A random heaping model of annual vehicle kilometers traveled considering heterogeneous approximation in reporting

Toshiyuki Yamamoto

Nagoya University

Annual vehicle kilometers traveled

VKT (vehicle kilometers traveled)

- has been used as an index of car use
  - The strongest indicator of car dependencies and household's travel patterns

- There have been many studies to make use of VKT for various purposes
  - Gasoline consumption, vehicle emissions, and crashes

Difficulty in modeling VKT

Generally, goodness-of-fit is low

- R^2: 0.11 (Train, 1986), 0.15 (Kockelman, 1997), 0.17 (Yamamoto et al., 2001)

Reason might be

- Variability among household's vehicle use
  - Factors to affect car use are not fully incorporated
- Inaccuracy in observation
  - Annual VKT reported by respondents
  - Short-period odometer readings

Literature review

Variability among household's vehicle use

- Discrete-continuous models of vehicle type and use (Bhat and Sen, 2006; Fang, 2008; Brownstone and Fang, 2009; Bhat et al., 2009) to incorporate interaction with vehicle type choice

Inaccuracy in observation

- Studies on departure and arrival time (Rietvelt, 2002; Bhat and Steed, 2002) and income (Bhat, 1994a, 1994b; Tong and Lee, 2009) assume either uniform distribution or fixed intervals, not applicable to VKT
- Heitjan and Rubin (1990, 1991) for reported children's age, applicable to VKT

Objectives

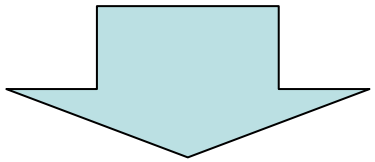
- Inaccuracy in observation is examined
- Annual VKT model is developed considering inaccuracy in observation
  - Efficiency is compared with conventional models
- Heterogeneity among respondents in inaccuracy of observation is also examined

Incomplete data

- *Missing data*: each data value is either perfectly known or entirely unknown
- *Coarse data*: only a subset of the complete-data sample space is observed
  - *Censoring*: in failure time data, if an item has not failed by the time observation ends, failure time is known only to lie beyond the last observation point
  - *Rounding*: data value is observed only to the nearest integer. Also called heaping if items reported with various levels of coarseness

Coarseness in VKT data

- Annual VKT reported by respondents includes some level of approximation
- Level of approximation may vary among respondents



VKT data is regarded as heaped

Methodology (Heitjan and Rubin, 1990, 1991)

- VKT $\ln y_i^* = \beta \mathbf{x}_i + \varepsilon_i$

- Relationship between true VKT, y_i^* and reported VKT, y_i

  y_i^* lies in the range

  - $y_i \pm 250$ if rounded as multiples of 500km
  - $y_i \pm 500$ if rounded as multiples of 1000km
  - $y_i \pm 2500$ if rounded as multiples of 5000km

- Coarseness $z_i^* = \alpha \ln y_i^* + \boldsymbol{\gamma}\mathbf{x}_i + \zeta_i$

$$z_i = 1 \quad \text{if } z_i^* < 0, \qquad \text{500km heaper}$$

$$= 2 \quad \text{if } 0 \le z_i^* < \theta, \qquad \text{1000km heaper}$$

$$= 3 \quad \text{if } \theta \le z_i^* \qquad \text{5000km heaper}$$

- Inclusion of VKT in coarseness function results in bivariate normal distribution

$$E\begin{pmatrix} \ln y_i^* \\ z_i^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}\mathbf{x}_i \\ \alpha\boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\gamma}\mathbf{x}_i \end{pmatrix} \quad V\begin{pmatrix} \ln y_i^* \\ z_i^* \end{pmatrix} = \begin{pmatrix} \sigma_\varepsilon^2 & \alpha\sigma_\varepsilon^2 \\ \alpha\sigma_\varepsilon^2 & \sigma_\zeta^2 + \alpha^2\sigma_\varepsilon^2 \end{pmatrix}$$

- We can define a region of possible values for (y_i^*, z_i^*) at given y_i

$$L_i = [y_i - 250, y_i + 250) \times (-\infty, 0) \text{ for 500km heaper}$$

$$M_i = [y_i - 500, y_i + 500) \times [0, \theta) \text{ for 1000km heaper}$$

$$H_i = [y_i - 2500, y_i + 2500) \times [\theta, \infty) \text{ for 5000km heaper}$$

- Coarseness of each respondent is not known, so

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

$$\begin{aligned} S(y_i) &= L_i \cup M_i \cup H_i && \text{if } y_i = 0 \bmod 5000 \\ &= L_i \cup M_i && \text{if } y_i = 0 \bmod 1000 \text{ and } y_i \neq 0 \bmod 5000 \\ &= L_i && \text{if } y_i = 0 \bmod 500 \text{ and } y_i \neq 0 \bmod 1000 \end{aligned}$$

Parc-Auto

- French households' car ownership panel data
- Conducted yearly since 1976, and continues today
- Sample size is maintained at about 7,000 households each year
- Includes characteristics of up to 3 cars in the household, vehicle use, general attitudes concerning transportation, etc.

