SURVEY DESIGN FOR THE ESTIMATION OF THE ECONOMIC PERFORMANCE OF THE ITALIAN FISHERIES SECTOR

Paolo Accadia[1], Federica Piersimoni[2], Dario Pinello[1], Evelina Sabatella[1]

[1] NISEA soc. coop. - Via Irno, 11 – 84135 Salerno, Italy

[2] ISTAT, Rome, Italy

ITACOSM2019

FLORENCE 5-7 June 2019 | 6th ITAlian COnference on Survey Methodology

EU Fisheries Data Collection (DCF)

- Since 2000, an EU framework for the collection and management of fisheries data is in place. This framework was reformed last in 2017 resulting in the EUMAP (European Multi-Annual Plan). Under this framework the Member States (MS) collect, manage and make available a wide range of fisheries data needed for scientific advice.

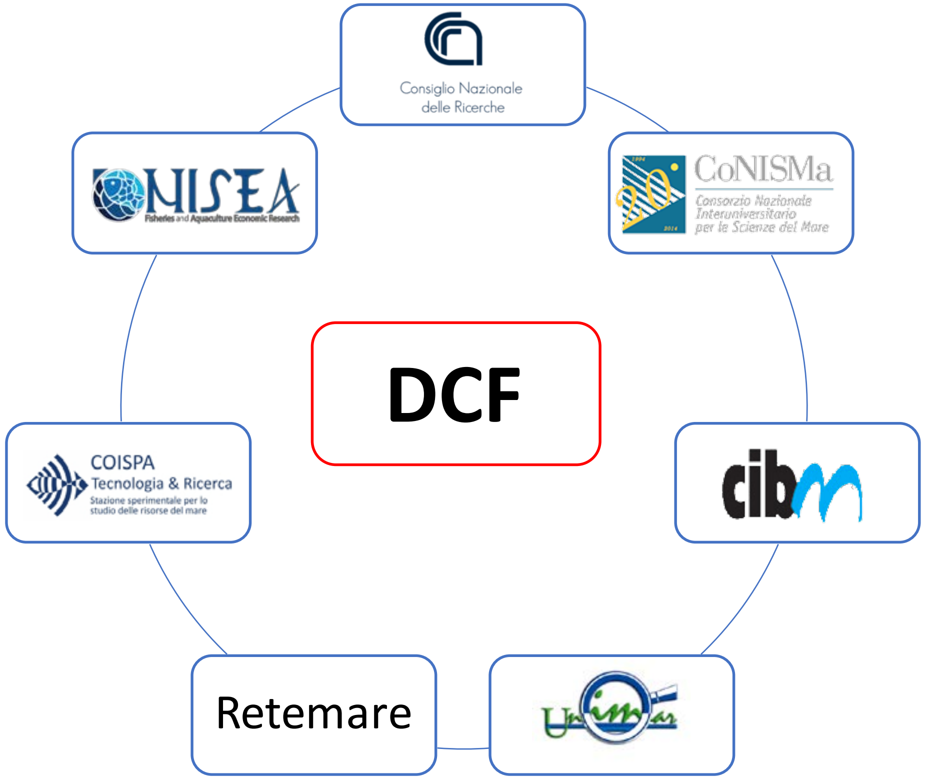
- The data is collected on the basis of National Work Plans in which the MS indicate which data is collected, the resources they allocate for the collection and how data is collected. MS must report annually to the European Commission on the implementation of their Work Plan.

- Part of the data collected by the MS is uploaded in databases managed by the Joint Research Centre (JRC) of the EC in response to data calls issued by DG MARE of the EC. This data forms the basis for scientific advice to inform the Common Fisheries Policy (CFP) decision making process.

- Data collection methods and quality shall be appropriate for the intended purposes defined in Article 25 of Regulation (EU) No 1380/2013 and shall follow the best practices and relevant methodologies advised by the relevant scientific bodies. To this end, the methods and the result of the application of the methods are examined at regular intervals by independent scientific bodies in order to verify their appropriateness with respect to the management of the common fisheries policy (CHAPTER II- EUMAP)

Italian Work Plan for data collection in the fisheries and aquaculture sectors 2017-2019

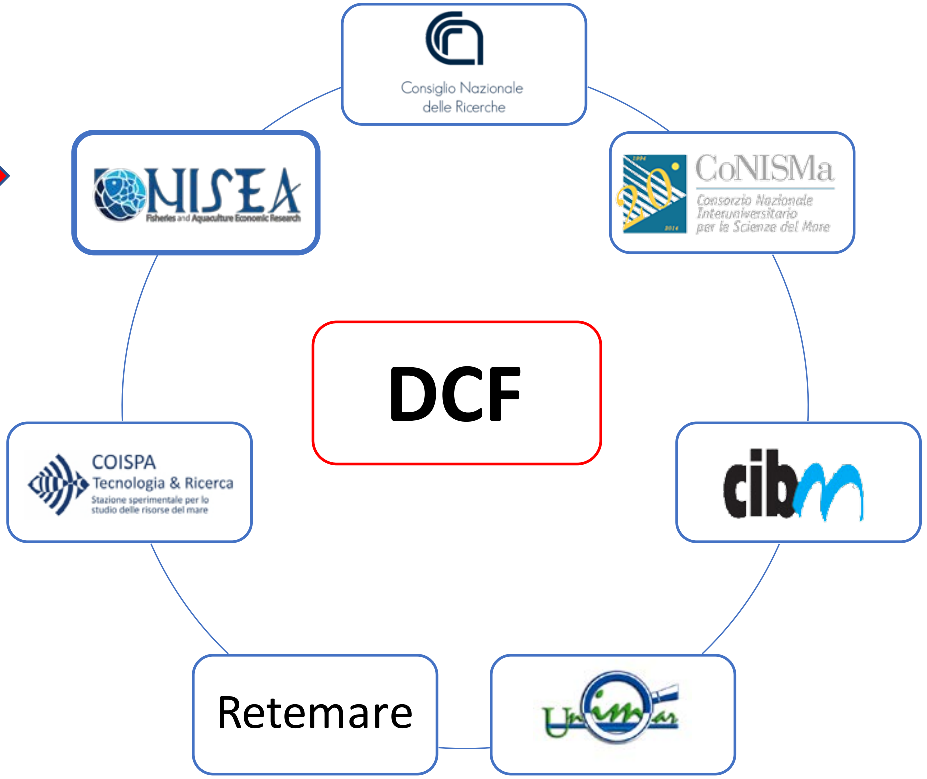
- Economic and social data:
  - Aquaculture
  - Fisheries
  - Processing
- Fishing activity data:
  - Production
  - Fishing effort (days of activity)
- Biological data:
  - Surveys at sea
  - Recreational fishery
- Ecosystem indicators:
  - Spatial distribution of fishing effort
  - By-catch of vulnerable species


DCF
Consiglio Nazionale delle Ricerche
CoNISMa
Consorzio Nazionale Interuniversitario per le Scienze del Mare
cibm
Unimar
Retemare
COISPA Tecnologia & Ricerca
Stazione sperimentale per lo studio delle risorse del mare
NISEA
Fisheries and Aquaculture Economic Research

mipaaft
ministero delle politiche agricole alimentari, forestali e del turismo

Italian Work Plan for data collection in the fisheries and aquaculture sectors 2017-2019

- Economic and social data:
  - Aquaculture
  - Fisheries
  - Processing
- Fishing activity data:
  - Production
  - Fishing effort (days of activity)
- Biological data:
  - Surveys at sea
  - Recreational fishery
- Ecosystem indicators:
  - Spatial distribution of fishing effort
  - By-catch of vulnerable species


Consiglio Nazionale delle Ricerche
NISEA Fisheries and Aquaculture Economic Research
CoNISMa Consorzio Nazionale Interuniversitario per le Scienze del Mare
DCF
COISPA Tecnologia & Ricerca Stazione sperimentale per lo studio delle risorse del mare
cibm
Retemare
Unimar

mipaaft
ministero delle politiche agricole
alimentari, forestali e del turismo

Economic and Social variables

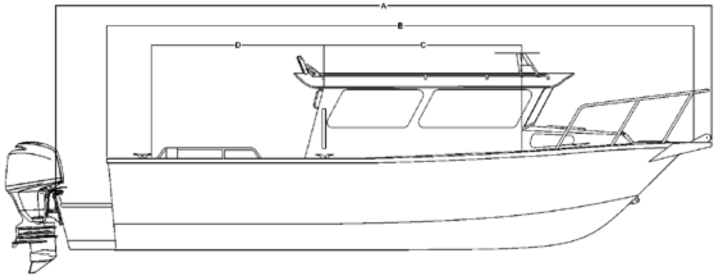
Different methodologies and data sources are used to collect data on the economic and social variables.

