SAMPLING SCHEMES FOR SHIPPER & TRANSPORT CHAIN SURVEY: THE EXAMPLE OF THE FRENCH ECHO SURVEY

Jimmy Armoogum, Michèle Guilbault
INRETS – DEST
jimmy.armoogum@inrets.fr

Abstract

Observation of freight transport statistics is segmented by transport mode, therefore we do not have any information about multimodal transport-chains. This is the main reason why France is performing a shipper survey ECHO. The goal of this paper is to explain the methodology that we used to cope with the optimisation problem in order to obtain a large enough non-truck sample with a maximum of accuracy on the data.

Keywords: Accuracy; Firms; Unequal selection probability; Sampling schemes; Stratification; Shipper
Topic Area: D5 Data Collection Methods

1. Introduction

Generally, the observation of freight transport statistics is segmented by transport mode. For example, in France, road freight transport is measured by a permanent survey (TRM) conducted by the ministry of transport. For the other modes, the observation relies mainly on administrative documents. The statistics on inland waterborne transport result from data collected by VNF (Inland waterways of France). It is the same for air transport as follows the DGAC (Directorate-General of the Civil aviation), or for port statistics which concern the DTMPL (Direction of the Maritime transport, the Ports and Littoral). For the statistics of the rail-bound transport, they depend entirely on the SNCF (the French National railways company). Information from these traditional statistics contains origin and destination of the journey of the vehicle, the weight in tons, the ton-kilometers, type of goods and so forth. In this manner it is possible to draw a picture of all transports performed in one year for each mode. But there is no available information of the reasons behind these transports and no information on intermodal transports because there are no interactions between modes in traditional statistics. The main problem with this type of observation is that we do not have any information about (multimodal) transport-chains. Therefore, we are performing a shipper survey ECHO, in France.

The data collection of the French shipper survey ECHO, that started in July 2003, have the following objectives:

- to identify the logistical determinants of the transport chain, both on the organizational and physical point of view, in order to develop a behavioural analysis of firms involved in the transport chains;
- thus to collect information directly from all the operators involved in the shipment or its organization: the shipper, but also the carriers, the customer, and all other involved parties as forwarders or logistical service providers;

- to focus the analysis on the modal choice and on its conditions to clarify the stakes of transport policy.

Therefore, the sample scheme is designed in order to obtain about one third of non-truck shipments (while a uniform random sample we should have only 5% of non-truck shipments), more precisely we want : 8 to 10% of railway shipments; 8 to 10% of maritime shipments; 8 to 10% of air shipments; 4 to 6% of combined rail-road shipments; 2 to 4% of inland waterborne shipments; 25% of international shipments an 9% of firms from the region Nord Pas-de-Calais.

The methodology is based on the experience of the previous shipper survey (1988). This survey incorporates two major components: the tracing of a selection of shipments from their departure from the plant up to their arrival to the consignee; the description of the main organizational features of the shippers' plants influencing the transport characteristics. We choose the same protocol as for the MYSTIC European project (see Rizet and al., 2000). Where a two phases sample was used: a first sample among the firms (see section 1), and then a selection of shipments (see section 2) sent from the selected firms. Per firm, 3 shipments are chosen (among the last 20 shipments of a firm) and tracked up to the final customer. Different types of questionnaire forms are used: for the shipper (plant), for the shipment, and for the operators (carriers, ancillary services).

The goal of this paper is to explain the methodology that we used to cope with the optimization problem in order to obtain a large enough non-truck sample with a maximum of accuracy on the data.

2. First sample phase: selection of firms

The first phase of the sample consist to select firms that allows us to gathered non-truck shipments in preserving accuracy. At this stage it's important to select firms that use non-truck shipment, because with a uniform random selection of firms almost all of them should be exclusive road users.

2.1 Non truck users firms profiles and exports firms profiles

The French National Institute of Statistics INSEE allow us to analyse the population data base of firms SIRET. Therefore at the level of firms we have some available variables in the SIRET data file which could help to find out "non exclusive road" firms. These variables are:

- the activity of the firm,
- the number of employees,
- the type of firms : head office or not,
- the origin : creation or part exchange,
- the date of creation,
- possibly the date of cessation,
- the Address,
- the SIRET number.

