Bhatta, B. & Larsen, Odd I. (2011). Errors in variables in multinomial choice modeling: a simulation study applied to a multinomial logit model of travel mode choice. *Transport policy, 18*(2), 326-335.

Errors in variables in multinomial choice modeling: a simulation study applied to a multinomial logit model of travel mode choice

"Forthcoming in Transport Policy"

Bharat P. Bhatta[±] and Odd I. Larsen[§]

[±]Sogn og Fjordane University College, P.O. Box 133, N-6851 Sogndal, Norway
Tel: +4757676000, Fax: +4757676301, Email: bharat.bhatta@hisf.no

[§]Molde University College, P.O. Box 2110, N-6402, Molde, Norway

Abstract

Modeling travel demand is a vital part of transportation planning and management. Level of service (LOS) attributes representing the performance of transportation system and characteristics of travelers including their households are major factors determining the travel demand. Information on actual choice and characteristics of travelers is obtained from a travel survey at an individual level. Since accurate measurement of LOS attributes such as travel time and cost components for different travel modes at an individual level is critical, they are normally obtained from network models. The network-based LOS attributes introduce measurement errors to individual trips thereby causing errors in variables problem in a disaggregate model of travel demand. This paper investigates the possible structure and magnitude of biases introduced to the coefficients of a multinomial logit model of travel mode choice due to random measurement errors in two variables, namely, access/egress time for public transport and walking and cycling distance to work. A model was set up that satisfies the standard assumptions of a multinomial logit model. This model was estimated on a data set from a travel survey on the assumption of without measurement errors. Subsequently random measurement errors were introduced and the mean values of the parameters from 200 estimations were presented and compared with the original estimates. The key finding in this paper is that errors in variables result in biased parameter estimates of a multinomial logit model and consequently leading to poor policy decisions if the models having biased parameters are applied in policy and planning purposes. In addition, the paper discusses some potential remedial measures and identifies research topics that deserve a detailed investigation to overcome the problem. The paper therefore significantly contributes to bridge the gap between theory and practice in transport.

Keywords: Level of service attributes; Errors in variables; Travel mode choice; Network models; Bias

1. Introduction

Modeling travel demand is a vital part of transportation planning and management. Level of service (LOS) attributes representing the performance of transportation system and characteristics of travelers including their households are typical factors determining the travel demand. Information on actual choice and characteristics of travelers is obtained from a travel survey at an individual level. Since accurate measurement of LOS attributes such as travel time and cost components for different travel modes at an individual level is critical (McFadden[1], 2000), they are normally obtained from zonal-based network models. A network model provides LOS attributes from the centroid of a zone to the centroid of another zone. But for the modeling purpose, the centroid to centroid LOS attributes are used as a proxy for LOS attributes related to individual trips which actually have origins and/or destinations located outside the centroids (but within the zone they represent). McFadden contends that the values derived from network models can show large and systematic biases which disrupt disaggregate models. The network-based LOS attributes can be, if correct, average values for a group of travelers for trips between different pair of zones rather than exact values for any one traveler (Daly and Ortuzar, 1990). We may treat the differences between the perceived LOS attribute values pertaining to individual trips and the network values as measurement errors and those measurement errors are random. The network-based LOS attributes, therefore, introduce errors in variables[2] (EIV) problem in disaggregate modeling of travel demand. Aggregation of all the trips having origins and/or destinations in a zone at the zone centroid is the key source of the measurement error dealt in this paper because the measurement error is directly related to aggregation error caused by zones and larger zones leads to larger measurement errors (see ibid.).

Errors due to the use of estimated values based on a model rather than accurate measurement can also be present in other settings. Another example from the transportation field is the estimation of traffic noise exposure for housing units based on volume and composition of traffic, distance to highways, topography and other factors. The reason for using model based estimates is that accurate measurement can be very costly or unavailable. In travel demand modeling particularly, the travel surveys do not give direct information on attributes of alternatives not chosen, so LOS attributes for all the alternatives, including those chosen and those not chosen, are normally obtained from the zonal-based network models (McFadden, 2000).

Building on ideas originally proposed by Alonso (1968), Daly and Ortuzar (1990) theoretically and empirically investigate data aggregation in travel demand modeling, different types of errors in modeling and forecasting, and the trade-off between model complexity and data accuracy with focus on the forecasting of mode and destination choice (see also Ortuzar and Willumsen, 2001). But we focus only on the effects of random measurement errors in estimation of a model.

The EIV problem has been extensively studied in regression models. The general conclusion is that EIV results in biased and inconsistent parameter estimates both for the variable measured with errors and the other variables in the model (Greene, 2003). Measures of accounting for the biases have also been investigated and proposed for both linear and nonlinear regression models (see Fuller, 1987; Carroll et al., 2006; Wansbeek and Meijer, 2000). The problem has not received the same attention in multinomial

[1] McFadden (2000) provides a detailed account of the evolution of the state-of-the-art in disaggregate travel demand modeling.

[2] We consistently use the term “errors in variables” to refer to the measurement errors in explanatory variables in a model.

choice modeling in literature. McFadden (1984, p. 1441), for example, comments on this problem: "How to handle measurement error in qualitative response models is an important unsolved problem". In general, this statement still seems valid.

It might be difficult to account for the EIV problem in travel demand models. The variance of measurement errors in network-based LOS attributes can be reduced by reducing the size of zones (and consequently increasing the number of zones) in a study area. The networks can also be coded more accurately. Increasing accuracy in this way comes at a cost, but may sometimes be worthwhile. However, at present we know a little about the trade-offs involved. Accounting for EIVs in model estimation may be another option. Before considering different measures to account for biases due to EIVs in travel demand models, it is probably worthwhile to investigate the direction and possible magnitude of biases due to the EIVs.

This paper has two specific objectives. The first is to empirically verify how EIV in a multinomial logit (MNL) model of mode choice may affect all the estimated coefficients of the model. The second is to explore some of the ways of overcoming the problem. In this way, this paper aims at bridging the gap between theory and practice in transport.

We have organized the remainder of the paper as follows. Section 2 presents a brief review of EIV in regression models. In Section 3, we discuss the sources of errors in network-based LOS attributes. Section 4 presents the data, model and method used in the study. In Section 5, we present and discuss the results. In Section 6, we discuss the potential solutions to overcome the problems due to the network-based LOS attributes. In Section 7, we conclude the paper.

2. A brief review of errors in variables in regression models

We expect that some of the conclusions from EIV in linear and other nonlinear regression models may carry over to multinomial choice models with some possible modifications. At least one conclusion seems intuitively reasonable: the greater the variance of the measurement error, the worse the biases of parameter estimates. We therefore briefly review the effects of EIV and approaches to account for those effects in linear and nonlinear models in this section.

