Building a Model of Freight Generation with a Commodity Flow Survey

Duy-Hung Ha, François Combes

Building a model of freight generation with a commodity flow survey

Duy-Hung Ha

Technical Division for Transportation Infrastructure and Materials, Cerema, Cité des Mobilités, B.P. 214, 77487 Provins Cedex – France

François Combes

Technical Division for Transportation Infrastructure and Materials, Cerema, Cité des Mobilités, B.P. 214, 77487 Provins Cedex - France

French Institute of Science and Technology for Transport, Development and Networks - East Paris University, Cité Descartes, 14-20 bvd. Newton, 77447 Marne-la-Vallée Cedex 2 - France

Abstract This study estimates a disaggregate freight generation model with the French shipper survey ECHO. This survey contains accurate information about French shippers, with variables describing their economic activity, the diversity of their production, their relationships with their clients and carriers, etc. These variables and their influence on production and attraction are first investigated sequentially. Then generation models are estimated using ordinary least squares, through various approaches: continuous explanatory variables only, continuous and qualitative variables and their interactions, and finally simple models for applications with limited data availability.

1 Introdution

In freight transport models, freight generation is the stage which estimates the amount of cargo generated or attracted by establishments or by geographic zones. The literature distinguishes two classes of models: on the one hand Freight Generation (FG) and Freight Attraction (FA) models, which are the production and attraction of cargo measured in tonnage (or volume), on the other hand Freight Trip Production (FTP) and Attraction (FTA) models, which regard the number of vehicle movements (Holguin-Veras et al., 2014).

Generation models can be estimated with aggregate or disaggregate data. Disaggregate data is interesting because it avoids aggregation biases. It also

allows, in some cases, to investigate the influence of variables which only make sense at the disaggregate level, or the presence of non-linear effects. Finally, disaggregate models can be a good basis to disaggregate aggregate data (for example, regional freight data could be disaggregated to the city level with the appropriate establishment dataset and a reliable disaggregate generation model.)

The estimation of disaggregate generation models requires disaggregate data at the establishment level. This data is obtained through surveys targeted at business establishments, such as commodity flow surveys. Establishments are typically described by the economic activity sector, economic size (workforce or turnover), location, and type (offices, plant, warehouse, etc.). Variables about production, logistics, relationships of the establishments with their business partners (providers, clients, carriers) are generally not described. With adequate data, it is possible to estimate both Freight Generation (FG) and Freight Trip generation (FTG) volumes, as in Holguin-Veras et al. (2012), who show that FG and FTG do not obey to the same logic.

The French shipper survey ECHO was realized in 2004-2005. This survey was designed to investigate the relationship between freight transport, production and supply chains, among other objectives. As a consequence, a limited number of establishments were surveyed, but a large number of variables were observed for each of them. In particular, this survey provides information on the economic characteristics of shippers (economic sector, turnover, workforce, etc.), production (number of product ranges, number of SKUs (stock keeping unit), etc.), logistics (share of transport costs in the product value, etc.) and economic relationships (number of clients, carriers, type of contract with carriers, etc.), as well as the total number of tons carried out or received per year, and the number of shipments sent per year. As such, this dataset offers the opportunity to statistically analyze the relationship between freight generation and many variables which are usually not observed. Shipment frequency is probably strongly correlated to FTG, but not identical: it is very likely that a unique vehicle can leave an establishment carrying many shipments sent to distinct destinations when the vehicle's destination is a break-bulk platform.

The objective of this study is to build a disaggregate generation model with the ECHO dataset. Generation was studied in the ECHO database by Rizet and Hémery (2008), who examined the relationships between generation, attraction, and some of the variables in the database, but did not investigate the interactions effects, and did not estimate models. In Section 2, the ECHO dataset is described, as well as the variables of interest for the paper. Section 3 describes a sequential analysis of the influence of the explanatory variables on generation using ANOVA and ANCOVA. Then, Section 4 presents generation and attraction models estimated by ordinary-least-squares, with a number of different specifications. Section 5 concludes the paper.

