Optimization of a Sampling Plan using R for Economic Data Collection

Application to the Atlantic French Fleet

Van Iseghem Sylvie1,* Demanèche Sébastien2, Daurès Fabienne1, Leblond Emilie2

1. IFREMER, Département d'Economie Maritime, Centre de Brest
2. IFREMER, Département STH, Centre de Brest

Système d'Informations Halieutiques

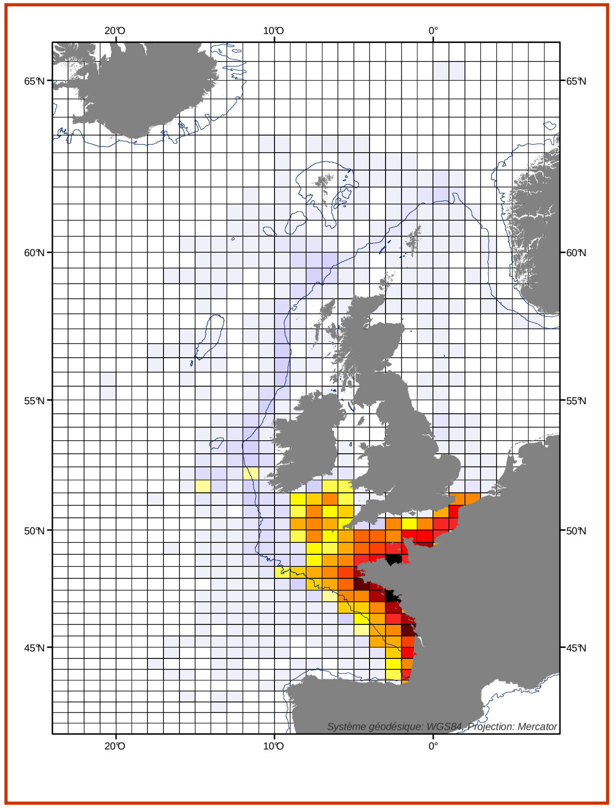
AMURE UMR CENTRE DE DROIT ET D'ECONOMIE DE LA MER

Context : Why to collect economic indicators on fisheries ?

UseR Conference 2009 – Agrocampus Rennes

Economic indicators on european fisheries : a necessity to conduct the Common Fisheries Policy (more details in the Community program for the collection of data in the fisheries sector *(EC) N°1639/2001*)

In France 70% of the fleet (<12 meters vessel) is miss-represented through official data.



The case study: The French fleet of the North Sea – Channel and Atlantic Coast

UseR Conference 2009 – Agrocampus Rennes

Optimization of a sampling plan for Economic Data Collection

Request of the community program :

Collection of Economic Indicators

by groups of vessels

with a “satisfactory” precision level L

Question :

How many vessels have to be interviewed ?…
Which vessels have to be interviewed ?…
… so that the Earning indicator is estimated
by groups of vessels
with a “satisfactory” precision

Optimization based on the Gross Revenue Indicator

Optimization of a sampling plan for Economic Data Collection

Preliminaries

Presentation of the population : the Atlantic French Fleet by groups of Vessels

Implementation in R

The link between the sampling plan and the precision defined in the community program

Optimal Sample size Estimation - *How many vessels have to be interviewed ?*

Estimated value 2006 of the Earning Parameter by segment - mean and variability

Implementation in R

Practical application of this Algorithm - *Which vessels have to be interviewed ?…*

Specificities of the Atlantic French Fleet – Spatial and Length considerations

Presentation of the systematic random sampling technique

Implementation in R

The example of the "Demersal Trawl 12-24m"

UseR Conference 2009 – Agrocampus Rennes

Ifremer

UseR Conference 2009 – Agrocampus Rennes

Optimization of a sampling plan for Economic Data Collection

Segmentation of the Atlantic French Fleet by groups of Vessels (data 2007)