In the previous shipper & transport chain survey conducted in 1988, the 3 most recent shipments were selected. Firms "rare" mode users are those where at least one of the shipments is made by the "rare" mode that we consider. As the number of "inland waterborne" shipments was very low we did not conceive this mode (only 4 shipment with this mode in 1988). For each of the "rare" mode we used a LOGIT model to find out non-truck users firm profiles (the outlines are in terms of activities, number of employees and location). We used the same methodology to find "export" firms, see result in table 1.

Table 1: Dimensions that explain the use of non-truck shipments.

Firms that used the following modes :	Activity (1)	Number of employees (1)	Firms localization (1)
Rail	Yes	No	No
Maritime	Yes	No	No
Air	Yes	Yes	No
Combined rail-road	Yes	No	No
International	Yes	Yes	No
>From Nord-Pas-de-Calais region	No	No	Yes

Source : Inrets calculations from Chargeur survey 1988.
(1) Parameter significativity at the level of 1%.

With a Logit model, we find that the activity of the firm allows us to capture (in the sample) non-exclusive road shipment firms and also exporting firms. Using firm location will obviously favour the inclusion in the sample of firms from the Nord-Pas-de-Calais region and also inland waterborne shipment. In the other hand, the number of employees doesn't bring any information for non-truck shipment, except for air shipment, but it is useful to identify exporting firms.

Thus, if we want non-truck users we have to use the firm' activity and if we want 9% of firms from the Nord-Pas-de-Calais region we have to treat the location variable. As we have access to the exhaustive SIRET database about French plants, we constructed 39 group of activities in order to keep the continuity and the homogeneity of the production process [cf. Guilbaut and al., 2002].

2.2 How to improve the accuracy

The budget of the ECHO survey allows a sample size of about 3230 firms (the sampling rate is about 4%). Optimizing the sample scheme is an important issue, specially in the field of firms because of its heterogeneity. For example, if we take a uniform random sample of 3230 firms without any optimization, the total number of shipments is known within a confidence interval of ± 60% at the level of 95% and we should get about 95% of truck shipment in the sample.

If we take a sample of 3230 firms, stratified on the activity, the accuracy of the estimates are (at the level of 95%):

with an optimization of the tonnage:

- the total number of shipments is known within ± 60% and total tonnage within ± 25%;

with an optimization of the number of shipments:

- the total number of shipments is known within ± 20% and total tonnage within ± 70%.

If we take a sample of 3230 firms, stratified on the number of employees of the firms , the accuracy of the estimates are (at the level of 95%):

with an optimization of the tonnage:

- the total number of shipments is known within ± 10% and total tonnage within ± 14%;

with an optimization of the number of shipments:

-the total number of shipments is known within ± 8% and total tonnage within ± 18%.

Thus, is then helpful to introduce the 5 groups of number of employees (firms with 6 to 19 employees, then from 20 to 49 employees, from 50 to 499 employees, from 500 to 999 employees and finally 1000 employees or more) as criteria of stratification to improve the precision of the estimators. Table 2 gives the optimal distribution for the number of employees in order to obtain the maximum of accuracy in term of tonnage.

Table 2: repartition of the sample for a maximum of accuracy in term of tonnage

Number of employees	Number of firms In the population	Number of firms in the sample	Sampling rate
6-19 employees	35572	385	1.1%
20-49 employees	26317	579	2.2%
50-499 employees	15319	1717	11.2%
500-999 employees	622	352	56.6%
1000 employees or more	197	197	100%
Population	78027	3230	4.1%

Sources : Inrets from SIRET (2002) of Insee and chargeur survey 1988.

2.3 Stratification of firms according to activity and number of employees

If we take a stratified sample according to activity (39), manpower (5 sizes of firms) and localization (3), we should obtain 585 strata. Then the average number of firms per stratum should be 5.7, which is too few. Studying the profile of non-road users (with firm activity at the finest level) with data from SNCF (the French rail firm), from customs (for maritime and international shipments) and from VNF (association of plants that used inland waterways), we obtain 79 modal subgroups of activity, which have to be crossed with the 5 groups of size (number of employees). Therefore we stratify the population into 300 non-empty strata (see table 8 in appendix). The allocation of the sample to the 300 strata is guided by the calibration on the margin for the activity and the margin for the number of employees. This methodology should lead us to achieve the objective (one third of non-truck shipments) with a maximum of accuracy in our estimates.

