Netw Spat Econ (2011) 11:393–418
DOI 10.1007/s11067-010-9143-6

On the Treatment of Repeated Observations in Panel Data: Efficiency of Mixed Logit Parameter Estimates

María Francisca Yáñez · Elisabetta Cherchi · Benjamin G. Heydecker · Juan de Dios Ortúzar

Published online: 27 August 2010

Abstract Travel demand models are usually estimated using cross-sectional data. Although the use of panel data has recently increased in many areas, there are still many aspects that have not been fully analyzed. Some examples of unexplored topics are: the optimal length of panel surveys and the resulting issue of how to model panel data correctly in the presence of repeated observations (for example, several trips per week, by people in a panel with waves every six months) and whether, and to which extent, this affects the efficiency of the estimated parameters and their capability to replicate the true situation. In this paper we analyse this issue and test the effect of including journeys made, with the same characteristics, several times in a week. A broad variety of models accounting for fixed parameters but also for random heterogeneity and correlation among individuals were estimated using real and synthetic data. The real data comes from the *Santiago Panel* (2006–2008), while the synthetic data were appropriately generated to examine the same problem in a controlled experiment. Our results show that having more observations per individual increases the probability of capturing more effects (i.e. different types of heterocedasticity), but having identical observations in a data panel reduces the

M. F. Yáñez (✉) · J. Ortúzar
Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Casilla 306, Cod. 105 Santiago 22, Chile
e-mail: mafrancisca@gmail.com

J. Ortúzar
e-mail: jos@ing.puc.cl

E. Cherchi
CRiMM – Department of Territorial Engineering, Università di Cagliari, Piazza d'Armi, 16 - 09123 Cagliari, Italy
e-mail: echerchi@unica.it

B. G. Heydecker
Centre for Transport Studies, University College London, Gower Street, WC1E 6BT London, UK
e-mail: ben@transport.ucl.ac.uk

capability to reproduce the true phenomenon. Consequently, the definition of panel survey length demands us to consider the implicit level of routine (i.e. the proportion of identical observations) in it.

Keywords Panel data with repeated observations · Discrete choice models · Parameter efficiency

1 Introduction

For many years, travel demand models have been estimated using mainly cross-sectional data involving the collection of information, at one point in time, over a relatively large number of individuals. One problem with this data structure is that it does not allow to model user's behaviour in the presence of temporal effects. Only recently, the increased need for better understanding of individual behaviour and the parallel advances in travel demand modelling techniques and computational power has made it possible to *re-discover* panel data. In this case, contrary to cross-sectional data, information is collected at a number of successive points in time retaining the same individuals for the entire series of surveys.

Panels can be classified into two categories: "long survey panels" and "short survey panels" (see Yáñez et al. 2010). The first are the most common in the literature and consists of repeating the same survey (i.e. with the same methodology and design) at "separate" times, for example once or twice a year for a certain number of years or before-and-after an important event. The second type of panel consists of multi-day data where repeated measurements on the same sample of units are gathered over a "continuous" period of time (e.g. seven or more successive days), but the survey is not repeated in subsequent years. Some recent examples of panel data gathered over a continuous period of time are the two-day time-use diary for the US National Panel Study of Income Dynamics (2002) and the six-week travel and activity diary data panels collected in Germany (Axhausen et al. 2002) and Switzerland (Axhausen et al. 2007). One of the most recent panels is that gathered in Santiago (2006–2008), which combines both the "short" and "long" survey panel approaches. In fact, the *Santiago Panel* (Yáñez et al. 2010) is a five-day pseudo diary which has also four waves before and after the implementation of a radically new and much maligned urban public transport system (Transantiago).

Panel data offer major advantages over cross-sectional data because repeated observations from the same individual generally allow for more accurate measurements of changes in individual mobility. In particular, a "long survey panel" including data for the same respondents at "separate" points in time allows studying dynamic effects over waves, such as habit formation, learning and the reaction to important policies (Yáñez and Ortúzar 2009; Yáñez et al. 2009). A "short survey panel", on the other hand, such as a multi-day panel over only one continuous period of time allows to detect effects such as rhythms of daily life (Axhausen et al. 2002), to explain current behaviour on the basis of the individuals' history and experience (Cirillo and Axhausen 2006) and to account for interpersonal variability and correlation across individuals over different time periods (Cherchi and Cirillo 2008).

Although panel data have many advantages, they can also suffer from certain specific problems, such as:

- Attrition bias: associated to the fact that some respondents may drop out among waves; the problem is that usually this does not occur randomly but might be depend on the person´s socio-economic characteristics, causing a bias in the results (Ruiz 2006).
- Panel effect: participating in a panel may affect individual decisions as panel members become more conscious about their own behaviour.
- Fatigue effect: respondents tend to show, for example, declining trip rates and omit short trips or journeys using slow modes; this error tends to increase with the number of days of the diary (Van Wisen and Meurs 1989).

Another problem (specific to successive multi-day panel data), which is important to take into account in order to establish the survey length, is the presence of repeated observations. It is normal to expect that individuals, in different days, may repeat exactly some trips (typical cases are the systematic trips to work that are often made every day with the same characteristics: time, cost, purpose, mode, and so on). The problem arises when these data are used for model estimation, as the way in which repeated information is treated may affect the estimation results.

In this paper we analyze the effect of repeated observations in a panel context, specifically, the problem of how repeated data should be treated in model estimation and whether, and to what extent, this affects the efficiency of the estimated parameters and their capability to replicate the true phenomenon. In fact, the way a panel data model is specified affects model estimation because the above assumptions imply substantial differences in model specification, data variability, degree of correlation over individuals and different sample dimensions. In particular, using simulated data we analyse the impact of multi-day-panel-survey length on the capability of models to recover the "true parameters".

The analyses were carried out with real data from the second wave of the *Santiago Panel* and with synthetic data generated to test, in a controlled experiment (free of unknown effects), the effect of the repeated observations on the efficiency of the model parameters. Efficiency is measured in this case by the Fisher information matrix; this is inversely related to sample size, to the values of the attributes associated to the estimated parameters and to the probability associated to the chosen alternative (McFadden 1974). Rose and Bliemer (2008) analysed the effect of the number of alternatives, attributes, and attribute levels on the optimal sample size for stated choice experiments in MNL models, as part of their search for the experimental design with highest asymptotic efficiency of the estimated parameters. They found that only the range of attribute levels could offer an explanation for some problems of convergence encountered in their experiments. In line with these analyses on experimental design, Cherchi and Ortúzar (2008) demonstrated that while efficiency clearly improves with sample size, data variability does not always increase it; rather, in some cases it might not even be beneficial to increase data variability or it might be better to have a smaller range of variation.