	EU length class	1. <12 m	2. [12 24m[	3. [24 40m[	4. >40m	Total	%	Total	%
EU large fleet segments	EU fleet segments								
Vessels using Activ gears	1. Beam Trawels		6	2		8	0%	1613	47%
	2. Demersal Trawels / Seiners	309	442	82	13	846	25%		
	3. Pelagic Trawels / Seiners	6	86	4	4	100	3%		
	4. Dredges	159	108			267	8%		
	6. Other Polyvalent Activ gears	84	53	2		139	4%		
	5. Others Activ gears	253				253	7%		
Vessels using Passiv gears	7. Hooks	346	16	6		368	11%	1642	48%
	8. Drift / Fixed Nets	516	134	19	1	670	19%		
	9. Pots / Traps	365	18			383	11%		
	10. Other Passiv gears	111				111	3%		
	11. Other Polyvalent Passiv gears	107	3			110	3%		
Vessels using Activ and Passiv gears	12. Activ and Passiv gears	179	14			193	6%	193	6%
Total	Total	2435	880	115	18	3448	100%	3448	100%
Pourcentage	Pourcentage	71%	26%	3%	1%	100%			

Source : Ifremer

Optimization of a sampling plan for Economic Data Collection

UseR Conference 2009 – Agrocampus Rennes

Segmentation of the Atlantic French Fleet by groups of Vessels (data 2007)

Implementation in R

1. Access data base

```
library(DBI)
library(RODBC)

entree = "FPC_COMPLETE_2008_MA";
nomBase  = "C://PECH2008.mdb"
#connexion à la base de données Access POP2006
chEntree = odbcConnectAccess(nomBase)
POP=selection(entree,chEntree)
odbcCloseAll()
```

2. Sql language to select data base

```
# table ACCESS selection
selection = function(entree,chEntree){
                    req=paste("select * from ",entree)
                    table = sqlQuery(chEntree,req)
                    return(table)
}
```

2. R programming

# vessels characteristics updates

# use of merge, match, is.element, which…

Source : Ifremer

Optimization of a sampling plan for Economic Data Collection

UseR Conference 2009 – Agrocampus Rennes

The link between the sampling plan and the “satisfactory” precision

What we are looking for : Mean Value of an Economic Indicator in a group of vessels of size N **m(Y)**

What is available : **Estimation of this** Mean Value of this Economic Indicator from a sample of size n n<N **m^eY**

According to some assumptions : **95% Confidence Interval I for mY around m^eY I=[m^eY-L.m^eY;m^eY+Lm^eY]**

I defines the interval in which the true mean has 95% of chance to be. It gives an indication of how much uncertainty there is in our estimate of the true mean

=> The narrower the interval, the more precise is our estimate

=> The smaller L, the more precise is our estimate

E.U. regulation - 3 values of L - Level 1: L=25% ***(minimum precision required)***- Level 2: L=15%- Level 3: L=5%

If the **sample is randomly chosen** in the population, an analytical formula can be established between **L** [precision], **N** [size of the group or population], **n**[sample size], **mY** [mean of the indicator] and **sY** [standart error of the indicator]

Ifremer

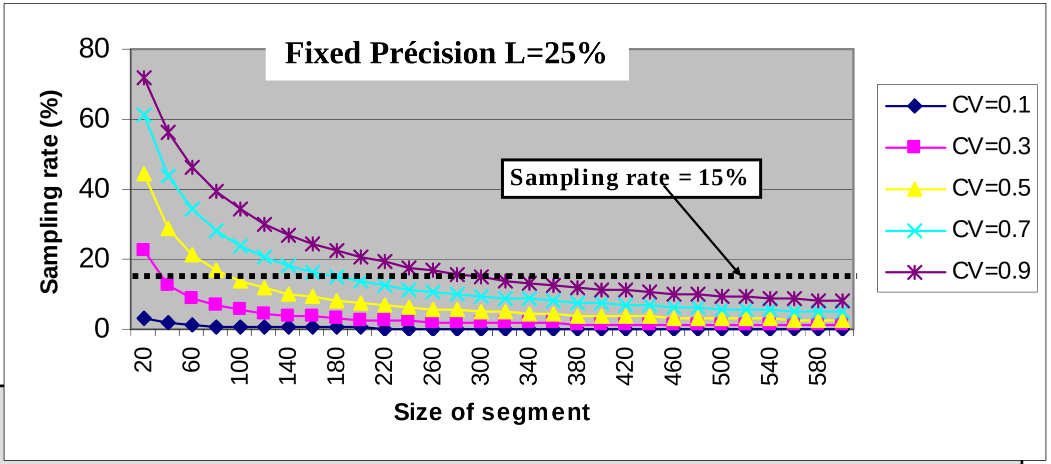
Optimization of a sampling plan for Economic Data Collection

