WD presented at ICES WGCATCH, 9–13 November 2015, Lisbon (Portugal)

# Developments in the "Northern and Southern Hake" Case Study of FishPi

Jose Rodríguez-Gutierrez[1], Nuno Prista[2], Lucia Zarauz[3], Manuela Azevedo[4] and Jose Castro[5*]

[*] on behalf of all participants and Institutes involved

[1] Instituto Español de Oceanografía (IEO), Promontorio de San Martin s/n, 39004, Santander, Cantabria, Spain
[2] Havsfiskelaboratoriet, Institutionen för akvatiska resurser, Sveriges lantbruksuniversitet, Sweden
[3] AZTI Tecnalia, Txatxarramendiirla z/g, 48395 Sukarrieta, Basque Country, Spain
[4] Instituto Português do Mar e da Atmosfera, Avenida de Brasilia, Lisboa, Portugal
[5] Instituto Español de Oceanografía (IEO), Subida Radio Faro 50, 36390 Vigo, Galicia, Spain

## Introduction and objectives

The overarching objective of the FishPi project is to strengthen regional cooperation in the area of fisheries data collection. The role of every case study within the project is to bring together the countries with the most involvement in the fisheries selected to coordinate and cooperate in the search of a probabilistic regional sampling design. To this aim, case study 4 (CS4) is focused in the Northern and Southern Hake stocks.

The work done in the case study include the description of the fishery at a regional level, the compilation of the present national sampling activity, the compilation of the logbooks and/or sales notes from 2013 and 2014 –to have a single regional data set of all trips of interest in the region– and different runs of simulations to test the selected sampling scenarios and stratifications. Finally, an objective evaluation of the performance of these regional designs is expected to understand the changes needed compared to the present situation.

This document presents the progress done in this case study, from the data compilation to the simulations, documenting the different steps taken and allowing an understanding of the pending tasks.

## Work planning and operational meetings

Work plan for CS4 was articulated in accordance with the calendar established for all case studies in the project which has included several online meetings –to determine aspects as the specific scope of the case studies or the content and format of the data needed – and two operational meetings.

First meeting was conceived for review and consolidation of data sets received, and start up of simulations was held in Aberdeen (22-26 June) with participants from all case studies.

A second meeting, specific for CS4, was held in Lisbon (14-16 October) with the participation of IPMA, IFREMER, AZTI and IEO. Nuno Prista, formerly at IPMA and recently employed at SLU/AQUA in Sweden participated by skype. The agenda included the following points: to check the final data compilation –fishPi CS4 matrix and the overall sampling resource allocation of institutes involved in the sampling–; to organize the descriptive documents

writing –fisheries and sampling descriptions–; to discuss and decide the final list of simulation scenarios; to present the results from the preliminary runs with the whole CS4 matrix; to split the CS4 matrix by stock and run new simulations; to discuss the results and write a draft with the main conclusions.

In both cases work was slowed down because of the time needed to compile and clean up proper data sets (explained below).

**Data compilation**

A data call for the Member States whose fleets operate in the area of interest of every case study was launched in May requesting trip level information in a new format (FishPi format). It was also the first time national institutes faced a data call requesting trip-level landings for all fleets involved in the fisheries and the first time this kind of analysis in the aforementioned disaggregated level was being undertaken at a regional level.

This context obliged to make some modifications in the data requested —for CS4 this meant to limit the data set to hake information—, and proved to make data compilation slower than initially expected as a data cleanup process was needed. Most part of problems were related to misunderstandings of the data call, missing values in key fields and wrong coding of variables.

After detailed checks of national data, CS4 produced in October a final harmonized and checked data set ready for subsequent analysis. In the cases of no answers from some laboratories CS4 moved on based on pragmatic decisions. In the case of French data, it has not been possible until now to work a proper file due to incorrect coding for location variables which were covered with IFREMER codes instead UNLOCODE. Among other possible analysis, field "*onShoreSampLoc*" is essential to disaggregate the matrix by country (responsible for sampling) as it is required in one of the scenarios for simulation decided (see later). France is the main country in the Northern hake stock fishery (52% of the landings), therefore this prevented any further development of a complete matrix for all areas and the subsequent work had to focus exclusively on Southern hake stock for the time being.

The fishpi CS4 Southern matrix was compiled, cleaned and formatted. The CS4 Southern data set enables simulation models of alternative sampling designs to be tested.

