CAR SURVIVAL IN A NATIONAL CAR FLEET: NON-PARAMETRIC AND PARAMETRIC APPROACHES APPLIED TO FRENCH DATA

Zéhir KOLLI, INRETS (French National Institute for Transport and Safety Research) / DEST (Department of Transport Economics and Sociology) , 2 rue de La Butte Verte 93166 Noisy le Grand, France, mail: kolli@inrets.fr Phone: +33 145925596
Ariane DUPONT-KIEFFER, INRETS/DEST, mail: dupont@inrets.fr Phone : +33 145925601
Laurent HIVERT, INRETS/DEST, mail: hivert@inrets.fr Phone : +33 145925595

ABSTRACT

We aim to determine the common demographic conditions and variables that affect light cars survival in the national fleet. The essential goal is to identify which variables determine cars' longevity and to understand how it can be measured. Therefore in this study two fundamental questions are asked:
-*"What are the main determinants of cars' longevity?"*
-"*Can we consider car mileage as a major dimension of cars' longevity?*"
To give an answer, we adopted a longitudinal analysis based on panel data. Yearly French `Parc-Auto' waves were linked together from 2000 up to 2006 to make a 7-year panel, with 6795 cars described. Variables like car age, total mileage, car status (i.e., whether the car is unique, main or secondary in the household) and motorization (i.e., gasoline or diesel) are used.
In order to have an overview on car survival, we consider a new approach by building and comparing Kaplan-Meier survival curves by car age and by average monthly mileage[1]. The idea is to compare car use measured by the average monthly mileage with car's life expectancy measured by age, and to conclude for each determinant (i.e., status and motorization)[2] about how intensity of use is linked to car longevity. As we do not have data concerning car scrapping or disappearing from the fleet; all observations are right censored.
We show that mortality risk increases for all car categories with age and mileage. Gasoline cars have a better survival rate than diesel cars. Car status is also a major determinant of car survival: secondary cars have better survival rates than main cars or than unique cars. This is explained by the fact that sole cars are used more extensively than main or secondary cars. Bivariate pdf (i.e., probability density functions) described by age and total mileage helps to show that this phenomenon has to be related to an extensive use of diesel cars and of cars declared as to be sole in the household (i.e., versus 'main' or 'secondary' cars). To complete this approach we choose to proceed to a parametric study in order to find the best functional shape for the car age and car mileage statistical distributions. Models were estimated using three-parameters- Beta, Gamma, Lognormal and Weibull distributions. Adequacy parameters and qq-plots show that the Beta and Weibull distributions are the best compared to all other models. Therefore, Beta and Weibull are good candidates to represent cars' survival laws. More importantly, this study shows that car mileage defined as a marker process is a major dimension to describe cars' longevity.

Keywords: Car survival, Kaplan-Meier estimator, car mileage, Beta and Weibull distributions, marker process, parent process.

[1] Calculated for each car as the total mileage declared divided by the age in months.

[2] During a preliminary work we also tested the influence of variables such as income, horse power and cylinder capacity without great success.

1 INTRODUCTION

A longstanding question concerns household car survival in the national fleet. Planners concerned with energy, resource consumption and environmental quality will, more and more, take into account time for a national fleet to achieve green standards, time of diffusion of a new technology in a fleet, or expected time for disappearances of the 5, 10 or 15 percent of oldest and most polluting cars from a given fleet. With the aim of having elements of answer, related forecasts on car fleet are required. In the majority of cases, car fleet statistics are not accurate enough. In order to have more accurate estimations and forecasts we can proceed to a dynamic modeling including car registration statistics that are usually well known and estimations of cars that are scrapped through a survival data analysis. A serious examination of the longevity of cars will also permit having better understanding of car ownership behavior, better estimations and assessments of GHG emissions, as well as estimations and forecasts of car fleet composition in the absence of direct taxes[3]. It can also help to simulate the reaction to a small, medium or massive introduction of a new technology like electric or hybrid cars.

Both at the EU and US levels critical forecasts are drawn for the next years concerning greenhouse gas emissions. The need for estimating vehicle population for assessing air quality has become increasingly important (Lynch and Lee, 1979; Southworth, 1986; Zachariadis et al., 1995; Bhat et al., 2008). In that way, we need a better understanding of car renewal and identify which are the economic and regulatory instruments packages that can best promote technological car renewal and technology deployment. For all these reasons, we need to have a better understanding of the forces that are driving car longevity through age, mileage accumulated, car status or motorization, etc…

Much existing research related to car longevity or survival in the U.S. market (or linked to cash for clunkers programs) has used a logistic functional form to estimate car survival rates per year or per vintage. Greene and Chen (1981) used a logistic equation separately for domestic U.S. cars and imported cars and for three vehicle types (i.e., trucks, light duty vehicles and automobiles). Their results show better survival expectancy for light trucks compared to light cars (i.e., median lifetimes respectively equal to 14.5 and 9.9 years)

Berkovec, (1985) estimated a scrappage model in which he regressed the log of the death rate on auto prices, prices squared, and dummy variables for vintage and class. Greenspan and Cohen, (1996) estimated car mortality as an input to the study of the demand for new cars. They captured engineering scrappage by estimating a pooled regression. In order to be consistent, they assumed that scrappage is equal to zero the first three years.

Gallez, (2000) fitted longitudinal annual rates of survival with the assumption that they follow a Lognormal survival function with a maximum age (28 years) after which the number of survivors is equal to zero. Using French data from 1968 to 1992, Gallez found that median lifetimes for diesel cars and gasoline cars are respectively equal to 10.82 and 11.23 and that mean lifetimes are equal to 11.74 and 12.22 years in France.

Hamilton and Macaulay, (1998) illustrated the nature and size of possible sources of improvements in cars' longevity, including improvements in intrinsic durability and other economic forces which induce consumers to maintain cars into older age. They followed the automobile-mortality literature in assuming that the pattern of car mortality could be characterized by a logistic function.

More recent research showed that drivers' travel behavioral attitudes also affect car use. Choo and Mokhtarian, (2004) found that travel attitudes, personality, lifestyle, and mobility factors are useful in predicting the most-used vehicle within a household.

Chen and Niemeier, (2004) applied a duration model of the Weibull form with two mass points to approximate the unobserved heterogeneity to estimate car survival rates on a sample of 678 cars. The main result of their vehicle scrappage model suggests that vehicle age and car mileage are important determinants of scrappage decision.

[3] In France a direct tax called `vignette' was aborted in 2000

In most researches, the traditional variable of lifetime measurement is the age of the car. In order to have an overview on car survival we choose to take into account car use intensity through the annual driven (i.e., declared) mileage variable or average driven (i.e., declared) mileage per month. This variable is treated in addition to the age of the car, the traditional variable of lifetime measure. Many researches show that the natural time scale of the survival process is not calendar or clock time. Mathematical research on different time scales has been carried out: Oakes, (1995) and Kordovsky and Gertsbakh, (1997) looked at multiple running time scales in survival data analysis. Duchesne and Lawless, (2000) and Duchesne and Rosenthal, (2003) described various advances in running time models for survival data. Cox and Oakes, (1984, Section 1.2, p. 3-4) pointed out that *"often the 'scale' for measuring time is clock time, although other possibilities certainly arise, such as the use of operating time of a system, mileage of a car, or some measures or cumulative load encountered"*.