After choosing the allocation in the 300 strata, the distribution of firm's modal subgroup is given in table 3. For example, to achieve the objective of having about 900 rail shipment we find that it's necessary to include about 70% of these firms in the sample.

Table 3 : Distribution of firm's modal subgroup in the population and in the sample

Firm's modal subgroup	Population (1)		Sample (2)		% of the sample in the population = (2) / (1)
Rail	404	0,5%	282	8,7%	69,8%
Inland waterborne	2676	3,4%	435	13,5%	16,3%
Inland waterborne + Rail	311	0,4%	257	7,9%	82,6%
Air	15045	19,3%	569	17,6%	3,8%
Others	59591	76,4%	1690	52,3%	2,8%
Total	78027	100%	3233	100%	4,1%

Sources : Inrets from INSEE (SIRET, 2002)

3. Second sample phase: selection of shipments in each firm

Once firms are selected, it's important to over-represent non-truck shipments to achieve the objective of 35% of non-truck shipments in the sample, because if we take a uniform random sample we should have less than 5% of non-truck shipments.

3.1 Different modes of transport and destinations

In each selected firm, we collect information on the last 20 shipments during an interview lasting no more than 15 mn (in fact 17 shipments were recorded in average). Then, 3 shipments are chosen and tracked up to the final customer. We chose these 3 shipments with unequal probabilities. Of course, shipments carried by road have the lowest probability to be selected at this stage.

We have 7 modes of transport and 3 combinations of modes identified as a proper mode, they are, in growing order of mode affectation's:

1 – Road on own account
2 – Hired road
3 – Rail
4 – Combined Rail-Road
5 – Inland waterborne
6 – Maritime
7 – Air
8 – Combination Rail + Maritime
9 – Combination River + Maritime
10 – Combined Rail-Road + Maritime

The calculation of over-representation coefficients for rare modes for the selection of shipments we need the following data :

From the initial interview in the firm (CAPI procedure) :
- The mode of transport of the M (M is equal to 20 in the general case) last shipments;
- The destination of these shipments (in France or Abroad).

From a data file where the value of the parameter may change:
- The coefficient of over-representation for each mode according to the destination.

To initialize the process, we have taken the following coefficient:

Mode:	coef (mode):
Road on own account	31
Hired road	82

Rail	630
Combined Rail-Road	355
Inland waterborne	1800
Maritime	410
Air	963
Combinaison Rail + Maritime	1040
Combinaison River + Maritime	2210
Combined Rail-Road + Maritime	765

Destination + mode:	coef (destination):
France (Every mode)	1.00
Abroad (every mode except maritime)	1.30
Abroad: GB + Ireland (mode maritime)	1.30
Abroad except (GB + Ireland) (mode maritime)	2.15

The coefficient of over-representation of a shipment is given by the multiplication of the coefficient of mode by the coefficient of destination. These figures come from computation according to clients profiles using the same data as the selection of firms process (see section 1.1) i.e. data from SNCF the French rail firm, the file from customs (for maritime and international shipments) and the file from VNF association of plants that used river. Let's note that these coefficients could be modified at any time along the data collection process.

3.2 Stages of the algorithm

To select 3 shipments among the M (M is equal to 20 in the general case) last shipments, we have written the following algorithm (see appendix **2**) :

Let:

1. c_i , be the multiplication of "transport mode" and "destination + mode" of the shipment, the result is in the bracket : 31 et 4751.5.

2. PG1, the step of spread $PG1 = \dfrac{1}{3}\sum_{i=1}^{M} c_i$

We may have the 3 following cases A, B, or C:

A) For any shipment i we have $c_i < PG1$:

In this case, we make a systematic sampling selection procedure of shipment in taking a random number and PG1 as step of spread.