From this compilation process institutes got experience with the fishPi format. Summary of main fields (interpretation/use):

- "*recType*": "at-sea scheme" means that landings sampling only is suited on-board (not that sampling is oriented to estimate discards). This situation only happens in the Gulf of Cadiz fleets.
- Fields "*onShoreSampLoc*" and "*atSeaSampLoc*" are essential to disaggregate the matrix by sampling country (instead of "*vslFlgCtry*").

The overall sampling resource allocation of institutes involved in the sampling was also compiled. In accordance with the project work this was to be quantified in terms of the number of market events and sampling trips achieved (Table 1).

**Scenarios**

Following the project objectives each case study has to set a list of scenarios that consist of differing stratifications and effort allocations. Simulations are supposed to allow calculation of different statistical measures of design performance, allowing for objective comparisons of the design effect.

Four sets of scenarios were defined for CS4, each of them representing different sampling designs (scheme below). The initial idea was to apply each sampling design to two different settings or populations:

- all fleets operating in the area of study (setting A),
- only the demersal fleets operating in the area (setting B)

The list of scenarios for CS4-Shake:

| **Scenario** | **Setting** | |
|---|---|---|
| | **A** | **B** |
| | Total fleets | Demersal fleets |
| **1** | Regional approach (RE):<br>a) Without stratification<br>b) Stratified by harbour (h)<br>c) Stratified by quarter (q)<br>d) Stratified by harbour and quarter (h*q) | Idem |
| **2** | National approach (MS):<br>a) Stratified by country<br>b) Stratified by country and harbour (h)<br>c) Stratified by country and quarter (q)<br>d) Stratified by country, harbour and quarter (h*q) | Idem |
| **3** | Current sampling scheme (CUR) | Idem |

The exploration of scenarios under sampling by stages can multiply the cases in more options:

- SRS with trip as PSU.
- SRS with harbour*day as PSU and all trips as SSU.
- SRS with harbour*day as PSU but a determined number of n trips in the harbour*day).

Demersal fleets: Vessels using demersal fishing gears (mainly trawl, set gillnet and set longline), as it was specified in the original data request (2 June 2015).

**Settings and Simulations**

The simulation analyses were organized in two steps: (1) definition of the different strata and the sampling effort by strata and (2) simulation of the different sampling scenarios. The results presented are provisional and work is under development.

1. Definition of the different strata and the sampling effort by strata

Two R scripts written in order to set the strata and the sampling effort by strata required to run the simulations under the different scenarios. This step was applied to setting A (total fleets) and setting B (demersal fleets). In both cases the general setting consisted in:

- Use of subset 2013 [design dataset].
- Geographical stratification (harbour): only individualize ports with >75% of hake landings. The rest of ports are aggregated in the same strata.
- Creation of time strata (quarter).
- Splitting by country from field "*onShoreSampLoc*" by extracting the first UNLOCODE two letters.
- Definition of the number of trips and sampling events (port*day) (total and by country).
- Definition of the sampling effort in number of trips and sampling events (port*day) (total and by country).

For setting B the selection of demersal fleets was done at level 5 of DCF metiers: "GNS_DEF", "GTR_DEF", "MIS_MIS", "LHM_DEF", "LLS_DEF", "OTB_DEF", "OTB_CRU", "OTB_MCD", "OTB_MPD", and "PTB_MPD".

As an example of the results scenario after the settings, Figure 1 and Figure 2 show the setting 1b and 2b for setting B (demersal fleets). In these examples, harbours comprising 75% of hake landings (at regional level and national level) were kept as main strata with all remaining harbours being considered together under a last strata.

2. Simulation of the different sampling scenarios.

Simulations are conducted on 2014 population data, including 555 968 trips (total). The annual numbers of sampled market events and trips per country are 582/684 and 508/809 for Portugal and Spain, respectively. Both countries carry out onshore sampling based on market event – vessel trip, while Spain applies at-sea sampling for the two Spanish metiers operating in the Gulf of Cadiz: purse seiner and bottom otter trawl. As a result, a total of 1493 trips were conducted, and we use n=1500 trips in our simulation.

The goal is to provide several examples on how to choose a good sampling design. For each simulation, we first select a random sample according to a specific sampling design, and estimate the total weight per domain (metier, area and quarter). We then repeat this process for n times.

Eventually, the performance of the sampling design can be judged by the mean and the variance of the estimate from the n replicates. For each scenario, the sampling designs are judged according to two aspects:

- Bias: a good sampling design should be unbiased. Thus, the mean of the estimated value should be the same as the true population value.
- Precision: a good sampling design should achieve low variance (high precision) of the estimated value. A ratio of variance of <1 between a sampling design and a reference sampling design indicates an improved sampling design.
- Frequency of occurrence of zero hake in the samples.