For the non parametric approach, the main idea is to compare car use (i.e., which can depend on its motorization, status, horse power, cylinder capacity, etc...) measured by an average mileage per month, calculated for each car as the total accumulated mileage declared divided by the age in months, with car life expectancy measured by age. This comparison show that car life duration is negatively correlated with use for each car category and that drivers' behavior is the principal determinant of cars' life duration. In order to illustrate the strong correlation between car age and car mileage and their ability for measuring car life duration some bivariate pdf for different categories of cars are shown.
Then, we show with a parametric approach that mileage of cars (i.e., besides car age) is a major variable to explain car life duration. This parametric approach is also useful to show the best pdf that fit our data and hence the most appropriate function to model survival curves by age or by mileage.

2 DATA

We used `Parc-Auto' an annual car fleet database built from annual postal surveys conducted with a panel of 10 000 French households[4].'Parc Auto' database is merged to 2006 with car and household's characteristics concerned back to 2000. Hence a panel data set from 2000 to 2006 was built up including 6795 observed cars and 4794 household (e.g., Table1). From this database we use some cars' specificity and characteristics like total mileage and age that are also well described. We study the impact of variables like motorization (e.g., Table 2) and car status in car life duration. Table 3 shows the proportion of cars by status. For each car, three principal categories of status can be considered. It can be a sole car (1), the main car (2) or a secondary car (3) with the respective proportions given as following: 63.88, 29.45 and 6.67 percents of our database (e.g., Table 3).

Table 1 - Parc Auto annual waves

Year	Number of cars	Number of households
2000	8115	6368
2001	8177	6438
2002	8249	6466
2003	8256	6523
2004	8170	6291
2005	8186	7162
2006	7152	5600
Total	56305	44848
Observations used	6795	4794

[4] About one hundred questions were asked in the survey about car ownership, car characteristics, main and secondary users, previous car characteristics, car use behavior, attitudes toward automobiles, etc...

Table 2 - Motorization

Motorization	Number of cars	Percents
Gasoline cars	3418	50.60
Diesel cars	3312	48.74
LPG (Liquefied Petroleum Gas)	30	0.44
Hybrid or Electric	2	0.03
Not described	33	0.49

Table 3 - Car Status in the household

Status	Motorization by status	Number of cars	Total	Percents
Sole car	Gasoline cars	2025	4341	63.88
	Diesel cars	2285		
	Others	31		
Main car	Gasoline cars	1154	2001	29.45
	Diesel cars	822		
	Others	25		
Secondary	Gasoline cars	239	453	6.67
	Diesel cars	205		
	Other	9		

3 NON PARAMETRIC ANALYSIS: Kaplan-Meier estimator

Usually, a first step in the analysis of survival data is the estimation of the distribution of the survival times. In biometrics, an analysis of Kaplan-Meier, (1958) is practical to calculate the empirical life expectancy of a population of individuals. In that way, if we consider a car like an individual with a 'birthdate' considered as the date of construction and if each car is followed until death, the curve may be estimated simply by computing the fraction surviving at each time. In our case study, we have no information about car destructions and consequently all observations are right censored. This is an important limitation but Kaplan-Meier analysis allows us to estimate survival trends over time, even when cars lifetime drop out or are studied for different lengths of time. For each interval, survival probability is calculated as the number of car still in the fleet divided by number of cars at risk. Cars that have been scrapped, abandoned, or that have not reached the time yet are not counted as 'at risk'. Probability of surviving to any point is estimated from cumulative probability of surviving each of the preceding time intervals (i.e., calculated as the product of preceding probabilities). The survival distribution provides the proportion of cars surviving at the start of any interval and is equal to the probability of surviving each of the preceding intervals multiplied together. If we let *S(t)* be the probability that an item from a given population will have a lifetime exceeding t. For a sample from this population of size N let the observed times until death of N sample members be :

$$t_1 < t_2 < t_3 < \cdots < t_n$$

Corresponding to each t_i is m_i; the number at risk just prior to time t_i (risk-set) and d_i the number of deaths at time t_i .
Note that the intervals between each time typically will not be uniform. The Kaplan-Meier estimator is the nonparametric maximum likelihood estimate of

$$S_t = \prod_{t_j<t} (1 - \theta_j)$$

Is a product given by:

$$\widehat{S_t} = \prod_{t_j<t}(1-\hat{\theta}_j) = \prod_{t_j<t}(\frac{r_j - d_j}{r_j})$$

When there is no censoring, n_i is just the number of survivors just prior to time t_i . With censoring, n_i is the number of survivors less the number of losses (i.e., censored cases). It is only those surviving cases that are still being observed (i.e., have not yet been censored) that are "at risk".

4 Results of the Non Parametric Estimation

4.1 Motorization

From 2000 to 2006 gasoline cars are, on average, 131 months old for an estimated average mileage of 115 112 kilometers. Diesel car drivers declare much more mileage for less old cars. They travel 148 286 kilometers and their cars have 108 months in average (e.g., Table 4). Gasoline cars travel monthly on average 1002 km, for diesel cars the value is much higher with a monthly value of 1681 km per month. These statistical analysis results show a substantial difference between use and age for gasoline and diesel cars. As a start, we try to compare car use for each type of motorization in order to identify their differences in terms of life expectancy (e.g., Figure 1).

Table 4 - Car use by motorization

Motorization	Average mileage (std-error)	Average age (std-error)
Gasoline cars	115112 (62517)	131 (70)
Diesel cars	148286 (78174)	108 (66)
Total sample	131472 (72563)	120 (69)

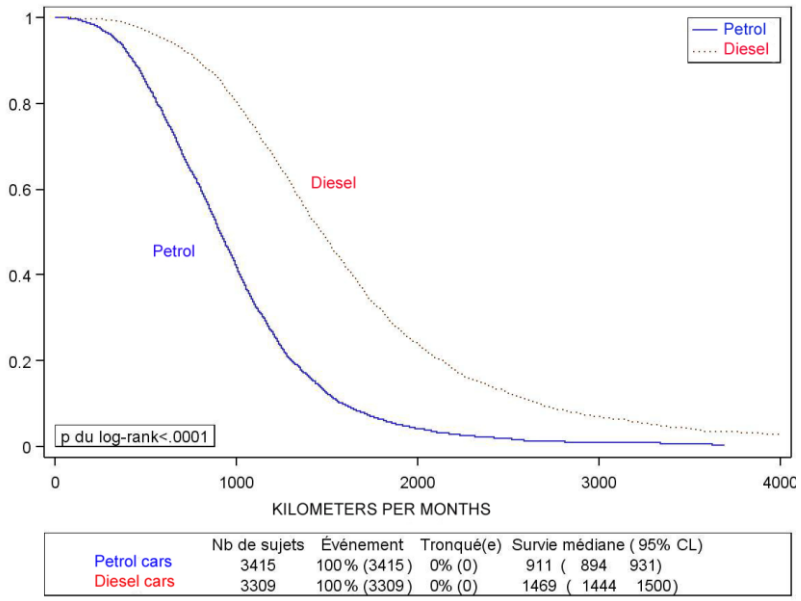


Figure 1 - Kaplan-Meier curves of survival rate by motorization and per monthly average kilometers

Figure 1 depicts results of the non parametric study. The diesel survival rate is always higher than gasoline car in terms of intensity of use. Hence, diesel cars are always submitted to a more intensive use than gasoline car population. Kaplan-Meier survival analysis results reveal that fifty percents of the diesel car “population” drive more than 1469 km per month

(95 percents confidence interval=1444-1500). For gasoline car it is much less with an average estimated at 911 km per months (95 percents confidence interval=894-931).