A probability sample survey is implemented to estimate the following variables:

- Income (Income from leasing out quota, Other income)
- Production in volume and value
- Fishing effort (days of activity)
- Labour costs (Personnel costs, Value of unpaid labour)
- Energy costs and energy consumption
- Repair and maintenance costs
- Other operating costs (variable costs, non-variable costs, Lease/rental payments for quota or other fishing rights)
- Value of quota and other fishing rights
- Investments in tangible assets
- Employment (engaged crew, unpaid labour, total hours per year).

The sample survey design

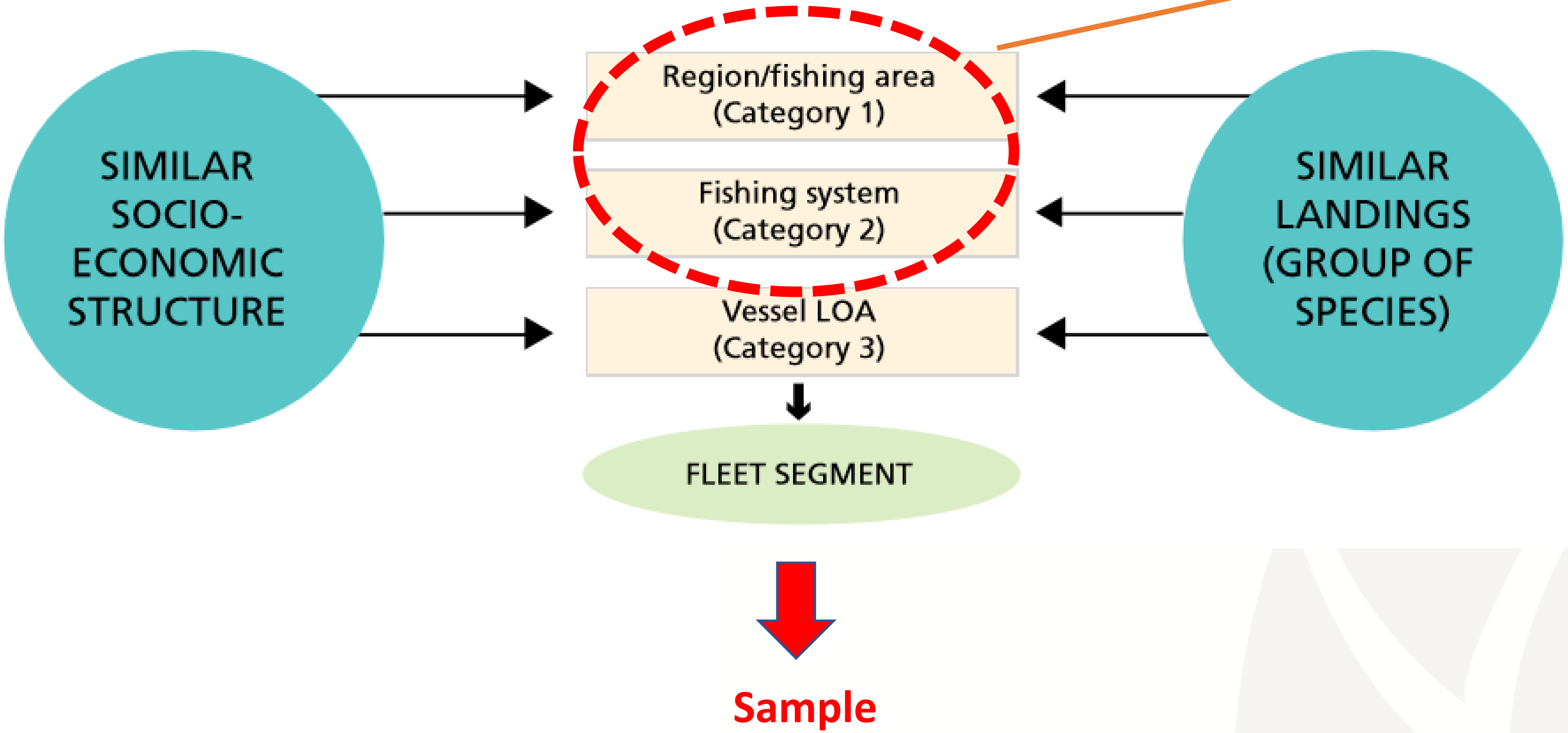
- The population consists of all active and inactive Italian vessels registered in the Union Fishing Fleet Register. The sampling unit is the fishing vessel.
- The fleet is segmented into homogeneous groups of vessels, according to the prevalent fishing technique and the vessel length, as requested by the EUMAP



- A further segmentation of the fleet based on a geographical criterion (combinations of administrative region and FAO geographical sub-area (GSA)) is used in the sampling design to improve the representativeness of the sample and provide a more detailed information at local level.
- The sample size and its allocation among strata are estimated by using the software MAUSS-R, Multivariate Allocation of Units in Sampling Surveys, developed by ISTAT (Italian National Institute of Statistics).
- A probability proportional to size (PPS) sample is extracted from the fleet. The probabilities of inclusion are proportional to the length over all (LOA) of the vessels.

Segmentation of the target population

- Fleet segmentation and sampling



Region/fishing area (Category 1)
Fishing system (Category 2)
Vessel LOA (Category 3)
SIMILAR SOCIO-ECONOMIC STRUCTURE
SIMILAR LANDINGS (GROUP OF SPECIES)
FLEET SEGMENT
Sample

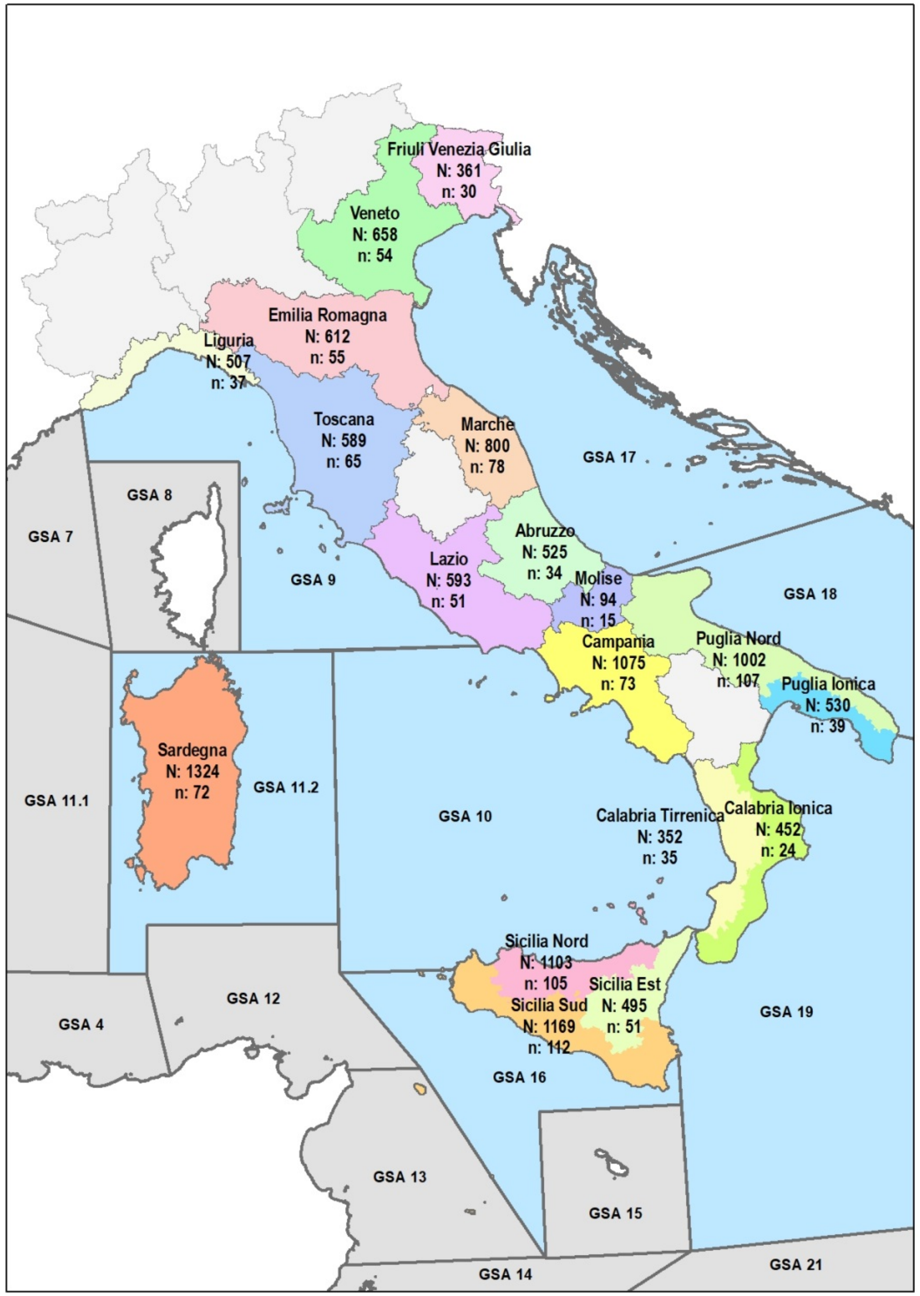
- **Fleet register**
- **Logbook**
- **VMS**
- **AIS**
- **Scientists/experts & data collectors**

Fleet Segmentation

Main fleet segmentation criteria:

Region	GSA	Fishing Technique	Length class
Abruzzo	9 (N. Tyrrhenian)	DRB	VL0006
Calabria	10 (S. Tyrrhenian)	DTS	VL0612
Campania	11 (Sardinia)	HOK	VL1218
E. Romagna	16 (S. Sicily)	PGP	VL1824
F. Venezia Giulia	17 (N. Adriatic)	PS	VL2440
Lazio	18 (S. Adriatic)	TBB	
Liguria	19 (Ionian)	TM	
Marche			
Molise			
Apulia			
Sardenia			
Sicily			
Tuscany			
Veneto			

Further segmentation is given by the 19 clams consortia for the dredges (DRB) and 3 special fisheries for the trawlers (DTS).



