Forthcoming in Astin Bulletin

Modelling and Estimating Individual and Firm Effects with Count Panel Data*

Jean-François Angers[a], Denise Desjardins[b]
Georges Dionne[b] and François Guertin[c]

9 January 2018

Abstract

We propose a new parametric model for the modelling and estimation of event distributions for individuals in different firms. The analysis uses panel data and takes into account individual and firm effects in a non-linear model. Non-observable factors are treated as random effects. In our application, the distribution of accidents is affected by observable and non-observable factors from vehicles, drivers, and fleets of vehicles. Observable and unobservable factors are significant to explain road accidents, which means that insurance pricing should take into account all these factors. A fixed effects model is also estimated to test the consistency of the random effects model.

Keywords: Individual effect, firm effect, non-linear model, panel data, Poisson, Dirichlet, insurance pricing, R code.

JEL codes: C23, C25, C55, G22.

[a] Department of Mathematics and Statistics, University of Montreal, Montreal
[b] Canada Research Chair in Risk Management, HEC Montréal, Montreal
[c] Calcul Québec, University of Montreal, Montreal

Contact author: Georges Dionne, Canada Research Chair in Risk Management, HEC Montréal, 3000, Chemin de la Côte-Sainte-Catherine, room 4.454, Montreal (Qc), Canada, H3T 2A7. georges.dionne@hec.ca.

*Acknowledgements
We thank the Société de l'assurance automobile du Québec for providing the data and Helmi Jedidi for his very useful comments. We also thank Victoire Michal for her excellent assistance in programming the R code.

1. INTRODUCTION

Since the early 1980s, several researchers have proposed different models to account for correlation resulting from temporal repetitions of observations. Indeed, the use of panel-type individual data has become popular in many economic applications in the fields of labor economics, health economics, firm productivity, patents, transportation, and education (Hausman and Wise, 1979; Gourieroux et al, 1984; Hausman et al, 1984; Cameron and Trivedi, 1996; Hsiao, 1986; Baltagi, 1995; Dionne et al, 1997, 1998). In the domain of count-data applications, the ground-breaking contribution is the article of Hausman et al (1984) that proposes a Maximum Likelihood Method (MLE) for estimating the parameters.[1] In this article we extend Hausman et al's (1984) parametric model to add a firm effect to the individual effect in the estimation of event distributions, and we apply the model to the accident distributions of trucks belonging to fleets of vehicles.[2]

To our knowledge there is no non-linear econometric model in the literature that estimates individual and firm effects with panel data. The matching of longitudinal individual and firm data is very important in environments where the observed results (here accidents) are a function of both parties' characteristics (here, individual and firm) and unobserved actions. For insurance companies, knowing all the sources of accidents involving vehicles belonging to a fleet is essential to develop a fair pricing scheme that takes into account the negligence of each actor. This is also important for the regulator, which has to compute the optimal fines of different infractions (driver, fleet owner) that affect accident distributions (Fluet, 1999).

In our application, we estimate the distribution of vehicle accidents for different fleets over time, by first decomposing the explanatory factors into heterogeneous factors linked to vehicles and their drivers, then into heterogeneous factors linked to fleets, and finally into residual factors.

[1] See also Gouriéroux et al (1984) who propose a pseudo-MLE treatment of the data.

[2] On insurance applications with non-parametric models, see Pinquet (2013), Fardilha et al (2016), and Desjardins et al (2001). On accident distribution estimation or insurance pricing see Purcaru and Denuit (2003), Boucher, Denuit, and Guillen (2008), Boucher and Denuit (2006), Angers et al (2006), Frangos and Vrontos (2001), Frees and Valdez (2011), and Cameron and Trivedi (2013a). Another class of models uses the hierarchical credibility approach with random effects in linear models (Norberg, 1986). It can be shown that the Negative Binomial model is a kind of hierarchical model (see Section 2.2 of this article). All these contributions do not consider separately trucks and fleets effects.

Factors linked to vehicles and drivers and those linked to fleets can be correlated. For example, a negligent manager may not spend enough money on mechanical repair of his trucks and might ask his employees to drive too fast. However, the employees may also exceed the speed limit without informing the manager.

As mentioned elsewhere, the modelling of Hausman et al (1984) is not directly applicable to the ex-post calculation of insurance premiums using a Bayesian model (Angers et al, 2006). However the extended model we propose can be used for the insurance pricing of vehicles that includes individual and firm effects. Our model can also be applied to any count modelling with random individual and common effects. We may think of different principal-agent output such as operational risk events in banks, innovations in teams, deaths in hospitals, airline accidents, or any other event involving many agents working for different principals under asymmetric information (Holmstrom, 1982; Laffont and Martimort, 2001).

This research uses parametric models exclusively. First, we want to compare our results with those of Hausman et al (1984), who use parametric estimation methods in their study. It is well known that parametric models involve explicit assumptions about the statistical distribution of the data and, thus, hypothesis testing involves estimation of the key parameters of the chosen distribution. Given that nonparametric models are distribution-free, they could be applied in a broader range of situations even where the parametric conditions of validity are not met. In our case we study accident distributions. The parametric Poisson family is known to satisfy the necessary conditions of validity for accident data.

Another advantage of the nonparametric test is its ability to handle various data types even if measured imprecisely or if they comprise outliers, anomalies widely known to seriously affect parametric tests. However, the foremost advantage of using parametric models is the statistical power of the estimations when the assumptions are satisfied. Under these circumstances, parametric tests produce more accurate and precise estimates than do nonparametric tests. Therefore, since our data set meets the sample size requirements, is very precise, and does not contain outliers that could not safely be removed from the dataset, we find it reasonable to consider the above statistical power argument as a third argument in favor of the use of parametric models.

Lastly, parametric models are very convenient when we wish to obtain close-form expressions for risk and premium forecasting.

In section 2, we propose a short literature review of count data models, and section 3 develops our econometric model. Sections 4 and 5 present the data and the results of our estimations. We also analyze our results based on various statistical performance criteria, including accidents prediction for the next year. Finally, we compare our random effects estimators with those obtained from a fixed effects model and we test the consistency of the random effects model. Section 6 concludes the paper.

2. LITERATURE

2.1 BASIC COUNT DATA MODELS

Our presentation is based on trucks accidents but the model can be applied to any other event involving count data. Most of the econometric models applied to count variables that takes nonnegative values start from the Poisson distribution, where the probability of truck *i* of fleet *f* being involved in y_{fit} observable accidents (or claims) in period *t* is estimated.

By definition of the Poisson law, the mathematical expectation of the number of accidents is equal to the variance, $E(Y_{fit}) = Var(Y_{fit}) = \lambda_{fit}$ where Y_{fit} is the random variable representing the number of accidents of truck *i*, fleet *f* in period *t* and $\lambda_{fit}(>0)$ is the Poisson parameter equal to the mean and the variance of the distribution. In fact, the parameter $\lambda_{fit} = \exp(X_{fit}\beta)$, where the vector $X_{fit} = (x_{fit1}, \cdots, x_{fitp})$ represents the *p* characteristics of truck *i* of fleet *f* observed in period *t* and β is a vector of parameters to be estimated. The exponential form of λ_{fit} introduces a non-linear relationship between accidents and control variables included in X_{fit}. X_{fit} can contain continuous variables and such variables can be non-linear. For example, x_{fit2} can be the kilometers driven and x_{fit3} the square of the kilometers driven. Moreover, X_{fit} can contain categorical variables with a fixed number of possible values such as size of the fleet or number of traffic violations. These variables can also introduce non-linear effects between accidents and different observable categories. All these characteristics are applicable in the usual Generalized Linear Model (GLM) setting. Poisson model can also incorporate data that are collected spatially by introducing a spatial autocorrelation term or in a Generalized Additive Model (GAM) setting by adding smooth functions. It is not clear however that such extensions would significantly improve our results for the type of basic data we used where spatial or strong non-linear effects are absent.

The Poisson model is an equidispersion model. This modelling implicitly supposes that the distribution of accidents can be explained entirely by observable heterogeneity. To take into account the overdispersion property in the data, we can suppose that the parameter λ_{fit} has a random term such that $\lambda_{\mathrm{fit}} = e^{X_{\mathrm{fit}}\beta+\varepsilon_i} = \alpha_i\gamma_{\mathrm{fit}}$ with $\alpha_i = e^{\varepsilon_i}$ and $\gamma_{\mathrm{fit}} = e^{X_{\mathrm{fit}}\beta}$ and where α_i is the random individual specific effect for truck *i*. Suppose that α_i follows a gamma distribution of parameter $\left(\delta^{-1},\delta^{-1}\right)$, we obtain the negative binomial distribution[3] (NB2):

$$P\left(Y_{\mathrm{fit}} \mid \gamma_{\mathrm{fit}},\delta\right) = \frac{\Gamma\left(\delta^{-1}+y_{\mathrm{fit}}\right)}{\Gamma\left(\delta^{-1}\right)\Gamma\left(y_{\mathrm{fit}}+1\right)}\left(\frac{\delta^{-1}}{\delta^{-1}+\gamma_{\mathrm{fit}}}\right)^{\delta^{-1}}\left(\frac{\gamma_{\mathrm{fit}}}{\delta^{-1}+\gamma_{\mathrm{fit}}}\right)^{y_{\mathrm{fit}}}, \qquad (1)$$

where $\Gamma(\bullet)$ is the gamma function. The mean remains equal to $\exp\left(X_{\mathrm{fit}}\beta\right)$ and the variance to mean ratio is equal to $(1+\delta)/\delta$. (Hausman et al, 1984; Cameron and Trivedi, 1986; Boyer, Dionne, and Vanasse, 1992). This modelling does not simply allow for overdispersion. It also lets us consider unobserved or latent heterogeneity that is absent from the Poisson model. Unobserved heterogeneity is very important for pricing insurance premiums under asymmetric information (Dionne and Vanasse, 1989, 1992). The above modelling is appropriate for independent observations, meaning those without individual and time effects, and cannot be appropriate for panel data. [4]

[3] For an analysis of the Poisson lognormal mixture see Greene (2005).

[4] Note that the Poisson model can also be estimated with panel data. We do not consider this possibility here. See Cameron and Trivedi (2013b) for a detailed analysis.

2.2 Taking time into account

Let us now consider data that contain observations where the same unit (individual or truck, for example) is observed over several successive periods but without firm or group effects. There are two treatments for panel data in the literature, the fixed effects and the random effects model. In this section we limit our discussion to the random effects Negative Binomial (NB) model applied to short periods of time where the number of periods is fixed and the number of individuals is large. Hausman et al (1984) propose an extension of the model expressed by equation (1), which is not designed to take into account repetitions of observations over time. The new model is a hierarchical model that comes directly from the Poisson model. Thus Y_{fit} would be distributed according to the NB2 model of parameters $\alpha_i \gamma_{fit}$ and ϕ_i, where α_i and ϕ_i vary across individuals. α_i is the random firm specific effect and ϕ_i is an additional random effect that permit the random firm specific effect to vary over time (Hausman et al, 1984). Suppose that $\left(1+\alpha_i/\phi_i\right)^{-1}$ follows a beta distribution of parameters (a,b), we obtain a closed form solution for the random effects negative binomial model:

$$P\left(Y_{fi1},\cdots,Y_{fiT_i}\right)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma\left(a+\sum_t \gamma_{fit}\right)\Gamma\left(b+\sum_t y_{fit}\right)}{\Gamma\left(a+b+\sum_t \gamma_{fit}+\sum_t y_{fit}\right)}\prod_t \frac{\Gamma\left(y_{fit}+\gamma_{fit}\right)}{\Gamma\left(\gamma_{fit}\right)\Gamma\left(y_{fit}+1\right)}. \tag{2}$$

The NB2 model can also be estimated with individual dummies (or other methods) in a fixed effects version. The β parameters can be inconsistent however because of the incidental parameters problem, but some contributions have shown that the inconsistency may be not important (Allison and Waterman, 2002; Green, 2004).

Estimating the random effects model in (2) can also yield inconsistent random effects estimators because α_i and the vector of observable individual characteristics may be correlated. We can apply the Hausman (1978) test statistic to determine whether or not we should reject the null hypothesis that the individual effects are not correlated with the variables in the regression component. The model in (2) is suitable for estimating parameters with individual effects but cannot take into account the firm or the fleet effect when individual observations belong to different firms with common characteristics that can affect accident distributions.

3. Methodology: Taking time and firm effects into account simultaneously

We now move on to the generalization of the model, which will allow us to account, simultaneously, for the individual effect, the firm effect and the time effect.[5] We are interested in observations that have common characteristics because they belong to the same firm, for example: workers in a firm, vehicles in a fleet, patients in a hospital or children attending the same school.

We consider a set $I=\left\{1,...,I^{\max}\right\}$ of individuals, a partition $\left\{I_1,...,I_f,...,I_F\right\}$ of I, a set $T=\left\{1,...,T^{\max}\right\}$ of dates and a collection $\left\{T_i \,|\, i\in I\right\}$ where T_i is a subset of T. I_f is the number of trucks in fleet *f*. For each *i* we may refer to it as $1,2,...,T_i$, keeping in mind that $T_i=1$ and $T_j=1$ does not necessarily refers to the same element of T whenever $i\neq j$. The vector $X_{fit}=\left(x_{fit1},\cdots,x_{fitk},\cdots,x_{fitp}\right)$ still represents the *p* characteristics of individual *i* from firm *f* observed in period *t*. Here we can have many different firms over a given number of periods. For example, the vector may contain specific information about the vehicle or the driver and other specific information about the fleet. β is a vector of *p* parameters to be estimated. Let α_f be the random effects associated with fleet *f* (i.e. the risk or non-observable characteristics attributable to the fleet), whereas $\theta_{(f)i}$ is the random effects of truck *i* of fleet *f* where $\sum_{i=1}^{I_f}\theta_{(f)i}=1$, I_f being the number of vehicles in fleet *f*. Finally $\eta_{(fi)t}$ is the time random effects of period *t* of truck *i* of fleet *f* such that $\sum_{t=1}^{T_i}\eta_{(fi)t}=1$ where T_i is the number of periods for truck *i*. Our model has an embedded structure which explains why the two above summations are equal to one. The random variable α_f is independent of other regressors including the $x_{fitk}, k=1,\cdots,p,$ while the $\theta_{(f)i}$ are dependent between themselves for a given fleet and the $\eta_{(fi)T}$ are

[5] Angers et al (2016) propose a model with individual and firm effects but their model cannot be applied to panel data.

dependent between themselves for the truck *i* of a given fleet *f*. Finally, the periods in T_i are not necessarily consecutive. An individual or a truck can leave the firm and come back.

Model assumptions: Let us suppose that $\lambda_{fit} = \gamma_{fit}\left(\alpha_f \theta_{(f)i} \eta_{(fi)t}\right) > 0$ with $\gamma_{fit} = e^{X_{fit}\beta}$. We posit that: 1) α_f follows a gamma distribution of parameters $\left(\sum_{i=1}^{I_f} T_i \, \kappa^{-1}, \kappa^{-1}\right)$; 2) the vector $\theta_{(f)} = \left(\theta_{(f)1}, \theta_{(f)2}, \cdots, \theta_{(f)I_f}\right)$ follows a Dirichlet distribution of parameters ($\nu_{(f)1}, \nu_{(f)2}, \cdots, \nu_{(f)I_f}$); and 3) the vector $\eta_{(fi)} = (\eta_{(fi)1}, \eta_{(fi)2}, \cdots, \eta_{(fi)T_i})$ follows a Dirichlet distribution of parameters $\left(\delta_{(fi)1}, \delta_{(fi)2}, \cdots, \delta_{(fi)T_i}\right)$ where T_i is the number of periods of vehicle *i*.

The Dirichlet distributions have been proposed because they naturally generalizes the beta distribution already used in the Negative Binomial model. They allow us to distribute the whole fleet effects on all trucks. Moreover, they permit to obtain an analytical solution for the model. It is clear that we can use other distributions than the Dirichlet for that purpose. Suppose that the variable X_i on $\mathbb{R}^+$ for $i = 1, ..., n$ follows any density. Then $\frac{X_i}{\sum_i X_i}, ..., \frac{X_n}{\sum_i X_i}$ will have the same properties as the Dirichlet.

Using a general distribution will add non-necessary complexities, however. Suppose, for example, that $Z = \frac{X_1}{X_1 + X_2}$ with X_i following an uniform distribution over the interval (0,1). We will obtain the following distribution:

$$f(z) = \frac{0.5}{(1-z)^2} \quad \text{if} \quad 0 < z \leq 0.5$$
$$f(z) = \frac{0.5}{z^2} \quad \text{if} \quad 0.5 < z < 1.$$

Even with such an easy case, computations will become much more complex than by using a Dirichlet distribution.

Proposition: The joint distribution of accidents of all vehicles in fleet *f* is given by:

$$P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\right)=\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\frac{\left(\gamma_{fit}\right)^{y_{fit}}}{\Gamma\left(y_{fit}+1\right)}\right]\left[\frac{\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)}{\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)}\right]\left[\frac{\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f}T_i\kappa^{-1}}}{\left(\kappa^{-1}+\overline{\gamma}_{g_2}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}}\right]\left[\frac{\Gamma\left(\sum_{i=1}^{I_f}\nu_{(f)i}\right)}{\Gamma\left(\sum_{i=1}^{I_f}S_i+\nu_{(f)i}\right)}\right]\left[\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)}{\prod_{i=1}^{I_f}\Gamma\left(\nu_{(f)i}\right)}\right]$$
$$\times\left[\frac{\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(y_{fit}+\delta_{(fi)t}\right)}{\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(\delta_{(fi)t}\right)}\right]\left[\frac{\prod_{i=1}^{I_f}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}{\prod_{i=1}^{I_f}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}\right]\times{}_2F_1\left(\sum_{i=1}^{g_1}\left(S_i+\nu_{(f)i}\right),S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1},\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right),\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right). \tag{3}$$

Proof: The conditional distribution of the number of accidents for all the vehicles in fleet *f* is given by:

$$\begin{aligned}P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\alpha_f,\theta_{(f)},\eta_{(fi)}\right)&=\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}P\left(Y_{fit}\mid\lambda_{fit}\right)\\&=\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\frac{e^{-\lambda_{fit}}\left(\lambda_{fit}\right)^{y_{fit}}}{\Gamma\left(y_{fit}+1\right)}\\&=\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\frac{1}{\Gamma\left(y_{fit}+1\right)}\right)\right]\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\lambda_{fit}\right)^{y_{fit}}\right]e^{-\sum_{i=1}^{I_f}\sum_{t=1}^{T_i}\lambda_{fit}}.\end{aligned} \tag{4}$$

Since $\lambda_{fit}=\gamma_{fit}\left(\alpha_f\theta_{(f)i}\eta_{(fi)t}\right)$ then:

$$\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\lambda_{fit}\right)^{y_{fit}}=\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\gamma_{fit}\right)^{y_{fit}}\right]\left[\left(\alpha_f\right)^{\sum_{i=1}^{I_f}\sum_{t=1}^{T_i}y_{fit}}\right]\left[\prod_{i=1}^{I_f}\left(\theta_{(f)i}\right)^{\sum_{t=1}^{T_i}y_{fit}}\right]\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\eta_{(fi)t}\right)^{y_{fit}}\right] \tag{5}$$

Let $S_i=\sum_{t=1}^{T_i}y_{fit}$ and $S_0=\sum_{i=1}^{I_f}\sum_{t=1}^{T_i}y_{fit}=\sum_{i=1}^{I_f}S_i$, equation (5) can be rewritten as follows

$$\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\lambda_{fit}\right)^{y_{fit}}=\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\gamma_{fit}\right)^{y_{fit}}\right]\left[\left(\alpha_f\right)^{S_0}\right]\left[\prod_{i=1}^{I_f}\left(\theta_{(f)i}\right)^{S_i}\right]\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\eta_{(fi)t}\right)^{y_{fit}}\right].$$

Moreover, the summation $\sum_{i=1}^{I_f}\sum_{t=1}^{T_i}\lambda_{fit}$ in equation (4) can be written as

$$\alpha_f\sum_{i=1}^{I_f}\theta_{(f)i}\sum_{t=1}^{T_i}\gamma_{fit}\eta_{(fi)t}\,.$$

Writing the general form of the joint distribution of the number of accidents for all the vehicles in fleet *f* as:

$$P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\right)=\int\cdots\int P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right)f\left(\eta_{(fi)}\right)d\eta_{(fi)} \tag{6}$$

with
$$P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right)=\int\cdots\int P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\theta_{(f)},\eta_{(fi)}\right)f\left(\theta_{(f)}\right)d\theta_{(f)} \tag{7}$$

and
$$P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\theta_{(f)},\eta_{(fi)}\right)=\int_0^\infty P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\alpha_f,\theta_{(f)},\eta_{(fi)}\right)f\left(\alpha_f\right)d\alpha_f\,, \tag{8}$$

and integrating equation (8) with respect to α_f, we obtain

$$\frac{\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\frac{\left(\gamma_{fit}\right)^{y_{fit}}}{\Gamma\left(y_{fit}+1\right)}\right]\left[\prod_{i=1}^{I_f}\left(\theta_{(f)i}\right)^{S_i}\right]\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\eta_{(fi)t}\right)^{y_{fit}}\right]\left[\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right]\left[\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]}{\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\left[\left(\kappa^{-1}+\sum_{i=1}^{I_f}\theta_{(f)i}\sum_{t=1}^{T_i}\gamma_{fit}\eta_{(fi)t}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right]}. \tag{9}$$