In contrast, the repeated observations in a panel will increase the number of observations but might reduce data variability, because observations that are identical do not bring new information about attribute trade-offs. Thus, when using panel data it is important to understand how efficiency is influenced by the repeated observations and up to what point these are beneficial. This last result is also crucial to determine the length of a multi-day panel survey which is something that has not been explored up to date. Moreover, as in the case of panel data the same individual provides more than one observation, we need to account for correlation among these, which obviously has a different effect depending on how the repeated observations are treated. Cherchi and Cirillo (2008) found that the effect of correlation is, to a large extent (at least 50%), given by the repeated observations; however, they only compared the case with and without repeated observations and did not explore the effect on the efficiency of the parameters and on the capability of a flexible model, such as Mixed Logit (ML), to reproduce the true phenomenon.

The rest of the paper is organised as follow. In section 2 we briefly discuss the issue of modelling with panel data and the effect of repeated observations in the efficiency of the estimated models. In section 3 the main features of the *Santiago Panel* data are described, and the results of several models estimated with these data are reported. Section 4 reports the same type of analyses but using synthetic data, in order to test also the capability of the ML model to reproduce the true phenomenon in the presence of panel data and repeated observations. Finally, section 5 summarizes our conclusions.

2 Modelling with panel data

Although panel data models have been estimated using fairly typical discrete choice functions, the presence of repeated observations makes it more appropriate to use a Mixed Logit (ML) formulation accounting for correlation among observations belonging to the same individual. As well known, a ML probability is the integral of standard logit probabilities over a density of parameters (Train 2003). In particular, when more than one observation per individual is available, we need to take into account the sequence of choices, made by the respondent; hence, if we assume the popular framework proposed by Revelt and Train (1998), which accommodates inter-respondent heterogeneity but assumes intra-respondent homogeneity in tastes, including the effect of the repeated choices by assuming that tastes vary across respondents in the sample, but stay constant across observations for the same respondent, the ML panel probability is given by the product of ML probabilities:

$$P_{qj}(\Omega) = \int\limits_{\mu_q} \prod_{t=1}^{T} \left(\frac{e^{V_{qjt}\left(\mu_q\right)}}{\sum\limits_{i \in A_q^t} e^{V_{qit}\left(\mu_q\right)}} \right) f\left(\mu_q | \Omega\right) d\mu_q \tag{1}$$

where V_{qi}^t is the observable component of the utility of option i for individual q at time t; and A_q^t is the choice set of individual q at time t. T is the number of periods, f is the so-called "mixing distribution" with means $\underline{\mu}$ and covariance matrix Ω (i.e. the so-called "population parameters") of the coefficients to be estimated in $\underline{V}$.

We also tested a generalised approach suggested by Hess and Rose (2009), which relaxes the assumption of intra-respondent homogeneity of tastes and where the choice probability is given by:

$$P_{qj}(\Omega) = \int\limits_{\alpha_q} \prod_{t=1}^{T} \left(\int\limits_{\gamma_{q,t}} \frac{e^{V_{qj}^{t}(\mu_q)}}{\sum\limits_{i \in A_q^t} e^{V_{qi}^{t}(\mu_q)}} g\left(\gamma_{q,t} \middle| \Omega_\gamma\right) d\gamma_{q,t} \right) h\left(\alpha_q \middle| \Omega_\alpha\right) d\alpha_q \tag{2}$$

where $\underline{\mu}$ is now a function of α_q which varies over respondents with density $h\left(\alpha_q \middle| \Omega_\alpha\right)$, and $\gamma_{q,t}$ which varies over all choices with density $g\left(\gamma_{q,t} \middle| \Omega_\gamma\right)$. This model has two integrals, inside and outside the product over choices; the outside integral accounts for inter-respondent heterogeneity as in the traditional model (Revelt and Train 1998), while the inside integral accounts for intra-respondent heterogeneity. Due to the limitations of currently available software, we used a simplified version (K_1 in the notation of Hess and Train 2009) available in BIOGEME (Bierlaire 2003).

To analyse the effects generated by variability and repeated observations in model structure we used the typical *t*-test and LR test (Ortúzar and Willumsen 2001). To check whether the estimated parameters differed significantly among specifications (i.e. different sample sizes), we used the *t**-test (Galbraith and Hensher 1982):

$$t^* = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\sqrt{\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2}} \tag{3}$$

where $\widehat{\beta}_i$ is the estimated parameter in specification i and $\widehat{\sigma}_i$ is its standard error.

The *t*-test is also inversely related to the Fisher information matrix that, as well known, measures the efficiency of the estimated parameters. In Maximum Likelihood estimation the expected value of the variance of the *kth* estimated parameter (i.e. the *kth* element of the diagonal of the Fisher information matrix) is given by:

$$E\left[\frac{\partial^2 \ell}{\partial \beta_k^2}\right] \cong \sum_{q=1}^{Q} \sum_{j \in A_q} \left[\frac{\partial^2 \left(y_{jq} \ln p_{jq}\left(x_{jq}; \beta\right)\right)}{\partial \beta_k^2}\right]_{\beta = \widehat{\beta}} \tag{4}$$

where $\ell = \ln \prod_q p_{qjt}^{y_{qjt}}$ is the log-likelihood function with respect to the parameters ($\boldsymbol{\beta}$) evaluated at their true values, p_{jq} is the probability that individual q chooses alternative j among the alternatives belonging to her choice set (A_q), $\boldsymbol{x_{jq}}$ are the level-of-service and socio-economic attributes, and y_{jq} equals one if j is the alternative actually chosen by individual q and zero otherwise.

Equation (4) shows that the efficiency of the estimated parameters depends on sample size, the values of the attributes associated to the estimated parameters and the probability of choosing the chosen alternative. In particular, the logit probability depends, among other things, on the data variability and on the variance of the error term (through the scale factor); thus, understanding the sensitivity of the efficiency of the estimated parameters is a complex task. Cherchi and Ortúzar (2008) analysed how the efficiency of the estimated parameters varied for RP and SP data. Here we extend their analysis to the case of panel data.

Equations (4), (5) and (6) below show the expression for one element of the Fisher information matrix. These expressions are convenient for our theoretical discussion because they help to understand easily what elements influence the matrix. However, in practice it would be better to try and measure the overall statistical efficiency of the expected outcomes of models estimated on a given dataset as in the experimental design literature (Rose and Bliemer 2009); this can be done by computing the negative inverse of the Fisher information matrix (i.e. the asymptotic covariance matrix, *AVC*), and then computing the D-error which is equal to det$(AVC)^{1/k}$ (with k the number of the parameters). Smaller D-error yields more efficient parameter estimates.

