for drastic reduction in time and cost or processing does not exist, sample level analyses are more precautionary and may provide for a more suitable starting point. Other circumstances such as the importance of the resource for national authorities, expectation on future research, etc. will also have to be considered in determining the exact amount of risk that can be taken when changing *status quo*. Throughout the process it is however important to keep clear to all intervening members that informed changes in sampling protocols, supported by analyses and discussions among all interested parties will lead to informed decisions. Decisions, that are superior in quality rather than maintaining goals and programmes that frequently were originally set ad-hoc and often based on round numbers (e.g. 1 kg sample, 200 individuals) or individual decisions and practical conveniences that may not have been evidence-supported.

Because optimization analyses frequently result in significant changes relative to previously established practices (e.g. in the number of samples, precision of the output data, time and costs spent involved on sampling programmes) it is fundamental that they are carried out under a clear road-map. Such road map should involve all people with a stake in the process (i.e. samplers, data analysts, end-users, and programme managers). It should include clarification (to all parties involved) of the analysis that will be carried out, their need, and proper consideration of the needs of both lab/field staff and the end-users of the data, to secure acceptance and support to its outcomes. A suggestion for one such roadmap is depicted in Figure 6.1.

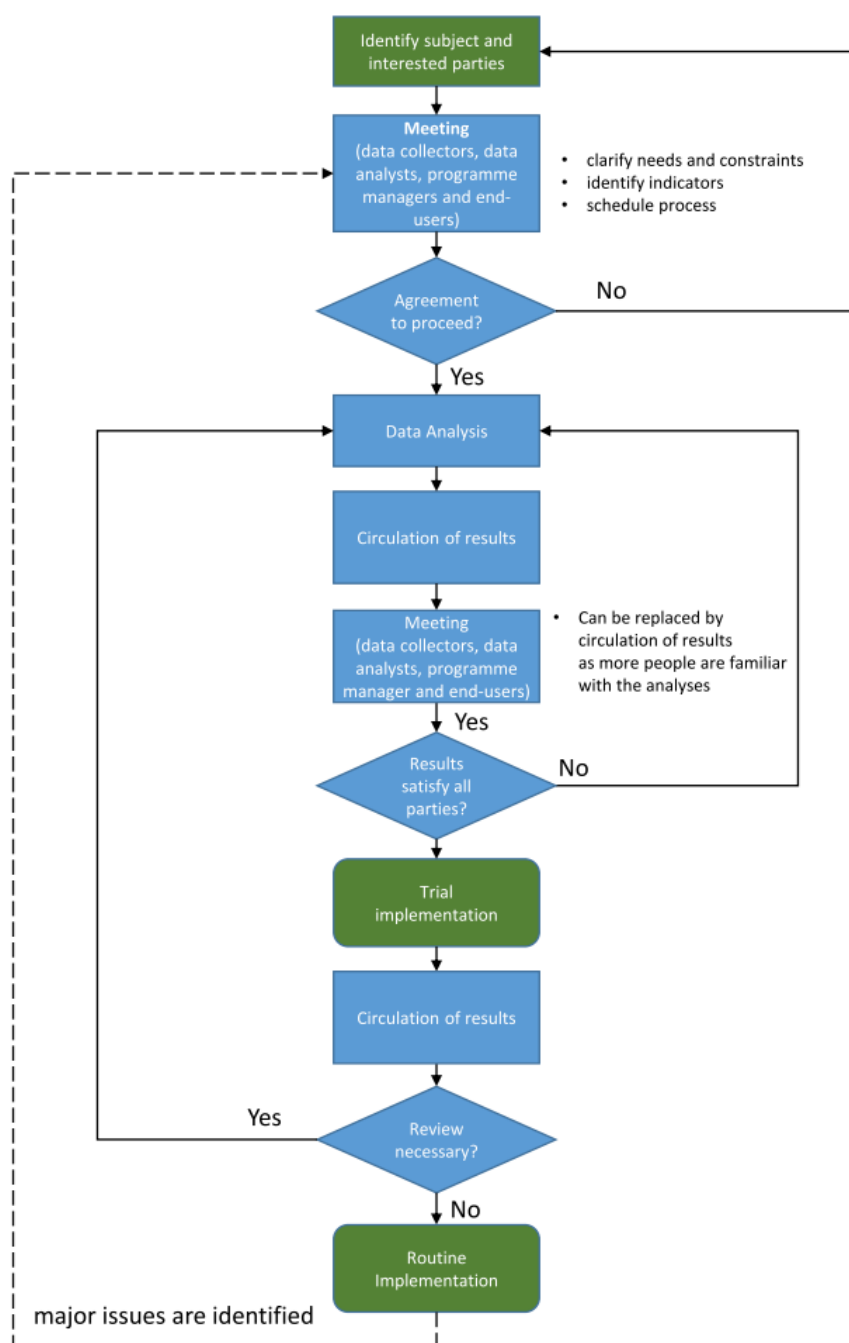


Figure 6.1 - Suggested roadmap for discussion of optimization processes.

## 7 Conclusions and future work

---

With an ever increasing focus on supplying data of a high quality to end-users each national laboratory is coming under additional pressure both in regards to time and costs involved with the sampling of biological data. WKBIOPTIM aims to look at how resources can be optimised and thus facilitate an overall improvement of data quality in respect to biological parameters. The workshop was found to be useful by the participants and steps have been taken at some national laboratories where the optimization procedures developed at the workshop have been used to begin dialogue with the end-users and form discussion on the possibilities of reducing the number of specimens taken in a sample.

In addition positive feedback was received from the RCM meetings, with a recommendation from the RCM NA supporting the continuation of this type of work. During the RCG Med & BS-LP, 2017 a practical session on sampling stratification and optimization was held. A case study was developed where both the *SamplingDesign* tool delivered under MARE/2014/19 Med & BS project and the sample level scripts developed at the WKBIOPTIM were tested in order to find common and/or complementary aspects to be further explored in the two different approaches (RCG Med & BS-LP Final Report, 2017 in prep). It was not anticipated that so much time would be spent at the workshop on data adaptation and code manipulation but through these processes the participants developed a better understanding of the optimization procedures and the steps involved. This will allow for an easier application of the optimization processes in their national laboratories at a later date. Many elements from the initial work plan were achieved and in the proposal for a follow up workshop on optimization of biological parameters (WKBIOPTIM2) additional analyses will include how to use sample weight to indicate sample size, how to consider maturity and sex ratio and/or combined vs. sex, and how to consider age when data are stratified by length. It is expected that there will be input from the Working Group on Biological Parameters (WGBIOP) and the Working Group on Commercial Catches (WGCATCH) on indicators of sample quality, which can be used in communicating the need and effects of statistical optimization of sampling to end-users (e.g. effective sample size, variability of mean length etc.). Both time and expertise was lacking to be able to fully answer the ToR on quality indicators but a recommendation is being put to WGBIOP and WGCATCH requesting a review of available indicators (including those already in the scripts) and to provide input on additional quality indicators. It is envisioned that this will be implemented as part of the new triannual work plan for WGBIOP. In addition to having input from WGBIOP, a recommendation is being put to WGCATCH to discuss, comment on and give feedback on the outcomes of the workshop in what concerns sample level and multi-level analyses.

In preparation for the next workshop close communication between WGBIOP and WGCATCH is expected with the aim to identify suitable case studies where there are clear examples of oversampling resulting from combinations of sampling strategies (stratified, two stage, etc.) and variables (length, age, sex, etc.) to be collected. There will be a stronger emphasis on the need for data quality checks and data preparation prior to the workshop to ensure that time allocated to working on case studies and answering the relevant ToR's is sufficient. It is recommended that the optimization procedures developed at WKBIOPTIM will be tested at the national laboratories in the coming months and the output will be used to begin discussions on optimization with the end-users (both the data collectors and those using the data). It is envisioned that there will be a Skype or WebEX meeting held six months following the workshop

where feedback will be given from the national laboratories. Further testing of the code will take place prior to WKBIOPTIM2 to ensure that the time spent at the workshop is used efficiently and participants should make themselves familiar with the required R packages (COST, R Markdown) and functions in advance.

## 8 References

---

- Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J. Horton. 2014. "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations in Statistics Education* 8 (1).
- Authors. (in prep). Workshop on Sampling Design and Optimization (Lysekil, 31 October–4 November 2016, unpublished)
- COMM. IMPL. DEC. (EU) 2016/1701 of 19 August 2016 laying down rules on format for the submission of work plans for data in the fisheries and aquaculture sectors.
- Gentleman, Robert, and Duncan Temple Lang. 2004. "Statistical Analyses and Reproducible Research." Bioconductor Project Working Papers. Working Paper 2. <http://www.bepress.com/bioconductor/paper2>.
- GFCM, 2017. GFCM Data Collection Reference Framework (DCRF). Version: 2017.1
- ICES. 2008. Report of the Workshop on Methods to Evaluate and Estimate the Accuracy of Fisheries Data used for Assessment (WKACCU), 27–30 October 2008, Bergen, Norway. ICES CM 2008\ACOM:32. 41 pp.
- ICES. 2010. Report of the Workshop on methods to evaluate and estimate the precision of fisheries data used for assessment (WKPRECISE), 8–11 September 2009, Copenhagen, Denmark. ICES CM 2009/ACOM:40. 43 pp.
- ICES. 2012. Report of the Working Group on Practical Implementation of Statistical Sound Catch Sampling Programs (WKPICS), 8–10 November 2011, Bilbao, Spain. ICES CM 2011 / ACOM:52. 55pp.
- ICES. 2012. Report of the Planning Group on Commercial Catches, Discards and Biological Sampling (PGCCDBS). 18–22 February 2013, Belfast, Northern Ireland. ICES CM 2013/ACOM: 49. 128pp.
- ICES. 2013. Report of the second Workshop on Practical Implementation of Statistical Sound Catch Sampling Programmes (WKPICS2), 6–9 November 2012, ICES Copenhagen. ICES CM 2012 / COM:54 71pp.
- ICES. 2014. Report of the third Workshop on Practical Implementation of Statistical Sound Catch Sampling Programmes (WKPICS3), 19–22 November 2013, ICES HQ, Copenhagen, Denmark. ICES M2013/ACOM: 54. 109 pp.
- ICES. 2014. Implementing the ICES Strategic Plan 2014–2018. Linking science, advice, data and information, and secretariat. ISBN 978-87-7482-146-5. Available at [https://is-suu.com/icesdk/docs/implementation\\_plans](https://is-suu.com/icesdk/docs/implementation_plans)
- ICES. 2016a. Report of the Working Group on Data Needs for Assessments and Advice (PGDATA), 30 June–3 July 2015, Lysekil, Sweden. 103 pp.
- ICES. 2017a. Report of the Workshop on cost benefit analysis of data collection in support of stock assessment and fishery management (WKCOSTBEN), 28 June -1 July 2016, ICES Headquarters, Copenhagen, Denmark. ICES CM 2016/SSGIEOM: 12. 41pp.
- ICES, 2017b. Report of the Working Group on Commercial Catches (WGCATCH), 7–11 November 2016, Oostende, Belgium. ICES CM 2016/SSGIEOM:03. 142pp.
- Jansen, Teunis, Henrik Degel, Joel Vigneau, and Ernesto Jardim. 2009. "Definition of Standard Data-Exchange Format for Sampling, Landings, and Effort Data Commercial Fisheries." ICES COOPERATIVE RESEARCH REPORT 296. ICES.
- MARE/2014/19 Med&BS project Final report, 2016 - Deliverable 2.5 - Scripts with a detailed document ensuring the reproducibility of the methods and the analyses realized with a special focus on the case studies tests - F. Gontrand, T. Rouyer, 2016.

- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>. <https://www.r-project.org>
- Report of the 2017 Regional Co-ordination Meeting Mediterranean and Black Sea Large Pelagic Fisheries (in prep)
- Rubner, Y., C. Tomasi, and L. J. Guibas. 1998. "A Metric for Distributions with Applications to Image Databases." *IEEE International Conference on Computer Vision*, 59–66.

## Annex 1: List of participants

NAME	ADDRESS	PHONE	E-MAIL
Ana Claudia Fernandes	Instituto Português do Mar e da Atmosfera, I.P., Rua Alfredo Magalhães Ramalho, 6 1495-006 Lisboa, Portugal	+351 21 302 70 00 (ext. 1385)	acfernandes@ipma.pt
Ana Costa	Instituto Português do Mar e da Atmosfera, I.P., Rua Alfredo Magalhães Ramalho, 6 1495-006 Lisboa, Portugal	+351 21 302 70 00	amcosta@ipma.pt
Annelie Hilvarsson	Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Turistgatan 5, 453 30 Lysekil, Sweden	+46 10 4784030	annelie.hilvarsson@slu.se
Carmen Hernandez	Instituto Español de Oceanografía (IEO), C. O. de Santander, Promontorio San Martín s/n, 39004 Santander, España	+34 942291716	carmen.hernandez@st.ieo.es
Corina Chaves	Instituto Português do Mar e da Atmosfera, I.P., Rua Alfredo Magalhães Ramalho, 6 1495-006 Lisboa, Portugal	+351 21 302 70 00	corina@ipma.pt
Dália Cristina Cunha Reis	Departamento de Oceanografia e Pescas – DOP, Rua Prof. Doutor Frederico Machado, 4, 9901-862, Horta, Faial – Açores, Portugal	+351 292 200 400	dalia.cc.reis@uac.pt
Estanis Mugerza	AZTI Sukarrieta Txatxarramendi ugarte a z/g, E-48395 Sukarrieta (Bizkaia), Spain		emugerza@azti.es
José Rodriguez	Instituto Español de Oceanografía (IEO), C. O. de Santander, Promontorio San Martín s/n, 39004 Santander, España	+34 942291716	jose.rodriguez@st.ieo.es
Julia Wischnewski	Institute of Sea Fisheries, Palmaille 9, 22767 Hamburg-Altona, Germany		julia.wischnewski@thuenen.de
Julie Coad Davies	DTU Aqua, Anker Engelunds Vej 1, Bygning 101A, 2800 Kgs. Lyngby, Denmark	+45 35 88 33 38	joco@aqua.dtu.dk
Kirsten Birch Håkansson	Denmarks Technical University, DTU Aqua, Anker Engelunds Vej 1, Bygning 101A, 2800 Kgs. Lyngby, Denmark	+45 35 88 30 34	kih@aqua.dtu.dk
Laurent Dubroca	Ifremer, Port-en-Bessin Station, PO Box 32, Avenue du Général De Gaulle, 14520 Port-en-Bessin, France		laurent.dubroca@ifremer.fr
María Soto Ruiz	Instituto Español de Oceanografía (IEO), Subdirección, General de Investigación, Corazón de Maria, 8 28002 Madrid, España	+34 91 510 7533	maria.soto@md.ieo.es

Maria Teresa Facchini	COISPA TECNOLOGIA & RICERCA, Stazione Sperimentale per lo Studio delle Risorse del Mare, Via dei Trulli, 18/20, 70126, Bari, Torre a Mare, Italy	+39 080 5433596	facchini@coispa.it
Natoya Jourdain	Institute of Marine Research, P.O.Box 1870 Nordnes, 5817 Bergen, Norway		natoya.jourdain@imr.no
Nuno Prista	Swedish University of Agricultural Sciences, Department of Aquatic Resources, Institute of Marine Research, Turistgatan 5, 453 30 Lysekil, Sweden	+46 10 478 40 22	nuno.prista@slu.se
Pedro Lino	Instituto Português do Mar e da Atmosfera,		plino@ipma.pt
Pedro Torres	Instituto Español de Oceanografía (IEO), C. O. de Málaga, Puerto Pesquero s/n, 29640 Fuengirola, España	+34 95 219 7124	pedro.torres@ma.ieo.es
Petra Bratina	Fisheries Research Institute of Slovenia, Sp. Gameljne 61a, SI-1211 Ljubljana-Šmartno, Slovenia	+ 386 (0)1 24 43 412	Petra.Bratina@zzrs.si
Rantanen Perttu	Natural Resources Institute Finland (Luke) Itäinen Pitkätatu 3, 20520 Turku, Finland	+29 532 7680	Perttu.Rantanen@luke.fi
Sofie Vandemaele	ILVO, Ankerstraat 1, 8400 Oostende, Belgium		Sofie.Vandemaele@ilvo.vlaanderen.be
Zachary Radford	Cefas, Pakefield Road, Lowestoft, Suffolk, NR33 0HT, UK	+44(0) 1502 524292	zachary.radford@cefas.co.uk





## Annex 2: Agenda

<b>Tuesday 20<sup>th</sup> June</b>		
9.00 – 9.30	Welcome and logistics	Chairs
9.30 – 10.00	Presentation of ToR's and adoption of agenda	Chairs
10.00 – 11.00	Presentation of code used for optimization at sampling level	Nuno Prista
11.00 – 11.30	<i>Coffee</i>	
11.30 – 12.00	Presentation of code used for optimization at trip level	Laurent Dubroca
12.00 – 12.30	Presentation of RCM Med &BS Training Workshop on Sampling Intensity	Maria Teresa Facchini
12.30 – 13.00	Questions and Answers	Plenary
13.00 – 14.00	<i>Lunch</i>	
14.00 – 15.00	Questions and Answers	Plenary
15.00 – 16.00	Discussion of case studies and subgroups	Plenary
16.00 – 16.30	<i>Coffee</i>	
16.30 – 18.00	Work on data modifications and set up of scripts and packages required	Plenary
<b>Wednesday 21<sup>st</sup> June</b>		
9.00 – 9.30	Discussion of issues with code adaptation	Plenary
9.30 – 11.00	Subgroup work	
11.00 – 11.30	<i>Coffee</i>	
11.30 – 13.00	Troubleshooting and Subgroup work	
13.00 – 14.00	<i>Lunch</i>	
14.00 – 16.00	Troubleshooting and Subgroup work	
16.00 – 16.30	<i>Coffee</i>	
16.30 – 17.30	Troubleshooting and Subgroup work	
17.30 – 18.00	Wrap up of the day	
Social Dinner		
<b>Thursday 22<sup>nd</sup> June</b>		
9.00 – 9.30	Discussion of issues with code adaptation	Plenary
9.00 – 11.00	Subgroup work	
11.00 – 11.30	<i>Coffee</i>	
11.30 – 13.00	Subgroup work	
13.00 – 14.00	<i>Lunch</i>	
14.00 – 15.00	Report writing of case study work	
15.00 – 16.00	Report preparation and discussion of future work	Plenary
16.00 – 16.30	<i>Coffee</i>	
16.30 – 17.30	Report writing	Plenary
17.30 – 18.00	Wrap up of the meeting	Chairs

### **Annex 3: List of oral presentations**

---

Maria Teresa Facchini (COISPA Italy) Input from Training Workshop on sampling intensity - use of tool developed by MARE/2014/19 Med & BS project Maria Teresa FACCHINI COISPA (Italy)

Nuno Prista (SLU Sweden): Sample level optimization and R-Code Manual (early draft) Version 1.0, 2017-06-20

Laurent Dubroca (Ifremer France): WKBIOPTIM: A tool to play with sampling program

## Annex 4: Case study documentation

---

### A – Optimization of the sampling plan for the sardine in divisions 8c and 9a using Spanish data

#### *Introduction*

The Southern stock of sardine (*Sardina pilchardus* in divisions 8.c and 9.a, sar-soth) is exploited by Spain and Portugal. This analysis focuses on the Spanish sampling data. In 2015, total Spanish catches were 6986 t. Using the commercial sampling data for the year 2015, this document presents a framework where length distribution estimates are built using different subsamples of the original database. The way how these subsamples are built are called 'scenario'.

This report, using numerical simulation, provides some indication on how sampling effort could be reallocated, keeping the same quality of information for ONE species and ONE country. The simulation were performed without taking into account the information related to other stock (concurrent sampling) and to other parameters (discards), and without any consideration to practical issues related to the organization of the sampling plan.

#### *Framework*

Sardine in 27.8.c and 27.9.a.

#### *Fleet*

Sardine is mainly captured by purse-seiners both in Spain and Portugal.

The Spanish fleet targets anchovy (*Engraulis encrasicolus*), mackerel (*Scomber scombrus*), horse-mackerel (*Trachurus trachurus*) and sardine (*Sardina pilchardus*) with some characteristic seasonal patterns for some of the species (anchovy in spring; mackerel in winter). The gear must have a minimum mesh size of 14 mm.

#### *Sampling design*

Overall Spanish sampling design in 2015 was based in métiers following the ranking system according to the DCF. For purse-seiners, sampling frame was constructed with a list of ports based on effort and landings where days and trips were selected. Frame was stratified by quarter with a systematic monthly allocation of effort. Full concurrent sampling (all species) was in place.

#### *Statistical analyses*

The framework developed for this study integrates the whole processing of fishery data to (1) build population estimates at the stock level (length distribution mainly) and (2) to analyse the impact of different sampling strategies in these estimates by simulation. This framework is aimed for reproducibility and transparency, following the recommendation of the reproducible research statement (Gentleman and Lang 2004). Consequently this report is self-consistent: the code used to process and to analyse the fishery data are embedded in the report itself.

#### *Software*

Analyses are carried out using the R environment (R Core Team 2012). R<sup>3</sup> is a free software environment for statistical computing and graphics. The reproducibility of the

---

<sup>3</sup><http://www.r-project.org/>.

results presented in this report relies on the use of a dialect of the Markdown language called Pandoc for word processing using the Knitr R package. Markdown is a plain text formatting syntax designed so that it can optionally be converted to HTML using a tool by the same name. Pandoc<sup>4</sup> is a Markdown dialect which extends the conversion capability to word processing file (docx, doc and odt) and pdf, among other formats. Pandoc understands a number of markdown syntax extensions, including document metadata (title, author, date), footnotes, tables, figures and references. Knitr<sup>5</sup> is an R package (a set of functions extending the R capabilities). With this package, the R code used to process and analyse the data are included directly in the report. Results are then produced dynamically. This framework has demonstrated the capacity to improve the conduct and the presentation of data analysis in a way that another person can understand and replicate (Baumer *et al.* 2014).

For example, if the calculus of 1+1 is needed, the code to compute it is written in the report using special hooks, as in this simple example:

```
'''{r test00,warn=FALSE,cache=TRUE,echo=TRUE}
#comment: addition example.
1+1
'''
```

This code is evaluated during the compilation of the report by the knitr command and it prints the following result:

```
#comment: addition example.
1+1

## [1] 2
```

The result is 2. All the numerical values related to the data, including tables and figures are produced following this procedure. Some hooks are appearing in the final report, other were removed to clarify the outputs.

### Data

Data inputs follow the Regional DataBase (RDB) Exchange Format version 2009 (Jansen *et al.* 2009). It's an older version of the actual format used in the Fishframe database (Jansen *et al.* 2016). Three data objects are used:

the object CL (Commercial Landings) brings population data related to landing,

the object CE (Commercial Effort) brings population data related to effort,

the object CS (Commercial Samplings) brings sampling data related to effort.