B) For 2 shipments we have $c_i \geq PG1$

We take the 2 shipments where $c_i \geq PG1$, and we select the third one with a systematic sampling selection procedure in the M-2 shipments, with the following inclusion

probability : $p_i = \dfrac{c_i}{\sum_{i=1}^{M-2} c_i}$

C) For only one shipment i we have $c_i \geq PG1$

We take the shipment i where $c_i \geq PG1$, then we calculate a new step of spread

$$PG2 = \sum_{i=1}^{M-1} c_i \, ,$$

And we may have two cases:

C-1) If : $\forall i = 1,...M-1 \quad c_i < PG2$

We select 2 shipments with a systematic sampling selection procedure with the

following inclusion probability (for the M-1 left shipments): $p_i = \dfrac{2\,c_i}{\sum\limits_{i=1}^{M-1} c_i}$

C-2) if we have for one shipment j, $c_j \geq PG2$

We take the shipment j, and select the third one with a systematic sampling selection

procedure with the following inclusion probability: $p_i = \dfrac{c_i}{\sum\limits_{i=1}^{M-2} c_i}$.

3.3 An iterative process to estimate the over-representation coefficient

At any time during the data collection, we have the possibility to modify the coefficients of "transport mode" and "destination + mode" and therefore the c_i , to achieve our objectives of non-road and abroad shipments.

Since the beginning of this paper, all we describe above is purely theoretically without considering the data collection background such as nonresponse and etc.

4. Data collection up to November 2003
4.1 The second phase sample (selection of non-truck shipments)

Up to November 2003, we have the response of about 18% of the overall sample (3230 firms), as first result but we did not achieve the objective of gathered about one third of non-truck shipments (Table 4). Let's take the example of rail shipments, where we had in the final sample only 65 shipments out of 103 possible. These 103 rail shipments where sent by 52 firms. Among these firms there are 44 firms where we selected all their rail shipments. In the 8 firms that we did not select all their rail shipments, 4 of them is because we have selected 3 rails shipments (which is the maximum). In the 4 firms left we have the following figures:

- 1 rail shipment + 2 maritime shipments;
- 2 rail shipments + 1 maritime shipments;
- 2 rail shipment + 1 Rail-Road shipment and
- 1 rail shipment + 1 road shipment + 1 maritime shipment.

For rail shipments, either we have collected the maximum we can do (according our methodology) either we have to consider other non-truck modes. Therefore the second phase of the sample scheme (the selection of shipments) has worked correctly.

Table 4 : Distribution shipments modes in the Shipment sample up to November 2003 (18% of the overall sample)

Shipments modal share	Objective	The last 20 Shipments sample		The 3 among 20 shipments sample	
Road	60% - 70%	9266	91.9%	1467	82.2%
Rail	8% - 10%	103	1.0%	65	3.6%
Rail-Road	4% - 6%	46	0.5%	16	0.9%
Inland waterborne	2% - 4%	15	0.1%	13	0.7%
Maritime	8% - 10%	297	2.9%	100	5.6%
Air	8% - 10%	360	3.6%	123	6.9%
Total	100%	10087	100%	1784	100%

Sources : Inrets from ECHO 2003-2004 (Sample up to November 2003)

4.2 The first phase sample (selection of firms)

Up to November 2003, we have interviewed 597 firms (about 18% of the overall sample), and the distribution according modal subgroup is as follow : 354 "Road" firms, 41 "Rail" firms, 70 "Inland waterborne" firms, 25 "Inland waterborne + Rail" firms and 107 "Air" firms. The distribution of shipments modal share according to firm's modal subgroup is given in the table 5. In each of the modal subgroup we found a massive use of road shipments (up to 80%). For example "Rail" shipments can be gathered in "Rail" and "Inland waterborne + rail" firm's modal subgroup which show that firm's modal subgroup is efficient (except for Inland waterborne but this is due to the very low quantity of inland waterborne shipments).

Table 5: Shipments modal share according to firm's modal subgroup

Firm's modal subgroup	Shipments modal share						
	Road	Rail	Rail + Road	Inland waterborne	Maritime	Air	Total
Road	93,7%	0,4%	0,6%	0,1%	2,8%	2,4%	100%
Rail	91,2%	5,5%	0,3%	0,4%	2,0%	0,7%	100%
Inland waterborne	96,7%	0,3%	0,3%	0,1%	2,3%	0,3%	100%
Inland waterborne + rail	80,1%	8,7%	0,7%	1,7%	4,9%	3,9%	100%
Air	85,4%	0,0%	0,2%	0,0%	3,8%	10,5%	100%

Sources : Inrets from ECHO 2003-2004 (Sample up to November 2003)

The main reason why the objectives are not acheives are because the very high level of nonresponse rates for certain stratums. For example, even if we interview all firm in the modal subgroup "Rail" we should have only about 105 "Rail" firms in the final sample (if the nonresponse rates keep the same level as before November 2003) (see table 6).