2.1. Effects of errors in variables

We begin with effects of measurement errors in a regressor in a simple linear regression model estimated by the ordinary least squares procedures (see, e.g., Gujarati, 2003; Greene, 2003). We assume a correct model:

$$Y_i = \alpha + \beta X_i^* + \varepsilon_i \tag{1}$$

where ε_i is independent of X_i^* for unbiased and consistent parameter estimates. Suppose instead of observing X_i^*, we observe X_i and we assume classical measurement errors[3] so:

[3] Measurement errors in a variable are said to be classical if they are uncorrelated with the true value of that variable, the true values of other variables in the model and any errors in measuring those variables (Bound et al., 2001). We also assume that the measurement errors are classical throughout the paper.

$$X_i = X_i^{*}+u_i \;\; \Rightarrow X_i^{*} = X_i-u_i \qquad (2)$$

Where u_i denotes measurement errors in the regressor. Therefore, instead of estimating (1), we estimate

$$\begin{aligned} Y_i &= \alpha + \beta(X_i-u_i)+ \varepsilon_i \\ &= \alpha + \beta X_i+(\varepsilon_i-\beta u_i) \\ &= \alpha + \beta X_i+ \acute{\varepsilon}_i \qquad (3) \end{aligned}$$

Where $\acute{\varepsilon}_i = \varepsilon_i-\beta u_i$ is a composite error term consisting of both equation and measurement errors.

$$\begin{aligned} \mathrm{Cov}\ (\acute{\varepsilon}_i, X_i) &= E[\acute{\varepsilon}_i -E(\acute{\varepsilon}_i)][X_i-E(X_i)] \\ &= E(u_i-\beta u_i\ )\ (u_i)\ \text{ assuming } E(u_i) = 0,\ \mathrm{cov}\ (u_i,\varepsilon_i) = 0 \\ &= E(-\beta u^2_i) = -\beta\sigma_u^2 \qquad (4) \end{aligned}$$

where σ_u^2 is variance of $u_{i.}$ Thus, the regressor and the composite error term are correlated violating the important assumption of the classical linear regression model. It can be shown (Gujarati, 2003, p. 527) that:

$$p\lim\hat{\beta} = \beta\left[\frac{1}{1+\sigma_u^2/\sigma_{x*}^2}\right] \qquad (5)$$

where plim $\hat{\beta}$ and σ^2_{x*} denote the probability limit of β and variance of X* respectively. Equation (5) is a simple formula and shows that $\hat{\beta}$ will not converge to true β. It is biased toward zero, called *attenuation*, since the term inside the brackets is less than one. The bias does not disappear even if the sample size increases indefinitely. EIV thus results in a biased and inconsistent parameter estimate. Clearly, the bias becomes worse when the variance of measurement error increases and the ratio of variances in the denominator is the critical factor. It is also clear from equation (5) that we can correct for the bias ex post in this simple case if we have á priori and independent information about the value of σ^2_u and var(X^*).

Expressions for biases in a multiple regression model with a single mismeasured regressor can similarly be derived, but matters get more complicated and the result is less transparent. In a multiple regression model, the coefficient of the mismeasured regressor is still biased toward zero and the other coefficients get biased as well, although in unknown directions (Greene, 2003; Bound et al., 2001). A mismeasured regressor thus can contaminate all the parameter estimates in a multiple regression model. Matters get even worse if more than one regressor in a model are mismeasured. We cannot conclude for sure about the direction and magnitudes of biases, which normally depend on numerous parameters whose signs and magnitudes are unknown and, seemingly unknowable (Greene, 2003).

While analytically tractable expressions for biases have been derived for linear regressions models, this has in general not been the case for a multinomial choice model. Nevertheless, results have been given for some special cases. Yatchew and Griliches (1985) derive expression for bias for the probit model with one mismeasured regressor assuming that all the variables are normally distributed and the measurement error is classical. According to their results, the coefficient is biased toward zero but it is compounded by an additional term compared to the bias in the linear regression model. They do not mention anything about the effects on the other coefficients. Similarly, Kao and Schnell (1987) show that EIVs in MNL model result in asymptotically biased parameter estimates. They do not mention anything about the distribution of variables and type and distribution of measurement errors. They also propose a bias-adjusted estimator. This seems an important contribution in the sparse literature about modeling measurement error in a multinomial choice model. However, probably surprisingly, this approach has never been applied in modeling measurement error in a multinomial choice model in general and an MNL model in particular in any field.

The effects of EIVs in a multinomial choice model can simply be demonstrated as follows (see, e.g., Gujarati, 2003). Assume that the accurate model is U = βX* + ε, where ε is assumed to be independent of X*. If X* is measured with error through the relationship X = X* + u, u ~ N(0, σ_u^2), the resulting utility function becomes U = β(X-u) + ε or U = βX + (ε -βu) = βX + έ where έ = (ε -βu) is a new error term consisting of both equation and measurement errors. Then we can derive:

$$\text{Cov. } (\acute{\varepsilon}, X) = -\beta\sigma_u^2 \qquad (6)$$

This shows that the error terms are no longer independent of explanatory variables in the model, violating the important assumption required for unbiased and consistent estimates.

More specifically we can look at the implication for maximization of a likelihood function. The likelihood function (L) for a general multinomial choice model is (cf. Ben-Akiva and Lerman, 1985):

$$L = \prod_{n=1}^{N} \prod_{i \in C_n} P_{in}^{y_{in}} \qquad (7a)$$

$$\text{where } P_{in} = \frac{e^{\beta' x_{in}}}{\sum_{j \in C_n} e^{\beta' x_{jn}}}$$

and y_{in} = 1 if the decision maker n chooses the alternative i, 0 otherwise. Taking the logarithm of equation (7a) gives the log-likelihood function as follows:

$$LL = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \left(\beta' x_{in} - \ln \sum_{j \in C_n} e^{\beta' x_{jn}} \right) \qquad (7b)$$

Differentiating equation (7b) with respect to β_k (the coefficient of a variable containing random measurement error) and setting the expression equal to zero gives the first order condition for maximization of the equation (7b).

$$\frac{\partial(LL)}{\partial\hat{\beta}_k}=\sum_{n=1}^{N}\sum_{i\in C_n} y_{in}\left(x_{ink} - \frac{\sum_{j\in C_n} e^{\beta' x_{jn}} x*_{jnk}}{\sum_{j\in C_n} e^{\beta' x_{jn}}}\right)$$

$$or \qquad (8)$$

$$\sum_{n=1}^{N}\sum_{i\in C_n}[\ y_{in}-P_{in}\ ]x*_{ink}=0$$

Here P_{in} is the probability calculated at the parameters that maximize the likelihood function. If $x*_{ink}$ is measured with error through the relation $x_{ink} = x*_{ink} + u_{ink,}$ the first order condition will actually be:

$$\sum_{n=1}^{N}\sum_{i\in C_n}[\ y_{in}-P_{in}\ ](x_{ink}-u_{ink})=0$$

$$or \qquad (9)$$

$$\sum_{n=1}^{N}\sum_{i\in C_n}[\ y_{in}-P_{in}\ ]x_{ink}-\sum_{n=1}^{N}\sum_{i\in C_n}[\ y_{in}-P_{in}\ ]u_{ink}=0$$

The first order condition for the accurately measured variable is:

$$\sum_{n=1}^{N}\sum_{i\in C_n}[\ y_{in}-P_{in}^{*}]x_{ink}^{*}=0 \qquad (10)$$

Where P_{in}^{*} is the probability calculated at the optimum parameter values estimated with the correctly measured variables. Deducting equation (10) from equation (9) we get:

$$\sum_{n=1}^{N}\sum_{i\in C_n}\left[P_{in}^{*}-P_{in}\right]x_{ink}^{*}+\sum_{n=1}^{N}\sum_{i\in C_n}[\ y_{in}-P_{in}]u_{ink}=0 \qquad (11)$$

The problem is caused by the second term.

