Published in "Transportation", 2020, vol. 47, pp. 1027-1045, which should be cited to refer to this work.
DOI : 10.1007/s11116-018-9933-0

A RANDOM HEAPING MODEL OF ANNUAL VEHICLE KILOMETRES TRAVELLED CONSIDERING HETEROGENEOUS APPROXIMATION IN REPORTING

by

Toshiyuki YAMAMOTO, Jean-Loup MADRE, Matthieu de LAPPARENT and Roger COLLET

Toshiyuki Yamamoto
Institute of Materials and Systems for Sustainability, Nagoya University
C1-3(651) Furo-cho, Chikusa-ku, Nagoya 464-8603 JAPAN
Tel: +81-52-789-4636, Fax: +81-52-789-5728
E-mail: yamamoto@civil.nagoya-u.ac.jp

Jean-Loup Madre
Lab. Economic and Social Dynamics of Transports (DEST), AME/IFSTTAR, French Institute of Science and Technology for Transport, Development and Networks
14-20 Boulevard Newton B524, Cité Descartes, Champs-sur-Marne
F-77447 Marne la Vallée Cedex 2, FRANCE
Tel: +33 1 81 66 86 24
E-mail: jean-loup.madre@ifsttar.fr

Matthieu de Lapparent
University of Applied Sciences of Western Switzerland
HEIG-VD -- School of Engineering and Management Vaud
Avenue des Sports 20, 1401 Yverdon-les-Bains, Switzerland
Tel: +41 (0) 76 429 72 01
E-mail: matthieu.delapparent@heig-vd.ch

Roger COLLET
Lab. Economic and Social Dynamics of Transports (DEST), AME/IFSTTAR, French Institute of Science and Technology for Transport, Development and Networks
14-20 Boulevard Newton B524, Cité Descartes, Champs-sur-Marne
F-77447 Marne la Vallée Cedex 2, FRANCE
Tel: +33 1 81 66 86 24
E-mail: roger.collet@laposte.net

Abstract:
Annual vehicle kilometres travelled (VKT) is a long used index of car use. Usually, the annual VKT, as reported by respondents, is used for the analysis. But the reported values almost systematically contain approximations such as rounding and heaping. We apply a latent class approach in modelling VKT to account for this problem. Our model takes the form of a mixture of ordered probit models. The level of coarseness in reporting is considered as a latent variable that determines a category the respondent may belong to. Ordered response probit models of VKT are developed for each category. Thresholds are predetermined and model the level of coarseness that relates to the category. Annual VKT is itself assumed to affect the level of coarseness in reporting, thus included as an explanatory variable of the latent coarseness model. It is also modelled by an ordered probit model. The data set used in this study is a panel data of French households' vehicle ownership (Parc-Auto panel survey). The results confirm that the longer VKT results in a larger coarseness in the report. The results also suggest that the coarseness in the report of VKT is larger for large car than others. The coefficient estimates on the VKT function are not statistically different from those estimated by conventional regression model of VKT, however, the estimated variance of the error term in the VKT function for the latent class model is smaller than that for conventional regression model, implying that the latent class model better represents VKT than the conventional regression model does.

Keywords: bivariate ordered probit model, coarseness, latent class model, rounding, vehicle use

Introduction

Long term trend of road traffic is a major determinant of CO2 emissions, with their consequences in terms of fossil fuel consumption and of Global Warming (c.f. EU White Paper of 2011). That is why a particular attention is paid to the balance of fuel consumption resulting from the number of vehicles in use, multiplied by their fuel efficiency and by their annual mileage. For instance, this exercise is conducted every year by the Commission of National Transport Accounts (CCTN, 2013) in France. Moreover, EUROSTAT is planning to generalise this approach all over Europe, taking advantage in particular of the generalisation of the compulsory periodical inspection of vehicles. Indeed, very few countries are conducting panel surveys on car ownership and use like in France.

Thus, annual vehicle kilometres travelled (VKT) is a crucial and long used indicator, which characterises car use and travel patterns of households. There have been many studies that model it for various purposes such as gasoline consumption, vehicle emissions, and exposure to road accidents (Musti and Kockelman, 2011). However, the goodness-of-fit in modelling VKT is relatively low in general. For example, R-square of standard linear regression models is about 0.11 in Train (1986), 0.15 in Kockelman (1997), 0.17 in Yamamoto et al. (2001). One of the reasons for this is difficulty in fully representing its large variability across households. VKT has also been analysed together with the vehicle type choice behaviour. Discrete-continuous model frameworks have been applied in several studies (Bhat and Sen, 2006; Fang, 2008; Bhat et al., 2009; Spissu et al., 2009; Eluru et al., 2010; Brownstone and Fang, 2014; Liu et al., 2014; Cirillo et al., 2016; Liu and Cirillo, 2016). Explicitly recognising interactions between vehicle type choice and use is one of the advances in analysing VKT. The discrete-continuous model framework enables to rigorously investigate the indirect effect of particular factors on the vehicle usage through the vehicle type choice. The increase in goodness-of-fit as compared to just modelling VKT is however not explicitly documented. Another reason for the low goodness-of-fit might directly come from disaggregate data. Usually, an individual self evaluates and reports his/her annual VKT. It is then directly used as a dependent variable in some empirical modelling, although reported values contain approximations such as rounding

and heaping. Here, rounding occurs when observed values are reported only in round numbers, and a data set is "heaped" if it includes various levels of coarseness (Heitjan and Rubin, 1991). For example, a data set is "heaped" when large values of VKT are reported as the multiples of 5000 km and when low values of VKT are reported as the multiples of 1000 km. Since the ability of mental accounting and memorizing of annual VKT may vary among drivers, the reported VKT may contain various levels of rounding among drivers. There are three positive effects in explicitly addressing such rounding effects, as mentioned by Rietveld (2002) in the context of departure and arrival times: it leads to a considerably better treatment of reported information; biases in the computation of average based on the data can be avoided; it overcomes the problem of erratic patterns in the data.