2017 survey: strata and sample

- Combining the segmentation criteria produced a total of 155 strata in the 2017 survey.
- Strata are characterised by a strong variability in the population size, from 2 to almost 800 vessels.
- A total of 1035 sample units were selected for the 155 strata, with a coverage by 8.5%.

Number of strata by Region-GSA and fishing technique

Region-GSA	**DRB**	**DTS**	**HOK**	**PGP**	**PS**	**TBB**	**TM**	**Total**
Abruzzo	2	3		2	1			8
Calabria Ionian		2		3				5
Calabria Tyrrhenian		2		3	2			7
Campania	1	2		3	2			8
E. Romagna	2	3		3		1	2	11
F. Venezia Giulia	2	1		2	1	1	1	8
Lazio	1	3		3	1			8
Liguria		2		3	2			7
Marche	4	4		2		1	1	12
Molise	1	3		2				6
Apulia Ionian		2		3	1			6
Apulia Adriatic	2	4	1	2	1		1	11
Sardinia		3		3	1			7
Sicily Ionian		2	2	3	1			8
Sicily Tyrrhenian		2	1	3	2			8
Sicily South		6	2	3	2		1	14
Tuscany		3		3	2			8
Veneto	4	3		3		1	2	13
Total	**19**	**50**	**6**	**49**	**19**	**4**	**8**	**155**

Methodological problems

The main methodological issues affecting the implementation of the survey are:

1. a very high number of strata requested by the European regulation for the provisions of final estimates;
2. the need to calibrate the estimates to exogenous variables (production and effort);
3. due to non-responses many strata have to be collapsed;
4. due to the low number of units, some strata must be censused and therefore n increases.

Compromise allocation as optimization problem

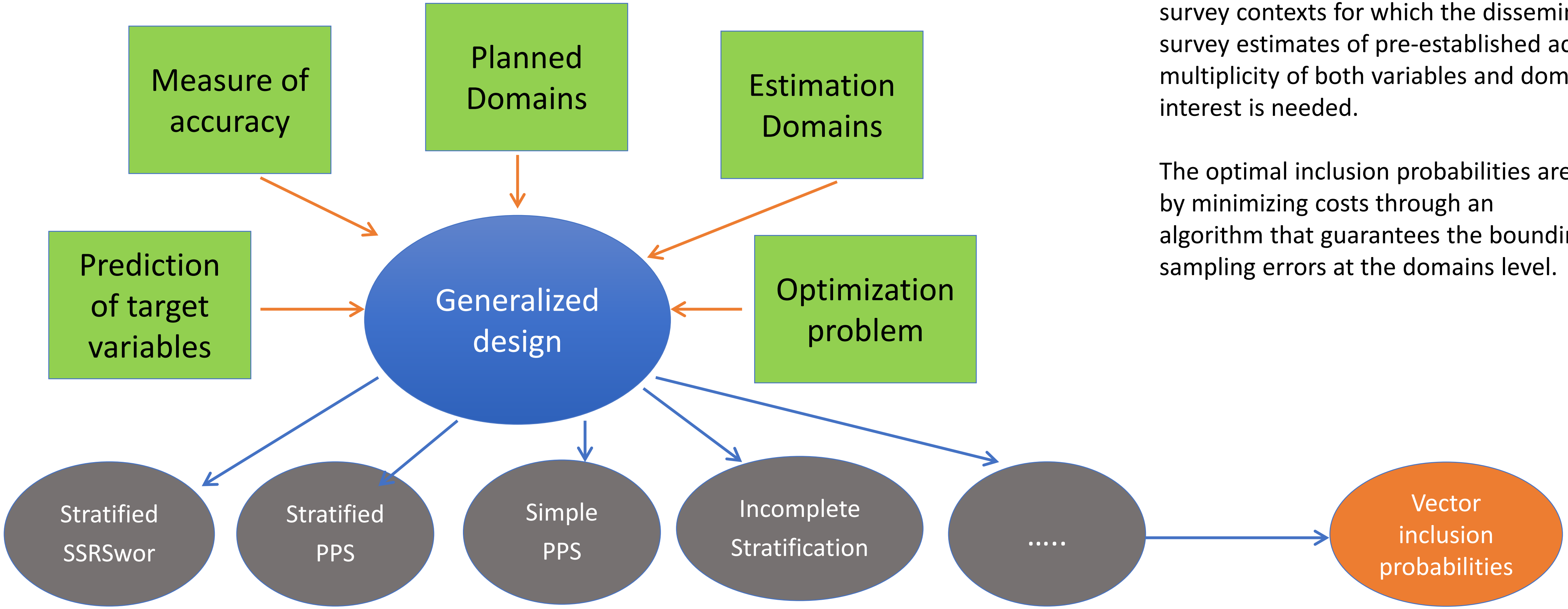
More than one variable and estimation domain

Precision constrained optimal allocation: it minimises the cost subject to the variance constraints

For stratified sampling

$$\begin{cases} \sum_h c_h \, n_h = \min \\ \sum_{h \in d} N_h^2 \sigma_{hr}^2 / n_h - V_{0rd} \le V_{rd}^* \quad r = 1,\ldots,R; d = 1,\ldots,D \\ 1 \le n_h \le N_h \end{cases}$$

Elements which characterize the general design



The method (Falorsi and Righi, 2015) derives the optimal inclusion probabilities for many survey contexts for which the dissemination of survey estimates of pre-established accuracy for a multiplicity of both variables and domains of interest is needed.

The optimal inclusion probabilities are obtained by minimizing costs through an algorithm that guarantees the bounding of sampling errors at the domains level.

Planned and estimation domains

An ***estimation domain*** U_d

is generic sub-population of *U* with N_d elements, for which separate estimates must be calculated.

Example for the economic performance of the Italian fishery sector survey, the estimates must be calculated separately considering two domain types:

- ✓ ***Fishing technique***(7 modalities: DRB , DTS,…)
- ✓ ***Length class*** (5 modalities: VL006, VL0612,….)

Fishing technique ***Length class***	DRB	DTS	…..	**Estimation Domain**
VL006				1
VL0612				2
……				
Estimation Domain	4	5		

The Estimation domains can overlap

The subpopulation in which the estimation domains 1 and 4 overlap

Planned and estimation domains

A ***planned domain*** **PD** $\boldsymbol{U_h}$

is generic sub-population of *U* with N_h elements, for which the sampling size is fixed at the design stage.

$$\sum_{k \in U} \pi_k \, \delta_{hk} = n_h = \text{fixed}$$

the planned domains can overlap; therefore, the unit *k* may have more than one value $\delta_{hk} = 1$

We focus on the designs for which $\boldsymbol{\pi}$ is such that
$\sum_{k \in U} \pi_k \, \boldsymbol{\delta}_k = \mathbf{n}$.

where $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{hk}, \ldots, \delta_{Hk})'$ and $\mathbf{n} = (n_1, \ldots, n_h, \ldots, n_H)'$ is a vector of integers

Planned and estimation domains

The estimation domains are defined as an aggregation of complete planned domains

The example from the **economic performance of the Italian fishery sector survey**

CASE 1: The Estimation domains coincide with planned domains

The PD **overlap**

Fishing technique ***Length class***	DRB	DTS	…..	**Estimation Domain**
VL006				1
VL0612				2
……				
Estimation Domain	4	5		

CASE 2: The Planned domains are defined by cross-classification of the estimation domains

The PD **do not overlap**

Planned and estimation domains

An example for the economic performance of the Italian fishery sector survey

Estimation domains:

- *region* (with 20 modalities),
- *Fishing technique*(2 modalities: DRB, DTS)
- and *Length class* (3 modalities: VL006, VL0612, VL1218)

There are *D*=20+3+2=25 possible overlapping estimation domains.