While $E\sum_{n}\sum_{i\in C_n} y_{in}u_{ink}=0$ we will in general have

$$E\sum_{n}\sum_{i\in C_n} P_{in}u_{ink}\neq 0\ and\ P_{in}^{*}\neq P_{in}\quad \forall i,n \qquad (12)$$

If the true parameter is positive, a positive value for u_{ink} will tend to increase P_{in} and conversely if the true parameter is negative. Due to the fact that probabilities at the optimum will be different, equations (9) and (11) together with (12) imply that the first order condition for all parameters will be more or less affected.

2.2. Accounting for errors in variables

Though the EIVs problem in regression models, particularly in linear regression models, is not new, it has recently attracted the increased attention of researchers. The efforts are largely about accounting for biases in parameter estimates. There are many approaches, for example, instrumental variables estimation, regression calibration, simulation extrapolation, likelihood approach[4], asymptotically corrected likelihood criterion (Li and Hsiao, 2004), finite bounds (Hu, 2006) and so on considered in contemporary literature to take into account for the biases due to EIVs in linear and nonlinear regression models.

Readers are referred to Fuller (1987) for linear models, Carroll et al. (2006) for nonlinear models, and Wansbeek and Meijer (2000) for both types of models for detailed study of these approaches. Buzas et al. (2003) also thoroughly discuss different types of measurement errors, their consequences, and solution methods in their recent work. Bound et al. (2001) contend that accounting for biases due to EIVs is more difficult in nonlinear models than in linear models. It might be even more difficult to account for biases in estimation of a multinomial choice model, a major class of nonlinear models. Consequently, far less attention has been paid to modeling EIVs in multinomial choice modeling. Brownstone (2001) uses multiple imputation to account for the biases due to EIVs introduced by the use of network-based LOS attributes when a small validation study is available to model the measurement error process. Recently, Walker et al. (2008) suggest a latent variable approach to correct for measurement errors introduced in a travel demand model (cf. Section 6).

3. Sources of errors due to network-based LOS attributes

Input data of transportation models may contain different types of errors. The network data of LOS attributes is perhaps the most crucial source in disaggregate modeling of travel demand. In this section, we summarize sources of errors relevant to the investigation in this paper (cf. Bhatta, 2009 for detailed account of sources of errors). Errors in network LOS attributes may stem from:

- Digitizing and coding of highways and transit networks and transit schedules. For example, the coded network generally does not contain all relevant links of the real network or attributes of the coded links may contain errors. Variations in transit services over a day can also be difficult to capture in coding of routes.
- The algorithms and assumptions used in the network model for ‘shortest path’ assignments. For example, travelers may consider other or additional attributes of travel paths than those included in the model.
- The use of centroids and centroid connectors and the degree of geographical dispersion of trip origins and destinations within a zone.

[4] The instrumental variables estimation, regression calibration, simulation extrapolation, likelihood approach and so on are discussed and reviewed in detail in Caroll et al. (2006). Additionally, Fuller (1987), Bound et al. (2001), and Wansbeek and Meijer (2000) are the excellent references for EIV in regression.

- Link and intersection delays. Most traffic assignment algorithms assume that delay occurs only on the links but not at intersections. Speeds assumed for links may not adequately account for delays at intersections.
- Inaccuracies in the assumed volume-delay functions, or the origin destination-matrices used, when travel times are taken from assignment in congested networks.
- Omission of external and intrazonal trips because these trips, particularly intrazonal trips, do not appear on networks in a centroid-to-centroid travel and are not always considered in model estimation (see Bhatta and Larsen, 2010).

For modeling purposes, all trips are assumed to begin and/or end at the centroid of each transportation analysis zone (TAZ). At best, the centroids can be placed at "the center of activities" in each TAZ and values for LOS attributes estimated by network models yield the "mean values" for LOS attributes related to travel between zones. The term "mean value" must here be interpreted as a value we get if we average the 'correct' values for all the potential trips between zones. The TAZ might have high variation internally indicating that centroids alone can't reflect it. Since a TAZ does not have consistent geographic size, the TAZ is subject to modifiable areal unit problem (MAUP) like other spatial studies (Zhang and Kukadia, 2005; Chang et al., 2002). Besides, single estimate of LOS attributes per mode for an origin-destination pair may mask meaningful variation of LOS attributes among the travelers. Our concern in this paper is therefore errors due to the use of network LOS attributes to estimate disaggregate demand models. These LOS attributes are average (or aggregate) estimates for all potential trips between zones. When applied for estimation of disaggregate choice models, these zonal averages introduce errors in the LOS attributes for individual trips (Brownstone, 2001). The errors are random by their very nature but their variance may depend on the geographical size of zones and the dispersion of population and activities within the zones.

Over the years, network LOS attributes are getting more correct due to more accurate and detailed coding of networks and improved network models generally. Sizes of zones are also getting smaller implying that the random measurement errors due to the use of zone centroids as origin and destination of all trips are also getting smaller. The number of zones in the transport model for the Oslo-region has, for example, increased from 430 in 1992 to 1940 zones at present. The first model developed in the mid 1960s had less than 100 zones for the same area in Norway. However, even if network LOS attributes are getting more accurate, there are at least two variables, namely, length for walking and cycling trips and access/egress time to get to public transport, still contain, in our opinion, relatively large measurement errors. Walking and cycling tend to be short trips. Consequently, the calculated distance between centroids used for model estimation may introduce relatively large errors in actual individual trips between origin and destination. Similarly, access/egress time also tends to involve short distance. In addition to the inaccuracies due to centroids and centroid connectors, network access/egress time will also be affected by the coding of public transport stops. In the Norwegian models, for example, bus stops are at present usually located at regular nodes in road networks. The other reason for considering these variables is the ratio of the variance of errors to the variance of the true variable because equation (5) indicates that magnitude of bias depends on the ratio rather than on absolute value of the measurement errors. The variances of the measurement errors for these variables are bigger compared to the error variances for other variables in a model.

Both variables can be more precisely estimated by using smaller zones and/or by calculating distances on an address to address basis, and by coding public transport stops more accurately. This is a costly exercise especially in large travel surveys. One objective of our work is to investigate the possible gains for model estimation of such an improved accuracy.

Generally, we can assume a combination of random and systematic measurement errors in a variable, particularly in a network LOS attribute. A linear variant, for example, could be:

$$X_i = aX_i^* + b + u_i \tag{13}$$

where X_i is the variable used in model estimation, X_i^* is the true variable, u_i is random variable, and a and b are constants. The consequences of the systematic part of the measurement error will depend on whether X is among variables that share a common parameter (e.g., cost of travel in our model) or has a separate parameter. In the first case, it will lead to a systematic bias in the generic parameter, but in the second case the systematic effect amounts to a scaling of the variable and the problem is minor as long as a and b remain the same in applications of the model. Thus, if we suspect that a variable has an error structure according to equation (13), avoiding the generic parameter for the variable could help.

4. Data, model and method

This paper uses data from the Survey for Transport in the Course of Work (PIA) and LOS attributes obtained from a zonal-based network model. The survey randomly selected 2654 employees working in Oslo and Akershus region, Norway, in 1996. The aim of the travel survey was to explore the factors that explain the extent of transport for work trips, modes of travel, trip length, travel costs, travel routes, and so on. The survey contains information on individual travel behavior; traveler and household characteristics including car availability and possession of a driving license, origin and destination location, mode choice, parking possibilities, and so on. A detailed description of the travel survey can be found in Stangeby (1997). The possible modes for the population for work trips in the study area consisted of six travel models, namely, walking (WK), cycling (CK), car driving (CD), car passenger (CP), public transport (PT) and taxi (TX). The PIA data set is a typical data set that is generally used to estimate a mode choice model. Besides, the sample truly represents the population under study.