Domain definition: selection of metiers taking into account the ranking of 2014 landings. Seven metiers contain 93.4% of total landings (Table 2), and they are currently used to provided fishing data to ICES (including OTB_MPD_>=55_0_0). Adding all the other metiers to "Other" and multiplying by 2 ICES Divisions (8c and 9a; but 2 metiers only operate in 9a) and 4 quarters, 56 raising domains are obtained.

At this point of the project, we can only present results related to cases in scenario 1 and scenario 2, applied to setting B (demersal fleet). However, work is under development for the rest of scenarios and settings.

- ***Scenario 1a. RE. Case: Simple Random Sampling (SRS) with trip as PSU***

For the sampling simulation the scripts make use of the "survey" package (Lumnley, 2014[1]), mainly the function "svydesign" to set de sampling design. 2000 iterations were run. Post stratification was applied to get the estimated landed weight by metier, area and quarter.

The highest bias is observed in metiers GNS_DEF_80-99_0_0 and PTB_MPD_>=55_0_0 in Division IXa (Figure 3). Simulations were not completed to draw conclusions in the CS4 Lisbon.

- ***Scenario 2a. Case: Two-stage Cluster Random Sampling stratified by country****. With harbour*day as PSU and trip as SSU (the number of trip to be sampled in each PSU has to be determined)*

It was necessary to create one extra variable "onShoreSampLoc&arvDate" in order to define PSU harbour *day. Sampling effort was 650 PSU for Portugal and 850 PSU for Spain.

The selection of the sample was performed with the function "mstage()" from R "sampling" package (Tillé and Matei, 2015 [2]). This function allows the selection of a sample with multistage stratified sampling design. The main advantage of this function is that it automatically calculates the probability of each sampling unit to be sampled, and that it is intuitive and easy to use. Its main drawback is that, with the volume of data needed for CS4, it is extremely slow. Another limitation was the definition of the number of trips to be sampled in each market*day. The more realistic option was to fix it to two trips selected with simple random sampling without replacement. However, this was not possible because some market*days had only 1 trip to sample. We then had two options:

- Number of trips per market*day= 2. Selection methods: simple random sampling with replacement
- Number of trips per market*day= 1. Selection methods: simple random sampling without replacement

The results presented here correspond to the second option.

Estimates were calculated using "svydesign()", "svytotal()" and "svyby()" functions of the R "survey" package (Lumnley, 2014). 200 iterations were run to calculate the total landed weight (Fig 4). Poststratification was applied to get the estimated weight by metier, area and quarter.

[1] T. Lumley (2014) "survey: analysis of complex survey samples". R package version 3.30.

[2] Tillé and Matei (2015). Package "sampling". Version 2.7

Due to the slowness of the code, only 50 iterations were run for the poststratification. Results are presented in Figure 5.

One of the main problems encountered when working on scenario 2, was the slowness of the functions needed to select a sample with a multistage stratified sampling design (mstage) and to apply poststratification to our estimates. To speed up this process, a custom function is being developed based on R "data.table" package. Parallel computing is also being used to keep simulation time within useful limits. This function would allow to develop scenarios with a sufficient number of iterations.

**Conclusions**

- Serious problem with the data call and data format. Problems with the data call and data format were major than expected. This made progress go slower or even prevented the case study to move on (Northern stock). This is an important point to think for future so detailed data calls in new data formats.
- Experience gained with the data format allowing to a better understanding of some of the variables requested. Main examples are fields "*onShoreSampLoc*" and "*atSeaSampLoc*", which are essential to disaggregate the matrix by sampling country (instead of "*vslFlgCtry*").
- One of the main problems encountered was the slowness of the functions needed to select a sample with a multistage stratified sampling design (mstage) and to apply poststratification to our estimates. To speed up this process, a custom function is being developed based on R "data.table" package and parallel computing is being used. This became necessary because some of the final matrices used in simulations have more than 500 000 lines.
- Poststrata considered in these exercises are based in the combination of metier, area and quarter, which is how the data are currently submitted to ICES. This results in a large number of poststrata (56) which sometimes are not covered with our sampling. This fact is not reflected in the results because of the large number of iterations. However, when just one sampling is done, there is a high probability that the strata with less harbor*days are not sampled. It would be interesting to somehow reflect this in future results.
- Future developments needed. Ongoing work to finalize the Southern stock scenarios and then plan feasible analysis for Northern stock.