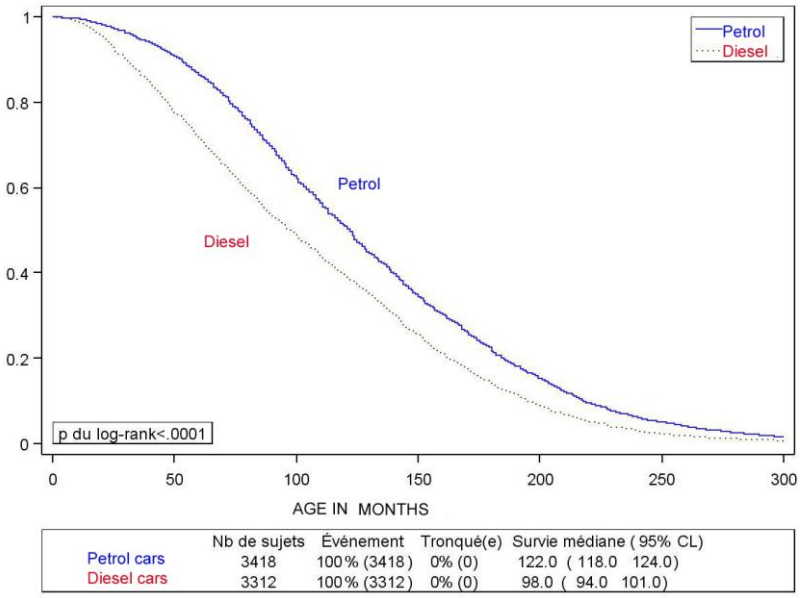


Figure 2 - Kaplan-Meier curves of survival rate by motorization and per months

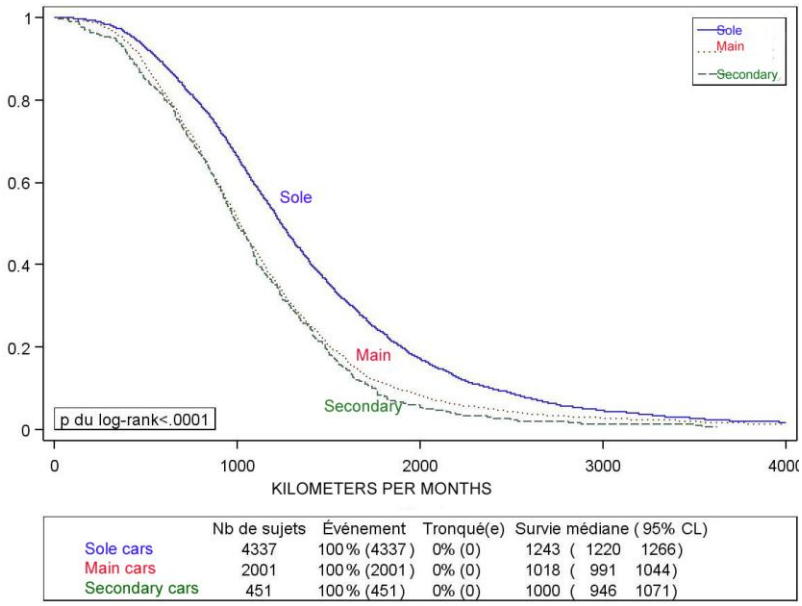
The life expectancy by type of motorization is shown in figure 2. Gasoline cars have the best survival rate during the whole period. Fifty percents of gasoline cars survive after 122 months (95 percents confidence interval=118-124) it is about ten years. For diesel cars it is only 98 months, about 8 years and 2 months (95 percents confidence interval=94-101). Results are not surprising yet we know that diesel car "population" is more intensively used. Diesel engine technical advantages and thermodynamic characteristics of longevity (compared to gasoline cars) are annihilated by a more intense use and their life expectancy is importantly shortened by this intensive use.

4.2 Car status

Considering motorized households only, the considerate cars can be the sole car in the household, the main car (i.e., as freely defined by respondent, usually household head) out of several cars, or a secondary car (i.e., second or more) out of several cars. This variable (i.e., sole, main, secondary) designates the car status.

Table 5 - Car use by status

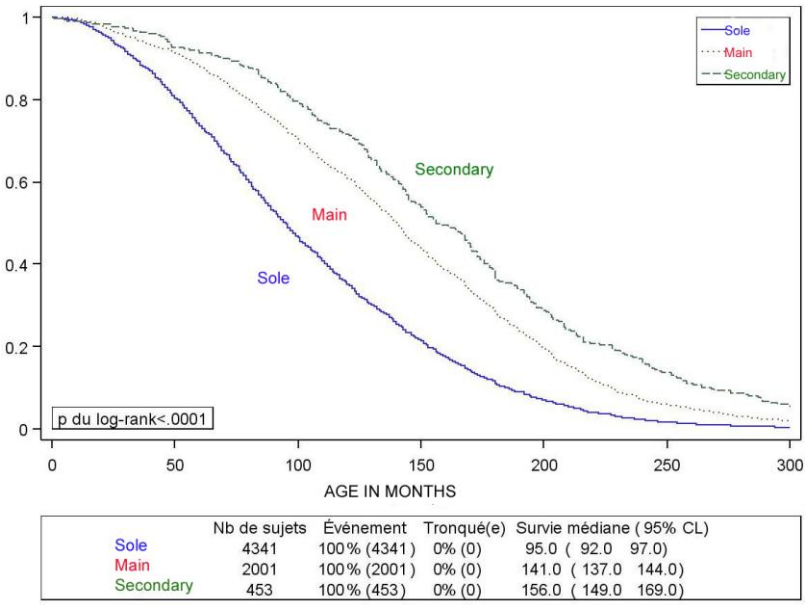
Status	Average mileage (std-error)	Average age (std-error)
Sole cars	124741 (70875)	104 (60)
Main cars	141162 (74056)	143 (71)
Secondary cars	153168 (72715)	168 (90)
Total	131472 (72563)	120 (69)



	Nb de sujets	Événement	Tronqué(e)	Survie médiane (95% CL)
Sole cars	4337	100 % (4337)	0% (0)	1243 (1220 1266)
Main cars	2001	100 % (2001)	0% (0)	1018 (991 1044)
Secondary cars	451	100 % (451)	0% (0)	1000 (946 1071)

Figure 3 - Kaplan-Meier curves of survival rate by car status and per monthly average kilometers

Sole cars can be considered as the most extensively used category in the fleet comparing to other categories. Households with a unique car travel in average 124741 km in 104 months (e.g., Table 5). They also do in average 141162 km in 143 months with their main car and 153168 km in 168 months with their secondary car (e.g., Table 5). A comparison of survival curves by status shows that sole cars' population is far away the most extensively used category. Fifty percents of households which have a unique car travel more than 1243 km per months (95 percents confidence interval=1220-1246). For main car category and secondary car categories, it is almost the same level of use: fifty percents of the main car "population" drive more than 1018 kilometers per months (95 percents confidence interval=991-1044). It is 1000 km per months (95 percents confidence interval=946-1071) for secondary car "population".