We performed several screening and consistency checks of the data set. As part of this screening process, we lost some observations that had unknown destinations and missing values on relevant variables. Since we did not consider intrazonal trips in model estimation, we also lost some observations having intrazone destination.

Table 1

Variable (including ASCs) definitions

Variable	Definition
CP_00	ASC for CP
TX_00	ASC for TX alternative
PT_00	ASC for PT alternative
WK_00	ASC for WK alternative
CK_00	ASC for CK alternative
GC_time	Generic travel time by different travel modes
GA_cost	Generic travel cost by different travel modes
CD_v2w	1 if the trip involves a visit on the way/back to work/home, 0 otherwise (s.t. CD)
CD_azone	1 if the traveler lives in a location with difficult parking but good PT services to get to work, 0 otherwise (s.t. CD)
CD_parking	1 if guaranteed free parking possibility at work , 0 otherwise (s.t. CD)
CD_carinb	1 if the car was used for business purposes, 0 otherwise (s.t. CD)
CP_female	1 if the traveler is female, 0 otherwise
CP_elderly	1 if the traveler is <50 years, 0 otherwise
PT_invht	Onboard time with PT (in minutes)
PT_wktm	Access/egress time to get to PT (in minutes)
PT_wait	Waiting time for PT (in minutes)
PT_transf	Number of transfers to get to the work place with PT
PT_freepark	Free parking at work, 0 otherwise
WK_dist	Walking distance to get to work
CK_dist	Cycling distance to get to work
CK_female	1 if female, 0 otherwise
CK_rzone	1 if the travelers lives in a location with mixed land use, 0 otherwise

Since MNL is tractable, by far the easiest and most widely used model of multinomial choice models (Train, 2003; Ben-Akiva and Lerman, 1985), we used the MNL model[5] of travel mode choice for work trips for this purpose. A similar model has previously been estimated on the same data set to investigate the factors influencing the mode choice for work trips (Rekdal, 1999). We coded the model, the analytic gradients and Hessian of the log-likelihood function in the GAUSS programming language. We used the Maxlik procedure available for GAUSS to estimate the model.

The factors influencing or correlated to the travel mode choice are classified into three categories (see, e.g., Ortuzar and Willumsen, 2001). The first are characteristics of the journey. We consider a round trip to get to work and back home. Some trips may have secondary destinations such as dropping kid/s to kindergarten or school on the way to work and picking them up and/or shopping on the way back home.

The characteristics of the traveler including the household, for example, age, gender, possession of a driving license, car availability, garage at home, accessibility status of different travel modes in the respondents' location, parking possibilities at work, and so on are the second category of factors. Possessing a driving license, car availability and garage at home determine the availability of CD for the respondents while the other

[5] In addition, most of the large scale real applications of multinomial choice models are still rely on the use of MNL models. We expect that other studies will follow up using nested and mixed logit models later on.

variables were included in the utility functions of relevant travel modes. Though income is an important variable affecting/correlated to travel mode choice, it was not used in the model due to inaccuracy of the variable.

Table 2
Descriptive statistics of LOS attributes (excluding unavailable alternatives)

LOS attribute	Mean	Std dev	Minimum	Maximum
Car travel time (hours)	0.4628	0.3358	0.0833	1.9567
Car travel cost (‘00NOK)	0.1565	0.1569	0.0000[1)]	1.1770
Travel cost by taxi (‘00NOK)	1.5717	1.1563	0.4010	9.1704
Travel time by taxi (hours)	0.5831	0.3022	0.2417	1.9277
Onboard time by PT (hours)	0.4266	0.3116	0.0167	2.3067
PT fare (‘00NOK)	0.1708	0.0493	0.1120	0.3210
Access/egress time for PT (hours)	0.3053	0.1635	0.0198	0.9910
Waiting time for PT (hours)	0.1115	0.0983	0.0113	0.7500
Number of transfers for PT	0.4616	0.5693	0.0000	3.0000
Walking distance (km $\leq$ 10)	5.2580	2.5721	0.4000	9.9500
Cycling distance (km $\leq$ 30)	10.6774	7.1919	0.4000	30.0000

1) Cost is assumed to be zero if the respondent uses a company car

The third category of factors relate to performance of the transportation system, measured by LOS attributes of the different travel modes. As we mentioned earlier, LOS attributes are obtained from network models. Travel time and travel cost by different travel modes such as car driving (CD), shared ride (CP), PT, TX, and walking and cycling distance to work for WK and CK modes are the major factors representing the performance of a transportation system. Travel time by PT is decomposed into three components, namely, access/egress time, onboard time, and waiting time. Since the use of PT may also involve changing the vehicle (or mode) to get to the final destination, we also use number of transfers for PT. Walking and cycling distances are obtained using the network driving distance between origin and destination. We arrived at the final specification based on the systematic process of model building. Table 1 presents the variables used in the final specification of the model including the alternative specific constant (ASCs) and their definitions. Further, Table 2 summarizes the descriptive statistics of LOS attributes used in the model.

Table 3
Simulated choices

Modes	Trips	Per cent
Car driver	1033	48.7
Car passenger	125	5.9
Taxi	6	0.3
Public transport	452	21.3
Walk	147	6.9
Cycle	358	16.9
Total	2121	100.0

As our purpose is to investigate the impacts of EIVs on the estimated parameters of the MNL model for choice of travel mode, we set up a model that fulfills the assumption of MNL model by construction, i.e., random terms in the utility functions are independently and identically Gumbel distributed (i.i.d. Gumbel). We maintained the original model specification with respect to utility functions, variables and criteria for availability of alternatives for all the simulations.

We estimated the MNL model on the observed choices and socioeconomic characteristics of PIA and network-based LOS attributes. We assumed a set of "true" parameters (column 2nd in Table 4) based on those estimated parameters and judgment of expected signs and relative magnitudes of the coefficients according to theory and previous studies. Consequently the "true" parameters are representative and the results can be generalized. Most of the coefficients were quite similar to the parameters originally estimated. The values of the deterministic utility were calculated with the original variables and the "true" parameters for each alternative and for each traveler. Random variables with Gumbel (0, 1) distribution were drawn for each alternative and for each traveler. These random variables were then added to the values of the deterministic utility. The choice was then taken as the alternative with the highest utility (including the random utility) provided that the alternative was available. If the alternative with the highest value of the utility function was unavailable, the alternative with the second highest value was chosen and so forth. Table 3 shows the simulated choices that were used for all the subsequent estimations. By this procedure we obtained a data set based on real observations, but with choices that were generated in order to fulfill the assumptions of the MNL model. This is the starting point for simulation. Random measurement errors were subsequently introduced to the two variables that we considered most prone to the EIVs problem, namely, walking and cycling distance to work and access/egress time.

Table 4 presents the estimated coefficients of the base model, i.e., model without the measurement errors. The first two letters indicate the utility function. The first five are alternative specific constants. GC_time is the generic coefficient for travel time. GC_cost is the generic coefficient of travel cost. As expected the estimated parameters deviate from the "true" parameters due to the sampling error. The mean value of coefficients estimated for 100 draws for the Gumbel variable came very close to the "true" parameters and confirmed that the model behaves as predicted by theory.