### Installation

The document needs the COST package (section software of the WKBIOPTIM share-point). To install them:

```
#the path to the library is local (aka where are the files on your hard
disk)
install.packages("/path/to/COSTcore_1.4-0.zip",repos=NULL)
install.packages("/path/to/COSTdbe_1.4-1.zip",repos=NULL)
install.packages("/path/to/COSTeda_1.4.0.zip",repos=NULL)
```

Some other packages are needed and can be installed using CRAN:

---

<sup>4</sup><http://johnmacfarlane.net/pandoc/>.

<sup>5</sup><http://yihui.name/knitr/>.

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("DATRAS")
install.packages("mapdata")
install.packages("pander")
install.packages('DATRAS', repos='http://www.rforge.net/', type='source')
devtools::install_github("ldbk/earthmovdist")
```

The library needs to be loaded in R. Moreover a set of functions are available in the `wkbioptim_fct.R` files and have to be sourced. Some options related to how the chunks of code are interpreted are fixed (size of the figures...):

```
#R general option:
options(stringsAsFactors = FALSE)
#working directory:
#setwd("C:/path/to/the/working/directory")
#chunk option
#knitr option
knitr::opts_chunk$set(cache=TRUE, echo=TRUE, warning=FALSE,
  message=FALSE, fig.height=6, progress=FALSE, verbose=FALSE,
  include=TRUE, dev='png', autodep=FALSE)
#package
library(COSTcore)
library(COSTdbe)
library(COSTeda)
library(dplyr)
library(ggplot2)
library(maps)
library(mapdata)
library(pander)
library(DATRAS)
library(earthmovdist)
#source local file
source("wkbioptim_fct.R")
```

### Data

National data are loaded and combined. Some corrections are made to homogenize the data, due to some discrepancies of variables type.

```
load("CSr.RData")
CSr@tr$vs1Id <- as.numeric(as.factor(CSr@tr$vs1Id))
load("CLr.RData")
```

Sardine (species id: 126421) is selected for the areas 8c and 9a:

```
#subset spp
CSr<-subsetSpp(CSr, spp=="126421", table="s1") # sardine
CLr<-subset(CLr, taxon=="126421", table="c1")
#subset in space
CSr<-subset(CSr, area%in%c("8c", "9a"), table="hh")
CLr<-subset(CLr, area%in%c("8c", "9a"), table="c1")
#subset technical
CSr<-subset(CSr, foCatEu6%in%unique(CLr@c1$foCatEu6), table="hh")

save(CSr, CLr, file="pipo.rdata")
```

### Exploratory data analysis

In this section, a short exploratory analysis of the data are provided.

#### Global figures

Landings and sampling effort are computed by country and ICES Subdivision.

### Sardine fishery for Spain in 2015

COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	FISH MEASURED
ESP	8c	1933	35	3225
ESP	9a	5053	80	5510

### Métier

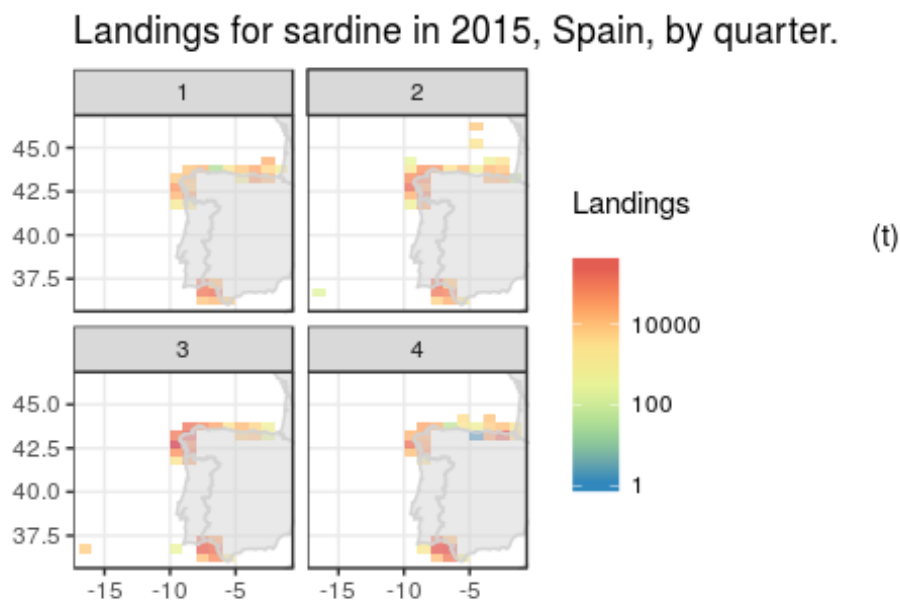
Landings and sampling effort by métier are summarized. The number of métier is reduced using the ICES rules on main métier: here métiers fishing less than 90% of the cumulated catches are coded as MIS\_MIS\_0\_0\_0.

Landings and sampling effort by main métier, country and subarea in 2016. MIS\_MIS\_métier is generated using the ICES rules.

MÉTIER	COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	HAULS SAMPLED	FISH MEASURED
PS_SPF-0_0_0	ESP	8c	1916	35	35	3225
MIS_MIS-0_0_0	ESP	8c	16.63	0	0	0
PS_SPF-0_0_0	ESP	9a	4902	72	72	5023
MIS_MIS-0_0_0	ESP	9a	151.4	8	8	487

### Landings

Landings are mapped by ICES rectangle, quarter and country.



### Stratification

The stratification defines the spatial, temporal and technical support on which population estimates are computed. The stratification follows the ICES datacall policy: quarter, ICES area and métier level 6.

Stratification

timeStrata	spaceStrata	techStrata
quarter	area	foCatEu6

### Validation and consolidation

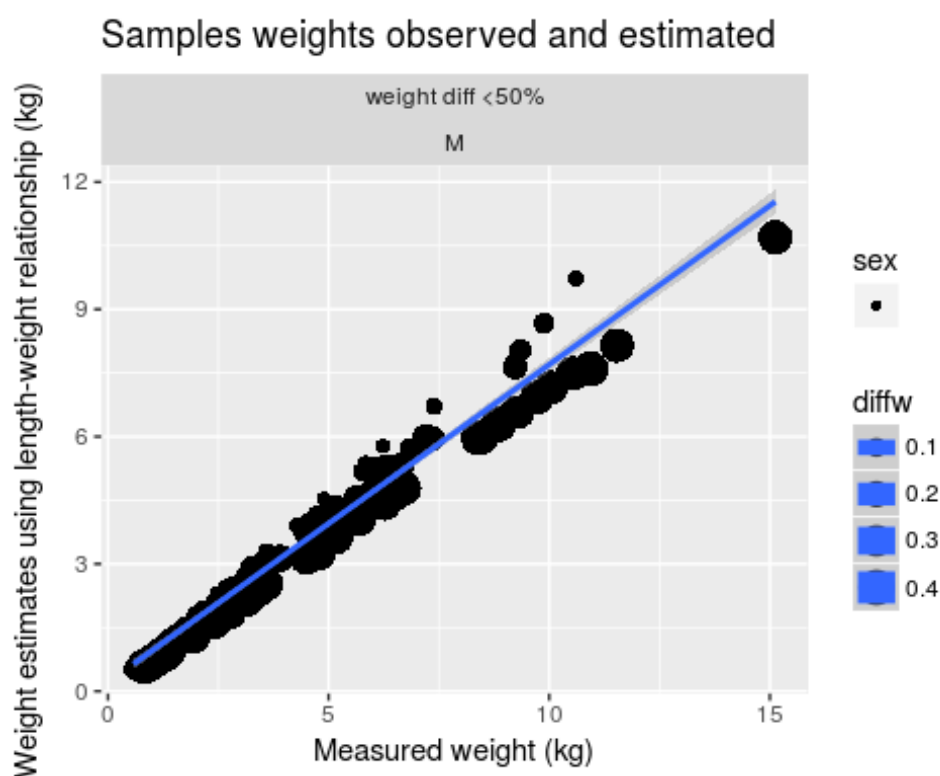
According to the stratification, the data are validated: samples are tested using quality checks.

### Quality checks

Sample weights

Subsample weights are checked using a generic sardine weight length relationship. If the sample weight differs from the theoretical weight by 50% then this sample is removed.

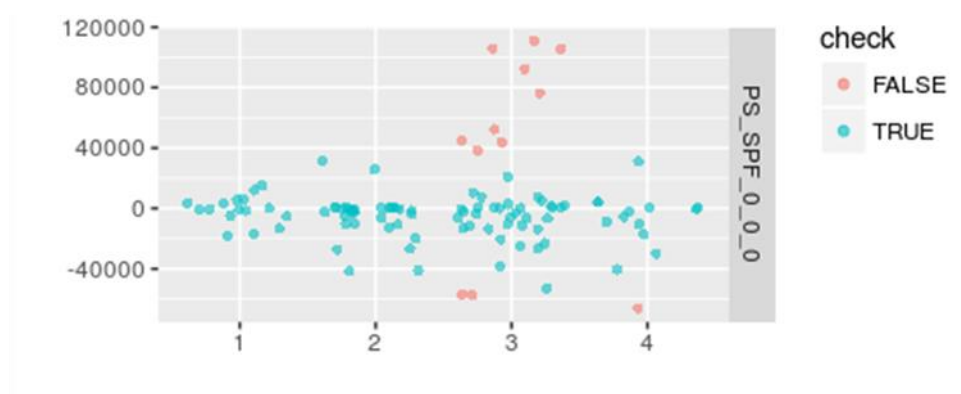
```
CSr<-corrbase(CSr)
rtp<-data.frame(sciname="Sardina pil",a=0.000006,b=3.0627)
rez<-corrsampw(CSr,rtp)
CSr<-rez[[1]];
pltsampw<-rez[[2]]
try(plot(pltsampw),silent=T)
```



### Sample outliers

Sample outliers are tested using the delta approach following Vigneau and Mahevas (2007). This index helps to detect sample outliers and sample heterogeneity when catch-at-length is estimated using the ratio estimator. In this report, in order to ensure the repeatability of our results, a fixed threshold is applied to detect outlier. In a strata, all the samples with delta values outside the quantile at 2.5% and 97.5% are flagged as outliers. The delta value outliers' detection for sample are computed but not used in this report:

```
rezdelta<-delta(CSr,myStr,idspecies="126421")
try(plot(rezdelta[[2]],silent=T))
```



### Validation and consolidation

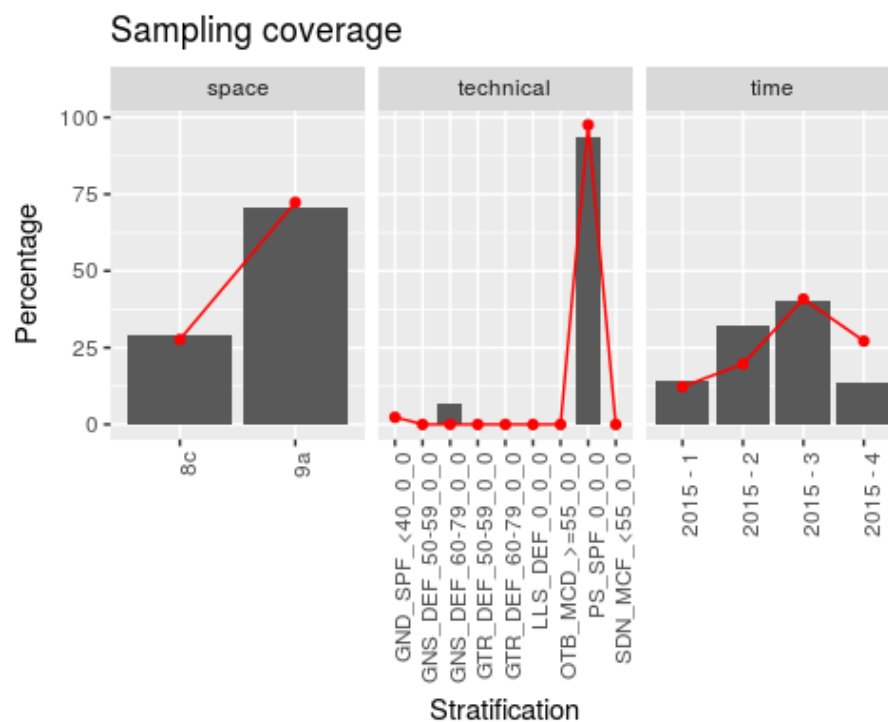
Validation and consolidation (i.e. application of the stratification to the data) are done using the ad-hoc COSTcore functions:

```
#CSr<-rezdelta[[1]]
CSv <- csDataVal(CSr) ; CSc <- csDataCons(CSv,myStr)
CLv <- clDataVal(CLv); CLc <- clDataCons(CLv,myStr)
#CEv <- ceDataVal(CEr); CEc <- ceDataCons(CEv,myStr)
if(F){
  CSc<-alkLgthRec(CSc,type='stepIncr',10,preview=FALSE,post-
view=FALSE,update=TRUE)
}
save(CSv,CLv,CSc,CLc,file="datavalcons.Rdata")
```

### Sample coverage

The sample coverage is checked according to the stratification: the percentage in numbers of samples by strata is presented against the percentage of landings in each modality (red line).





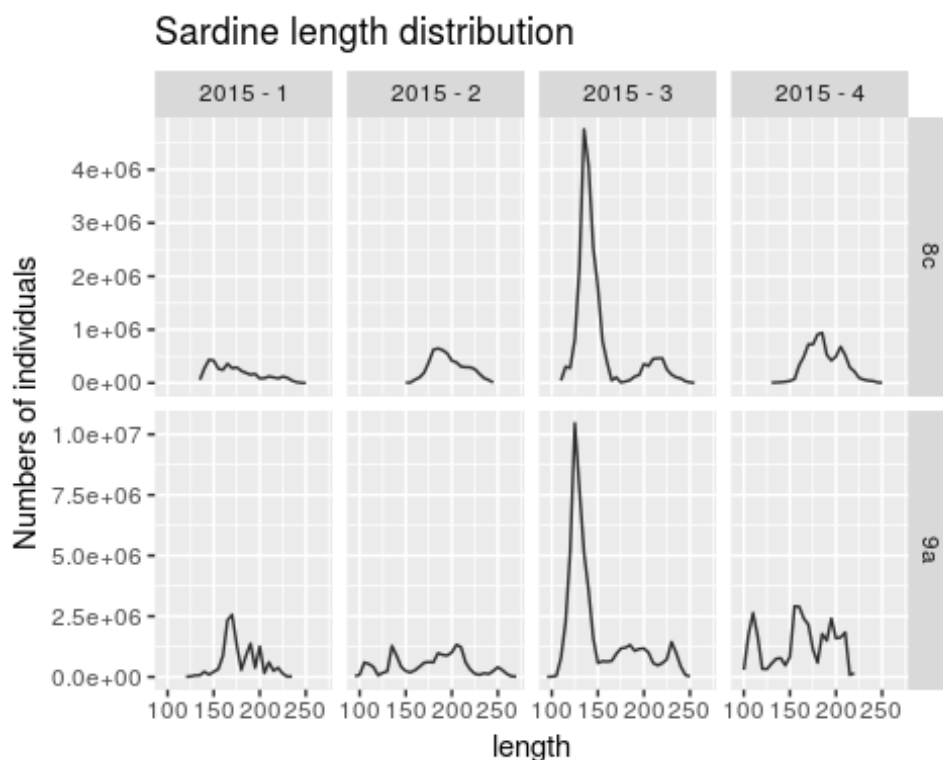
### *Raised length composition of the stock*

The length composition is estimated using ratio-estimator based on landings.

```
rezdolen<-dolen(CSc,CLc,myStr,idspecies="126421")
dbelan<-rezdolen$dbelan
save(dbelan,file='dbelan.Rdata')
pltlenlan<-graphlen(dbelan,idstock="pil.8c.9a",rtp,checkn=50,checks=3)
#plot(pltlenlan)
```

To summarize our graphical output, length distributions are combined by ICES Sub-area and quarter.

```
sizefish<-dbelan@lenStruc$estim%>%group_by(space=sub-
str(space,1,4),time,length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="true estimates")
plt1<-ggplot(rbind(sizefish),aes(x=length,y=value))+
  facet_grid(space~time,scales="free_y")+#,space="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Sardine length distribution")
plt1
```



### Scenarios

From the first analyses, the length distribution of the sardine in the areas 27.8.c and 27.9.a are estimated using the dataset collected by Spain in 2015. These length distributions are considered the "true" estimates of the information caught. This section is now dedicated to the estimation of the same parameters, with changes in the sampling plan. These changes are scenario testing some modification of the sampling plan and reallocation of the sampling effort. To do so, samples are removed according to the scenario, in a fixed or random way, and length distribution are estimated based on the new samples base. The results are compared to the "true" estimates using the original number of samples. To compare the 2 distribution, the Earth Mover's Distance (EMD) is used. In statistics, the (EMD) is a measure of the distance between two probability distributions over a region  $D$  (names Wasserstein metric in mathematics) (Rubner, Tomasi, and Guibas 1998). Informally<sup>6</sup>, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region  $D$ , the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.

The 2 scenario tested in this report are:

Scenario 1: decrease the number of sampled trips.

Scenario 2: decrease the number of fish sampled by trip.

<sup>6</sup>from [https://en.wikipedia.org/wiki/Earth\\_mover%27s\\_distance](https://en.wikipedia.org/wiki/Earth_mover%27s_distance)

## Scenario 1

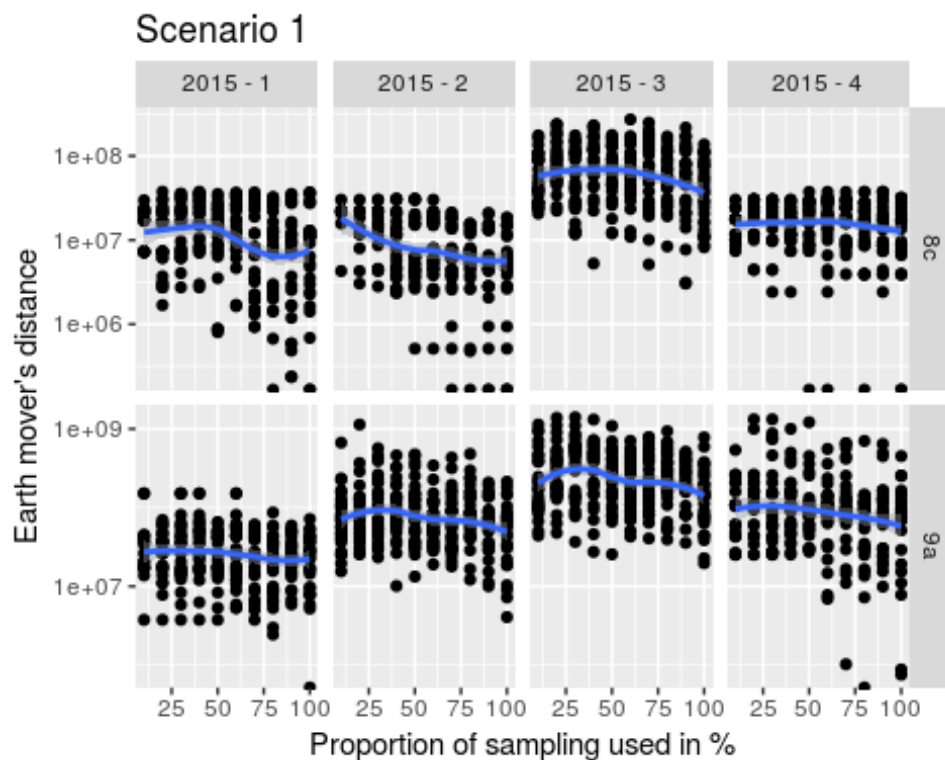
In this scenario, the sampling effort was decreased, from 100% to 10% of the number of original trip sampled. Trips are sampled randomly with replacement in 30 simulations.