2 Presentation of the data

The ECHO dataset provides information on 10,462 shipments sent by 2,935 French shippers, obtained by face-to-face and phone interviews, and based on closed questionnaires. It is similar to a commodity flow survey or CFS; its main particularity is that it provides very detailed information on the shipper-receiver relationship, and on the way the shipments were transported (Guilbault, and Soppe, 2009).

In the ECHO survey, a shipper is an establishment. Each shipper is described by a large number of variables, some typical (economic activity, workforce, turnover) and others not. In this study, the dependent variables are: the freight volume generated by the establishments in tons per year E_i, the freight volume attracted by the establishments in tons per year A_i, and the number of shipments sent by the establishment per year S_i. In the following, these variables are transformed into logarithms.

The explanatory variables are categorized into four groups:

- Economic activity: shippers are described by their economic activity group *G*, and by their turnover *T* (turnover is not available directly in the ECHO database, it was discretized into nine classes).
- Relationship with the economic environment: shippers are described by the type of contract *TC* they most often have with carriers (three levels: long period contracts, occasional contracts, or both); the number of clients *Ncl* which constitute 80% of their activity; and the number of carriers or freight forwarders *CR* with which they worked during the year.
- Organization of the production: the number of distinct product ranges *Npr*, the number of references or SKU *Nr*, and the share of transport cost in the product value *CT*.
- Employment: shippers are described by the number of employees *N* and by their main qualification level *L* (four levels: unskilled, without certification, skilled, highly skilled).

Many of these variables are completely absent from classic freight transport databases: freight transport databases, targeted at carriers, typically do not observe shippers; while commodity flow surveys, targeted at shippers, do not cover the same range of information.

Table 1: Explanatory variables for shipper *i*

Category	Qualitative variables	Quantitative variables
Economic activity	Shipper activity group *G*	
	Slices of turnover *T*	
Relations with economic agents	Type of contract with carriers or freight forwarders *TC*	Number of clients log(*Ncl*)
		Number of carriers *CR*
Production and logistics characteristics		Number of references log(*Nr*)
		Share of transport cost in product value log(*CT*)
Employment	Labour qualification level *L*	Number of employees log(*N*)

The objective of the paper is to analyze the relationship between these explanatory variables and the dependent variables, and then to estimate freight generation models.

3 Analysis of the explanatory variables

3.1 Methodology

The main tools used in this section are the analysis of variance (ANOVA) and the analysis of covariance (ANCOVA); their principles are briefly summarized below.

Analysis of variance

Many of the explanatory variables in the ECHO database are categorical ones; the first step to determine whether they have an influence on the dependent variable is the analysis of variance, or ANOVA (Tenenhaus, 1986). The ANOVA methodology requires that the dependent variables are normally distributed, which is why E_i, A_i and S_i are transformed into logarithms. It also requires that for each sub-group defined by the categorical explanatory variables, the distribution of the dependent variable is normal, and that the variance is the same among the sub-groups (homoscedasticity).

The one-way ANOVA models is generally used and formulated as follows:

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik} (1)$$

Where Y_{ik} is the continuous dependent variable of the k^{th} value in the sub-population i of independent variable ; μ is the average level value of Y ; α_i is the effect of the sub-population i of X on Y ; and ϵ_{ik} is the error term.

The ANOVA procedure also allows examining the effect of interaction between categorical variables on the dependent variable$Y$. In particular, the two-way ANOVA model is written as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} (2)$$

Where Y_{ijk} is the k^{th} value in the sub-group corresponding mutually to the sub-population j of the second independent variable X_2 and the sub-population i of the first independent variable X_1; μ is the average level value of Y ; $\alpha_i$is the effect of the sub-population i of X_1 on Y ; β_j is the effect of the sub-population j of X_2 on Y ;$(\alpha\beta)_{ij}$is the effect of the interaction between the i^{th}sub-population of $X_1$and the j^{th} sub-population of X_2 on Y; and ϵ_{ijk} is the error term.

In practice, the F-test is applied to verify the null hypothesis of the equality of means among the distinct sub-populations. However, this test is only valid under the hypothesis of homoscedasticity. This null hypothesis can be tested using Levene's statistic. If the test fails, i.e. if there is heteroscedasticity, then other tests can be applied, such as Welch's test to test the equality of means (Welch, 1951).