5. Simulation of measurement errors

To investigate the impact of EIVs on the estimated parameters, we introduced random measurement errors with varying variances and distributions to access/egress time (PT_wktm) and walking and cycling distance (WC_dist) to work. The results presented in subsequent sections are the average values of the coefficients based on 200 estimations of the model, each with different random draws for the "measurement error". We have only presented the selected coefficients in subsequent sections. In succeeding sections, we will compare the simulation results with the base case in Table 4.

Table 4
Base case – estimation results without measurement errors

Variable	"True" parameters	Est. (base)	Std. err.	t-stat.
CP_00	-3.20	-3.1676	0.3355	-9.443
TX_00	-1.40	-1.6177	0.6297	-2.569
PT_00	1.30	1.6176	0.2934	5.514
WK_00	2.10	2.5498	0.3269	7.799
CK_00	0.70	0.8978	0.2658	3.378
GC_time	-2.50	-2.5158	0.4613	-5.454
GA_cost	-2.60	-2.5299	0.6666	-3.795
CD_v2w	1.20	1.0798	0.1669	6.471
CD_azone	-0.90	-0.9564	0.1955	-4.892
CD_parking	1.20	1.3031	0.1947	6.694
CD_carinb	1.30	1.3897	0.2021	6.876
CP_female	1.30	1.3016	0.2629	4.951
CP_elderly	0.70	0.2952	0.3301	0.895
PT_invht	-2.00	-1.9442	0.5132	-3.788
PT_wktm	-3.50	-4.0987	0.6826	-6.005
PT_wait	-3.50	-4.5446	1.3809	-3.291
PT_transf	-0.70	-0.7443	0.1732	-4.297
PT_freepark	-1.00	-1.2506	0.1564	-7.999
WK_dist	-0.80	-0.8752	0.0743	-11.785
CK_dist	-0.20	-0.2163	0.0171	-12.633
CK_female	-1.00	-1.0247	0.1405	-7.292
CK_rzone	0.50	0.5373	0.1405	3.824
Final log-likelihood = -1445.2		Number of observations =2035		

5.1. Measurement errors in access/egress time

Measurement errors can have different error structures. The standard results (e.g., equations (1)-(5) in Section 2.1) for linear regression models are based on additive errors with normal distribution. We also simulated additive errors having normal distribution to access/egress time in this section. Estimation results of selected coefficients are reported in Table 5. The fourth and the last column show the mean value of 200 estimations divided by the coefficient values of the base case (without measurement error) in Table 4. We take this as a measure of bias even though the estimates for the base case were affected by the sampling error.

First, we assumed an additive error having normal distribution with a variance corresponding to 10% of the variance of access/egress time. The variance of access/egress time is 0.0258 (the square of the standard deviation in Table 2) and the error consequently gets a variance of 0.00258 and a standard deviation of 0.0508. Notice that the coefficient in a simple linear regression model would get a downward bias of 9% (1/1.1=0.909) for this error structure according to equation (5). Like in the linear regression model, we see that the variable with measurement error (PT_wktm) was biased toward zero and the mean value of the parameter was 83.3% of the base value, i.e., a downward bias of 16.7%. Both the coefficients of car driving time

(GC_time) and on board time for public transport (PT_invht) got a downward bias of approximately 5%, while waiting time for public transport got an upward bias of 6%. The coefficient of travel cost (GA_cost) was hardly affected implying that the implied value of riding time both for car and public transport got biased downward by approximately 5%. The implicit weight of waiting time for public transport got an upward bias of 11% (1.061/0.954). We also notice that the simulated result for PT_wktm was much closer to the true parameter than the base case.

Table 5
Simulation results with normally distributed errors in access/egress time (Selected coefficients)

	Standard deviation = 0.0508			Standard deviation = 0.0803		
Variable	Mean	Std. Dev.	Mean/base	Mean	Std. Dev.	Mean/base
GC_time	-2.3775	0.0554	0.945	-2.2499	0.0746	0.894
GA_cost	-2.5484	0.0161	1.007	-2.5632	0.0215	1.013
PT_invht	-1.8551	0.0415	0.954	-1.7676	0.0610	0.909
PT_wktm	-3.4135	0.2372	0.833	-2.7926	0.3149	0.681
PT_wait	-4.8208	0.1189	1.061	-5.0966	0.1674	1.122
PT_transf	-0.7658	0.0126	1.029	-0.7905	0.0186	1.062
WK_dist	-0.8696	0.0025	0.994	-0.8640	0.0034	0.987
CK_dist	-0.2114	0.0020	0.977	-0.2068	0.0027	0.956

Second, we increased the variance of the error to 25% of the variance of the variable. This gives a standard deviation of 0.0803. A normal distribution for the error may cause some negative values for the recalculated access/egress time and also some of the recalculated values may exceed the limit set for availability (one hour). To avoid negative values, any negative value was replaced by 0.005 and availability was set according to the recalculated values. The bias increased to 32%. The pattern with respect to the other coefficients affected and the direction of the bias persisted, but the magnitude of the biases increased. For this particular model, it seems that the bias introduced by an additive measurement error was at least as severe as we could expect from the formula derived for simple linear regression. The coefficients of onboard time for public transport and car driving time got downward bias, while that of waiting time for public transport got an upward bias. Consequently we got a downward bias in the implied value of time.

Are 200 simulations sufficient to draw firm conclusion? Figure 1 shows the development of the mean value of PT_wktm for two runs of 200 estimations with a standard deviation of 0.0508. The figure shows that the mean value for both runs tends to converge to approximately 3.415 with 200 estimations. We thus conclude that 200 estimations should be sufficient for our purpose and used 200 estimations in all subsequent simulations.

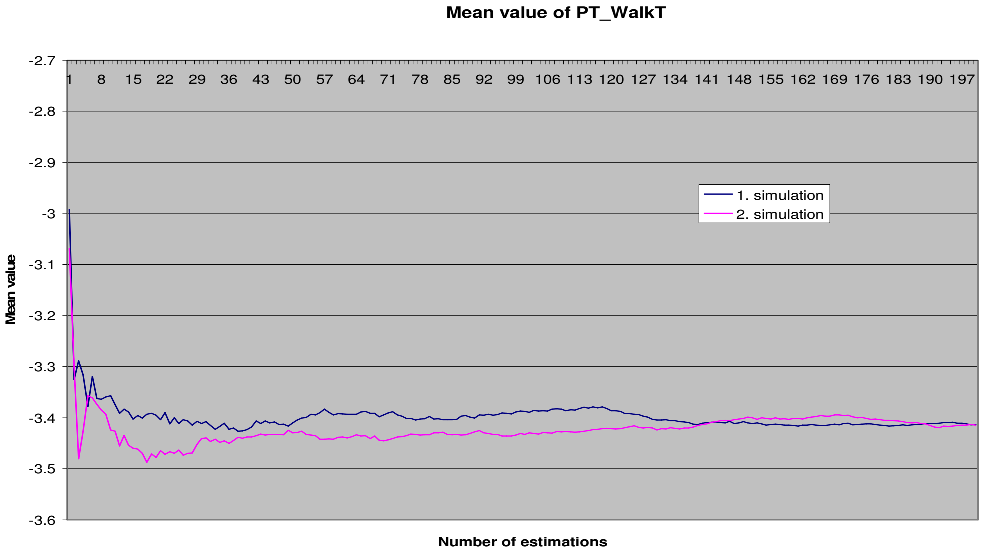

Fig. 1. Development of the mean value with the number of estimations

A LOS variable may not contain an additive normal and homoscedastic error in many situations. We tested the impacts of errors having multiplicative and triangular distributions among possible alternatives.