By replacing $P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\theta_{(f)},\eta_{(fi)}\right)$ in equation (7) by its value given in (9) and by replacing the density function $f\left(\theta_{(f)}\right)$by the density of a parametric Dirichlet distribution of parameters $\left(\nu_{(f)1},\nu_{(f)2},\cdots,\nu_{(f)I_f}\right)$, we obtain the following expression:

$$\begin{aligned}&P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right)\\&=\frac{\left[\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]\left[\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right]\left[\Gamma\left(\sum_{i=1}^{I_f}\nu_{(f)i}\right)\right]\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\gamma_{fit}\eta_{(fi)t}\right)^{y_{fit}}\right]}{\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(y_{fit}+1\right)\right]\left[\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]\left[\prod_{i=1}^{I_f}\Gamma\left(\nu_{(f)i}\right)\right]}\int\cdots\int\frac{\prod_{i=1}^{I_f}\left(\theta_{(f)i}\right)^{S_i+\nu_{(f)i}-1}}{\left(\kappa^{-1}+\sum_{i=1}^{I_f}\theta_{(f)i}\sum_{t=1}^{T_i}\gamma_{fit}\eta_{(fi)t}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}}d\theta_{(f)}\end{aligned} \tag{10}$$

We must estimate the multidimensional integral of equation (10) to obtain the model parameters. We analyze two possibilities.

i) All trucks of the same fleet have identical *a priori* risk

This first possibility, which greatly simplifies the estimations, is to suppose that all the γ_{fit} of the I_f vehicles are identical and equal to γ_f, for all periods. Under this hypothesis, the multidimensional integral of equation (10) is reduced to:

$$\int\cdots\int \frac{\prod_{i=1}^{I_f}\left(\theta_{(f)i}\right)^{S_i+\nu_{(f)i}-1}}{\left(\kappa^{-1}+\gamma_f\sum_{i=1}^{I_f}\theta_{(f)i}\sum_{t=1}^{T_i}\eta_{(fi)t}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}}\,d\theta_{(f)} = \frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)}{\left(\left(\kappa^{-1}+\gamma_f\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right)\Gamma\left(\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)\right)} \qquad (11)$$

and the joint distribution of the number of accidents at period *t* for the I_f vehicles in fleet *f* is given by the following expression:

$$P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right) = \frac{\left[\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]\left[\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right]\left[\Gamma\left(\sum_{i=1}^{I_f}\nu_{(f)i}\right)\right]\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\gamma_{\mathrm{fit}}\eta_{(fi)t}\right)^{y_{\mathrm{fit}}}\right]}{\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(y_{\mathrm{fit}}+1\right)\right]\left[\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]\left[\prod_{i=1}^{I_f}\Gamma\left(\nu_{(f)i}\right)\right]}\,\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)}{\left(\left(\kappa^{-1}+\gamma_f\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right)\Gamma\left(\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)\right)} \qquad (12)$$

Further, by replacing $P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right)$ in equation (6) by the expression in (12) and by replacing the density function $f\left(\eta_{(fi)}\right)$by the density of a parametric Dirichlet distribution of parameters ($\delta_{(fi)1},\delta_{(fi)2},\cdots,\delta_{(fi)T_i}$), we obtain the following approximation for (6), the joint distribution of the number of accidents for all the vehicles in all fleets:

$$P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\right)=$$

$$\left[\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\frac{\left(\gamma_{fit}\right)^{y_{fit}}}{\Gamma\left(y_{fit}+1\right)}\right]\frac{\left[\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]\left[\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right]\left[\Gamma\left(\sum_{i=1}^{I_f}\nu_{(f)i}\right)\right]}{\left[\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\right]\left[\prod_{i=1}^{I_f}\Gamma\left(\nu_{(f)i}\right)\right]}\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)}{\left(\left(\kappa^{-1}+\gamma_f\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right)\Gamma\left(\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)\right)} \quad (13)$$

$$\times\left[\frac{\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(y_{fit}+\delta_{(fi)t}\right)}{\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(\delta_{(fi)t}\right)}\right]\left[\frac{\prod_{i=1}^{I_f}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}{\prod_{i=1}^{I_f}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}\right]$$

This is an approximation because the main working hypothesis for this first scenario supposes implicitly that all the vehicles in the fleet represent identical *a priori* risks, which is probably a very strong hypothesis because, as we shall see, several variables distinguishing vehicles and driver behavior are significant in estimating the probabilities of accidents of different vehicles. Another possibility is to divide the vehicles into homogeneous risk groups, as insurers do when classifying risks.

ii) Trucks belong to different groups

Under this second possibility, we suppose that $\gamma_{fit}=\overline{\gamma}_{fi}\ \forall t=1,\cdots,T_i$ where $\overline{\gamma}_{fi}=\frac{1}{T_i}\sum_{i=1}^{T_i}\gamma_{fit}$. We can separate the vehicles into two groups (high risk and low risk) and define $G_1=1,\cdots,g_1$ as the set of all vehicles of the first group with $\overline{\gamma}_{g_1}=\frac{\sum_{i=1}^{g_1}\overline{\gamma}_{fi}}{g_1}$, and $G_2=g_1+1,\cdots,I_f$, as the set of all vehicles of the second group with $\overline{\gamma}_{g_2}=\frac{\sum_{i=g_1+1}^{I_f}\overline{\gamma}_{fi}}{g_2}$.

The integral of equation (10) thus becomes:

$$\int\cdots\int \frac{\left[\prod_{i=1}^{g_1}\left(\theta_{(f)i}\right)^{S_i+\nu_{(f)i}-1}\right]\left[\prod_{i=g_1+1}^{I_f}\left(\theta_{(f)i}\right)^{S_i+\nu_{(f)i}-1}\right]}{\left(\kappa^{-1}+\overline{\gamma}_{g_1}\sum_{i=1}^{g_1}\theta_{(f)i}+\overline{\gamma}_{g_2}\sum_{i=g_1+1}^{I_f}\theta_{(f)i}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}}\,d\theta_{(f)}\,. \qquad (14)$$

Let $v=\sum_{i=1}^{I_f}\theta_{(f)i}$, $u_i=\frac{\left(\theta_{(f)i}\right)^{\chi_{1i}}}{v}$ and $w_i=\frac{\left(\theta_{(f)i}\right)^{\chi_{2i}}}{1-v}$. $\chi_{si}=1$ if truck *i* belongs to group *s* $\left(s=1,2\right)$ and 0 otherwise. The integral of equation (14) can be rewritten as follows:

$$\int\cdots\int\frac{\left[\left[\prod_{i\neq g_1}\left(vu_i\right)^{S_i+\nu_{(f)i}-1}\right]\left[\prod_{i\neq I_f}\left(\left(1-v\right)w_i\right)^{S_i+\nu_{(f)i}-1}\right]\left[v\left(1-\sum_{i\neq g_1}u_i\right)\right]^{S_{g_1}+\nu_{(f)g_1}-1}\left[\left(1-v\right)\left(1-\sum_{i\neq I_f}w_i\right)\right]^{S_{I_f}+\nu_{(f)I_f}-1}\right]}{\left(\left(\kappa_f^{-1}+\overline{\gamma}_{g1}\right)v+\left(\kappa_f^{-1}+\overline{\gamma}_{g2}\right)\left(1-v\right)\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}}v^{g_1-1}\left(1-v\right)^{I_f-g_1-1}dudwdv.$$

By integrating we obtain:

$$=\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)}{\left[\Gamma\left(\sum_{i=1}^{I_f}S_i+\nu_{(f)i}\right)\right]\left[\left(\kappa^{-1}+\overline{\gamma}_{g_2}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}\right]}\,{}_2F_1\left(\sum_{i=1}^{g_1}S_i+\nu_{(f)i},S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1},\sum_{i=1}^{I_f}S_i+\nu_{(f)i},\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right). \qquad (15)$$

Thus, by replacing the integral in equation (14) by its value given in (15) we obtain the following approximation for $P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right)$ in (10):

$$\frac{\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f}T_i\kappa^{-1}}\Gamma\left(\sum_{i=1}^{I_f}\nu_{(f)i}\right)\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\left(\gamma_{fit}\eta_{(fi)t}\right)^{y_{fit}}}{\prod_{i=1}^{I_f}\prod_{t=1}^{T_i}\Gamma\left(y_{fit}+1\right)\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)\prod_{i=1}^{I_f}\Gamma\left(\nu_{(f)i}\right)}\left(\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)}{\Gamma\left(\sum_{i=1}^{I_f}S_i+\nu_{(f)i}\right)}\right)\frac{1}{\left(\kappa^{-1}+\overline{\gamma}_{g_2}\right)^{S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}}}$$
$$\times{}_2F_1\left(\sum_{i=1}^{g_1}S_i+\nu_{(f)i},S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1},\sum_{i=1}^{I_f}S_i+\nu_{(f)i},\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right) \qquad (16)$$

where ${}_2F_1$ is a hypergeometric function whose value is equal to:

$$1+\sum_{\ell=1}^{\infty}\left[\frac{\left(\sum_{i=1}^{g_1} S_i+\nu_{(f)i}\right)^{[\ell]}\left(S_0+\sum_{i=1}^{I_f} T_i\kappa^{-1}\right)^{[\ell]}}{\left(\sum_{i=1}^{I_f} S_i+\nu_{(f)i}\right)^{[\ell]}}\frac{\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)^{\ell}}{\ell!}\right], \tag{17}$$

with $h^{[\ell]}=h\left(h+1\right)\cdots\left(h+\ell+1\right)$, being an increasing factorial function.

Further, by replacing $P\left(Y_{f11},\cdots,Y_{fI_fT_{I_f}}\mid\eta_{(fi)}\right)$ in equation (6) by the expression in (16) and by replacing the density function $f\left(\eta_{(fi)}\right)$by the density of a parametric Dirichlet distribution of parameters ($\delta_{(fi)1},\delta_{(fi)2},\cdots,\delta_{(fi)T_i}$), we obtain (3), which completes the proof.

This procedure in estimating the integral can be generalized to several homogeneous groups, but it is not obvious that the precision gained would be greater than that corresponding to a Monte Carlo approximation of the multivariate integral of equation (10), which is not presented here.[6] In Section 5, we present the econometric results obtained from equation (3).

3.3 Parameters estimation

Let $\nu_{(f)i}=\nu\ \forall i$ and $\delta_{(fi)t}=\delta\ \forall t$. We can apply the maximum likelihood method to estimate the unknown parameters, ν,κ^{-1},δ and β of the log likelihood corresponding function of equation (3).[7] In the application, presented in Section 5, we will apply the quasi-Newton method of estimation (known as a variable metric algorithm). We use the package Optim available in R (see Appendix D for more details). The initial values of the vector β are the maximum likelihood estimates of the NB2 model, and the initial values for ν,κ^{-1},δ parameters are set to one. To determine the variance-covariance matrix of the asymptotic distribution, we solve the Hessian

[6] This third possibility of estimating the integral in (10) by the Monte Carlo method is presented in Angers et al (2006). It is shown that the results are very similar to the two groups method.

[7] We could have used the Monte Carlo method with importance sampling to perform the parameters estimation. However, since the likelihood function is highly skewed and given the large number of parameters to estimate, we have chosen to use the maximum likelihood method.

matrix at $\hat{\nu}, \hat{\kappa}^{-1}, \hat{\delta}$ and $\hat{\beta}$. The size of the data is quite large; to reduce the computation time drastically, we compute the log likelihood function with a homemade C program inside the R system.

To divide the trucks of a fleet into two homogeneous groups as shown in Section 3.2, we take the maximum likelihood estimates $\left(\hat{\beta}\right)$ of the NB2 model to estimate $\hat{\gamma}_{\mathrm{fit}} = e^{X_{\mathrm{fit}}\hat{\beta}}$ for all the vehicles. Given that a truck has an estimate by year of follow-up, we calculated its mean $\hat{\hat{\gamma}}_{\mathrm{fi}} = \frac{1}{T_i}\sum_{i=1}^{T_i}\hat{\gamma}_{\mathrm{fit}}$. We sorted $\hat{\hat{\gamma}}_{\mathrm{fi}}$ for $i = 1, \cdots, I_f$ and calculated the difference $\left(\hat{\hat{\gamma}}_{\mathrm{fi+1}} - \hat{\hat{\gamma}}_{\mathrm{fi}}\right)$ for $i = 1, \cdots, I_f - 1$. After, we choose a cut-off point γ_{fc} where c is such that $\arg\max\left(\hat{\hat{\gamma}}_{\mathrm{fi+1}} - \hat{\hat{\gamma}}_{\mathrm{fi}}\right)$. The truck *i* is in group 1 if $\hat{\hat{\gamma}}_{\mathrm{fi}} < \gamma_{\mathrm{fc}}$ or is in group 2 if $\hat{\hat{\gamma}}_{\mathrm{fi}} \geq \gamma_{\mathrm{fc}}$ for all the observations of the truck *i*.

For example, for a fleet of 8 trucks and 20 observations (truck-years) as shown in Table 1, $c = \arg\max\left(0.08732\right) = 6$. Then the cutoff point $\gamma_{\mathrm{fc}} = \hat{\hat{\gamma}}_{\mathrm{f}6} = 0.23357$. All observations of truck 1 to truck 5 will therefore be in group 1 (low risk group) and all the others will be in group 2 (high risk group) . If we use the median or the mean instead of the maximum difference then the cut-off point will be respectively 0.12067 ((0.09509+0.14625)/2) and 0.14605, and truck 5 will change to group 2. However, it is more appropriate to be in group 1 because $\hat{\hat{\gamma}}_{\mathrm{f}5}$ is nearer to those in group 1.

Table 1
Example of group division, fleet of 8 trucks and 20 truck-years

Truck	Year	$\hat{\gamma}_{\mathrm{fit}}$	$\hat{\hat{\gamma}}_{\mathrm{fi}}$	Difference	Group
1	94	0.02527			1
	95	0.06524	0.04526		1
2	91	0.02417			1
	92	0.07178			1
	93	0.06422			1
	94	0.07340			1
	95	0.06423	0.05956	0.01430	1
3	91	0.09947			1
	92	0.09067	0.09067	0.03111	1

4	91	0.09677			1
	92	0.09817			1
	93	0.09033	0.09509	0.00442	1
5	91	0.15184			1
	92	0.14065	0.14625	0.05116	1
6	91	0.22807			2
	92	0.23906	0.23357	0.08732	2
7	91	0.25807			2
	92	0.23906	0.24857	0.01500	2
8	91	0.25989			2
	92	0.23906	0.24948	0.00091	2
Mean			0.14605		

4. DATA

The *Société de l'assurance automobile du Québec* (SAAQ) provided the files for our data set. The SAAQ is in charge of monitoring whether vehicles engaged in road transportation of people or merchandise comply with applicable laws and regulations. The SAAQ is also the insurer for bodily injuries linked to traffic accidents for individuals and fleets of vehicles and collects information on all truck accidents. Our starting point is the whole population of carriers registered in a SAAQ file on July 1997. To be in that file of carriers, a carrier must be the owner of a vehicle that meets different administrative conditions. Linked to each carrier, the data contain: (1) information on violations (with convictions) committed by the carrier during the 1990-1998 period, either for non-compliance with the Highway Safety Code's provisions on mechanical inspection; with rules on vehicles and their equipment; with codes on driving and hours of service or for oversize or poor load securement, etc., and (2) information identifying the carrier.

We also have access to information on vehicles registered in Quebec for the period of January 1, 1990 to December 31, 1998. We can link vehicles to carriers in each period. From the authorization status, we obtained information describing vehicles and plates. For each plate number, we have data covering the 1990-1998 period drawn from the files on mechanical inspection of vehicles and from the record of drivers' violations with conviction and demerit points for speeding, failure to stop at a red light or stop sign, and illegal passing, and for accidents.

The type of insurance coverage we consider is for property damages of the trucks. The premium for these losses is paid by the owner of the fleet to a private insurer. A truck can be driven by different drivers and a driver can drive different trucks during a policy period. We assume that the owner knows who drives each truck of the fleet at any point in time. The insurer does not observe the safety actions of both the driver and the fleet owner and there is also asymmetric information on prevention between the driver and the owner of the fleet. In the application of the model, the individuals are the trucks and their accidents are function of both observable and non-observable characteristics or actions from drivers, vehicles and fleets. The owner observes accidents and the drivers' traffic violations committed while driving a truck of the company. The choice and the description of the variables used in this study are presented in Appendix A.

5. RESULTS

5.1 Descriptive statistics

5.1.1 By fleet

We have 17,542 fleets with at least two trucks with a follow-up of at least two periods. In December 31, 1998, the average number of years each carrier is in the sample is seven, the minimum is one year and three months, and the maximum is 20 years and 9 months. Table 2 shows the distribution of fleets according to their main economic sector: 76.75% of 17,542 carriers are independent trucking firms, 13.09% are bulk public trucking firms and 8.70% are general public trucking firms. The sector is unknown for only a few firms. In addition, a few fleets also transport passengers or are short-term leasing firms.

Table 2
Distribution of firm's main activity

Firm's main activity	N	%
Unknown (sect_00)	63	0.36
Transporting passengers (sect_14)	71	0.40
General public trucking (sect_05)	1,526	8.70
Independent trucking (sect_07)	13,464	76.75
Short-term leasing firm (sect_08)	121	0.69
Bulk public trucking (sect_06)	2,297	13.09
Total	17,542	100.00

We note in Table 3 that approximately 4% of 17,542 fleets have over 20 trucks. On average, a truck has 3.87 observation periods ranging from 3.38 (a truck from a fleet of size 2) to 4.30 (a truck from a fleet of 10 to 20 trucks).

Table 3
Size of fleet distribution

Size of fleet	N	%	Average observation period per truck
2	6,888	39.27	3.38
3	3,203	18.26	4.07
4 to 5	3,285	18.73	4.18
6 to 9	2,171	12.38	4.29
10 to 20	1,298	7.40	4.30
21 to 50	496	2.83	4.24
More than 50	201	1.15	4.03

We note in Table 4 that a quarter of the 17,542 carriers have eight years of follow-up.

Table 4
Number of years of follow-up of the firm

Number of years of follow-up	N	%
2	3,649	20.80
3	2,512	14.32
4	2,075	11.83
5	1,654	9.43
6	1,645	9.38
7	1,567	8.93
8	4,440	25.31
Total	17,542	100.00

Table 5 shows that there are 3,629 fleets for which we have two consecutive years of follow-up, which is 99.5% (3,629/3,649) of fleets with two observation periods (Table 4). This percentage varies from 98.96 (3 periods) to 87.17% (7 periods). The higher the number of years of follow-up, the higher the percentage of carriers with absences during the reporting period.

Table 5
Number of consecutive years of follow-up of the fleets by year of follow-up start, Quebec 1991 to 1997

Number of years of follow-up	Year of follow-up start							Total
	1991	1992	1993	1994	1995	1996	1997	
2	949	296	263	272	332	325	1,192	3,629
3	708	251	178	180	231	938		2,486
4	619	174	166	169	807			1,935
5	448	181	131	739				1,499
6	565	144	783					1,492
7	483	883						1,366
8	4,440							4,440
Total	8,212	1,929	1,521	1,360	1,370	1,263	1,192	16,847

5.1.2 By truck-years

There are 43,037 trucks in 1991. This number increased to 63,749 in 1996. It decreases to 52, 392 in 1998 for a total of 456,177 truck-years, 15% of which had an accident during one year (Table 6). In 1991, nearly 86 out of 100 vehicles had no accident; this percentage rises to 88 out of 100 in 1997. Other statistics are presented in Appendix B.

Table 6
Number of truck accidents distribution according to year of observation

Number of truck accidents	% (by year of observation)								Total
	1991	1992	1993	1994	1995	1996	1997	1998	
0	85.61	86.07	87.19	86.66	86.47	87.04	88.44	86.52	86.78
1	12.22	11.93	11.05	11.47	11.53	11.14	10.11	11.56	11.36
2	1.82	1.67	1.47	1.58	1.65	1.49	1.24	1.60	1.56
3	0.28	0.27	0.23	0.23	0.28	0.26	0.17	0.26	0.24
4 and more	0.07	0.06	0.05	0.06	0.07	0.06	0.04	0.06	0.06
Number of trucks	43,037	55,388	57,795	59,347	61,917	63,749	62,552	52,392	456,177
Means truck crash	0.1696	0.1632	0.1489	0.1556	0.1596	0.1515	0.1327	0.1578	0.1541

5.2 Estimation of the models

For comparison we first estimate the Hausman (1994) random effects model that cannot take into account the firm-specific effect. The results are presented in columns 2 and 3 of Table 7. Several variables measure observable heterogeneity. Some of these variables (type of fuel, number of cylinders, number of axles, type of vehicle used) are characteristics concerning vehicles, whereas others (sector, fleet size, etc.) have to do with the fleet. We also include the number of violations of the trucking standards the year before the accidents and the number of violations of the road safety code leading to demerit points the year before the accidents**.** The first group of violations is more related to fleet behavior, and the second group is more related to drivers' behavior**.** Almost all coefficients are significant at 1%.

Table C1.1 (see appendix C), present the corresponding estimates of the Poisson model in columns 2 and 3 and the NB2 model in columns 4 and 5. The estimate of δ is equal to 0.8135 with the standard error of 0.0282. The implied variance to mean ratio $(1+\delta)/\delta$ is 2.23, which is greater than 1. Thus, the NB2 model specification allows for overdispersion in accidents distribution so we reject the Poisson model. Otherwise, the coefficients of the observable characteristics are very stable between the two models. All these results do not control for firm-specific effect so the serial correlation of residuals may be a problem having panel data. Columns 4 and 5 in Table 7 present the estimates of our Gamma-Dirichlet model when we add random firm-specific effect. The estimated coefficients are also very stable between the two models in the table, with the exception of the year variables. In the Hausman model, the year coefficients reflect the statistics provided for truck accidents in Table 6, where the fleet effect is not present. When we look at Table B3 in

Appendix B, we see that the average truck accident by fleet size does not have the same pattern during many years, as in Table 6. Since the year variables in the Gamma-Dirichlet model are for trucks of a given fleet, this may explain the difference. In Table C1.2, we present three other estimations of the Gamma-Dirichlet model by omitting different categorical variables, including the year variables. We observe that all parameters, including the random effects parameters, are stable.