	Nb de sujets	Événement	Tronqué(e)	Survie médiane (95% CL)
Sole	4341	100 % (4341)	0% (0)	95.0 (92.0 97.0)
Main	2001	100 % (2001)	0% (0)	141.0 (137.0 144.0)
Secondary	453	100 % (453)	0% (0)	156.0 (149.0 169.0)

Figure 4 - Kaplan-Meier curves of survival rate by car status and per months

Figure 4 depicts increasingly poorer survival rates with lower status category. As expected the subset of sole cars demonstrates low survival rates and the survival curve is the only one that seems to be convex. The half-population of sole cars disappear in only 95 months (i.e., nearly 8 years) (95 percents confidence interval=92-97). Survival curve decreasing is not so

fast for main and secondary cars that spend respectively 141(almost 12 years) (95 percents confidence interval=137-144) and 156 months (i.e., 13 years) (95 percents confidence interval=149-169) to see their half population disappear.

5 PARAMETRIC ANALYSIS

5.1 Preliminary

If we suppose that T is a continuous random variable. Then, we can define F(t) = Pr(T < t) as the probability that the random variable T is less than some value t. The corresponding density function is given by f(t) =$dF(t)/dt$. It is also useful to define the survival function $S(t)$ = 1 – $F(t)$ = $P(t \leq T)$ that correspond to the probability that the random variable T will equal or exceed the value t. A particular useful function for duration analysis is the hazard function defined as $h(t) = f(t)/S(t)$ and gives the probability of exit from the state immediately after time t given that the state is still occupied at t. These relations shows that if either $h(t), S(t), f(t)\ or\ F(t)$ is given, the others can be derived.
As we are computing Beta, Gamma, Lognormal and Weibull three-parameters distributions that are usually written with two-parameters, a theoretical background of these distributions written with three parameters can be useful.

5.2 Beta distribution

The Beta probability distribution function is useful for modeling random probabilities and proportions, particularly in the context of Bayesian analysis. The Beta pdf has a left parameter *α* and a right parameter β. The Beta three-parameters distribution is given by:

$$f(t,\alpha,\beta,\theta) = \begin{cases} \dfrac{(t-\theta)^{\alpha-1}\left(1-(t-\theta)\right)^{\beta-1}}{\int_0^1 (u-\theta)^{\alpha-1}\left(1-(u-\theta)\right)^{\beta-1} du} & if\ \ t \geq \theta \\ 0 & if\ \ t < \theta \end{cases}$$

where α and β, respectively right and left parameters are positive.

5.3 Gamma distribution

The Gamma three-parameters pdf of a Gamma-distributed random variable *t* is given by :

$$f(t,\alpha,\beta,\theta) = \begin{cases} \left(\dfrac{t-\theta}{\beta}\right)^{\alpha-1} \beta^{-\alpha} \dfrac{e^{\left(-\frac{t-\theta}{\beta}\right)}}{\beta\,\Gamma(\alpha)} & if\ \ t \geq \theta \\ 0 & if\ \ t < \theta \end{cases}$$

With $\alpha \geq 0$ and $\beta \geq 0$ and $\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t}\, dt$
α: shape parameter β: scale parameter and θ the location parameter (θ = 0 yields to the two-parameter Gamma distribution).

5.4 Lognormal distribution

The random variable *t* follows the Lognormal distribution if and only if $T = Ln(t)$ follows a standard normal distribution N(α; β). Thus, the probability density function for the three-parameters Lognormal distribution is:

$$f(t,\alpha,\beta,\theta) = \begin{cases} \dfrac{1}{t.\alpha\sqrt{2\pi}} e^{-\frac{[Log(t-\theta)-\beta]^2}{2\alpha^2}} & if \quad t \geq \theta \\ 0 & if \quad t < \theta \end{cases}$$

5.5 Weibull and exponential distributions

The Weibull distribution has been found to provide a reasonable model for lifetimes of several types of units. The three-parameters Weibull pdf is defined as :

$$f(t,\alpha,\beta,\theta) = \begin{cases} \dfrac{\alpha}{\beta}\left(\dfrac{t-\theta}{\beta}\right)^{\alpha-1} e^{-\left(\frac{t-\theta}{\beta}\right)^{\alpha}} & if \quad t \geq \theta \\ 0 & if \quad t < \theta \end{cases}$$

where α and β respectively shape and scale parameters are positive and θ the location parameter (θ = 0 yields to the two-parameter Weibull distribution). Note that If α = 1 we fall into an exponential model.

6 MODELISATION

6.1 Modelisation by age

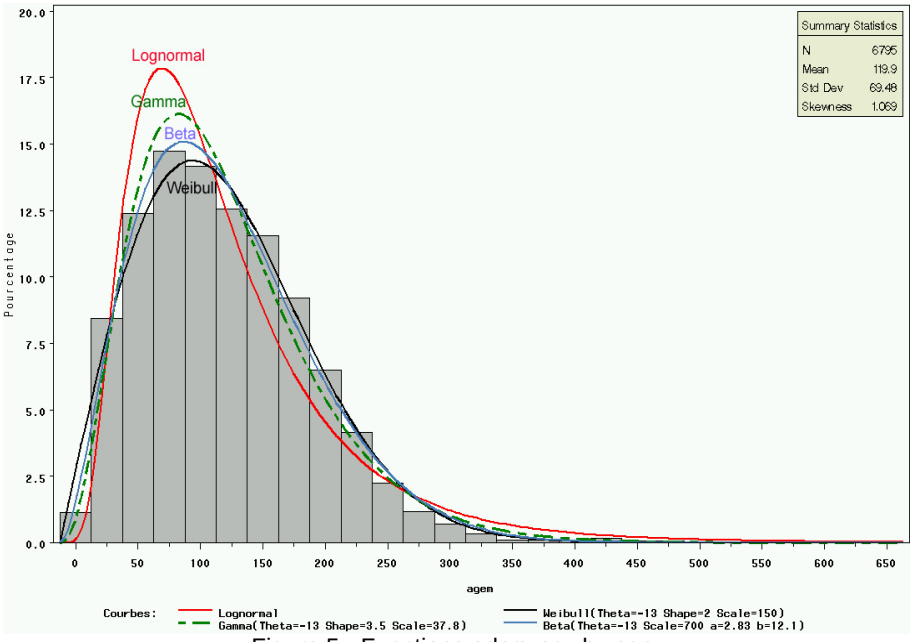


Figure 5 - Functions adequacy by age

Table 6 - Estimated parameters for distribution functions given by car age

Status	Location (θ)	Scale (β)	Shape (α)	Mean	Variance
Beta($t,\theta,\alpha,\beta,\sigma(scale)$)	-12.5	σ=700	α =2.829 β=12.074	120.39	68.84
Gamma(t,θ,α,β)	-12.5	37.826	3.502	119.93	70.78
Lognormal(t,θ,α,β)	-12.5	4.731	0.579	122.37	85.17
Weibull(t,θ,α,β)	-12.5	149.731	2.011	120.18	69.01