Table 6 summarizes the results. As can be seen, the two alternatives gave very similar results with respect to the pattern and magnitude of the biases. The same coefficients got most severely affected as in additive normal distribution with homoscedastic errors, but the biases were smaller. This may indicate that the consequences of heteroscedastic errors with variance increasing with the true value of the variable are less severe than homoscedastic errors producing the same variance for the error.

Table 6
Simulation results with log-normally and triangularly distributed error in access/egress time (Selected coefficients)

	Lognormal σ=0.2027			Triangular distribution		
Variable	Mean	Std. Dev.	Mean/base	Mean	Std. Dev.	Mean/base
GC_time	-2.2420	0.0734	0.938	-2.3658	0.0650	0.940
GA_cost	-2.5647	0.0225	1.004	-2.5421	0.0244	1.005
PT_invht	-1.7615	0.0599	0.944	-1.8460	0.0531	0.950
PT_wktm	-2.7554	0.3130	0.782	-3.3102	0.2514	0.808
PT_wait	-5.1148	0.1672	1.097	-4.9107	0.1461	1.081
PT_transf	-0.7907	0.0170	1.047	-0.7720	0.0160	1.037
WK_dist	-0.8639	0.0034	0.992	-0.8687	0.0025	0.993
CK_dist	-0.2066	0.0027	0.972	-0.2108	0.0021	0.974

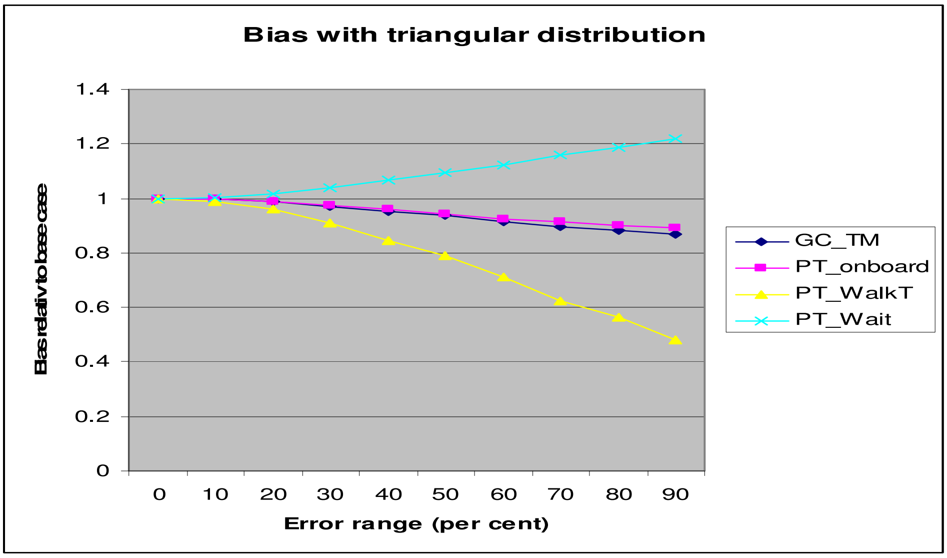

Fig. 2. Bias as a function of error range

We also raised the magnitude of errors with the triangular distribution in an interval of 10% of the true value of access/egress time. Figure 2 shows the results. It seems to be a non-linearity involved as the lines get steeper when the error range exceeds 20 – 30% of the true value. The biases are rather insignificant for this error structure when the range of the error stays below ±20-30% of the true variable

In our opinion, the "measurement errors" caused by the use of centroids and centroid connectors in conjunction with inaccurate coding of public transport stops can easily involve an error range that exceeds ± 50% of the true value of access/egress time (see also Talvitie and Dehghani, 1980).

5.2. Measurement errors in walking and cycling distance to work

While access/egress time deals with error in one component of a journey by one mode, walking and cycling distance to work deals with a variable that is relevant for the whole trip. As walking and cycling tend to follow the same paths in the coded networks, the measurement errors are highly correlated (perfectly correlated in our case). However, the maximum distance for availability is different for the two modes.

We began with an additive normal error. First, we assumed that the variance of the errors was 10% of the sample variance of walking distance. The variance of walking distance is 6.62 so we used 0.662 for variance and a standard deviation of 0.8134 for the errors. As expected both WK_dist and CK_dist got biased toward zero (Table 7). The bias for WK_dist is greater than the expected bias in linear regression with a similar error. GC_time and PT_invht also got severely biased. The implicit values of in-vehicle time got biased downward while the implicit weights on access/egress time and wait time of public transport got biased upward.

Table 7
Simulation results with normally distributed errors in walking/cycling distance (Selected coefficients)

	Additive normal σ=0.8134			Additive normal σ=1.286		
Variable	Mean	Std. Dev.	Mean/base	Mean	Std. Dev.	Mean/base
GC_time	-2.2838	0.0545	0.908	-1.9993	0.0793	0.795
GA_cost	-2.4532	0.0256	0.970	-2.3607	0.0346	0.933
PT_invht	-1.7075	0.0590	0.878	-1.4199	0.0893	0.730
PT_wktm	-3.9208	0.0603	0.957	-3.7151	0.0785	0.906
PT_wait	-4.3769	0.0732	0.963	-4.1518	0.0982	0.914
PT_transf	-0.7467	0.0099	1.003	-0.7480	0.0150	1.005
WK_dist	-0.7343	0.0274	0.839	-0.6160	0.0307	0.704
CK_dist	-0.2048	0.0035	0.947	-0.1905	0.0053	0.881

We also simulated errors with the variance of 25% of the variance of walking distance. Now the standard deviation of the errors is 1.286. If the recalculated variable initially became negative, the distance was set to 0.1 km. Availability was also set according to the value of the recalculated variable. The results of 200 simulations are summarized in Table 7. We see that GC_time and PT_invt got a strong downward bias in addition to a downward bias (in absolute value) of CK_dist and WK_dist. PT_transf however did not get a downward bias.

Last, we simulated additive heteroscedastic errors with the symmetric triangular distribution. If the true distance is 5 km or less, the range of the error term is ±2/3 of the distance. If the true distance is more than 5 km, the range is set at ± 3.33 km. Thus for more than 5 km distance, the error has a constant variance. This choice of error structure produces a slightly smaller error variance than for N(0, 1.286). Table 8 shows the same pattern as in Table 7 with rather strong impacts also on GC_time and PT_invht. One interesting point that emerges from both tables is that the measurement error in distance for walking and cycling tends to give a downward bias in the implicit values of in-vehicle time and strongest for public transport

Table 8
Simulation results with triangularly distributed errors in walking/cycling distance (Selected coefficients)

Variable	Mean	Std. Dev.	Mean/base
GC_time	-2.0975	0.0773	0.834
GA_cost	-2.3947	0.0353	0.947
PT_invht	-1.5111	0.0867	0.777
PT_wktm	-3.7901	0.0727	0.925
PT_wait	-4.2357	0.0989	0.932
PT_transf	-0.7471	0.0156	1.004
WK_dist	-0.7219	0.0313	0.825
CK_dist	-0.1948	0.0051	0.900

5.3. Errors in both variables

In most models estimated on network LOS-variables, both variables are prone to errors. We therefore introduced errors simultaneously to both variables and used the assumption of additive normal with the lowest standard deviation. One interesting result was that the bias in the coefficients of PT_wktm and WK_dist were smaller than in Tables 6 and 7 respectively, but the bias got quite large for GC_time and PT_invht (Table 9). Compared to the base model, the implicit value of in-vehicle time of car got a downward bias of 30%t while the implicit value of public transport on-board time got a downward bias of 74%! Thus when we combine rather moderate measurement errors in access/egress time and walking and cycling distance to work, it seemingly caused large biases to coefficients of in-vehicle time and implicit values of time, most notably for public transport.