Table 7
Estimation of the parameters of the distribution of the number of annual truck accidents for the 1991-1998 period (fleet of two trucks or more and trucks with two periods or more), with the Hausman and Gamma-Dirichlet models.

Explanatory variables	Hausman model		Gamma-Dirichlet model	
	Coefficient	Standard error	Coefficient	Standard error
Constant	-0.1254	0.0819	-3.9070*	0.0573
Number of years as carrier at 31 December	-0.0436*	0.0031	-0.0464*	0.0044
Sector of activity in 1998				
Other sector	-0.2484*	0.0929	-0.1426	0.1163
General public trucking	0.1003*	0.0252	0.1685*	0.0304
Bulk public trucking	Reference group		Reference group	
Private trucking	0.1574*	0.0213	0.2290*	0.0256
Short-term rental firm	0.4480*	0.0336	0.5633*	0.0483
Size of fleet				
2	Reference group		Reference group	
3	0.1260*	0.0180	0.0801*	0.0205
4 to 5	0.1941*	0.0172	0.1385*	0.0205
6 to 9	0.2798*	0.0171	0.2137*	0.0210
10 to 20	0.3617*	0.0166	0.2937*	0.0209
21 to 50	0.3574*	0.0177	0.3010*	0.0223
More than 50	0.3591*	0.0167	0.3077*	0.0217
Number of days authorized to drive in previous year	1.6878*	0.0300	2.0537*	0.0300
Number of violations of trucking standards in year before				
For overload	0.1216*	0.0117	0.0966*	0.0115
For excessive size	0.1456***	0.0883	0.1480***	0.0860
For poorly secured cargo	0.2522*	0.0363	0.2054*	0.0354
For failure to respect service hours	0.2585*	0.0663	0.1984*	0.0664
For failure to pass mechanical inspection	0.2383*	0.0308	0.1778*	0.0298
For other reasons	0.2678*	0.0779	0.1754**	0.0743
Type of vehicle use				
Commercial use including transport of goods without C.T.Q. permit	-0.1407*	0.0213	-0.1938*	0.0212
Transport of other goods	-0.0513**	0.0244	-0.1148*	0.0243
Transport of "bulk" goods	Reference group		Reference group	

Explanatory variables	Hausman model		Gamma-Dirichlet model	
	Coefficient	Standard error	Coefficient	Standard error
Type of fuel				
Diesel	Reference group		Reference group	
Gas	-0.4089*	0.0145	-0.3973*	0.0136
Other	-0.3109*	0.0775	-0.3079*	0.0736
Number of cylinders				
1 to 5 cylinders	0.3591*	0.0440	0.2167*	0.0403
6 to 7 cylinders	0.3778*	0.0136	0.3780*	0.0126
8 or more than 10 cylinders	Reference group		Reference group	
Number of axles				
2 axles (3,000 to 4,000 kg)	-0.1620*	0.0210	-0.2916*	0.0208
2 axles (more than 4,000 kg)	-0.1715*	0.0150	-0.2850*	0.0150
3 axles	-0.1559*	0.0151	-0.1278*	0.0149
4 axles	-0.1896*	0.0199	-0.1321*	0.0190
5 axles	-0.2182*	0.0173	-0.1973*	0.0174
6 axles or more	Reference group		Reference group	
Number of violations with demerit points year before				
For speeding	0.2585*	0.0105	0.1946*	0.0103
For driving with suspended license	0.4494*	0.0426	0.3830*	0.0422
For running a red light	0.3838*	0.0247	0.3094*	0.0239
For ignoring stop sign or traffic officer	0.4264*	0.0267	0.3597*	0.0258
For not wearing a seat belt	0.2044*	0.0304	0.1568*	0.0294
Observation period				
1991	0.0187	0.0251	0.0760**	0.0332
1992	-0.0183	0.0226	0.0548***	0.0293
1993	-0.0837*	0.0208	0.0806*	0.0259
1994	-0.0201	0.0190	0.1845*	0.0226
1995	0.0014	0.0175	0.2073*	0.0197
1996	-0.0426*	0.0165	0.1198*	0.0175
1997	-0.1583*	0.0163	-0.0791*	0.0163
1998	Reference group		Reference group	
$\hat{a}$	56.9383*	3.4587		
$\hat{b}$	1.8274*	0.0384		
$\hat{\nu}$			2.0086*	0.0422
$\hat{\kappa}$			12.6597*	0.2508
$\hat{\delta}$			4.6690*	0.3102
Number of observations:	456,117		456,117	

* significant at 1%; ** significant at 5%; *** significant at 10%

The random effects parameters are significant in both models. Let us concentrate on the Gamma-Dirichlet model proposed in this article. The significance of the three random effects parameters means that the random effects associated with the fleets (or the non-observable risk of the fleets) $(\hat{\kappa})$, as well the random effects of the trucks including the drivers $(\hat{\nu})$ and the random time effect $(\hat{\delta})$ significantly affect the truck distribution of accidents even when we control for many observable characteristics.

Suppose we are proposing a parametric model to rate insurance for vehicles belonging to a fleet. According to the results in Table 7, this premium will be a function of observable characteristics of the vehicle and fleet of the vehicle, as well a function of violations of the road-safety code committed by drivers and carriers.[8] This will not be enough because many unobservable characteristics of trucks, drivers and carriers also affect the trucks' distribution of accidents. The premiums will also have to be adjusted using the parameters of the random effects so as to account for the impact that the unobservable characteristics or actions of carrier, truck and drivers and even time can have on the truck accident rate. This form of rating makes it possible to visualize the impact (observable and non-observable) of behaviors of owners and drivers on the predicted rate of accidents, and consequently on premiums under potential moral hazard (see Angers et al, 2006, for more details). We show below how we can predict the number of truck accidents.

5.3 Fit statistics of different models according to fleet size

We now analyze the performance of the models. Model fits are based on the log likelihood statistics as well on other measures of information criteria such as the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). One advantage of using these two information criterion measures is that they can compare non-nested models.

Table 8
Fit statistics of the two models with two data sets

Statistics	Hausman model	Gamma-Dirichlet model	Hausman model	Gamma-Dirichlet model
	17,542 fleets having more than 1 truck		5,423 fleets having more than 4 trucks	
Log L	-197,165.23	-197,116.18	-155,634.46	-154,793.07
BIC	394,903.67	394,818.74	311,803.46	310,133.41
AIC	394,418.46	394,322.36	311,352.92	309,672.14
Number of trucks	111,106	111,106	79,609	79,609
Number of observations	456,177	456,177	336,772	336,772
Number of parameters	44	45	42	43

Bayesian Information Criterion $(\mathrm{BIC}) = -2\ln \mathrm{L} + \mathrm{k}\ln(\mathrm{N})$; Akaikes Information Criterion $(\mathrm{AIC}) = -2\ln \mathrm{L} + 2\mathrm{k}$ where k and N are the number of parameters and observations respectively, LogL = Log Likelihood ratio.

[8] It can also be a function of observable characteristics of the drivers but we do not consider them here.

For two models estimated from the same data set, the model with the smaller BIC and AIC is preferable. We note in Table 8 that the Gamma-Dirichlet model is preferred to the Hausman model wathever the fleet size.

The same results were obtained for fleets with more than two trucks and fleets with more than three trucks. Because the large majority of trucks belong to fleets that have more than two trucks, it is clear that our model permits better estimation of accident distributions than the Hausman model does. Detailed estimation results of the models with fleets having more than four trucks are presented in Appendix C2.

5.4 Predicted numbers of accidents

In order to check how the Gamma-Dirichlet model performs in predicting the number of truck accidents per fleet at time *t+1*, we assess an out-of-sample performance of the model in 1998 and we compare its forecasting performance with the observed accidents in 1998. From a methodological point of view, we proceed as follows: We partition the original sample period into two subsamples: an estimation sample for the 1991-1997 period and a forecasting sample for the year 1998.

The estimating sample consists of 16,344 fleets with at least two trucks and 393,634 trucks with a follow-up of at least two periods from 1991 to 1997. We obtain the coefficient estimates for the Gamma-Dirichlet model presented in Table C1.3, where we observe that the coefficients are similar to those presented in Table 7. We should mention that the year variable is continuous in table C1.3 to simplify the computation of the predictive accident distribution. This modification does not affect the estimation results.

One interesting feature of the Bayesian parametric model is to compute a parametric predictive distribution of accidents, $P\left(Y_{f1T_1+1},\cdots,Y_{fI_fT_{I_f}+1} \mid Y_{f11},\cdots,Y_{f1T_1},\cdots,Y_{fI_f1},\cdots,Y_{fI_fT_{I_f}}\right)$, which is equal to:

$$P\left(Y_{f1T_1+1},\cdots,Y_{fI_fT_{I_f}+1} \mid Y_{f11},\cdots,Y_{f1T_1},\cdots,Y_{fI_f1},\cdots,Y_{fI_fT_{I_f}}\right)=\frac{P\left(Y_{f11},\cdots,Y_{f1T_1},Y_{f1T_1+1},\cdots,Y_{fI_f1},\cdots,Y_{fI_fT_{I_f}},Y_{fI_fT_{I_f}+1}\right)}{P\left(Y_{f11},\cdots,Y_{f1T_1},\cdots,Y_{fI_f1},\cdots,Y_{fI_fT_{I_f}}\right)} \quad (18)$$

Using the Gamma-Dirichlet model, we obtain:

$$
\begin{aligned}
&P\left(Y_{f1T_1+1},\cdots,Y_{fI_fT_{I_f}+1} \mid Y_{f11},\cdots,Y_{f1T_1},\cdots,Y_{fI_f1},\cdots,Y_{fI_fT_{I_f}}\right)= \\
&\prod_{i=1}^{I_f^{t+1}}\frac{\left(\gamma_{fiT_i+1}\right)^{y_{fiT_i+1}}}{\Gamma\left(y_{fiT_i+1}+1\right)}\times\frac{\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}+S_0^{t+1}+\sum_{i=1}^{I_f^{t+1}}\kappa^{-1}\right)}{\Gamma\left(S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)}\times\frac{\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}\right)}{\Gamma\left(\sum_{i=1}^{I_f}T_i\kappa^{-1}+\sum_{i=1}^{I_f^{t+1}}\kappa^{-1}\right)} \\
&\times\frac{\left(\kappa^{-1}\right)^{\sum_{i=1}^{I_f^{t+1}}\kappa^{-1}}}{\left(\kappa^{-1}+\overline{\gamma}_{g_2}\right)^{S_0^{t+1}+\sum_{i=1}^{I_f^{t+1}}\kappa^{-1}}}\times\frac{\Gamma\left(\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)\right)}{\Gamma\left(\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)+S_0^{t+1}\right)}\times\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}+y^{*}_{fiT_i+1}\right)}{\prod_{i=1}^{I_f}\Gamma\left(S_i+\nu_{(f)i}\right)} \\
&\times\frac{\prod_{i=1}^{I_f^{t+1}}\Gamma\left(y_{fiT_i+1}+\delta_{(fi)T_i+1}\right)}{\prod_{i=1}^{I_f^{t+1}}\Gamma\left(\delta_{(fi)T_i+1}\right)}\times\frac{\prod_{i=1}^{I_f}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}{\prod_{i=1}^{I_f}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}+\left(y^{*}_{fiT_i+1}+\delta^{*}_{(fi)T_i+1}\right)\right)}\times\frac{\prod_{i=1}^{I_f}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}+\delta^{*}_{(fi)T_i+1}\right)}{\prod_{i=1}^{I_f}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}\right)} \\
&\times\frac{{}_2F_1\left(\sum_{i=1}^{g_1}\left(S_i+\nu_{(f)i}\right)+S_{g_1}^{t+1},S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}+S_0^{t+1}+\sum_{i=1}^{I_f^{t+1}}\kappa^{-1},\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)+S_0^{t+1},\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right)}{{}_2F_1\left(\sum_{i=1}^{g_1}\left(S_i+\nu_{(f)i}\right),S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1},\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right),\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right)}
\end{aligned}
\tag{19}
$$

where

* means that the truck *i* is present in the forecasting sample;

I_f^{t+1} is the number of trucks of fleet *f* in the forecasting sample;

γ_{fiT_i+1} are calculated for the forecasting sample with the coefficient estimates for the years 1991 to 1997 presented in Table C1.3;

S_0 is the total number of trucks accidents of fleet *f* before t+1.

S_0^{t+1} is the total number of truck accidents of fleet *f* at t+1;

$S_{g_1}^{t+1}$ is the total number of truck accidents in group 1 of fleet *f* at t+1.

The forecasting sample consists of 8,401 fleets with 2 trucks or more for a total of 41,614 trucks. In Table 9, we observe that out of the 8,401 fleets, 5,670 of them had no accident in 1998 (i.e. 67.5%). The average predictive probability of having zero accident is equal to 68.9% for the same year, in supposing that the number of accidents of truck *i* at time T_i+1, y_{fiT_i+1}, is equal to zero for all trucks of fleet *f* and so on for all fleets. Consequently there is 68.9% chance that a fleet will have no accident during the next year.

The average predictive probability that a fleet has 1 accident during the next year is 18.6% while the observed one is 19.1%. For two accidents, the respective probabilities are 6.1% and 6.5%, while for three accidents, they are 2.6% and 2.7%. Results for more than three accidents are available from the authors. Details of the computations are presented in Appendix E.

We used a paired *t*-test to compare the observed percentages and the predicted ones from the Gamma-Dirichlet model. First we need to check whether the differences between the two percentages follow a Normal distribution (i.e. Shapiro-Wilk test of normality). In Table 9 we observe large *p*-values for the normality test, thus, we do not reject the Normal distribution. Moreover, since the *p*-values of paired *t*-tests are greater than 0.05, we do not reject H_0 that the mean difference between the observed and the predicted percentages of accidents do not differ from zero at the 5% level of significance.

Table 9

Percentage of 8,401 fleets having no accident, 1 accident, 2 accidents or 3 accidents in forecasting sample and the average predictive probability of having *n* accidents from the Gamma-Dirichlet model by size of fleet and for all firms.

Size of fleet	% of firms with 0 accident		% of firms with 1 accident		% of firms with 2 accidents		% of firms with 3 accidents	
	Observed	Gamma-Dirichlet	Observed	Gamma-Dirichlet	Observed	Gamma-Dirichlet	Observed	Gamma-Dirichlet
2	84.5	84.8	14.4	13.1	2.6	2.0	0.4	0.4
3	78.8	80.4	20.5	19.9	5.8	4.7	2.1	1.3
4 to 5	71.3	74.3	25.6	26.3	9.4	8.7	3.8	2.9
6 to 9	60.7	62.3	31.0	30.7	13.0	15.8	5.9	7.4
10 to 20	39.5	41.2	22.2	25.7	21.3	21.4	11.2	14.8
21 to 50	20.3	18.5	7.4	8.8	11.7	12.8	9.2	14.4
More than 50	5.7	5.4	0.0	0.2	5.8	0.6	3.3	2.1
All firms	67.5	68.9	19.1	18.6	6.5	6.1	2.6	2.7
Shapiro-Wilks normality test	0.946		0.930		0.915		0.863	
p-value	0.689		0.549		0.433		0.162	
Paired *t*-test	-1.453		-0.860		0.553		-1.121	
df	6		6		6		6	
p-value	0.197		0.423		0.600		0.305	

5.5 Fixed and random effects models

We now analyze the consistency of the random effects estimators for the Gamma-Dirichlet model. One important assumption in the random effects model is that the random effects are uncorrelated with the observed explanatory variables used in the estimations. One way to verify the consistency of the random effects model is to compare the results with those obtained from a fixed effects model by applying the Hausman (1978) test. In the Gamma-Dirichlet model, we assume that Y_{fit} follows a Poisson distribution with parameter λ_{fit}. Let λ_{fit} be a log-linear function of the explanatory variables:

$$\log \lambda_{fit} = \beta_0 + \beta x_{fit} + \xi z_{fi} + \alpha_f + \theta_{(f)i} + \eta_{(fi)t} \tag{20}$$

where x_{fit} represents the time-varying explanatory variables, z_{fi} represents the time-invariant explanatory variables, α_f denotes the firm effects, $\theta_{(f)i}$ the truck effects with $\sum_{i=1}^{I_f} \theta_{(f)i} = 1$ and $\eta_{(f)i}$ the time effects with $\sum_{t=1}^{T_i} \eta_{(fi)t} = 1$.

In the random effects model, α_f is assumed to be an independent and identically distributed (iid) random variable following the Gamma distribution implying no correlation with the other regressors. In the fixed effects model, such an assumption is not needed because α_f is estimated using dummy variables. In the Gamma-Dirichlet model, the vector $\theta_{(f)i}$ follows a Dirichlet distribution. Hence, its components are not independent from one another. The same situation holds for the vector $\eta_{(f)i}$. When the random effects model is correctly specified, both the fixed and the random effects estimators would be consistent. The difference between the two estimators can be used as the basis for a Hausman test. Cameron and Trivedi (2013b) propose the following representation of the test:

$$T_H = \left(\hat{\beta}_{RE} - \tilde{\beta}_{FE}\right)' \left[\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}]\right]^{-1} \left(\hat{\beta}_{RE} - \tilde{\beta}_{FE}\right) \tag{21}$$

where T_H is the Hausman test statistic, $\tilde{\beta}_{FE}$ are the estimated parameters obtained from the fixed effects model and $\hat{\beta}_{RE}$ are the estimated parameters obtained from the random effects model. To estimate the variance term $\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}]$ we can use a panel bootstrap method that resamples over the 5 423 firms of the sample:

$$\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}] = \frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{\beta}_{FE}^{(b)} - \hat{\beta}_{RE}^{(b)}\right)\left(\tilde{\beta}_{FE}^{(b)} - \hat{\beta}_{RE}^{(b)}\right) \tag{22}$$

where $\tilde{\beta}_{FE}^{(b)}$ and $\hat{\beta}_{RE}^{(b)}$ are the estimates obtained from the bth bootstrap replication (see Appendix F for more details). If $T_H < \chi^2_{p,0,05}$ then, at the 5% level of significance, we do not reject the null hypothesis that the random effects are uncorrelated with the regressors and there is no need to use the fixed effects estimation.

In the fixed effects model, all characteristics that are not time-varying are captured by the fixed effects variables and have to be removed from the model. So we carry out the Hausman test only on the coefficients of the time-varying variables. We estimated the fixed effects Poisson regression model with the conventional Poisson model using 5,423 dummy variables for the fleets of four trucks and more.[9] Greene (2004) has demonstrated the computational feasibility of this approach. Table 10 shows the estimated coefficients and standard deviations of the time-varying variables of both the fixed effects Poisson model and the Gamma-Dirichlet random effects model. The estimates of the two models are likewise quite similar, with few exceptions.

We must mention that the Gamma-Dirichlet model has a constant term while the Poisson model does not by construction. Moreover, as for Table 7, the coefficients of the year variables differ between the two models. These differences, again, seem to be explained by the presence of the fleets effects in the Gamma-Dirichlet model. But the main point in this section is to test if the random effects are uncorrelated with the regressors. As Equation (21) above shows, the test consists in verifying whether the coefficients between the two regressions are statistically different.

[9] We used this group of fleets to reduce the number of dummies. It is clear that the methodology can be used for all groups of fleets.

Table 10
Estimation of the parameters of the distribution of the number of annual truck accidents for the period 1991-1998 (fleets of four trucks or more and with trucks having at least two periods) of the fixed effects Poisson model with 5,423 dummy variables (coefficients not presented here) and the Gamma Dirichlet model.

Explanatory variables	Fixed effects Poisson model		Gamma-Dirichlet model	
	Coefficient	Standard error	Coefficient	Standard error
Constant			-1.9611 *	0.0213
Number of violations of trucking standards in previous year				
For overload	0.1584*	0.0125	0.2006*	0.0127
For excessive size	0.2828*	0.0975	0.2675*	0.0993
For poorly secured cargo	0.2321*	0.0387	0.2770*	0.0397
For failure to obey service hours	0.2245*	0.0693	0.2777*	0.0717
For failure to pass mechanical inspection	0.1583*	0.0334	0.2012*	0.0347
For other reasons	0.2331*	0.0784	0.2369*	0.0807
Number of violations with demerit points in previous year				
For speeding	0.2248*	0.0112	0.2584*	0.0116
For driving under suspension	0.3857*	0.0479	0.4245*	0.0502
For running a red light	0.3068*	0.0268	0.3804*	0.0274
For ignoring a stop sign or traffic agent	0.3443*	0.0287	0.4105*	0.0295
For not wearing a seat belt	0.1219*	0.0337	0.1651*	0.0342
Size of fleet				
5 trucks	Reference group		Reference group	
6 to 9 trucks	0.0283	0.0235	0.0168	0.0198
10 to 20 trucks	0.0532*	0.0307	0.0864*	0.0219
20 to 50 trucks	0.0347	0.0403	0.0829*	0.0244
More than 50 trucks	0.0841*	0.0511	0.0849*	0.0243
Observation period				
1991	0.0586*	0.0189	0.0995*	0.0188
1992	0.0292	0.0178	0.0823*	0.0178
1993	-0.0584*	0.0178	0.0877*	0.0178
1994	-0.0209	0.0173	0.1694*	0.0173
1995	-0.0157	0.0170	0.1631*	0.0170
1996	-0.0608*	0.0170	0.0672*	0.0170
1997	-0.1914*	0.0175	-0.1759*	0.0174
1998	Reference group		Reference group	
$\hat{\nu}$			1.7438*	0.0360
$\hat{\kappa}$			23.2452*	0.5708
$\hat{\delta}$			4.7249*	0.3431
Log L	-150,397.2		-159,255.61	
Number of observations:	336,772		336,772	

* Significant at 1%.