Analysis of covariance

The analysis of covariance is a technique treating both continuous and categorical explanatory variables in relationship with a continuous dependent variable. The categorical explanatory variables in ANCOVA models are called independent *factors* while the continuous explanatory variables are called *covariates*. The ANCOVA is in fact a combination between the ANOVA analysis and the linear regression.

ANCOVA analysis allows increasing the statistic explicative power of the model, because the effects of the factors are adjusted after considering the variability of the covariates. The interaction between the factors and the covariates are also analysed and estimated in ANCOVA.

In general, the ANCOVA models are formulated as follows:

$$Y_{ij} = \mu + \tau_i + \beta X_{ij} + \phi_i X_{ij} + \epsilon_{ij} (3)$$

Where $Y_{ijk}$is the j^{th}observed response value of Y in the i^{th} sub-population of the independent variable X; μ is the average level value of Y; τ_i is the effect of the sub-population i of Xon Y; β is the overall slope of the model; $\phi_i$is the effect of the i^{th} sub-population of Xon the slope of Y; and $\epsilon_{ij}$is the error term.

For both the ANOVA and ANCOVA, a series of statistical tests exist, testing against the null hypothesis of the absence of effect of a given explanatory variable on the mean values of the dependent variable. One of the main advantages of ANOVA and ANCOVA methodologies is that they allow to quickly test not only

whether the explanatory variables have a significant effect, but also to analyze the pairwise comparisons between subgroups of that variable.

3.2 Results

In this section, the influence of explanatory variables on shippers' emissions, attractions and shipments is analyzed. Beforehand, the dependent variables are described with a bit more detail.

Table 2: Dependent variables descriptive statistics

Variable	N	Min	Median	Mean	Max	Std
Generation volume	2935	1	4 600	52 773	6 414 000	235 716
(in natural logarithm)		0	8.434	8.299	15.654	2.641
Attraction volume	2935	1	4 614	38 588	7 500 000	187 421
(in natural logarithm)		0	8.437	8.179	15.830	2.628
Shipment frequency	2935	3	3 900	21 350	3 000 000	90 737
(in natural logarithm)		1.099	8.269	8.260	14.910	1.838

Table 2 shows that the distributions of the generation and attraction volumes and the shipment frequency span very wide ranges, and are extremely skewed. The generation and attraction distributions are relatively similar. By contrast, the logarithm distributions are symmetric, and, incidentally, the distribution of the logarithm of the shipment frequency is similar to the other two (although the standard deviation is substantially smaller). Finally, the normal qq-plots show that the distributions of the three variables are reasonably close to normal (as confirmed by the histograms in Fig 1).