Departure and arrival times are also known as vulnerable to rounding in conventional travel surveys. Madre and Armoogum (1997, 1998) have shown that arrival and departure time are more heaped when reported in a survey on long distance travel than on daily mobility, and more heaped when obtained by interview than through a car-diary with possible checking by the clock on the dashboard. Stopher et al. (2002) compared reported departure and arrival times in the standard trip-based CATI (computer aided telephone interview) survey with those obtained by GPS survey from the same respondents, and found that about 55% of the reported departure and arrival times are within 5 min of the correct time, but that the departure and arrival times have probably been rounded by most respondents with rounding to the nearest 5 or 10 min in most cases. Rietvelt (2002) estimated rounding models of departure and arrival times using a standard trip-based survey data. Without obtaining the correct times, he estimated the probabilities of various levels of rounding including 5, 15, 30 and 60 min assuming the equal probability of actual departure or arrival times within an hour. The results suggest that rounding is a rule rather than an exception, although that the reported arrival time is more accurate than departure time. Bhat and Steed (2002) considered the rounding of the reported departure time in developing a hazard-based duration model of departure time choice, but 5 min multiple of clock time is used as predetermined midpoint of the interval, and the possible larger levels of rounding reported in Stopher et al. (2007) were not considered.

Other than VKT and departure and arrival times, household income is also not precisely reported usually. However, the household income is in general measured in a discrete number of categories or intervals with fixed thresholds. In this case, ordered response models with fixed thresholds can be applied. Bhat (1994a, 1994b) applied ordered probit models, and Tong and Lee (2009) applied a hazard-based duration model for grouped income data. One additional problem in income data is the missing cases. The ordered response models can be used to impute an income measure for the missing data, but the systematic variations in unobserved characteristics between respondent and nonrespondent of income variable may exist, resulting biased imputations if not correctly considered. Bhat (1994a, 1994b) considered this problem by applying sample selection approach with bivariate ordered probit model. As stated above, income is usually measured with fixed thresholds, but the thresholds in reporting VKT is not fixed and may vary across respondents. Thus, the modelling of VKT needs approaches different from income.

In our modelling approach, rounding and coarseness in reporting VKT values is explicitly accounted for. At the disaggregate level, it is not feasible to assume that the level of coarseness does not vary across respondents. These levels of coarseness are latent outcomes: we do not observe them. For example, if the reported VKT is 15000 km, we know that the value is not rounded as the multiples of 10000 km, but we do not know it is rounded as the multiples of 5000 km, 1000 km, or smaller numbers. To this extent, we apply a model mixing ordered probit specifications. The latter models the observed reporting of VKT, given a rounding behaviour in reporting VKT. The mixture distribution, also based on an ordered probit model, models the latent behaviour of the respondent as it regards the willingness to round VKT when being

surveyed. We also consider that VKT itself may affect the coarseness in that the longer VKT may have a higher probability of higher levels of coarseness. Heitjan and Rubin (1990, 1991) developed a statistical model explicitly dealing with such various levels of rounding, called as heaping, in the context of anthropometric data on children's age from rural Tanzania. The statistical model is applied in this study for the reported VKT. Contrast to the anthropometric study where only age was treated as the factor to affect the coarseness, the effects on the coarseness is structuralised in this study, and socio-demographic characteristics as well as VKT are incorporated as the explanatory variables of the coarseness of the report.

Modelling methodology

The approach is built up on Heitjan and Rubin model (1990, 1991). It takes here the form of a discrete mixture of ordered probit models. Continuous and discrete probabilistic mixtures of (probabilistic) models become more and more standard practice. Conventionally, the model of VKT is given as

$$\ln(y_i^*) = \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i \tag{1}$$

where y_i^* is the VKT of vehicle i, $\boldsymbol{\beta}$ is a parameter vector, $\mathbf{x}_i$ is a vector of explanatory variables, and ε_i is a random variable following a normal distribution. Here, the VKT is not precisely reported, but the reported VKT is rounded. From the preliminary analysis, we assume that the reported VKT is rounded as multiples of 500km, 1000km, or 5000km. It means that y_i^* lies in the range $[y_i - 250, y_i + 250)$ if the reported VKT, y_i is rounded as multiples of 500km, that the range $[y_i - 500, y_i + 500)$ if multiples of 1000km, and that the range $[y_i - 2500, y_i + 2500)$ if multiples of 5000km.

The coarseness of the reported VKT by individual i is modelled as a latent variable. It is defined as a function of the actual VKT and of other determinants:

$$z_i^* = \alpha \ln(y_i^*) + \boldsymbol{\gamma}\mathbf{x}_i + \zeta_i \tag{2}$$

where $\alpha, \boldsymbol{\gamma}$ are parameters and ζ_i is a normally distributed random variable.

It is assumed that the coarseness of the report can be discretized:

$$\begin{aligned} z_i &= 1 \quad \text{if } z_i^* < 0, \\ &= 2 \quad \text{if } 0 \le z_i^* < \theta, \\ &= 3 \quad \text{if } \theta \le z_i^* \end{aligned} \tag{3}$$

where z_i is the coarseness of the report and θ is a threshold. The report is heaped as multiples of 500km if $z_i = 1$, multiples of 1000km if $z_i = 2$, and multiples of 5000km if $z_i = 3$. Eq. (2) shows that not only VKT but also socio-demographic characteristics affect the coarseness of the report. Note also that one of the thresholds is normalized at 0 for identification of an intercept term in the latent coarseness process as shown in Eq. (3). Following Heitjan and Rubin (1990), taking into account that VKT itself is included in the explanatory variables of the function of the coarseness of the report, $\ln(y_i^*)$ and z_i^* are given as bivariate normal with mean

$$E\begin{pmatrix} \ln y_i^* \\ z_i^* \end{pmatrix} = \begin{pmatrix} \beta\mathbf{x}_i \\ \alpha\beta\mathbf{x}_i + \gamma\mathbf{x}_i \end{pmatrix} \tag{4}$$

and covariance matrix

$$V\begin{pmatrix} \ln y_i^* \\ z_i^* \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 & \alpha\sigma_\varepsilon^2 \\ \alpha\sigma_\varepsilon^2 & \sigma_\zeta^2 + \alpha^2\sigma_\varepsilon^2 \end{pmatrix} \quad (5)$$

where σ_ε^2 and σ_ζ^2 are variances of ε_i and ζ_i, respectively. Without loss of generality, and for identification of α, $\sigma_\zeta^2 + \alpha^2\sigma_\varepsilon^2$ is normalized as 1, and correlation between y_i^* and z_i^* is given as $\alpha\sigma_\varepsilon$.