Figure 1 presents the values of Hausman test statistic, T_H, based on the number of bootstrap replications (for the bootstrap variance matrix estimated in the Hausman test). We observe that after 300 replications, $T_H < \chi^2_{22,0,05}$ where $\chi^2_{22,0,05} = 33.9$. So, at 5% or any lower level of significance, we do not reject the null hypothesis that the random effects are uncorrelated with the regressors. Consequently, there is no statistical difference between the coefficients of the Gamma-Dirichlet model and those of the fixed effects Poisson model presented in Table 10.

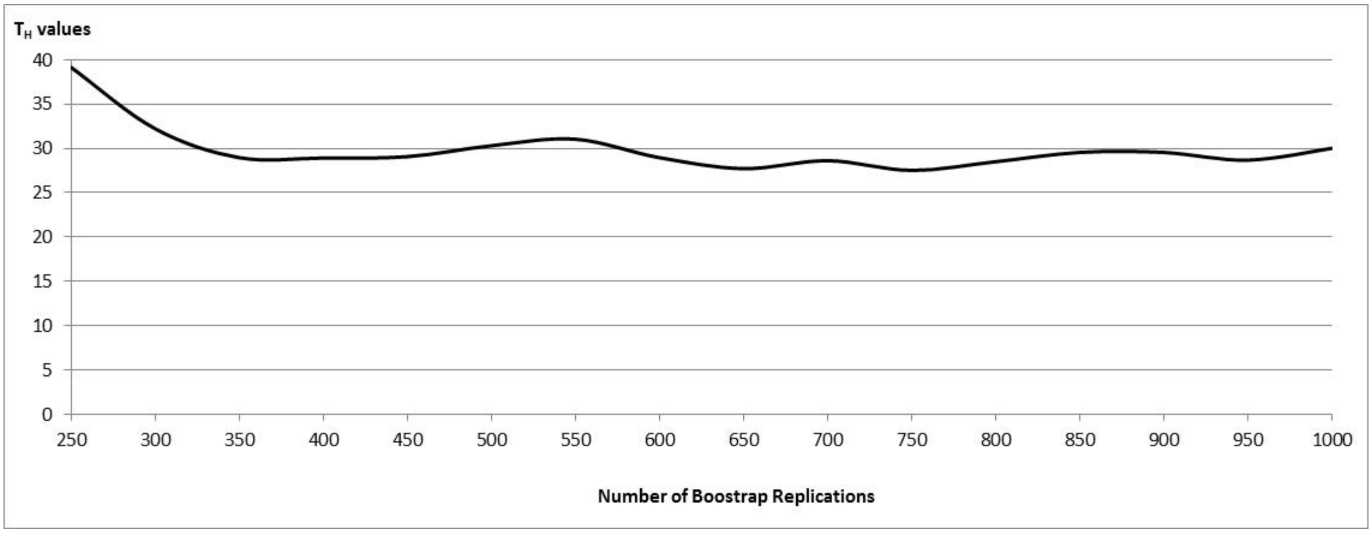

Figure 1: T_H values of the Hausman test is based on the number of bootstrap replications for the firm effects

6. CONCLUSION

In this article, we propose a new parametric model with random effects for the estimation of accidents distribution in the presence of individual and firm effects. Non-observable factors are treated as random effects. A Poisson fixed effects model is estimated to verify the consistency of the random effects model. We do not reject the null hypothesis that the random effects are uncorrelated with the regressors.

This type of model can be used to compute insurance premiums for drivers or vehicles belonging to a fleet because the characteristics and the management behavior of the fleets can affect the accident rate of vehicles and their drivers. For example, the manager of a given fleet may have a high risk appetite and ask their drivers to drive faster or to work more than the regulated number of hours during a week. He may also ask them to transport poorly secured cargo. A pricing rule that includes the observable and non-observable characteristics of all parties that affect accident distributions should consequently be fairer, and introduce the appropriate incentives of all parties under asymmetric information. Our results show that the Gamma-Dirichlet model performs well in predicting out-of-sample accidents.

The methodology developed in this study can be applied to estimating event distributions in many other domains than insurance pricing. Since 2004, banks are regulated by Basel II for keeping capital for operational risk. The operational risk of different banks is a function of the observable characteristics and the non-observable behavior of the personnel and of the management. A similar environment is present for the default risk of different firms or for the accident risk of any public institution or transportation firm including airline accidents. Other domains of applications include the failure or success rate of hospitals, universities, or any institution with principal-agent situations with teams.

In this study, we used a parametric model to estimate accident distributions. The main motivation was to obtain explicit parameter estimates for the insurance pricing of vehicles that includes individual and firm effects. Since we have a very large dataset, we could also have used the Classification and Regression Tree (CART) approach which does not require any ex-ante relationship between dependent and independent variables (Chang and Chen, 2005). It would be interesting to extend our analysis to such data mining techniques and see their advantages and disadvantages with respect to our pricing objectives.

REFERENCES

Allison, P.D., Waterman, R.P., 2002. Fixed-effects negative binomial regression models. *Sociological Methodology* 32, 247-265.

Angers J-F., Desjardins D., Dionne G., Guertin F., 2006. Vehicle and fleet random effects in a model of insurance rating or fleets of vehicles. *Astin Bulletin* 36, 25-77.

Baltagi B.H., 1995. *Econometric Analysis of Panel Data.* Wiley, Chichester.

Boucher J.P., Denuit M., 2006. Fixed versus random effects in Poisson regression models for claim counts: A case study with motor insurance. *Astin Bulletin* 36, 285-301.

Boucher J-P, Denuit, M., Guillen, M., 2008. Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Casualty Actuary Society* 2, 135-162.

Boyer, M., Dionne, G., Vanasse, C., 1992. Econometric Models of Accident Distributions. In Dionne, G. (Ed.), *Contributions to Insurance Economics*, Springer, 169-213.

Cameron A.C., Trivedi P.K., 2013a. *Regression Analysis of Count Data*, Second Edition, Cambridge University press.

Cameron, A.C., Trivedi, P.K., 2013b. Count panel data. Mimeo, University of California.

Cameron, A.C., Trivedi, P.K., 1986. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1, 29–54.

Chang, L., Chen, W., 2005. Data mining of tree-based models to analyse freeway accident frequency. *Journal of Safety Research* 36, 365-375.

Desjardins, D., Dionne, G., Pinquet, J., 2001. Experience rating schemes for fleets of vehicles. *Astin Bulletin* 31, 1, 81-106.

Dionne, G., Gagné, R., Gagnon, F., Vanasse, C., 1997. Debt, moral hazard and airline safety: An empirical evidence. *Journal of Econometrics* 79, 379-402.

Dionne G., Gagné, R., Vanasse, C., 1998. Inferring technological parameters from incomplete panel data. *Journal of Econometrics* 87, 303-327.

Dionne G., Vanasse, C., 1992. Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics* 7, 149-165.

Dionne G., Vanasse, C., 1989. A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *Astin Bulletin* 19, 199-212.

Fardilha, T., de Lourdes Centeno, M., Esteves, R., 2016. Tariff systems for fleets of vehicles: a study on the portfolio of Fidelidade. *European Actuarial Journal* 6, 331-349.

Frangos N., Vrontos, S.D., 2001. Design of optimal bonus-malus systems with a frequency and a security component on an individual basis in automobile insurance. *Astin Bulletin* 31, 1-22.

Fluet C., 1999. Commercial vehicle insurance: Should fleet policies differ from single vehicle plans? In Dionne, G., Laberge-Nadeau, C. (Eds.), *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*. Kluwer Academic Press, pp. 101-117.

Frees, E.W., Valdez, E., 2011. Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103, 1457-1469.

Gouriéroux C., Monfort A., Trognon A., 1984. Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* 52, 701-720.

Gradshteyn I.S., Ryzhik, I.M., 1980. *Table of integrals, series and products*. Academic Press Inc., New York.

Greene, W.H., 2005. Functional form and heterogeneity in models for count data. In W. Greene (Ed.), *Foundations and Trends in Econometrics*, vol. 1, no 2, 115-218.

Greene, W.H., 2004. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal* 7, 98-119.

Hausman J.A., Hall, B.H., Griliches, Z., 1984. Econometric models for count data with an application to the patents – R&D relationship. *Econometrica* 52, 909-938.

Hausman, J.A., 1978. Specification Tests in Econometrics. *Econometrica*, 46, 1251-1271.

Hausman J.A., Wise, D.A., 1979. Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica* 47, 455-473.

Holmstrom, B., 1982. Moral hazard in teams. *The Bell Journal of Economics* 13, 324-340.

Hsiao C., 1986. Analysis of panel data. *Econometric Society Monographs*, no 11, Cambridge University Press, Cambridge.

Laffont, J.J., Martimort, D., 2001. *The theory of incentives: The principal-agent model*. Princeton University Press, Princeton.

Lange K., 1999. *Numerical analysis for statisticians*. Springer, New York.

Norberg, N., 1986. Hierarchical credibility: analysis of a random effect linear model with nested classification. *Scandinavian Actuarial Journal*, 204-222.

Pinquet J., 2013. Experience rating in non-life insurance. In Dionne, G. (Ed.), *Handbook of Insurance*. Springer, New York, 471-586.

Purcaru, O., Denuit, M., 2003. Dependence in dynamic claim frequency credibility models. *Astin Bulletin* 33, 1, 23-40.

Scheffé H., 1999. The Analysis of Variance. Wiley, New York.

Modelling and Estimating Individual and Firm Effects with Count Panel Data

Online Appendix

Jean-François Angers, Denise Desjardins
Georges Dionne and François Guertin

APPENDIX A: CHOICE AND DESCRIPTION OF VARIABLES

The unit of observation is an eligible vehicle with authorization to circulate at least one day in year *t*, and which has been followed up for at least two years. We analyze the accident totals found in SAAQ files. These totals include all the traffic accidents causing bodily injuries and all accidents causing material damage reported by the police in Quebec.

Dependent variable

Y_{fit} = the number of accidents in which vehicle *i* of fleet *f* has been involved during year *t*.

Y_{fit} can take the values 0, 1, 2, 3, 4 and over.

Explanatory variables

We have two types of explanatory variables: those concerning the carrier and those concerning vehicles and drivers.

Variables concerning the carrier

- *Size of fleet for year t:* 7 dichotomous variables have been created.
  The two-vehicle size is used as the reference category.

- *Sector of economic activity*: 5 dichotomous variables have been created for vehicles transporting goods :

  sect_14 = 1 if the main sector of activity is transporting passengers;
  sect_05 = 1 if the sector of activity is general public trucking;
  sect_06 = 1 if the sector of activity is public bulk trucking;
  sect_07 = 1 if the sector of activity is independent trucking;
  sect_08 = 1 if the sector of activity is a short-term leasing firm.

  The "public bulk trucking" sector is used as the reference category.

- Seven (7) variables have been created for vehicles engaged in the *transportation of goods,* to measure the number of convictions per vehicle in the year preceding year *t* for each carrier:

  - *Number of violations per vehicle for overweight committed by a carrier in the year preceding year t.* A positive sign is predicted, because more overweight violations should, on average, generate more accidents.
  - *Number of violations per vehicle for oversize committed by a carrier in the year preceding year t:* A positive sign is predicted, because more violations for oversize should, on average, generate more accidents.
  - *Number of violations per vehicle for poorly secured loads committed by a carrier in the year preceding year t:* A positive sign is predicted, because more violations for poorly secured loads should, on average, generate more accidents.
  - *Number of violations per vehicle of Highway Safety Code provisions regarding transportation of hazardous materials committed by a carrier in the year preceding*

year t: A positive sign is predicted, because more violations of regulations for the transportation of hazardous materials should, on average, generate more accidents.

- *Number of violations per vehicle of hours-of-service regulations committed by a carrier in the year preceding year t:* A positive sign is predicted because more violations of hours-of-service regulations should, on average, generate more accidents.
- *Number of violations per vehicle of Highway Safety Code provisions regarding mechanical inspection committed by a carrier in the year preceding year t:* A positive sign is predicted, because more violations of regulations regarding mechanical inspection should, on average, generate more accidents.
- *Number of violations per vehicle, other than those already mentioned, committed by a carrier in the year preceding year t:* A positive sign is predicted, because more violations other than those already mentioned should, on average, generate more accidents.

Variables concerning vehicles and drivers (a vehicle may have more than one driver)

- *Vehicle's number of cylinders:* 4 dichotomous variables have been created:
  cyl_0 = 1 if the vehicle's number of cylinders is not known;
  cyl1_5 = 1 if the vehicle has 1 to 5 cylinders;
  cyl6_7 = 1 if the vehicle has 6 to 7 cylinders;
  cyl_8p = 1 if the vehicle has 8 or more than 10 cylinders.

  The group of vehicles with 8 or more than 10 cylinders is used as the reference category.

- *Vehicle's type of fuel:* 3 dichotomous variables have been created:
  diesel = 1 if the vehicle uses diesel as fuel;
  fuel = 1 if the vehicle uses gas as fuel;
  other = 1 if the vehicle uses another type of fuel.

  The group of vehicles using diesel as fuel is considered the reference category.

- *maximum number of axles:* 7 dichotomous variables have been created:
  ess_0 = 1 if the maximum number of axles does not apply to this type of vehicle;
  ess_2 = 1 if the vehicle has two axles and a mass of between 3,000 and 4,000 kg;
  ess_2p = 1 if the vehicle has two axles and a mass higher than 4,000 kg;
  ess_3 = 1 if the vehicle is supported by a maximum of three axles;
  ess_4 = 1 if the vehicle is supported by a maximum of four axles;
  ess_5 = 1 if the vehicle is supported by a maximum of five axles;
  ess_6p = 1 if the vehicle is supported by six or more axles.

  The group of vehicles with two axles and a mass of between 3 000 and 4 000 kg is used as the reference category.

- *Vehicle's type of use:* 3 dichotomous variables for vehicles transporting goods have been created:

compr = 1 if the vehicle is meant for commercial use, including transportation of goods without a CTQ permit;
tbrgn = 1 if the vehicle is meant for transportation of goods but other than in bulk, which requires a CTQ permit;
tbrvr = 1 if the vehicle is meant for transportation of bulk goods.

The group of vehicles transporting bulk goods is used as the reference category.

- Six (6) variables have been created to measure the number of convictions per vehicle accumulated in the year preceding year *t* by one or more drivers:
  - *Number of violations for speeding per vehicle, committed in the year preceding year t.* A positive sign is predicted because more speeding violations should, on average, generate more accidents.
  - *Number of violations for driving with a suspended license per vehicle, committed in the year preceding year t.* A positive sign is predicted because more driving with a suspended license should, on average, generate more accidents.
  - *Number of violations for running a red light per vehicle, committed in the year preceding year t.* A positive sign is predicted because more incidences of running a red light should, on average, generate more accidents.
  - *Number of violations for failure to obey a stop sign or a signal from a traffic officer per vehicle, committed in the year preceding year t.* A positive sign is predicted because more incidents of failure to respect a stop sign or a signal from a traffic cop should, on average, generate more accidents.
  - *Number of violations for failure to wear a seat belt per vehicle, committed in the year preceding year t.* A positive sign is predicted because more incidents of failure to wear a seat belt should, on average, generate more accidents.
  - *Number of violations other than those mentioned per vehicle, committed in the year preceding year t.* A positive sign is predicted because a greater number of violations other than those mentioned should, on average, generate more accidents.

APPENDIX B: ADDITIONAL DESCRIPTIVE STATISTICS

Table B1 shows the distribution of the size of the fleet by year and for a total of 8 years. In 1991, we have 8,650 fleets, this number increases to 11,965 fleets in 1996 and decreases to 10,321 in 1998 for a total of 87,771 fleet-years. Among the 87,771 fleet-years, 46.51% have two vehicles and about 3% have over 20 vehicles.

Table B1
Size of fleet distribution (in %) by year

Size of fleet	% by year								% total
	1991	1992	1993	1994	1995	1996	1997	1998	
2	47.86	46.31	46.60	46.82	45.98	46.08	45.83	47.03	46.51
3	19.63	19.75	19.92	19.71	19.45	19.30	19.28	19.10	19.51
4 to 5	15.26	15.99	15.75	15.54	16.28	16.02	16.40	16.26	15.96
6 to 9	9.16	9.40	9.19	9.46	9.62	9.88	9.62	9.27	9.47
10 to 20	5.45	5.72	5.76	5.74	5.76	5.68	5.95	5.64	5.72
21 to 50	1.97	2.06	2.00	1.89	2.05	2.14	2.08	2.01	2.03
More than 50	0.68	0.78	0.78	0.83	0.86	0.89	0.85	0.70	0.80
Number of fleets	8,650	10,691	11,132	11,445	11,733	11,965	11,834	10,321	87,771

From Table B2, we observe that 9,963 fleets remain in the same class of fleet size during the eight years of observation (sum of the diagonal of the table), which is 56.80% of 17,542 fleets. There are 2,722 fleets whose size varies between 2 and 3 trucks, and 1,423 fleets whose size varies between 2 trucks to 4-5 trucks.

Table B2
Minimum and maximum fleet size distribution during the follow-up, Québec 1991-1998

Minimum size	Maximum size of fleet							Total	
of fleet	2	3	4 to 5	6 to 9	10 to 20	21 to 50	+ 50	N	%
2	7,884	2,722	1,423	365	101	19	2	12,561	71.35
3		790	946	368	85	14	2	2,205	12.57
4 to 5			551	644	172	22	1	1,390	7.92
6 to 9				320	394	57	12	783	4.46
10 to 20					268	152	22	442	2.52
21 to 50						93	56	149	0.85
More than 50							57	57	0.32
Total N	7,884	3,512	2,920	1,697	1,020	357	152	17,542	100.00

Minimum size	Maximum size of fleet							Total	
of fleet	2	3	4 to 5	6 to 9	10 to 20	21 to 50	+ 50	N	%
%	44.94	20.20	16.65	9.67	5.81	2.04	0.84	100.00	

We observe from Table B3 that the average accident rate of trucks per fleet is lowest for the year 1997, followed by 1993, 1998, 1996 and 1994. In the years 1991, 1992 and 1995, the highest average rates of truck accidents per fleet were recorded. These observations are almost stable for different fleet sizes.

Table B3
Average truck accidents per fleet according to size of fleet and year.

Size of fleet	Average truck accident per fleet by year								Total
	1991	1992	1993	1994	1995	1996	1997	1998	
2	0.2626	0.2480	0.2219	0.2215	0.2219	0.2155	0.1809	0.2186	0.2224
3	0.4370	0.4154	0.3811	0.4007	0.4194	0.3712	0.3129	0.4049	0.3909
4 to 5	0.6689	0.6864	0.6030	0.6296	0.6408	0.5863	0.5507	0.6490	0.6239
6 à 9	1.3914	1.2259	1.0909	1.1311	1.1833	1.0981	1.0018	1.1996	1.1550
10 to 20	2.6730	2.6127	2.3744	2.5099	2.4527	2.4824	2.0767	2.5223	2.4497
21 to 50	5.9176	5.3818	5.0448	5.3565	5.8875	5.2461	4.8618	5.7681	5.4094
More than 50	22.6780	22.4096	21.7701	22.0421	22.0198	21.0935	18.4700	22.5417	21.5014
Average truck accidents per fleet	0.8575	0.8561	0.7824	0.8157	0.8531	0.8153	0.7106	0.8120	0.8109

We have 111,106 different trucks in the database, nearly three-quarters of which are for commercial use, including transportation of goods. As indicated in Table B4, 17.52% of the trucks are used for transportation of goods other than bulk and 8% for transportation of goods in bulk.

Table B4
Vehicle use distribution.

Vehicle use	N	%
Commercial use, including transport of goods without CTQ permit. (combr)	82,798	74.52
Transport of goods other than in bulk (tbrgn)	19,470	17.52
Transport of goods in bulk (tbrvr)	8,838	7.95
Total	111,106	100.00

Table B5 presents the variation of average annual accidents per truck relative to the number of driver's violations of the Highway Safety Code during the year preceding the accidents. Violations committed by drivers are very powerful in explaining truck accidents during the next year. Indeed, we observe that the year *t* accident rate is an increasing function of previous year violations committed by drivers.

Table B5
Average truck accidents according to the driver's violations committed the previous year.