The link between the sampling plan and the “satisfactory” precision

If the **sample is randomly chosen** in the population, an analytical formula can be established between
n [sample size], **N** [size of the group or population],

L [precision], **mY** [Mean of the indicator] and **sY** [standart error of the indicator]

$$n = N\frac{1}{1+\frac{N L^2}{4(\frac{sY}{mY})^2}} = N\frac{1}{1+\frac{N L^2}{4[CV(Y)]^2}} \quad (1)$$



Rapid analysis of this formula

If L => 0, then n => N **so, “greater” precision implies a larger sample rate**

If CV(Y) =>infinity, then n=>N **so, higher variability** of the parameter of interest **leads to a larger sample rate**

If N=>0, then n=>N **so, smaller segments implies a larger sample rate**

Sample size estimation

To apply formula (1), we need estimation of the Gross Revenue Parameter 2007 by fleet segment (mean and coefficient of variation)

Estimations are based on

- The gross revenue parameter collected in 2006 on a sample
- A revenue model to estimate gross revenue parameter **on the whole population.**

Revenue model : ***ln(CA)=5.34+0.88 ln(Pfact) -0.08 ln(Age)*** ***(Daurès Eafe 2003)***

based on explanatory variables **available for each vessel:**

- *the production factor* **(product of length of vessel, crew size and number of fishing months)**

- *the age of the vessel*.

Sample size estimation

Revenue model : ***ln(CA)=5.34+0.88 ln(Pfact) -0.08 ln(Age)*** ***(Daurès Eafe 2003)***

Implementation in R

2. Linear Model

```
library(stats);
res=lm(CA_l~FILEMO_l+AGE_l+AQ+BN+HN+NB+NPC+PC+PL+CHnex+SE+DR+TA+FI+F
       Ica+FIha+CAS+CAha+HA+DI,data=Tt)#+Nb_met5_l
  res2=step(res,direction= c("both"));
  summary(res2)
```

2. Hypotheses Tests on residuals;

```
   # bptest & dwtest : H0 homoscedastics /autocorrelation

library(lmtest);library(MASS);

bptest(CA_l~FILEMO_l+AGE_l,data=Tt);
dwtest(CA_l~FILEMO_l+AGE_l,data=Tt);
```

Residuals have satisfactory properties, model is considered valid

Sample size estimation

Optimization of the sample size for the sample data 2007 in each group of vessels

The example of 2 groups of vessels

Example 2 : Group of vessels “Mobile Gears – Dredges – <12m”

N=136 and $CV^{n-1}Y$: 53% [Coefficient of variation of the **Earning** indicator in 2006] = [**Estimator** of the Coefficient of variation of the **Earning** indicator in 2007]

According to Formula (1) we find “Optimal sample size for this group” : n=23 and n/N=16%

More important variability of the Earning Indicator implies larger sample rate

Example 3 : Group of vessels “Passive Gears – Pots and Traps– 12-24m”

N=24 and $CV^{n-1}Y$: 44.5% [Coefficient of variation of the **Earning** indicator in 2006] = [**Estimator** of the Coefficient of variation of the **Earning** indicator in 2007]

According to Formula (1) we find “Optimal sample size for this group” : n=11 and n/N=45%

Smaller segment entails a larger the sample rate [for a given variability]