Let us consider, for simplicity, a binary choice logit model (i.e. with "fixed" parameters). The variance of the parameters estimated with panel data is analogous to the case with SP data:

$$\mathrm{var}\left(\widehat{\beta}\right) = -\frac{1}{\sum_q \sum_t \Delta x^2_{qjt}\widehat{p}_{qjt}\left(1-\widehat{p}_{qjt}\right)} \tag{5}$$

where Δx^2_{qjt} is the attribute difference among both alternatives in period t. However, in contrast to the case of SP data, when using information from a "short survey panel" the attribute values do not vary over periods as they are identical for the same individual. Thus, in this case we have that $\Delta x_{qjt} = \Delta x_{qj} \forall t$ and the variance becomes:

$$\mathrm{var}\left(\widehat{\beta}\right) = -\frac{1}{\sum_q \sum_t \Delta x^2_{qj}\widehat{p}_{qjt}\left(1-\widehat{p}_{qjt}\right)} \tag{6}$$

These equations show that the variance depends clearly on the number of repeated observations as well as on the data variability and number of observations. However, as noted by Cherchi and Ortúzar (2007) "... the efficiency of the parameter increases with the variability of the attribute but only for scale factors over 0.5. This effect, that might seem counterintuitive, is due to the effect that the scale factor has on the variability of the data, because efficiency reduces as data variability diminishes; and is also due to the second order function of the probability that tends to zero as the probability of the chosen alternative approximates one." It is important to note that a panel with identical repeated observations for each individual is a special case. In fact, in terms of the above discussion having equal observations repeated a certain number of time increases only marginally the variability of the attributes. In particular, if N is the number of observations and R is the number of times these are repeated for each individual, the variance of the attributes (Δx_{qj}) for N and RN observations is related by the following expression:

$$\frac{\mathrm{var}\left(R\Delta x_{qj}\right)}{\mathrm{var}\left(\Delta x_{qj}\right)} = \frac{(RN-R)}{(RN-1)} \tag{7}$$

Hence, identical repeated observations should not infuence (at least in theory) the efficiency of the estimated parameters. This result was also confirmed by computing the D-error typically used to generate efficient experimental designs (Rose and Bliemer 2008).

The extension of this result to the MNL and ML cases is not difficult. In the ML model, the variance of the mean of the random parameters is more complex, but the structure is basically the same (Cherchi and Ortúzar 2008). It is still inversely related to the square value of the attributes associated to each parameter (as in the case of the fixed parameters model), to the number of repeated observations, and is also a function of the probabilities (Bliemer and Rose 2010).

3 Experiments with real data

The real data set used in this paper comes from the second wave of the *Santiago Panel*, which included four waves around the introduction of Transantiago (full details can be found in Yáñez et al. 2010). One important characteristic of this panel is that people were asked only about their work trips in the morning peak hour, but it was also enquired if this trip was repeated identically during other working days of the week and, if not, respondents were asked to report on the alternative trips for other days of the week.

In the *Santiago Panel* the sampling unit is the individual. The initial sample consisted of 303 individuals who lived in Santiago and worked full-time at one of the six campuses of the Pontificia Universidad Católica de Chile (see Fig. 1). The information sources used in the panel were:

(a) Face-to-face interviews with the aid of palms. The questionnaire had the following sections:

  i. Socioeconomic characteristics: age, sex, income, education, car ownership, work hours in a week, possession of a driving licence.

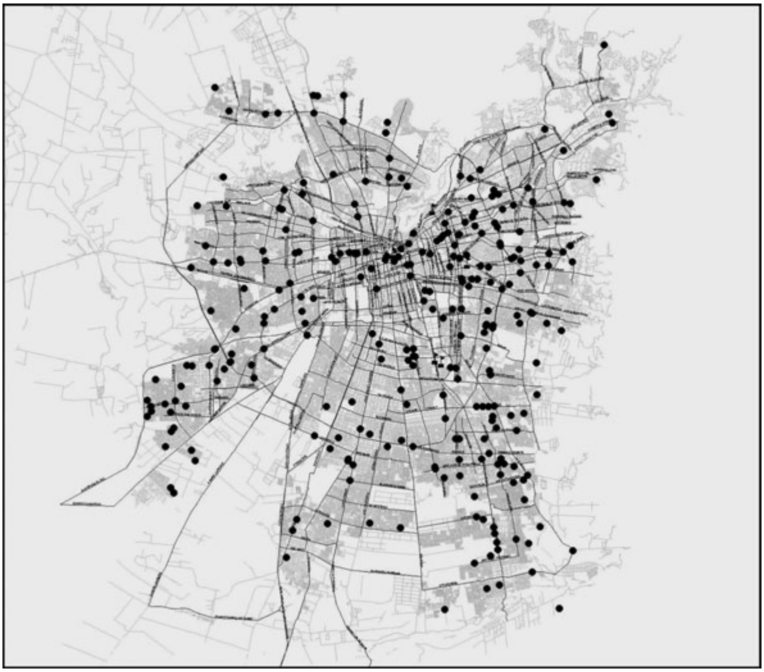

Fig. 1 Home locations of the panel respondents

ii. Characteristics of the trip to work: mode (could differ in different days), departure and arrival times, travel times per trip stage, fares.

iii. From the second wave onwards, subjective perception of the performance of the new public transport system (Transantiago).

(b) Very precise measurement of level-of service variables using state-of-the-art technology, such as GPS and geocoding of origin-destination pairs.

The design of the survey was based on the 2002–2006 Great Santiago Origin-Destination survey (DICTUC 2003; Ampt and Ortúzar 2004) and considered 12 modes. As illustrated in Fig. 2, the mode chosen most frequently was car driver followed by bus.

3.1 Modelling results with real data

We are interested in the effects of the multi-day-panel-survey length. Thus, using data from the second wave of the *Santiago Panel*, we considered five sample specifications, each one having 1, 2, 3, 4 and 5 observations (working days) per individual respectively. It means that the first specifications lose part of the information contained in the data, as 12% of the users selected more than one mode per week.

In summary, in the first specification, as in any cross-sectional dataset, we have only one available observation per individual, while in the other four specifications we have a panel data context, due to the multiple available observations per individual. However, each group of observations from the same individual has a rather low variation. This was expected as we have travel information for one working week and also because the members of the *Santiago Panel* have a fairly static routine. The main reason for this is that all of them work at least eight hours per day at the same place. This behaviour is similar to what was found by Cherchi and Cirillo (2008) in the six-week data panel from the Mobidrive (2000) survey, i.e. choices are much more persistent for tours the main activity of which is work or study. This feature reminds us of an important issue in panel data: the presence of habit or inertia in choice making behaviour (Cantillo et al. 2007).