Violations committed by the driver the previous year	Year								Total
	1991	1992	1993	1994	1995	1996	1997	1998	
For speeding									
0	0.1642	0.1586	0.1432	0.1486	0.1516	0.1435	0.1240	0.1498	0.1472
1	0.2974	0.2592	0.2640	0.2723	0.2631	0.2523	0.2161	0.2609	0.2556
2	0.2701	0.3410	0.3045	0.4000	0.3566	0.3249	0.3207	0.3281	0.3337
3 and more	0.4194	0.5000	0.2424	0.4651	0.5506	0.4821	0.3973	0.4600	0.4505
For driving with a suspended license									
0	0.1696	0.1629	0.1485	0.1547	0.1584	0.1507	0.1321	0.1574	0.1535
1 and more	0.7500	0.5217	0.3750	0.4076	0.3549	0.3426	0.3017	0.3265	0.3566
For running a red light									
0	0.1679	0.1617	0.1473	0.1538	0.1571	0.1491	0.1308	0.1555	0.1521
1	0.2726	0.2846	0.2764	0.2999	0.3350	0.3135	0.2981	0.3413	0.3036
2 and more	0.5294	0.6667	0.3846	0.2727	0.6000	0.7272	0.2308	0.2727	0.5040
For disobeying stop signs or police signals									

Violations committed by the driver the previous year	Year								Total
	1991	1992	1993	1994	1995	1996	1997	1998	
0	0.1677	0.1618	0.1474	0.1541	0.1572	0.1498	0.1315	0.1561	0.1524
1	0.3204	0.3140	0.2797	0.2823	0.3570	0.2931	0.2411	0.3100	0.2993
2 and more	0.5000	0.2857	0.2500	0.5833	0.2941	0.5263	0.3125	0.5000	0.4016
For failing to wear a seat belt									
0	0.1689	0.1626	0.1481	0.1554	0.1588	0.1508	0.1316	0.1576	0.1534
1	0.2304	0.2246	0.2293	0.1770	0.2376	0.2100	0.2096	0.2124	0.2164
2 and more	0.4138	0.4333	0.2571	0.2750	0.1774	0.2653	0.3137	0.1200	0.2741
For overweight									
0	0.1649	0.1583	0.1448	0.1517	0.1544	0.1461	0.1293	0.1540	0.1497
1	0.2430	0.2764	0.2410	0.2501	0.2432	0.2383	0.1889	0.2631	0.2394
2 and more	0.3387	0.2956	0.3364	0.2926	0.3552	0.3026	0.2155	0.3874	0.3065
For oversize									
0	0.1695	0.1632	0.1488	0.1554	0.1596	0.1515	0.1326	0.1577	0.1540
1 and more	0.2836	0.1000	0.2917	0.2603	0.1574	0.1545	0.2269	0.2821	0.2119
For poorly secured loads									
0	0.1688	0.1625	0.1482	0.1550	0.1587	0.1509	0.1323	0.1570	0.1534
1 and more	0.3185	0.3198	0.2667	0.2665	0.2656	0.2621	0.2214	0.3778	0.2791
For exceeding hours of service									
0	0.1696	0.1632	0.1486	0.1556	0.1592	0.1513	0.1325	0.1575	0.1539
1 and more	0.5714	0.3000	0.6333	0.1951	0.3529	0.2743	0.3881	0.3571	0.3496
For failure to undergo mechanical inspection									
0	0.1691	0.1626	0.1474	0.1546	0.1578	0.1509	0.1321	0.1572	0.1532
1 and more	0.2890	0.3180	0.2388	0.2534	0.3024	0.2251	0.2168	0.2768	0.2591

We note in Table B6 that 78.80% of the 111,106 trucks use diesel as fuel.

Table B6
Type of fuel distribution

Type of fuel	N	%
Diesel	87,546	78.80
Gas	22,999	20.70
Other	561	0.50
Total	111,106	100.00

Table B7 illustrates that 21.15% of the 111,106 trucks have six axles or more and 28.57% have two axles and weigh more than 4,000 kg, and Table B8 shows that 64.95% of the 111,106 trucks have 6 to 7 cylinders. Only 1.15% has 5 cylinders or fewer.

Table B7
Number of axles distribution

Number of axles	N	%
2 axles (3,000 to 4,000 kg)	15,960	14.36
2 axles (More than 4,000 kg)	31,747	28.57
3 axles	21,856	19.67
4 axles	7,377	6.64
5 axles	10,666	9.60
6 axles and more	23,500	21.15
Total	111,106	100.00

Table B8
Number of cylinders distribution

Number of cylinders	N	%
Unknown	501	0.45
1 to 5 cylinders	1,283	1.15
6 to 7 cylinders	71,159	64.05
8 or more than 10 cylinders	38,163	34.35
Total	111,106	100.00

Table B9 indicates that 10.64% of the 111,106 trucks have 8 years of follow-up, which represents 10.64% of the population.

Table B9
Number of years of follow-up of the truck

Number of years of follow-up	N	%
2	30,716	27.65
3	23,270	20.94
4	17,831	16.05
5	11,998	10.80
6	9,241	8.32
7	6,225	5.60
8	11,825	10.64
Total	111,106	100.00

We note in Table B10 that there are 30,432 trucks for which we have two consecutive years of follow-up, which corresponds to 99.07% (30,432/30,716) of trucks with two observation periods. This percentage varies from 98.43 (3 periods) to 97.65 (7 periods).

Table B10
Number of consecutive years of follow-up of the trucks by year of follow-up start, Quebec 1991 to 1997.

Number of year of follow-up	Year of follow-up start							Total
	1991	1992	1993	1994	1995	1996	1997	
2	8,326	2,581	2,193	2,081	2,844	2,351	10,056	30,432
3	6,421	2,291	1,624	1,855	1,947	8,766		22,904
4	5,273	1,535	1,711	1,524	7,304			17,347
5	3,967	1,289	1,067	5,226				11,549
6	3,680	818	4,441					8,939
7	2,630	3,449						6,079
8	11,825							11,825
Total	42,122	11,963	11,036	10,686	12,095	11,117	10,056	109,075

APPENDIX C1: POISSON AND NB2 MODELS ESTIMATION RESULTS

Table C1.1: Poisson negative binomial estimates
Estimation of the parameters of the distribution of the number of annual truck accidents for the 1991-1998 period (fleet of two trucks or more and trucks with two periods or more), Poisson and NB2 models

Explanatory variables	Poisson model		NB2 model	
	Coefficient	Standard error	Coefficient	Standard error
Constant	-3.5846*	0.0415	-3.5895*	0.0438
Number of years as carrier at 31 December	-0.0424*	0.0026	-0.0432*	0.0028
Sector of activity in 1998				
Other sector	-0.2766*	0.0804	-0.2694*	0.0839
General public trucking	0.0933*	0.0210	0.0977*	0.0226
Bulk public trucking	Reference group		Reference group	
Private trucking	0.1548*	0.0177	0.1595*	0.0190
Short-term rental firm	0.4055*	0.0275	0.4185*	0.0299
Size of fleet				
2	Reference group		Reference group	
3	0.1245*	0.0161	0.1246*	0.0171
4 to 5	0.1900*	0.0151	0.1926*	0.0160
6 to 9	0.2764*	0.0148	0.2797*	0.0158
10 to 20	0.3704*	0.0142	0.3761*	0.0152
21 to 50	0.3698*	0.0151	0.3782*	0.0161
More than 50	0.3837*	0.0142	0.3892*	0.0151
Number of days authorized to drive in previous year	1.6703*	0.0290	1.6765*	0.0300
Number of violations of trucking standards in year before				
For overload	0.1456*	0.0104	0.1502*	0.0117
For excessive size	0.1607***	0.0825	0.1615***	0.0910
For poorly secured cargo	0.2927*	0.0329	0.2991*	0.0380
For failure to respect service hours	0.2771*	0.0598	0.2880*	0.0710
For failure to pass mechanical inspection	0.2819*	0.0280	0.2977*	0.0316
For other reasons	0.2812*	0.0699	0.2602*	0.0807
Type of vehicle use				
Commercial use including transport of goods without C.T.Q. permit	-0.1167*	0.0177	-0.1249*	0.0191
Transport of other than "bulk" goods	-0.0325	0.0203	-0.0387***	0.0220
Transport of "bulk" goods	Reference group		Reference group	
Type of fuel				
Diesel	Reference group		Reference group	
Gas	-0.3922*	0.0124	-0.3939*	0.0130
Other	-0.3169*	0.0684	-0.3161*	0.0713
Number of cylinders				
1 to 5 cylinders	0.3536*	0.0360	0.3527*	0.0385
6 to 7 cylinders	0.3752*	0.0114	0.3763*	0.0121
8 or more than 10 cylinders	Reference group		Reference group	

Explanatory variables	Poisson model		NB2 model	
	Coefficient	Standard error	Coefficient	Standard error
Number of axles				
2 axles (3,000 to 4,000 kg)	-0.1603*	0.0177	-0.1616*	0.0188
2 axles (more than 4,000 kg)	-0.1505*	0.0122	-0.1541*	0.0132
3 axles	-0.1156*	0.0124	-0.1203*	0.0133
4 axles	-0.1818*	0.0163	-0.1817*	0.0175
5 axles	-0.2040*	0.0145	-0.2056*	0.0156
6 axles or more	Reference group		Reference group	
Number of violations with demerit points year before				
For speeding	0.2961*	0.0092	0.3098*	0.0106
For driving with suspended license	0.4895*	0.0350	0.5590*	0.0433
For running a red light	0.4549*	0.0226	0.4723*	0.0256
For ignoring stop sign or traffic officer	0.4953*	0.0244	0.5107*	0.0277
For not wearing a seat belt	0.2295*	0.0281	0.2386*	0.0310
Observation period				
1991	0.0099	0.0222	0.0142	0.0239
1992	-0.0225	0.0202	-0.0195	0.0217
1993	-0.0881*	0.0189	-0.0876*	0.0203
1994	-0.0228	0.0174	-0.0218	0.0187
1995	-0.0012	0.0163	-0.0011	0.0175
1996	-0.0463*	0.0157	-0.0453*	0.0168
1997	-0.1605*	0.0158	-0.1597*	0.0168
1998	Reference group		Reference group	
$\hat{\delta}$			0.8135	0.0282
Number of observations:	456,117		456,117	

* significant at 1%; ** significant at 5%; *** significant at 10%

Table C1.2: Robustness analysis of the estimations

Estimation of the parameters of the distribution of the number of annual truck accidents for the 1991-1998 period (fleet of two trucks or more and trucks with two periods or more), Gamma-Dirichlet models (excluding the variable fleet size at left, cylinders in the middle, and observation period at right).

Explanatory variables	Gamma-Dirichlet model Fleet size omitted		Gamma-Dirichlet model Cylinders omitted		Gamma-Dirichlet model Period omitted	
	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Constant	-3.8990*	0.0572	-3.5304*	0.0563	-3.7451*	0.0408
Number of years as carrier at 31 December	-0.0331*	0.0044	-0.0467*	0.0044	-0.0578*	0.0019
Sector of activity in 1998						
Other sector	-0.0802	0.1173	-0.1170	0.1169	-0.1505	0.1159
General public trucking	0.2429*	0.0303	0.1843*	0.0305	0.1658*	0.0303
Bulk public trucking	Reference group		Reference group		Reference group	
Private trucking	0.2775*	0.0257	0.2358*	0.0258	0.2201*	0.0256
Short-term rental firm	0.6832*	0.0482	0.5976*	0.0488	0.5655*	0.0482
Size of fleet						
2			Reference group		Reference group	
3			0.0830*	0.0206	0.0811*	0.0205
4 to 5			0.1467*	0.0206	0.1403*	0.0205
6 to 9			0.2237*	0.0211	0.2184*	0.0209
10 to 20			0.3093*	0.0211	0.2981*	0.0208
21 to 50			0.3137*	0.0225	0.3051*	0.0222
More than 50			0.3182*	0.0219	0.3193*	0.0215
Number of days authorized to drive in previous year	2.0586*	0.0299	2.0414*	0.0299	2.0408*	0.0297
Number of violations of trucking standards in year before						
For overload	0.0949*	0.0115	0.1015*	0.0114	0.0983*	0.0114
For excessive size	0.1434***	0.0862	0.1378	0.0859	0.1514***	0.0861
For poorly secured cargo	0.2011*	0.0357	0.2048*	0.0355	0.2227*	0.0355
For failure to respect service hours	0.1963*	0.0667	0.2012*	0.0664	0.2141*	0.0663
For failure to pass mechanical inspection	0.1685*	0.0300	0.1807*	0.0299	0.1994*	0.0300
For other reasons	0.1709**	0.0744	0.1779**	0.0743	0.1717**	0.0744
Type of vehicle use						
Commercial use including transport of goods without C.T.Q. permit	-0.1802*	0.0212	-0.2139*	0.0212	-0.1901*	0.0212
Transport of other than "bulk" goods	-0.0900*	0.0243	-0.1233*	0.0243	-0.1138*	0.0242
Transport of "bulk" goods	Reference group		Reference group		Reference group	

Explanatory variables	Gamma-Dirichlet model Fleet size omitted		Gamma-Dirichlet model Cylinders omitted		Gamma-Dirichlet model Period omitted	
	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
Type of fuel						
Diesel	Reference group		Reference group		Reference group	
Gas	-0.3993*	0.0137	-0.4945*	0.0134	-0.3976*	0.0136
Other	-0.2933*	0.0738	-0.4617*	0.0734	-0.3058*	0.0736
Number of cylinders						
1 to 5 cylinders	0.2171*	0.0406			0.2162*	0.0402
6 to 7 cylinders	0.3865*	0.0127			0.3782*	0.0126
8 or more than 10 cylinders	Reference group				Reference group	
Number of axles						
2 axles (3,000 to 4,000 kg)	-0.3091*	0.0209	-0.5535*	0.0189	-0.2854*	0.0208
2 axles (> 4,000 kg)	-0.3003*	0.0151	-0.3884*	0.0147	-0.2806*	0.0150
3 axles	-0.1371*	0.0150	-0.1603*	0.0149	-0.1238*	0.0149
4 axles	-0.1302*	0.0191	-0.1551*	0.0190	-0.1315*	0.0190
5 axles	-0.2048*	0.0175	-0.2027*	0.0174	-0.1954*	0.0174
6 axles or more	Reference group		Reference group		Reference group	
Number of violations with demerit points year before						
For speeding	0.1898*	0.0103	0.1974*	0.0103	0.1918*	0.0103
For driving with suspended license	0.3725*	0.0423	0.3816*	0.0421	0.4026*	0.0421
For running a red light	0.3014*	0.0239	0.3178*	0.0239	0.3130*	0.0239
For ignoring stop sign or traffic officer	0.3519*	0.0258	0.3626*	0.0258	0.3620*	0.0258
For not wearing a seat belt	0.1484*	0.0295	0.1559*	0.0294	0.1536*	0.0295
Observation period						
1991	0.1575*	0.0332	0.0442	0.0334		
1992	0.1299*	0.0293	0.0271	0.0295		
1993	0.1431*	0.0260	0.0592**	0.0261		
1994	0.2348*	0.0226	0.1689*	0.0227		
1995	0.2483*	0.0197	0.1978*	0.0198		
1996	0.1502*	0.0175	0.1146*	0.0175		
1997	-0.0590*	0.0163	-0.0813*	0.0163		
1998	Reference group		Reference group			
$\hat{\nu}$	2.0152*	0.0424	1.9876*	0.0415	2.0077*	0.0422
$\hat{\kappa}$	13.3070*	0.2580	13.0769*	0.2556	12.6287*	0.2503
$\hat{\delta}$	4.6682*	0.3100	4.6666*	0.3098	4.6683*	0.3101
Number of observations:	456,117		456,117		456,117	

* Significant at 1%; ** Significant at 5%; *** Significant at 10%

Table C1.3: Estimating sample results for predicted numbers of accidents
Estimation of the parameters of the distribution of the number of annual truck accidents for the 1991-1997 period (fleet of two trucks or more and trucks with two periods or more), Gamma-Dirichlet models.

Explanatory variables	Gamma-Dirichlet model	
	Coefficient	Standard error
Constant	-3.6903*	0.0438
Number of years as carrier at 31 December	-0.0468*	0.0051
Sector of activity in 1998		
Other sector	-0.1260	0.1201
General public trucking	0.1792*	0.0326
Bulk public trucking	Reference group	
Private trucking	0.2336*	0.0275
Short-term rental firm	0.5916*	0.0508
Size of fleet		
2	Reference group	
3	0.0774*	0.0219
4 to 5	0.1357*	0.0219
6 to 9	0.2071*	0.0224
10 to 20	0.2871*	0.0222
21 to 50	0.2817*	0.0237
More than 50	0.3070*	0.0229
Number of days authorized to drive in previous year	2.0100*	0.0317
Number of violations of trucking standards in year before		
For overload	0.0965*	0.0120
For excessive size	0.1423	0.0901
For poorly secured cargo	0.2081*	0.0377
For failure to respect service hours	0.2213*	0.0757
For failure to pass mechanical inspection	0.1877*	0.0315
For other reasons	0.1568***	0.0812
Type of vehicle use		
Commercial use including transport of goods without C.T.Q. permit	-0.1936*	0.0228
Transport of other than "bulk" goods	-0.1061*	0.0261
Transport of "bulk" goods	Reference group	
Type of fuel		
Diesel	Reference group	
Gas	-0.3819*	0.0142
Other	-0.3830*	0.0810
Number of cylinders		
1 to 5 cylinders	0.2319*	0.0433
6 to 7 cylinders	0.3702*	0.0133
8 or more than 10 cylinders	Reference group	
Number of axles		
2 axles (3,000 to 4,000 kg)	-0.2738*	0.0221
2 axles (more than 4,000 kg)	-0.2809*	0.0160
3 axles	-0.1303*	0.0160
4 axles	-0.1421*	0.0203
5 axles	-0.1958*	0.0186
6 axles or more	Reference group	

Explanatory variables	Gamma-Dirichlet model	
	Coefficient	Standard error
Number of violations with demerit points year before		
For speeding	0.1961*	0.0113
For driving with suspended license	0.4088*	0.0449
For running a red light	0.3041*	0.0256
For ignoring stop sign or traffic officer	0.3495*	0.0277
For not wearing a seat belt	0.1684*	0.0306
Observation period	-0.0238*	0.0053
$\hat{\nu}$	2.0657*	0.0480
$\hat{\kappa}$	11.7490*	0.2480
$\hat{\delta}$	4.7158*	0.3412
Number of observations:	393,634	

* Significant at 1%; ** Significant at 5%; *** Significant at 10%.

APPENDIX C2: ESTIMATION RESULTS FOR FLEETS OF MORE THAN FOUR TRUCKS

Table C2.1
Estimation of the parameters of the distribution of the number of annual truck accidents for the 1991-1998 period (fleet of more than four trucks and trucks with two periods or more): Poisson and NB2 models.

Explanatory variables	**Poisson model**		**NB2 model**	
	Coefficient	**Standard deviation**	**Coefficient**	**Standard deviation**
Constant	-3.5145*	0.0495	-3.5211*	0.0524
Number of years as carrier at 31 December	-0.0372*	0.0032	-0.0377*	0.0034
Sector of activity in 1998				
Other sector	-0.3248*	0.0923	-0.3180*	0.0964
General public trucking	0.0913*	0.0242	0.0964*	0.0262
Bulk public trucking	Reference group		Reference group	
Private trucking	0.1714*	0.0214	0.1776*	0.0232
Short-term rental firm	0.4264*	0.0301	0.4416*	0.0329
Size of fleet				
5	Reference group		Reference group	
6 to 9	0.0654*	0.0140	0.0661*	0.0150
10 to 20	0.1622*	0.0132	0.1649*	0.0142
21 to 50	0.1596*	0.0142	0.1644*	0.0153
More than 50	0.1705*	0.0133	0.1720*	0.0142
Number of days authorized to circulate year before	1.7167*	0.0328	1.7231*	0.0339
Number of violations of trucking standards year before				
For overload	0.1375*	0.0119	0.1413*	0.0135
For excessive size	0.1725***	0.0964	0.1786***	0.1071
For poorly secured cargo	0.2669*	0.0374	0.2720*	0.0433
For failure to respect service hours	0.2507*	0.0668	0.2557*	0.0785
For failure to pass mechanical inspection	0.2330*	0.0327	0.2449*	0.0374
For other reasons	0.3083*	0.0758	0.2846*	0.0885
Type of vehicle use				
Commercial use including transport of goods without C.T.Q. permit	-0.0748*	0.0210	-0.0813*	0.0229
Transport of other than "bulk" goods	-0.0065	0.0232	-0.0118	0.0253
Transport of "bulk" goods	Reference group		Reference group	
Type of fuel				
Diesel	Reference group		Reference group	
Gas	-0.3387*	0.0140	-0.3400*	0.0148
Others	-0.2869*	0.0735	-0.2859*	0.0769
Number of cylinders				
1 to 5 cylinders	0.3369*	0.0424	0.3352*	0.0454
6 to 7 cylinders	0.3725*	0.0130	0.3732*	0.0137
8 or more than 10 cylinders	Reference group		Reference group	
Number of axles				
2 axles (3,000 to 4,000 kg)	-0.1840*	0.0202	-0.1859*	0.0215
2 axles (more than 4,000 kg)	-0.1308*	0.0134	-0.1344*	0.0145
3 axles	-0.0678*	0.0137	-0.0723*	0.0148

Explanatory variables	Poisson model		NB2 model	
	Coefficient	**Standard deviation**	**Coefficient**	**Standard deviation**
4 axles	-0.1951*	0.0178	-0.1951*	0.0191
5 axles	-0.1850*	0.0159	-0.1864*	0.0171
6 axles or more	Reference group		Reference group	
Number of violations with demerit points year before				
For speeding	0.2819*	0.0105	0.2930*	0.0122
For driving under suspension	0.5355*	0.0461	0.5713*	0.0558
For running a red light	0.4070*	0.0262	0.4200*	0.0299
For ignoring stop sign or traffic agent	0.4735*	0.0280	0.4843*	0.0321
For not wearing a seat belt	0.1910*	0.0331	0.1969*	0.0367
Observation period				
1991	0.0109	0.0268	0.0146	0.0290
1992	-0.0221	0.0242	-0.0188	0.0262
1993	-0.0817*	0.0224	-0.0811*	0.0241
1994	-0.0147	0.0204	-0.0129	0.0220
1995	0.0044	0.0188	0.0050	0.0202
1996	-0.0373**	0.0177	-0.0355***	0.0191
1997	-0.1443*	0.0176	-0.1438*	0.0189
1998	Reference group		Reference group	
$\hat{\delta}$			0.8032*	0.0203
Number of observations:	336,772		336,772	

* significant at 1%;   ** significant at 5%;   *** significant at 10%

Table C2.2

Estimation of the parameters of the distribution of the number of annual truck accidents for the 1991-1998 period (fleet of more than four trucks and trucks with two periods or more): Hausman's model and Gamma-Dirichlet model.