```

if(file.exists("sizefish4.rdata")){
  load("sizefish4.rdata")
}else{
  sizefish2<-data.frame()
  idtrip<-unique(CSc@hh$trpCode)
  for(i in 1:30){
    for(j in seq(0.1,1,0.1)){
      print(paste(i,j))
      idtrip0<-sample(idtrip,length(idtrip)*j,replace=T)
      CSctmp<-subset(CSc,trpCode%in%idtrip0,table="hh")
      rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="126421")
      dbelantmp<-rezdolen$dbelan
      sizefishtmp<-dbelantmp@lenStruc$estim%>%group_by(time,space=substr(space,1,4),length)%>%
        summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
        mutate(length=as.numeric(length),type="scenario
4",prop=j,rep=i)
      sizefish2<-rbind(sizefish2,sizefishtmp)
    }
  }
  save(sizefish2,file="sizefish4.rdata")
}

#compute emd
emd1<-left_join(sizefish2%>%transmute(space,time,length,prop,rep,val1=value),
  sizefish%>%transmute(space,time,length,val0=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time,prop,rep)%>%summarise(emd=emdL1(val1,val0))
plt1<-ggplot(emd1,aes(x=100*prop,y=emd))+geom_point()+
  scale_y_log10()+facet_grid(space~time,scale="free_y")+geom_smooth()+
  xlab("Proportion of sampling used in %")+
  ylab("Earth mover's distance")+
  ggtitle("Scenario 1")
plt1

```



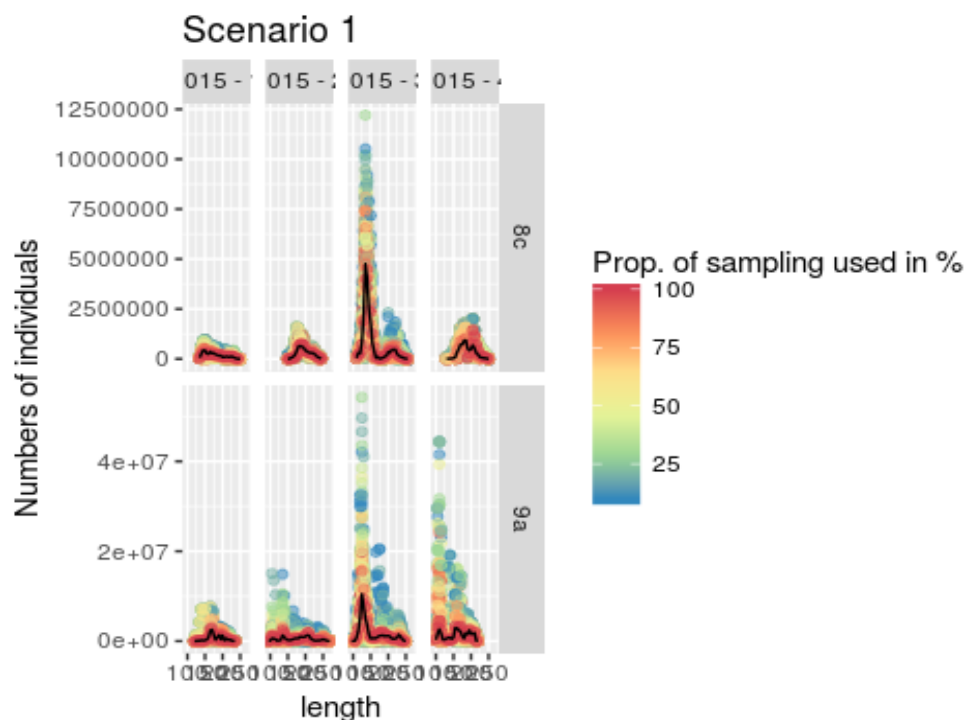
```
#extra info
sizefish2avg<-sizefish2%>%group_by(time,space,length,prop)%>%
  summarise(value=mean(value,na.rm=T))
#compare
plt2<-ggplot(sizefish2,aes(x=length,y=value))+
  #stat_bin2d(bins=50)+

geom_point(data=sizefish2,aes(x=length,y=value,group=prop,color=100*prop),alpha=.5)+
  scale_colour_distiller(palette='Spectral',name="Prop. of sampling
used in %") +

#geom_line(data=sizefish2avg,aes(x=length,y=value,group=prop,color=prop,size
=0.1),alpha=.5)+
  geom_line(data=sizefish,aes(x=length,y=value),color="black",alpha=1)+

  facet_grid(space~time,scale="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  ggtitle("Scenario 1")

plt2
```



The earth mover's distance (EMD) decrease logically with the increase of number of trips used to estimate the length distribution (the blue line is a LOESS smoother) for some strata (quarter 2 in 8c for example). For other strata the EMD curve is flat: decreasing the number of trips sampled did not change the shape of the length distribution. For the quarter 1 in 9c, the sampling effort could be reduced to 50% without loss of information. This is confirmed by the visual inspection of the length distribution. For the other strata, the length distribution computed with 50% to 100% of the trips (red to orange dots) convey the same amount of information than the original length distribution (black line) for the subarea 27.8c.

## Scenario 2

In this scenario, the impact of the number of fish sampled by haul inside a trip is investigated. For each haul, 10 to 100 fish are taken randomly with replacement from the original measurements, and length distributions are computed using the new samples. 30 simulations were performed using this scenario.

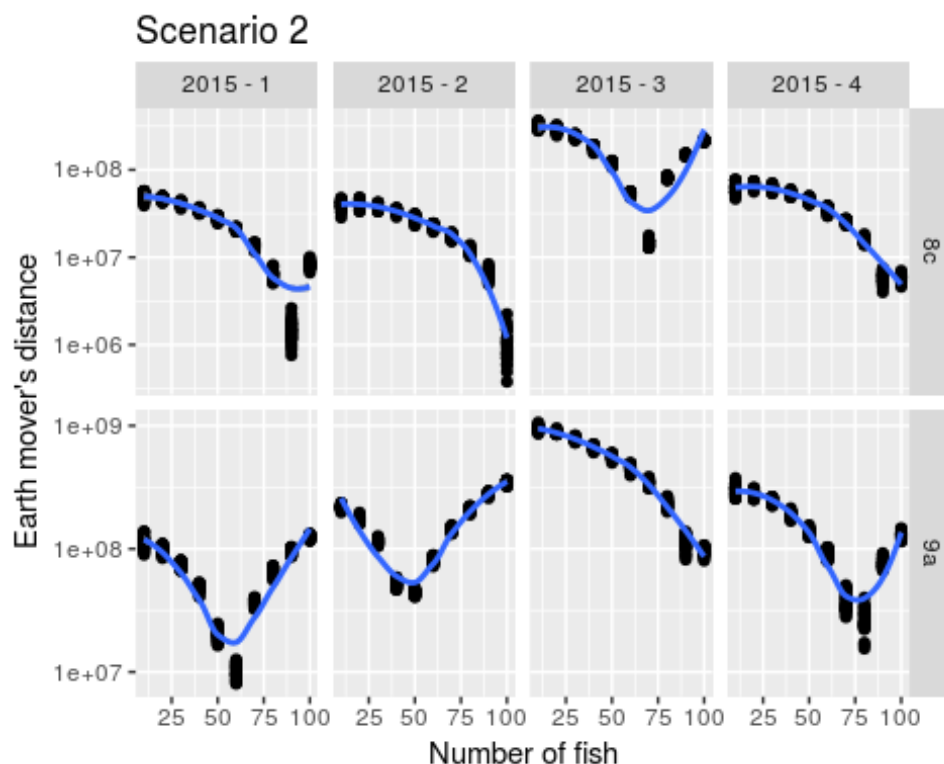
```
if(file.exists("sizefish6.rdata")){
  load("sizefish6.rdata")
}else{
  sizefish2<-data.frame()
  #expand the hl table
  hlexp <- CSc@hl[rep(row.names(CSc@hl), CSc@hl$lenNum), ]
  hlexp<-hlexp%>%mutate(lenNum=1,idhaul=paste(trpCode,staNum))
  aa<-hlexp%>%group_by(idhaul,time,space,technical)%>%summarise(n=sum(lenNum))%>%group_by(time,space,technical)%>%
    summarise(m=mean(n))
  for(i in 1:30){
    for(j in seq(10,100,10)){
      print(paste(i,j))
      aa<-stratified(hlexp,"idhaul",j,replace=TRUE)
      CSctmp<-CSc
      CSctmp@hl<-aa%>%group_by(PSUid,SSUId,TSUId,time,space,technical,sort,sampType,landCtry,
                              vs1FlgCtry,proj,trpCode,staNum,spp,sex,lenCls)%>%
```

```

      summarise(lenNum=sum(lenNum))%>%ungroup()
      rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="126421")
      dbelantmp<-rezdolen$dbelan
      sizefishtmp<-dbelantmp@lenStruc$estim%>%group_by(time,space=sub-
str(space,1,4),length)%>%
      summarise(value=sum(as.numeric(value),na.rm=T))%>%un-
group()%>%
      mutate(length=as.numeric(length),type="scenario
6",nbfish=j,rep=i)
      sizefish2<-rbind(sizefish2,sizefishtmp)
    }
  }
  save(sizefish2,file="sizefish6.rdata")
}

#compute emd
emd1<-left_join(sizefish2%>%trans-
mute(space,time,length,nbfish,rep,val1=value),
  sizefish%>%transmute(space,time,length,val0=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time,nbfish,rep)%>%summa-
rise(emd=emdL1(val1,val0))
plt1<-ggplot(emd1,aes(x=nbfish,y=emd))+geom_point()+
  scale_y_log10()+facet_grid(space~time,scale="free_y")+geom_smooth()+
  xlab("Number of fish")+
  ylab("Earth mover's distance")+
  ggtitle("Scenario 2")
plt1

```



```

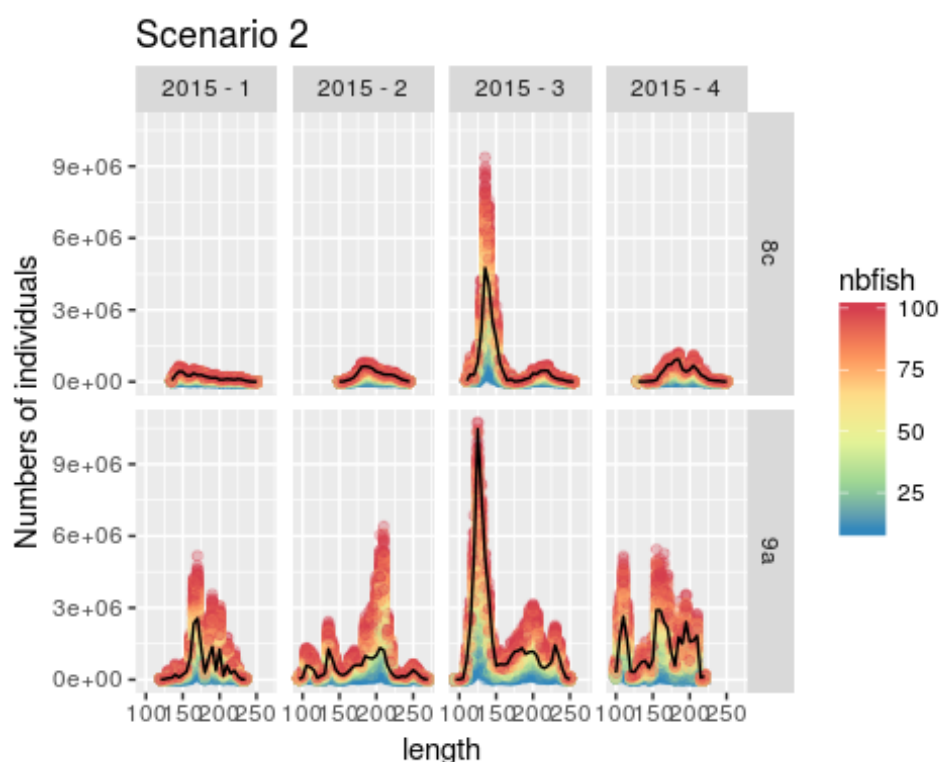
sizefish2avg<-sizefish2%>%group_by(time,space,length,nbfish)%>%
  summarise(value=mean(value,na.rm=T))
#compare
plt2<-ggplot(sizefish2,aes(x=length,y=value))+
  #stat_bin2d(bins=50)+
  geom_point(data=sizefish2,aes(x=length,y=value,group=nbfish,color=nbfish),al-
pha=.3)+

```

```

scale_colour_distiller(palette='Spectral',name="nbfish") +
#geom_line(data=sizefish2avg,aes(x=length,y=value,group=nbhaul,color=nbhaul,
size=1.1),alpha=.5)+
geom_line(data=sizefish,aes(x=length,y=value),color="black",alpha=1)+
facet_grid(space~time)+
ylab("Numbers of individuals")+xlab("length")+
ggtitle("Scenario 2")
plt2

```



The earth mover's distance (EMD) reaches local minima, in some strata, indicating an optimal number of fish to be measured. It is not clear what variables are determining these minima. While sampling protocols are the same in all the area, samplers and port procedures (access to the fish, sale, etc.) differs. The simulation using replacement explain why in the area 9c, around the local minima, an increase or a decrease in number of fish modify the length distribution. Some latitude at + or - 10 fish is possible. Graphically the estimated length distributions are similar to the original one when the number of fish is high.

### Discussion

Optimization analyses prove to be a valuable tool to explore sampling design and implementation. Detection of sample outliers with the delta index is already included in the Spanish procedure; implementing a common fixed threshold would be a step forwards a general quality improvement. No modifications are needed in relation to the sampling coverage of this stock, which shows a correct connection between sampling effort and landings in all the three stratifications. Nevertheless, the scenario analyses for the sampling effort show that a decrease in number of trips could be performed without a substantial impact in the information provided. A better understanding of this EMD analyses can be used for adjacent years thus ensuring a clear indication for change. As a second factor for consideration, this analyses need to be done considering

the other species covered under the sampling of this fleet. With the current concurrent sampling scheme the reduction of trips have to ensure similar information for all the sampled stocks.

## References

- Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J. Horton. 2014. "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations in Statistics Education* 8 (1).
- Gentleman, Robert, and Duncan Temple Lang. 2004. "Statistical Analyses and Reproducible Research." Bioconductor Project Working Papers. Working Paper 2. <http://www.bepress.com/bioconductor/paper2>.
- Jansen, Teunis, Henrik Degel, Joel Vigneau, and Ernesto Jardim. 2009. "Definition of Standard Data-Exchange Format for Sampling, Landings, and Effort Data Commercial Fisheries." ICES COOPERATIVE RESEARCH REPORT 296. ICES.
- — —. 2016. "The Regional DataBase (RDB) Exchange Format." ICES COOPERATIVE RESEARCH REPORT 296. ICES.
- R Core Team, R. 2012. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Rubner, Y., C. Tomasi, and L. J. Guibas. 1998. "A Metric for Distributions with Applications to Image Databases." *IEEE International Conference on Computer Vision*, 59–66.
- Vigneau, Joel, and Stephanie Mahevas. 2007. "Detecting Sampling Outliers and Sampling Heterogeneity When Catch-at-Length Is Estimated Using the Ratio Estimator." *ICES Journal of Marine Science: Journal Du Conseil* 64 (5): 1028–32. doi:[10.1093/icesjms/fsm077](https://doi.org/10.1093/icesjms/fsm077).



## **B – Hake in area 27.9.a**

Southern hake stock comprises the Atlantic coast of Iberian Peninsula corresponding to ICES division 8.c and 9.a. Hake in these divisions is caught in the trawl and artisanal mixed fisheries together with other fish species and crustaceans. In 2014, total landings were 2443 t. using the commercial at-market sampling data for the year 2014, this document presents a framework where length distribution estimates for landings are built using different subsamples of the original database. The way how these subsamples are built are called 'scenario'.

This report, using numerical simulation, provides some indication on how sampling effort could be reallocated, keeping the same quality of information for ONE species and ONE countries. The simulations were performed without taking into account the information related to other stocks (concurrent sampling) and to other parameters (discards), and without any consideration to practical issues related to the organization of the sampling plan.

### ***Framework***

Hake in 27.9.a.

### ***Fleet***

The Portuguese fleets targeting hake operate mainly in division 9.a. Trawl fleet comprises two distinct components – the trawl targeting demersal fish (OTB\_DEF) that operates along the entire Portuguese coast and the trawl fleet targeting crustaceans (OTB\_CRU) that operates mainly in the southwest and south coasts. Artisanal fisheries include gillnetters, longliners and a MIS\_MIS métier group of vessels is also present. This is related to the fact that a major part of Portuguese polyvalent fleet can operate with more than one gear in the same trip making the métier level 6 definition difficult to allocate for each trip. During the last year IPMA and DGRM (Portuguese administration) are joining efforts to come up with common criteria for the classification of trips from multigear vessels.

### ***Sampling design***

The sampling of commercial landings is carried out by the Portuguese Institute of Sea and Atmosphere (IPMA) and the length composition sampling design follows a multistage stratified random scheme by quarter, harbour and fleet (or métier) where, following the DCF requirements, less significant fleets are not sampled (e.g. beach-seines and dredges). The primary sampling unit (PSU) is 'harbour\*day' and the sampling effort is based on number of trips that constitute the secondary sampling unit (SSU). The sampling effort is allocated to harbour and quarter according to landings proportions in the most recent year. Trips are sampled using concurrent sampling meaning that all the species and size categories landed from each trip are sampled. For this exercise size category was not considered in the analysis.

### ***Statistical analyses***

The framework developed for this study integrates the whole processing of fishery data to (1) build population estimates at the stock level (length distribution mainly) and (2) to analyse the impact of different sampling strategies in these estimates by simulation. This framework is aimed for reproducibility and transparency, following the recommendation of the reproducible research statement (Gentleman and Lang 2004). Consequently this report is self-consistent: the code used to process and to analyse the fishery data are embedded in the report itself.

## Software

Analyses are carried out using the R environment (R Core Team 2012). R7 is a free software environment for statistical computing and graphics. The reproducibility of the results presented in this report relies on the use of a dialect of the Markdown language called Pandoc for word processing using the Knitr R package. Markdown is a plain text formatting syntax designed so that it can optionally be converted to HTML using a tool by the same name. Pandoc8 is a Markdown dialect which extends the conversion capability to word processing file (docx, doc and odt) and pdf, among other formats. Pandoc understands a number of markdown syntax extensions, including document metadata (title, author, date), footnotes, tables, figures and references. Knitr9 is an R package (a set of functions extending the R capabilities). With this package, the R code used to process and analyse the data are included directly in the report. Results are then produced dynamically. This framework has demonstrated the capacity to improve the conduct and the presentation of data analysis in a way that another person can understand and replicate (Baumer *et al.* 2014).

For example, if the calculus of 1+1 is needed, the code to compute it is written in the report using special hooks, as in this simple example:

```
'''{r test00,warn=FALSE,cache=TRUE,echo=TRUE}
#comment: addition example.
1+1
'''
```

This code is evaluated during the compilation of the report by the knitr command and it prints the following result:

```
#comment: addition example.
1+1
## [1] 2
```

The result is 2. All the numerical values related to the data, including tables and figures are produced following this procedure. Some hooks are appearing in the final report, other were removed to clarify the outputs. Moreover for the case study analysed here, the original Rmarkdown document will be available on the WKBIOPTIM SharePoint.

## Data

Data inputs follow the Regional DataBase (RDB) Exchange Format version 2009 (Jansen *et al.* 2009). It's an older version of the actual format used in the Fishframe database (Jansen *et al.* 2016). Three data objects are used:

- the object CL (Commercial Landings) brings population data related to landing,
- the object CE (Commercial Effort) brings population data related to effort,
- the object CS (Commercial Samplings) brings sampling data related to effort.

---

<sup>7</sup><http://www.r-project.org/>.

<sup>8</sup><http://johnmacfarlane.net/pandoc/>.

<sup>9</sup><http://yihui.name/knitr/>.

### Installation

The document needs the COST package (section software of the WKBIOPTIM Share-Point). To install them:

```
#the path to the library is local (aka where are the files on your hard disk)
install.packages("/path/to/COSTcore_1.4-0.zip",repos=NULL)
install.packages("/path/to/COSTdbe_1.4-1.zip",repos=NULL)
install.packages("/path/to/COSTeda_1.4.0.zip",repos=NULL)
```

Some other packages are needed and can be installed using CRAN:

```
install.packages("dplyr")
install.packages("ggplot2")
install.packages("DATRAS")
install.packages("mapdata")
install.packages("pander")
install.packages('DATRAS',repos='http://www.rforge.net/',type='source')
devtools::install_github("ldbk/earthmovdist")
```

The library needs to be loaded in R. Moreover a set of functions are available in the `wkbioptim_fct.R` files and have to be sourced. Some options related to how the chunks of code are interpreted are fixed (size of the figures...):

```
#R general option:
options(stringsAsFactors = FALSE)
#working directory:
#setwd("C:/path/to/the/working/directory")
#chunk option
#knitr option
knitr::opts_chunk$set(cache=TRUE,echo=TRUE, warning=FALSE,
  message=FALSE, fig.height=6,progress=FALSE,verbose=FALSE,
  include=TRUE,dev='png',autodep=FALSE)
#package
library(COSTcore)
library(COSTdbe)
library(COSTeda)
library(dplyr)
library(ggplot2)
library(maps)
library(mapdata)
library(pander)
library(DATRAS)
library(earthmovdist)
#source local file
source("wkbioptim_fct.R")
```

National data are loaded and combined.

```
#read CS
tr<-read.csv("./data/CSTR_hke_9a.csv",sep=";")
hh<-read.csv("./data/CSHH_hke_9a.csv",sep=";")
hl<-read.csv("./data/CSHL_hke_9a.csv",sep=";")
sl<-read.csv("./data/CSSL_hke_9a.csv",sep=";")
ca<-read.csv("./data/CSCA_hke_9a.csv",sep=";")
#integrity pb
sl0<-semi_join(sl,hh)
CSr<-csData(desc="",tr=tr,hh=hh,sl=sl0,hl=hl,ca=ca)
#read CL
cl15<-read.csv("./data/CL15.csv",sep=";",header=F,dec=",")[, -1]
cl14<-read.csv("./data/CL14.csv",sep=";",header=F,dec=",")[, -1]
CLr<-cldata(rbind(cl14,cl15))
```

Hake (*Merluccius merluccius*) is selected for the areas 9a in 2014:

```
#subset spp
CSr<-subsetSpp(CSr,spp=="Merluccius merluccius",table="sl") # hake
CLr<-subset(CLr, taxon=="Merluccius merluccius",table="cl")
#subset in space
CSr<-subset(CSr,area%in%c("27.9.a"),table="hh")
CLr<-subset(CLr,area%in%c("27.9.a"),table="cl")
#subset in time
CSr<-subset(CSr,year%in%c("2014"),table="hh")
CLr<-subset(CLr,year%in%c("2014"),table="cl")
save(CSr,CLr,file="pipo.rdata")
```

### Exploratory data analysis

In this section, a short exploratory analysis of the data are provided.

### Global figures

Total landings and sampling effort for hake, landed in Portugal, are computed in Table 1.

**Table 1 - Hake fishery for Portugal in 2014**

COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	FISH MEASURED
PRT	9.a	2443	522	26166

### Métier

Landings and sampling effort by métier are summarized in Table 2. The number of métiers is reduced: only métier catching hake and sampled are kept (with at least 7 trips sampled). The other métiers are labelled as "Other".

**Table 2 - Landings and sampling effort by main métier, country and subarea in 2014.**

MÉTIER	COUNTR	SUBARE	LANDINGS(T)	TRIPS SAMPLED	FISH MEASURED
	Y	A			
MISMIS-0_0_0	PRT	9.a	899.1	202	9308
OTBDEF- >=55_0_0	PRT	9.a	657.9	134	9795
GTRDEF- >=100_0_0	PRT	9.a	326.4	78	1892
Other	PRT	9.a	197.8	8	112
LLSDEF-0_0_0	PRT	9.a	72.7	8	616
GNSDEF-60- 79_0_0	PRT	9.a	10.7	31	1311
OTBCRU- >=55_0_0	PRT	9.a	6.1	62	3132

*Landings*

Landings are mapped by ICES rectangle, quarter and country.

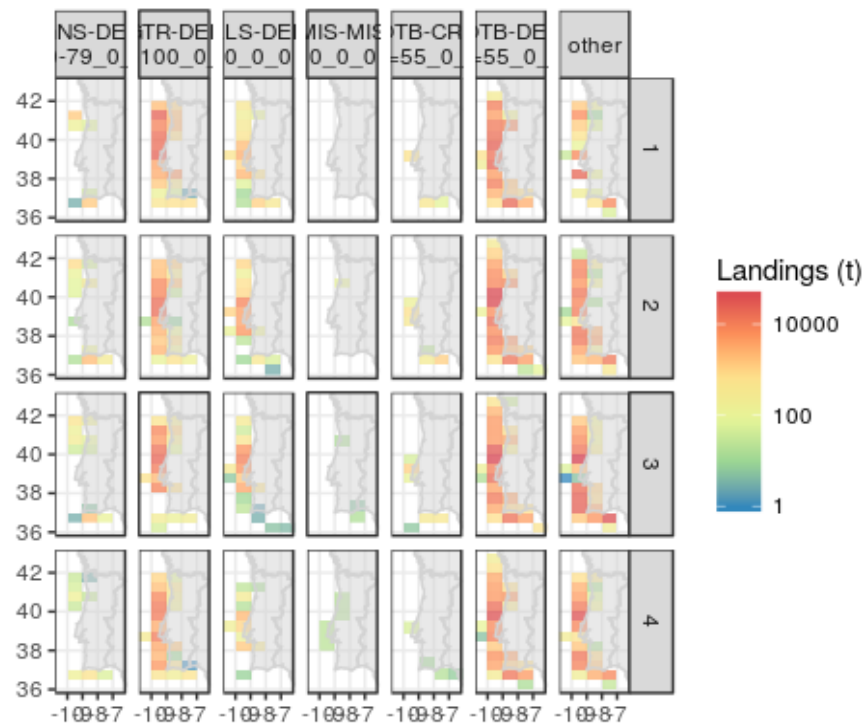


Figure 1 – Spatial distribution of hake landings in 2014, by métier and quarter.

*Stratification*

The stratification defines the spatial, temporal and technical support on which population estimates are computed. The stratification follows the ICES datacall policy: quarter, ICES area and métier level 6.

Stratification

TIMESTRATA	SPACESTRATA	TECHSTRATA
quarter	area	foCatEu6

*Validation and consolidation*

According to the stratification, the data are validated: samples are tested using quality checks.

### Quality checks

#### Sample weights

Subsample weights are checked using the hake weight length (WL) relationship. If the sample weight differs from the theoretical weight by 50% then this sample is removed. In Portuguese at-market sampling, only individual lengths are collected and the weight that is uploaded to RDB is already converted using WL relationship.