Optimization of a sampling plan for Economic Data Collection

Optimal Sample size estimation in each group of vessels

	Vessel length	<12m	12-24m	24-40m	>=40m	Total
Types of Fishing Techniques						
Mobile Gears	Beam Trawl	7 46%	5 50%	1 50%		13 48%
	Demersal Trawl	38 10%	54 10%	15 17%	8 40%	115 11%
	Pelagic Trawl and Seiners	6 42%	12 10%	3 33%	2 50%	23 16%
	Dredges	23 16%	14 10%			37 13%
	Other Mobile Gears : « Tamis »	42 15%				42 15%
	Polyvalent	25 37%	11 35%	3 50%		39 37%
Passive Gears	Gears using Hooks	43 12%	9 69%			52 15%
	Drift and Fixed Nets	55 10%	21 12%	8 61%		84 11%
	Pots and Traps	55 14%	11 45%			66 16%
	Other Passive Gears	16 17%				16 17%
	Polyvalent	52 38%	2 40%			54 38%
Polyvalent Gears	Combining Mobile and Passive Gears	24 10%	7 63%			31 13%
	Total	386 15%	146 14%	30 25%	10 41%	572 15%

Optimization of a sampling plan for Economic Data Collection

A minimum sample size by group of vessels has been estimated

so that the so that the Earning indicator is estimated

by groups of vessels

with a precision L of 25% inside all groups

Total sample size : 587 fishing vessels

This sample size equals about 15% of the population is very variable between segments

⇨

In each group of vessels this percentage is

all the more important as the CV is important

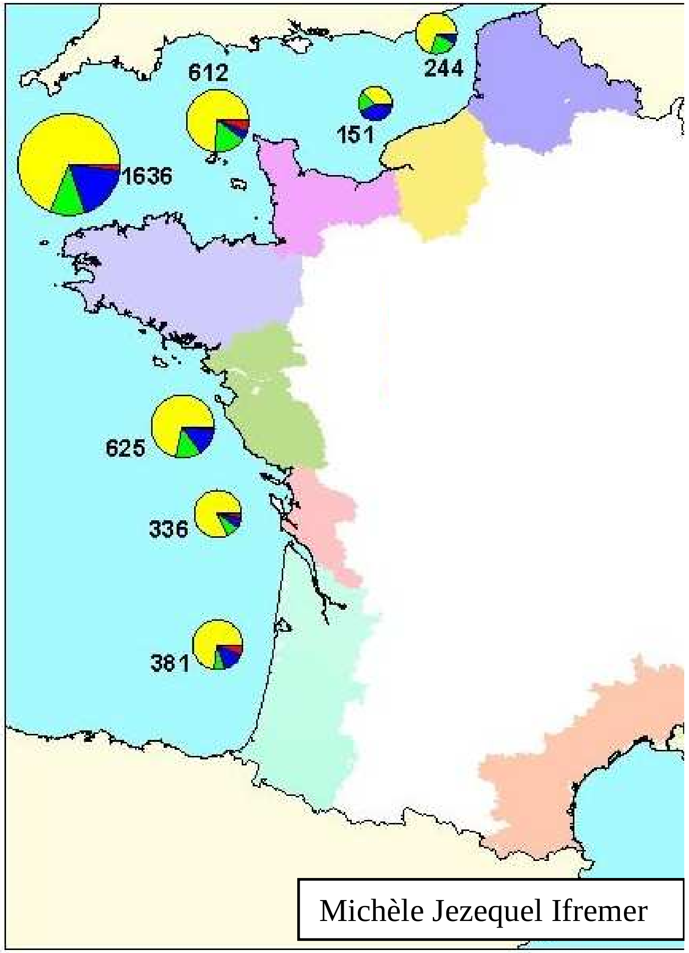
all the more important as the group is small

Remaining question : How to choose fishing vessels inside each group of vessels?

- **randomly?** *Not optimum*
- **so that the sample is representative of National Specificities**

Optimization of a sampling plan for Economic Data Collection

Specificities of the Atlantic French Fleet



Michèle Jezequel Ifremer

In order to have a good knowledge of the Atlantic French Fleet, it is important to have information about

Variability between maritime districts

Variability in length (even inside a group of vessels)

The sample can not be randomly chosen inside a segment.