**Table 1.** Number of market events and trips sampled by National stratum and Institute, 2014.

| MS | Institute | Program | National stratum name | Total number of market events | Totals number of trips sampled |
|---|---|---|---|---|---|
| ESP | IEO | Observer on-shore | IEO_BACA_APN | 13 | 13 |
| ESP | IEO | Observer on-shore | IEO_BACA_CN | 128 | 191 |
| ESP | IEO | Observer at-sea | IEO_BACA_GC | - | 48 |
| ESP | IEO | Observer on-shore | IEO_BETA_CN | 113 | 123 |
| ESP | AZTI | Observer on-shore | AZTI_CERCO_CN | 255 | 664 |
| ESP | IEO | Observer on-shore | IEO_CERCO_CN | 286 | 435 |
| ESP | IEO | Observer at-sea | IEO_CERCO_GC | - | 62 |
| ESP | IEO | Observer on-shore | IEO_JURELERA_CN | 52 | 54 |
| ESP | AZTI | Observer on-shore | AZTI_LIN_CABALLA | 36 | 106 |
| ESP | IEO | Observer on-shore | IEO_LIN_CABALLA | 17 | 22 |
| ESP | IEO | Observer on-shore | IEO_NASAPULP_CN | 55 | 64 |
| ESP | IEO | Observer on-shore | IEO_PALANGRE_CN | 79 | 100 |
| ESP | AZTI | Observer on-shore | AZTI_PAREJA_CN | 20 | 20 |
| ESP | IEO | Observer on-shore | IEO_PAREJA_CN | 83 | 95 |
| ESP | IEO | Observer on-shore | IEO_RASCO_CN | 57 | 60 |
| ESP | IEO | Observer on-shore | IEO_SABLE_GC | 24 | 24 |
| ESP | IEO | Observer on-shore | IEO_TRASMALL_CN | 17 | 17 |
| ESP | IEO | Observer on-shore | IEO_VOLANTA_CN | 44 | 56 |
| ESP | IEO | Observer on-shore | IEO_VORACERA_GC | 20 | 20 |
| ESP | AZTI | Observer on-shore | AZTI_ART | 126 | 208 |
| PRT | IPMA | Observer on-shore | DRB_DRH | 0 | 0 |
| PRT | IPMA | Observer on-shore | FPO_MOL | 101 | 202 |
| PRT | IPMA | Observer on-shore | GNS_GTR | 330 | 659 |
| PRT | IPMA | Observer on-shore | LLD_LPF | 36 | 36 |
| PRT | IPMA | Observer on-shore | LLS_DWS | 47 | 47 |
| PRT | IPMA | Observer on-shore | OTB_CRU | 60 | 99 |
| PRT | IPMA | Observer on-shore | OTB_DEF | 150 | 150 |
| PRT | IPMA | Observer on-shore | OTHER | 28 | 55 |
| PRT | IPMA | Observer on-shore | OUT_OF_FRAME | 0 | 0 |
| PRT | IPMA | Observer on-shore | PS_SPF | 100 | 141 |
| PRT | IPMA | Observer on-shore | SB | 0 | 0 |
| PRT | IPMA | Observer on-shore | TBB_CRU | 20 | 20 |
| PRT | IPMA | Observer at-sea | DRB_DRH | - | 0 |
| PRT | IPMA | Observer at-sea | FPO_MOL | - | 0 |
| PRT | IPMA | Observer at-sea | GNS_GTR | - | 24 |
| PRT | IPMA | Observer at-sea | LLD_LPF | - | 8 |
| PRT | IPMA | Observer at-sea | LLS_DWS | - | 12 |
| PRT | IPMA | Observer at-sea | OTB_CRU | - | 28 |
| PRT | IPMA | Observer at-sea | OTB_DEF | - | 36 |
| PRT | IPMA | Observer at-sea | OTHER | - | 0 |
| PRT | IPMA | Observer at-sea | OUT_OF_FRAME | - | 0 |
| PRT | IPMA | Observer at-sea | PS_SPF | - | 24 |
| PRT | IPMA | Observer at-sea | SB | - | 0 |
| PRT | IPMA | Observer at-sea | TBB_CRU | - | 12 |