```
CSr<-corrbase(CSr)
rtp<-data.frame(sciname="hke",0.00000513,b=3.0744)
rez<-corrsampw(CSr,rtp)
CSr<-rez[[1]];
pltsampw<-rez[[2]]
try(plot(pltsampw),silent=T)
```

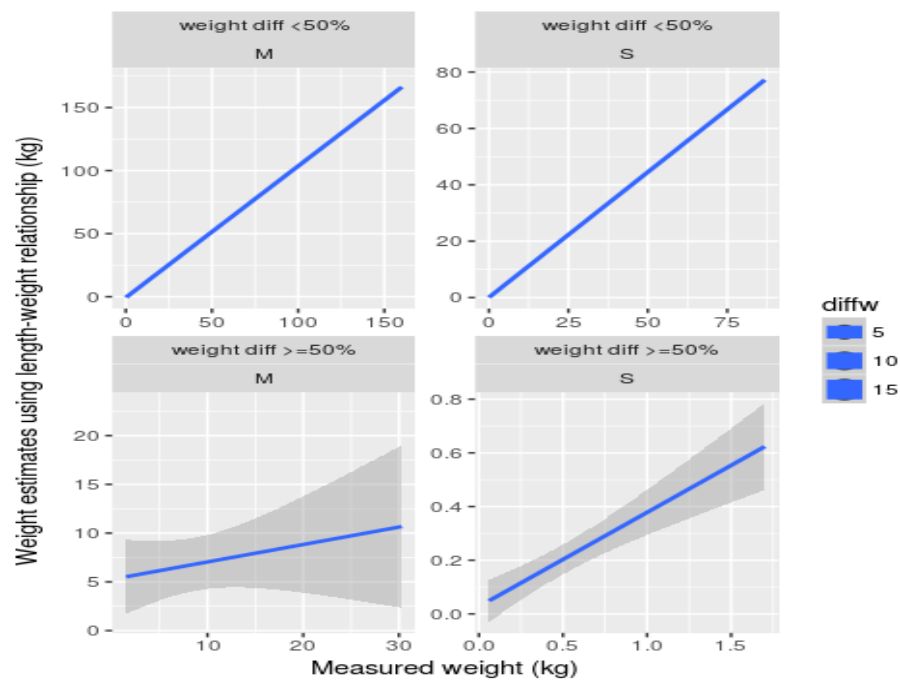


Figure 2 – Samples weight observed and estimated. (M=market sampling; S=sea sampling)

#### Sample outliers

Sample outliers are tested using the delta approach following Vigneau and Mahevas (2007). This index helps to detect sample outliers and sample heterogeneity when catch-at-length (includes both landings and discards data) is estimated using the ratio estimator. In this report, in order to ensure the repeatability of our results, a fixed threshold is applied to detect outlier. In a strata, all the samples with delta values outside the quantile at 2.5%% and 97.5% are flagged as outliers and removed from the analysis.

```
rezdelta<-delta(CSr,myStr,idspecies=unique(CLn@c1$taxon))
try(plot(rezdelta[[2]],silent=T))
```

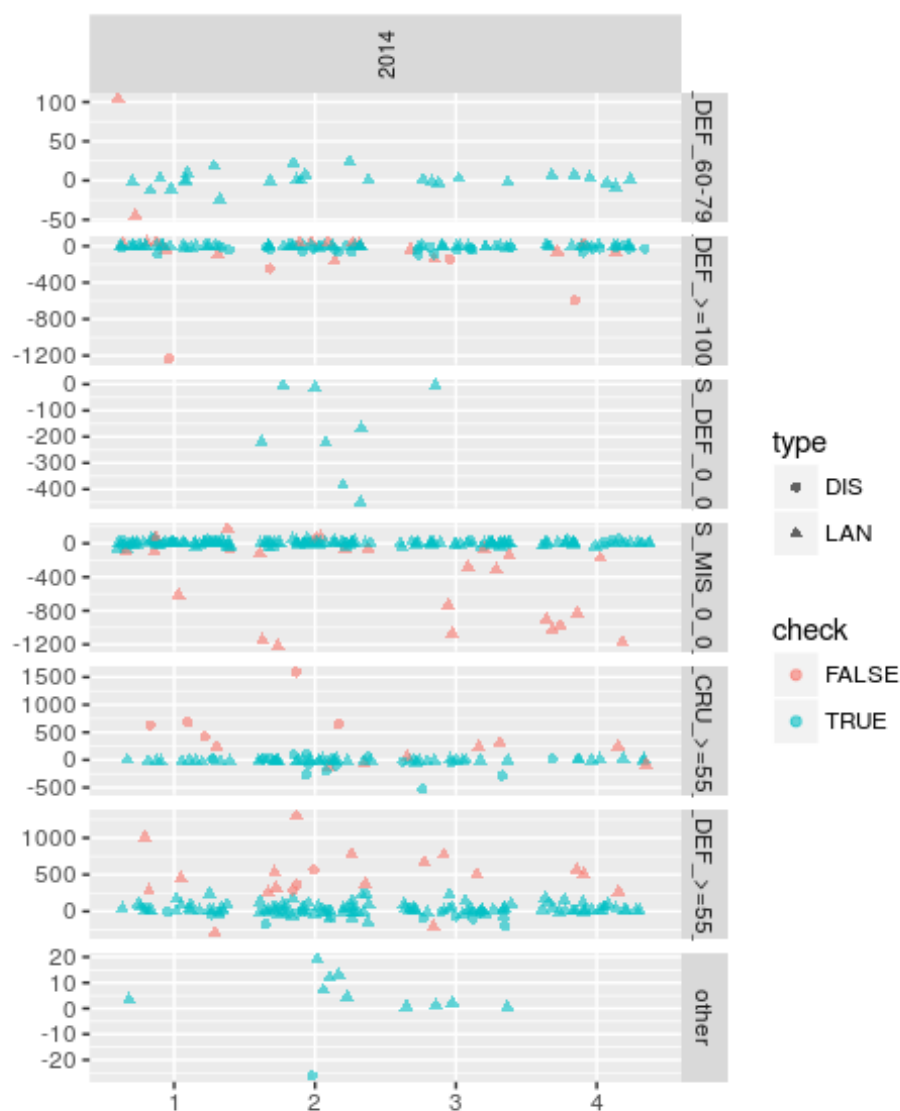


Figure 3 - Analysis of the sample outliers using the Delta index approach.

```
CSr<-rezdelta[[1]]
```

#### *Validation and consolidation*

Validation and consolidation (i.e. application of the stratification to the data) are done using the ad-hoc COSTcore functions:

```
#CSr<-rezdelta[[1]]
CSv <- csDataVal(CSr) ; CSc <- csDataCons(CSv,myStr)
CLv <- clDataVal(CLv); CLc <- clDataCons(CLv,myStr)
#CEv <- ceDataVal(CEr); CEc <- ceDataCons(CEv,myStr)
if(F){
  CSc<-alkLgthRec(CSc,type='stepIncr',10,preview=FALSE,post-
view=FALSE,update=TRUE)
}
save(CSv,CLv,CSc,CLc,file="datavalcons.Rdata")
```

### Sample coverage

The sample coverage is checked according to the stratification: the percentage in numbers of samples by strata is presented against the percentage of landings in each modality.

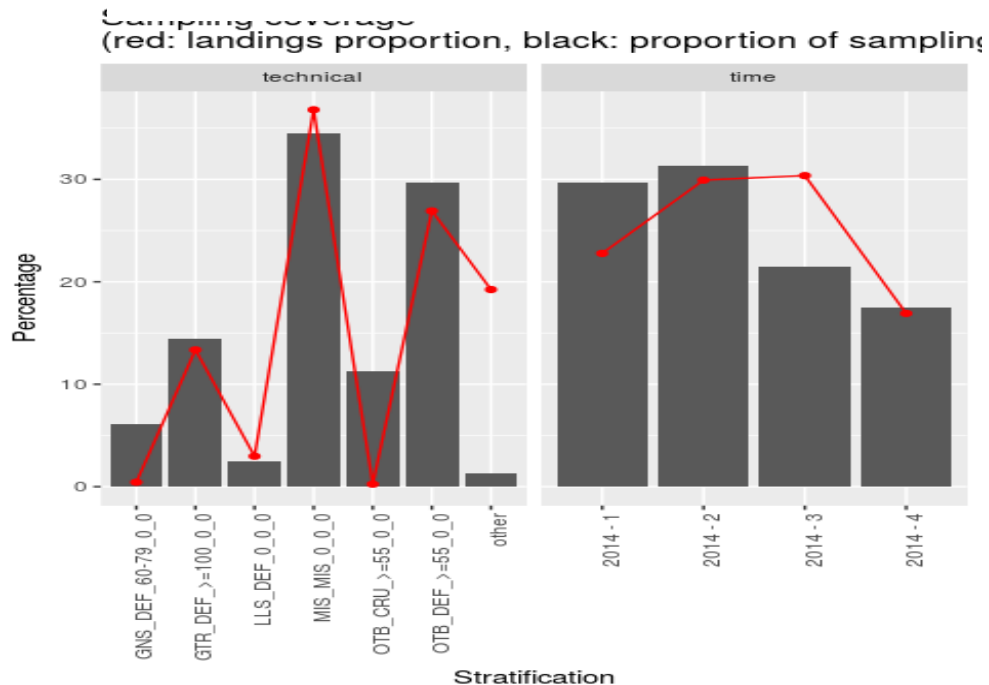


Figure 4 - Analysis of the sample coverage by métier and quarter in 2014. (Dark columns are the percentage of samples and the red lines the percentage of landings)

### Raised length composition of the stock

The length composition is estimated using ratio-estimator based on landings for each métier and quarter.

```
rezdolen<-dolen(CSc,CLc,myStr,idspecies=unique(CLr@c1$taxon))
dbelan<-rezdolen$dbelan
save(dbelan,file='dbelan.Rdata')
pltlenlan<-graphlen(dbelan,idstock="hke.9a",rtp,checkn=50,checks=3)
#plot(pltlenlan)
sizefish<-dbelan@lenStruc$estim%>%group_by(space=sub-
str(space,1,4),time,length, technical)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="true estimates")
pipo<-sizefish%>%mutate(technical=sub("_","\n",sub("_","-",technical)))
plt1<-ggplot(pipo,aes(x=length,y=value))+
  facet_grid(technical~time,scales="free_y")+#,space="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Hake length distribution by quarter
and metier")
plt1
```



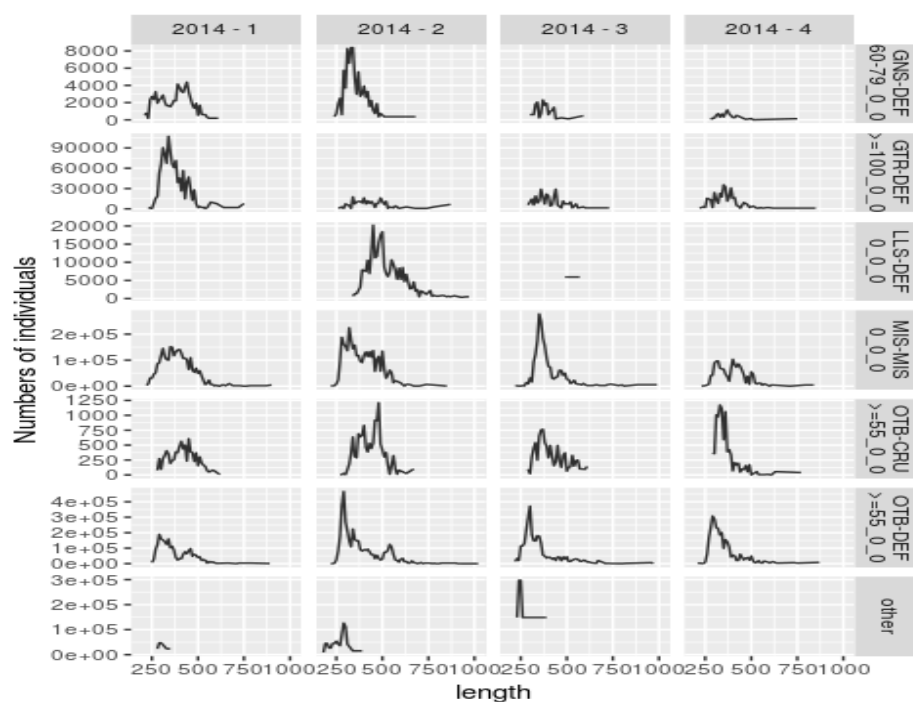


Figure 5 – Landings length distribution for hake by quarter and métier.

To summarize our graphical output, length distributions are combined by quarter.

```
sizefish<-dbelan@lenStruc$estim%>%group_by(space=sub-
str(space,1,4),time,length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="true estimates")
plt2<-ggplot(rbind(sizefish),aes(x=length,y=value))+
  facet_wrap(~time,scales="free_y")+#,space="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Hake length distribution by quar-
ter")
plt2
```

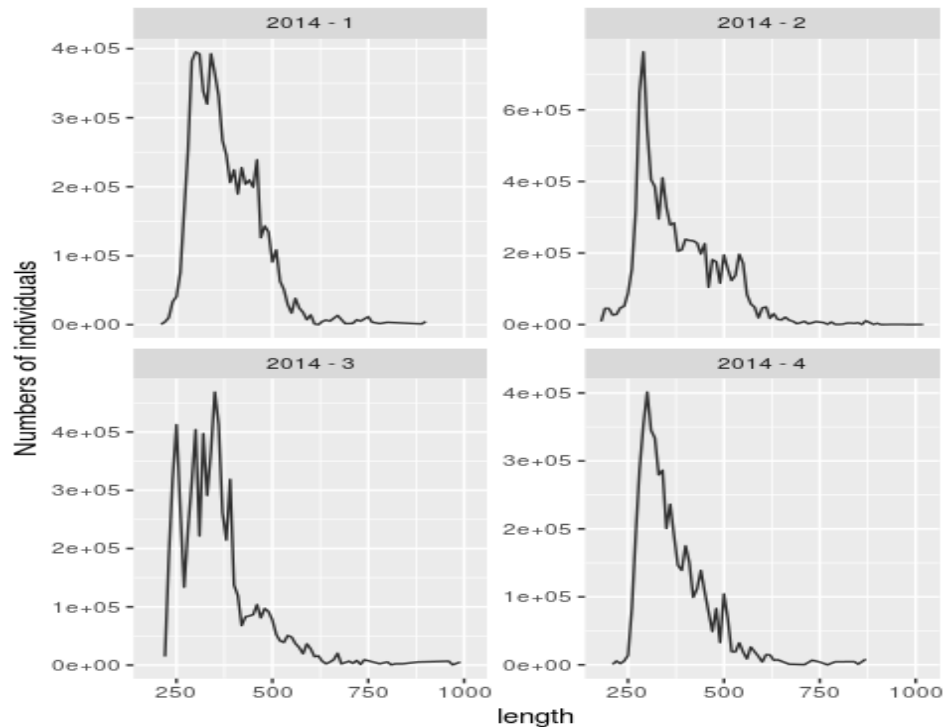


Figure 6 – Hake length distribution grouped for each quarter.

### Scenarios

From the first analyses, the hake length distributions in area 27.9.a are estimated using the dataset collected by Portugal in 2014 (Figure 6). These length distributions are considered the "true" estimates of the information landed by national sampling program. This section is now dedicated to the estimation of the same parameters, with changes in the sampling plan. These changes are for scenario testing and some modification of the sampling plan and reallocation of the sampling effort are performed. To do so, samples are removed according to the scenario, in a fixed or random way, and length distribution are estimated based on the new sampling base. The results are compared to the "true" estimates using the original number of samples. To compare the 2 distributions, the Earth Mover's Distance (EMD) is used. In statistics, the (EMD) is a measure of the distance between two probability distributions over a region  $D$  (names Wasserstein metric in mathematics) (Rubner, Tomasi, and Guibas 1998). Informally<sup>10</sup>, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region  $D$ , the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be the amount of dirt moved times the distance by which it is moved.

The 2 scenario tested in this report are:

- Scenario 1: decrease the number of sampled trips.
- Scenario 2: decrease the number of fish sampled by trip.

<sup>10</sup>from [https://en.wikipedia.org/wiki/Earth\\_mover%27s\\_distance](https://en.wikipedia.org/wiki/Earth_mover%27s_distance)

### Scenario 1

In this scenario, the sampling effort was decreased, from 100% to 10% of the number of original trips sampled. Trips are sampled randomly with replacement in 30 simulations.

```

    if(file.exists("sizefish4.rdata")){
      load("sizefish4.rdata")
    }else{
      sizefish2<-data.frame()
      idtrip<-unique(CSc@hh$trpCode)
      for(i in 1:30){
        for(j in seq(0.1,1,0.1)){
          print(paste(i,j))
          idtrip0<-sample(idtrip,length(idtrip)*j,replace=T)
          CSctmp<-subset(CSc,trpCode%in%idtrip0,table="hh")
          rezdolen<-dolen(CSctmp,CLc,myStr,idspecies=unique(CLr@cl$taxon))
          dbelantmp<-rezdolen$dbelan
          sizefishtmp<-dbelantmp@lenStruc$estim%>%group_by(time,space=sub-
str(space,1,4),length)%>%
            summarise(value=sum(as.numeric(value),na.rm=T))%>%un-
group()%>%
            mutate(length=as.numeric(length),type="scenario
4",prop=j,rep=i)
          sizefish2<-rbind(sizefish2,sizefishtmp)
        }
      }
      save(sizefish2,file="sizefish4.rdata")
    }

    #compute emd
    emd1<-left_join(sizefish2%>%trans-
mute(space,time,length,prop,rep,val1=value),
      sizefish%>%transmute(space,time,length,val0=value))
    emd1[is.na(emd1)]<-0
    emd1<-emd1%>%group_by(space,time,prop,rep)%>%summa-
rise(emd=emdL1(val1,val0))
    plt1<-ggplot(emd1,aes(x=100*prop,y=emd))+geom_point()+
      scale_y_log10()+facet_grid(space~time,scale="free_y")+geom_smooth()+
      xlab("Proportion of sampling used in %")+
      ylab("Earth mover's distance")+
      ggtitle("Scenario 1: distances (y-axis) between the original length
distribution\nand the length distribution computed with less sampling (x-
axis)\nby quarter (dots: distance for each replication, blue line: polyno-
mial fit), ")
    plt1

```

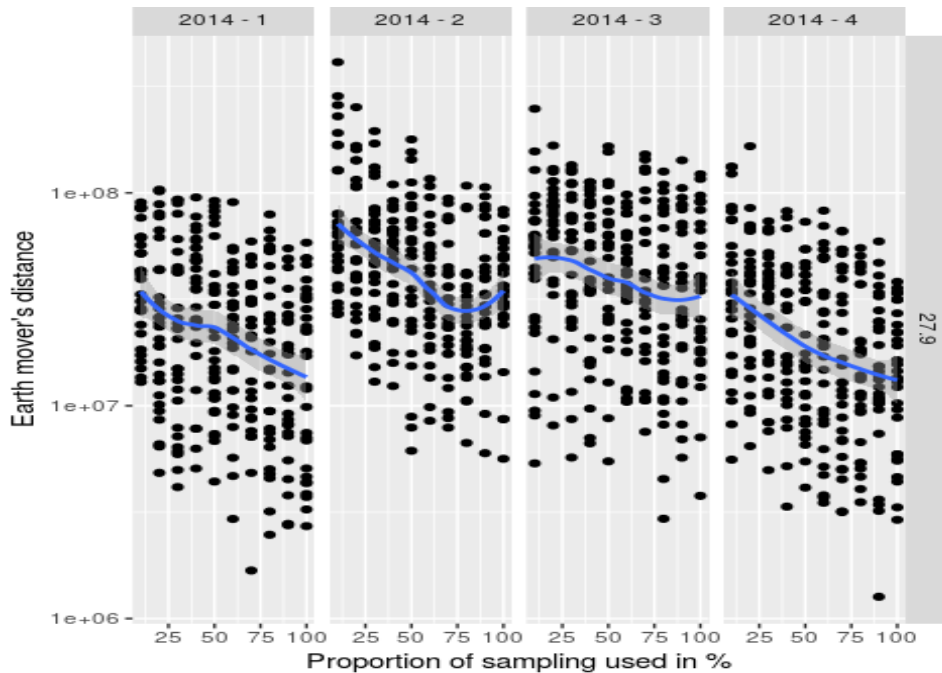


Figure 7 - Scenario 1: distances (y-axis) between the original length distribution and the length distribution computed with less samples (x-axis) by quarter (dots: distance for each replication, blue line: polynomial fit)

```
#extra info
gridsizefish2avg<-sizefish2%>%group_by(time,space,length,prop)%>%
  summarise(value=mean(value,na.rm=T))

#compare
plt2<-ggplot(sizefish2,aes(x=length,y=value))+
  #stat_bin2d(bins=50)+

geom_point(data=sizefish2,aes(x=length,y=value,group=prop,color=100*prop),alpha=.5)+
  scale_colour_distiller(palette='Spectral',name="Prop. of sampling
used in %") +

#geom_line(data=sizefish2avg,aes(x=length,y=value,group=prop,color=prop,size
=0.1),alpha=.5)+
  geom_line(data=sizefish,aes(x=length,y=value),color="black",alpha=1)+
  facet_wrap(space~time,scale="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  ggtitle("Scenario 1: original length distribution (black
line)\nand simulated one using less sampling (colored dots)")
plt2
```

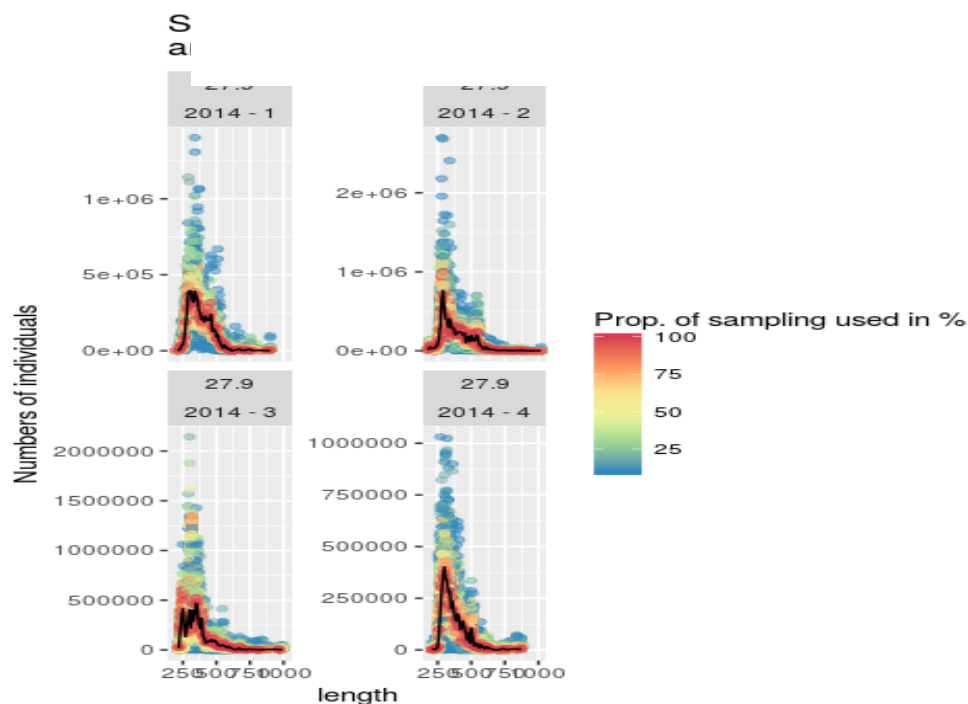


Figure 8 - Scenario 1: original length distribution (black line) and simulated one using less samples (colored dots)

The earth mover's distance (EMD) decrease logically with the increase of number of trips used to estimate the length distribution (the blue line is a LOESS smoother) for some strata (quarter 1 and 4). For the third quarter the EMD curve is rather flat: decreasing the number of trips sampled did not change the shape of the length distribution. For the quarter 2 a local minimum is reached at 80% meaning a reduction of the sampling effort is possible without loss of information. This is confirmed by the visual analysis of the length distribution. For the other strata, the length distribution computed with 75% to 100% of the trips (red to dark orange dots) convey the same amount of information than the original length distribution (black line).

### Scenario 2

In this scenario, the impact of the number of fish sampled by haul inside a trip is investigated. For each haul, 10 to 100 fish are taken randomly with replacement from the original measurements, and length distributions are computed using the new samples. 30 simulations were performed using this scenario.