Explanatory variables	Hausman's model		Gamma-Dirichlet model	
	Coefficient	Standard deviation	Coefficient	Standard deviation
Constant	-0.0290	0.0963	-3.8350*	0.0829
Number of years as carrier at 31 December	-0.0381*	0.0038	-0.0401*	0.0068
Sector of activity in 1998				
Other sector	-0.3001*	0.1068	-0.1768	0.1561
General public trucking	0.0988*	0.0293	0.1442*	0.0402
Bulk public trucking	Reference group		Reference group	
Private trucking	0.1761*	0.0261	0.2470*	0.0358
Short-term rental firm	0.4730*	0.0369	0.5967*	0.0626
Size of fleet				
5	Reference group		Reference group	
6 to 9	0.0648*	0.0161	-0.0004	0.0918
10 to 20	0.1466*	0.0158	0.0522**	0.0219
21 to 50	0.1396*	0.0170	0.0489**	0.0244
More than 50	0.1373*	0.0160	0.0515**	0.0245
Number of days authorized to circulate in year before	1.7256*	0.0338	2.1354*	0.0338
Number of violations of trucking standards in year before				
For overload	0.1099*	0.0133	0.0828*	0.0219
For excessive size	0.1570	0.1030	0.1571	0.0992
For poorly secured cargo	0.2282*	0.0411	0.1786*	0.0397
For failure to respect service hours	0.2329*	0.0732	0.1709**	0.0728
For failure to pass mechanical inspection	0.1807*	0.0359	0.1141*	0.0346
For other reasons	0.2982*	0.0853	0.1800**	0.0806
Type of vehicle use				
Commercial use including transport of goods without C.T.Q. permit	-0.1004*	0.0254	-0.1646*	0.0251
Transport of other than "bulk" goods	-0.0300	0.0280	-0.1009*	0.0278
Transport of "bulk" goods	Reference group		Reference group	
Type of fuel				
Diesel	Reference group		Reference group	
Gas	-0.3521*	0.0167	-0.3509*	0.0152
Others	-0.2782*	0.0840	-0.2652*	0.0789
Number of cylinders				
1 to 5 cylinders	0.3366*	0.0526	0.1441*	0.0462
6 to 7 cylinders	0.3724*	0.0156	0.3695*	0.0141
8 or more than 10 cylinders	Reference group		Reference group	

Explanatory variables	Hausman's model		Gamma-Dirichlet model	
	Coefficient	Standard deviation	Coefficient	Standard deviation
Number of axles				
2 axles (3,000 to 4,000 kg)	-0.1852*	0.0242	-0.3573*	0.0237
2 axles (more than 4,000 kg)	-0.1505*	0.0166	-0.3135*	0.0167
3 axles	-0.1088*	0.0170	-0.0963*	0.0167
4 axles	-0.2013*	0.0218	-0.1174*	0.0208
5 axles	-0.1968*	0.0190	-0.1768*	0.0193
6 axles or more	Reference group		Reference group	
Number of violations with demerit points year before				
For speeding	0.2433*	0.0118	0.1719*	0.0115
For driving under suspension	0.4705*	0.0519	0.3862*	0.0495
For running a red light	0.3392*	0.0286	0.2697*	0.0272
For ignoring stop sign or traffic agent	0.4042*	0.0306	0.3337*	0.0292
For not wearing a seat belt	0.1659*	0.0357	0.1211*	0.0341
Observation period				
1991	0.0231	0.0309	0.1004**	0.0496
1992	-0.0151	0.0276	0.0833***	0.0433
1993	-0.0761*	0.0251	0.1252*	0.0375
1994	-0.0111	0.0225	0.2314*	0.0316
1995	0.0091	0.0203	0.2498*	0.0262
1996	-0.0319***	0.0188	0.1615*	0.0216
1997	-0.1409*	0.0182	-0.0464**	0.0188
1998	Reference group		Reference group	
$\hat{a}$	57.9375*	4.0818		
$\hat{b}$	1.8363*	0.0420		
$\hat{\nu}$			1.9181*	0.0416
$\hat{\kappa}$			22.0245*	0.5539
$\hat{\delta}$			4.7245*	0.3430
Number of observations:	336,772		336,772	

* significant at 1%; ** significant at 5%; *** significant at 10%

APPENDIX D: MODEL ESTIMATIONS, R-CODE

```
#
#Read the dataset
#
library(foreign, pos=15)
Dataset <- read.table("donnee.csv", header=TRUE, sep=",",na.strings="NA", dec=".", strip.white=TRUE)
library(abind, pos=16)
library(e1071, pos=17)
library("BMS")
library("spuRs")
library("MASS")

# Poisson model estimation
GLM.2 <- glm(NB_ATOT ~ AN_TRANS + SECT_14 + SECT_05 + SECT_07 + SECT_08 +
                  N_VH3 + N_VH45 + N_VH69 + N_VH20 + N_VH50 + N_VH51 +
                  DUREE_AT + NB_INF1 + NB_INF2 + NB_INF3 + NB_INF6 + NB_INF7 + NB_INF89 +
                  COMPR + TBRGN + ESSENCE + CARB_AUT + CYL1_5 + CYL6_7 +
                  ESS_02 + ESS_02P + ESS_03 + ESS_04 + ESS_05 +
                  VIT + SANCT + ROUGE + ARRET + CEINTURE +
                  an_91 + an_92 + an_93 + an_94 + an_95 + an_96 + an_97,
          family=poisson(log), data=Dataset)
est=GLM.2$coefficients

# Negative Binomial model estimation
GLM.3 <- glm.nb(NB_ATOT ~ AN_TRANS + SECT_14 + SECT_05 + SECT_07 + SECT_08 +
               N_VH3 + N_VH45 + N_VH69 + N_VH20 + N_VH50 + N_VH51 +
               DUREE_AT + NB_INF1 + NB_INF2 + NB_INF3 + NB_INF6 + NB_INF7 + NB_INF89 +
               COMPR + TBRGN + ESSENCE + CARB_AUT + CYL1_5 + CYL6_7 +
               ESS_02 + ESS_02P + ESS_03 + ESS_04 + ESS_05 +
               VIT + SANCT + ROUGE + ARRET + CEINTURE +
               an_91 + an_92 + an_93 + an_94 + an_95 + an_96 + an_97,
          start=est,init.theta=1, data=Dataset)
estNB=GLM.3$coefficients

# Negative Binomial model with random effect  (Hausman’s model) estimation
#Create the vector y: number of accident of truck i at time t
y <- as.matrix(cbind(Dataset$NB_ATOT))
max_y <- max(y)
s_y <- sum(y)


#Create the matrix x : Variables concerning the carriers, vehicles and the drivers (a vehicle may have more than one
#driver
x <- as.matrix(cbind(1, Dataset$AN_TRANS, Dataset$SECT_14, Dataset$SECT_05,  Dataset$SECT_07,
                   Dataset$SECT_08, Dataset$N_VH3, Dataset$N_VH45, Dataset$N_VH69, Dataset$N_VH20,
                   Dataset$N_VH50, Dataset$N_VH51, Dataset$DUREE_AT, Dataset$NB_INF1,
                   Dataset$NB_INF2, Dataset$NB_INF3, Dataset$NB_INF6,  Dataset$NB_INF7,
                   Dataset$NB_INF89, Dataset$COMPR, Dataset$TBRGN, Dataset$ESSENCE,
                   Dataset$CARB_AUT,  Dataset$CYL1_5, Dataset$CYL6_7, Dataset$ESS_02, Dataset$ESS_02P,
                   Dataset$ESS_03, Dataset$ESS_04, Dataset$ESS_05, Dataset$VIT, Dataset$SANCT,
                   Dataset$ROUGE, Dataset$ARRET, Dataset$CEINTURE, Dataset$an_91, Dataset$an_92,
                   Dataset$an_93, Dataset$an_94, Dataset$an_95,  Dataset$an_96, Dataset$an_97))

n <- nrow(x)       # Total number of observations
p <- ncol(x)       # Number of parameters
p1 <- p+1
```

```
p2 <- p+2

nper<- as.matrix(cbind(Dataset$n_period,Dataset$camion))
nper1<-nper[Dataset$camion == 1,]
n_period<-cbind(nper1[,1])

#number of trucks
ki <- nrow(n_period)

#Initial values
r_beta <- c(est, 57, 1.8)

#Log likelihood function
llf <- function (r_beta) {

  parp <- r_beta[1:p]
  a <- r_beta[p1]
  b <- r_beta[p2]

  r_llf <- 0
  nx <- 0
  for (j in 1:ki){
    rl <- 0
    ni <- n_period[j]
    i_deb <- nx+1
    i_fin <- nx+ni
    ri <- i_deb : i_fin
    nx <- i_fin
    yi <- y[ri]
    xi <- x[ri,]
    zi <- xi%*%parp
    mui <- exp(zi)
    s_mui <- sum(mui)
    s_yi <- sum(yi)
    ter_1 <- lgamma(a+b) + lgamma(a+s_mui) + lgamma(b+s_yi)
    ter_2 <- lgamma(a) + lgamma(b) + lgamma(a+b+s_yi+s_mui)
    ter_3 <- lgamma(mui+yi) - lgamma(mui) - lgamma(yi+1)
    s_ter_3=sum(ter_3)
    rl <- ter_1-ter_2+s_ter_3
    r_llf=r_llf+rl
  }
 return(r_llf)
}

#Gradient function
llg <- function (r_beta) {
  parp <- r_beta[1:p]
  a <- r_beta[p1]
  b <- r_beta[p2]
  r_llg <- matrix(0,1,p2)
  lla <- 0
  llb <- 0
  llp <-  matrix(0,1,p)
  nx <- 0
  for (j in 1:ki){
    ter1 <- 0
    ter2 <- 0
    ter3 <- matrix(0,p,1)
    ni <- n_period[j]
    i_deb <- nx+1
    i_fin <- nx+ni
    ri <- i_deb : i_fin
    nx <- i_fin
    yi <- y[ri]
    xi <- x[ri,]
    zi <- xi%*%parp
    mui <- exp(zi)
    s_mui <- sum(mui)
    s_yi <- sum(yi)
    ter1 <- digamma(a+b) +digamma(a+s_mui) -digamma(a) - digamma(a+b+s_yi+s_mui)
    lla = lla +ter1
    ter2 <- digamma(a+b) + digamma(b+s_yi) - digamma(b) - digamma(a+b+s_yi+s_mui)
    llb = llb+ter2
    ter3 <- (t(xi)%*%mui)*(digamma(a+s_mui)) - (t(xi)%*%mui)*(digamma(a+b+s_yi+s_mui))
    ter3mui <- matrix(0,ni,p)
    for (iii in 1:p) {
      ter3mui[,iii] <- (xi[,iii]*mui[,1])
    }

    ter3b <- t(ter3mui)%*%(digamma(mui+yi) - digamma(mui))
    ter3c <- ter3 + ter3b
    llp=llp+t(ter3c)

  }
  r_llg[1:p]=llp
  r_llg[p1]=lla
  r_llg[p2]=llb
  return(r_llg)

}

#Lower bound of the parameters
Low <- c( "-inf", "-inf", "-inf", "-inf",  "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf",
          "-inf", "-inf",  "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf",
          "-inf", "-inf",  "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf",
          "-inf", "-inf",  "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf", "-inf",
           1e-8, 1e-8)
#Optimization
hausman <- optim(r_beta, fn=llf, gr=llg, method="L-BFGS-B", lower=Low, control=list(trace=5, fnscale=-1,
report=1, maxit=1000), hessian=TRUE)

hausman$convergence  # the method  converge
#Result
parlistL<-hausman$par
print(parlistL)

#Hessian at the optimal values
hessL<-hausman$hessian
inv_hessL<- solve(hessL)

#Standard errors
 r_stder=sqrt(abs(diag(inv_hessL)))
print(r_stder)

t_ratio=parlistL/r_stder
```

```
#P-values
n_df=n-p2
p_value=2*(1-pt(abs(t_ratio),df=n_df))
print(p_value)

# Poisson fixed effect model estimation
#
GLMmboot.2 <- glmmboot(NB_ATOT ~ NB_INF1 + NB_INF2 + NB_INF3 + NB_INF6 + NB_INF7 +
                       NB_INF89+ VIT + SANCT + ROUGE + ARRET + CEINTURE + N_VH69 + N_VH20
                       + N_VH50 + N_VH51 +an_91 + an_92 + an_93 + an_94 + an_95 + an_96 + an_97,
                       family = poisson(log), data=Dataset, cluster=TRNIP,
                       start.coef = NULL, control = list(epsilon = 1e-08, maxit = 200, trace = FALSE))
summary(GLMmboot.2)
#
#Gamma-Dirichlet model estimation#
#Create the vector y: number of accident of truck i at time t
y <- as.matrix(cbind(Dataset$NB_ATOT))
max_y <- max(y)

#Create the matrix x : Variables concerning the carriers, vehicles and the drivers (a vehicle may have more than one
#driver
x <- as.matrix(cbind(1, Dataset$AN_TRANS, Dataset$SECT_14, Dataset$SECT_05,  Dataset$SECT_07,
                Dataset$SECT_08, Dataset$N_VH3, Dataset$N_VH45, Dataset$N_VH69, Dataset$N_VH20,
                Dataset$N_VH50, Dataset$N_VH51, Dataset$DUREE_AT, Dataset$NB_INF1,
                Dataset$NB_INF2, Dataset$NB_INF3, Dataset$NB_INF6,  Dataset$NB_INF7,
                Dataset$NB_INF89, Dataset$COMPR, Dataset$TBRGN, Dataset$ESSENCE,
                Dataset$CARB_AUT,  Dataset$CYL1_5, Dataset$CYL6_7, Dataset$ESS_02, Dataset$ESS_02P,
                Dataset$ESS_03, Dataset$ESS_04, Dataset$ESS_05, Dataset$VIT, Dataset$SANCT,
                Dataset$ROUGE, Dataset$ARRET, Dataset$CEINTURE, Dataset$an_91, Dataset$an_92,
                Dataset$an_93, Dataset$an_94, Dataset$an_95,  Dataset$an_96, Dataset$an_97))

head(x)
n <- nrow(x)      # Total number of observations
p <- ncol(x)      # Number of parameters
p1=p+1
p2=p+2
p3=p+3

per_max <- 8
n_kappa <- 1

#Create indcam  matrix: Equal to 1 if the truck i is present at the year  t; 0 otherwise
indc <- as.matrix(cbind(Dataset$ind1, Dataset$ind2,  Dataset$ind3, Dataset$ind4, Dataset$ind5, Dataset$ind6,
        Dataset$ind7, Dataset$ind8, Dataset$camion))
indc1<-indc[Dataset$camion == 1,]
indcam<-indc1[,1:8]

ncamion <- nrow(indcam) # Total number of trucks
pcam <- ncol(indcam)  # Maximum number of  observed periods

#Create the vector period: Equal to 1 if the truck i is present at the year  1991 ; ….; Equal to 8 1 if the truck i is
present at the year  1998
period <- as.matrix(cbind(Dataset$PERIOD))
nperf <- as.matrix(cbind(Dataset$nper_f,Dataset$FLOTTE))
nperf1<-nperf[Dataset$FLOTTE == 1,]
nper_f<-cbind(nperf1[,1]) #Number of periods per firm
nflotte <- nrow(nper_f) # Total number of firms
```

```
taillet<- as.matrix(cbind(Dataset$taille_t,Dataset$FLOTTE))
taillet1<-taillet[Dataset$FLOTTE == 1,]
tt<-cbind(taillet1[,1])  # Number of trucks per firm

taillec<- as.matrix(cbind(Dataset$taille_c,Dataset$FLOTTE))
taillec1<-taillec[Dataset$FLOTTE == 1,]
ttc<-cbind(taillec1[,1]) # Number of truck-years per  firm

nper<- as.matrix(cbind(Dataset$n_period,Dataset$camion))
nper1<-nper[Dataset$camion == 1,]
n_period<-cbind(nper1[,1]) # Number of  periods per truck

# We divide the vehicles into two groups (high risk and low risk)
grp <- as.matrix(cbind(Dataset$grp,Dataset$camion))
grp1<-grp[Dataset$camion == 1,]
grpc<-cbind(grp1[,1])

#A vector of starting values
nu<- 2.06
kap<- 12.65
del<- 4.67
parlist<-c(nu,kap,del,betaest)
newparlist<-c(4,8,3,estNB)

#Beginning of the R and C++ interface
library(Rcpp)
library(inline)
foo <- paste(readLines("flotte_8periode_1v_1k_1d_ttf_maxdiff_chg_grp.cc"),collapse="\n")
fx <- cxxfunction(signature(),plugin="Rcpp",include=foo)
tclass <- Module("test",getDynLib(fx))
Vraisemblance <- tclass$Vraisemblance

#Initialisation Vraisemblance object
vsemblance <- new(Vraisemblance, ttc, tt, nflotte, p, n, per_max, y)
vsemblance$init_x(x)
vsemblance$init_indcam(ncamion, indcam)

#Optimisation, A quasi-Newton medthod (BFGS)
essaiL<-optim(newparlist,vsemblance$r_llf,method="BFGS",control=list(trace=5, fnscale=-1, report=1,
maxit=1000), hessian=TRUE)

#Optimation results
essaiL$convergence  # The method converged
parlistL<-essaiL$par
hessL<-essaiL$hessian
print(parlistL)
inv_hessL<- solve(hessL)  # Invert the  Hessian matrix
r_stder=sqrt(abs(diag(inv_hessL))) #Standard errors
print(r_stder)
t_ratio=parlistL/r_stder
n_df=n-p3
p_value=2*(1-pt(abs(t_ratio),df=n_df))
print(p_value)
```