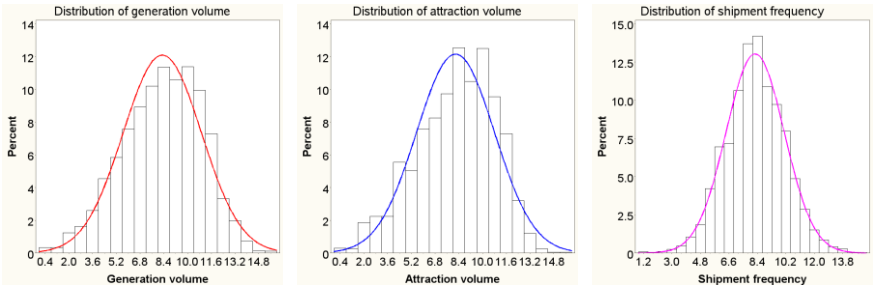


Fig. 1: The normality of distribution of the dependent continuous variables

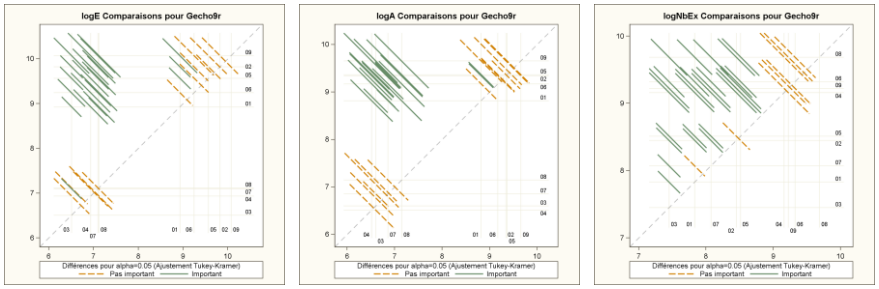
Economic activity

In France, the economic activity of an establishment is described on the basis of the NAP classification of the French Statistics Institute INSEE. This classification, which distinguishes 700 classes, was used to design the sampling pattern of the ECHO survey. In the ECHO database, these classes have been grouped in nine broad categories, to ensure significance:

1. Intermediate goods industry
2. Intermediate goods wholesale
3. Productive assets industry
4. Productive assets wholesale
5. Agri-food industry
6. Agri-food wholesale
7. Consumer goods industry
8. Consumer goods wholesale
9. Warehouses

One-way ANOVA shows that this classification has a significant effect on generation, attraction, and shipment frequency. Levene's test rejects the null hypothesis of homoscedasticity *it* with a p-value lower than 0.001. The Welch test is then applied, and rejects the hypothesis of equality of means.

In addition to this global conclusion, pairwise comparisons can be made. Figure 2 presents a diffogram of Tukey's multiple comparison adjustement, which allows to examine quickly and efficiently which groups differ and which do not. In a diffogram, each line corresponds to a pairwise comparison between two subgroups, indexed by the projection of the line's midpoint to the vertical and horizontal axes. Furthermore, the projection of each line on each axis allows us to obtain the corresponding confidence interval of the subgroups. Hence, if the line crosses the diagonal line, the difference is not significant. In that case, the line is orange and dotted. In the contrary case, the line is green and solid, and the difference between the two sub-groups is significant.



Generation volume *Attraction volume* *Shipment frequency*

Fig. 2: Pairwise comparison of the groups of shipper activities

Figure 2 shows that generation and attraction volumes share similarities. Two groups can be distinguished: the group including (03, 04, 07, 08) from the group including (01, 02, 05, 06, 09). Broadly speaking, shippers of the agri-food, intermediate goods or warehousing sectors behave similarly with respect to freight generation and attraction, and differently from shippers of the productive assets or consumer goods sectors. In both cases, industry and wholesale are grouped together.

In the case of the shipment frequency variable, the sub-populations are more segmented. There are similarities between activity groups 04, 06, 08, 09, i.e. wholesale and warehousing, except for intermediate goods wholesale.

The other economic variable in the ECHO dataset is the turnover. The exact turnover is not available: the ECHO dataset provides a categorical variable with nine levels. The ANOVA concludes that the four lower tiers of turnover are similar in terms of generation and attraction, while all the others are distinct from this first group and from one another.

Relations with economic agents

As explained above, the three variables examined here are the main type of contract between the shipper and its carriers (three values: long period contracts, occasional contracts, or both); the number of clients *Ncl* which constitute 80% of their activity; and the number of carriers or freight forwarders *CR* with which they worked during the year. For the first one, an ANOVA analysis is made. For the two others, correlations between them and the explanatory variables are calculated. The results are summarized below:

Table 3. Correlation (Pearson's coefficient) of the business relationships of the shipper and generation

Variable name	Variable signification	Generation volume	Attraction volume	Shipment frequency
TC	Type of contract with carrier	Significant influence	Significant influence	Significant influence
Ncl	Number of clients	Not correlated	Not correlated	9.7 %
Log(*Ncl*)	Number of clients (logarithmic scale)	Not correlated	Not correlated	30.5 %
CR	Number of carriers	29.40 %	29.70 %	16.60 %
log(*CR* + 1)	Number of carriers (logarithmic scale)	27.54 %	27.67 %	14.94 %

Again, freight generation and attraction are similar, but shipment frequency is different. In all three cases, the relationship between the type of contract and all the dependent variables is significant. This is the same for the number of carriers.