VKT data in Parc-Auto

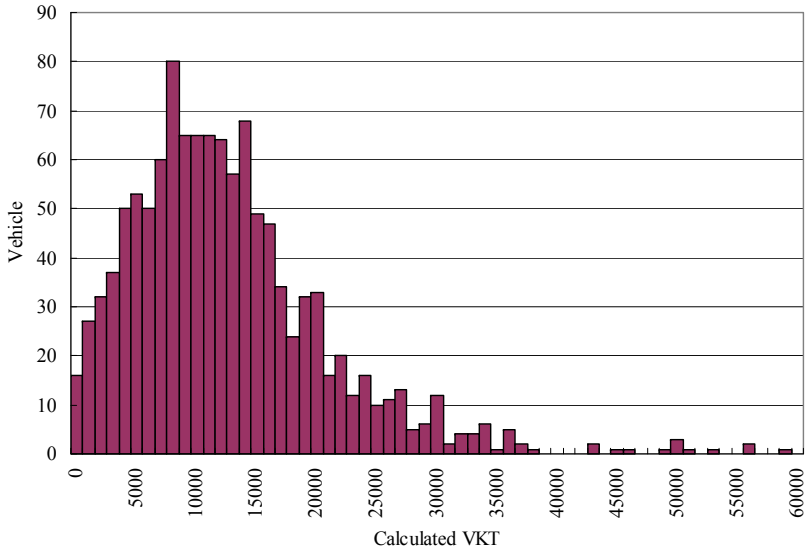
2 types of information

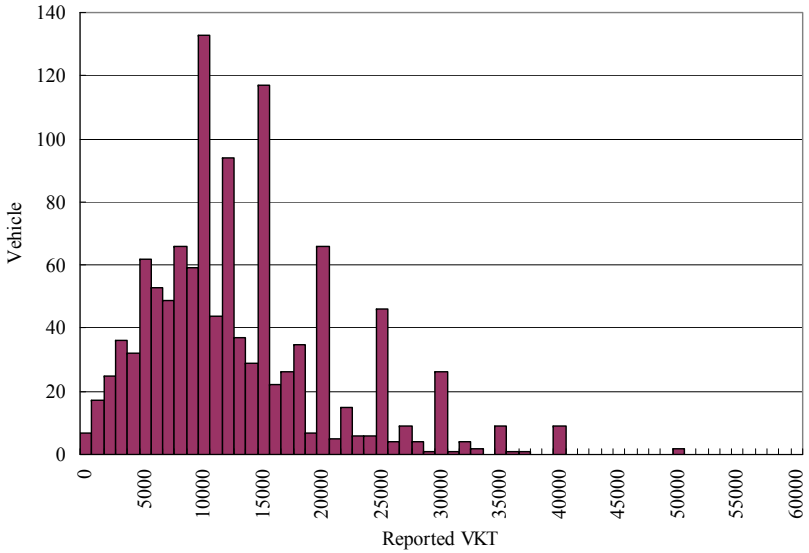
- Difference in odometer readings at 2 successive years -> *Calculated VKT*
- Annual mileage in kilometers reported by respondent -> *Reported VKT*

We use for analysis 1167 sample cases

- 1998 VKT data
- Sub-sample who answered both 1997 & 1998 survey to get *Calculated VKT*