Option 1. The single PD is identified by the intersection of the categories of the estimation domains.

⇨ $H = 20 \times 2 \times 3 = 120$ with $\sum_h \delta_{hk} = 1$

Option 2. The PD coincide with the estimation domains.

⇨ $H = D = 25$ with $\sum_h \delta_{hk} = 3$

Option 3. The PD U_h are defined as (*i*) **region by fishing technique** and (*ii*) **fishing technique by length class**.

⇨ $H = (20 \times 2) + (2 \times 3) = 46$ with $\sum_h \delta_{hk} = 2$.

Options….

Sample allocation problem

Sample allocation problem (or definition of π-values) can be formalized by the following general optimization problem

$$\begin{cases} Min\left(\sum_{k\in U}\pi_k\, c_k\right) \\ V(\hat{t}_{(dr)}) \leq \overline{V}_{(dr)} \quad (d=1,\dots,D;\, r=1,\dots,R) \\ 0 < \pi_k \leq 1 \qquad\qquad (k=1,\dots,N) \end{cases}$$

Prior to sampling, the y_{rk} values are not known and the variance expressed in the formula cannot be used for planning the sampling precision at the design phase.

In practice, it is necessary to either obtain some proxy values or predict the y_{rk} values based on superpopulation models that exploit auxiliary information.

Prediction of target variables

Under a model-based inference, the y_{rk} values are assumed to be the realization of a superpopulation model *M*.

The model we study has the following form:

$$\begin{cases} y_{rk} = f_r(\mathbf{x}_k;\boldsymbol{\beta}_r) + u_{rk} \\ E_M(u_{rk}) = 0 \quad \forall k;\ E_M(u_{rk}^2) = \sigma_{rk}^2;\quad E_M(u_{rk}, u_{rl}) = 0\ \forall k \neq l \end{cases} ,$$

$\mathbf{x}_k$ is a vector of predictors (available in the sampling frame), $\boldsymbol{\beta}_r$ is a vector of regression coefficients and

$f_r(\mathbf{x}_k;\boldsymbol{\beta}_r) = \tilde{y}_{rk}$ → **predictors** are double logarithmic functions,

u_{rk} is the error term and $E_M(\cdot)$ denotes the expectation under the model.

The parameters $\boldsymbol{\beta}_r$ and the variances σ_{rk}^2 are assumed to be known, although in practice they are usually estimated.

Prediction of target variables

$$\begin{cases} log\, y_{rk} = \beta_0 + \beta_{hr}\, log\, x_k + u_{rk} & \forall k \in U_h \\ E_M(u_{rk)=}0 & \forall k \\ E_M(u_{rk}^2) = \sigma_{hr}^2; \qquad E_M(u_{rk}, u_{rl}) = 0 & \forall k \neq l \end{cases}$$

The measure of accuracy is the Anticipated variance

$$\boxed{AV(\hat{t}_{dr}) = E_M E_p(\hat{t}_{dr} - t_{dr})^2}$$

$$\boxed{= E_p[V_M(\hat{t}_{dr})] + V_p[E_M(\hat{t}_{dr})] - V_M(t_{dr})}$$

$$\boxed{= V_{Pois}(\text{predictions}) + V_{Pois}(\text{model variance}) + R}$$

List of the operational steps

1. **PREPARE THE INPUTS FOR THE ALGORITHM**
   - **DEFINE THE ED**
   - **DEFINE THE PD**
   - **DEFINE** $\tilde{y}_{rk}$ **,** σ^2_{rk} **,** c_k **,** $\overline{V}_{(dr)}$

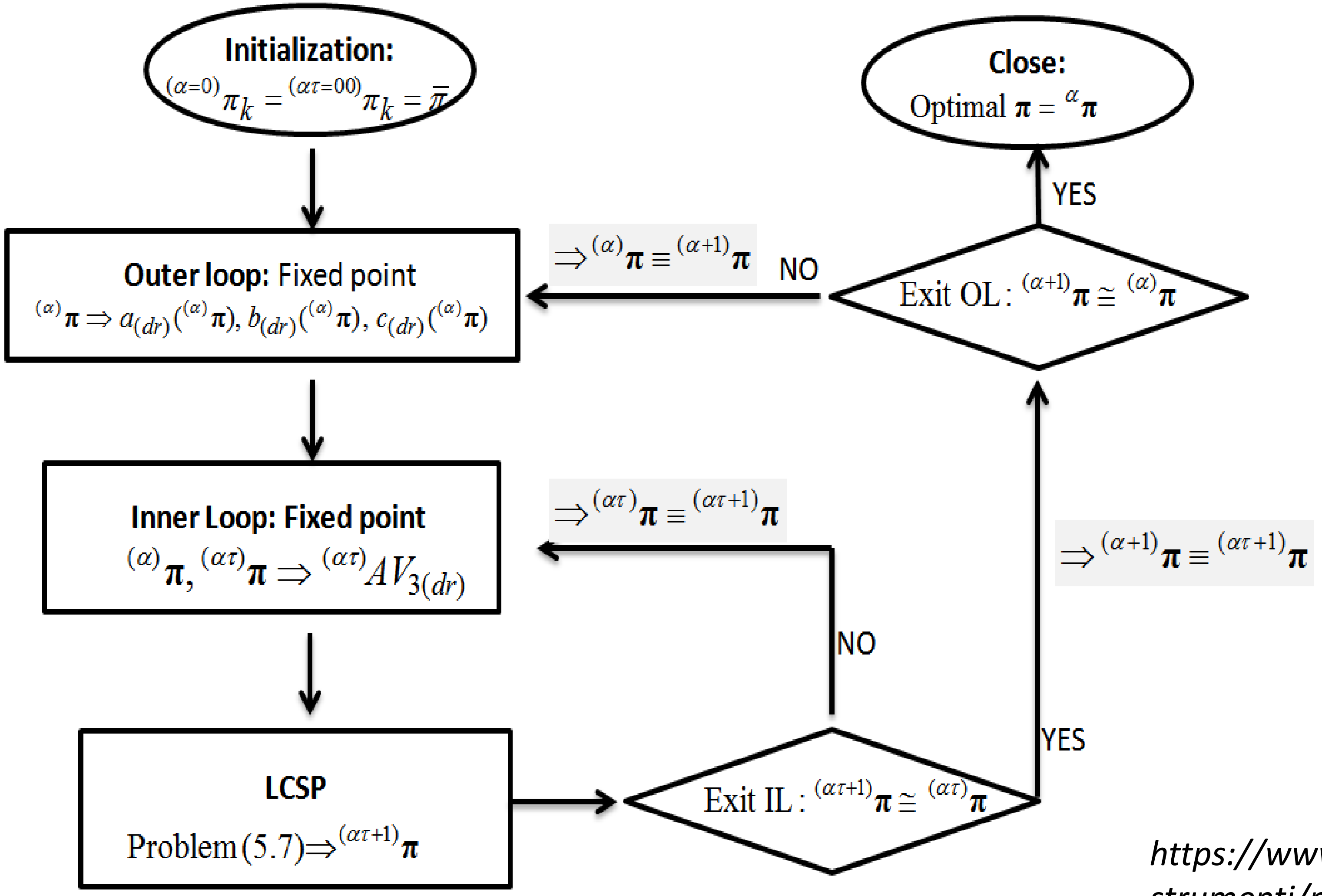
2. **RUN THE ALGORITHM**

3. **CALIBRATE THE RESULTING INCLUSION PROBABILITIES FOR ENSURING THE EXPECTED SAMPLE SIZES FOR EACH ED ARE INTEGERS: NOT DONE/FUTURE STEP**

4. **SELECT THE SAMPLE either with cube algorithm or with traditional methods (e.g. R….):L NOT DONE/FUTURE STEP**

- **The algorithm defines the optimal inclusion probabilities in various survey contexts, which are characterized by the need to disseminate survey estimates of prefixed accuracy, for many variables and domains of interest.**
- **suitable for a general multi-way sampling design**
- **the algorithm and the final computation are domain- and variable-driven**

Algorithm flowchart



Initialization: ${}^{(\alpha=0)}\pi_k = {}^{(\alpha\tau=00)}\pi_k = \bar{\pi}$

Outer loop: Fixed point ${}^{(\alpha)}\boldsymbol{\pi} \Rightarrow a_{(dr)}({}^{(\alpha)}\boldsymbol{\pi}), b_{(dr)}({}^{(\alpha)}\boldsymbol{\pi}), c_{(dr)}({}^{(\alpha)}\boldsymbol{\pi})$