A region $S(y_i)$ of possible values for (y_i^*, z_i^*) can be defined that all map to y_i. First define the regions $L_i = [y_i - 250, y_i + 250)\times(-\infty, 0)$ corresponding to heaped as multiples of 500km, $M_i = [y_i - 500, y_i + 500)\times[0, \theta)$ corresponding to heaped as multiples of 1000km, and $H_i = [y_i - 2500, y_i + 2500)\times[\theta, \infty)$ corresponding to heaped as multiples of 5000km. Now, we don't know the levels of coarseness for a part of the respondents. If we observe y_i at multiples of 5000km (e.g., 10000km), it could be the result of heaped as multiples of 5000km, that of 1000km, and that of 500km. On the other hand, if we observe y_i at multiples of 1000km but not at multiples of 5000km (e.g., 8000km), it could be the results of the latter two. Thus, we have the region given as

$$\begin{aligned} S(y_i) &= L_i \text{ È } M_i \text{ È } H_i && \textbf{if } y_i = 0 \textbf{ mod 5000} \\ &= L_i \text{ È } M_i && \textbf{if } y_i = 0 \textbf{ mod 1000 and } y_i \text{ ¹ } 0 \textbf{ mod 5000} \\ &= L_i && \textbf{if } y_i = 0 \textbf{ mod 500 and } y_i \text{ ¹ } 0 \textbf{ mod 1000} \end{aligned} \quad (6)$$

The log-likelihood function for the parameters is given as

$$LL = \sum_{i=1}^{n} \ln \int_{S(y_i)} f\left(\ln y_i^*, z_i^*\right) dy_i^* \, dz_i^* \quad (7)$$

where $f(\ln y_i^*, z_i^*)$ is the bivariate normal given by Eqs. (4) and (5). The specification of the conditional probability that the latent coarseness level is of a given level and that the reported VKT belongs to a given interval takes the form of a bivariate ordered Probit (e.g. Bhat 1994a, 1994b). It is then further integrated to obtain its expected value with respect to the distribution of the latent coarseness level. More specifically, the log-likelihood function of Eq. (7) can be written as

$$LL = \sum_{i=1}^{n} \ln \left\{ \begin{array}{l} \Phi_2 \left\{ \left(y_i + 250 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ -\alpha\boldsymbol{\beta}\mathbf{x}_i - \boldsymbol{\gamma}\mathbf{x}_i ,\ \alpha\sigma_\varepsilon \right\} \\ -\Phi_2 \left\{ \left(y_i - 250 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ -\alpha\boldsymbol{\beta}\mathbf{x}_i - \boldsymbol{\gamma}\mathbf{x}_i ,\ \alpha\sigma_\varepsilon \right\} \\ +m_i \left[\begin{array}{l} \Phi_2 \left\{ \left(y_i + 500 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ \theta - \alpha\boldsymbol{\beta}\mathbf{x}_i - \boldsymbol{\gamma}\mathbf{x}_i ,\ \alpha\sigma_\varepsilon \right\} \\ -\Phi_2 \left\{ \left(y_i - 500 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ \theta - \alpha\boldsymbol{\beta}\mathbf{x}_i - \boldsymbol{\gamma}\mathbf{x}_i ,\ \alpha\sigma_\varepsilon \right\} \\ -\Phi_2 \left\{ \left(y_i + 500 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ -\alpha\boldsymbol{\beta}\mathbf{x}_i - \boldsymbol{\gamma}\mathbf{x}_i ,\ \alpha\sigma_\varepsilon \right\} \\ +\Phi_2 \left\{ \left(y_i - 500 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ -\alpha\boldsymbol{\beta}\mathbf{x}_i - \boldsymbol{\gamma}\mathbf{x}_i ,\ \alpha\sigma_\varepsilon \right\} \end{array} \right] \\ +h_i \left[\begin{array}{l} \Phi_2 \left\{ \left(y_i + 2500 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ \alpha\boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\gamma}\mathbf{x}_i - \theta,\ -\alpha\sigma_\varepsilon \right\} \\ -\Phi_2 \left\{ \left(y_i - 2500 - \boldsymbol{\beta}\mathbf{x}_i \right) / \sigma_\varepsilon ,\ \alpha\boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\gamma}\mathbf{x}_i - \theta,\ -\alpha\sigma_\varepsilon \right\} \end{array} \right] \end{array} \right\} \quad (8)$$

where $m_i = 1$ if $y_i = 0$ mod 1000, and 0 otherwise, and $h_i = 1$ if $y_i = 0$ mod 5000, and 0 otherwise. Maximum likelihood estimation can be applied to obtain parameter estimates.

Data

The data set used in this study is a panel data of French households' vehicle ownership obtained by the panel survey called Parc-Auto (Hivert, 2000). The survey adopted mail-out and mail-back self-administered questionnaires on vehicle ownership. The sample size has been maintained at about 7000 households each year. Rotation panel system is employed by the survey, where the participants were originally assigned to stay on the panel for four years. The questionnaire includes questions concerning the characteristics of up to three vehicles in the household, vehicle use in terms of odometer reading, annual mileage, main purposes of vehicle use, etc. Also included are the attributes of main driver and household. The data set includes a rough estimate of the annual VKT by the respondent, who is not always the main driver of the car (denoted reported VKT hereafter). In addition, the estimated annual VKT can be calculated from the difference in odometer readings reported much more precisely by the respondent at successive two surveys which is one year apart (denoted calculated VKT hereafter). Sample used for the empirical analysis of this study is 2257 vehicles for which the odometer readings were reported by respondent on both 2010 and 2011, and the reported VKT was also obtained at the survey in 2011. Hivert (2000) has shown that there is no significant bias between the reported VKT for 1998, and the odometer readings at the fall of 1998 and of 1997, which is no more the case as tested for 2009-10 and 2010-11. Indeed, the reported VKT has become significantly lower than the difference between odometer readings. Thus, the data is a little dated, but the dataset used in the study is well prepared, and sufficient for the investigation of the coarseness of the reported VKT without interfering with the analysis of the bias appeared since then. When the odometer reading is precise, there is no significant gap between the reported VKT for 2010, and the odometer readings at the fall of 2010 and of 2009, which is not the case when odometer is rounded as a multiple of 1000 km. Although the gap has become significant according to more recent surveys (2010 to 2015), it suggests the comparison between calculated and reported annual mileage is more meaningful.

Sample distributions of calculated VKT and reported VKT are shown in Figure 1 and Figure 2, respectively. As shown in Figure 1, calculated VKT follows approximately lognormal distribution, though some fluctuations can be admitted. On the other hand, as shown in Figure 2, reported VKT has several concentrations in the distribution. It is clear that the concentrations are located at multiples of 5000km. It confirms that the reported VKT contains rounding effects.