```
if(file.exists("sizefish6.rdata")){
  load("sizefish6.rdata")
}else{
  sizefish2<-data.frame()
  #expand the hl table
  hlexp <- CSc@hl[rep(row.names(CSc@hl), CSc@hl$lenNum), ]
  hlexp<-hlexp%>%mutate(lenNum=1,idhaul=paste(trpCode,staNum))
  aa<-hlexp%>%group_by(idhaul,time,space,technical)%>%summarise(n=sum(lenNum))%>%group_by(time,space,technical)%>%
    summarise(m=mean(n))
  for(i in 1:30){
    for(j in seq(5,50,5)){
      print(paste(i,j))
      aa<-stratified(hlexp,"idhaul",j,replace=TRUE)
      CSctmp<-CSc
```

```

CSctmp@hl<-aa%>%group_by(PSUid,SSUid,TSUid,time,space,technical,sort,sampType,landCtry,
                          vs1FlgCtry,proj,trpCode,staNum,spp,sex,lenCls)%>%
  summarise(lenNum=sum(lenNum))%>%ungroup()
rezdolen<-dolen(CSctmp,CLc,myStr,idspecies=unique(CLR@cl$taxon))
dbelantmp<-rezdolen$dbelan
sizefishtmp<-dbelantmp@lenStruc$estim%>%group_by(time,space=substr(space,1,4),length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="scenario
6",nbfish=j,rep=i)
sizefish2<-rbind(sizefish2,sizefishtmp)
  }
}
save(sizefish2,file="sizefish6.rdata")
}

#compute emd
emd1<-left_join(sizefish2%>%transmute(space,time,length,nbfish,rep,val1=value),
               sizefish%>%transmute(space,time,length,val0=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time,nbfish,rep)%>%summarise(emd=emdL1(val1,val0))
plt1<-ggplot(emd1,aes(x=nbfish,y=emd))+geom_point()+
  scale_y_log10()+facet_grid(space~time,scale="free_y")+geom_smooth()+
  xlab("Number of fish")+
  ylab("Earth mover's distance")+
  ggtitle("Scenario 2: distances (y-axis) between the original length
distribution\nand the length distribution computed with less measurements
(x-axis)\nby quarter (dots: distance for each replication, blue line: polynomial fit), ")
plt1

```

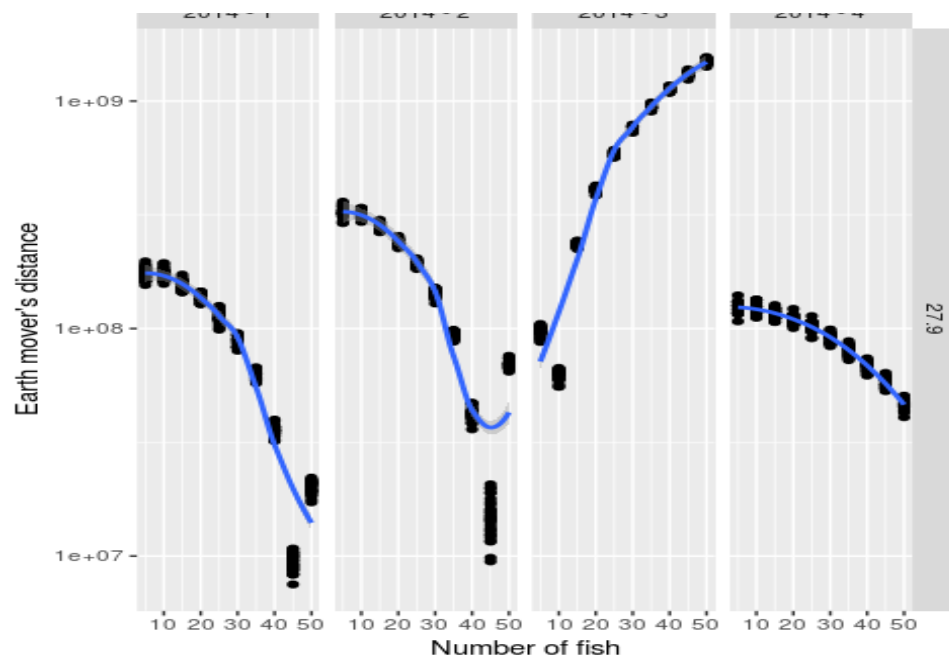


Figure 9 - Scenario 2: distances (y-axis) between the original length distribution and the length distribution computed with fewer measurements (x-axis) by quarter (dots: distance for each replication, blue line: polynomial fit).

```
sizefish2avg<-sizefish2%>%group_by(time,space,length,nbfish)%>%
  summarise(value=mean(value,na.rm=T))
#compare
plt2<-ggplot(sizefish2,aes(x=length,y=value))+
  #stat_bin2d(bins=50)+

geom_point(data=sizefish2,aes(x=length,y=value,group=nbfish,color=nbfish),alpha=.3)+
  scale_colour_distiller(palette='Spectral',name="nbfish") +

#geom_line(data=sizefish2avg,aes(x=length,y=value,group=nbfish,color=nbfish),
size=1.1,alpha=.5)+
  geom_line(data=sizefish,aes(x=length,y=value),color="black",alpha=1)+
  facet_wrap(space~time,scale="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  ggtitle("Scenario 2: original length distribution (black
line)\nand simulated one using less sampling (colored dots)")
plt2
```

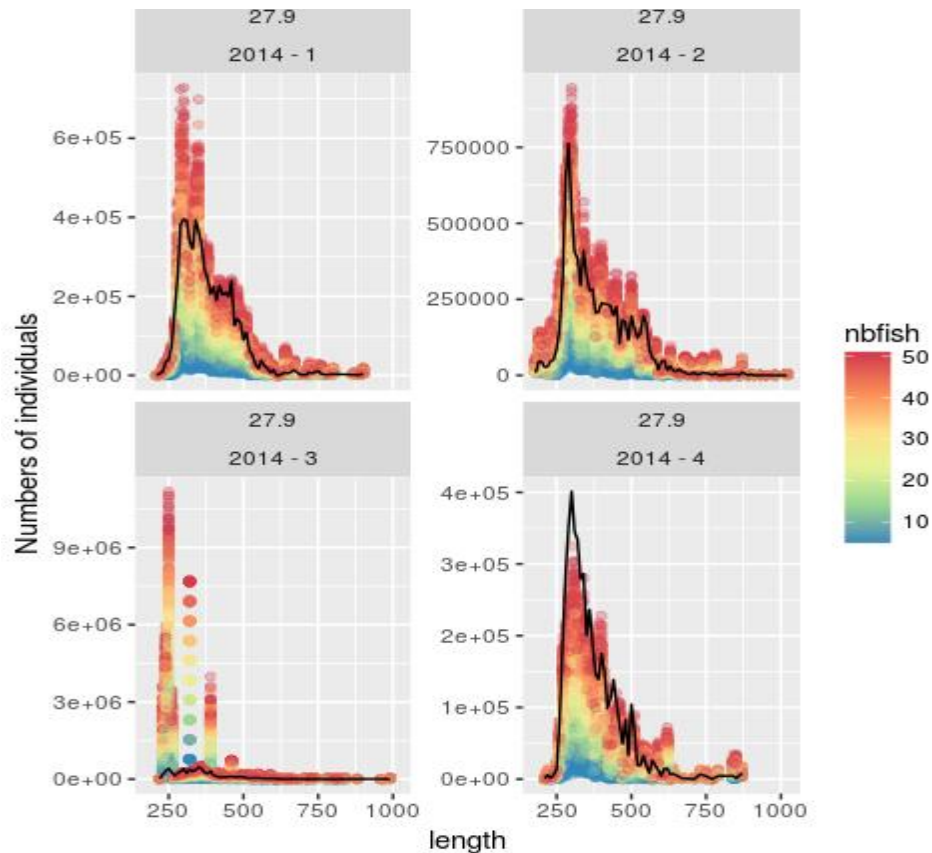


Figure 10 - Scenario 2: original length distribution (black line) and the simulated one using less samples (colored dots)

The earth mover's distance (EMD) increase sharply with the number of fish measured for the quarter 1 and 2. In this case, decreasing the amount of fish by sample changes the length distribution. For the quarter 2 the increase is smoother, and the number of fish can be diminished. For the quarter 3, the trend is the opposite, due to original number of fish available (less than 50). Some length classes are not well represented in the original samples, as seen in the graphical representation of the simulated distribution. Bootstrapping the fish measurements bring noise in the final length distribution.

### Discussion

This procedure seems to be a good tool to be used or adapted to when analyses of the sampling effort is the aim. It also includes data pre-screening, validation and consolidation procedures. Scenarios results show that considering the sampling effort in number of trips there is some space for reduction without much loss of information. In what concerns to the second scenario, it seems that the reduction of the number of fish measured must be taken with more caution. Results show that some differences in the length distribution may compromise the final estimates using this procedure. Taking this into account, maybe for this case the analysis should also consider species spatial distribution and size categories as levels of the analysis. Indications exist that these variables may influence the optimization procedure for this species (Silva *et al.*, 2017).



Also effective sample size procedures in terms of number of trips, must take in consideration the kind of sampling strategies in use (concurrent vs. species focus), not to compromise the information collected for other species/stocks targeted by that same fleets.

### References

- Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J. Horton. 2014. "R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations in Statistics Education* 8 (1).
- Gentleman, Robert, and Duncan Temple Lang. 2004. "Statistical Analyses and Reproducible Research." Bioconductor Project Working Papers. Working Paper 2. <http://www.bepress.com/bioconductor/paper2>.
- ICES. 2011. Report of the Working Group on the Assessment of Southern Shelf stocks of Hake, Monk and Megrim (WGHMM), 5–11 May 2011, ICES Headquarters, Copenhagen. ICES CM 2011/ACOM:11.625 pp
- Jansen, Teunis, Henrik Degel, Joel Vigneau, and Ernesto Jardim. 2009. "Definition of Standard Data-Exchange Format for Sampling, Landings, and Effort Data Commercial Fisheries." ICES COOPERATIVE RESEARCH REPORT 296. ICES.
- — —. 2016. "The Regional DataBase (RDB) Exchange Format." ICES COOPERATIVE RESEARCH REPORT 296. ICES.
- R Core Team, R. 2012. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Rubner, Y., C. Tomasi, and L. J. Guibas. 1998. "A Metric for Distributions with Applications to Image Databases." *IEEE International Conference on Computer Vision*, 59–66.
- Silva C, Azevedo M, Chaves C, Coelho R, Costa AMC, Dinis D, Dorés S, Fernandes AC, Gonçalves P, Lino PG, Mendes H, Moura T, Nunes C, Oroszlányová M, Pinto D, Silva MC, 2017. Report of the Workshop on Sampling Effort for Biological Parameters (WKSEBP), IPMA, Lisbon 18–20 April. Relat. Cient. Téc. do IPMA (<http://ipma.pt>), nº 17, 55 p + 4 Anexos.
- Vigneau, Joel, and Stephanie Mahevas. 2007. "Detecting Sampling Outliers and Sampling Heterogeneity When Catch-at-Length Is Estimated Using the Ratio Estimator." *ICES Journal of Marine Science: Journal Du Conseil* 64 (5): 1028–32. doi:[10.1093/icesjms/fsm077](https://doi.org/10.1093/icesjms/fsm077).

## C – Sole in areas 28.8a and 27.4

The stock sol.27.8ab is mainly exploited by France and Belgium. In 2016, total catches were 3346 t (91% from France and 9% from Belgium) according the ICES Advice 2016. Using the commercial sampling data for the year 2016 for the countries involved in the exploitation of this stock, this section presents a framework where length distribution estimates are build using different subsamples of the original database. The way how these subsamples are built are called 'scenario'.

This section, using numerical simulation, provides some indication on how sampling effort could be reallocated, keeping the same quality of information for ONE species and TWO countries. The simulation were performed without taking into account the information related to other stock (concurrent sampling) and to other parameters (discards), and without any consideration to practical issues related to the organization of the sampling plan.

### *Belgium*

#### Belgian fleet

The beam trawl fleet (TBB) targeting demersal species (DEF) is the most important fishery for Belgium, covering 74% of the total Belgian fishing hours. The TBB\_DEF fleet comprises of 2 fleet segments: a > 221 kW fleet segment and a ≤221 kW fleet segment:

The TBB\_DEF\_>221 kW fleet segment comprises beam trawl vessels with a capacity of more than 221 kW, operating in North Sea, the English Channel, the Irish Sea, the Celtic Sea, the Irish sea and in summertime the Bay of Biscay. The TBB\_DEF\_>221 kW trip duration is on average 8–10 days and one trip can cover several areas.

- The TBB\_DEF\_≤221 kW fleet segment comprises beam trawl vessels with a maximum power of 221 kW (coastal fleet segments and “eurocutters”), operating in the Southern North Sea and the Eastern English Channel (4 and 6.d respectively). In contrast to TBB\_DEF\_>221 kW, this fleet segments has also access to the 12 mile zone. A coastal vessel has a trip duration of less than 24 hours and an eurocutter of around 4 days.

#### Belgian sampling design

ILVO is only collecting fishery-dependent data for the beam trawl fleet targeting demersal species. Catch information (all catch fractions are covered) is obtained through on-board observation. Four ILVO observers assure a sampling coverage of on average 1 % (expressed in fishing hours). The two fleet segments (TBB\_DEF\_>221 kW and TBB\_DEF\_≤221 kW) are treated as two separate strata in the at sea sampling programme. The sampling effort targets for one year are set at 8 trips for the TBB\_DEF\_≤221 kW fleet segment and 32 trips for the TBB\_DEF\_>221 kW fleet segment.

The primary sampling unit (PSU) in the Belgian at sea sampling programme is vessel x trip (as a proxy for trip). A haul (within a trip) is defined as the secondary sampling unit (SSU). Selecting a vessel x trip (PSU) for the TBB\_DEF\_>221 kW fleet segment is done by a random draw from a vessel list (with replacement). Selecting a vessel x trip (PSU) for the TBB\_DEF\_≤221 kW fleet segment is done ad hoc. The vessel lists are so small (mainly because of logistic issues on-board certain vessels) that a random PSU selection is not feasible.

In the TBB\_DEF\_>221 kW fleet segment, every second haul (systematic sampling of SSU) is sampled by an observer so sampling takes place around the clock to reflect

typical working conditions. In the TBB\_DEF\_≤221 kW fleet segment, all hauls are sampled by an observer (only when large amounts of small fish are caught, the next haul might be skipped for sampling in order to be able to process the entire catch). The crew is sorting the marketable fish from the conveyor belt and they store the different species in different baskets for the observer to sample later on. In the meantime, the observer is taking care of the discarded fraction of the catch. The observer sorts all the discarded species of commercial importance and determines the total weight in a haul for almost all the species in the catch (a.o.). For a selected set of species (a.o. *Solea solea*), the observer also takes length measurements. Usually, the length of all individual fish in the discarded part of the tow is measured. Only when a species is extremely abundant, a smaller representative subsample (TSU) is measured. The retained part of the catch is treated in the same way as the discarded part of the catch.

#### Optimization of the Belgian at sea sampling design

In the upcoming years, ILVO will invest in optimizing the design of the at sea sampling programme. The goal is to design a new statistically sound sampling programme that will follow the guidelines for good practice provided by ICES and the European Commission (making the most efficient use of sampling resources in order to collect unbiased and precise catch data) and implement this new programme in practice. This will be a work in progress and depending on the outcomes of different types of statistical analyses (a.o. a random effects analysis using the at sea sampling data from the last decade), the design will be optimized step by step during 2017-2019.

### *France*

#### French fleet

The fleets targeting sole in the Bay of Biscay are separated in four main segments: inshore and offshore gillnetters and inshore and offshore trawlers. Vessel length characterize inshore (≤12 m) and offshore (>12 m) operations, while gears define the main métier.

#### French sampling design

Landings are collected using the national fishery declaration system (combining and cross-checking the logbook information and the sells notes), for all the vessels trips. At-sea and in auction samplings cover both large and small vessels related to this fishery. The sampling coverage is given each year in relation with the importance of each fleet segment.

The primary sampling unit (PSU) is vessel x trip (as a proxy for trip). For at-sea observation, a haul (within a trip) is defined as the secondary sampling unit (SSU). Selecting a vessel x trip (PSU) is done by a random draw from a vessel list, linked to a geographical area and a gear use. Haul selection follows some hierarchical rules linked to working conditions (at least 1/3 of the hauls of a métier have to be sampled).

## Data

National data are loaded and combined. Some corrections are made to homogenise the data, due to some discrepancies between variables type.

### French data

```
pathfrabob<-"./data/bob_fra/"
load(paste0(pathfrabob,"CSr.rdata"))
load(paste0(pathfrabob,"CLr.rdata"))
load(paste0(pathfrabob,"CEr.rdata"))
CSrfrabob<-CSr
CErfrabob<-CEr
CLrfrabob<-CLr
pathfrans<-"./data/ns_fra/"
load(paste0(pathfrans,"CSr.rdata"))
load(paste0(pathfrans,"CLr.rdata"))
load(paste0(pathfrans,"CEr.rdata"))
CSrfrans<-CSr
CErfrans<-CEr
CLrfrans<-CLr
CSrfra<-rbind2(CSrfrabob,CSrfrans)
CSrfra@tr$vs1Id<-as.numeric(as.factor(CSrfra@tr$vs1Id))
CLrfra<-rbind2(CLrfrabob,CLrfrans)
CErfra<-rbind2(CErfrabob,CErfrans)
```

### Belgian data

```
pathbelbob<-"./data/bob_bel/"
tr<-read.csv(paste0(pathbelbob,"TR_2016.csv"))
hh<-read.csv(paste0(pathbelbob,"HH_2016.csv"),colClasses=sap-
ply(csData())@hh,typeof),quote='')
sl<-read.csv(paste0(pathbelbob,"SL_2016.csv"))
hl<-read.csv(paste0(pathbelbob,"HL_2016.csv"))
CSrbelbob<-csData(tr,hh,sl,hl)
cl<-read.csv(paste0(pathbelbob,"CL_2016.csv"),colClasses=sap-
ply(clData())@cl,typeof),quote='')
CLrbelbob<-clData(cl)
ce<-read.csv(paste0(pathbelbob,"CE_2016.csv"))
CErbelbob<-ceData(ce)
pathbelns<-"./data/ns_bel/"
tr<-read.csv(paste0(pathbelns,"TR_2016.csv"))
hh<-read.csv(paste0(pathbelns,"HH_2016.csv"))
sl<-read.csv(paste0(pathbelns,"SL_2016.csv"))
hl<-read.csv(paste0(pathbelns,"HL_2016.csv"))
CSrbelns<-csData(tr,hh,sl,hl)
cl<-read.csv(paste0(pathbelns,"CL_2016.csv"))
CLrbelns<-clData(cl)
ce<-read.csv(paste0(pathbelns,"CE_2016.csv"))
CErbelns<-ceData(ce)
CSrbel<-rbind2(CSrbelbob,CSrbelns)
CSrbel@tr$vs1Id<-as.numeric(as.factor(CSrbel@tr$vs1Id))
CLrbel<-rbind2(CLrbelbob,CLrbelns)
CErbel<-rbind2(CErbelbob,CErbelns)