Inner Loop: Fixed point ${}^{(\alpha)}\boldsymbol{\pi}, {}^{(\alpha\tau)}\boldsymbol{\pi} \Rightarrow {}^{(\alpha\tau)}AV_{3(dr)}$

LCSP Problem (5.7) $\Rightarrow {}^{(\alpha\tau+1)}\boldsymbol{\pi}$

Exit IL : ${}^{(\alpha\tau+1)}\boldsymbol{\pi} \cong {}^{(\alpha\tau)}\boldsymbol{\pi}$

NO $\Rightarrow {}^{(\alpha\tau)}\boldsymbol{\pi} \equiv {}^{(\alpha\tau+1)}\boldsymbol{\pi}$

YES $\Rightarrow {}^{(\alpha+1)}\boldsymbol{\pi} \equiv {}^{(\alpha\tau+1)}\boldsymbol{\pi}$

Exit OL : ${}^{(\alpha+1)}\boldsymbol{\pi} \cong {}^{(\alpha)}\boldsymbol{\pi}$

NO $\Rightarrow {}^{(\alpha)}\boldsymbol{\pi} \equiv {}^{(\alpha+1)}\boldsymbol{\pi}$

YES

Close: Optimal $\boldsymbol{\pi} = {}^{\alpha}\boldsymbol{\pi}$

Falorsi P. D., Righi P. (2015)

https://www.istat.it/it/metodi-e-strumenti/metodi-e-strumenti-it/progettazione/strumenti-di-progettazione/multiwaysampleallocation

Application

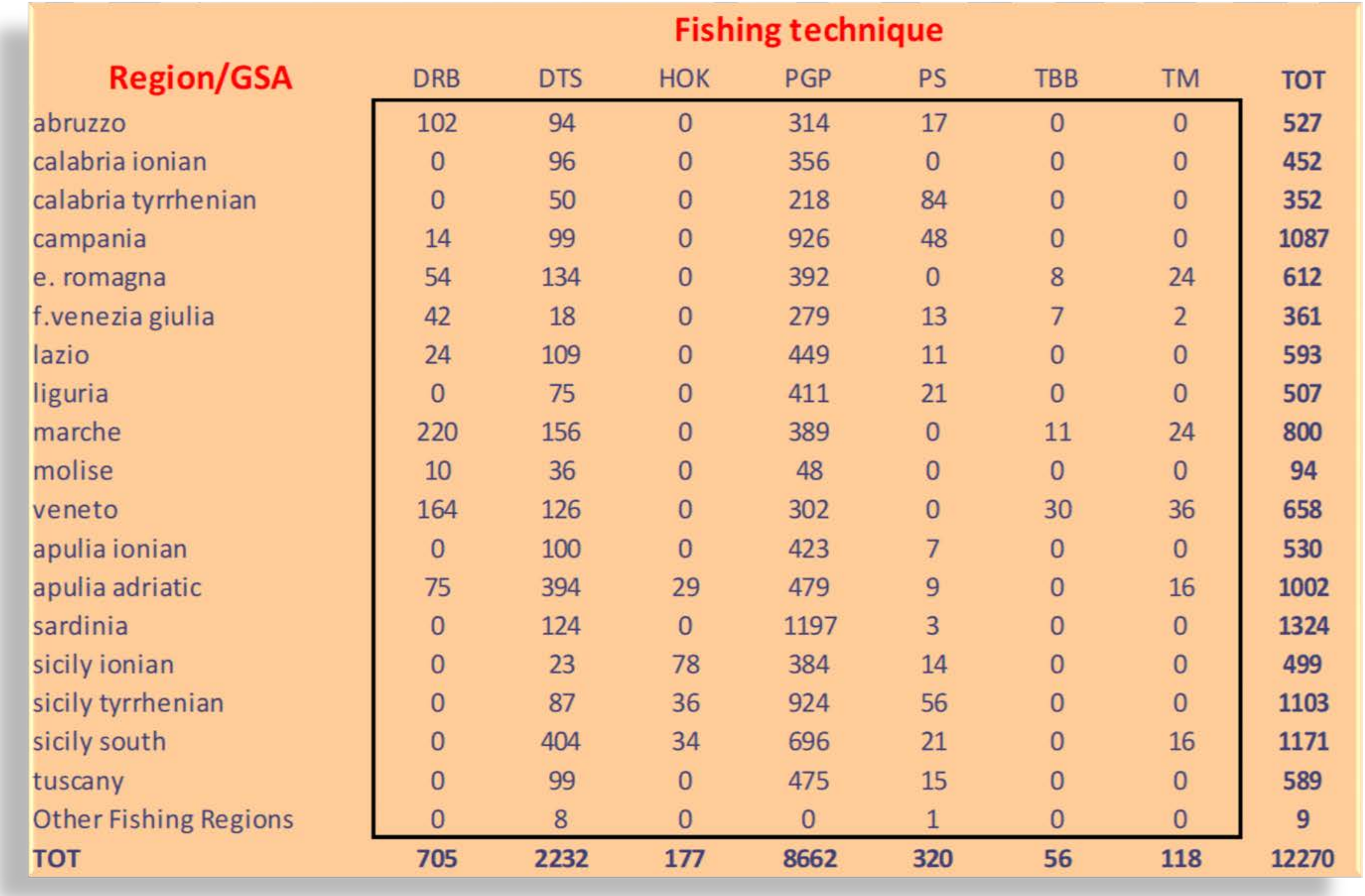
- N= 12.270
- 10 variables:
  - Energy cost
  - Labour cost
  - Repair and maintenance cost
  - Commercial cost incluse in Variable costs in Other operating costs
  - Non-variable cost incluse in Other operating costs
  - Variable cost incluse in Other operating costs
  - Energy consumption
  - Employment
  - Production in value
  - Days at sea (Fishing effort)
- 4 domains:

  1: GSA and groups of regions (8 codes)

  2: Region

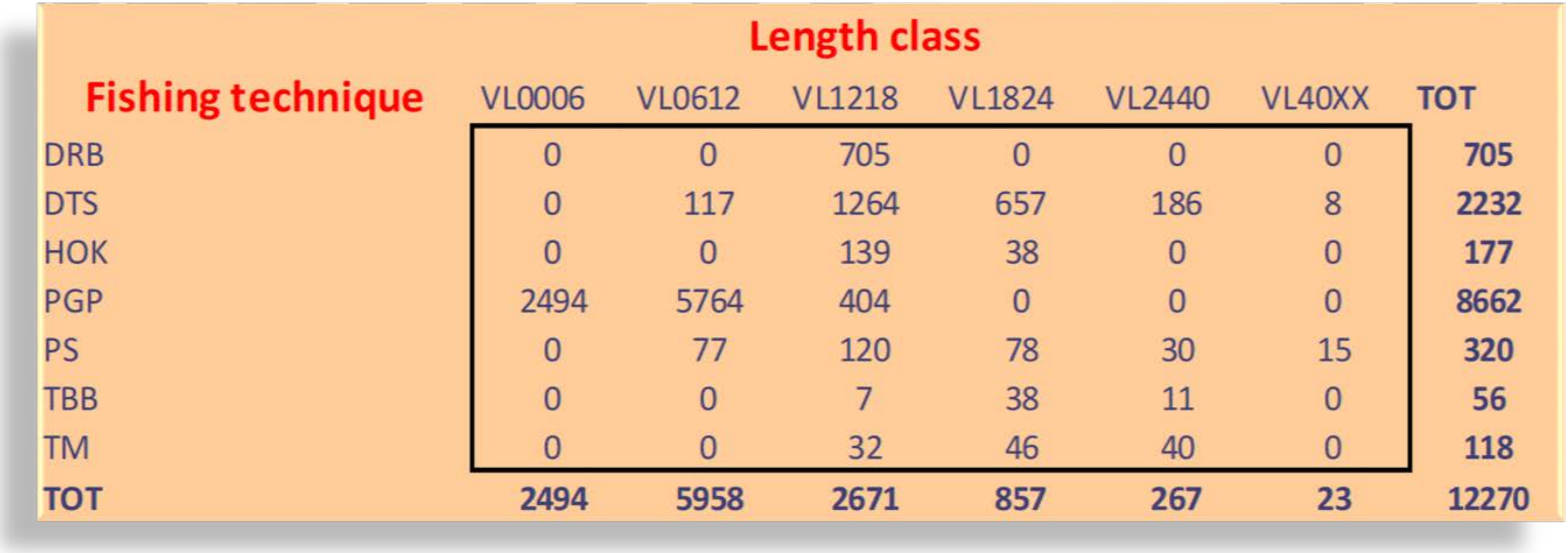
  3: Fishing technique (7 codes)