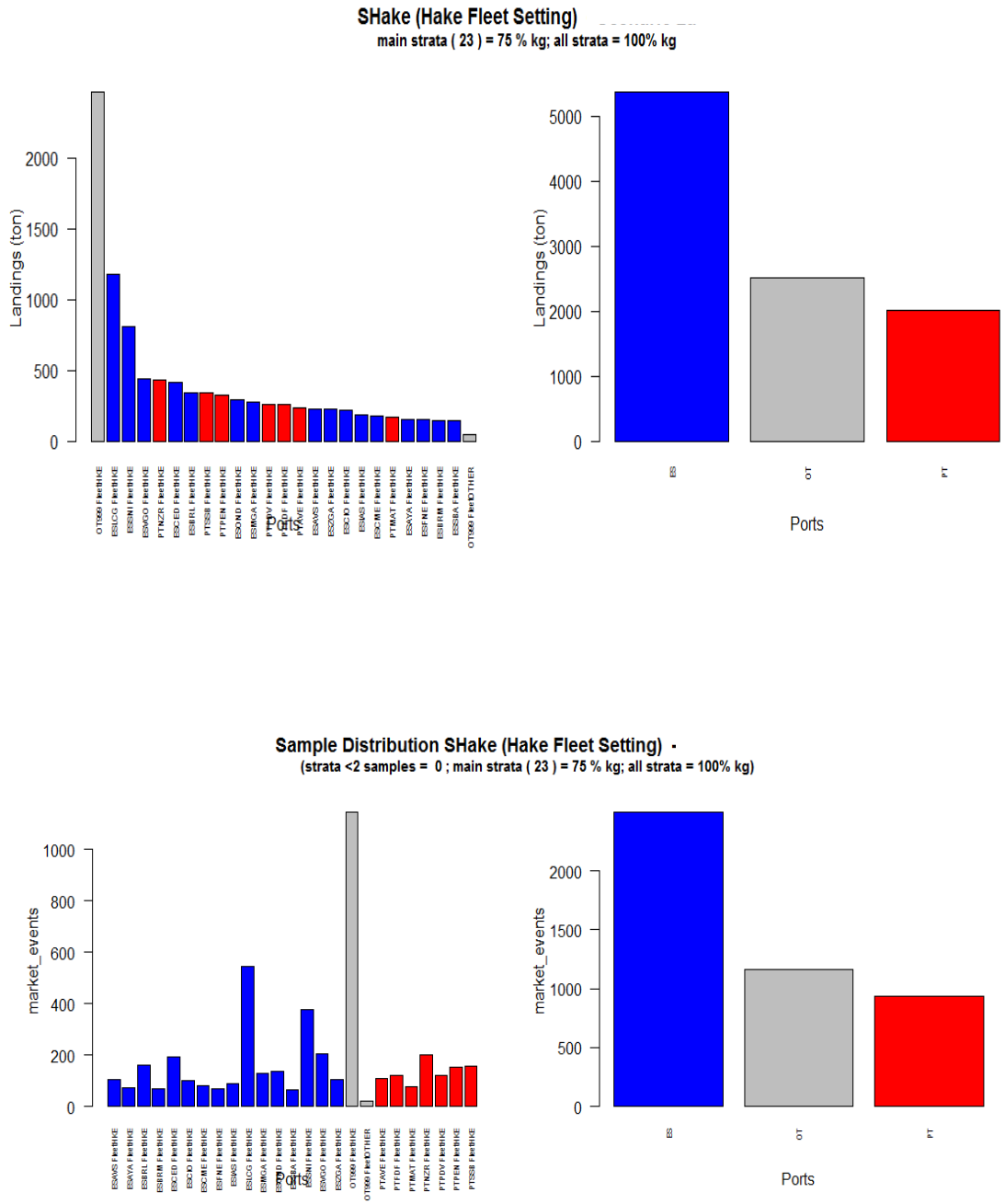


**Figure 1.** Scenario 1b: Regional stratified by port. Southern stock.

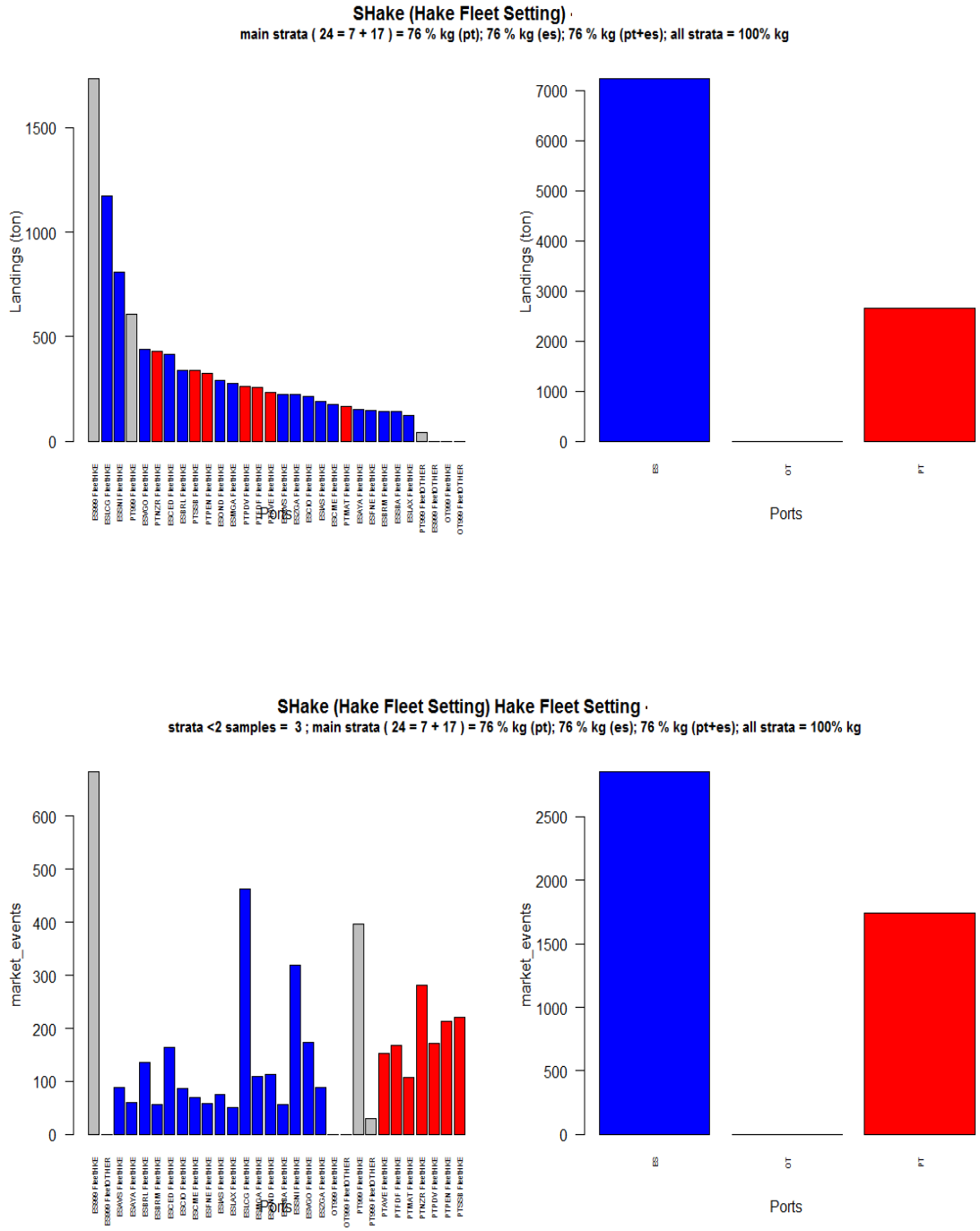


**Figure 2.** Scenario 2b: Stratified by country and port. Southern hake.

**Table 2.** Landing per metier.

| ID | metier | tons |
|---|---|---|
| 1 | PTB_MPD_>=55_0_0 | 1714 |
| 2 | OTB_DEF_>=55_0_0 | 1670 |
| 3 | MIS_MIS_0_0_0 | 1599 |
| 4 | LLS_DEF_0_0_0 | 1598 |
| 5 | GNS_DEF_80-99_0_0 | 1456 |
| 6 | OTB_MCD_>=55_0_0 | 539 |
| 7 | GNS_DEF_60-79_0_0 | 399 |
| 8 | GTR_DEF_60-79_0_0 | 241 |
| 9 | OTB_MPD_>=55_0_0 | 165 |
| 10 | OTB_CRU_>=55_0_0 | 72 |
| 11 | FPO_MOL_0_0_0 | 70 |
| 12 | LHM_DEF_0_0_0 | 31 |
| 13 | GNS_DEF_40-59_0_0 | 18 |
| 14 | GTR_DEF_40-59_0_0 | 14 |
| 15 | GNS_DEF_>=100_0_0 | 6 |
| 16 | PS_SPF_0_0_0 | 4 |
| 17 | TBB_CRU_<55_0_0 | 3 |
| 18 | LLS_DWS_0_0_0 | 2 |
| 19 | TBB_MOL_<55_0_0 | 2 |
| 20 | DRB_MOL_0_0_0 | 0 |
| 21 | SDN_MCF_<55_0_0 | 0 |
| 22 | SB_FIF_0_0_0 | 0 |
| 23 | LHM_DWS_0_0_0 | 0 |
| 24 | LLD_LPF_0_0_0 | 0 |
| 25 | FPO_CRU_0_0_0 | 0 |
| 26 | FPO_FIF_0_0_0 | 0 |
| 27 | LHM_CEP_0_0_0 | 0 |
| 28 | LHM_SPF_0_0_0 | 0 |
| TOTAL | TOTAL | 9607 |