CSr<-rbind2(CSrfra,CSrbel)
CLr<-rbind2(CLrfra,CLrbel)
CEr<-rbind2(CErfra,CErbel)
save(CSr,CLr,CEr,file="pipo.rdata")
```

### Exploratory data analysis

In this section, a short exploratory analysis of the data are provided.

### Global figures

Landings and sampling effort are computed by country and ICES Subdivision.

#### Sole fishery for Belgium and France in 2016

COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	HAULS SAMPLED	FISH MEASURED
BEL	27.4	776.8	12	117	23409
FRA	27.4	365.6	23	41	1393
BEL	27.8	287.7	4	95	16723
FRA	27.8	3100	340	620	25553

### Métier

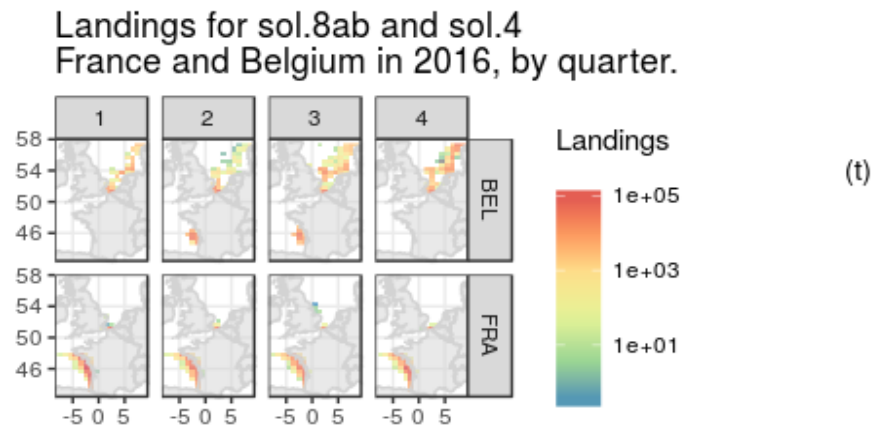
Landings and sampling effort by métier are summarized. The number of métier is reduced using the ICES rules on main métier: here métier fishing less than 90% of the cumulated catches are coded as MIS\_MIS\_0\_0\_0.

Landings and sampling effort by main métier, country and subarea in 2016. MIS\_MIS métier is generated using the ICES rules.

MÉTIER	COUNTRY	SUBAREA	LANDINGS(T)	TRIPS SAMPLED	HAULS SAMPLED	FISH MEASURED
GTRDEF-90_99_0	FRA	27.4	335.5	19	34	1385
MISMIS-0_0_0	FRA	27.4	30.14	5	7	8
GTRDEF->=100_0	FRA	27.8	2083	156	268	16508
MISMIS-0_0_0	FRA	27.8	371.8	102	193	3682
OTBDEF->=70_0	FRA	27.8	292.3	23	41	1694
OTTCRU->=70_0	FRA	27.8	228.9	48	81	2258
OTBCEP->=70_0	FRA	27.8	124.6	23	33	1231
TBBDEF-70-99_0_0	BEL	27.4	570.9	10	88	21717
TBBDEF->=120_0_0	BEL	27.4	120.6	2	29	1692
MISMIS-0_0_0	BEL	27.4	85.32	0	0	0
TBBDEF-70-99_0_0	BEL	27.8	287.7	4	95	16723

### Landings

Landings are mapped by ICES rectangle, quarter and country.



### *Stratification*

The stratification defines the spatial, temporal and technical support on which population estimates are computed. The stratification follows the ICES datacall policy: quarter, ICES area and métier level 6.

### *Métier*

For métier, métier level 6 area recoded using the vessel length and the gear:

Inshore-trawlers: trawlers  $\leq 12$  m (TBB, OTB, PTB)

Offshore-trawlers: trawlers  $> 12$  m (TBB, OTB, PTB)

Inshore-Gillnets: gillnets  $\leq 12$  m

Offshore-Gillnets: gillnets  $> 12$  m

This stratification is the one used by France for intercatch submission, and is in accordance with the Belgian fleet.

### *Space*

The spatial stratification takes into account some undetermined area for some French samples.

*Spatial stratification*

OLD	NEW
27.8.a	27.8.a
27.8.b	27.8.b
27.8.b,27.8.a	27.8.a
27.8.a,27.8.b	27.8.a
27.4.c	27.4.c
27.4.a	27.4.a
27.4.b	27.4.b

*Final stratification***Stratification**

TIMESTRATA	SPACESTRATA	TECHSTRATA
quarter	area	foCatEu6

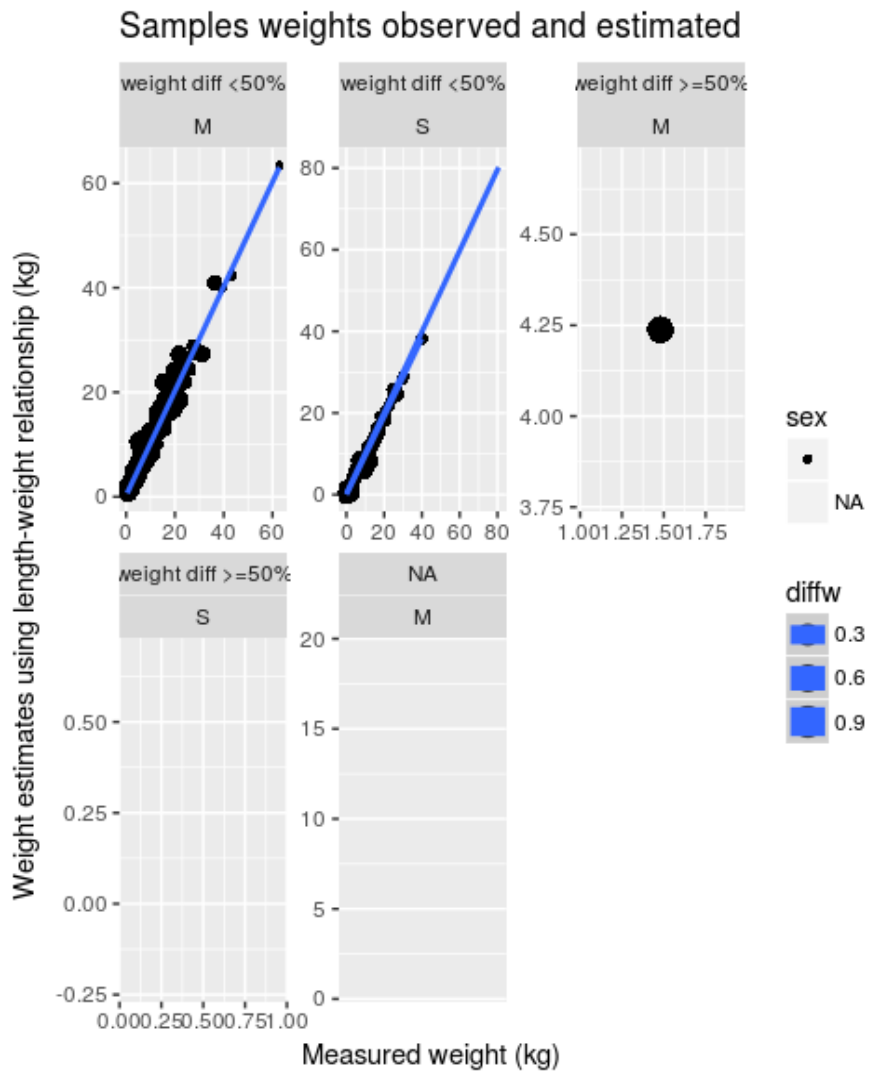
*Validation and consolidation*

According to the stratification, the data are validated: samples are tested using quality checks.

*Quality checks***Sample weights**

Sub samples weights are checked using the sole weight length relationship. If the sample weight differs from the theoretical weight by 50% then this sample is removed.

```
CSr<-corrbase(CSr)
rtp<-data.frame(sciname="Solea solea",a=3.638e-6,b=3.283)
rez<-corrsampw(CSr,rtp)
CSr<-rez[[1]];pltsampw<-rez[[2]]
try(plot(pltsampw),silent=T)
```

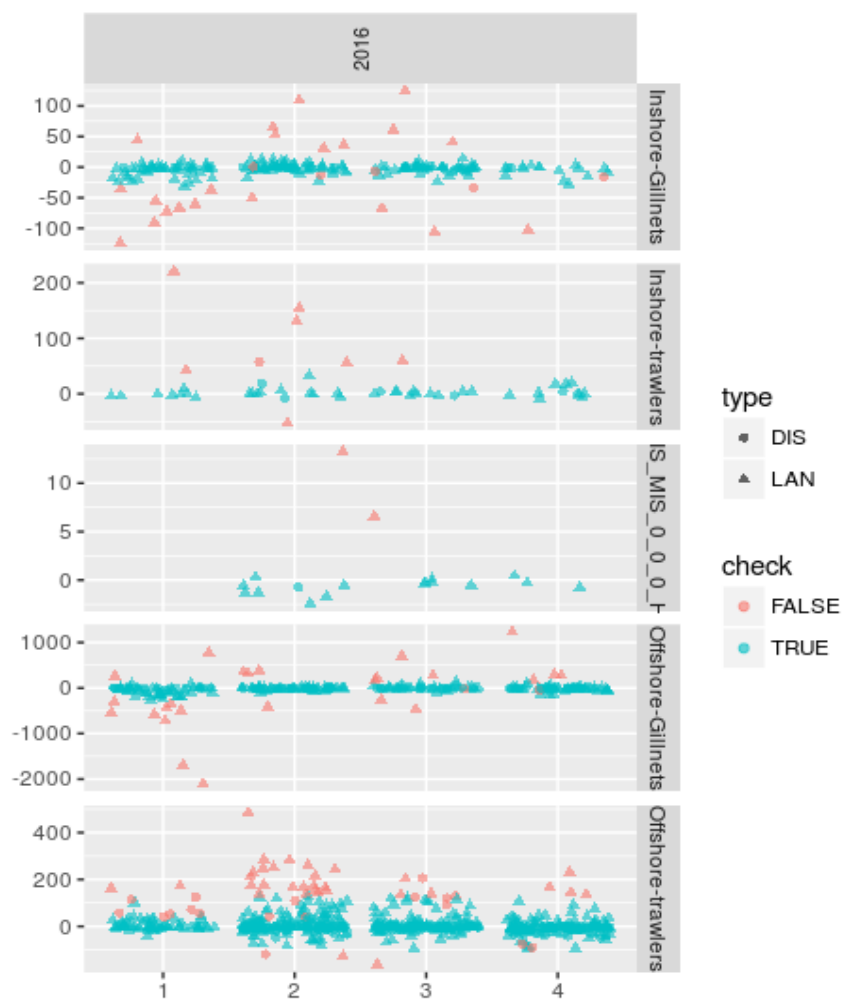


### Sample outliers

Sample outliers are tested using the delta approach following Vigneau and Mahevas (2007). This index helps to detect sample outliers and sample heterogeneity when catch-at-length is estimated using the ratio estimator. In this study, in order to ensure the repeatability of our results, a fixed threshold is applied to detect outliers. In a strata, all the samples with delta values outside the quantile at 2.5% and 97.5% are flagged as outliers. The delta value outlier detection for samples are computed but not used here:

```
rezdelta<-delta(CSr,myStr,idspecies="Solea solea")
try(plot(rezdelta[[2]],silent=T))
```





### Validation and consolidation

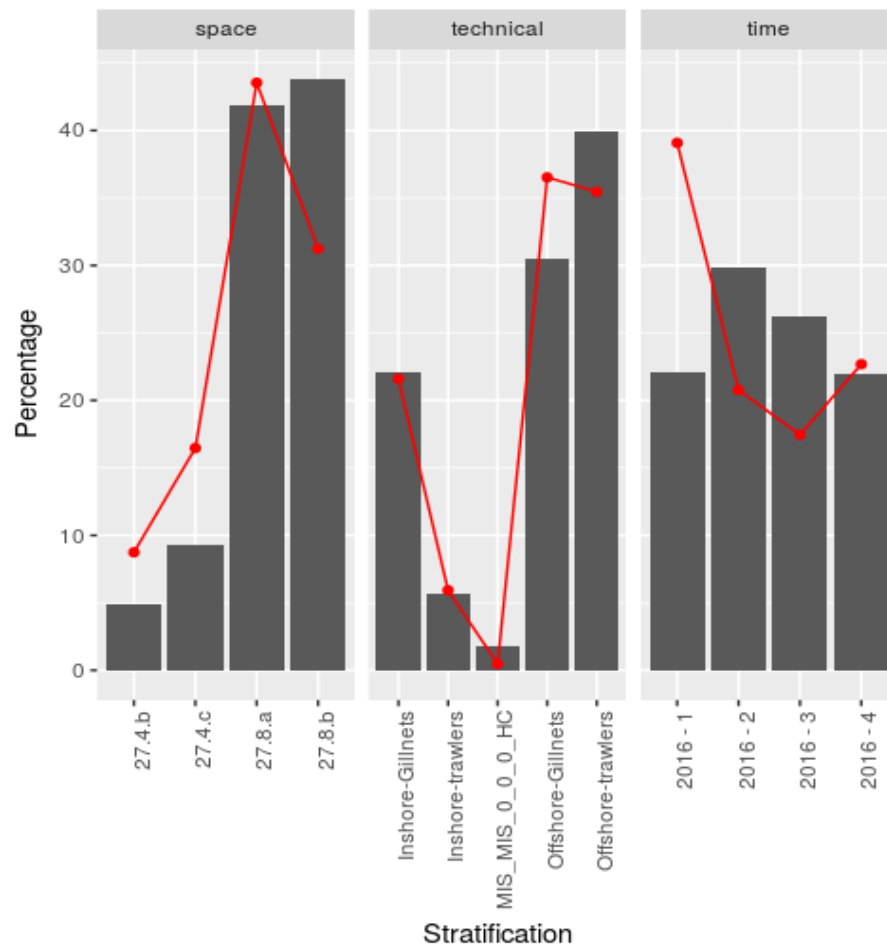
Validation and consolidation (i.e. application of the stratification to the data) are done using the ad-hoc COSTcore functions:

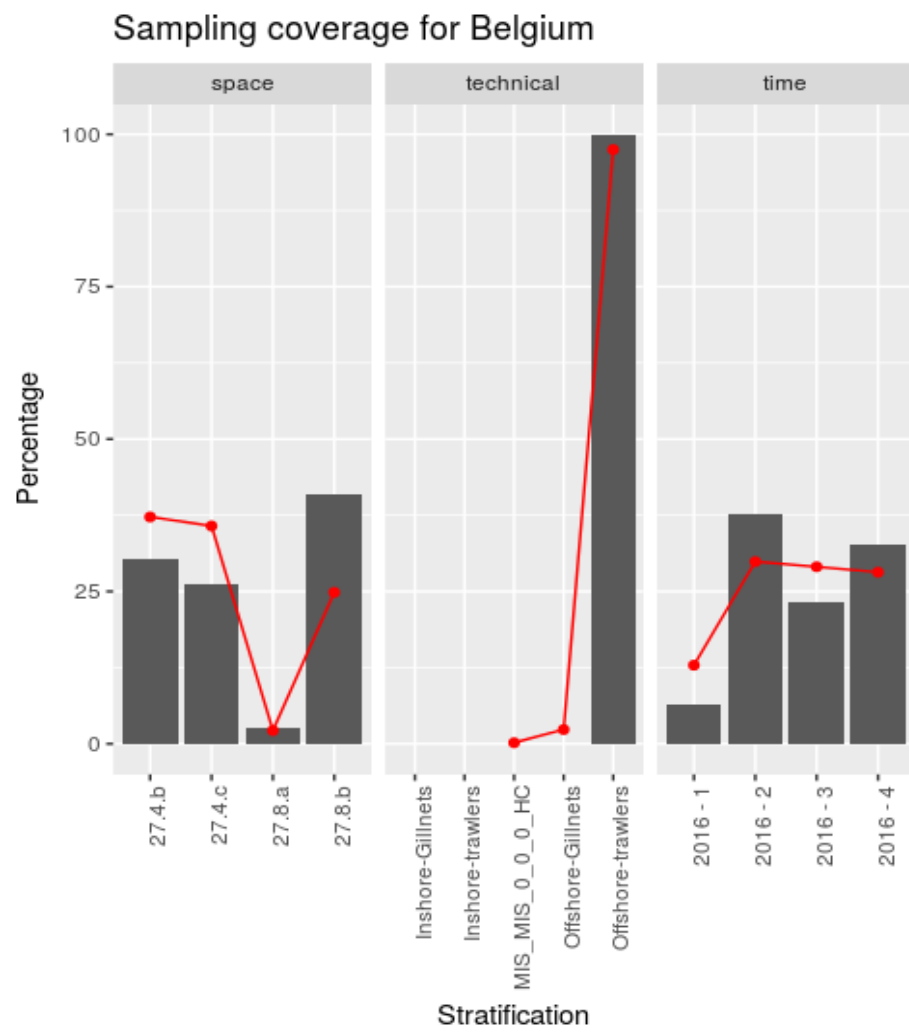
```
#Csr<-rezdelta[[1]]
CSv <- csDataVal(CSr) ; CSc <- csDataCons(CSv,myStr)
CLv <- clDataVal(CLv); CLc <- clDataCons(CLv,myStr)
CEv <- ceDataVal(CEr); CEc <- ceDataCons(CEv,myStr)
if(F){
  CSc<-alkLgthRec(CSc,type='stepIncr',10,preview=FALSE,post-
view=FALSE,update=TRUE)
}
save(CSv,CLv,CEv,CSc,CLc,CEc,file="datavalcons.Rdata")
```

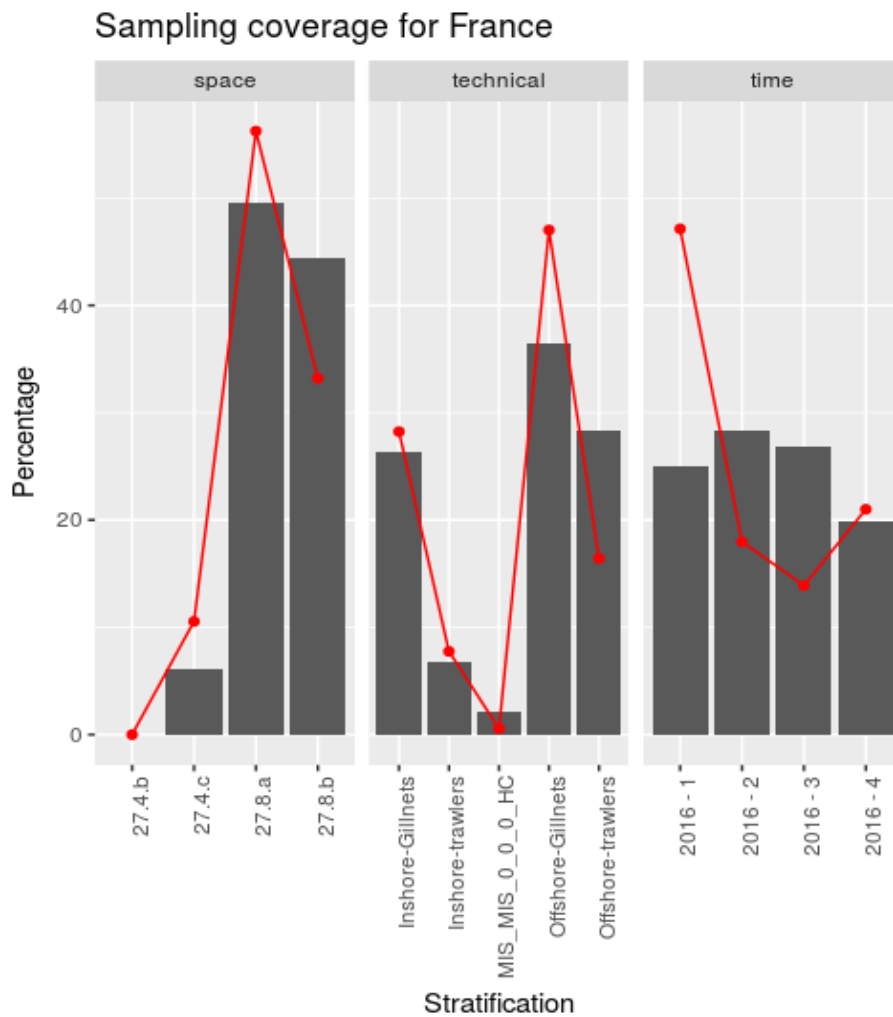
### Sample coverage

The sample coverage is checked according to the stratification: the percentage in numbers of samples by strata is presented against the percentage of landings in each modality (red line).

### Sampling coverage for Belgium and France







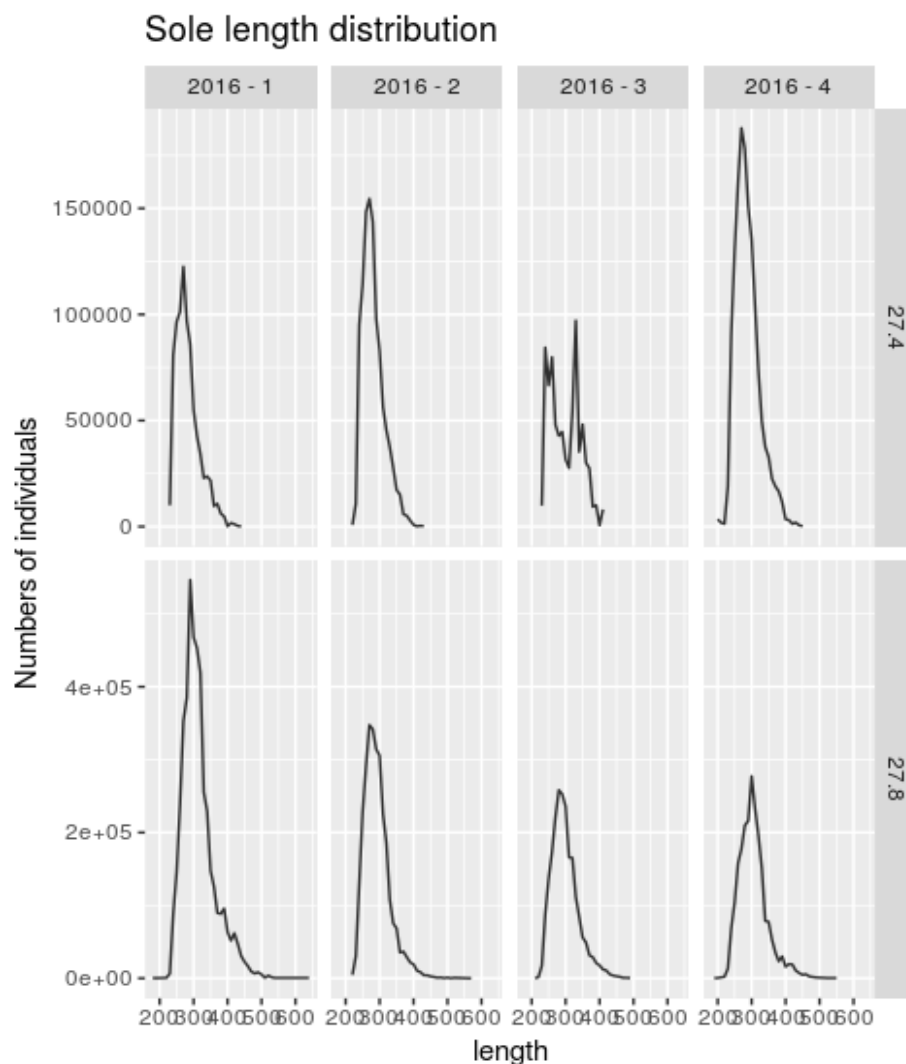
### *Raised length composition of sole stock*

The length composition is estimated using ratio-estimator based on landings.

```
rezdolen<-dolen(CSc,CLc,myStr,idspecies="Solea solea")
dbelan<-rezdolen$dbelan
save(dbelan,file='dbelan.Rdata')
pltlenlan<-graphlen(dbelan,id-
stock="sol.8ab&sol.4",rtp,checkn=50,checks=3)
#plot(pltlenlan)
```

To summarize our graphical output, length distributions are combined by ICES Sub-area and quarter.