  4: Length class (6 codes)
- CV = 0.05
- Sample design varying the number of variables (from 2 to 10) and domains (from 2 to 4): all combinations of variables from 2 to 10 and domain codes from 2 to 4 for a total of 27 different sample designs: 10 variables under control for 4 domains
- with almost the same *n*, i.e. costs, it allows to control much more the estimates and the relative CVs
- n<1.000

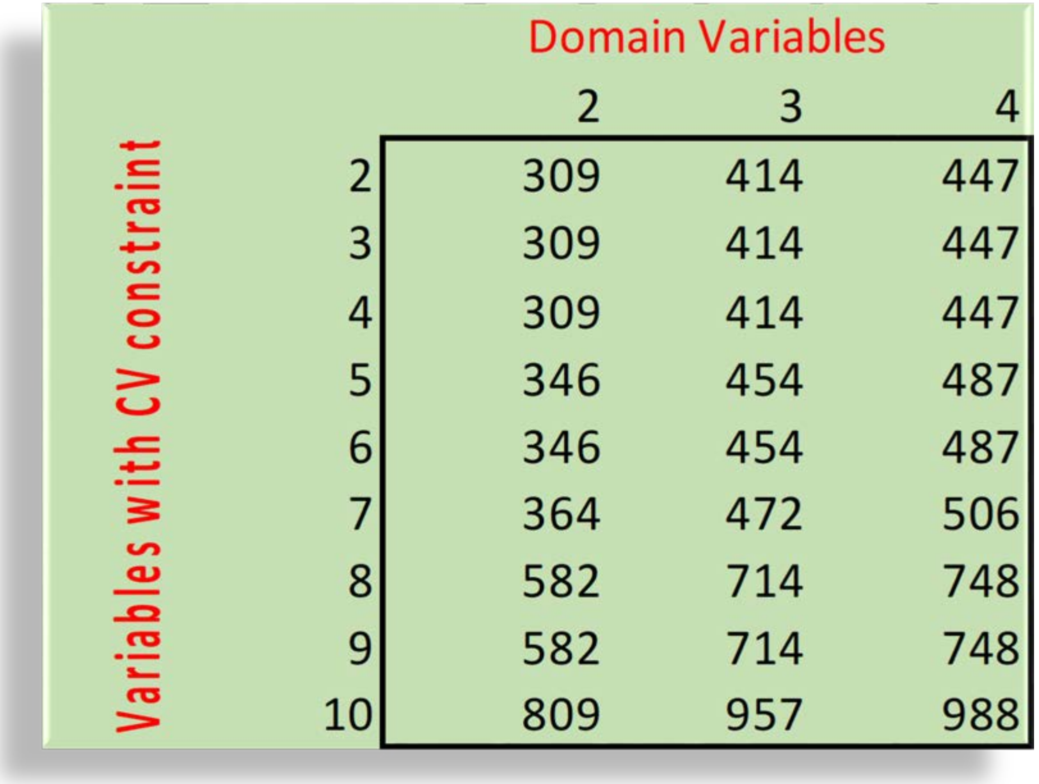


	Fishing technique							
Region/GSA	DRB	DTS	HOK	PGP	PS	TBB	TM	**TOT**
abruzzo	102	94	0	314	17	0	0	**527**
calabria ionian	0	96	0	356	0	0	0	**452**
calabria tyrrhenian	0	50	0	218	84	0	0	**352**
campania	14	99	0	926	48	0	0	**1087**
e. romagna	54	134	0	392	0	8	24	**612**
f.venezia giulia	42	18	0	279	13	7	2	**361**
lazio	24	109	0	449	11	0	0	**593**
liguria	0	75	0	411	21	0	0	**507**
marche	220	156	0	389	0	11	24	**800**
molise	10	36	0	48	0	0	0	**94**
veneto	164	126	0	302	0	30	36	**658**
apulia ionian	0	100	0	423	7	0	0	**530**
apulia adriatic	75	394	29	479	9	0	16	**1002**
sardinia	0	124	0	1197	3	0	0	**1324**
sicily ionian	0	23	78	384	14	0	0	**499**
sicily tyrrhenian	0	87	36	924	56	0	0	**1103**
sicily south	0	404	34	696	21	0	16	**1171**
tuscany	0	99	0	475	15	0	0	**589**
Other Fishing Regions	0	8	0	0	1	0	0	**9**
TOT	**705**	**2232**	**177**	**8662**	**320**	**56**	**118**	**12270**

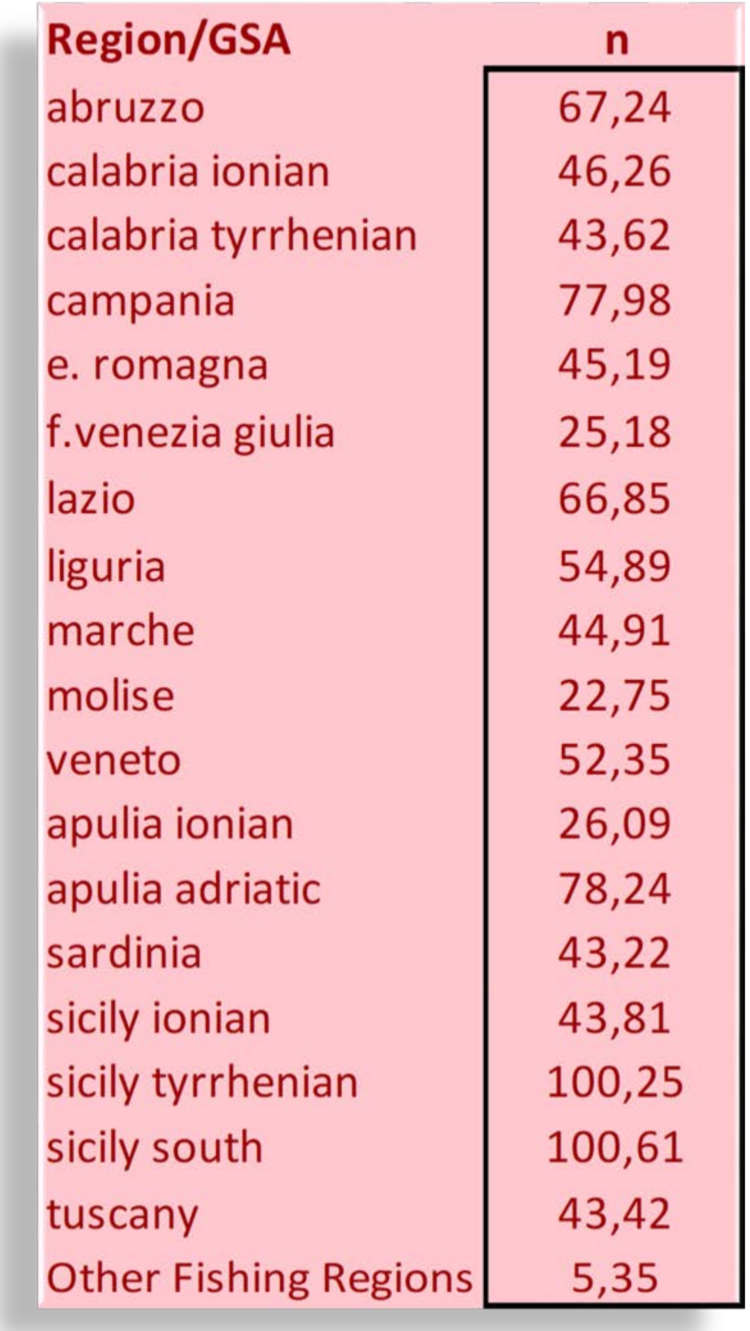
particularly sparse (with many 0s) and therefore it is necessary to stratify on the marginals and not on the intersections of all the strata codes



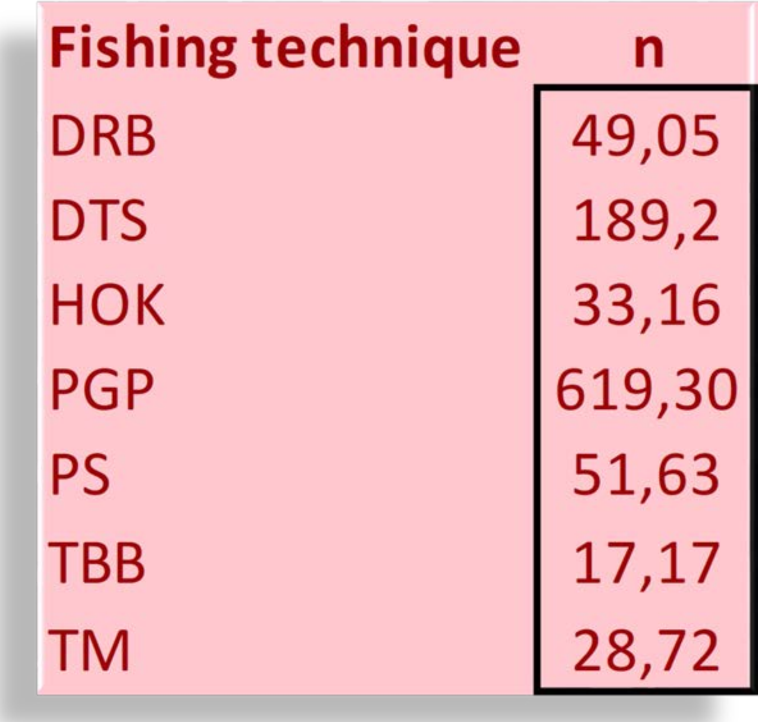
	Length class						
Fishing technique	VL0006	VL0612	VL1218	VL1824	VL2440	VL40XX	TOT
DRB	0	0	705	0	0	0	**705**
DTS	0	117	1264	657	186	8	**2232**
HOK	0	0	139	38	0	0	**177**
PGP	2494	5764	404	0	0	0	**8662**
PS	0	77	120	78	30	15	**320**
TBB	0	0	7	38	11	0	**56**
TM	0	0	32	46	40	0	**118**
TOT	**2494**	**5958**	**2671**	**857**	**267**	**23**	**12270**