Models of increasing complexity were estimated, starting with the simplest multinomial logit (MNL), followed by nested logit (NL) structures ending with a flexible random parameters mixed logit (ML) specification. We assumed that

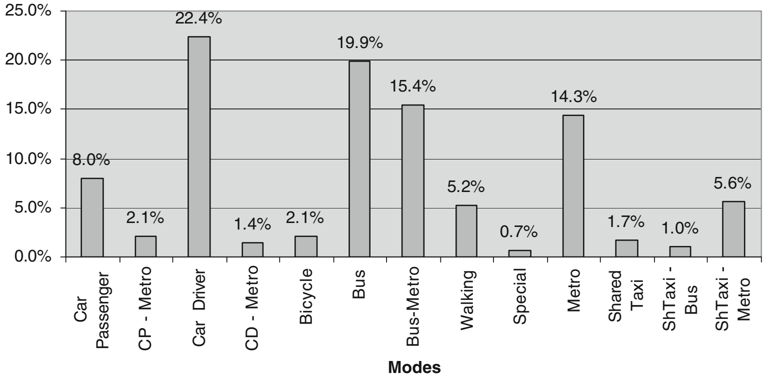

Fig. 2 Distribution of the mode chosen

individuals chose their mode among a finite set of alternatives and that this choice set could vary over working days and individuals. For space reasons we will only present here the best models found, and we will leave out several specifications tested (i.e. MNL models with systematic taste variations and models featuring the introduction of additional activities[1]) as they came out as no improvement (i.e. using the likelihood ratio test) over their restricted versions.

Tables 1 and 2 show the model results obtained for each data specification. In particular, and with the aim of simplifying the analysis, Table 1 presents the results of simple models which assume fixed parameters (*MNL* and *NL*), while Table 2 presents the results of the two selected ML models, based on the previous simple models, for each sample specification. The c*orrected* ρ^2 index (Ortúzar and Willumsen 2001) is provided for each case in order to account for the differences in terms of number of estimated parameters.

All of models incorporate systematic heterogeneity around the alternative specific constants (ASC) through the introduction of two variables related to the start of Transantiago.[2] The first was needed to represent the effect of the new payment system, which featured an integrated fare for bus and Metro trips (which did not exist in the old system) and which allows a time window of two hours to make an interchange. For this we decided to introduce a dummy variable that indicates if the individual was a new Metro user (Transantiago relies on Metro as its backbone), with an expected positive effect in the utility function. The second variable was needed to represent the comfort perceived by public transport users; this became especially important after the inauguration of Transantiago as the levels of overcrowding in the buses and the underground were extreme when the system started (although this has improved a lot after two years of operation, at least in the Metro, it is still higher than before). Regrettably, the initial survey design did not consider the inclusion of latent variables so we could not predict the importance of the above effects. Nevertheless, we obtained valuable information about the subjective perceptions of individuals, and we used these to create a dummy variable that took the value of one for respondents stating that comfort was the attribute that worsened the most after Transantiago. The expected sign of this variable is positive, as users chose the mode despite declaring a negative change in comfort. It is important to remark that these two new variables obtained high t-ratios for every model and data specification. Moreover, when they were not considered the models presented problems with the signs of key policy variables.

To facilitate comparisons, all models presented have the same "root" defined by *MNL*. Therefore, they all started with the same utility function. The differences are:

– the *NL* model also considers a hierarchical structure, nesting the modes involving the use of a bus,

[1] The panel survey also enquired about all activities involved in the tour that had *work* as its main purpose; this allowed us to incorporate the effect of additional activities in the choice of mode.

[2] It is important to note that the second wave was taken just three months after the introduction of Transantiago, when the situation was still chaotic in the city.

Table 1 Estimation results for MNL and HL models

	Specification 1 (1 obs.)		Specification 2 (2 obs.)		Specification 3 (3 obs.)		Specification 4 (4 obs.)		Specification 5 (5 obs.)	
	MNL	HL	MNL	HL	MNL	HL	MNL	HL	MNL	HL
No. of observations	231	231	462	462	693	693	923	923	1151	1151
Number of cars	1.67	1.68	1.67	1.68	1.67	1.68	1.7	1.7	1.79	1.8
t-test vs 0	2.3	2.35	3.26	3.33	3.99	4.07	4.67	4.76	5.51	5.6
CW (mean)	−0.024	−0.019	−0.024	−0.019	−0.024	−0.019	−0.024	−0.020	−0.023	−0.020
t-test vs 0	−2.28	−2.15	−3.22	−3.04	−3.95	−3.73	−4.64	−4.41	−5.15	−4.88
TRAVEL	−0.015	−0.008	−0.015	−0.008	−0.015	−0.008	−0.013	−0.008	−0.013	−0.0081
t-test vs 0	−1.36	−1.12	−1.93	−1.58	−2.36	−1.93	−2.49	−2.14	−2.69	−2.39
WAITING	−0.06	−0.039	−0.064	−0.039	−0.064	−0.039	−0.061	−0.040	−0.057	−0.038
t-test vs 0	−2.71	−2.09	−3.84	−2.95	−4.7	−3.62	−5.22	−4.41	−5.58	−4.78
WALKING	−0.0656	−0.0619	−0.0656	−0.0619	−0.0656	−0.0619	−0.0654	−0.0622	−0.0639	−0.061
t-test vs 0	−3.1	−3.27	−4.39	−4.63	−5.37	−5.67	−6.27	−6.61	−6.93	−7.25
INTERCHANGES	−0.47	−0.28	−0.47	−0.28	−0.47	−0.28	−0.44	−0.27	−0.46	−0.31
t-test vs 0	−1.52	−1.38	−2.15	−1.95	−2.63	−2.39	−2.84	−2.62	−3.41	−3.24
COMFORT	1.92	1.88	1.92	1.88	1.92	1.88	1.77	1.75	1.68	1.66
t-test vs 0	3.93	3.93	5.55	5.55	6.8	6.8	7.33	7.34	7.77	7.8
TRANSANTIAGO	1.92	2.87	3.68	2.87	3.68	2.87	3.48	2.59	3.32	2.44
t-test vs 0	4.6	3.33	6.51	4.7	7.97	5.76	9.11	6.42	10.02	7.02
CAR DRIVER	−0.553	−0.233	−0.553	−0.233	−0.553	−0.233	−0.509	−0.274	−0.524	−0.317
t-test vs 0	−0.95	−0.43	−1.34	−0.61	−1.64	−0.75	−1.77	−1.05	−2.06	−1.36
CAR PASSENGER	−0.229	0.177	−0.229	0.178	−0.229	0.178	−0.191	0.119	−0.16	0.117
t-test vs 0	−0.4	0.32	−0.57	0.45	−0.69	0.55	−0.67	0.44	−0.64	0.49
SHARED TAXI	−0.351	−0.387	−0.351	−0.387	−0.351	−0.387	−0.497	−0.552	−0.637	−0.698
t-test vs 0	−0.6	−0.69	−0.85	−0.98	−1.04	−1.2	−1.65	−1.92	−2.33	−2.66