Table 6: Distribution of firm's modal subgroup in the population and in the sample

Firm's modal subgroup	Population		Theoretical sample		Probable sample	
Road	59591	76,4%	1690	52,3%	2133	66,0%
Rail	404	0,5%	282	8,7%	105	3,2%
Inland waterborne	2676	3,4%	435	13,5%	335	10,4%
Inland waterborne + Rail	311	0,4%	257	7,9%	90	2,8%
Air	15045	19,3%	569	17,6%	570	17,6%
Total	78027	100%	3233	100%	3233	100%

Sources : Inrets from INSEE (SIRET, 2002)

5. Conclusion

The overall conclusion of this project is : even if from a theoretical point of view the sample schemes seems perfect we should take great care of the data collection to reduce the nonresponse and the burden of the firms.

Since we did not achieve the objectives do we have worked for nothing? The comparison of a shipment random sample and the ECHO sample is given in table 7, and the ECHO sample multiply by 14 the number of "rail" shipment, by 4 the number of "Rail-Road" shipments, by 24 the number of "Inland waterborne" shipments, by 5 the number of "Maritime" shipments and by 3 the number of "Air" shipments.

Table 7: Comparison of a shipments random sample and the ECHO shipments sample

Number of shipments per mode	Random sample [1]	ECHO sample (up to November 2003) [2]	Gain = [2] / [1]
Road	1719.6	1467	0,9
Rail	4.6	65	14,0
Rail+Road	4.3	16	3,7
Inland waterborne	0.5	13	24,3
Maritime	18.9	100	5,3
Air	36.2	123	3,4
Total	1784	1784	

Sources : Inrets from ECHO 2003-2004 (Sample up to November 2003)

Acknowledgements

Reference

Armoogum, J., 2002. Correction de la non-réponse et de quelques erreurs de mesure dans une enquête par sondage : Application à l'enquête Transports et Communication 1993-94, INRETS report n° 239.

Armoogum, J., Madre, J.L., 2003. Sample Selection, in Axhausen K., Madre J.-L., Polak J., Toint P. (Eds.), Capturing Long-Distance Travel, Research Science Press, Baldock, pp 205-222.

Besnard, F., 2002. Optimisation du plan de sondage pour une enquête sur les transports de marchandises, Rapport de stage, IUT de Vannes.

Guilbault, M., Armoogum, J., and Rizet, C., 2002. Enquête ECHO- Rapport méthodologique d'étape, INRETS - METL, 45 p. + annexes

Rizet, C., Guilbault, M., van Meijeren, J. C., and Houée, M., 2003. Tracking along the Transport Chain via the Shipper Survey, in Transport Study Quality and Innovation, P. Stoher and P. Jones (eds), Pergamon Elsevier Science, chapter 25, pp. 427-441

Rizet, C., 2003. Energy consumed in freight transport - Estimates from shipper surveys. paper for the European Transport Conference Strasbourg, October, 2003.

Rizet, C., and Keïta, B., 2002. The logistical Choices of Companies and Energy Consumption, INRETS – ADEME report, 91 p.

APPENDIX

Table 8: Distribution of the firms population: Firms activities (39) according to modal subgroup (5)