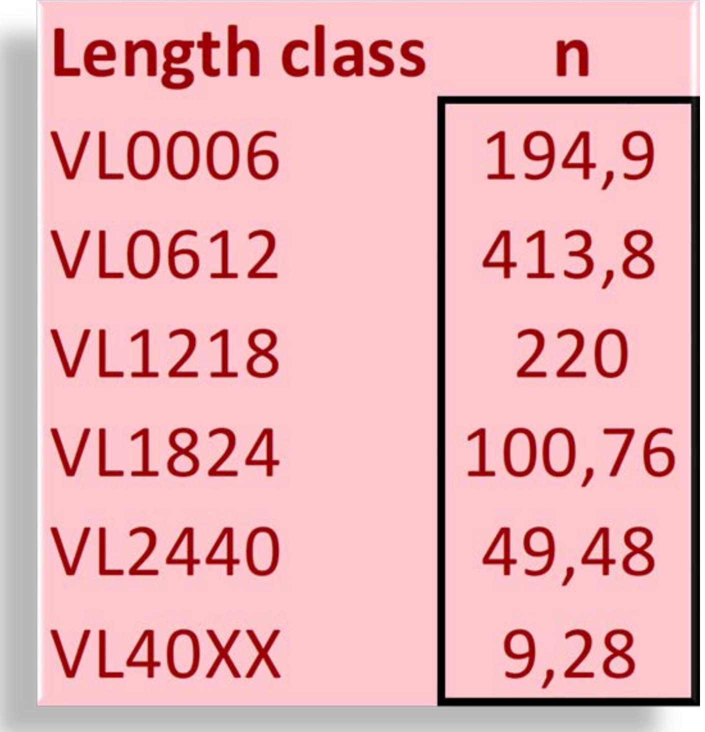
		Domain Variables		
		2	3	4
Variables with CV constraint	2	309	414	447
	3	309	414	447
	4	309	414	447
	5	346	454	487
	6	346	454	487
	7	364	472	506
	8	582	714	748
	9	582	714	748
	10	809	957	988



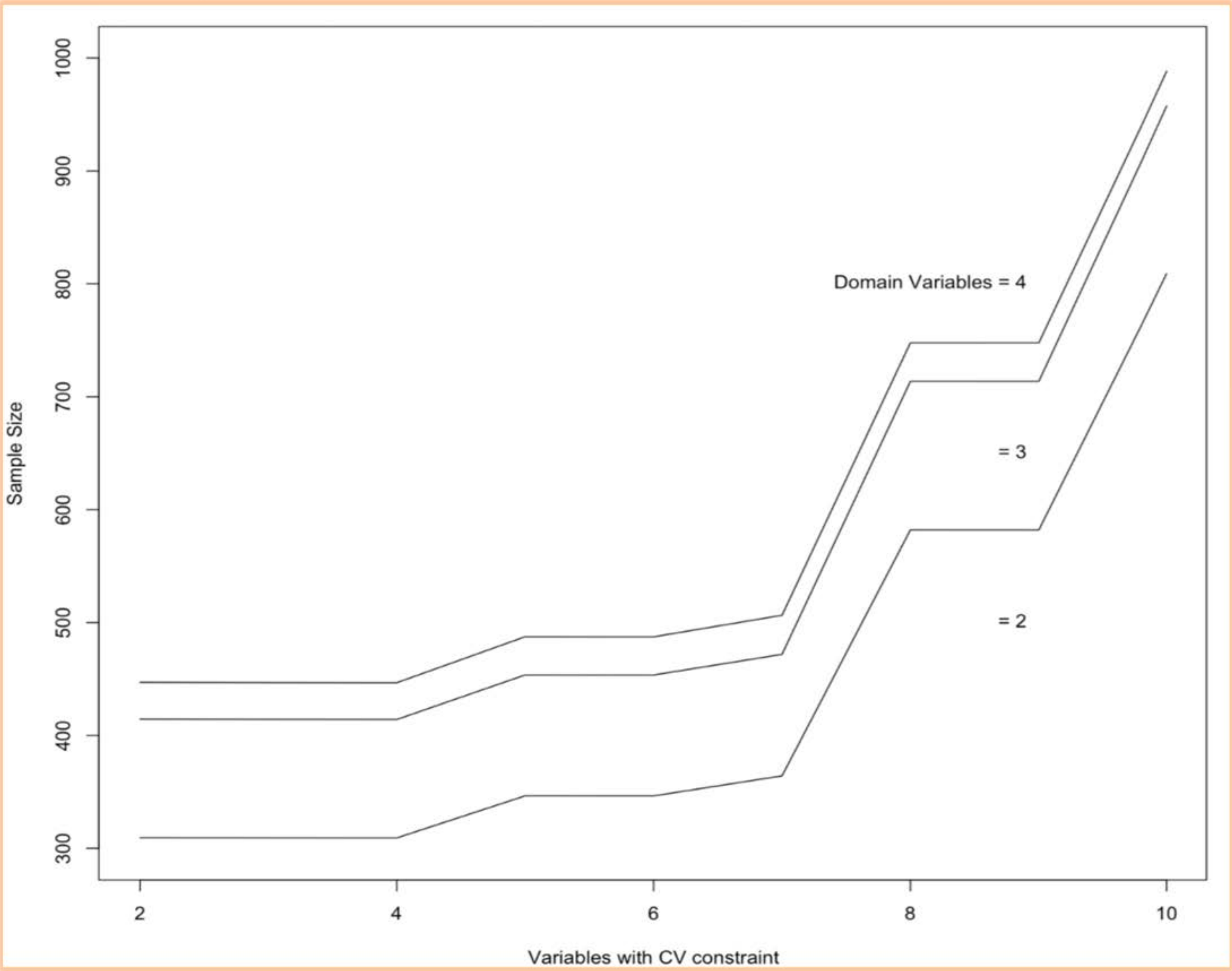
Region/GSA	n
abruzzo	67,24
calabria ionian	46,26
calabria tyrrhenian	43,62
campania	77,98
e. romagna	45,19
f.venezia giulia	25,18
lazio	66,85
liguria	54,89
marche	44,91
molise	22,75
veneto	52,35
apulia ionian	26,09
apulia adriatic	78,24
sardinia	43,22
sicily ionian	43,81
sicily tyrrhenian	100,25
sicily south	100,61
tuscany	43,42
Other Fishing Regions	5,35