Scatter plot of sample cases on the calculated VKT and the reported VKT is shown in Figure 3. At first, although most of the plots are located closer to 45 degree line, several plots are found as located far from 45 degree line. It means some problems in the data set. The possible reasons for such outliers are that respondent reported totally wrong annual VKT, or totally wrong odometer readings in 2010 or 2011, or that the odometer readings for different vehicles obtained in 2010 and 2011 are mismatched resulting the wrong calculated VKT. However, we are not sure about the reason, so the dataset is used for the empirical analysis without any amendment. Figure 3 shows that many plots lie in horizontal lines at multiples of 5000km, the same as the concentrations shown in Figure 2. Also, similar horizontal lines can be found at other values of reported VKT, and it suggests that there exist the concentrations at multiples of smaller values than 5000km.

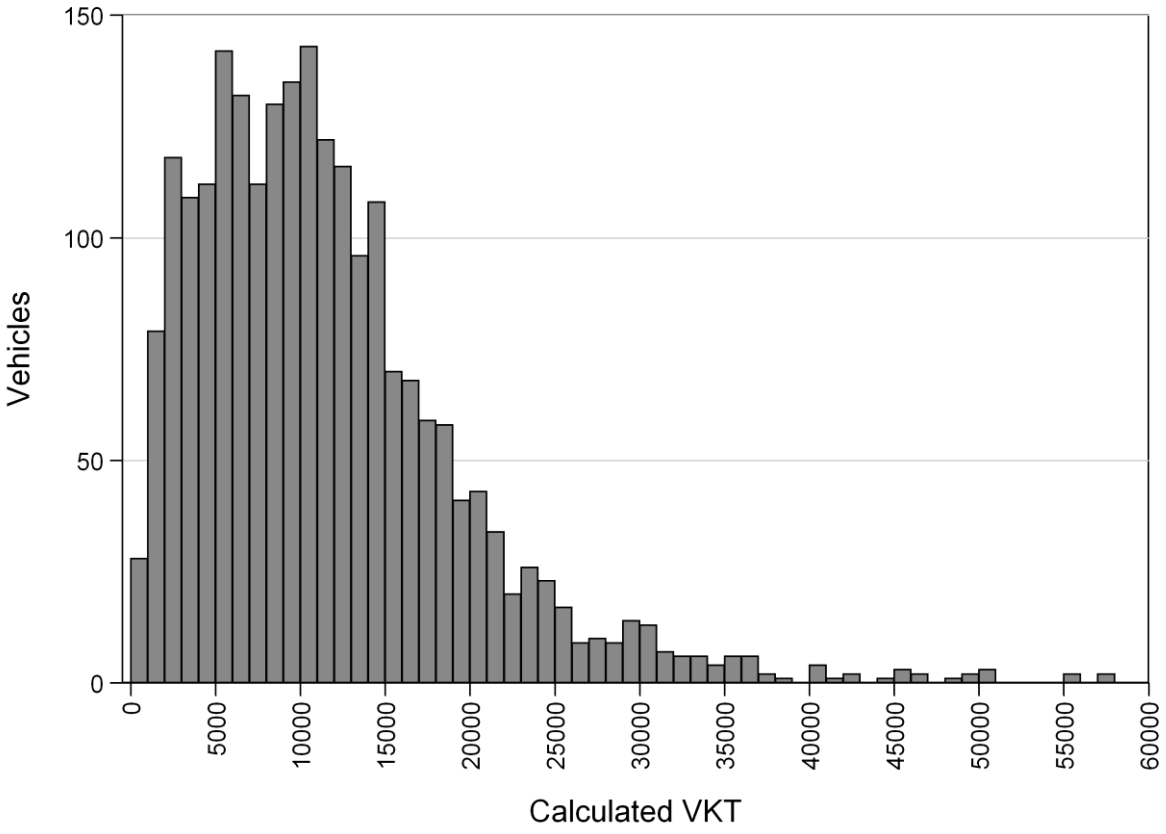

Figure 1 Sample distribution of calculated VKT

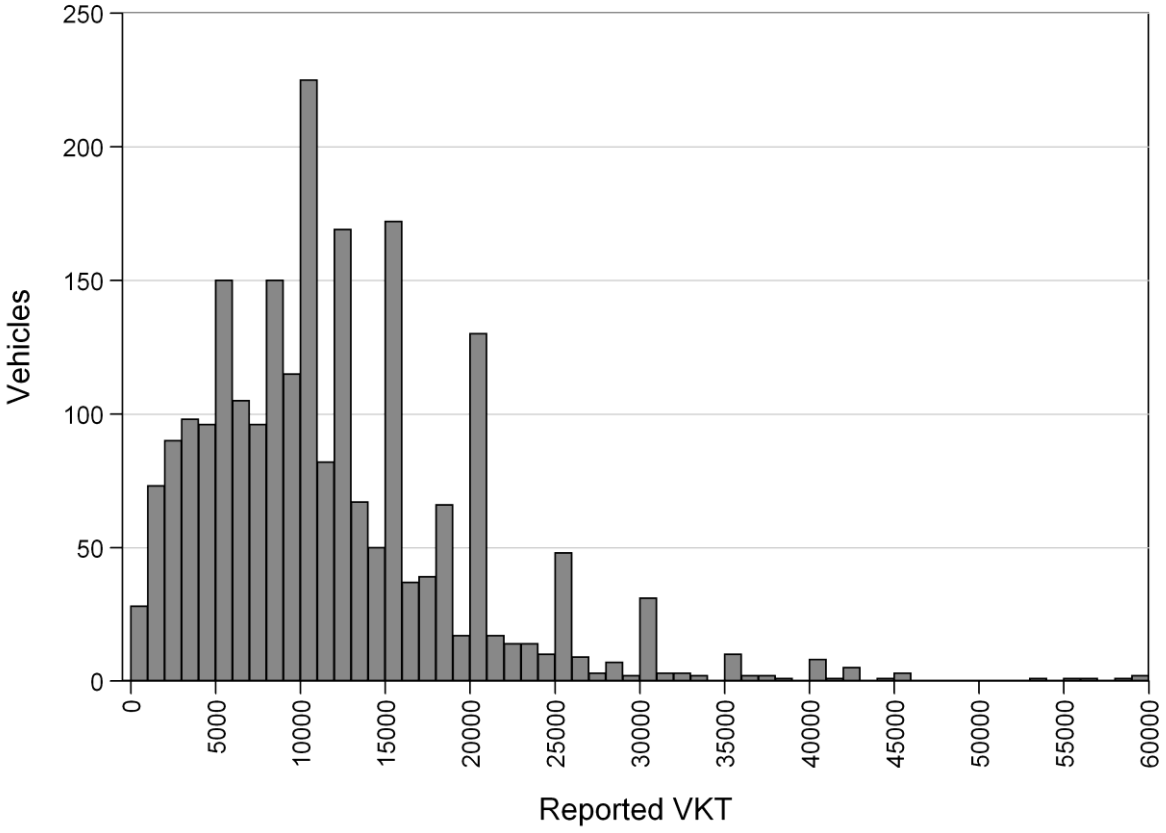

Figure 2 Sample distribution of reported VKT

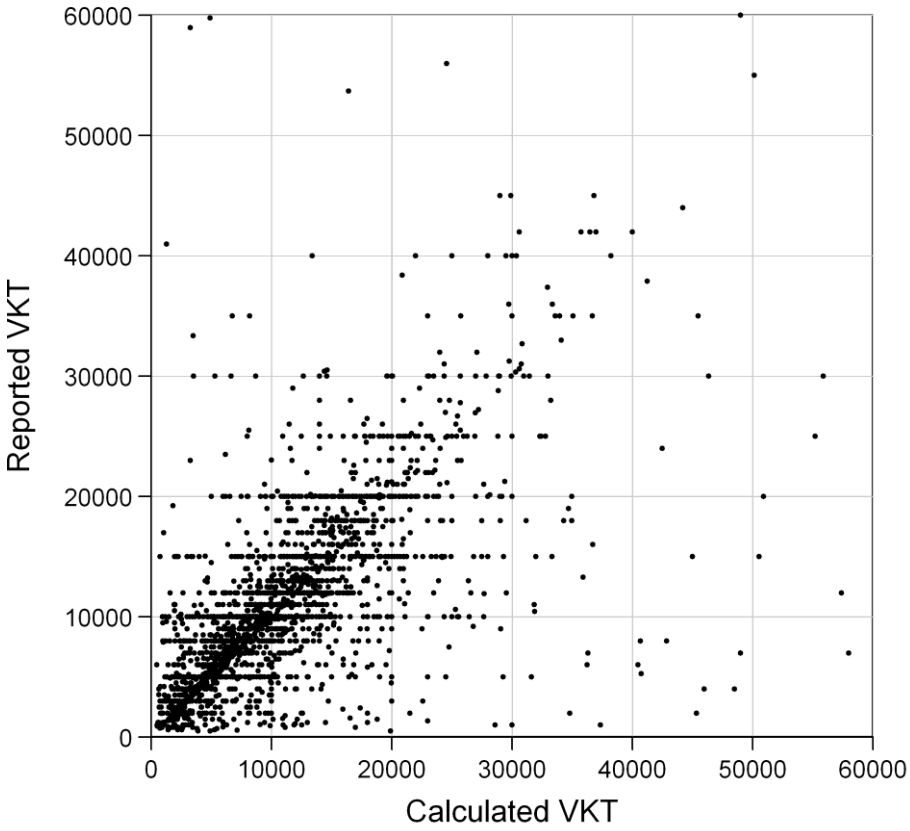

Figure 3 Scatter plot of calculated and reported VKT for sample cases

Sample distribution of reported VKT in terms of rounding is shown in Table 1. It confirms that many reported VKT are rounded at multiples of 5000km, but that more reports are rounded at multiples of 1000km excluding multiples of 5000km. The former contains the cases where the VKT is rounded as multiples of 1000km, so the cases where the VKT is rounded as multiples of 1000km dominate the cases where the VKT is rounded as multiples of 5000km. Table 1 shows that there also exist the cases where the VKT is rounded as multiples of 500km and the cases with smaller rounding than 500km. The latter might include the cases with no rounding, but we are not sure about how accurately the respondents can answer the VKT. The cases with smaller rounding than 500km are treated as rounded as multiples of 500km in the empirical analysis of this study for the simplicity of the model although it might result in biased estimation results.