Table 9
Simulation results with normally distributed errors in both variables (Selected coefficients)

Variable	Mean	Std. Dev.	Mean/base
GC_time	-1.7362	0.0826	0.6901
GA_cost	-2.4962	0.0324	0.9867
PT_invht	-0.6962	0.0786	0.3581
PT_wktm	-3.7206	0.2325	0.9077
PT_wait	-3.3416	0.1353	0.7353
PT_transf	-0.8344	0.0189	1.1211
WK_dist	-0.7606	0.0295	0.8691
CK_dist	-0.1947	0.0040	0.9001

6. Potential remedial measures

The above results in the paper clearly demonstrate that EIV result in biased parameter estimates of an MNL model and consequently leading to poor policy decisions if the models having biased parameters are applied for policy and planning purposes. It is therefore important to develop methods that take into account the errors in modeling or to find the ways of obtaining disaggregate LOS attributes. We therefore briefly discuss some of the ways of overcoming the problem in this section[6].

Alonso (1968) investigates the implications of imperfect data on modeling and prediction. His investigation is not related to transportation but his conclusions are generally applicable to all fields including transportation using statistical analysis and modeling. Based on simple numerical exercises, he generalizes a few rules of thumb for model building as follows: (i) avoid inter-correlated variables, (ii) add if possible, (iii) multiply or divide if addition is not applicable, and (iv) avoid taking differences or raising variables to powers as far as possible. He concludes in general that it is the correlation of input variables that causes large errors in outcome variables so he suggests avoiding the correlated variables. Most of the LOS attributes used in travel demand modeling are highly correlated. Given Alonso's prescription, we could

[6] Each issue below deserves a separate research paper (possibly several) in its own so we only briefly mention the methods here in this section. We expect that subsequent studies will follow later.

somewhat reduce the output errors if we could exclude the correlated variables in the model. He also suggests using simpler models if the input data are not accurate enough. Given Alonso's thesis, formulation of a model may also help minimize the output errors. Unfortunately, we cannot exclude cost and time, which are highly correlated variables, to estimate a travel demand model. Later Daly and Ortuzar (1990) and Ortuzar and Willumsen (2001) apply Alonso's original ideas in transportation. Daly and Ortuzar theoretically and empirically explore data aggregation in travel demand modeling, different types of errors in modeling and forecasting, and the trade-off between model complexity and data accuracy with focus on the forecasting of mode and destination choice. They recommend that (i) the model building should take into account the efficient allocation of modeling resources, (ii) errors, especially those which violate basic assumptions of the model, should be minimized, and (iii) since measurement error is an important component of the overall error in modeling, it should be minimized given the budget. They thus emphasize the most efficient allocation of modeling resources.

Multiple imputation, originally developed to handle missing data (c.f. Rubin, 1987), may also help to solve the problem of measurement errors introduced by network-based LOS attributes in a disaggregate model of travel demand. Multiple imputation creates multiple imputed values and weights, and then combines the estimators using each set of values into a final consistent estimators that accounts for the errors in the imputation process. Brownstone (2001) uses multiple imputation approach to correct for the measurement errors due to the use of network data of key variables such as travel time and travel cost in transportation when a small validation study is available to model the measurement error process. Recently, Walker et al. (2008) suggest a latent variable approach to correct for measurement errors introduced in travel demand modeling. They treat LOS attributes as latent variables. They deduce latent LOS attributes by combining the measurement equations with the mode choice model. Since we do not know the distribution of measurement errors, we have to make strong assumptions about their distributions to apply this approach. Importantly, the assumptions made about the distribution of measurement errors may have a direct influence on model results. We generally expect that any method accounting for EIVs in the estimation process must be based on some prior information about the variance and the distribution of measurement errors.

Obtaining disaggregate LOS attributes with the help of new technology such as geographic information system (GIS) and global positioning system (GPS) can be another way of overcoming the problem. However there are only a few recent studies that use disaggregate LOS attributes. Sacramento model (Bowman et al., 2006) can be considered as one of them because the model does not rely entirely on zone-to-zone attributes. The destination choices are parcels and the models use attributes such as distance from the parcel to the nearest transit stop. This is not an entirely accurate measure because a person might want to go in a direction other than that of the nearest transit stop. Nevertheless, it improves considerably over relying only on zone-to-zone attributes. Similarly, Bierlaire and Frejinger (2007) use network-free data to estimate a long distance route choice model quite recently. The data were collected by asking the travelers about the origins and destination of their trips as well as intermediate locations in the path in the telephone interviews. The reported data supplemented by GPS data were later reconciled with a network based model. The reported LOS attributes in the interview, often called perceived LOS attributes, can be thus another potential source.

The stated preference (SP) survey gives the truly disaggregate data of LOS attributes of transportation system. The power of these data is that we can also collect information on respondents' attitudes and perceptions including qualitative attributes such as comfort, flexibility, reliability and so on of different travel modes. There are a large number of travel demand studies using the SP data. Hensher (1994) has a review article regarding the development in the applications of SP data in travel behavior research before the early 1990s. Similarly, Brownstone and Train (1999), Hess et al. (2007) and Catalano et al. (2008) (among others) are a few of the applications of the SP data to model travel demand.

7. Summary and conclusions

The network-based LOS attributes introduce random measurement errors for individual trips thereby causing EIVs problem in a disaggregate model of travel demand. We investigated the EIVs modeling in linear and nonlinear regression models with special emphasis on discrete choice models, particularly travel demand models. As mentioned in Section 2, far less attention has been paid to the EIV problem in multinomial choice modeling in literature. McFadden (1984) pointed out this as an important unsolved problem. This is still an important unsolved problem although it is obviously a vital issue in travel demand modeling where network-based LOS attributes introduce EIV problem, potentially causing wrong policy decisions due to the application of the models having biased parameters for policy and planning purposes.

The analytically tractable expressions for biases and biased adjusted estimators do not exist for multinomial choice models, making an á priori assessment of possible biases due to random measurement errors difficult. Simulation may be the possible option. The purpose of the paper was to investigate the possible magnitude and direction of biases of parameter estimates of the MNL model for travel mode choice using simulation. We chose access/egress time and walking and cycling distance to work for the purpose. These two variables were chosen because we believe that the ratio between the variance of measurement errors and the variance of the explanatory variable is among the highest for LOS-variables (given equation 5 in Section 2.1). Errors in these variables are inherent when we use LOS attributes obtained from zonal based network model.

We simulated measurement errors of varying variances and distributions to these variables. The simulation results show that the possible bias of coefficient of variable measured with error was at least as severe as in simple linear regression model with mismeasured variable having a comparable error structure. Other coefficients also got affected, both of the variables in the utility function containing the mismeasured variable and the other variables in other utility functions. As expected, the simulation results show that the absolute value of biases tended to increase with the variance of the measurement error given the variance of the true variable in sample. For our particular mode choice model, a persistent result over all simulations was that measurement errors in access/egress time and walking and cycling distance to work caused the coefficients of in-vehicle time both of car and public transport bias downward in absolute value, with the largest bias for public transport. Measurement errors had comparatively small impact on the generic coefficient of travel cost implying that implicit values of time got downward biased with the largest bias for public transport. Measurement errors introduced simultaneously to both the variables greatly amplified the effects, while the biases in the coefficient of the mismeasured variables actually got smaller surprisingly. At present, we can not say whether biases with the same pattern and magnitudes will also show up in similar models estimated on other data sets and in models for simultaneous choice of mode and destination or in models with different types of nesting structures. If they do, the implications are rather severe.