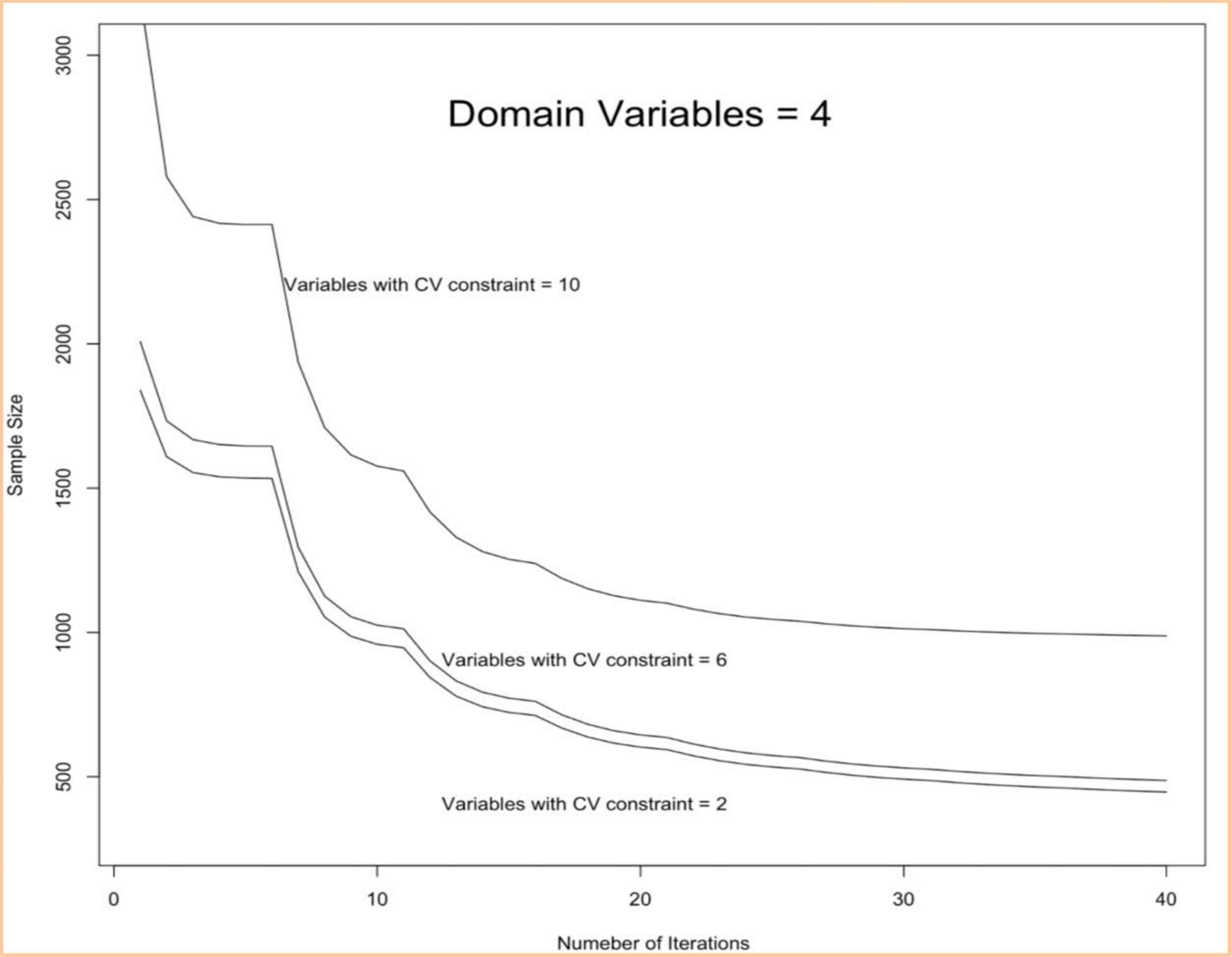


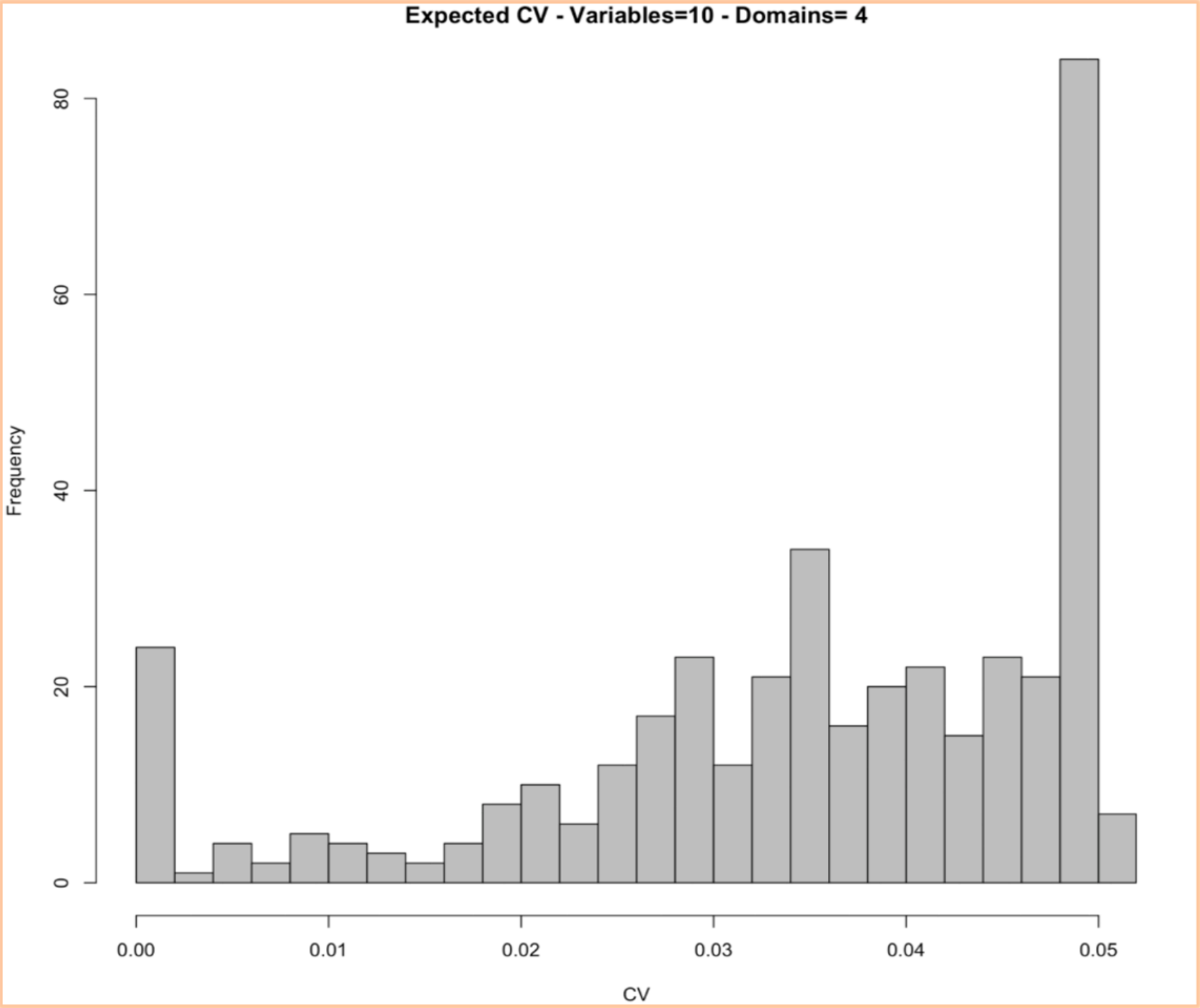
Fishing technique	n
DRB	49,05
DTS	189,2
HOK	33,16
PGP	619,30
PS	51,63
TBB	17,17
TM	28,72



Length class	n
VL0006	194,9
VL0612	413,8
VL1218	220
VL1824	100,76
VL2440	49,48
VL40XX	9,28


Sample Size
1000
900
800
700
600
500
400
300
2
4
6
8
10
Variables with CV constraint
Domain Variables = 4
= 3
= 2


Domain Variables = 4
Variables with CV constraint = 10
Variables with CV constraint = 6
Variables with CV constraint = 2
Sample Size
500
1000
1500
2000
2500
3000
0
10
20
30
40
Numeber of Iterations



Matrix containing information on the expected CV for each variable in each domain.

All combinations between variables and domains = 400.

7 > 0,05 of no more than 0,0065

Conclusions

The design adopted is very promising as it allows to respect all the constraints on CVs with much less sampling units than are used in the current survey.

Future research

- Avoid rounding the inclusion probability by applying a variant of the CUBE method, based on the distance from the constraints and not on the Landing Phase;
- attempt to extend this sample design for panel surveys, consequently studying the coordination of this complex design;
- use indirect estimates for those sub-domains that are too small (cut-off sampling).

References

Chromy J. (1987). Design Optimization with Multiple Objectives, *Proceedings of the Survey Research Methods Section. American Statistical Association*, 194-199.

Choudhry, G. H., J. N. K. Rao, and M. A. Hidiroglou (2012). On sample allocation for efficient domain estimation. *Survey Methodology* 18, 23-29.

Deville J.-C., Tillé Y. (2004). Efficient Balanced Sampling: the Cube Method, *Biometrika*, 91, 893-912.

Deville J.-C., Tillé Y. (2005). Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569-591.

Falorsi P. D., Righi P. (2008). A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation, *Survey Methodology*, 34, 223-234.

Falorsi P. D., P. Righi. 2016. A flexible tool for defining optimal sampling designs. *The Survey Statistician*, 73:21-31.

Falorsi P. D., Righi P. (2015). Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-domain Surveys,*Survey Methodology* , 41.

Winkler, W. E. (2001). Multi-Way Survey Stratification and Sampling, *Research Report Series*, Statistics #2001-01. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233

Software

https://www.istat.it/it/metodi-e-strumenti/metodi-e-strumenti-it/progettazione/strumenti-di-progettazione/multiwaysampleallocation

Thanks for your attention