However, the number of clients making up to 80% of the shipper's turnover has no visible influence on freight generation and attraction. On the contrary, this variable is clearly correlated with shipment frequency. As a matter of fact, shipment frequency is much more closely related to the structure and constraints of the logistic chains than commodity flows measured in tons per year. For a given establishment, more clients means a more dispersed supply chain, with more destinations, and the need to send smaller and thus more frequent shipments. There is no such relationship between number of clients and commodity flows.

Production and logistics characteristics

The variables examined here are the number of distinct references *Nr*, and the share of transport cost in the product value *CT*.

Table 4. Correlation (Pearson's coefficient) between production and logistic characteristics and the explanatory variables.

Variable name	Variable signification	Generation volume	Attraction volume	Shipment frequency
Nr	Number of SKUs	Not correlated	Not correlated	19.60 %
log(*Nr*)		Not correlated	Not correlated	33.93 %
Npr	Number of product ranges	Weak correlation (< 5 %)	Weak correlation (< 5 %)	9.4 %
log(*NPr*)		Weak correlation	Weak correlation	12.8 %

		(< 5 %)	(< 5 %)	
CT	Transport cost share in total sale price	14.53 %	13.42 %	7.04 %
log(*CT*)		16.03 %	13.76 %	8.30 %

Table 4 shows, again the similarity between generation and attraction volumes: there is no clear relationship between them and the number of SKUs or the number of product ranges. However, there is a clear relationship between them and the share of transport costs in the product's value. Shipment frequency works very differently: more SKUs or product ranges clearly means more frequent shipments. From a logistic perspective, this is understandable: each SKU is distinct from the perspective of clients; and they are most often not easily substitutable. Therefore, each SKU needs its own supply chain, which means more frequent shipments than for a supply chain of homogenous products. Note shipments may be carried together; more frequent shipments do not necessarily mean more frequent vehicle movements (or at least not proportionally).

Employment

Two variables regard employment in the ECHO dataset: the number of employees, and a qualitative appreciation of their overall skill. The number of employees is strongly correlated to the generation and attraction volumes, as well as to the shipment frequency. An ANOVA analysis also concludes that the overall skill level has a significant influence on freight generation and attraction volumes. More precisely, a pairwise comparison of the different levels show that there are two groups: all the shippers where the employees are less than 'highly skilled', and the others. For the shipment frequency, the relationship is significant, but it is less easy to interpret the pairwise comparison.

As a conclusion to this section, the ECHO dataset contains a large amount of information about shippers, their economic activities, production, workforce, logistic characteristics, and relationships with other establishments and carriers. The analyses presented in this section help to draw first conclusions about the relationships between all these variables and the dependent variables of freight generation. Besides, the literature has shown that freight generation and freight trip generation work very differently (Holguin-Veras et al., 2014); this study shows that freight generation and shipment frequency also work very differently. This is not that surprising, given the fact that shipment frequency and freight trip generation are probably closely correlated.

4 Generation models

The second objective of the paper is the estimation of generation models. Two types of models are estimated: exploratory models, making the most of the information available in the ECHO dataset, and pragmatic models, using only variables which are expected to be reasonably easily available to a freight transport modeler. In each case, generation, attraction and shipment frequency are analyzed and compared.

In practice, three groups of models are examined: first, only quantitative exploratory variables are introduced. Second, quantitative and qualitative variables are both taken into account: in this category, the most complete specifications are examined. In the third category, simpler models are presented and discussed.

4.1 Models with quantitative explanatory variables

Regarding generation and the characteristics of establishments, the continuous variables in the ECHO database are the number of employees (log(*N*)), the number of SKUs (log(*Nr*)), and the number of clients (log(*Ncl*), number of carriers (*CR*), and share of transport costs in the commodity sale price (log(*CT)*).

Table 5 presents models for generation, attraction and shipment frequency. In each case, there are two models: one with the number of employees as an explanatory variable, and one with all the significant continuous variables.