```
sizefish<-dbelan@lenStruc$estim%>%group_by(space=sub-
str(space,1,4),time,length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="true estimates")
plt1<-ggplot(rbind(sizefish),aes(x=length,y=value))+
  facet_grid(space~time,scales="free_y")+#,space="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Sole length distribution")
plt1
```



### Scenarios

From the first analyses, the length distribution of the sole in the subarea 27.4 and 27.8 are estimated using the dataset collected by Belgium and France in 2016. These length distributions are considered the "true" estimates of the information captured by two national sampling programs. This section is now dedicated to the estimation of the same parameters, with changes in the sampling plan. These changes are scenario testing some modification of the sampling plan and reallocation of the sampling effort between two countries. To do so, samples are removed according to the scenario, in a fixed or random way and length distribution are estimated based on the new samples base. The results are compared to the "true" estimates using the original number of samples. To compare the 2 distribution, the Earth Mover's Distance (EMD) is used. In statistics, the (EMD) is a measure of the distance between two probability distributions over a region  $D$  (names Wasserstein metric in mathematics) (Rubner, Tomasi, and Guibas 1998). Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region  $D$ , the EMD is the minimum cost of turning one pile into the other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.

The 6 scenario tested in this report are:

Scenario 1: remove a semester of French market samples in the Bay of Biscay.

Scenario 2: remove the Belgian samples in the Bay of Biscay.

Scenario 3: remove the French samples in North Sea.

Scenario 4: decrease French sampling effort in Bay of Biscay.

Scenario 5: decrease the number of sampled hauls in the samples.

Scenario 6: decrease the number of fish sampled by hauls in the sample.

### Scenario 1

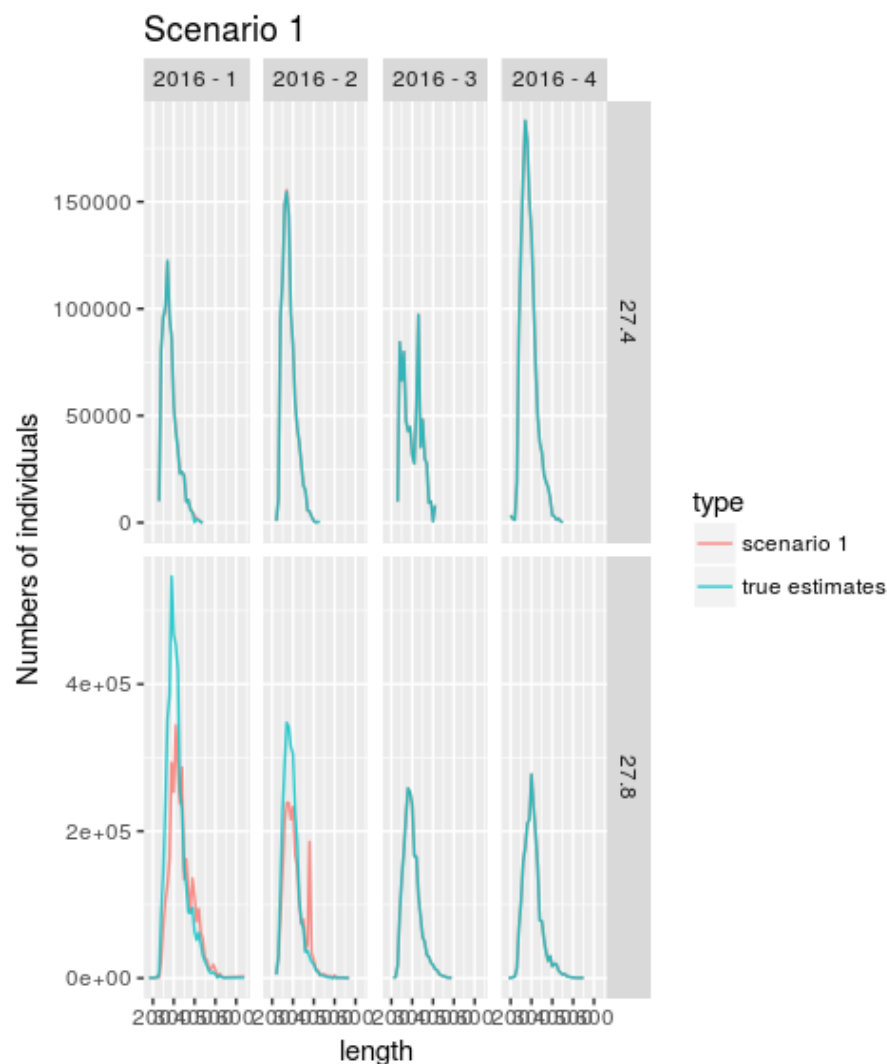
In this scenario, the French market samples from the first semester are removed. This scenario tests practical issue in sampling plan: the realization of a part of the sampling plan is compromised during a given amount of time.

```
test<-as.numeric(substr(as.character(CSc@hh$time),8,8)) %in% c(1:2) &
CSc@hh$sampType=="M" & CSc@hh$vs1FlgCtry=="FRA"
idtrip<-unique(CSc@hh$trpCode[test])
CSctmp<-subset(CSc,! (trpCode%in%idtrip),table="hh")
rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="Solea solea")
dbelan1<-rezdolen$dbelan
sizefish1<-dbelan1@lenStruc$estim%>%group_by(space=sub-
str(space,1,4),time,length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="scenario 1")
#compute emd
emd1<-full_join(sizefish%>%transmute(space,time,length,val0=value),
  sizefish1%>%transmute(space,time,length,val1=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time)%>%summarise(emd=emdL1(val0,val1))
names(emd1)<-c("Subarea","Quarter","Earth movers distance")
set.caption("Earth mover's distance between distributions by temporal
and space strata for scenario 1")
pander(as.data.frame(emd1),style="simple")
```

Earth mover's distance between distributions by temporal and space strata for scenario 1

SUBAREA	QUARTER	EARTH MOVERS DISTANCE
27.4	2016 - 1	47009
27.4	2016 - 2	62482
27.4	2016 - 3	0
27.4	2016 - 4	0
27.8	2016 - 1	34557857
27.8	2016 - 2	14803275
27.8	2016 - 3	0
27.8	2016 - 4	0

```
#compare
plt1<-ggplot(rbind(sizefish,sizefish1),aes(x=length,y=value,col-
our=type))+
  facet_grid(space~time,scales="free_y")+#,space="free_y")+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Scenario 1")
plt1
```



As expected, only the distribution for the quarter 1 and 2 change. For the subarea 27.8, the changes are visible: if the shape of the distributions are similar, they peak above the original value. For the subarea 27.4, the distributions keep the same shape. Only the Earth mover's distance indicates a slight modification in the distribution.

### Scenario 2

In this scenario, the Belgian samples in the Bay of Biscay are removed from the original data. This area is far away from the Belgian coast and complicates the organization of the on board sampling.

```
test<-grepl("27.8",CSc@hh$space) & CSc@hh$vs1FlgCtry=="BEL"
idtrip<-unique(CSc@hh$trpCode[test])
CSctmp<-subset(CSc,! (trpCode%in%idtrip),table="hh")
rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="Solea solea")
dbelan1<-rezdolen$dbelan
sizefish1<-dbelan1@lenStruc$estim%>%group_by(space=sub-
str(space,1,4),time,length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="scenario 1")
#compute emd
emd1<-full_join(sizefish1%>%transmute(space,time,length,val0=value),
  sizefish1%>%transmute(space,time,length,val1=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time)%>%summarise(emd=emd1(val0,val1))
```

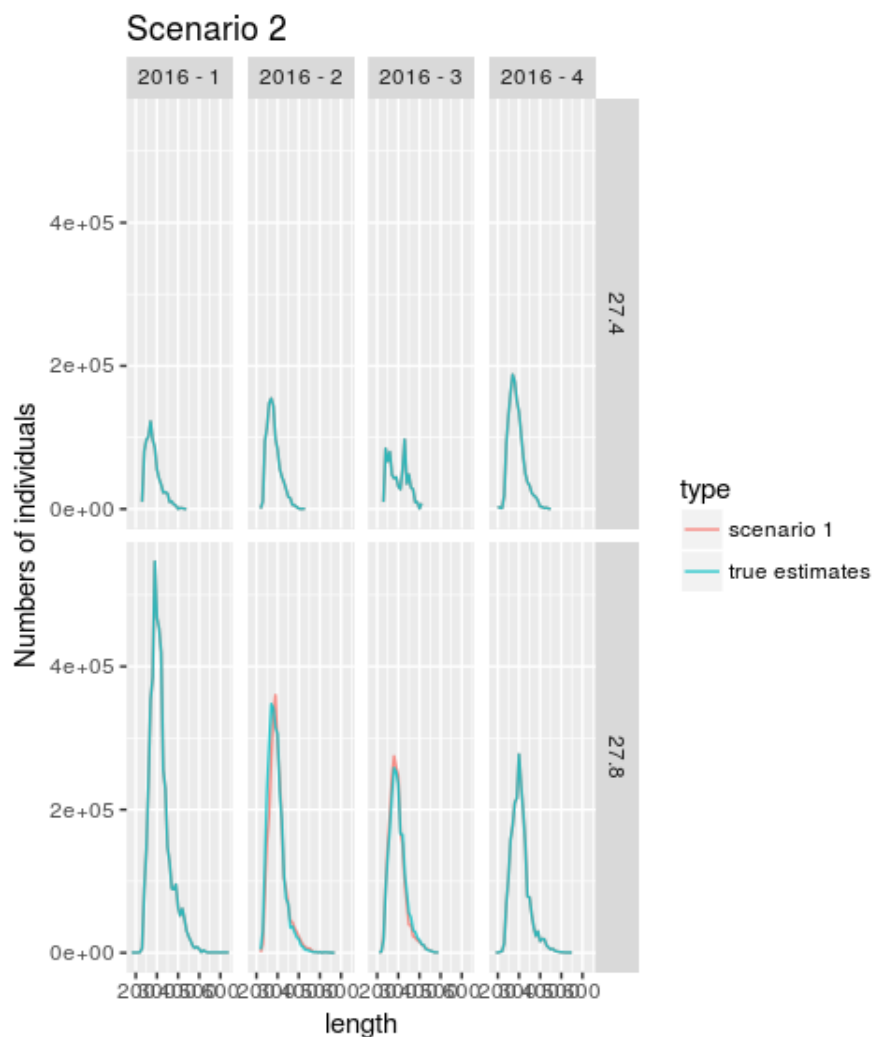
```
names(emd1)<-c("Subarea","Quarter","Earth movers distance")
set.caption("Earth mover's distance between distributions by temporal and
space strata for scenario 2")
pander(as.data.frame(emd1),style="simple")
```

Earth mover's distance between distributions by temporal and space strata for scenario 2

SUBAREA	QUARTER	EARTH MOVERS DISTANCE
27.4	2016 - 1	0
27.4	2016 - 2	0
27.4	2016 - 3	0
27.4	2016 - 4	0
27.8	2016 - 1	0
27.8	2016 - 2	7240174
27.8	2016 - 3	1050283
27.8	2016 - 4	0

```
#compare
plt1<-ggplot(rbind(sizefish,sizefish1),aes(x=length,y=value,col-
our=type))+
  facet_grid(space~time)+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Scenario 2")
plt1
```





As expected, only subarea 27.8 shows differences in length distribution. These differences are only on the quarter 2 and 3 (corresponding to the Belgian samples). The differences between the distributions of the quarter 2 and 3 are small.

### Scenario 3

In this scenario, the French samples in the North Sea are removed from the original data. This scenario is complementary to the scenario 2.

```
#test<-grepl("27.8",CSc@hh$space) & CSc@hh$vsLFlgCtry=="BEL"
#idtripbel<-unique(CSc@hh$trpCode[test])
test<-grepl("27.4",CSc@hh$space) & CSc@hh$vsLFlgCtry=="FRA"
idtripfra<-unique(CSc@hh$trpCode[test])
#idtrip<-c(as.character(idtripbel),as.character(idtripfra))
idtrip<-c(as.character(idtripfra))

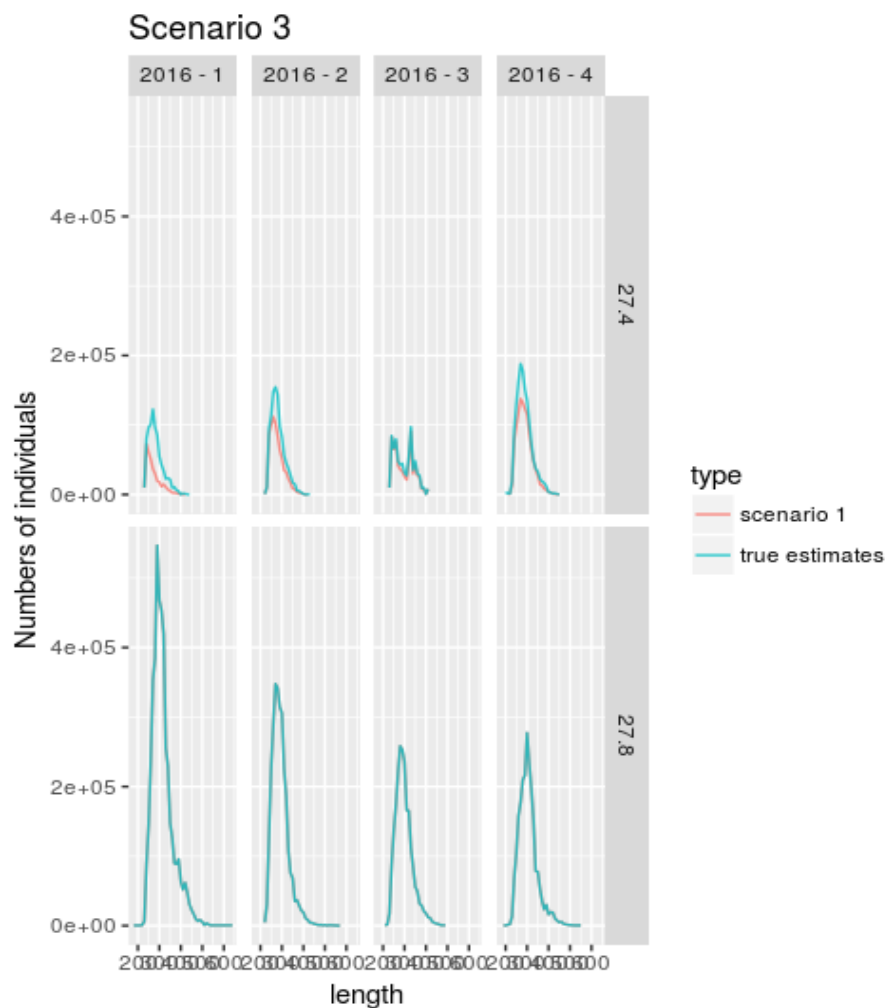
CSc$tmp<-subset(CSc,! (trpCode%in%idtrip),table="hh")
rezdolen<-dolen(CSc$tmp,CLc,myStr,idspecies="Solea solea")
dbelan1<-rezdolen$dbelan
sizefish1<-dbelan1@lenStruc$estim%>%group_by(space=substr(space,1,4),time,length)%>%
  summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
  mutate(length=as.numeric(length),type="scenario 1")
#compute emd
emd1<-full_join(sizefish1%>%transmute(space,time,length,val0=value),
  sizefish1%>%transmute(space,time,length,val1=value))
emd1[is.na(emd1)]<-0
```

```
emd1<-emd1%>%group_by(space,time)%>%summarise(emd=emdL1(val0,val1))
names(emd1)<-c("Subarea","Quarter","Earth movers distance")
set.caption("Earth mover's distance between distributions by temporal
and space strata for scenario 3")
pander(as.data.frame(emd1),style="simple")
```

Earth mover's distance between distributions by temporal and space strata for scenario 3

SUBAREA	QUARTER	EARTH MOVERS DISTANCE
27.4	2016 - 1	7504434
27.4	2016 - 2	4755967
27.4	2016 - 3	821073
27.4	2016 - 4	5400240
27.8	2016 - 1	0
27.8	2016 - 2	0
27.8	2016 - 3	0
27.8	2016 - 4	0

```
#compare
plt1<-ggplot(rbind(sizefish,sizefish1),aes(x=length,y=value,col-
our=type))+
  facet_grid(space~time)+
  ylab("Numbers of individuals")+xlab("length")+
  geom_line(alpha=.8)+ggtitle("Scenario 3")
plt1
```



Changes in the length distribution of the four quarter are clearly visible. The shapes are quite similar for the quarter 2 to 4, with peaks above the original values for the quarter 2 and 4. For the quarter 1, the distribution is very different from the original one.

#### Scenario 4

In this scenario, the French sampling effort was decreased in the Bay of Biscay, from 100% to 10% of the number of original trip sampled. Trips are sampled randomly with replacement in 30 simulations.

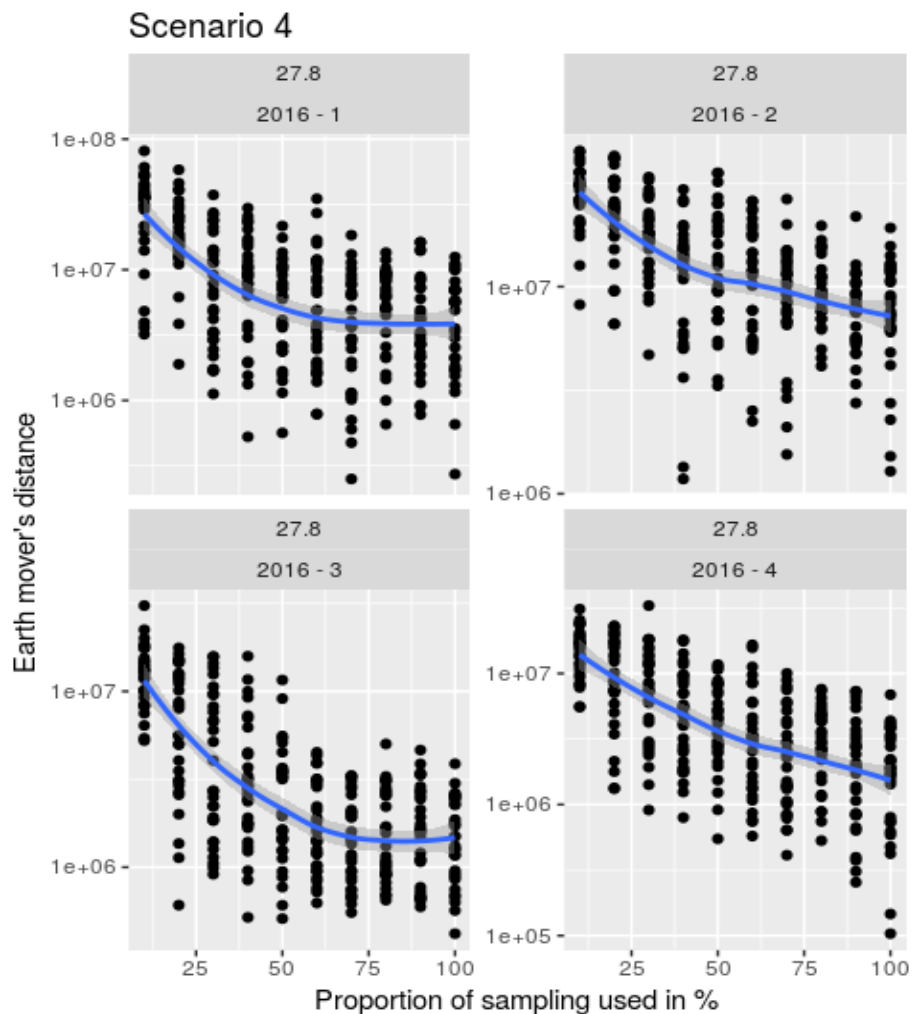
```
if(file.exists("sizefish4.rdata")){
  load("sizefish4.rdata")
}else{
  sizefish2<-data.frame()
  idtrip<-unique(CSc@hh$trp-
Code[CSc@hh$vs1FlgCtry=="FRA"&grepl("27.8",CSc@hh$space)])
  for(i in 1:30){
    for(j in seq(0.1,1,0.1)){
      print(paste(i,j))
      idtrip0<-sample(idtrip,length(idtrip)*j,replace=T)
      CSctmp<-subset(CSc,trpCode%in%idtrip0,table="hh")
      rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="Solea solea")
      dbelantmp<-rezdolen$dbelan
      sizefishtmp<-dbelantmp@lenStruc$estim%>%group_by(time,space=sub-
str(space,1,4),length)%>%
```

```

      summarise(value=sum(as.numeric(value),na.rm=T))%>%un-
group()%>%
      mutate(length=as.numeric(length),type="scenario
4",prop=j,rep=i)
    sizefish2<-rbind(sizefish2,sizefishtmp)
  }
}
save(sizefish2,file="sizefish4.rdata")
}

#compute emd
emd1<-left_join(sizefish2%>%trans-
mute(space,time,length,prop,rep,val1=value),
sizefish%>%transmute(space,time,length,val0=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time,prop,rep)%>%summa-
rise(emd=emdL1(val1,val0))
plt1<-ggplot(emd1,aes(x=100*prop,y=emd))+geom_point()+
scale_y_log10()+facet_wrap(space~time,scale="free_y")+geom_smooth()+
xlab("Proportion of sampling used in %")+
ylab("Earth mover's distance")+
ggtitle("Scenario 4")
plt1

```