METRO	0.229	0.462	0.229	0.462	0.229	0.462	0.264	0.439	0.216	0.369
t-test vs 0	0.58	1.31	0.83	1.86	1.01	2.27	1.36	−2.53	1.25	2.4
WALK	2.08	2.38	2.08	2.38	2.08	2.38	1.89	2.12	1.73	1.93
t-test vs 0	2.69	3.26	3.8	4.61	4.65	5.65	5.03	5.99	5.23	6.2
BIKE	−1.33	−1.03	−1.33	−1.03	−1.33	−1.03	−1.37	−1.14	−1.46	−1.25
t-test vs 0	−1.91	−1.51	−2.7	−2.13	−3.31	−2.61	−3.98	−3.43	−4.73	−4.2
PARK'N'RIDE	−3.07	−2.82	−3.07	−2.82	−3.07	−2.82	−2.96	−2.75	−2.9	−2.71
t-test vs 0	−2.53	−2.45	−3.58	−3.46	−4.38	−4.24	−4.9	−4.78	−5.37	−5.26
KISS'N'RIDE	−1.58	−1.32	−1.58	−1.32	−1.58	−1.32	−1.4	−1.21	−1.18	−1.04
t-test vs 0	−2.24	−2.05	−3.16	−2.9	−3.88	−3.55	−4.17	−3.94	−4.09	−3.94
SH. TAXI-METRO	−0.383	−0.559	−0.383	−0.559	−0.383	−0.559	−0.369	−0.551	−0.329	−0.527
t-test vs 0	−0.86	−1.49	−1.22	−2.1	−1.49	−2.57	−1.66	−2.89	−1.67	−3.07
BUS –METRO	−0.227	−0.135	−0.227	−0.135	−0.227	−0.135	−0.209	−0.137	−0.169	−0.114
t-test vs 0	−0.68	−0.74	−0.96	−1.05	−1.17	−1.28	−1.26	−1.44	−1.15	−1.33
BUS-SH. TAXI	−1.62	−0.675	−1.62	−0.675	−1.62	−0.675	−1.49	−0.68	−1.38	−0.679
t-test vs 0	−2.36	−1.46	−3.34	−2.07	−4.09	−2.53	−4.48	−2.96	−4.84	−3.38
Phi		0.42		0.42		0.42		0.46		0.48
t-test vs 0		2.55		*3.60*		4.41		5.54		6.29
t-test vs 1		−3.47		−4.90		−6.00		−6.54		−6.86
L(max)	−193.527	−190.874	−387.053	−381.748	−580.58	−572.622	−654.144	−778.484	−999.301	−988.15
L(max) / Sample	−0.838	−0.826	−0.838	−0.826	−0.838	−0.826	−0.709	−0.843	−0.868	−0.859
L(max) / Residual degree of freedom (*)	−0.913	−0.905	−0.874	−0.864	−0.861	−0.851	−0.724	−0.862	−0.883	−0.874
Corrected ρ^2	0.239	0.245	0.273	0.281	0.284	0.293	0.276	0.284	0.268	0.275
Number of parameters	19	20	19	20	19	20	19	20	19	20
LR test		−5.31		−10.61		−15.92		248.68		−22.30

Details: *CW* cost/wage rate; Travel, Walking and Waiting Time [min]

(*) L(max) / Residual degree of freedom = L(max) / (Sample - Number of parameters). This index is close to L(max) / Sample for large sample sizes, but it provides a valuable measure of goodness of model fit that is independent of the sample size.

Table 2 Estimation results for ML models

	Specification 2 (2 obs.)		Specification 3 (3 obs.)		Specification 4 (4 obs.)		Specification 5 (5 obs.)	
	ML	HL	ML	HL	ML	HL	ML	HL
No. of observations	462	462	693	693	923	923	1151	1151
Number of cars	4.10	4.94	5.38	5.29	4.33	5.11	4.98	4.67
t-test vs 0	2.34	2.61	3.11	3.20	2.76	3.53	3.84	3.42
CW (mean)	−0.53	−0.48	−0.88	−0.74	−0.75	−0.69	−0.65	−0.92
t-test vs 0	−4.20	−3.71	−6.19	−4.80	−7.26	−6.15	−9.14	−6.40
CW (st.dev.)	0.94	0.74	1.64	1.35	1.37	1.22	1.21	1.05
t-test vs 0	3.78	3.92	5.82	5.10	7.29	6.18	9.48	6.30
TRAVEL	−0.04	−0.01	−0.06	−0.02	−0.05	−0.03	−0.05	−0.03
t-test vs 0	−2.58	−1.68	−3.57	−1.95	−3.92	−2.82	−4.46	−3.27
WAITING	−0.12	−0.06	−0.14	−0.10	−0.15	−0.10	−0.13	−0.10
t-test vs 0	−3.82	−2.80	−4.61	−3.57	−5.68	−4.21	−5.80	−4.42
WALKING	−0.11	−0.07	−0.13	−0.10	−0.12	−0.11	−0.12	−0.09
t-test vs 0	−3.86	−2.77	−5.00	−2.62	−5.53	−5.36	−5.95	−5.97
INTERCHANGES	−0.87	−0.35	−0.72	−0.39	−0.64	−0.41	−0.78	−0.56
t-test vs 0	−2.56	−1.58	−2.17	−1.87	−2.23	−1.96	−3.23	−2.73
COMFORT	2.53	1.70	1.98	1.41	2.29	1.54	2.47	1.86
t-test vs 0	3.05	2.32	2.67	1.99	3.43	2.58	3.91	3.60
TRANSANTIAGO	3.81	2.83	3.43	2.66	3.00	2.32	2.82	2.31
t-test vs 0	4.91	3.76	5.23	3.83	6.06	4.58	6.52	5.44
CAR DRIVER	−0.27	0.701	−0.617	−0.0135	−0.482	−0.202	−0.823	−0.259
t-test vs 0	−0.32	0.89	−0.8	−0.21	−0.76	−0.33	−1.44	−0.5
CAR PASSENGER	−2.32	−1.15	−3.66	−2.51	−2.77	−2.27	−2.77	−2.1
t-test vs 0	−2.43	−1.21	−4.07	−2.04	−3.88	−3.21	−4.14	−3.43
SHARED TAXI	−0.542	−0.169	−0.43	−0.408	−0.519	−0.353	−0.674	−0.488
t-test vs 0	−0.65	−0.22	−0.6	−0.53	−0.84	−0.59	−1.09	−1