Table 5. Generation models, quantitative explanatory variables

Estimated model	R^2
$log(E_i) = 4.63 + 0.86log(N_i)$	0.156
$log(E_i) = 3.95 + 0.77log(N_i) + 0.49log(CT_i) + 0.046CR_i$	0.223
$log(A_i) = 4.19 + 0.93log(N_i)$	0.183
$log(A_i) = 3.58 + 0.85log(N_i) + 0.45log(CT_i) + 0.044CR_i$	0.239
$log(S_i) = 5.98 + 0.53log(N_i)$	0.124
$log(S_i) = 4.20 + 0.52log(N_i) + 0.22log(Ncl_i) + 0.15log(Nr_i)$	0.281

The generation and the attraction models are similar: the same variables are significant, the coefficients share similar orders of magnitudes, and the R^2 are equivalent. In both cases, the commodity flows are a bit less than proportional to the number employees. Furthermore, generation and attraction increase with the share of transport costs in the products' sales price: intuitively, larger commodity flows imply higher transport costs, and this cost increase is not necessarily

compensated by an increase of the market price of these commodities. There is also a significant correlation between the number of carriers and the commodity flows, although the explanation is less clear. In both cases, the R^2 coefficient is rather low, just below 0.25.

The shipment frequency model differs strongly from the two other models. While it seems to be proportional to the number of employees according to the first model, this does not hold with the second, more complete specification. This is consistent with the theory and empirics about the relationship (or lack thereof) between commodity flow and shipment frequency, as theorized in Baumol and Vinod (1970) and explained in Holguin-Veras et al. (2014). In addition, in the second model, the other explanatory variables are the number of SKUs and the number of clients, two variables which, as explained above, are intimately related to the structure of the supply chain of the shipper. Both variables have a positive impact on shipment frequency. They also have a substantial explanatory power, bringing the R^2 up to 0.281 from 0.124.

4.2 Models with quantitative and qualitative explanatory variables

This section takes more complete models from the previous section and introduces the following qualitative variables: the economic activity sector *G*, the turnover category *T*, the labor qualification *L*, and finally the main type of contract between the shipper and its carriers *TC*.

These variables are introduced using the ANCOVA methodology, which means they modify the models' intercepts and the coefficient of explicative variable log(*N*). For all three models (generation, attraction and shipment frequency), the variables are introduced sequentially. The starting points are the models estimated in subsection 3.1. Tables 6, 7 and 8 report the models' R^2, the coefficient of the number of employees (on a logarithmic scale) and its share in the model variability, the contribution of the interactions between the qualitative variables to the model variability, and the number of non-significant subgroups, for generation, attraction and shipment frequency respectively.

Table 6. Generation models, quantitative and qualitative explanatory variables

	Linear regression (LR)	(LR) and G	(LR) and G*T	(LR) and G*T*L	(LR) and G*T*L*TC
Coefficient of $log(N_i)$	0.7735	0.9745	0.4970	0.4755	0.4418
Contribution Type 1 SS of $log(N_i)$		31.48 %	17.79 %	15.91 %	13.64 %
Contribution Type 1 SS of interactions		53.75 %	66.38 %	70.05 %	74.34 %
R^2	0.223	0.484	0.511	0.576	0.672

The introduction of the qualitative variables and their interactions increases significantly the model's explanatory power. The best generation model without these variables has a R^2 of 0.223; the R^2 jumps to 0.672 with all the qualitative variables. The most important improvement is due to the introduction of G, i.e. the economic sector (in the log specification, this is akin to modifying the model's slope with respect to G). With the introduction of turnover, the R^2 does not increase much, but the coefficient of log(*N*) decreases substantially: this is to be expected; both variables are correlated, and correlated with the economic activity of shippers, and thus to the amount of commodity they generate. Labor qualification is also very significant: establishments with unskilled workers have very different generation patterns than those with highly skilled workers. Finally, the type of contract bound with carriers CT also brings information about freight generation; although in this case the opportunity of using this variable in a simulation model is questionable: there is a real risk of endogenous bias (cf. Table 7).

Table 7. Attraction models, quantitative and qualitative explanatory variables.