The pdf given by age in month is particularly well fitted[5] by a Beta distribution (e.g., Appendix A for qq-plot tests; Appendix B Table 8 for quantile distributions given for Beta and Weibull distributions) although the Weibull distribution give similar adequacy results (e.g., Appendix B Table 8). The Lognormal distribution doesn't provide a good fit of our sample. It has been pointed out Lawless, (1982) that the hazard function for the Lognormal model decreases for longer values of cycles. This does not agree with our knowledge concerning cars' progressive deterioration resulting from fatigue process, nevertheless the Lognormal pdf is often used satisfactorily to match shorter values of time-to-failure Lawless, (1982).
The Lognormal pdf cannot represent cars' lifetime exactly. The Lognormal is quite limited to represent lifetime data compared to the Weibull. For example, the Weibull can be either positively or negatively skewed but the Lognormal can only be positively skewed.
The Gamma distribution gives better adequacy parameters than the Lognormal but still far behind the Weibull or the Beta (e.g., Appendix A). It is known that, the Weibull distribution is based on more physically convincing arguments than the Gamma distribution. Moreover, the Weibull pdf is well suited for certain procedures of statistical extrapolation to large systems and has the advantage of dealing easily with our censored data.
More importantly, from the view of statistics, as the data are well fitted by the Beta's and Weibull's pdf, these two distributions might be applicable to model our data hazard function. Therefore cars' longevity (in age) might be matched accurately by Beta or Weibull survival functions.

6.2 Modelisation by mileage

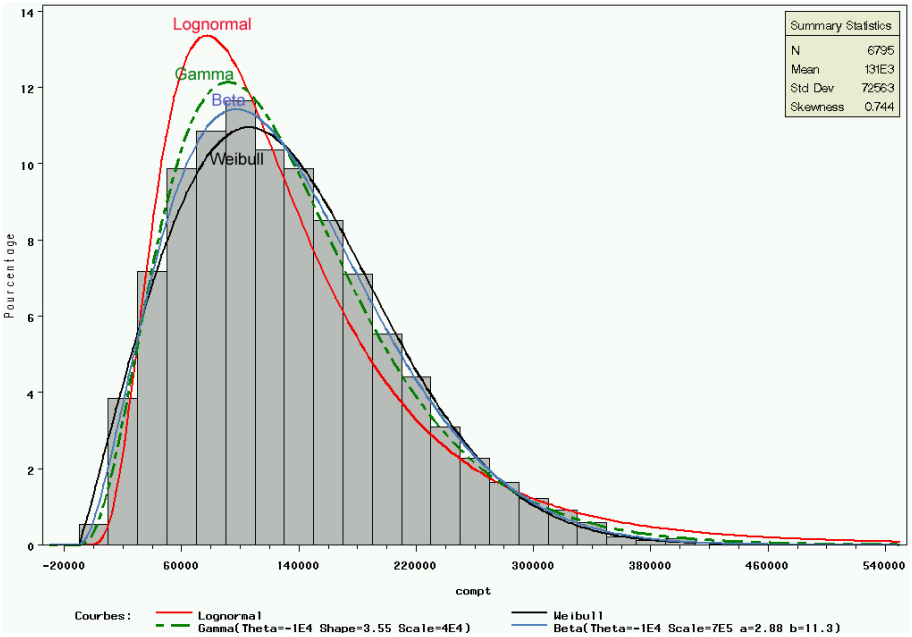


Figure 6 - Functions adequacy by mileage

Table 7 - Estimated parameters for distribution functions given by car mileage

Status	Location (θ)	Scale (β)	Shape (α)	Mean	Variance
Beta($t, \theta, \alpha, \beta, \sigma(scale)$)	-10000	σ=700000	α =2.876 β=11.345	131565	72068
Gamma(t, θ, α, β)	-10000	39896	3.546	131472	75127
Lognormal(t, θ, α, β)	-10000	11.711	0.578	137790	105881
Weibull(t, θ, α, β)	-10000	160053	2.064	131780	72048

[5] The maximum likelihood method of estimation was used to fit all four models. The Newton-Raphson algorithm, having a quadratic rate of convergence, was used to optimize the likelihood equations solving for the unknown parameters with maximum likelihood estimates and for each fixed location parameter.

The pdf given by mileage is also particularly well fitted by the Beta distribution (e.g., Appendix A for qq-plot tests and Appendix B Table 8). The superiority of the Beta distribution is obvious against the Lognormal's and the Gamma's. But like previously, the Weibull distribution is not far of these results and from the empirical quantile distribution (e.g., Appendix B Table8). Like for car age , as the data are well fitted by the Beta's and Weibull's pdf then our cars' longevity given in mileage might be matched accurately by Beta or Weibull survival functions.
These results from parametric estimation shows that both the Beta and Weibull pdf often seems to provide a good fit of cycles to failure data for our marker process Y(r) which is car mileage and for our parent process X(r) which is car age.
More importantly this distribution given by mileage shows the interest of a comparison by total driven mileage versus the classic comparison by age. We can also use mileage as a determinant of car life duration. More precisely, if we consider that car age is a parent process, car mileage can be called marker process defined by E.A. Peña, (2006): *"A marker process refers to an external process that covaries with the parent process and hence, It assists in tracking progress of the parent process if the parent process is latent or only infrequently observed. Marker processes may also be in scientific interest in their own right. As markers of the parent process, they offer potential insights into the causal forces that are generating the movements of the parent process"*.

The basic analytical framework for a marker process conceives of a bivariate stochastic process [X(r),Y(r)] where the parent process X(r) would be car age and the marker process Y(r) the mileage (Both are one dimensional and defined in the running time scale r). In the next section we work on these bivariate distributions to catch the link between our marker and parent process.