METRO	0.694	1.27	0.894	1.31	0.792	1.11	0.588	0.893
t-test vs 0	0.438	3.57	2.13	2.84	2.23	3.36	1.9	3.14
WALK	−0.275	0.56	−1.47	−0.912	−1.1	−0.95	−1.28	−0.754
t-test vs 0	−0.24	0.5	−1.34	−0.78	−1.22	−1.14	−1.31	−1.01
BIKE	−10.4	−11.5	−9.03	−16.9	−15.3	−16.8	−14.8	−11.1
t-test vs 0	−3.22	−2.44	−4.69	−3.58	−2.85	−4.5	−6.24	−2.89
PARK’N’RIDE	−3.1	−2.14	−2.92	−2.32	−2.69	−2.21	−3.13	−2.17
t-test vs 0	−2.06	−1.67	−2.13	−1.98	−2.5	−2.16	−3.15	−2.41
KISS’N’RIDE	−1.62	−0.815	−1.45	−0.89	−1.5	−1.07	−1.27	−0.89
t-test vs 0	−2.46	−1.47	−2.58	−1.53	−3.18	−2.5	−3.18	−2.44
SH. TAXI-METRO	−1.32	−0.175	−0.952	−0.231	−1.06	−0.478	−1.57	−0.548
t-test vs 0	−1.58	−0.28	−1.37	−0.32	−1.82	−0.91	−3.09	−1.05
BUS –METRO	0.7	0.794	1.27	1.32	1.15	1.09	1.02	0.946
t-test vs 0	1.62	2.93	3.01	3.3	3.3	3.84	3.58	4.07
BUS-SH. TAXI	−3.92	−0.716	−3.17	−0.565	−2.86	−1.21	−3.69	−1.61
t-test vs 0	−2.96	−1.04	−3.1	−0.49	−3.27	−1.44	−5.36	−2.52
υ_q^d							0.213	0.0551
t-test vs 0							1.99	1.98
Phi		0.18		0.27		0.51		0.61
t-test vs 0		1.97		1.89		3.47		4.76
t-test vs 1		−8.90		−5.08		−3.30		−3.09
L(max)	−292.946	−289.383	−367.971	−360.665	−500.887	−497.22	−623.488	−627.189
L(max) / Sample	−0.634	−0.626	−0.531	−0.520	−0.543	−0.539	−0.542	−0.545
L(max) / Residual degree of freedom	−0.663	−0.656	−0.547	−0.537	−0.555	−0.551	−0.552	−0.556
Corrected ρ^2	0.44	0.444	0.536	0.544	0.533	0.535	0.537	0.533
Number of parameters	20	21	20	21	20	21	21	22
LR test		−7.126		−14.612		−7.334		7.402

Details: *CW* cost/wage rate; Travel, Walking and Waiting Time [min]

(*) L(max) / Residual degree of freedom = L(max) / (Sample - Number of parameters). This index is close to L(max) / Sample for large sample sizes, but it provides a valuable measure of goodness of model fit that is independent of the sample size

- the *ML_INTER* model allows for random variation in the cost attribute,[3] which also includes inter-respondent heterogeneity,[4] and
- the *ML_INTER_INTRA* model is based on *ML_INTER*, but also includes intra-respondent heterogeneity via the error components.

The signs of all estimated parameters are consistent with microeconomic theory. However, the expected relationship between time coefficients ($|\beta_{travel\ time}| < |\beta_{walk\ time}| < |\beta_{wait\ time}|$) is achieved only in the case of ML models for the specification with five observations per individual, and in the case of ML models that do not include nested structures for the other sample specifications.

As well-known, panel correlation is usually accommodated via (i) random parameters (RP) or (ii) error components (EC). The main difference between these two approaches is that the EC are applied to constants (thus, in this case it is mandatory to set up an error-component reference alternative, i.e. an alternative without error component) while the RP apply to variables that vary across the sample. This gives an advantage to the RP approach because it does not suffer from identification problems per se (Walker 2001).

Both methods generate correlation among alternatives, but while the RP-panel generates correlation among all the alternatives, the EC-panel generates correlation among a sub-set of (*n*-1) alternatives, generating confounding effects with NL forms. This is the reason why in the literature the RP-panel is usually preferred over the EC-panel method. However, in the RP-panel version correlation and random heterogeneity are confounded and the effect of correlation can only be appreciated comparing the log-likelihood and eventually the *t*-test of the random parameters estimated with and without panel correlation. But we know that these measures do not allow the two effects to be disentangled. Moreover, if there is no random heterogeneity in the parameters associated to the level of service attributes, the RP-panel version cannot be used as it generates a false random heterogeneity, which is even worse than having a false NL structure, because it involves the level of service attributes.

Thus, although the influence of repeated observations (i.e. inter-respondent heterogeneity in tastes) can be considered via the estimation of random coefficients (i.e. RP-panel version), we believe there might be extra correlation across repeated observations besides the effect of the random cost coefficients. Therefore, even though random parameters and error components might induce confounding effects they might also account for slightly different effects. In fact, as long as both effects are significant the pure error-panel component (i.e. EC-panel version) will account for correlation in the preference for the alternative, while the random cost coefficient will account for correlation in tastes; we provide an analytical explanation below.

We postulated the following utility function:

$$U_{iq}^{d} = \alpha_i + \sum_j X_{ijq}^{d} \cdot \beta_{ijq} + \varepsilon_{iq}^{d} \tag{8}$$

[3] The Cost attribute was specified divided by the wage rate. This parameter has a Log-Normal distribution because this gave a superior fit. Nevertheless, we first used a Normal distribution and checked that the proportion of individuals with incorrect sign was in fact minimal (Sillano and Ortúzar 2005).

[4] We also tested for the presence of inter-respondent heterogeneity in tastes for the rest of the explanatory variables, but this effect was only significant for cost.

Here the error component has the form $\varepsilon_{iq}^{d} = \upsilon_{q}^{d} + \zeta_{iq}^{d}$ where ζ_{iq}^{w} is a random term distributed independent and identically Gumbel, and υ_{q}^{d} is a random effect which may be specific to the individual (υ_{q}), in which case we assume panel correlation as inter-respondent heterogeneity, and/or variable among observations (υ_{q}^{d}), in which case we assume intra-respondent heterogeneity.

As $\boldsymbol{\beta}$ have mean β_{ij} and standard deviation μ_{ijq} we can rewrite the utility function as:

$$U_{iq}^{d} = \left(\alpha_i + \upsilon_{q}^{d}\right) + \sum_{j} \left(\beta_{ij} + \mu_{ijq}\right) X_{ijq}^{d} + \zeta_{iq}^{d} \tag{9}$$

where X_{ijq}^{d} stands for level-of-service attribute j of option i for individual q on day d.