	Linear regression (LR)	(LR) and G	(LR) and G*T	(LR) and G*T*L	(LR) and G*T*L*TC
Coefficient of $log(N)$	0.8507	1.0362	0.5803	0.5328	0.4940
Contribution Type 1 SS of $log(N)$		37.49 %	24.10 %	20.68 %	17.70 %
Contribution Type 1 SS of interactions		48.93 %	62.00 %	67.41 %	72.12 %
R^2	0.239	0.471	0.495	0.577	0.675

Once again, the attraction and the generation models behave in remarkably similar ways. The log(N) coefficients are consistently but marginally larger; the R^2 coefficients are very similar. The introduction of additional variables and interactions increases the models' explanatory power at the same pace.

Table 8. Shipment frequency models, quantitative and qualitative explanatory variables.

	Linear regression (LR)	(LR) and G	(LR) and G*T	(LR) and G*T*L	(LR) and G*T*L*TC
Coefficient of $log(N)$	0.5160	0.5918	0.3920	0.4042	0.3896
Contribution Type 1 SS of $log(N)$		33.58 %	22.43 %	18.68 %	14.70 %
Contribution Type 1 SS of interactions		23.62 %	33.74 %	44.77 %	56.55 %
R^2	0.281	0.369	0.373	0.445	0.566

The introduction of the qualitative variables does not increase the shipment frequency models' explanatory power as much as the generation and attraction ones, with a maximum R^2 at 0.57 instead of 0.67. However, the improvements brought by each new variable to the shipment frequency models are comparable, in relative terms, to those of the other two groups of models.

In all these models, the explanatory power comes at the cost of the introduction of a very large number of subgroups. G introduces 9 subdivisions; with the three other variables, there are 571 to 641 subgroups, depending on the model (a half to two thirds of these models are not significant). This raises the question of the model's robustness, and of its usefulness. The main conclusion of this part is that regularities can be found between freight generation and shipment frequency and variables regarding such different fields as economic activity, labor's level of skill, carriers contracts, and so on. Another conclusion is that employment and the type of activity are solid explanatory variables, fortunately often available, and are a good basis to build a pragmatic freight generation model.

4.3 Simple models

In this section, a third group of models is introduced. In order to develop models which can be used with limited data, the explanatory variables are limited to the sector of activity, and to the number of employees. In these models, the interaction between the number of employees and the sector of activity is examined. The estimations are reported in Table 9.

Table 9. Simple models (number of employees and sector of activity)

Variable	Generation	Attraction	Shipment frequency
Intercept	4.20***	3.68***	5.65***
1 (Intermediate good industry)	0.0033	0.20	-0.54*
2 (Intermediate good wholesale)	2.88***	2.47***	-0.027
3 (Productive asset industry)	-2.01***	-1.90***	-1.46***
4 (Productive asset wholesale)	-1.79*	-0.95	-1.45*
5 (Agri-food industry)	1.44***	2.19***	0.50
6 (Agri-food wholesale)	0.50	0.92	0.67
7 (Consumer good industry)	-2.48***	-2.61***	0.048
8 (Consumer good wholesale)	-0.25	-1.19	0.98°
Intercept	4.20***	3.68***	5.65***
log(N)	0.99***	1.08***	0.72***
1 x log(N)[a]	0.070	0.031	-0.11°
2 x log(N)	-0.26*	-0.23.	0.035
3 x log(N)	-0.016	0.019	0.016***
4 x log(N)	0.14	-0.13	0.52*
5 x log(N)	-0.080	-0.26**	-0.17
6 x log(N)	0.15	0.079	0.032***
7 x log(N)	0.19**	0.20**	-0.20
8 x log(N)	-0.20	0.11	0.045
# Observations	*2935*	*2935*	*2935*
R^2	*0.454*	*0.449*	*0.291*
Adjusted R^2	*0.451*	*0.445*	*0.287*

Significance levels: *** p-value < 0.001; ** p-value < 0.01; * p-value < 0.05; ° p-value < 0.1

[a] (and below) interaction between economic sector and number of employees : for example, the coefficient of log(N) in the generation model of the first economic sector is not significantly different from 0.99; the coefficient of the second economic sector is significantly lower.