```

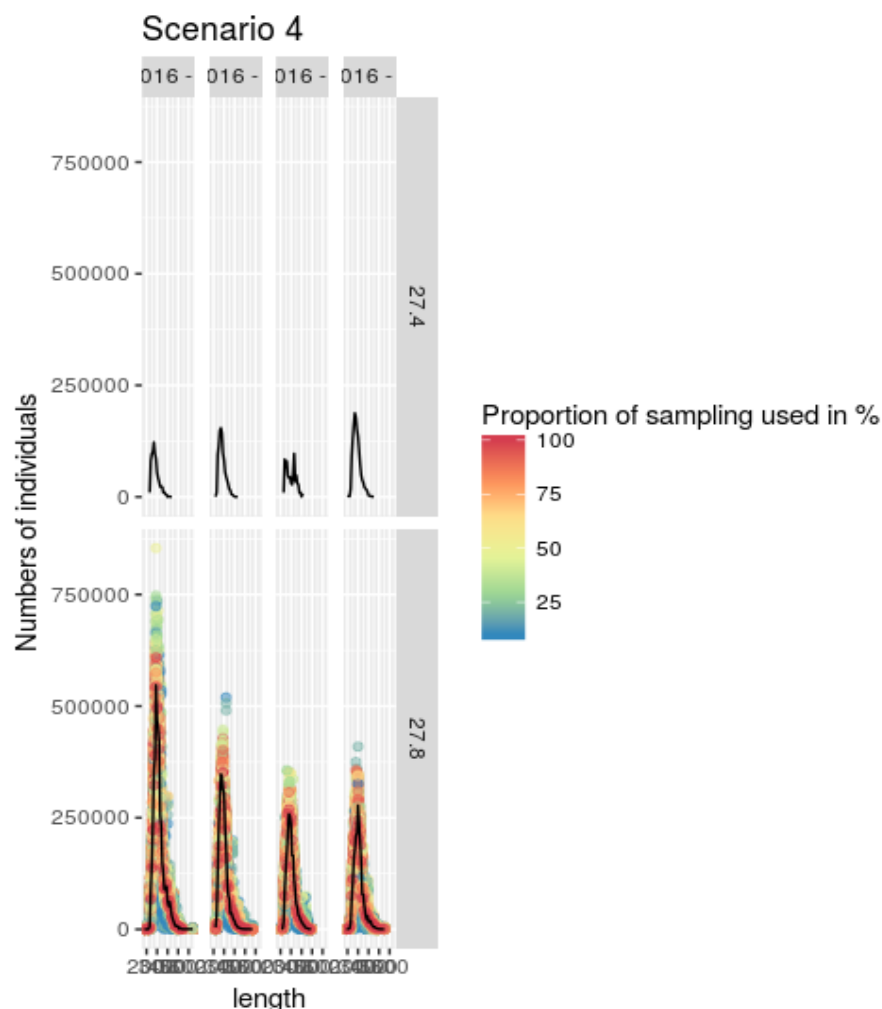
#extra info
sizefish2avg<-sizefish2%>%group_by(time,space,length,prop)%>%
  summarise(value=mean(value,na.rm=T))
#compare
plt2<-ggplot(sizefish2,aes(x=length,y=value))+
  #stat_bin2d(bins=50)+

geom_point(data=sizefish2,aes(x=length,y=value,group=prop,color=100*prop),alpha=.5)+
  scale_colour_distiller(palette='Spectral',name="Proportion of sampling used in %") +

#geom_line(data=sizefish2avg,aes(x=length,y=value,group=prop,color=prop,size=0.1),alpha=.5)+
geom_line(data=sizefish,aes(x=length,y=value),color="black",alpha=1)+
  facet_grid(space~time)+
  ylab("Numbers of individuals")+xlab("length")+
  ggtitle("Scenario 4")

plt2

```



The earth mover's distance (EMD) decrease logically with the increase of number of trips used to estimates the length distribution (the blue line is a LOESS smoother). A plateau between 75% to 100% for quarter 1 and 3, and to a less extent for the other quarters, indicates that length distributions maintain some similarities while the number of trips decrease. Visually, the length distribution computed with 75% to 100% of

the trips (red to orange dots) conveys the same amount of information as the original length distribution (black line) for the subarea 27.8.

### Scenario 5

In this scenario, the impact of the number of hauls sampled by trip is investigated. For each trip, 1 to 10, 15 and 20 hauls are selected if the number of hauls available in a given trip is more than the number tested. For the trips with a number of hauls smaller than the number requested they are kept as they are. The length distributions are then computed using the new samples. Hauls are sampled randomly with replacement in 30 simulations.

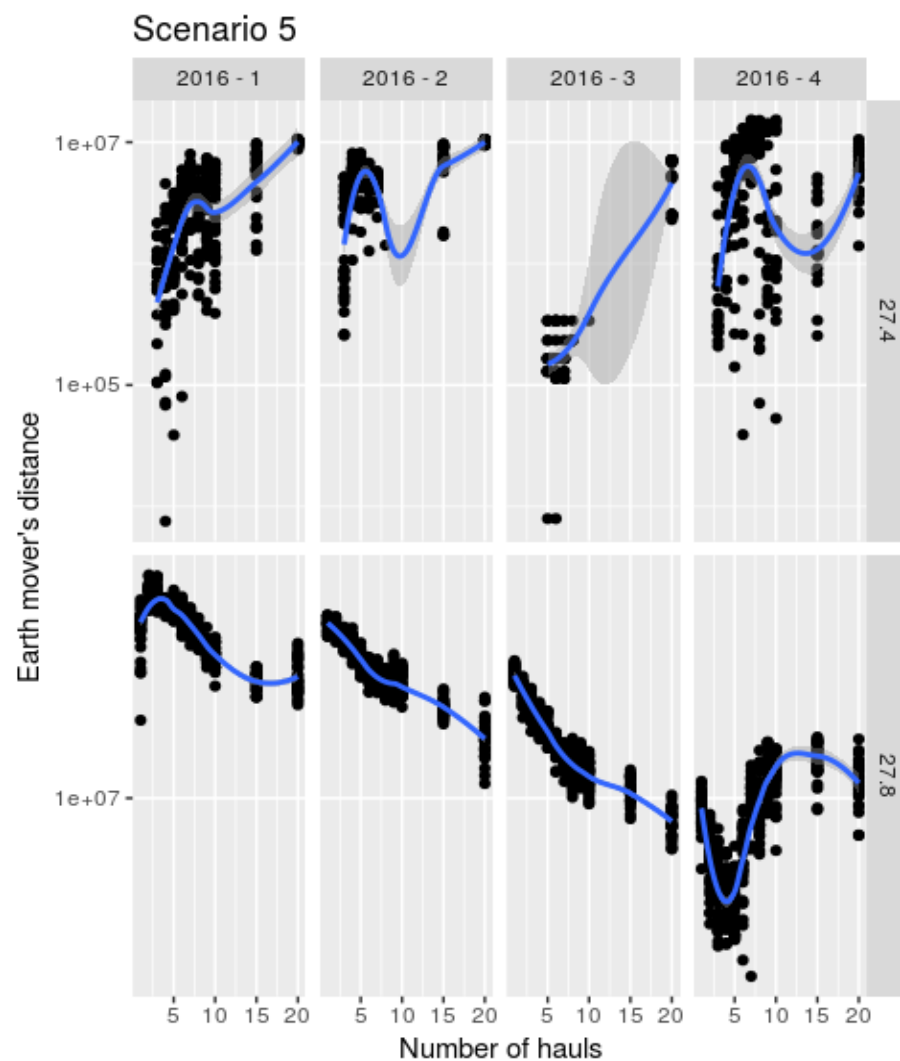
```

if(file.exists("sizefish5.rdata")){
  load("sizefish5.rdata")
}else{
  sizefish2<-data.frame()
  idhaul<-CSc@hh%>%select(trpCode,staNum)%>%distinct()%>%group_by(trp-
Code)%>%mutate(nb=n())%>%ungroup()%>%
  mutate(idhaul=paste(trpCode,staNum))
  for(i in 1:30){
    for(j in c(1:10,15,20,30)){
      print(paste(i,j))
      idhaulout<-idhaul%>%filter(nb<j)
      idhaultin<-stratified(idhaul%>%filter(nb>=j), "trpCode", j, re-
place=TRUE)
      CSctmp0<-subset(CSc,paste(CSc@hh$trpCode,CSc@hh$sta-
Num)%in%idhaulout$idhaul,table="hh")
      CSctmp1<-subset(CSc,paste(CSc@hh$trpCode,CSc@hh$sta-
Num)%in%idhaultin$idhaul,table="hh")
      CSctmp<-rbind2(CSctmp0,CSctmp1)
      rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="Solea solea")
      dbelantmp<-rezdolen$dbelan
      sizefishtmp<-dbelantmp@lenStruc$estim%>%
        group_by(time,space=substr(space,1,4),length)%>%
        summarise(value=sum(as.numeric(value),na.rm=T))%>%un-
group()%>%
        mutate(length=as.numeric(length),type="scenario
5",nbhaul=j,rep=i)
      sizefish2<-rbind(sizefish2,sizefishtmp)
    }
  }
  save(sizefish2,file="sizefish5.rdata")
}

#compute emd
emd1<-left_join(sizefish2%>%trans-
mute(space,time,length,nbhaul,rep,val1=value),
  sizefish%>%transmute(space,time,length,val0=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time,nbhaul,rep)%>%summa-
rise(emd=emdL1(val1,val0))
emd1<-emd1%>%filter(nbhaul<=20)
plt1<-ggplot(emd1,aes(x=nbhaul,y=emd))+geom_point()+
  scale_y_log10()+facet_grid(space~time,scale="free_y")+geom_smooth()+
  xlab("Number of hauls")+
  ylab("Earth mover's distance")+
  ggtitle("Scenario 5")

plt1

```

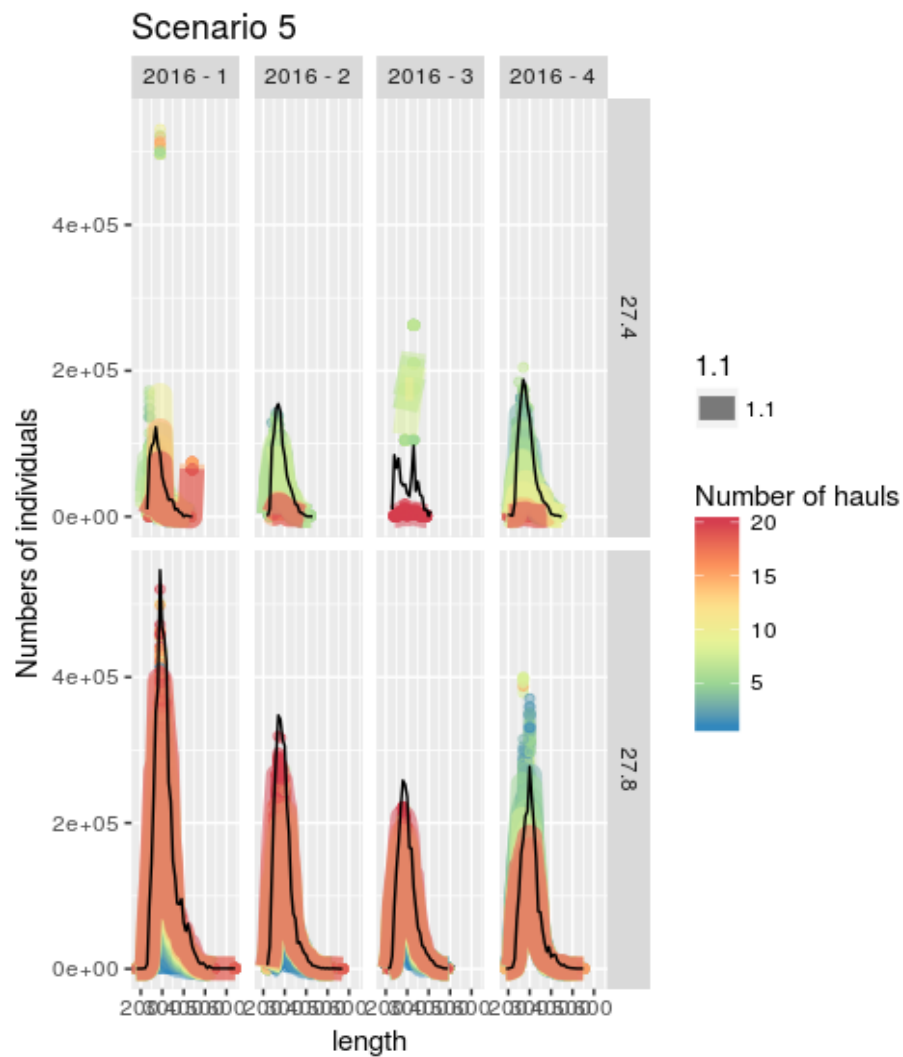


```

sizefish2avg<-sizefish2%>%group_by(time,space,length,nbhaul)%>%
  summarise(value=mean(value,na.rm=T))%>%filter(nbhaul<=20)
#compare
plt2<-ggplot(sizefish2%>%filter(nbhaul<=20),aes(x=length,y=value))+
  #stat_bin2d(bins=50)+
  geom_point(data=sizefish2%>%fil-
ter(nbhaul<=20),aes(x=length,y=value,group=nbhaul,color=nbhaul),alpha=.5)+
  scale_colour_distiller(palette='Spectral',name="Number of hauls") +

geom_line(data=sizefish2avg,aes(x=length,y=value,group=nbhaul,color=nbhaul,s
ize=1.1),alpha=.5)+
  geom_line(data=sizefish,aes(x=length,y=value),color="black",al-
pha=1)+
  facet_grid(space~time)+
  ylab("Numbers of individuals")+xlab("length")+
  ggtitle("Scenario 5")
plt2

```



The earth mover's distance (EMD) decrease with the increase of number of hauls used to estimates the length distribution for the subarea 27.8 in quarter 1–3, while some noisy patterns are seen for the other quarters. The type of the trips could influence the simulation, and no clear conclusion can be made regarding these results. From the visual comparison of the length distribution, the number of sampled hauls seems to play an important role in the shape of the length distribution. Further investigations, outside the scope of this report, are needed.

#### **Scenario 6**

In this scenario, the impact of the number of fish sampled by haul inside a trip is investigated. For each haul, 10 to 100 fish are taken randomly with replacement from the original measurements, and length distributions are computed using the new samples. 30 simulations were performed using this scenario.



```

if(file.exists("sizefish6.rdata")){
  load("sizefish6.rdata")
}else{
  sizefish2<-data.frame()
  #expand the hl table
  hlexp <- CSc@hl[rep(row.names(CSc@hl), CSc@hl$lenNum), ]
  hlexp<-hlexp%>%mutate(lenNum=1,idhaul=paste(trpCode,staNum))
  for(i in 1:30){
    for(j in seq(10,100,10)){
      print(paste(i,j))
      aa<-stratified(hlexp,"idhaul",j,replace=TRUE)
      CSctmp<-CSc
      CSctmp@hl<-aa%>%group_by(PSUId,SSUId,TSUId,time,space,technical,sort,sampType,landCtry,
                             vs1FlgCtry,proj,trpCode,staNum,spp,sex,lenCls)%>%
        summarise(lenNum=sum(lenNum))%>%ungroup()
      rezdolen<-dolen(CSctmp,CLc,myStr,idspecies="Solea solea")
      dbelantmp<-rezdolen$dbelan
      sizefishtmp<-dbelantmp@lenStruc$estim%>%group_by(time,space=substr(space,1,4),length)%>%
        summarise(value=sum(as.numeric(value),na.rm=T))%>%ungroup()%>%
        mutate(length=as.numeric(length),type="scenario
6",nbfish=j,rep=i)
      sizefish2<-rbind(sizefish2,sizefishtmp)
    }
  }
  save(sizefish2,file="sizefish6.rdata")
}

#compute emd
emd1<-left_join(sizefish2%>%transmute(space,time,length,nbfish,rep,val1=value),
               sizefish%>%transmute(space,time,length,val0=value))
emd1[is.na(emd1)]<-0
emd1<-emd1%>%group_by(space,time,nbfish,rep)%>%summarise(emd=emdL1(val1,val0))
plt1<-ggplot(emd1,aes(x=nbfish,y=emd))+geom_point()+
  scale_y_log10()+facet_grid(space~time,scale="free_y")+geom_smooth()+
  xlab("Number of fish")+
  ylab("Earth mover's distance")+
  ggtitle("Scenario 6")
plt1

sizefish2avg<-sizefish2%>%group_by(time,space,length,nbfish)%>%
  summarise(value=mean(value,na.rm=T))
#compare
plt2<-ggplot(sizefish2,aes(x=length,y=value))+
  #stat_bin2d(bins=50)+
  geom_point(data=sizefish2,aes(x=length,y=value,group=nbfish,color=nbfish),alpha=.3)+
  scale_colour_distiller(palette='Spectral',name="nbhaul") +
  #geom_line(data=sizefish2avg,aes(x=length,y=value,group=nbhaul,color=nbhaul,
  size=1.1),alpha=.5)+
  geom_line(data=sizefish,aes(x=length,y=value),color="black",alpha=1)+
  facet_grid(space~time)+
  ylab("Numbers of individuals")+xlab("length")+
  ggtitle("Scenario 6")
plt2

```

The earth mover's distance (EMD) reaches local minima, in some strata, indicating an optimal number of fish to be measured. But these simulation mix data with different sampling protocol (marketvs.on-board etc.), and no conclusion should be made according to this simulation. These simulations should be performed inside a well-defined part of the sampling sharing the same protocol.

### *Discussion*

Scenarios 1 to 3 investigate some fixed effects on the sampling plan to the length distribution. The dependence on market sampling for France is strong for the sole stocks, and the reallocation to sampling effort from market to at-sea sampling should be done carefully. Scenarios 2 and 3 tested if sampling plan could be reorganized at the regional level. Removing the Belgian samples in the Bay of Biscay did not change the length distribution estimates. Moreover there is a dependence on French samples for the estimates of the sole length distribution in the North Sea, but these samples are mainly on board samples. Sampling fleets doing long fishing trips are costly: reallocation of sampling effort in the closest area by country (Bay of Biscay for France and North Sea for Belgium), should be possible in some way. Scenario 4 highlights some possibility to decrease sampling effort on the very well sampled fleet in the Bay of Biscay for France: with 75% of the trips, the length estimates are very similar to the original one.

Scenarios 5 and 6 were not adapted to the dataset and the scope of this study. At the hauls and the number of fish measured level, simulations have to be performed on samples sharing the same protocol. But code is now available and will be used in the next WKBIOPTIM workshop.

## Annex 5: Proposal: The second Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM 2)

---

The **second Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM 2)** chaired by Ana Cláudia Fernandes (Portugal) and Maria Teresa Facchini (Italy) will meet in Ifremer Port en Bessin, May 2018 to:

- a ) Development of R-Toolbox: Based on selected case studies continue to develop, test and document R-scripts extending the analysis to other biological variables (e.g. age and maturity); discuss the calculation of the effective sample size for these variables.
- b ) Discuss methods for the objective selection of the biological parameters to optimise in view of the simulated distribution outputs and in line with end-users' needs.
- c ) Development of quality indicators: evaluate a second set of quality indicators (after feedback from WGBIOP and WGCATCH)
- d ) Road-Map and implementation: discuss progress achieved in implementation at national level. The road map should involve all people with a stake in the process (i.e. samplers, data analysts, end-users, and programme managers), include clarification (to all parties involved) of the analysis that will be carried out, their need, and proper consideration of the needs of both lab/field staff and the end-users of the data.

WKBIOPTIM 2 will report by XXX to the attention of the SSGIEOM Committee.

### Supporting Information

---

Priority	This workshop is considered to have a high priority for already established and new commercial fishery and survey sampling programmes developed under the MAUP. The expectation is that the time and costs that will be saved by the development and implementation of the R-toolbox will be fundamental to increase data provision on data-limited stocks and environmental variables. The basic toolbox was developed by WKBIOPTIM and in order for the full potential of this tool to be realised further testing and input are required under a wider range of scenarios.
----------	---

---

Scientific justification	<p>Statistical sound sampling is a requirement of the new EU-MAUP that now specifies that “where data are to be collected by sampling, Member States shall use statistically sound designs” (COM IMPL DEC 2016/1701). One important component of a “statistically sound design” is that sampling effort is optimized and fit for purpose, i.e. that time and costs spent in sampling can be effectively justified in terms of quality of the information finally provided to end-users. There is an increasing demand to determine MSY reference points for an increasing number of stocks, including many data-limited stocks, and, at the same time, to collect additional environmental and biological information. This makes optimisation of the number of length measurements, age and maturity estimation a priority since these tasks involve costs and time that could alternatively be spent in data collection of other stocks and/or variables. It is important that the national laboratories of MS have common tools to quantify the effects, advantages and disadvantages of different sampling intensities and sampling designs so they can optimise sampling in terms of time and costs savings. Several ICES EG’s, including e.g. WKPRECISE 2009, PGCCDBS 2012, PGDATA 2015 and WKCOSTBEN 2016 have pointed out that clustering effects in multistage catch sampling programmes may lead to effective sample sizes much lower than the number of units sampled, e.g. fish caught during one trip or haul often have more similar characteristics than the general population of fish they came from. This effect highlights the likely existence of oversampling in the lower stages of many national catch sampling programmes (e.g. trips, hauls within trips, samples within hauls), where an excessive number of individuals may be being sampled and not accruing significant additional information to estimates provided to end-users.</p> <p>The Workshop on Optimization of Biological Sampling at Sample Level (WKBIOPTIM) developed and tested a set of simple R-scripts (based on the RBD exchange format) which produce a range of simple statistical and graphical outputs to be used for discussion of appropriate levels of biological sampling of different stocks. Data quality indicators of the biological variables under the optimization procedures carried out at the workshop were discussed and a roadmap for future discussions with end-users outlined. Given the positive feedback both from national labs and RCM’s it is recommended that a second workshop takes place to continue the work initiated. It is envisioned that WKBIOPTIM should be a joint workshop bringing together experts from WGCATCH and WGBIOP and that the main results will be brought to further discussion by these two groups. Case studies will be carefully selected and developed to calculate the effective sample size for length, age and maturity (ToR a); combine data from different on board and onshore sampling programs (ToR b) and discuss the consequences of pooling strata and low sample sizes under optimisation (ToR c). Outputs from these case studies will assist discussions on the objective selection of the biological parameters to optimise in view of the simulated distribution outputs, in line with end-users needs (ToR d). Testing and documentation of the code and the R-tool box will be ongoing and implemented via the case studies.</p>
Resource requirements	<p>The data collection programmes which provide the main input to this group are already underway, and resources are already committed. All EU countries already have the datasets required for analysis available in the RDB format. Some preparation of R-scripts and selection of case-studies will be required prior to the meeting. It is expected that a progress meeting will take place 6 months following the meeting where feedback from the national laboratories will be required.</p>

Participants	The Workshop is expected to attract wide interest from those involved in WGCATCH and WGBIOP and should include a subset of participants familiar with R-code to the level of “loop coding” and “function building” and a subset of participants experienced in age and reproduction analysis. In view of its relevance to data collection within ICES, the EU-MAUP and regional sampling designs, it should include those involved in the annual planning of sampling and and laboratory analysis, including e.g. number of trips to be sampled and fish to be measured and aged/sexed. Members of survey groups located under SSGIEOM are also among the probable participants.
Secretariat facilities	Some secretarial support will be needed. The WK should take place in 2018. Therefore it will need to be approved by ACOM and SCICOM in early 2018.
Financial	Member States may fund this through their EMFF programme
Linkages to advisory committees	ACOM and SCICOM
Linkages to other committees or groups	WGCATCH, WGBIOP, PGDATA, SSGIEOM
Linkages to other organizations	RCGs

## Annex 6: Recommendations

RECOMMENDATION	ADDRESSED TO
1. Give input on quality indicators of biological parameters; identify possible case studies based on commonly used sampling strategies and collected parameters”	WGBIOP and WGCATCH
2 Discuss and comment on outcomes of the workshop in what concerns sample level and multi-level analyses.	WGCATCH
3. Provide guidelines which identify various end-users needs under optimisation procedures	PGDATA