Equation (9) shows that both random coefficients and error components are separable. Indeed, the random coefficients assume inter-respondent heterogeneity.[5] It means, they allow tastes to vary across respondents in the sample, but stay constant across observations for the same respondent (Revelt and Train 1998). On the other hand, the "pure" error components (which also capture heterogeneity) affect the values of the ASC. Thus, the error component υ_{q}^{d} has the power to increase/decrease the relative weight of the ASC in relation to the explanatory variables in the utility function. Now, confounding effects are implicit in the ML structure and should not strictly depend on whether we account or not for random tastes. On the contrary, our experience (Cherchi and Ortúzar 2008) is that decomposing randomness in as many components as possible helps to reveal the confounding effects.

Another important issue regarding panel correlation has to do with estimation using available software. The standard form to incorporate panel correlation under the pure EC approach consists of adding an error component to (n-1) of the available alternatives, otherwise for identifiability reasons the model cannot be estimated (Walker 2001). However, this methodology may lead to biased results as it requires choosing, arbitrarily, a reference alternative for the error-component (i.e. one not having a pure panel error-component). Equation (8) allows us to see that this is equivalent to assuming that this reference alternative has the same ASC for all observations, while the remaining alternatives have different values for the ASC among observations. Moreover, even using the best normalization (i.e. selecting as error-component reference alternative that with the minimum variance), this methodology leads to a type of NL model, as it correlates the (n-1) alternatives including an error component. To avoid this problem, in this paper we modified the traditional estimation method by randomly selecting the error-component reference alternative for each individual.[6]

The ML models were estimated varying the number of draws to test for the presence of empirical identification problems (Walker 2001). It is important to note

[5] It is also possible to accommodate intra-respondent heterogeneity in the random parameters (Hess and Rose 2009), but in this case the best models accounted for inter-respondent heterogeneity by the cost parameter, and intra-respondent heterogeneity via the error components which, at the end, affect the ASC.

[6] Note that if the reference alternative varies across individuals in the EC-panel version, it becomes equivalent to the RP-panel version, at least in terms of correlation, with the (non-negligible) advantage that the variance does not vary across alternatives. However, it is important to note that this method is just a practical simplification which, while allowing to estimate panel models with EC using any estimation software (without increasing estimation time), overcomes only partially the problems of the standard method. In fact, our method does not generate systematically false NL correlation among two specific alternatives, but it is a model with randomly distributed heteroskedasticity across the sample. We wish to thank an anonymous referee for having made us clarify this point.

that in this case we did not find any such problems. All ML models estimated with repeated observations accounted correctly for correlation over a given individual.

Although the repeated observations problem has been widely studied in the literature, because it is especially relevant for SP data with multiple answers (Abdel-Aty et al. 1997; Cirillo et al. 2000; Ortúzar et al. 2000; Revelt and Train, 1998; Yen et al. 1998), in this case we are working with RP data. Moreover, as every data specification effectively represented the real series of choices observed in the sample (considering different number of days), the amplification effect was not artificial.

As expected, the models with repeated observations have parameters with higher t-ratios in all the model formulations estimated. However, the key travel time variable is not significant in the first two specifications.

Table 1, which presents the results of the simple models, shows that *NL* is significantly better than the restricted *MNL* for all sample specifications (the LR values are clearly larger than the critical value $\chi^2_{95\%,1} = 3.84$). Furthermore, the likelihood improvement is even larger for specifications with more observations. Contrariwise, Table 2 shows two singularities for the larger sample specification (i.e. 5 observations per individual). First, the *NL* version of the best ML model is not significantly better than its restricted version. Second, only the best model (*ML_INTER_INTRA*) allows us including not only inter-respondent heterogeneity via the cost parameter, but also intra-respondent heterogeneity via the error component. This particular result suggests that the model capability of including simultaneously inter and intra-respondent heterogeneity depends on the number of observations. Actually, we suspect that it is related to empirical identifiability issues, as all five specifications are theoretically able to accommodate both types of heterogeneity, but the empirical results show that the contribution of the intra-respondent heterogeneity is significant only for the largest sample specification. Additionally, the inclusion of intra-respondent heterogeneity improves model fit only in the presence of inter-respondent heterogeneity. Indeed, the models that allow only intra-respondent variation are not significantly better (according to the LR test) than their restricted versions, even for the largest sample specification.

In conclusion, the number of observations does not affect significantly the structural definition of models which do not include panel correlation. In all cases, no matter the number of observations per individual, the LR test indicates that the *NL* model is significantly better than its restricted (*MNL*) version. However, the presence of more observations seems to provide more evidence to discriminate between models if heterogeneity is accounted for correctly.

Agreeing with other findings reported in the recent literature (Cherchi and Cirillo 2008; Hess and Rose 2009), we confirm that the introduction of correlation is the main improvement factor in terms of fit. Moreover, this difference becomes higher with the number of repeated observations. Complementary, if we compare the likelihood value for the market share models (i.e. only with ASC) we could, in some sense, isolate our analysis taking out the influence of the other variables. In this way, we could confirm that the difference between the various models tested is mainly due to panel correlation.

Regarding the error component there are two important points to mention: first, the inclusion of inter-respondent heterogeneity via an error component was tested using the individual effect (υ_q), but for every specification the contribution was not significant. Second, the error component υ_q^d, which accommodates intra-respondent heterogeneity was found to distribute Normal among observations with mean zero and standard deviation as reported in Table 2.

In line with Hess and Rose (2009), we also observed a higher significance for inter-respondent heterogeneity. Anyway, we expected a lower level of intra-individual variation, as the second wave of the *Santiago Panel* shows a strong presence of routine. So, a higher variation in tastes across respondents rather than variation across observations is consistent with our previous assumptions.

As the final aim of this paper was to analyse the impact of the length of multi-day-panel-surveys, it is also important to compare the values of the estimated parameters. Thus, on the basis of Eq. (3), we found that:

- If we vary the sample specification of any model, we only get significant differences in the parameter value of the variable associated to new Metro users (i.e. TRANSANTIAGO), and only between the first specification (i.e. one observation per individual) and the rest of the sample specifications.
- Only a few parameters show significant differences among models for a given sample specification; for example, comparing the ML models with their restricted versions (*MNL* and *NL*), the specifications with 3, 4 and 5 observations per individual present significant differences for the parameter values of mean cost and the time-related variables (travel, waiting and walking time).

4 Simulated experiments

Following the tradition of Williams and Ortúzar (1982), a collection of datasets were generated in which pseudo-observed individuals behaved according to a choice process determined by the analyst. Simple samples, with three alternatives, two generic attributes (travel time and cost) and a Gumbel error (ε_q) term were generated. The marginal utility of travel time (β_{qt}) was varied such that the generated sample showed random heterogeneity in tastes which was generic among alternatives. In all experiments, the attributes and the travel time parameter were generated according to a censored Normal distribution to avoid mass points on the truncations that can induce estimation problems (Cherchi and Polak 2005).