Table 1 Rounding of reported VKT

	Cases
Multiples of 5000km	677
Multiples of 1000km excluding multiples of 5000km	876
Multiples of 500km excluding multiples of 1000km	187
Not multiple of 500km	517
Total	2257

Source: Parc-Auto panel survey 2011.

The explanatory variables used in the model of this study are summarized in Table 2. The explanatory variables contain attributes of household, attributes of main driver and vehicle attributes. The focus of the study is on the effects of the rounding on the estimation of VKT function, so the exploration of the explanatory variables has not gone beyond the basic set of the

variables.

Table 2 Variables used in the analysis

Variable	Definition	Mean	SD
Characteristics of main driver's household			
Children	Number of children under 15	0.343	0.734
PT access	Dummy: 1 if public transport is accessible from residence by less than 5 minutes on foot, 0 otherwise	0.621	0.485
Large City	Dummy: 1 if hh located in an urban area >200 000 inhabitants, 0 otherwise	0.354	0.478
Fleet size	Number of vehicles held by household	1.612	0.668
Low income	Dummy: 1 if household income is less than 15 000 euros per year, 0 otherwise	0.092	0.289
High income	Dummy: 1 if household income is over 45 000 euros per year, 0 otherwise	0.181	0.385
Characteristics of main driver			
Under 40	Dummy: 1 if main user is less than 40, 0 otherwise	0.206	0.405
Over 60	Dummy: 1 if main user is 60 or over, 0 otherwise	0.452	0.498
Worker	Dummy: 1 if main user is active on labour market	0.522	0.500
Male	Dummy: 1 if main user is a man, 0 othserwise	0.593	0.491
Vehicle attributes and VKT			
Commuting	Dummy: 1 if car is used for commuting, 0 otherwise	0.402	0.490
Diesel	Dummy: 1 if Diesel car , 0 otherwise	0.570	0.495
Small	Dummy: 1 if small car, 0 otherwise	0.482	0.500
Large	Dummy: 1 if large car, 0 otherwise	0.039	0.193
Light truck	Dummy: 1 if SUV or light truck, 0 otherwise	0.037	0.189
Car age	Vehicle age	7.669	5.124
Reported VKT	Annual VKT reported by respondents	11551	7719
Calculated VKT	Annual VKT calculated by the difference in odometer readings	11513	7966

Source: Parc-Auto panel survey 2010-11.