7 BIVARIATE DISTRIBUTIONS

In the following, figure (a) to (e) represents bivariate distributions given by age and by mileage. If we let $X_i = (x_i, y_i), i \in [\![1, n]\!]$ be a sample of size *n* drawn from this distribution. The kernel density estimate of $\mathrm{f}(\mathrm{x}, \mathrm{y})$ based on this sample will be given by:

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \varphi(x - X_i, y - Y_i) = \frac{1}{n h_x h_y}\sum_{i=1}^{n} \varphi\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right)$$

where $(x,y) \in \Re^2, h_x > 0, h_y > 0$ are the bandwidth and $\varphi_h(x,y)$ the rescaled normal density:

$$\varphi_h(x,y) = \frac{1}{h_x h_y}\varphi\left(\frac{x}{h_x}, \frac{y}{h_y}\right)$$

Where $\varphi(x,y)$ is the standard normal density function :

$$\varphi(x,y) = \frac{1}{2\pi}\, e^{\left(-\frac{x^2+y^2}{2}\right)}$$

7.1 Bivariate distributions by motorization

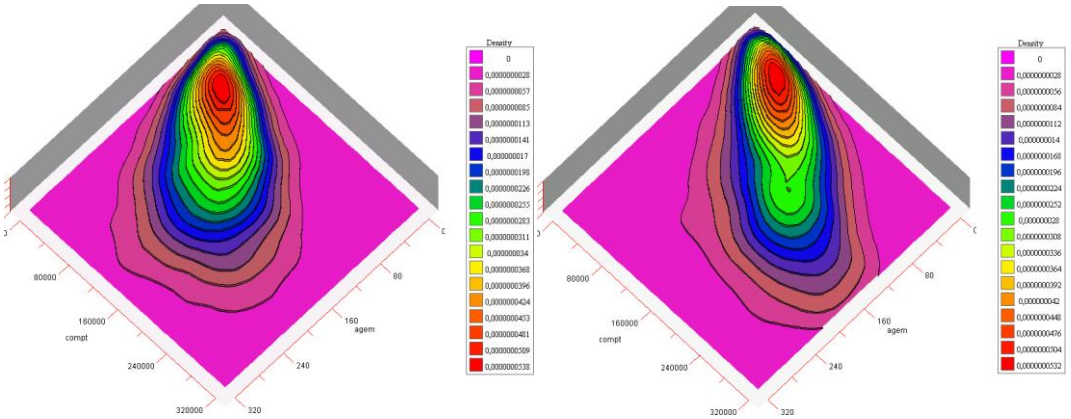


Figure 7 - (a) Bivariate distribution for gasoline cars - (b) Bivariate distribution for diesel cars

Figures 7 (a) and (b) show that for the gasoline car population the bivariate pdf is centered compared to diesel car one. For practical reasons, our bivariate distributions are given in a grid with an upper bound corresponding to 320 000 kilometers and to 320 months (26 years and 8 months). Gasoline car population reach its maximum for 81 months and 76520 km as for diesel cars it is around 54 months and 87625 km. The spread is also more important for diesel than for gasoline car population.
In addition to that a substantial proportion of the diesel car population still in the fleet after 240 000 kilometers (some after 320000 km) while for gasoline cars the bivariate pdf ends around 260 000 kilometers.
Figures 7 (a) and (b) enlighten us on a heterogeneousness of the behavior in automobile usage depending on motorization: diesel cars population travel more than gasoline cars population. Moreover, bivariate distributions are equilibrated both by marker process (car mileage: Y(r)) and parent process (car age: X(r)).

7.2 Bivariate distributions by car status

The bivariate pdf is given by category of status in Figures 8, (c) (d) and (e) is showing that the spreading of the density increase in a radical way with the status.
Moreover, age and mileage of maximums and dispersion of points increases with status. We can say that secondary cars stay longer than main cars, and that main cars stay longer than secondary car. But we also know that a quite important proportion of cars are intended to change of status one, two or more times during their lifetime in the same or in an other household.

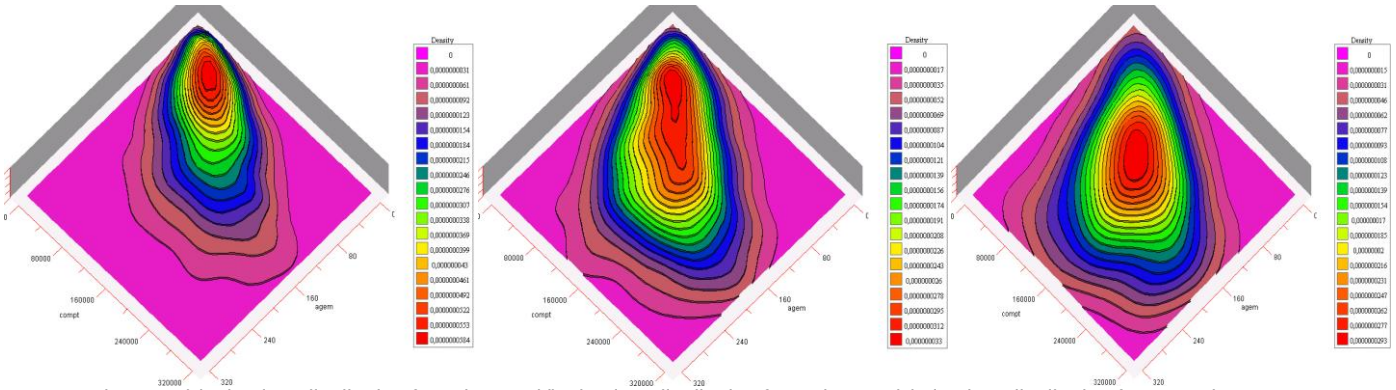


Figure 8 -(c) Bivariate distribution for sole cars (d) Bivariate distribution for main cars (e)Bivariate distribution for secondary cars

8 CONCLUSION

Car status (i.e., sole, main or secondary), an indicator of the intensity of use, appears to be a major determinant of car longevity in our sample of the French car fleet. We noticed better survival rates for cars declared as secondary cars. With the assumption that the status remains the same during the whole life of each car, secondary cars are maintained longer than main cars in the French fleet. And respectively, main cars are maintained longer in the fleet than sole cars. But we know that a car status is not fixed during a whole life duration and that status can move from one to another of our three categories. Consequently we have to be careful with these findings in a dynamic context of car transaction market. Motorization has a big impact too. Gasoline cars are maintained longer in the fleet than diesel cars and here too this is explained by an heterogeneity of car use behavior.

The reason of relatively good survival rates observed for some categories are mainly found in car use behavior. In that sense, this study underlined the negative correlation between car use and longevity. However, car survival duration could also be influenced by other variables: Bhat and Sen (2006), Bhat et al.(2008) conclude that vehicle-holdings and miles of travel vary with demographic characteristics, vehicle attributes, fuel costs etc... In our case study, car use and miles of travel depends mostly of car status and car motorization determinants. As car life duration depends mostly of car use, it also depends on status and motorization. The correlation between age and mileage vary by status and by motorization precisely because there is an heterogeneity in use behavior and miles of travel in each of these subpopulations.