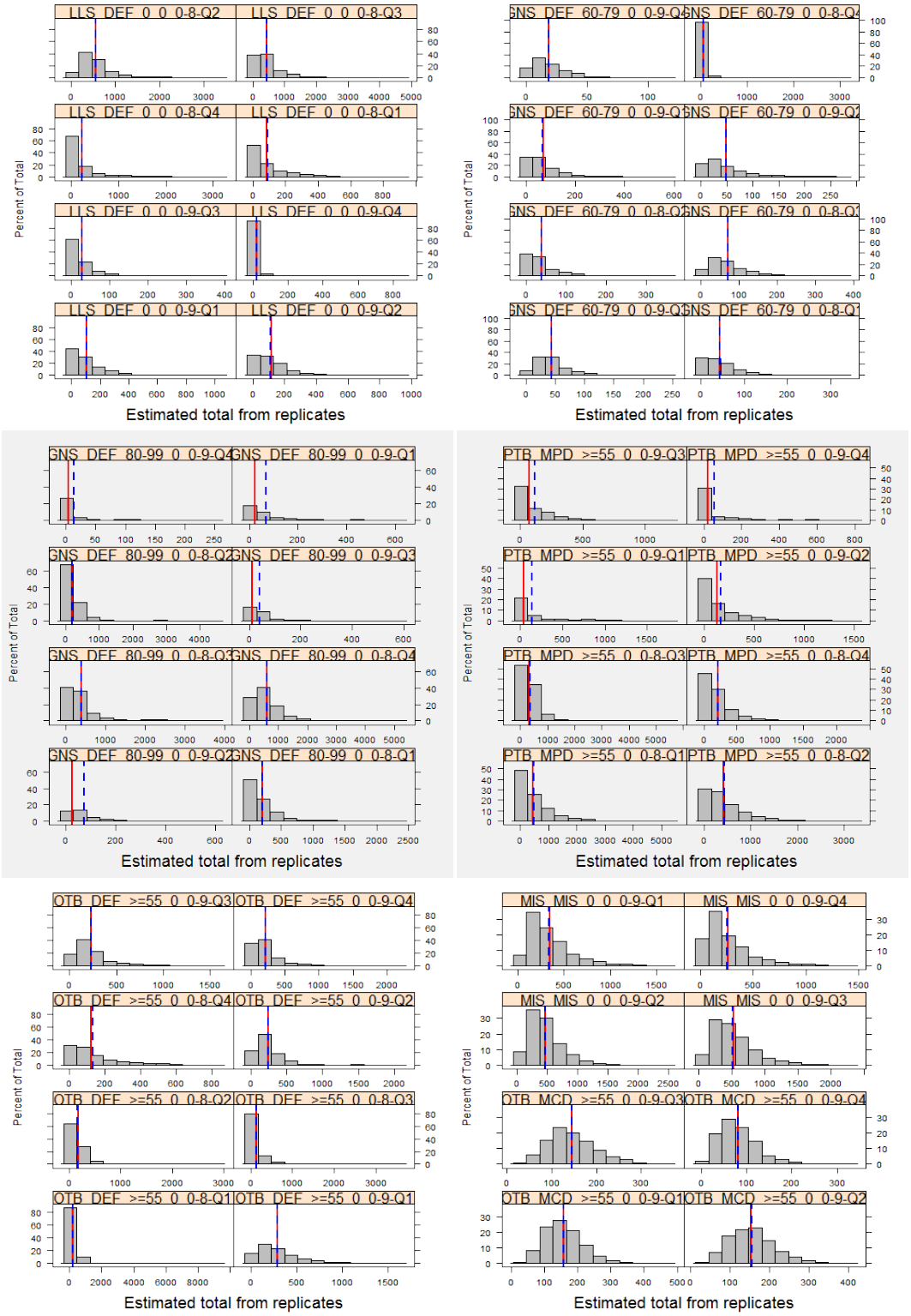


**Figure 3.** SRS of trips with domain estimation.

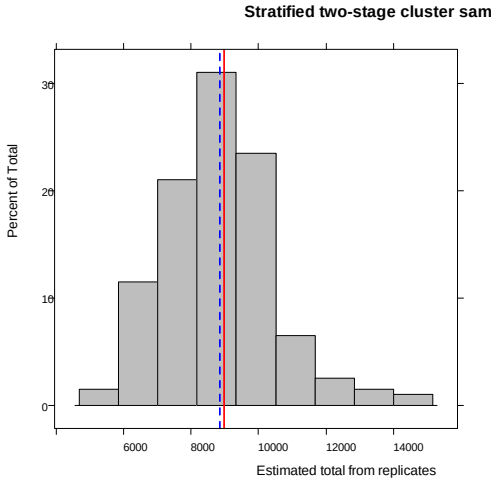


**Figure 4** Estimated total from 200 replicates applying a two-stage sampling design stratified