It has to be representative of

The spatial variability **(priority 1)**

The length variability **(priority 2)**

Optimization of a sampling plan for Economic Data Collection

Presentation of the systematic random sampling technique

Systematic random sampling
Inside each segment :

1. List of fishing vessels ordered by
   - **priority 1 : maritime districts** to ensure spatial representativity
   - **priority 2 : vessels length** inside each maritime districts to ensure length representativity

2. Estimation of the sample size by Formula (1) in the group of vessels

3. Random number to identify the first vessel of the sample

4. Pull Vessels at regular intervals so that the number of vessels pulled at the end of the list equals the sample size estimated in (2)

The obtained sample has the optimum size defined before.
It is representative of the spatial and length variability of the group of vessels

Vessel Identification	Maritime District	Length	Sample?
*******	BA	12.8	
*******	BA	13.5	**1**
*******	BA	16.5	
*******	BA	16.8	
*******	BA	19.4	
*******	BA	19.5	**1**
*******	BA	19.6	
*******	BA	20.4	
*******	BA	20.7	
*******	AC	15.7	**1**
*******	AC	15.9	
*******	AC	16.0	
*******	AC	16.3	
*******	AC	16.5	**1**
*******	AC	16.8	
*******	AC	18.99	
*******	AC	12.0	
Etc …			

Optimization of a sampling plan for Economic Data Collection

UseR Conference 2009 – Agrocampus Rennes

Presentation of the systematic random sampling technique

Implementation in R

List of vessels ordered

```
o=order(nQAM_iseg,long_iseg);
Panel_segment_trie=Panel_segment[o,];
```

Statistical Unit definition N/n

```
pas_panel=max(N_panel_iseg/n_opt_panel,1);
unit_stat_panel[i]=ceiling(i/pas_panel)
```

Random Number to identify the first number of the sample

```
iseg_depart=max(1,runif(1)*pas_panel);
```

Identification of the other vessels (take into account priorities relative to vessels…)

Two independent sample Panel Vessels / Structrural vessels

Optimization of a sampling plan for Economic Data Collection

Comparison of the distribution in Space [Maritime quarters] and Length [12 – 24] between the Sample and the Population

The example of the fleet segment "Demersal Trawl 12-24m"

Population N=535

SRG	12	13	14	15	16	17	18	19	20	21	22	23	T
AQ	0	0	0	0	2	0	0	1	0	0	0	0	4
BN	1	1	0	2	1	1	1	0	1	0	2	0	11
HN	0	0	0	0	0	0	0	0	0	0	0	1	2
NB	1	0	0	1	0	0	0	1	1	1	1	1	7
NPC	0	0	0	0	0	0	0	0	2	1	2	2	9
PC	1	1	1	1	1	1	1	1	1	0	0	0	8
PL	1	1	1	1	2	1	0	1	1	0	1	0	11
SB	3	2	4	9	7	1	0	3	7	1	6	5	49
Total	8	5	7	15	13	4	3	6	14	3	12	10	100

Sample n=54 **n/N=10%**

SRG	1 2	13	1 4	15	16	17	18	19	20	21	22	23	T
AQ	0	0	0	0	4	0	0	0	0	0	0	0	4
BN	2	0	0	4	1	2	2	0	1	0	2	0	11
HN	0	0	0	0	0	0	0	2	0	0	0	0	2
NB	2	0	1	2	0	0	0	0	0	0	0	2	7
NPC	0	0	0	1	0	2	0	0	2	2	0	2	9
PC	0	4	0	2	0	2	2	0	0	0	0	0	9
PL	0	0	4	2	0	0	0	0	4	0	0	0	9
SB	4	2	2	9	7	0	0	4	6	0	7	7	48
Total	7	6	7	20	11	6	4	6	11	2	9	11	100

Results about the sample :

1. **Spatial representativity is very good**
2. **Length representativity is satisfactory but not as precise**

This Algorithm is a compromise to represent both length and space variability

Concluding Remarks

A methodology using R has been proposed to

Optimize the sample size of a sample when estimation and precision of economic indicators are required by group of vessels

- This optimization is based on the Gross Revenue parameter
- This optimization makes use of previously collected data – size of segments and relative variability

Choose the vessels in each segment to respect the specificities of the Atlantic French Fleet;

Distribution in space [Maritime Districts] and in length of vessels

Work on going in the Marine Economics Service

What would have been the results if an other Economic Indicator had been considered?

What are the qualities of the precision estimation given by Bootstrap algorithm?

Graphical restitutions with R