Sample distribution



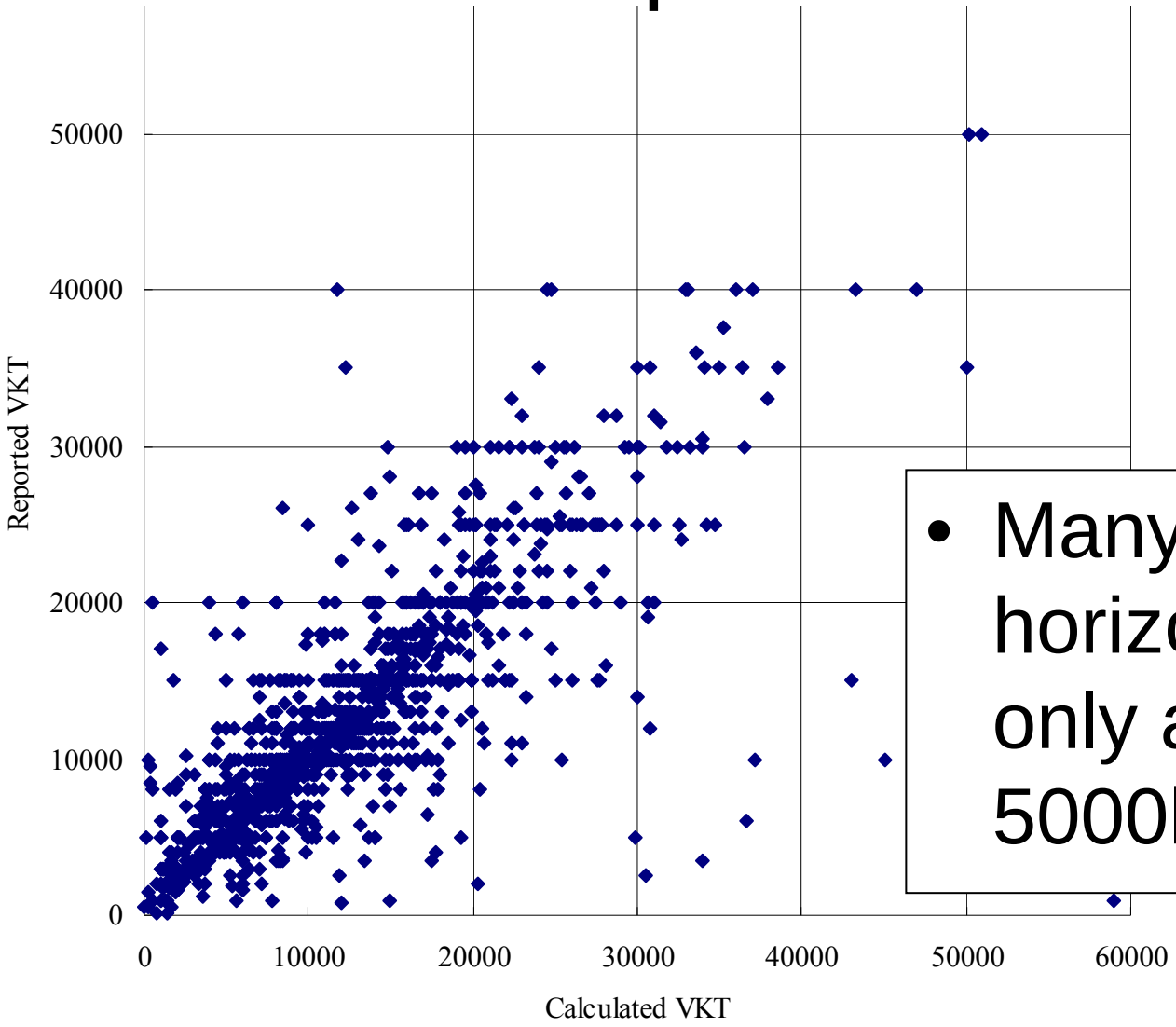


Calculated VKT

Reported VKT

- Reported VKT is obviously rounded at multiples of 5000km

Scatter plots of calculated and reported VKT



- Many plots lie in horizontal lines not only at multiples of 5000km

Rounding of reported VKT

	Cases
Multiples of 5000km	430
Multiples of 1000km (excluding multiples of 5000km)	488
Multiples of 500km (excluding multiples of 1000km)	109
Not multiples of 500km	140
Total	1167

Explanatory variables

- Household’s attribute
  - #children (15-), PT access., large city (300,000+), #cars, low income (F75,000-), high income (F200,000+)
- Personal attribute
  - Young (39-), old (60+), worker, male, car commute
- Car attribute
  - Diesel car, small car, large car, truck, car age

Estimation results

Coarseness function

- Longer VKT results in a larger coarseness
- Larger cars have a larger coarseness
  - Large car owners are not sensitive to fuel use?

VKT function

- Coefficient estimates are not significantly different from conventional regression models
- Estimated variance of the error term is smaller than conventional models

Conclusions

- The proposed model is suggested as superior to conventional models, though coefficient estimates are not different with the data used in this study

- Further investigations are needed to confirm the superiority with different data
- Multiple imputations should be applied to obtain smoother histograms than original sample distribution with the estimated parameters