by country

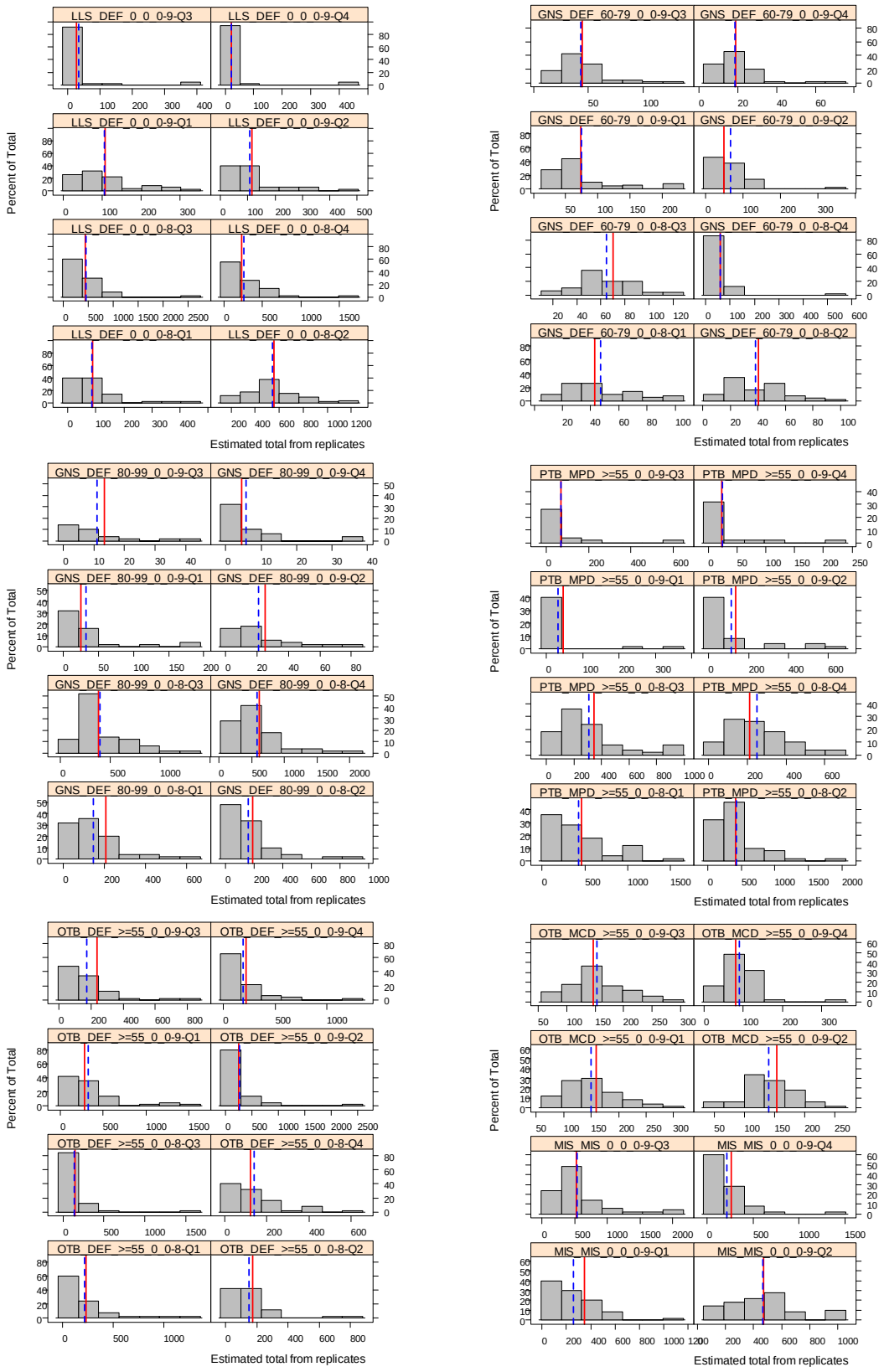


**Figure 5.** Estimated total by domain (metier*quarter*area) calculated with poststratification from 50 replicates applying a two-stage sampling design stratified by country.