Results

First and before estimating the proposed model, the difference between reported and calculated VKT is examined in order to identify possible explanatory variables that may be included in the coarseness function. In this perspective, three standard linear regression models are estimated to evaluate the effects of vehicle attributes and driver characteristics on the absolute deviation of reported VKT from calculated VKT. The first model is estimated using the full sample of observations, the second is implemented using the subsample reporting an underestimated VKT with respect to calculated VKT, and the third model relies on the subsample reporting an overestimated VKT with respect to calculated VKT. In these models, the absolute deviation is taken in logarithm to be consistent with the dependent variable in the VKT function, which is also expressed in logarithm. As a consequence of this log-transformation, the cases with no difference between reported and calculated VKT are discarded from the sample.[1] The estimation results for these three models are shown in Table 3. They are globally consistent with each other and suggest that the deviation is larger for diesel than for petrol cars, shorter for small cars compared to medium-sized cars and larger for commuting cars than for others. Concurrently, annual VKT is expected to be higher for diesel, medium-sized and commuting cars compared to petrol, small and non-commuting cars respectively. Therefore, higher VKT

[1] 79 observations over 2257 in the dataset.

should induce larger deviations, thus supporting the inclusion of VKT as explanatory variable in the coarseness function. Moreover, annual VKT is usually lower on average in densely populated areas than in low-density areas, but the estimate related to the "large city" dummy is not significantly negative in the deviation models of Table 3, suggesting that urban cars might be subject to larger coarseness. Thus, the final set of explanatory variables retained in the coarseness function includes car fuel type and size, commuting use, owner's location and annual VKT.

Table 3 Regression models of the deviation of reported VKT from calculated VKT

	Absolute difference (a)		Absolute underestimate (b)		Overestimate (c)	
Variable	coefficient	t-stat.	coefficient	t-stat.	coefficient	t-stat.
Children	-0.013	-0.26	0.025	0.30	-0.060	-0.97
PT access	-0.052	-0.75	-0.084	-0.69	-0.022	-0.27
Large City	0.072	1.02	0.075	0.61	0.065	0.77
Fleet size	0.088	1.71	0.169	1.86	0.030	0.50
Low income	0.176	1.58	0.385	2.03	0.026	0.20
High income	-0.131	-1.53	-0.249	-1.70	-0.037	-0.36
Under 40	0.178	1.94	0.098	0.65	0.262	2.30
Over 60	-0.124	-1.00	-0.395	-1.81	0.105	0.73
Worker	0.052	0.44	0.014	0.07	0.062	0.43
Male	0.126	1.89	0.106	0.92	0.127	1.59
Commuting	0.264	2.95	0.097	0.67	0.419	3.74
Diesel	0.293	4.26	0.259	2.18	0.321	3.93
Small	-0.166	-2.42	-0.130	-1.09	-0.182	-2.25
Large	0.047	0.28	-0.416	-1.43	0.389	1.96
Light truck	-0.007	-0.04	-0.090	-0.33	0.036	0.18
Car age	-0.002	-0.32	0.012	1.06	-0.012	-1.64
Constant	7.109	38.46	7.175	21.48	7.028	33.00
Sample size	2178		908		1270	
R-squared	0.051		0.054		0.073	
RMSE	1.433		1.603		1.29	

Note: OLS estimates. The endogenous variable is ln(|reported VKT - calculated VKT|). Model (a) is estimated using the full dataset, model (b) relies on the subsample reporting an underestimated VKT compared to calculated VKT, and model (c) relies on the subsample reporting an overestimated VKT.
Source: Parc-Auto panel survey 2010-11.

The bivariate ordered response probit model has been implemented using GAUSS, a matrix-programming software which provides routines for maximum likelihood estimation. The likelihood function of the proposed model has been coded by the author and the estimation results are shown in Table 4 (column a). The coefficient estimates for the coarseness function are first discussed. Subsequently, the results for the VKT function are commented and compared to alternative models.

As expected, greater VKT results in significantly higher coarseness in reporting. Indeed, the estimate for α is positive and presents a very large t-statistic. Other than VKT, only the "commuting car" and "large city" dummies have significant estimated coefficients in the coarseness function, calling here for explanations. Commuting cars are driven on a longer annual distance but they are also much more often used than others in terms of number of trips. Thus, higher car use frequency is possibly another source of approximation in reporting annual VKT. In addition, drivers living in large cities can more easily substitute public transports for car use during a whole year. This may lead to more coarseness in reporting annual VKT than for drivers residing in low-density areas, where the choice of car is more systematic due to a lack of

efficient public alternatives, making annual VKT easier to determine accurately.

Table 4 Estimation results of VKT models

	Proposed model of reported VKT (a)		Regression model of reported VKT (b)		Regression model of calculated VKT (c)	
	coefficient	t-stat.	coefficient	t-stat.	coefficient	t-stat.
Coarseness function						
Log-VKT (α)	0.693	12.22				
Large City	0.151	2.10				
Commuting	0.337	4.70				
Diesel	0.135	1.72				
Small	-0.008	-0.10				
Large	0.166	1.03				
Light truck	0.218	1.28				
Constant	-6.876	-13.01				
Threshold (θ)	0.630	11.42				
VKT function						
Children	-0.015	-0.64	-0.015	-0.67	0.003	0.14
PT access	-0.071	-2.30	-0.069	-2.23	-0.116	-3.66
Large City	-0.083	-2.67	-0.085	-2.68	-0.040	-1.24
Fleet size	-0.026	-1.16	-0.026	-1.14	0.003	0.14
Low income	-0.155	-3.68	-0.157	-3.17	-0.051	-1.01
High income	0.052	1.32	0.054	1.42	0.033	0.84
Under 40	0.033	0.77	0.031	0.76	0.055	1.31
Over 60	-0.059	-0.97	-0.060	-1.09	-0.178	-3.17
Worker	-0.081	-1.38	-0.082	-1.52	-0.104	-1.91
Male	0.153	5.15	0.152	5.09	0.164	5.39
Commuting	0.456	11.16	0.458	11.42	0.380	9.34
Diesel	0.338	11.24	0.339	11.02	0.343	10.98
Small	-0.252	-8.41	-0.253	-8.23	-0.232	-7.43
Large	0.151	1.86	0.152	2.05	0.024	0.32
Light truck	-0.149	-2.11	-0.148	-1.98	-0.077	-1.01
Car age	-0.036	-13.32	-0.036	-12.56	-0.034	-11.61
Constant	9.226	106.70	9.225	111.42	9.223	109.61
Std. error (σ_ε)	0.641		0.652		0.663	
Sample size	2257		2257		2257	
R-squared	-		0.30		0.29	
Log-likelihood at convergence	-8215		-		-	

Note: the endogenous variable in the VKT function is expressed in logarithm. FIML estimates in column (a), OLS estimates in columns (b) and (c).
Source: Parc-Auto panel survey 2010-11.