This is a simulation study with its own limitations. We focused only on normal, lognormal, and triangular distributions of measurement errors. We do not know the true distribution and true variance of measurement errors. This could obviously be too simple in most real situations. Further tests are needed with other data sets from other areas before we conclude that our findings carry over to other mode choice models and to models of multidimensional choice of mode and destination. However, we think this can contribute to analysis of EIVs problems in discrete choice models where the literature is very sparse. We also believe that the results from this study may be relevant to other situations where the explanatory variables used in estimation of discrete choice models are estimated rather than directly observed. By no means have we been able to provide answers to this rather complex problem, but we believe that our contribution may give some insight into these issues anyway.

An extensive discussion of possible solutions for EIVs problem in discrete choice analysis goes beyond the scope of the study. On possible measure is of course to invest more in accurate measurement with special focus on critical variables. The variance of measurement errors in network LOS attributes can be reduced by reducing the size of zones (and consequently increasing the number of zones) of area under study. The networks can also be coded more accurately. We can also use LOS attributes estimated on address to address (given that this possibility is available) for model estimation even if the models are applied for trips between zones later. Multiple imputation could be another option to estimate the values of LOS attributes (see Brownstone, 2001). The other option may involve accounting for EIVs in model estimation, for example, by treating the true value as a latent variable (see Walker et al., 2008). However, we need to know more about the distribution of errors to properly account for EIVs in the modeling process. The key finding in this paper is that EIV result in biased parameter estimates of an MNL model and consequently leading to poor policy decisions and we need to develop methods that explicitly recognize and correct for such errors or else resulting policy decision will be wrong. In addition, the paper discusses some of the ways of overcoming the problem. So the paper significantly contributes to link between theory and practice in transport.

Acknowledgement

We are thankful to Antti P. Talvitie and two anonymous reviewers for their help and compliments on earlier draft of this paper. The authors are fully responsible for errors and omissions if any.

References

Alonso, W., 1968. Predicting best with imperfect data. Journal of the American Planning Association 34, 248-255.

Ben-Akiva, M., Lerman, S., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT press, Massachusetts, Cambridge.

Bhatta, B.P. 2009. Discrete Choice Analysis with Emphasis on Network-Based Level of Service Attributes in Travel Demand Modeling, Ph.D. thesis, Molde University College, Molde, Norway.

Bhatta, B.P., Larsen, O.I., 2010. Are intrazonal trips ignorable? doi:10.1016/j.transnspol.2010.04.04.

Bierlaire, M., Frejinger, E., 2008. Route choice modeling with network-free data. Transportation Research Part C: Emerging Technologies 16, 187-198

Bound, J, Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In Hekmann J.J. and Learner, E. (eds.) Handbook of Econometrics Vol. 5. Elsevier Science, North-Holland.

Bowman, J., Bradley, M., Gibb, J., 2006. The Sacramento activity-based travel demand model: estimation and validation results presented at the European Transport Conference, 18-20 September 2006, Strasbourg, France.

Brownstone, D., 2001. Discrete choice modeling for transportation. In Hensher, D. (ed.). Travel Behavior Research: the Leading Edge. The proceeding of the 9th International Association for Travel Behavior Research. Pergamon, Oxford, UK.

Brownstone, D., Train, K., 1999. Forecasting new product penetration with flexible substitution patterns. Journal of Econometrics 89, 109-129.

Buzas, T.S., Tosteson, T.D., Stefanski, L.A., 2003. Measurement Error. Institute of Statistics, Mimeo Series No. 2544.

Carroll, R. J., Ruppert, D., Stefanski, L.A. Criniceanu, C.M., 2006. Measurement Error in Nonlinear Models, 2nd edition. Chapman & Hall/CRC, London.

Chang, K-t, Khatib, J., Ou, Y., 2002. Effects of zoning structure and network detail on traffic demand modeling. Environment and Planning B 29, 37-52.

Catalano, M., Casto, B., Migliore, M., 2008. Car sharing demand estimation and urban travel demand modeling using state preference techniques. European Transport 40, 33-58.

Daly, A., Ortuzar, J.D., 1990. Forecasting and data aggregation: theory and practice. Traffic Engineering and Control 31, 632-643.

Fuller, W.A., 1987. Measurement Error Models. John Wiley, New York, USA.

Greene, W.H., 2003. Econometric Analysis 5th edition. Pearson Education International, New Jersey.

Gujarati, D.N., 2003. Basic Econometrics 4th edition. McGraw-Hill Higher Education, New York.

Hensher, D., 1994. Stated preference analysis of travel choices: the state of practice. Transportation 21, 107-133.

Hess, S., Polak, J.W, Daly, A., Hayman, G., 2007. Flexible substitution patterns in models of mode and time of day choice: new evidence from the UK and the Netherlands. Transportation 34, 213-238.

Hu, Y., 2006. Bounding parameters in a linear regression model with a mismeasured regressor using additional information. Journal of Econometrics 133, 51-70.

Kao, C., Schnell, J. F., 1987. Errors in variables in the multinomial response model. Economics Letters 25, 249-254.

Li, T., Hsiao, C., 2004. Robust estimation of generalized linear models with measurement errors. Journal of Econometrics 118, 51-65.

McFadden, D., 2000. Disaggregate behavioral travel demand's RUM side: a 30-year perspective. Paper presented at Conference on the International Association of Travel Behavior Research, Brisbane, July 2, 2000, Australia.

McFadden, D., 1984. Econometric analysis of qualitative response models. In Griliches, Z., Intriligator, M.D. (eds.). Handbook of Econometrics, Volume 2. Elsevier Science Publishers BV, North-Holland, Amsterdam.

Ortuzar, J.D., Willumsen, L.G., 2001. Modeling Transport 3rd edition. John Wiley & Sons Ltd, England.

Rekdal, J., 1999. Yrkesaktives reiseaktivitet. TØI report 444/1999 (In Norwegian with English summary). Institute of Transport Economics, Oslo, Norway.

Rubin, D.B., 1987. Multiple Imputation for Non-response in Surveys. John Wiley, New York.

Stangeby, I., 1997. Transport in the course of work. TOI report 375/1997 (in Norwegian with English summery). Institute of Transport Economics, Oslo, Norway.

Talvitie, A., Dehghani, Y., 1980. Models for transportation level of service. Transportation Research B 14, 87-99.

Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, UK.

Wansbeek, T., Meijer, E., 2000. Measurement Error and Latent Variables in Econometrics. Elsevier, Amsterdam.

Yatchew, A., Griliches, Z., 1985. Specification error in probit models. The Review of Economics and Statistics 67, 134-139.

Walker, J. L., Jieping, L., Sirinivasan, S., Bolduc, D., 2008. Travel demand models in the developing world: correcting for measurement errors. TRB 87th Annual Meeting Compendium of Papers DVD.

Zhang, M., Kukadi, N., 2005. Metrics of urban form and the modifiable areal unit problem. Transportation Research Record 1902, 71-79.