The idea of integrating car mileage in survivor function is not new but was not widely explored mainly because of practical and technical reasons. First because odometer data are not always available or well described, as explained by Zachariadis et al[6], (1995)*: "Thinking about vehicle populations, one could reasonably assume that the most suitable unit to express their age is the mileage accumulated per vehicle since the beginning of its useful life. Expressing age in terms of years since entering the market is nevertheless equally appropriate and more practical (…) Statistics always group vehicles according to their model year or year of entering the market, not according to their accumulated mileage".* Secondly because of the problem of unobserved heterogeneity, as explained by Chen and Niemeier, (2004): *"One difficulty in specifying this model is the inclusion of a variable such as odometer reading, which is clearly important. Variables such as odometer readings are more complicated to handle during the estimation process compared to variables such as vehicle age. This is because as long as these variables (e.g., odometer readings) are observed, the vehicle is not scrapped."*
Contrary to what is advanced by Zachariadis et al., (1995) cars' age and cars' mileage would enlighten us differently on car use behavior mostly because their correlation is different for each subpopulation of car status and motorization and could be different for each country. As defined by E.A. Peña, (2006): *"Marker processes may also be in scientific interest in their own right. As markers of the parent process, they offer potential insights into the causal forces that are generating the movements of the parent process"*. In that sense car survival time can be explained both by its mileage which could be defined as a marker process or by car age defined as a classic calendar time which would be the parent process. A perfect knowledge of car longevity require to capture the correlation between marker and parent processes of each determinant that can affect cars' longevity. Moreover, the correlation structure between age and mileage is quite important to distinguish engineering scrappage which refers to vehicle scrappage due to engineering failures or physical deterioration from cyclical scrappage which refers to non-engineering related vehicle scrappage (e.g., cash for clunkers, primes à la casse, etc..).

[6] The study gives a median lifetime of 11.5 years for France and Spain, 9.6 years for the former West Germany, 12.9 years for Italy, 20.6 years for Portugal and 9.7 years for the United Kingdom.

The literature did not provide models integrating car mileage in the survival functions (i.e., as well as car status or rank, cylinder capacity or horsepower etc…). This research rests on the extensive existing literature, by examining the link between car use, car age and car longevity. But also by giving answers concerning functional forms of pdf, survival functions and main determinants (i.e., motorization and car status) that should be used to model and forecast car fleet.

There are several areas that remain open for further exploration. Car scrapping data are not as readily available as car registration data but it would be valuable to collect such data in upcoming studies. Their contribution to further study could help to bypass the problem of right censored data and helpful to compute robust Kaplan-Meier survival curves. Recognition of vehicle use patterns along with unexpected vehicle loss (due to crashes and theft, for example) would enhance the information content and reliability of results.

In order to give robust survival parametric model we also discussed the problem of comparative distributions. Parametric models are not always robust mainly because their adjustment is performed on a small amount of annuals survival rates or because the choice of the pdf is not justified mostly when a logistic function characterized by its lack of flexibility (i.e., versus Lognormal, Gamma etc…) is employed to model. But we showed that Beta and Weibull distributions could be good choices to model car's longevity. These two distributions fit the data very well mostly because they are well suited to model cars' progressive deterioration resulting from fatigue process. Although it is well known that they are flexible distributions.

In the near future, It would be of great importance to focus on determinants of multi-motorization and diesel cars' diffusion on car life expectancy in France, E.U. and U.S. for example. Moreover calculated survival rates for different car categories should take into consideration the complex relationship among motorization and cars status of the car fleet population. An integration of declared mileage in the hazard specification by collecting mileage of cars send to scrappage to solve the problem of internal time dependent covariates (Chen and Niemeier, 2004) would be an important breakthrough in the field of cars' longevity, car fleet and GHG forecasts.

ACKNOWLEDGMENTS: The authors are grateful to two anonymous referees for valuable suggestions. Thanks are also due to Jean-Luc Wingert for careful reading of a first version of the manuscript.

REFERENCES

Berkovec, J.: New Car Sales and Used Car Stocks: A Model of the Automobile Market, Rand Journal of Economics (1985).

Bhat, C.R. Sen, S. Eluru, N. : The Impact of Demographics, Built Environment Attributes, Vehicle Characteristics and Gasoline Prices on Household Vehicle Holdings and Use, Transportation Research Part B (2008), 43(1): 1-18.

Chen, C. Niemeier, D.: A mass point vehicle scrappage model, Transportation Research Part B (2004), 39: 402-414.

Choo, S. Mokhtarian, P.: What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice, Transportation Research Part A (2004), 38(3): 201-222.

Cox, D.R. Oakes, D.: Analysis of survival data, Chapman and Hall (1984), London.

Duchesne, T. Lawless,J. P: Alternative time scales and failure time models, Lifetime Data Anal. 6 (2000),157-179.

Duchesne, T. Lawless,J. P: On the collapsibility of lifetime regression models. Alternative time scales and failure time models, Adv. in Appl. Probab.35 (2003),755-772.

Gallez, C. Couturier, N.: Renouvellement du parc automobile et prospective transport-environnement, Rapport Ademe-Inrets (2000).

Greene, D. Chen, C. K. E.: Scrappage and Survival Rates of Passenger Cars and Light Trucks in the U.S. 1966-77, Transport Research Part A (1981) , 15 383-389.

Greenspan, A. Cohen, D.: Motor Vehicle Stocks, Scrappage, and Sales, Finance and Economics Discussion Series (1996), 40 (Washington, D.C.: Federal Reserve Board).

Hamilton, B. Macauley, M.: Heredity or Environment: Why is automobile longevity increasing?, Journal of Industrial Economics (1999), 47 (3), 251-261.

Kaplan, E.L Meier, P: Nonparametric estimation from incomplete observations, Journal of the American Statistical Association 53,(1958), 457-481.

Kiefer N.M. : Economic Duration Data and Hazard Functions, Journal of Economic Literature vol. XXVI (june 1988), 646-679.

Lawless, J.F.: Statistical Models and Methods of life time data, J WILEY, New York, (1982).

Leemis L.M.: Computing the nonparametric estimator of the survivor function when all observations are either left-or right-censored with tied observation times, Computer and Operations Research 29 (2002), 423-431.

Lynch, R.A., Lee, L.F. A Statewide Aggregate Model for Forecasting Vehicle Miles of Travel and Fuel Consumption in California. Division of Transportation Planning, California Department of Transportation (1979), pp.1–44.

Madre J.-L., Armoogum J.: Accuracy of data and memory effects in home based surveys on travel behavior, 76th annual meeting of Transportation Research Board, Washington.

Oakes D.: Multiple time scale in survival analysis, Lifetime Data Anal. 1 (1995), 7-18.

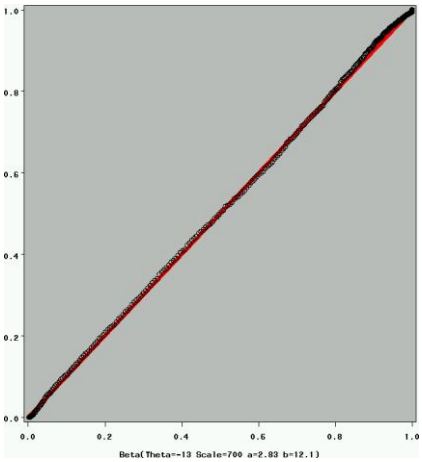
Peña E.A. : Dynamic Modeling and Statistical Analysis of Event Times, Statistical Science 21 (2006),487-500.