#C++ files : flotte_8periode_1v_1k_1d_ttf_maxdiff_chg_grp.cc

```
#include <Rcpp.h>
using namespace Rcpp;
```

```
#include <iostream>
#include <iomanip>
#include <fstream>
#include <vector>
#include <numeric>
#include <algorithm>

#include <math.h>
#include <values.h>

#include <unistd.h>
#include <stdlib.h>
#include <sys/types.h>
#include <sys/stat.h>
#include <fcntl.h>
#include <strings.h>

#include <gsl/gsl_sf_hyperg.h>
#include <gsl/gsl_sf_psi.h>
#include <gsl/gsl_sort_double.h>
#include <gsl/gsl_statistics.h>
#include <gsl/gsl_errno.h>

/************************
To compile the program we must include this file:
~/.R/Makevars ceci

PKG_LIBS=-L/home/apps/Logiciels/GSL/1.16/lib -lgsl -lgslcblas $(shell "/home/apps/Logiciels/R/3.2.1-gcc/bin/Rscript" -e "Rcpp:::LdFlags()")

It is also necessary to load these modules
 1) R/3.2.1-gcc   2) GSL/1.16
*******************************/

#ifdef _OPENMP
#include <omp.h>
#endif

typedef std::vector<double> My_Vector;

class Vraisemblance
{
public:
  Vraisemblance(My_Vector ttc_, My_Vector tt_, double nflotte_, double p_, double n_, double per_max_, My_Vector y_);
  void init_x(NumericVector x_);
  void init_indcam(int ncamion, NumericVector indcam_);

  double   r_llf(My_Vector& r_beta);
  double   r_llf_autrestermes(double delta);

  void     chg_grp_moyenne(My_Vector& r_beta);
  void     chg_grp_mediane(My_Vector& r_beta);
  void     chg_grp_maxdiff(My_Vector& r_beta);


private:
  //variables from  R
  int *ttc;     //number of truck per firm
  int *tt;      //number of truck-years per firm
  int p;        //number of parameters
  int n;        // Total number of observations
  int nb_alpha; //number of  kappa (1/alpha)
  int per_max; // maximum number of observed periods
  double *y;
  double **x;   //truck characteristics
  double** indcam;   //indicate the period (year) the truck is present
  double *d;
  int nflotte;  // total number of firms
  int *grp;     / It indicate in  which group the truck is

  //Variables internes
  int *vect_nx;
  int *vect_nxc;
  int *nb_camion_par_annee;

  double  *mui;    // mui = exp ( x * beta)
  double *zi;      // for the intermediate calculation of mui;
  double *s_yi;     // sum of yi for each fleet
  double *s_vij;   // sum of vij for each fleet
};

Vraisemblance::Vraisemblance( My_Vector ttc_, My_Vector tt_, double nflotte_, double p_, double n_, double
per_max_, My_Vector y_ ) :
  nflotte(static_cast<int> (nflotte_)),
  p(static_cast<int> (p_)),
  n(static_cast<int> (n_)),
  nb_alpha(1),
  per_max(static_cast<int> (per_max_))
{
  gsl_set_error_handler_off ();

  tt = new int[nflotte];
  ttc = new int[nflotte];
  std::cout << "nflotte " << nflotte << '\n';
  std::cout << "p " << p << '\n';
  std::cout << "n " << n << '\n';
  std::cout << "permax " << per_max << '\n';

  for (int i=0; i<nflotte; i++)
    {
      tt[i] = static_cast<int>(tt_[i]);
      ttc[i] = static_cast<int>(ttc_[i]);
    }

  int y_size = y_.size();
  y = new double[y_size];
  for (int i=0; i<y_size; i++)
      y[i]=y_[i];

  vect_nx = new int[nflotte];
  vect_nx[0] = 0;
  for (int i=1; i<nflotte; i++)
    vect_nx[i] = vect_nx[i-1] + tt[i-1];
  vect_nxc = new int[nflotte];
  vect_nxc[0] = 0;
  for (int i=1; i<nflotte; i++)
    vect_nxc[i] = vect_nxc[i-1] + static_cast<int>(ttc[i-1]);

  mui = new double[n];
  zi = new double[n];
  s_yi = new double[nflotte];
  s_vij = new double[nflotte];

  grp = new int[n];

  for (int f=0; f<nflotte; f++)
    {
      int tt_f = tt[f];
      int nx_t=vect_nx[f];

      s_yi[f] = std::accumulate(y+nx_t, y+nx_t+tt_f, 0.0);
    }

  nb_camion_par_annee = new int[nflotte];
  for (int i=0; i<nflotte; i++)
    {
      int tt_f =static_cast<int>(tt[i]);
      int nx_t = vect_nx[i];

      nb_camion_par_annee[i] = tt_f;

      if (tt_f==2)
        {
                grp[nx_t] = 0;
                grp[nx_t+1] = 1;
        }
    }
}

//conversion of NumericVector in the matrix x
void Vraisemblance::init_x(NumericVector x_)
{
  x = new double*[n];
  for (int i=0; i<n; i++)
    x[i] = new double[p];

  int k=0;
  for (int j=0; j<p; j++)
    for (int i=0; i<n; i++)
      {
              x[i][j] = x_[k];
              k++;
      }
}

//conversion of NumericVector in the matrix indam
void Vraisemblance::init_indcam(int ncamion, NumericVector indcam_)
{
  indcam = new double*[ncamion];
  for (int i=0; i<ncamion; i++)
    indcam[i] = new double[per_max];
  int k=0;
  for (int j=0; j<per_max; j++)
    for (int i=0; i<ncamion; i++)
      {
              indcam[i][j] = indcam_[k];
              k++;
      }
}

//Function to evaluate
double Vraisemblance::r_llf(My_Vector& r_beta)
{
  double *v = &r_beta[0];
  double *alpha = &r_beta[1];
  double delta = r_beta[2];
  int    debut_beta = 1 + nb_alpha + 1;
  double *beta = &r_beta[debut_beta];

  /* reading the observations of carrier f */

  double r_ll_f1=0;
  double r_ll_f2=0;
  double r_ll_f3=0;
  double r_ll_f=0;

  //int size_x_col = x[0].size();

  //My_Vector zi(ni);
  //zi=xi*par`;
  //zi += x_ptr[j] * beta[j];
  for (int f=0; f<nflotte; f++)
    {
      int ni = static_cast<int>(tt[f]);
      int nx =  vect_nx[f];
      for (int i=0; i<ni; i++)
              {
                zi[nx+i]=0.0;
                for (int k=0; k<p; k++)
                  zi[nx+i] += x[nx+i][k] * beta[k];
              }
    }

#pragma omp parallel for
  //My_Vector mui(ni);
  //=(di#exp(zi));
  for (int i=0; i<n; i++)
    mui[i] = exp (zi[i]);
#pragma omp parallel for reduction(+:r_ll_f1, r_ll_f2, r_ll_f3) schedule (static, 10)
  for (int f=0; f<nflotte; f++)
    {
      int ttc_f=static_cast<int>(ttc[f]);
      int tt_f =static_cast<int>(tt[f]);
      int nx_c = vect_nxc[f];
      int nx_t = vect_nx[f];

      //there is only one  alpha
      int ind_kappa = 0;
```

```
double min_mui= *std::min_element(mui+nx_t, mui+nx_t+tt_f);//min(mui);
double max_mui= *std::max_element(mui+nx_t, mui+nx_t+tt_f);//max(mui);
double s_mui = std::accumulate(mui+nx_t, mui+nx_t+tt_f, 0.0);//sum(mui);

s_vij[f]=0;
double lgamma_v = 0;
double present_v;
for (int i=0; i<ttc_f; i++)
        {
          s_vij[f] += v[0];
          lgamma_v += lgamma(v[0]);
        }

double s_y_mui = 0;
for (int i=0; i<tt_f; i++)
        s_y_mui += y[nx_t+i]*log(mui[nx_t+i]);


{
        double g1=0;
        double g2=0;
        double s_yi1=0;
        double s_yi2=0;
        double s_vi1=0;
        double s_vi2=0;
        double s_mui1=0;
        double s_mui2=0;

        int ind = 0;
        for (int i=0; i<ttc_f; i++)
          {
            double somme_y=0;
            double somme_v=v[0];
            double somme_mui=0;
            double nb = 0;
            int    groupe;
            for (int j=0; j<per_max; j++)
              if (indcam[nx_c+i][j]>0)
                      {
                        groupe = grp[nx_t+ind];
                        somme_y += y[nx_t+ind];
                        somme_mui += mui[nx_t+ind];
                        nb++;
                        ind++;
                      }

            if (groupe == 0)
              {
                      g1++;
                      s_yi1 += somme_y;
                      s_vi1 += somme_v;
                      s_mui1 += somme_mui/nb;
              }
            else
              {
                      g2++;
                      s_yi2 += somme_y;
                      s_vi2 += somme_v;
                      s_mui2 += somme_mui/nb;
              }
          }


        double mui1, mui2;
        if (g1 > 0)
          mui1=s_mui1/g1;
        else
          mui1=s_mui2/g2;
        if (g2 > 0)
          mui2=s_mui2/g2;
        else
          mui2=mui1;

        double par1, par2;
        double ter_11;
        double par3 = s_yi[f] + s_vij[f];
        double par4 = s_yi[f]+nb_camion_par_annee[f]/alpha[ind_kappa];

        if (mui1 <= mui2)
          {
            if (g1>0)
              par2=s_yi1+s_vi1;
            else
              par2=s_yi2+s_vi2;
            par1=((mui2-mui1)/(1/alpha[ind_kappa]+mui2));
            ter_11=par4*log(1+alpha[ind_kappa]*mui2);
          }
        else if (mui1 > mui2)
          {
            par2=s_yi2+s_vi2;
            par1=((mui1-mui2)/(1/alpha[ind_kappa]+mui1));
            ter_11=par4*log(1+alpha[ind_kappa]*mui1);
          }

        double f_hyp1=gsl_sf_hyperg_2F1(par2,par4,par3,par1);
        double ter_1=
          s_yi[f]*log(alpha[ind_kappa])
          +lgamma(par4)
          -lgamma(nb_camion_par_annee[f]/alpha[ind_kappa])
          +lgamma(s_vij[f])
          +s_y_mui
          -lgamma(par3)
          -ter_11
          +log(f_hyp1);


        double s_ter_3= -lgamma_v;
        for (int i=0; i<tt_f; i++)
          s_ter_3 -= lgamma(y[nx_t+i]+1);

        ind = 0;
        for (int i=0; i<ttc_f; i++)
          {
            double syv = v[0];
            for (int j=0; j<per_max; j++)
```

```
                    if (indcam[nx_c+i][j]>0)
                            {
                              syv += y[nx_t+ind];
                              ind++;
                            }
                  s_ter_3 += lgamma(syv);
                }

              double rl3 = ter_1 + s_ter_3;
              r_ll_f3 += rl3;
      }
    }

  r_ll_f = r_ll_f1 + r_ll_f2 + r_ll_f3 + r_llf_autrestermes(delta);

  if (!finite(r_ll_f))
      r_ll_f=-1000000000000.5;

  return (r_ll_f);

  //end  r_llf
}


double Vraisemblance::r_llf_autrestermes(double delta)
{
  double r_ll_f_temps=0;

#pragma omp parallel for reduction(+:r_ll_f_temps) schedule (static, 10)
  for (int f=0; f<nflotte; f++)
    {
      int ttc_f=static_cast<int>(ttc[f]);
      int tt_f =static_cast<int>(tt[f]);
      int nx_c = vect_nxc[f];
      int nx_t = vect_nx[f];

      double lgamma_d = 0;
      for (int i=0; i<ttc_f; i++)
              {
                for (int j=0; j<per_max; j++)
                  if (indcam[nx_c+i][j]>0)
                    lgamma_d += lgamma(delta);
              }

      int ind = 0;
      double lgamma_sd=0;
      double lgamma_sy_plus_sd=0;
      for (int i=0; i<ttc_f; i++)
              {
                double somme_delta = 0;
                double somme_y = 0;
                for (int j=0; j<per_max; j++)
                  {
                    if (indcam[nx_c+i][j]>0)
                            {
                              somme_delta += delta;
                              somme_y += y[nx_t+ind];
                              ind++;
                            }
                  }
                lgamma_sd += lgamma(somme_delta);
                lgamma_sy_plus_sd += lgamma(somme_y + somme_delta);
              }


      ind = 0;
      double lgamma_y_plus_d = 0;
      for (int i=0; i<ttc_f; i++)
              for (int j=0; j<per_max; j++)
                if (indcam[nx_c+i][j]>0)
                  {
                    lgamma_y_plus_d += lgamma(y[nx_t+ind] + delta);
                    ind++;
                  }

      r_ll_f_temps += (lgamma_y_plus_d + lgamma_sd - lgamma_d - lgamma_sy_plus_sd);
          }

  return (r_ll_f_temps);
}
void Vraisemblance::chg_grp_moyenne(My_Vector& r_beta)
{
  double *v = &r_beta[0];
  double *alpha = &r_beta[1];
  double delta = r_beta[2];
  int    debut_beta = 1 + nb_alpha + 1;
  double *beta = &r_beta[debut_beta];

  //My_Vector zi(ni);
  //zi=xi*par`;
  //zi += x_ptr[j] * beta[j];
  for (int f=0; f<nflotte; f++)
    {
      int ni = static_cast<int>(tt[f]);
      int nx =  vect_nx[f];
      for (int i=0; i<ni; i++)
              {
                zi[nx+i]=0.0;
                for (int k=0; k<p; k++)
                  zi[nx+i] += x[nx+i][k] * beta[k];
              }
    }

#pragma omp parallel for
  //My_Vector mui(ni);
  //=(di#exp(zi));
  for (int i=0; i<n; i++)
    mui[i] = exp (zi[i]);

  //Compute the mean of  mui
  for (int f=0; f<nflotte; f++)
    {
      double somme_mui = 0;
      int ttc_f=static_cast<int>(ttc[f]);
      int tt_f =static_cast<int>(tt[f]);
      int nx_c = vect_nxc[f];
      int nx_t = vect_nx[f];

      for (int cam=0; cam<tt_f; cam++)
              somme_mui += mui[nx_t+cam];

      double moyenne_mui = somme_mui / tt_f;

      //Sum of mui
      int ind1 = 0;
      int ind2 = 0;
      for (int i=0; i<ttc_f; i++)
              {
                somme_mui = 0.0;
                int nb_annee = 0;
                for (int j=0; j<per_max; j++)
                  if (indcam[nx_c+i][j]>0)
                    {
                            somme_mui += mui[nx_t+ind1];
                            nb_annee++;
                            ind1++;
                    }
                double moyenne_mui_cam = somme_mui / nb_annee;

                for (int j=0; j<per_max; j++)
                  if (indcam[nx_c+i][j]>0)
                    {
                            if (moyenne_mui_cam <= moyenne_mui)
                              grp[nx_t+ind2] = 0;
                            else
                              grp[nx_t+ind2] = 1;
                            ind2++;
                    }
              }
    }
}
  //end chg_grp


void Vraisemblance::chg_grp_mediane(My_Vector& r_beta)
{
  double *v = &r_beta[0];
  double *alpha = &r_beta[1];
  double delta = r_beta[2];
  int    debut_beta = 1 + nb_alpha + 1;
  double *beta = &r_beta[debut_beta];

  //My_Vector zi(ni);
  //zi=xi*par`;
  //zi += x_ptr[j] * beta[j];
  for (int f=0; f<nflotte; f++)
    {
      int ni = static_cast<int>(tt[f]);
      int nx =  vect_nx[f];
      for (int i=0; i<ni; i++)
              {
                zi[nx+i]=0.0;
                for (int k=0; k<p; k++)
                  zi[nx+i] += x[nx+i][k] * beta[k];
```

```
       }
  }

#pragma omp parallel for
 //My_Vector mui(ni);
 //=(di#exp(zi));
 for (int i=0; i<n; i++)
  mui[i] = exp (zi[i]);

 //Compute the mean of  mu
 for (int f=0; f<nflotte; f++)
  {
   int ttc_f=static_cast<int>(ttc[f]);
   int tt_f =static_cast<int>(tt[f]);
   int nx_c = vect_nxc[f];
   int nx_t = vect_nx[f];

   int ind1 = 0;
   std::vector<double> mediane;
   for (int i=0; i<ttc_f; i++)
       {
        double somme_mui = 0.0;
        int nb_annee=0;
        for (int j=0; j<per_max; j++)
         if (indcam[nx_c+i][j]>0)
          {
              somme_mui += mui[nx_t+ind1];
              nb_annee++;
              ind1++;
          }
        mediane.push_back(somme_mui/nb_annee);
       }

   std::sort(mediane.begin(), mediane.end());

   double val_mediane = mediane[mediane.size()/2];

   //calcul des somme des mui sur les camions
   ind1=0;
   int ind2 = 0;
   for (int i=0; i<ttc_f; i++)
       {
        double somme_mui = 0.0;
        int nb_annee=0;
        for (int j=0; j<per_max; j++)
         if (indcam[nx_c+i][j]>0)
          {
              somme_mui += mui[nx_t+ind1];
              ind1++;
          }

        for (int j=0; j<per_max; j++)
         if (indcam[nx_c+i][j]>0)
          {
              if (somme_mui/nb_annee < val_mediane)
               grp[nx_t+ind2] = 0;
              else
               grp[nx_t+ind2] = 1;
```

```
                            ind2++;
                    }
              }
    }
}

template <class random_iterator>
class IndexedComparison
{
public:
  IndexedComparison (random_iterator begin,
                                 random_iterator end)
    : p_begin (begin), p_end (end) {}
  bool operator () (unsigned int a, unsigned int b) const
  { return *(p_begin + a) < *(p_begin + b); }

private:
  random_iterator const p_begin;
  random_iterator const p_end;
};

void Vraisemblance::chg_grp_maxdiff(My_Vector& r_beta)
{
  double *v = &r_beta[0];
  double *alpha = &r_beta[1];
  double delta = r_beta[2];
  int    debut_beta = 1 + nb_alpha + 1;
  double *beta = &r_beta[debut_beta];

  //My_Vector zi(ni);
  //zi=xi*par`;
  //zi += x_ptr[j] * beta[j];
  for (int f=0; f<nflotte; f++)
    {
      int ni = static_cast<int>(tt[f]);
      int nx =  vect_nx[f];
      for (int i=0; i<ni; i++)
              {
                zi[nx+i]=0.0;
                for (int k=0; k<p; k++)
                  zi[nx+i] += x[nx+i][k] * beta[k];
              }
    }

#pragma omp parallel for
  //My_Vector mui(ni);
  //=(di#exp(zi));
  for (int i=0; i<n; i++)
    mui[i] = exp (zi[i]);

  //Compute the mean of mu
  for (int f=0; f<nflotte; f++)
    {
      int ttc_f=static_cast<int>(ttc[f]);
      int tt_f =static_cast<int>(tt[f]);
      int nx_c = vect_nxc[f];
      int nx_t = vect_nx[f];
int ind1 = 0;
std::vector<double> moymui;
for (int i=0; i<ttc_f; i++)
    {
     double somme_mui = 0.0;
     int nb_annee=0;
     for (int j=0; j<per_max; j++)
       if (indcam[nx_c+i][j]>0)
        {
            somme_mui += mui[nx_t+ind1];
            nb_annee++;
            ind1++;
        }
     moymui.push_back(somme_mui/nb_annee);
    }

std::vector<unsigned int> indices(ttc_f);
for (int i = 0; i < indices.size (); i++)
    indices [i] = i;

std::sort (indices.begin (), indices.end (),
            IndexedComparison<std::vector<double>::const_iterator>
            (moymui.begin(), moymui.end()));

double val_mui = 0;
double maxdiff = -1e300;
for (int i=0; i<ttc_f-1; i++)
    {
     double diff = moymui[indices[i+1]] - moymui[indices[i]];
     if (diff > maxdiff)
       {
        val_mui = moymui[indices[i+1]];
        maxdiff = diff;
       }
    }

float nbCamTot=0;
float nbCamGrp1=0;
//Compute the sum of  mui on trucks
ind1=0;
int ind2 = 0;
for (int i=0; i<ttc_f; i++)
    {
     double somme_mui = 0.0;
     int nb_annee=0;
     for (int j=0; j<per_max; j++)
       if (indcam[nx_c+i][j]>0)
        {
            somme_mui += mui[nx_t+ind1];
            nb_annee++;
            ind1++;
        }

     for (int j=0; j<per_max; j++)
       if (indcam[nx_c+i][j]>0)
        {
            nbCamTot += 1.0;
```

```
                            if (somme_mui/nb_annee < val_mui)
                              {
                                grp[nx_t+ind2] = 0;
                                nbCamGrp1 += 1.0;
                              }
                            else
                              grp[nx_t+ind2] = 1;
                            ind2++;
                    }
              }
    }

}


RCPP_MODULE(test){
  class_<Vraisemblance>( "Vraisemblance" )
    //.constructor()
    .constructor< My_Vector, My_Vector, double, double, double, double, My_Vector >()
    .method( "init_x", &Vraisemblance::init_x)
    .method( "init_indcam", &Vraisemblance::init_indcam)
    .method( "r_llf", &Vraisemblance::r_llf )
    .method( "chg_grp_moyenne", &Vraisemblance::chg_grp_moyenne )
    .method( "chg_grp_mediane", &Vraisemblance::chg_grp_mediane )
    .method( "chg_grp_maxdiff", &Vraisemblance::chg_grp_maxdiff )
    ;
}
```

APPENDIX E: PREDICTIVE PROBABILITIES

Table E1 presents an example of predictive probabilities calculated for a fleet *f* of three vehicles in 1998. In the estimating sample, the same fleet *f* had five vehicles, so in equation (19), $I_f = 5$ and $I_f^{t+1} = 3$. The estimated values of the random effects parameters are equal to $\hat{\kappa}^{-1} = 1/11.749 = 0.0851$, $\hat{\nu} = 2.0657$, and $\hat{\delta} = 4.7158$ (see Table C1.3). Suppose that fleet *f* will have no accident at time t+1, then $S_0^{t+1} = 0$ which means that the three vehicles of the fleet *f* in the forecasting sample will have no accident. In applying the formula in (19) with the estimated parameters, the predictive probability of fleet *f* to have no accident at t+1 is equal to 80.2% (1×1.2656×0.8979×0.7839×1×1×1×0.875×1.0290)×100. The calculations are given below.

$$\prod_{i=1}^{3} \frac{\left(\gamma_{fiT_i+1}\right)^{y_{fiT_i+1}}}{\Gamma\left(y_{fiT_i+1}+1\right)} = 1$$

$$\frac{\Gamma\left(S_0 + \sum_{i=1}^{5} T_i \kappa^{-1} + S_0^{t+1} + \sum_{i=1}^{3} \kappa^{-1}\right)}{\Gamma\left(S_0 + \sum_{i=1}^{5} T_i \kappa^{-1}\right)} = \frac{\Gamma\left(1+22\times 0.0851+0+3\times 0.0851\right)}{\Gamma\left(1+22\times 0.0851\right)} = \frac{\Gamma\left(3.1275\right)}{\Gamma(2.8722)} = 1.2656$$

$$\frac{\Gamma\left(\sum_{i=1}^{5} T_i \kappa^{-1}\right)}{\Gamma\left(\sum_{i=1}^{5} T_i \kappa^{-1} + \sum_{i=1}^{3} \kappa^{-1}\right)} = \frac{\Gamma\left(22\times 0.0851\right)}{\Gamma\left(22\times 0.0851+3\times 0.0851\right)} = \frac{\Gamma\left(1.8722\right)}{\Gamma\left(2.1275\right)} = 0.8979$$

$$\frac{\left(\kappa^{-1}\right)^{\sum_{i=1}^{I3} \kappa^{-1}}}{\left(\kappa^{-1} + \overline{\gamma}_{g_2}\right)^{S_0^{t+1} + \sum_{i=1}^{3} \kappa^{-1}}} = \frac{\left(0.0851\right)^{3\times 0.0851}}{\left(0.0851+0.1357\right)^{0+3\times 0.0851}} = 0.7839$$

$$\frac{\Gamma\left(\sum_{i=1}^{5}\left(S_i + \nu_{(f)i}\right)\right)}{\Gamma\left(\sum_{i=1}^{5}\left(S_i + \nu_{(f)i}\right) + S_0^{t+1}\right)} = \frac{\Gamma\left(0+5\times 2.0657\right)}{\Gamma\left(0+5\times 2.0657+0\right)} = 1$$

$$\frac{\prod_{i=1}^{5} \Gamma\left(S_i + \nu_{(f)i} + y^{*}_{fiT_i+1}\right)}{\prod_{i=1}^{5} \Gamma\left(S_i + \nu_{(f)i}\right)} = \frac{\Gamma\left(0+2.0657+0\right)\times\Gamma\left(1+2.0657+0\right)\times\Gamma\left(0+2.0657+0\right)\times\Gamma\left(0+2.0657\right)\times\Gamma\left(0+2.0657\right)}{\Gamma\left(0+2.0657\right)\times\Gamma\left(1+2.0657\right)\times\Gamma\left(0+2.0657\right)\times\Gamma\left(0+2.0657\right)\times\Gamma\left(0+2.0657\right)} = 1$$

$$\frac{\prod_{i=1}^{3} \Gamma\left(y_{fiT_i+1} + \delta_{(fi)T_i+1}\right)}{\prod_{i=1}^{3} \Gamma\left(\delta_{(fi)T_i+1}\right)} = \frac{\Gamma\left(0+4.7158\right)\times\Gamma\left(0+4.7158\right)\times\Gamma\left(0+4.7158\right)}{\Gamma\left(4.7158\right)\times\Gamma\left(4.7158\right)\times\Gamma\left(4.7158\right)} = 1$$

$$\frac{\prod_{i=1}^{5}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}{\prod_{i=1}^{5}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}+\left(y^{*}_{fiT_i+1}+\delta^{*}_{(fi)T_i+1}\right)\right)}=\frac{\Gamma\left(0+7\times 4.7158\right)\times\Gamma\left(1+7\times 4.7158\right)\times\Gamma\left(0+2\times 4.7158\right)\times\Gamma\left(0+4\times 4.7158\right)\times\Gamma\left(0+2\times 4.7158\right)}{\Gamma\left(0+7\times 4.7158+0+4.7158\right)\times\Gamma\left(1+7\times 4.7158+0+4.7158\right)\times\Gamma\left(0+2\times 4.7158+0+4.7158\right)\times\Gamma\left(0+4\times 4.7158\right)\times\Gamma\left(0+2\times 4.7158\right)}$$