Firms activities on the target population of ECHO	Number of firms by modal subgroup					Total	%
	Other	Rail	Inland waterborne	Inland waterborne + Rail	Air		
01-Metallurgie	597	15	122	51	0	785	1,0
02-Verre,Ceramique,Mat.Constr.	1388	0	137	0	0	1525	2,0
03-Raffinage Cokefaction	39	6	5	25	0	75	0,1
04-Chimie de base	367	38	132	58	0	595	0,8
05-Parachimie	1991	0	0	0	0	1991	2,6
06-Travail des Metaux	5079	0	0	0	0	5079	6,5
07-Filature,Tissage,Tannerie	1103	0	0	0	0	1103	1,4
08-Ind. du Bois et du Papier	1975	4	165	0	0	2144	2,8
09-Recuperation et Traitement des dechets	1006	23	0	0	0	1029	1,3
10-CG Biens intermediaires non classes	286	0	0	0	0	286	0,4
11-CG agricole	716	135	192	121	0	1164	1,5
12-CG Bois, Mat.constr., Sanitaires	2298	0	928	0	0	3226	4,1
13-CG Comb,Metaux,Mineraux,Chimie	1072	93	442	36	0	1643	2,1
14-Elements finis de Construction	2466	0	0	0	0	2466	3,2
15-Fab. Eqpt industriel, Machines	6434	0	0	0	0	6434	8,3
16-Outillage mecanique, Ouvrages metaux	1551	0	0	0	0	1551	2,0
17-Mach.electriq, Compos.electroniques	744	0	0	0	1378	2122	2,7
18-Articles techniques, Mat.precision	0	0	0	0	2911	2911	3,7
19-CG Biens de production	7593	0	0	0	0	7593	9,7
20-IAA,prdts alim. Frais	2692	0	248	0	0	2940	3,8
21-IAA,prdts alim. secs; Parfumerie; Entretien	1830	0	0	0	354	2184	2,8
22-Ind. de Boissons	588	32	0	0	0	620	0,8
23-Travail du grain	475	1	79	3	0	558	0,7
24-Fab. art. textile et cuir, Confection	3740	0	0	0	0	3740	4,8
25-Ind. Automobile et cycles	639	57	25	17	1527	2265	2,9
26-Ind. Pharmaceutique	0	0	0	0	475	475	0,6
27-Electromenager, Hifi, Telephonie, Informatique	224	0	0	0	655	879	1,1
28-Fab. de Meubles	1551	0	0	0	0	1551	2,0
29-Edition	0	0	0	0	3638	3638	4,7
30-Metaux precieux Joaillerie	0	0	0	0	235	235	0,3
31-Fab. Objets divers	2558	0	0	0	0	2558	3,3
32-CG Produits Epicerie et entretien	1212	0	0	0	0	1212	1,6
33-CG Boissons	953	0	0	0	0	953	1,2
34-CG Pdts alim frais	2038	0	201	0	0	2239	2,9
35-CG Biens consom. Non alim.	2519	0	0	0	3872	6391	8,2
36-VPC	417	0	0	0	0	417	0,5
37-Entrepots Industrie	222	0	0	0	0	222	0,3
38-Entrepots du Commerce	254	0	0	0	0	254	0,3
39-Entrepots Prestataires	974	0	0	0	0	974	1,3
Total	59591	404	2676	311	15045	78027	100
Percentage	76,4%	0,5%	3,4%	0,4%	19,3%	100%	

Sources : Inrets from SIRET file of the French National Institute of statistics and Economics Studies, 2002.

Graph of shipments selection algorithm

Selection of 3 shipments among
the M last shipments
(M = 20 in the general case)

We calculate :
- the ci and
- $PG1 = \dfrac{1}{3}\sum\limits_{i=1}^{M} c_i$

A

If: $\forall i = 1,...,M$
$$c_i < PG1$$

In this case we make a
systematic sampling selection
procedure of shipment, with the
following inclusion probability:
$$p_i = \frac{3\,c_i}{}$$

B

If for 2 shipments we have:
$$c_i \geq PG1$$

We take the 2 shipments where:
$$c_i \geq PG1$$
and we select the third one with a
systematic sampling selection
procedure, with the following
inclusion probability
$$p_i = \frac{c_i}{\sum\limits_{i=1}^{M-2} c_i}$$

C

If for one shipment we have:
$$c_i \geq PG1$$

We take the shipment where:
$$c_i \geq PG1$$

We calculate: $PG2 = \sum\limits_{i=1}^{M-1} c_i\,/2$

if we have for one shipment:
$$c_j \geq PG2$$

If: $\forall i = 1,...M-1$
$$c_i < PG2$$

We take the shipment j where:
$$c_j \geq PG2$$
and select the third one with a
systematic sampling selection
procedure with the following
inclusion probability:
$$p_i = \frac{c_i}{\sum\limits_{i=1}^{M-2} c_i}$$

We select 2 shipments with a
systematic sampling selection
procedure with the following
inclusion probability (for the M-1
left shipments):
$$p_i = \frac{2\,c_i}{\sum\limits_{i=1}^{M-1} c_i}$$