Southworth, F: VMT Forecasting for National Highway Planning: A Review of Existing Approaches, Oak Ridge National Laboratory, Oak Ridge (1986), TN 37831.

Zachariadis, T. Samaras, Z. Zierock, K.H.: Dynamic Modeling of Vehicle Populations: An Engineering Approach for Emissions Calculations, Technological Forecasting and Social Change 50 (1995), 135-149.

APPENDIX A- QQ-plots

QQ-plots are quite useful to compare data distribution with each family of distribution that vary in location and scale. They are here computed for each distribution family to compare the quantiles of the data distribution with the quantiles of the theoretical distribution family. In the following the reference line (in red) represent the particular theoretical distribution that depends on the location and scale parameters. This theoric distribution have an intercept and a slope that are equal to the location and scale parameters of that distribution.



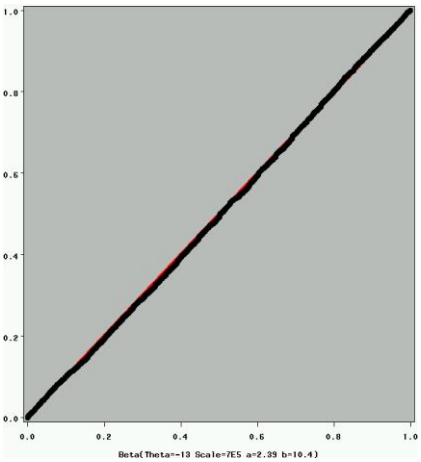
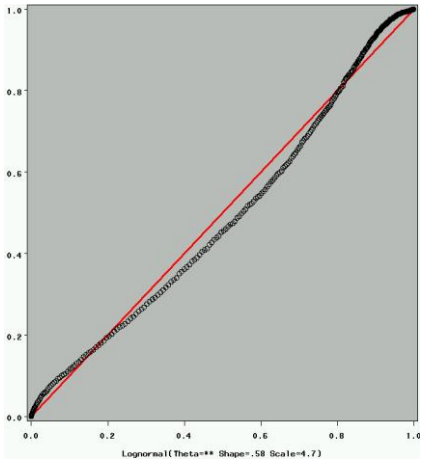


Figure 9 - (h) Beta QQ-plot by AGE - (i) Beta QQ-plot by Mileage



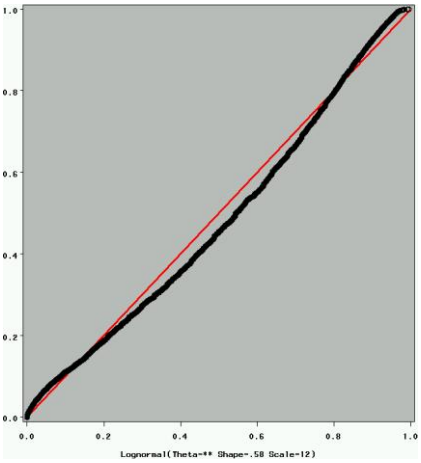
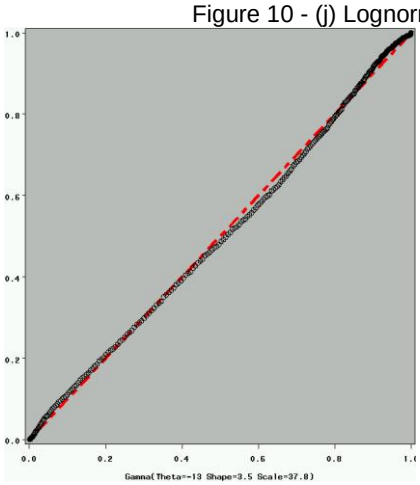


Figure 10 - (j) Lognormal QQ-plot by AGE - (k) Lognormal QQ-plot by Mileage



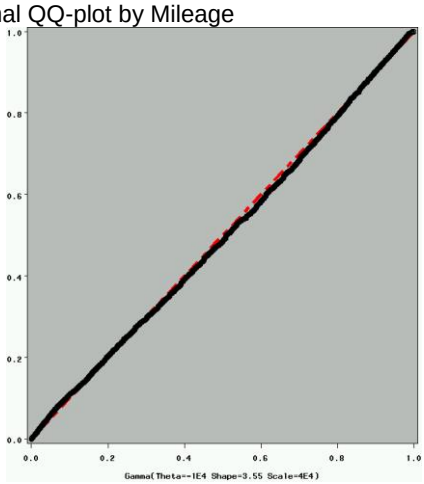
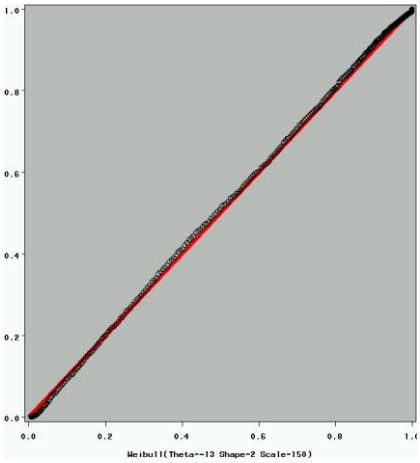


Figure 11 - (l) Gamma QQ-plot by AGE - (m) Gamma QQ-plot by Mileage



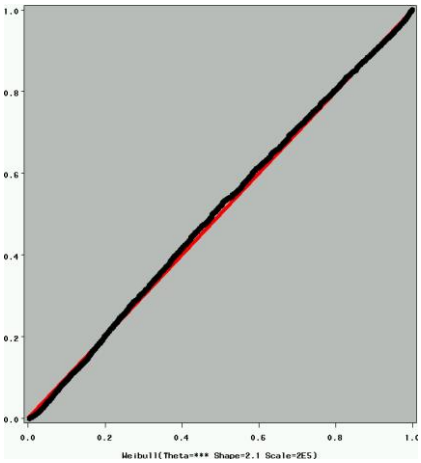


Figure 12 - (n) Weibull QQ-plot by AGE - (o) Weibull QQ-plot by Mileage

APPENDIX B-Quantile distributions

Table 8 - Quantile by age for observed sample, Beta and Weibull distributions

Percents	Observed	Estimated Beta	Estimated Weibull
1.0	12	7.7	2.7
5.0	25.0	25.9	21.7
10.0	38.0	39.4	36.4
25.0	68.0	68.5	68.1
50.0	111.0	110.6	112.3
75.0	162.0	162.1	163.6
90.0	209.0	215.0	214.2
95.0	238.0	248.7	245.9
99.0	316.0	314.0	307.4

Table 9 - Quantile by mileage for observed sample, Beta and Weibull distributions

Percents	Observed	Estimated Beta	Estimated Weibull
1.0	14380	12154	7228
5.0	32000	31817	27952
10.0	45625	46265	43791
25.0	76000	77202	77515
50.0	120145	121645	124010
75.0	175059	175667	177499
90.0	230710	230627	229759
95.0	268340	265425	262365
99.0	332934	332306	325457