$$=\frac{\Gamma(33.0106)\times\Gamma\left(34.0106\right)\times\Gamma\left(9.4316\right)}{\Gamma\left(37.7264\right)\times\Gamma\left(38.7264\right)\times\Gamma\left(14.1474\right)}$$

$$\frac{\prod_{i=1}^{5}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}+\delta^{*}_{(fi)T_i+1}\right)}{\prod_{i=1}^{5}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}=\frac{\Gamma\left(7\times 4.7158+4.7158\right)\times\Gamma\left(7\times 4.7158+4.7158\right)\times\Gamma\left(2\times 4.7158+4.7158\right)\times\Gamma\left(4\times 4.7158\right)\times\Gamma\left(2\times 4.7158\right)}{\Gamma\left(7\times 4.7158\right)\times\Gamma\left(7\times 4.7158\right)\times\Gamma\left(2\times 4.7158\right)\times\Gamma\left(4\times 4.7158\right)\times\Gamma\left(2\times 4.7158\right)}$$

$$=\quad\frac{\Gamma\left(37.7264\right)\times\Gamma\left(37.7264\right)\times\Gamma\left(14.1474\right)}{\Gamma\left(33.0106\right)\times\Gamma\left(33.0106\right)\times\Gamma\left(9.4316\right)}$$

$$\frac{\prod_{i=1}^{5}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}{\prod_{i=1}^{5}\Gamma\left(S_i+\sum_{t=1}^{T_i}\delta_{(fi)t}+\left(y^{*}_{fiT_i+1}+\delta^{*}_{(fi)T_i+1}\right)\right)}\times\frac{\prod_{i=1}^{5}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}+\delta^{*}_{(fi)T_i+1}\right)}{\prod_{i=1}^{5}\Gamma\left(\sum_{t=1}^{T_i}\delta_{(fi)t}\right)}=\frac{\Gamma\left(34.0106\right)}{\Gamma\left(38.7264\right)}\times\frac{\Gamma\left(37.7264\right)}{\Gamma\left(33.0106\right)}=0.875$$

$$\frac{{}_2F_1\left(\sum_{i=1}^{g_1}\left(S_i+\nu_{(f)i}\right)+S^{t+1}_{g_1},S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1}+S^{t+1}_0+\sum_{i=1}^{I_f^{t+1}}\kappa^{-1},\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right)+S^{t+1}_0,\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right)}{{}_2F_1\left(\sum_{i=1}^{g_1}\left(S_i+\nu_{(f)i}\right),S_0+\sum_{i=1}^{I_f}T_i\kappa^{-1},\sum_{i=1}^{I_f}\left(S_i+\nu_{(f)i}\right),\left(\frac{\overline{\gamma}_{g_2}-\overline{\gamma}_{g_1}}{\kappa^{-1}+\overline{\gamma}_{g_2}}\right)\right)}$$

$$=\frac{{}_2F_1\left(0+2\times 2.0657+0;1+22\times 0.0851+0+3\times 0.0851;1+5\times 2.0657+0;\left(\frac{0.1357-0.0749}{0.0851+0.1357}\right)\right)}{F_1\left(0+2\times 2.0657;1+22\times 0.0851;1+5\times 2.0657;\left(\frac{0.1357-0.0749}{0.0851+0.1357}\right)\right)}$$

$$=\frac{{}_2F_1\left(4.1314;3.1275;11.3285;0.2752\right)}{{}_2F_1\left(4.1314;2.8722;11.3285;0.2752\right)}=\frac{1.4084}{1.3687}=1.0290$$

If now we suppose that fleet *f* will have 1 accident at t+1, then $S_0^{t+1} = 1$ in table E1. Since fleet *f* has three vehicles, there are 3 possibilities for the fleet to accumulate 1 accident. In applying the formula in (19) we obtain that the predictive probability of fleet *f* to have 1 accident is equal to 12.7% (Table E1). There are 6 possibilities for the fleet *f* of three vehicles to accumulate 2 accidents at t+1 (Table E1). The predictive probability that the fleet *f* will have 2 accidents during the next year is then 1.5%. And so on…

We observe that the predictive probabilities in Table E1 differ from the average predictive probabilities in Table 9 for fleets of 5 trucks. The fleet in this example represents a lower risk than the average fleet of this size: during the last 7 years, the fleet had only one accident by assumption meaning that the implied accident rate for a truck in this fleet is 3% while the mean is 13% for this size of fleet.

Table E1: Example of predictive probabilities calculated for a fleet *f* of three trucks.

Fleet *f*	Estimating sample			Forecasting sample										
					$S_0^{t+1}=0$	$S_0^{t+1}=1$			$S_0^{t+1}=2$					
					1 possibility	3 possibilities			6 possibilities					
Truck *i*	S_i	Group	T_i	$\hat{\gamma}_{fiT_i+1}$	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}	y_{fiT_i+1}
1	0	2	7	0.1304	0	1	0	0	2	0	0	1	1	0
2	1	2	7	0.0960	0	0	1	0	0	2	0	1	0	1
3	0	1	2	0.0633	0	0	0	1	0	0	2	0	1	1
4	0	1	4	.	.	.	.	.	.	.	.	.	.	.
5	0	2	2	.	.	.	.	.	.	.	.	.	.	.
	$S_0 = 1$	$\sum_{i=1}^{g_1} S_i = 0$	$\sum_{i=1}^{5} T_i = 22$		$S_{g1}^{t+1}=0$	$S_{g1}^{t+1}=0$	$S_{g1}^{t+1}=0$	$S_{g1}^{t+1}=1$	$S_{g1}^{t+1}=0$	$S_{g1}^{t+1}=0$	$S_{g1}^{t+1}=2$	$S_{g1}^{t+1}=0$	$S_{g1}^{t+1}=1$	$S_{g1}^{t+1}=1$
Predictive probabilities						3.6% +3.9%+5.2%			0.17%+0.17%+0.35%+0.21%+0.29%+0.31%					
					80.2%	12.7%			1.5%					

$\hat{\bar{\gamma}}_{g_1} = 0.0749$ and $\hat{\bar{\gamma}}_{g_2} = 0.1357$ are the means of group 1 and group 2 respectively.

$\hat{\gamma}_{fiT_i+1}$ is calculated from the forecasting sample with the estimated coefficients presented in Table C1.3.

APPENDIX F: BOOSTRAP REPLICATIONS, R-CODE

/************************************

Cameron and Trivedi (2013b) propose the following representation of the Hausman test:

$$T_H = \left(\hat{\beta}_{RE} - \tilde{\beta}_{FE}\right)' \left[\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}]\right]^{-1} \left(\hat{\beta}_{RE} - \tilde{\beta}_{FE}\right)$$

where T_H is the Hausman test statistic, $\tilde{\beta}_{FE}$ are the estimated parameters obtained from the fixed effects model and $\hat{\beta}_{RE}$ are the estimated parameters obtained from the random effects model (Gamma-Dirichlet model). To estimate the variance term $\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}]$ we use a panel bootstrap method that resamples over the 5,423 firms of the sample:

$$\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}] = \frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{\beta}_{FE}^{(b)} - \hat{\beta}_{RE}^{(b)}\right)\left(\tilde{\beta}_{FE}^{(b)} - \hat{\beta}_{RE}^{(b)}\right)$$

where $\tilde{\beta}_{FE}^{(b)}$ and $\hat{\beta}_{RE}^{(b)}$ are the estimates obtained from the *bth* bootstrap replication

***********************************/

Estimation of the $\hat{\beta}$

```
library(foreign, pos=15)

#
#Read the data set
Dataset <- read.table("donneePlusDe4.csv", header=TRUE, sep=",",na.strings="NA", dec=".", strip.white=TRUE)

#Read  B random samples with replacement,  the firm can be selected more than once.
SampleURS data set include Three variables
                    1. Replication number
                    2. Firm identification
                    3. Number of hit refers to the number of times a firm  is selected
SampleURS <- read.table("SampleUSR.csv", header=TRUE, sep=",",na.strings="NA", dec=".", strip.white=TRUE)

library(abind, pos=16)
library(e1071, pos=17)
library("BMS")
library("spuRs")
library("MASS")
library(Rcpp)
library(inline)
library(splitstackshape)

#Write the 1,000 coefficient estimations in OURRandom  file
outputfile = "OUTRandom"
cat("replicate", "nu", "kap", "del", "Intercep", "NB_INF1", "NB_INF2", "NB_INF3", "NB_INF6", "NB_INF7",
          "NB_INF89", "VIT", "SANCT", "ROUGE", "ARRET", "CEINTURE", "N_VH69", "N_VH20",
          "N_VH50", "N_VH51", "an_91", "an_92", "an_93", "an_94", "an_95",  "an_96", "an_97",
sep="\t",file=outputfile, append=T)
cat("\n",file=outputfile , append=T)
```

```
B = 1000
for (repl  in 1:B){

sample <- SampleURS[SampleURS$Replicate==iter,]
sample_1 <- merge (Dataset, sample, by=c("TRNIP"))
sample_2  <- expandRows(sample_1, "NumberHits",  drop=FALSE)
sample_2$frac <- as.numeric(row.names(sample_2))
sample_2$integ <- trunc(sample_2$frac)
sample_2$Hits <- round((sample_2$frac - sample_2$integ)*10,digits=1)
sample_2$TRNIP_1 <- as.character(paste(sample_2$TRNIP, sample_2$Hits, sep="" ))
attach(sample_2)
sort.sample_2 <- sample_2[order(TRNIP_1,VEH_1,AN),]
detach(sample_2

/*************************
Estimate $\hat{\beta}_{RE}^{(iter)}$ with sort.sample_2 dataset
*************************/

parlistL<-essaiL$par
cat(i, parlistL,sep="\t",file=outputfile, append=T)
cat("\n",file=outputfile , append=T)

}

#Write the 1,000 coefficient estimations in OUTFixed file
outputfile = "OUTFixed"
cat("replicate ", "NB_INF1", "NB_INF2", "NB_INF3", "NB_INF6", "NB_INF7",
            "NB_INF89", "VIT", "SANCT", "ROUGE", "ARRET", "CEINTURE", "N_VH69", "N_VH20",
            "N_VH50", "N_VH51", "an_91", "an_92", "an_93", "an_94", "an_95",  "an_96", "an_97",
sep="\t",file=outputfile, append=T)
cat("\n",file=outputfile , append=T)

B = 1000
for (repl in 1:B){

sample <- SampleURS[SampleURS$Replicate==iter,]
sample_1 <- merge (Dataset, sample, by=c("TRNIP"))
attach(sample_1)
sort.sample_1 <- sample_1[order(TRNIP_1,VEH_1,AN),]
detach(sample_1

/*************************
Estimate $\tilde{\beta}_{FE}^{(iter)}$ with sort.sample_1 dataset
If the same firm f appears twice in a bootstrap resample iter then $\tilde{\beta}_{FE}^{(iter)}$ needs to treat the fixed effect $\alpha_f$ as being
the same for both observations f.
*************************/

#Poisson model  to estimate the initial values of the parameters
GLM <- glm(NB_ATOT ~ N_VH69 + N_VH20 + N_VH50 + N_VH51 + VIT + SANCT + ROUGE + ARRET +
CEINTURE + an_91 +
          an_92 + an_93 + an_94 + an_95 + an_96 + an_97 + NB_INF1 + NB_INF2 + NB_INF3 + NB_INF6 +
          NB_INF7 + NB_INF89, family=poisson(log), data=sort.sample_1)
param=GLM$coefficients
```

```
  x <- as.matrix(cbind(sort.sample_1$N_VH69, sort.sample_1$N_VH20, sort.sample_1$N_VH50,
                    sort.sample_1$N_VH51,
                    sort.sample_1$VIT, sort.sample_1$SANCT, sort.sample_1$ROUGE, sort.sample_1$ARRET,
                    sort.sample_1$CEINTURE,
                    sort.sample_1$an_91, sort.sample_1$an_92, sort.sample_1$an_93, sort.sample_1$an_94,
                    sort.sample_1$an_95,  sort.sample_1$an_96, sort.sample_1$an_97,
                    sort.sample_1$NB_INF1,sort.sample_1$NB_INF2, sort.sample_1$NB_INF3,
                   sort.sample_1$NB_INF6, sort.sample_1$NB_INF7, sort.sample_1$NB_INF89))

y= as.matrix(cbind(sort.sample_1$NB_ATOT))
max_y=max(y)

parm=c('n_vh69', 'n_vh20', 'n_vh50', 'n_vh51', 'vit', 'sanct', 'rouge', 'arret', 'ceinture', 'an_91', 'an_92', 'an_93',
           'an_94', 'an_95', 'an_96', 'an_97', 'nb_inf1', 'nb_inf2', 'nb_inf3', 'nb_inf6', 'nb_inf7', 'nb_inf89' )

p=ncol(x)
n=nrow(x)
p1=p+1

#t is number of periods per  truck,
nper=as.matrix(cbind(sort.sample_1$n_period, sort.sample_1$camion))
nper1=nper[sort.sample_1$camion==1,]
t=cbind(nper1[,1])

#fl is the number of trucks per fleet,
taillec=as.matrix(cbind(sort.sample_1$taille_c, sort.sample_1$FLOTTE))
taillec1=taillec[sort.sample_1$FLOTTE==1,]
fl=cbind(taillec1[,1])

#fl_an is the number of year-trucks per fleet.
taillet=as.matrix(cbind(sort.sample_1$taille_t, sort.sample_1$FLOTTE))
taillet1=taillet[sort.sample_1$FLOTTE==1,]
fl_an=cbind(taillet1[,1])

# Number of time the firm had been selected
Hits<- as.matrix(cbind(sort.sample_1$NumberHits,sort.sample_1$FLOTTE))
 Hits1<-Hits[sort.sample_1$FLOTTE == 1,]
NHits<-cbind(Hits1[,1])

#ki is  total number of trucks
ki=nrow(t)

# kf  is total  number of fleets.
kf=nrow(fl)

#Total number of parameters including firm fixed effects.
n_parm=p+kf

#Initial values of the parameters.
ini_fix=matrix(0,1,kf)
ini_p=param[2:p1]
r_beta= c(ini_p, ini_fix)

eps=1e-8
diff=1
x_sol=matrix(0,nrow=1, ncol=n_parm)
```

```
for(iter  in 1:100){

    par=r_beta[1:p]
    r_ll_ff=0
    r_ll_gf=matrix(0, nrow=1, ncol=p)
    r_ll_hf=matrix(0, nrow=p, ncol=p)
    ster_the=matrix(0, nrow=kf, ncol=1)
    xpar=matrix(0, nrow=kf, ncol=p)
    nx=0
    for(f in 1:kf){
      ind_f=f+p
      thetaf=r_beta[ind_f]

      #Number of trucks in the fleet f
      nf=fl[f]

      #Number of year-trucks in the fleet f
      snf=fl_an[f]

      #rf is the vector of subscripts for each year-trucks in the fleet f
      i_deb=nx+1
      i_fin=nx+snf
      rf=i_deb:i_fin

      #nx brings us to the next fleet
      nx=i_fin

       yf=y[rf]
      NHitsf <- NHits[f]
       xf=x[rf,]
      zf=xf%*%par
      muf=exp(zf+thetaf)
      s_muf=sum(muf)

      # log-likelihood (for the fleet f).
      ter_t=-muf+yf*(thetaf+zf)- lgamma(yf+1)

      #r_ll_ff is the sum of all fleets
      ster_t= NHitsf *sum(ter_t)
      r_ll_ff=r_ll_ff+ster_t

       ter_1f=matrix(0,snf,p)
      for(i in 1:p){
        ter_1f[,i]=xf[,i]*muf[,1]
      }
      hgf= colSums(ter_1f)

      xparf=hgf/s_muf
      xpar[f,]=xparf

      ter_2f=matrix(0, nrow=snf, ncol=p)
      for(ii in 1:snf){
          ter_2f[ii,]=xf[ii,]-xparf
      }

 ter_3=t(ter_2f)%*%(yf-muf)
      llpf=t(ter_3)
      r_ll_gf=r_ll_gf+NHitsf*llpf

      ter2muf=matrix(0,snf,p)
      for(iii in 1:p){
        ter2muf[,iii]=ter_2f[,iii]*muf[,1]
      }

      ter_p=-t(ter_2f)%*%(ter2muf)
      r_ll_hf=r_ll_hf+NHitsf*ter_p


      ter_the_f=(yf-muf)
      ster_the_f=sum(ter_the_f)
      ster_the[f,]=ster_the_f/s_muf

    }
#end for (f in 1: kf)

    inv_r_ll_hf=solve(r_ll_hf)

    delta=inv_r_ll_hf%*%t(r_ll_gf)
    delta_theta=-ster_the+xpar%*%delta

    #We iterate on the solution to make it converge to the final estimators.
    x_sol[1:p]=t(delta)
    x_sol[p1:n_parm]=t(delta_theta)
    r_beta=r_beta-x_sol
    diff1=which.max(abs(x_sol))
    diff=max(abs(x_sol))

    if(diff<=eps) {
      break }

}
#end  for(iter in 1:100)

# Optimation results
beta_sol=r_beta[1:p]

parlistL<- beta_sol
cat(i, parlistL,sep="\t",file=outputfile, append=T)
cat("\n",file=outputfile , append=T)

}
# end for (repl in 1:B)
```

Hausman test R-code

```
random=read.csv(OUTRandom.csv", header=TRUE, sep=";")
fixed=read.csv("OUTFixed", header=TRUE, sep="\t")


N_VH69_r=random$N_VH69
N_VH69_f=fixed$N_VH69
N_VH20_r=random$N_VH20
N_VH20_f=fixed$N_VH20
N_VH50_r=random$N_VH50
N_VH50_f=fixed$N_VH50
N_VH51_r=random$N_VH51
N_VH51_f=fixed$N_VH51
VIT_r=random$VIT
VIT_f=fixed$VIT
SANCT_r=random$SANCT
SANCT_f=fixed$SANCT
ROUGE_r=random$ROUGE
ROUGE_f=fixed$ROUGE
ARRET_r=random$ARRET
ARRET_f=fixed$ARRET
CEINTURE_r=random$CEINTURE
CEINTURE_f=fixed$CEINTURE
an_91_r=random$an_91
an_91_f=fixed$an_91
an_92_r=random$an_92
an_92_f=fixed$an_92
an_93_r=random$an_93
an_93_f=fixed$an_93
an_94_r=random$an_94
an_94_f=fixed$an_94
an_95_r=random$an_95
an_95_f=fixed$an_95
an_96_r=random$an_96
an_96_f=fixed$an_96
an_97_r=random$an_97
an_97_f=fixed$an_97
NB_INF1_r=random$NB_INF1
NB_INF1_f=fixed$NB_INF1
NB_INF2_r=random$NB_INF2
NB_INF2_f=fixed$NB_INF2
NB_INF3_r=random$NB_INF3
NB_INF3_f=fixed$NB_INF3
NB_INF6_r=random$NB_INF6
NB_INF6_f=fixed$NB_INF6
NB_INF7_r=random$NB_INF7
NB_INF7_f=fixed$NB_INF7
NB_INF89_r=random$NB_INF89
NB_INF89_f=fixed$NB_INF89

n_vh69=N_VH69_f-N_VH69_r
n_vh20=N_VH20_f-N_VH20_r
n_vh50=N_VH50_f-N_VH50_r
n_vh51=N_VH51_f-N_VH51_r
vit=VIT_f-VIT_r
```

```
sanct=SANCT_f-SANCT_r
rouge=ROUGE_f-ROUGE_r
arret=ARRET_f-ARRET_r
ceinture=CEINTURE_f-CEINTURE_r
an91=an_91_f-an_91_r
an92=an_92_f-an_92_r
an93=an_93_f-an_93_r
an94=an_94_f-an_94_r
an95=an_95_f-an_95_r
an96=an_96_f-an_96_r
an97=an_97_f-an_97_r
nb_inf1=NB_INF1_f-NB_INF1_r
nb_inf2=NB_INF2_f-NB_INF2_r
nb_inf3=NB_INF3_f-NB_INF3_r
nb_inf6=NB_INF6_f-NB_INF6_r
nb_inf7=NB_INF7_f-NB_INF7_r
nb_inf89=NB_INF89_f-NB_INF89_r

diff=cbind(n_vh69, n_vh20, n_vh50, n_vh51, vit, sanct, rouge, arret, ceinture, an91, an92, an93, an94, an95, an96,
an97, nb_inf1, nb_inf2, nb_inf3, nb_inf6, nb_inf7, nb_inf89)

# $\tilde{\beta}_{FE}$ , Table 10 column 2 page 26
FE=c(0.0283,0.0532,0.0347,0.0841,0.2248,0.3857,0.3068,0.3443,0.1219,0.0586,0.0292,-0.0584,-0.0209,-0.0157,
      -0.0608,-0.1914,0.1584,0.2828,0.2321,0.2245,0.1583,0.2331)

# $\hat{\beta}_{RE}$ , Table 10 column 4 page 26
RE=c(0.0168,0.0864,0.0829,0.0849,0.2584,0.4245,0.3804,0.4105,0.1651,0.0995,0.0823,0.0877,0.1694,0.1631,
      0.0672,-0.1759,0.2006,0.2675,0.2770,0.2777,0.2012,0.2369)

# $\hat{\beta}_{RE} - \tilde{\beta}_{FE}$
diff_true=RE-FE

# $\hat{V}[\tilde{\beta}_{FE} - \hat{\beta}_{RE}] = \frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{\beta}_{FE}^{(b)} - \hat{\beta}_{RE}^{(b)}\right)\left(\tilde{\beta}_{FE}^{(b)} - \hat{\beta}_{RE}^{(b)}\right)$
V=cov(diff[1:B,])
V_1=solve(V)

# Hausman test statistic
TH[=t(diff_true)%*%V_1%*%diff_true
```