The estimation results for the VKT function are standard and do not diverge from other studies dealing with car use in France. While the car size and fuel type dummies turned out to be non-significant in the coarseness function, the “diesel” and “small car” dummies have statistically significant coefficient estimates in the VKT function. Given that higher VKT entails significantly more coarsened report, this result is consistent with the regression estimates of the deviation models (Table 3), in which both these car attributes have significant estimates. Because of a lower fuel price per litre and a better fuel efficiency on average, the annual distance travelled is significantly longer for diesel than for petrol cars. The result on the positive relationship between the fuel efficiency and the annual distance travelled is consistent

with previous studies, where the driving cost per mile had a negative effect on the usage (Liu et al., 2014; Cirillo et al., 2016). It is significantly shorter for small cars than for medium-sized or large cars and decreasing as the vehicle age increases. As expected, drivers living in large cities make significantly lower use of their car than drivers from low-density areas, while working drivers make greater use than unemployed drivers only if they commute by car. In addition, French driving men make significantly greater use of their car than women, but the results show no evidence of a difference according to the driver's age.[2] The gender difference on vehicle use is also consistent with the literature. Spissu et al. (2009), Liu et al. (2014) and Cirillo et al. (2016) suggested male drivers have longer annual vehicle miles than female except that Spissu et al. (2016) have an opposite result only for coupé drivers. Drivers from low-income households make lower car use than those living in medium or high-income households, and an access to public transports near drivers' home induces a significant decrease in their car mileage. Both results are again consistent with the literature. Train (1986), Brownstone and Fang (2014), Liu et al. (2014) and Cirillo et al. (2016) suggested high-income households had a longer annual vehicle miles. Also, Train (1986) suggested the number of trips by public transit had a negative effect on annual vehicle miles, and Liu et al. (2014) and Cirillo et al. (2016) suggested the negative relationship between the residential density and annual vehicle miles. Lastly, the total number of cars and the number of children in drivers' households have no significant impact on their car use.

The coefficient estimates discussed above for the VKT function are not statistically different from those estimated by conventional linear regression models (Table 4, columns b and c), This suggests that not considering heterogeneous coarseness in the VKT reports does not result in a significant estimation bias. However, the estimated variance of the error term is lower in the VKT function of the proposed model than in the conventional regression models, implying a better representation of VKT if coarseness is taken into account. The proposed model can also be compared to univariate ordered response models. In these, it has been successively assumed a fixed coarsened level of 500, 1000 and 5000 km, whichever the VKT value reported by respondents. In the case of a presumed coarseness level of 500 km for example, the real VKT is supposed to lie in the interval around the reported VKT plus or less 250 km for all the observations. Again, the coefficient estimates of these models do not significantly differ from those of the proposed model. Thus, univariate ordered response probit models may also be applied to investigate VKT under the assumption of a fixed coarseness in the reports, even though they may contain several levels of rounding. The standard deviation of the error term in these univariate models is decreasing as the assumed coarseness level increases.[3] However, assuming the largest coarseness level is not consistent with the collected data because only 30% of the sample has reported a VKT rounded to a multiple of 5000 km. In addition, setting the coarseness level to the smallest value (500 km) yields an estimated variance for the error term similar to the proposed model, but this assumption is unlikely for large values of reported VKT. Indeed, coarseness has been shown to increase with VKT in the proposed model, which turns out to be a better statistical framework in our context than univariate ordered response probit models with predetermined homogenous coarseness.

Conclusions

[2] Despite appearances, this conclusion is not inconsistent with the expected tail-off in car use on the last part of life cycle: while retired drivers over 60 years old do not use their car anymore to commute, the induced decrease in car use is here captured by the "commuting car" dummy in the VKT function.

[3] The estimated standard deviations are 0.642, 0.628 and 0.527 for predetermined coarseness levels of 500 km, 1000 km and 5000 km respectively.

Annual vehicle kilometres travelled is analysed in this study, particularly focusing on the coarseness in the data resulting from the reports by survey respondents. The reports are regarded as heaped where various levels of rounding are included, and VKT itself is assumed to affect the coarseness of the report. Bivariate ordered response probit model is developed to represent the reported VKT and the coarseness of the report simultaneously. The coarseness of the report by each respondent is not perfectly known to analyst, thus the latent class approach is used to represent the probabilities the reported VKT could be rounded as multiples of 5000km, 1000km and 500km. One of the limitations of the proposed analysis is the selection of the levels of the coarseness. In our empirical data, about 23% of the reports are not multiples of 500 km, which means the actual coarseness of these reports are smaller than 500 km. The appropriate selection of the coarseness levels remains as a further research topic.

The model is applied for the French panel data called Parc-Auto, and the results suggest that longer VKT results in a larger coarseness in the report as expected. The results also suggest that a commuting car has a larger coarseness in the report of VKT. Commuting cars are driven on a longer annual distance but they are also much more often used than others in terms of number of trips. Thus, higher car use frequency is possibly another source of approximation in reporting annual VKT. The high-end in-vehicle technology getting popular especially among expensive vehicles might help recording the VKT at any duration and it can support the improvement in the VKT survey. The results suggest that such a technology provides a larger benefit from the view point of data collection when installing into commuting cars than others.

The estimation results on the VKT function suggest that the estimates by the proposed model are not statistically significantly different from conventional regression model with the data set used in the empirical analysis of this study, but the estimated variance of the error term is smaller than conventional model, implying that the proposed model better represents VKT than the conventional regression model does. The results support that the proposed model is superior to conventional regression models, so the estimates might become statistically significantly biased if the conventional models are used for different data set. Thus, further investigations are needed to clarify the advantage of the proposed model.

The extension of the proposed random heaping model by integrating with discrete-continuous models of vehicle type choice and use may be the next step. Although our model treats the heaping appropriately when estimating the VKT function, the indirect effects of explanatory variables through the vehicle type choice cannot be considered. The integration may provide better estimates of the indirect as well as direct effects of particular factors on the vehicle usage.