From Table 9, a number of conclusions appear: first, generation and attraction can be considered as proportional to the number of employees, except for a few cases (generation increases more slowly in intermediate good wholesale, attraction in the agri-food industry; both increase faster in the consumer good industry).

Second, this is not the case for shipment frequency. Shipment frequency increases less than proportionately to the number of employees. There are significant differences in the productive asset industry and in the agri-food wholesale sector, but the orders of magnitude of the coefficients are similar. Third, generation and attraction are, once more, similar, and the R^2 coefficient is acceptable, at 0.45. The situation is far less satisfying in the shipment frequency model. The last model loses a lot of information compared with the models in Section 4.2; however those models rely on measures which are usually not available.

5 Conclusion

This study took the opportunity offered by the French shipper survey to estimate a disaggregate freight generation model, with a distinction of generation, attraction and shipment frequency. It confirmed that while generation and attraction work in similar ways, this is not the case of shipment frequency, which is not driven by the same economic and logistic mechanisms. Three categories of models were presented, illustrating the potential of using rich datasets to model statistically the behavior of shippers, but also the limitations of models relying on less variables.

The ECHO dataset contains variables that are usually unavailable. This study examined how they impacted statistically the dependent variables. Consistently with the literature, the number of employees and the economic sector were identified as very important explanatory variables. However, other variables also have a substantial explanatory power, such as the share of transport prices in products' sales price on generation and attraction, or the number of product ranges and product references (SKUs) on shipment frequency.

This study is part of an ongoing work, of which the next step is to increase the accuracy of the economic segmentation, and to also introduce a distinction of the types of generated and attracted products (the models developed in this paper only consider the tons sent and received without distinguishing commodity types). In the long term, the objective is to use these results to disaggregate French aggregate freight generation and attraction data at fine spatial levels.

Acknowledgments This study is funded by the CGDD, the public office of sustainable development in France in the French Misnistry of Transport. The data used in this paper is supplied by the French Institut IFSTTAR, who has conducted the survey and processed the data. The authors would like to thank IFSTTAR for providing the ECHO dataset.

References

Baumol, W. J., & Vinod, H. D. (1970). An inventory theoretic model of freight transport demand. Management science, 16(7), 413-421.

Chin S.-M. and Hwang H.-L. (2007) National freight demand modeling – bridging the gap between freight flow statistics and U.S. economic patterns, in *Proceedings of the Annual Meeting of the Transportation Research Board*, Washing-ton D.C., U.S.A.

de Jong G. and Ben-Akiva M. (2007) A micro-simulation model of shipment size and transport chain choice, *Transportation Research Part B*, 41 pp. 950-965.

Guilbault M., Soppe M. (2009) *Apports des enquêtes chargeurs – Connaissance des chaînes de transport de marchandises et de leurs déterminants logistiques*, Actes INRETS N°121,

Holguin-Veras, J., Jaller, M., Sanchez-Diaz, I., Wojtowicz, J., Campbell, S., Levinson, D. M., et al. (2012a). *Freight trip generation and land use*. National Cooperative Freight Research Program, 19. from ‚http://onlinepubs.trb.org/onlinepubs/ncfrp/ncfrp_rpt_019.pdf.

Holguin-Veras, J., Jaller M., Sanchez-Diaz I., Campbell S., and Lawson C. T. (2014) Freight generation and freight trip generation models. In *Modelling Freight Transport*, L . Tavasszy, G. de Jong (eds.), London: Elsevier, 2014.

Novak D.C., Hodgdon C., Guo F., and Aultman-Hall L. (2011) Nationwide Freight Generation Models: A spatial Regression approch, in *Network and Spatial Economics*, 11:23-41.

Rizet, C. and Hémery, C. (2008) Génération de trafic. In *Enquête ECHO «Envois – Chargeurs - Opérateurs de transport », Résultats de référence*. Synthèse INRETS n°58.

Tenenhaus, M. (1996) *Méthodes statistiques en Gestion*, Dunod, 1996.

Welch B.L. (1951) On the Comparison of Several Mean Values: An Alternative Approach, *Biometrika* 38, p 330-336

Introduction to Logistic Engineering, edited by G. Don Taylor, CRC Press, 2008