Another direction of the further analysis is the multiple imputations used in Heitjan and Rubin (1990). They applied multiple imputations to the data with the estimated parameter estimates, and obtained smoother histograms than original sample distribution in the context of children's age. The same imputation technique can be applicable to the data set used in this study, and is expected to provide smoother histograms than original sample distribution of reported VKT.

Nowadays, compulsory periodical inspection of vehicles, which is generalising in Europe, provides a new data source on odometer reading, allowing much larger sample sizes, but with fewer information on the car and on its driver. Moreover new tools are emerging, which allow a much more accurate measurement of distance travelled (e.g. Global Positioning Systems). However, especially for analysing long term trends (e.g. for GHG emissions or infrastructure building) it is crucial to compare actual more precise data with data collected in the past with conventional survey methods (e.g. for the analysis of peak car travel (Grimal et al., 2013; Madre et al., 2012)).

Acknowledgements

The data set used in the empirical analysis of this study was provided by SOFRES, and the

survey was funded by ADEME (Agency for Environment and Energy Savings) and CCFA (Committee of French Car Manufacturers). The authors thank Noritaka Nakagawa who provided computational assistance.

References

Bhat, C.R. (1994a) Estimation of travel demand models with grouped and missing income data, Transportation Research Record, No. 1443, pp. 45-53.

Bhat, C.R. (1994b) Imputing a continuous income variable from grouped and missing income observations, Economics Letters, Vol. 46, pp. 311-319.

Bhat, C.R. and Sen, S. (2006) Household vehicle type holdings and usage: and application of the multiple discrete-continuous extreme value (MDCEV) model, Transportation Research Part B, Vol. 40, pp. 35-53.

Bhat, C.R. and Steed, J.L. (2002) A continuous-time model of departure time choice for urban shopping trips, Transportation Research Part B, Vol. 36, 207-224.

Bhat, C.R., Sen, S. and Eluru, N. (2009) The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use, Transportation Research Part B, Vol. 43, pp. 1-18.

Brownstone, D and Fang, H. (2014) A vehicle ownership and utilization choice model with endogenous residential density, The Journal of Transport and Land Use, Vol. 7, pp. 135-151.

Cirillo, C., Liu, Y. and Tremblay, J.-M. (2016) Simulation, numerical approximation and closed forms for joint discrete continuous models with an application to household vehicle ownership and use, Transportation, Online First. doi:10.1007/s11116-016-9696-4

Commission des Comptes des Transports de la Nation (2013), 50ème rapport: les comptes des transports en 2012.

Eluru, N., Bhat, C.R., Pendyala, R.M. and Konduri, K.C. (2010) A joint flexible econometric model system of household residential location and vehicle fleet composition/usage choices, Transportation, Vol. 37, pp. 603-626.

European Commission (2011), White paper: Roadmap to a Single European Transport Area - Towards a competitive and resource efficient transport system.

Fang, H.A. (2008) A discrete-continuous model of households' vehicle choice and usage, with an application to the effects of residential density, Transportation Research Part B, Vol. 42, pp. 736-758.

Grimal, R., Collet, R. and Madre, J.-L. (2013) Is the stagnation of individual car travel a general phenomenon in France? A time-series analysis by zone of residence and standard of living, Transport Reviews, Vol. 33, No. 3, pp. 291-309.

Heitjan, D.F. and Rubin, D.B. (1990) Inference from coarse data via multiple imputation with application to age heaping, Journal of American Statistical Association, Vol. 85, No. 410, pp. 304-314.

Heitjan, D.F. and Rubin, D.B. (1991) Ignorability and coarse data, The Annals of Statistics, Vol. 19, No. 4, 2244-2253.

Hivert, L. (2000) Le Parc Automobile des Ménages: Étude en fin d'année 1998, à partir de la source "Parc Auto" SOFRES. Report for the French Agency for Environment and Energy Savings (ADEME) (in French).

Kockelman, K. (1997) Travel behavior as a function of accessibility, land use mixing, and land use balance: evidence from the San Francisco Bay Area, Transportation Research Record, No. 1607, pp. 117-125.

Liu, Y. and Cirillo, C. (2016) Small area estimation of vehicle ownership and use, Transportation Research Part D, Vol. 47, pp. 136-148.

Liu, Y., Tremblay, J.-M. and Cirillo, C. (2014) An integrated model for discrete and continuous

decisions with application to vehicle ownership, type and usage choices, Transportation Research Part A, Vol. 69, pp. 315-328.

Madre,J.-L. and Armoogum, J. (1997) Accuracy of data and memory effects in home based surveys on travel behavior, Presented at the 76th Annual Meeting of Transportation Research Board, Washington D.C.

Madre, J.-L. and Armoogum, J. (1998) Weighting or imputations? The example of non-responses for daily trips in the French NPTS, Journal of Transportation and Statistics, Vol. 1, No. 3, pp. 51-63.

Madre, J. L., Bussiere, Y.D., Collet, R. and Villareal, I.T. (2012) Are we heading towards a reversal of the trend for ever-greater mobility? Discussion Papers 2012-16, International Transport Forum, OECD.

Musti, S. and Kockelman, K.M. (2011) Evolution of the household vehicle fleet: Anticipating fleet composition, PHEV adoption and GHG emissions in Austin, Texas, Transportation Research Part A, Vol. 45, pp. 707-720.

Rietveld, P. (2002) Rounding of arrival and departure times in travel surveys: an interpretation in terms of scheduled activities, Journal of Transportation and Statistics, Vol. 5, No. 1, pp. 71-81.

Spissu, E., Pinjari, A.R., Pendyala, R.M. and Bhat, C.R. (2009) A copula-based joint multinomial discrete-continuous model of vehicle type choice and miles of travel, Transportation, Vol. 36, pp. 403-422.

Stopher, P., FitzGerald, C. and Xu, M. (2007) Assessing the accuracy of the Sydney Household Travel Survey with GPS, Transportation, Vol. 34, pp. 723-741.

Tong C.O. and Lee, J.K.L. (2009) The use of a hazard-based duration model for imputation of missing personal income data, Transportation, Vol. 36, pp. 565-579.

Train, K. (1986) Qualitative Choice Analysis: theory, econometrics, and an application to automobile demand, The MIT Press, Cambridge, MA.

Yamamoto, T., Kitamura, R. and Kohmoto, I. (2001) An analysis of vehicle type choice, allocation and use by households, Journal of Infrastructure Planning and Management, No. 674/IV-51, pp. 63-72 (in Japanese).