The datasets were generated according to the following utility functions:

$$\begin{aligned} U_{q3} - U_{q1} &= -\beta_{qt}\left(Time_{q3} - Time_{q1}\right) - 1.5 \cdot \left(Cost_{q3} - Cost_{q1}\right) + \varepsilon_{q3} - \varepsilon_{q1} \\ U_{q2} - U_{q1} &= -\beta_{qt}\left(Time_{q2} - Time_{q1}\right) - 1.5 \cdot \left(Cost_{q2} - Cost_{q1}\right) + \varepsilon_{q2} - \varepsilon_{q1} \end{aligned} \tag{10}$$

where individuals q are assumed to evaluate only one choice task as in a RP data set. Several samples were generated with the characteristics illustrated in Table 3, but varying the sample dimensions: 5,000; 1,000; 500; 250 and 125 observations, the last four corresponding to random samples of the original 5,000 observations (and these also were varied up to four times). Finally, each sub-sample was randomly generated 10 times with different seeds for the random terms, yielding 40 samples in the last four cases and 10 in the first one.

Table 3 Characteristics of the synthetic samples

	Mean	Standard deviation	Coefficient of variation	Limits
$Time_3$-$Time_1$	2.66	0.79	0.30	[1.0; 4.0]
$Time_2$-$Time_1$	2.85	1.23	0.43	[1.0; 6.0]
$Cost_3$-Cost	−2.34	0.80	0.34	[−4.0; −1.0]
$Cost_2$-$Cost_1$	−2.67	0.79	0.28	[−4.0; −1.0]
Time parameter- 5000 obs. (**)	−0.86	0.55	0.64	[−0.01; −2.00]

(**) These values were computed as the average values over the 10 samples with 5000 individuals

The mean and standard deviation of the random travel time parameters were computed for all subsamples of 1000, 500, 250 and 125 members (with 1–5 observations per individual); figures did not change

In a second stage we assumed more than one observation for each individual, as in panel data, but that all observations belonging to the same individual were identical (i.e. as in 88% of cases in the *Santiago Panel*). Starting from the above samples (excepting that with 5,000 observations), 16 synthetic panel datasets were generated varying the number of identical observations from the equivalent to a 2-day panel up to a 5-day panel. The data for each panel was generated repeating the same observations *t* times (*t*=2, 3, 4, 5). Note that the sample size and the number of identical observations were appropriately chosen in order to test, also, the effect of identical observations independently from the effect of the panel length.

4.1 Modelling results

For each dataset above, several ML models allowing for random heterogeneity in the travel time parameter were estimated. In particular, Table 4 shows the values of the mean, standard deviation and coefficient of variation (CV) of the estimated parameters over the 10 samples generated with different seeds (we tested if there were variations for the four subsamples of each case with less than 5,000 observations, finding that these occurred only in one of the 40 sets for the 125 size case with no repeated observations). Table 5 in turn, shows the minimum and maximum values (over the 10 samples) of the t-tests against zero; for space reasons we do not report the results with three repeated observations, which followed the trend.

The models shown allow us to appreciate how results vary with the number of repeated observations. They also allow us comparing the effects on models estimated with samples of comparable size (i.e. a sample of 1,000 individuals with five identical observations each is comparable to a sample of 5,000 individuals with only one observation). All models were estimated varying the number of Halton draws of the simulated maximum likelihood procedure between 125 and 4,000 and, analogously to our real data set, we did not find estimation problems except for the single case mentioned above where we encountered an empirical identification problem. As results did not change much, those reported in the tables are based on only one of the four subsamples generated in each case.

Looking at the models estimated with only one observation per individual, the *t*-tests against zero correctly diminish with sample size because the standard deviation is inversely related to the dimension of the sample. This result is in line with the

Table 4 Mean and standard deviation of the estimated parameters over 10 samples

		ML1	ML2	ML3	ML4	ML5	ML6	ML7	ML8	ML9	ML10	ML11	ML12	ML13	ML14	ML15	ML16	ML17
No. of repeated obs.		1	1	1	1	1	2	4	5	2	4	5	2	4	5	2	4	5
No. of individuals		5000	1000	500	250	125	1000			500			250			125		
Mean	Ttime (mean)	−1.2286	−1.2262	−1.1497	−1.2386	−1.5156	−1.2263	−1.2263	−1.2262	−1.1497	−1.1497	−1.1498	−1.2384	−1.2383	−1.2383	−1.5160	−1.5149	−1.5150
	Ttime (st.dev.)	0.8478	0.8656	0.7891	0.8582	1.2435	0.8658	0.8658	0.8657	0.7890	0.7891	0.7890	0.8576	0.8576	0.8577	1.2435	1.2426	1.2428
	Cost (mean)	−2.1599	−2.1516	−2.0759	−2.1594	−2.8265	−2.1518	−2.1518	−2.1517	−2.0759	−2.0759	−2.0760	−2.1588	−2.1587	−2.1588	−2.8268	−2.8252	−2.8256
St.dev	Ttime (mean)	0.0460	0.0986	0.1544	0.2493	1.0475	0.0985	0.0985	0.0984	0.1543	0.1543	0.1544	0.2490	0.2489	0.2489	1.0503	1.0485	1.0494
	Ttime (st.dev.)	0.0460	0.1361	0.1659	0.3425	0.6152	0.1360	0.1359	0.1358	0.1658	0.1659	0.1660	0.3420	0.3420	0.3420	0.6177	0.6159	0.6164
	Cost (mean)	0.0712	0.1879	0.2115	0.4233	1.4848	0.1877	0.1876	0.1875	0.2112	0.2112	0.2114	0.4227	0.4227	0.4227	1.4892	1.4865	1.4885
CV	Ttime (mean)	−0.0375	−0.0804	−0.1343	−0.2013	−0.6911	−0.0803	−0.0803	−0.0803	−0.1342	−0.1342	−0.1343	−0.2010	−0.2010	−0.2010	−0.6928	−0.6921	−0.6927
	Ttime (st.dev.)	0.0543	0.1573	0.2102	0.3991	0.4948	0.1571	0.1570	0.1569	0.2101	0.2102	0.2104	0.3987	0.3988	0.3987	0.4967	0.4957	0.4960
	Cost (mean)	−0.0330	−0.0873	−0.1019	−0.1960	−0.5253	−0.0872	−0.0872	−0.0872	−0.1017	−0.1017	−0.1018	−0.1958	−0.1958	−0.1958	−0.5268	−0.5261	−0.5268

Table 5 Maximum and minimum *t*-test of the estimated parameters over 10 samples

		ML1	ML2	ML3	ML4	ML5	ML6	ML7	ML8	ML9	ML10	ML11	ML12	ML13	ML14	ML15	ML16	ML17
No. of repeated obs.		1	1	1	1	1	2	4	5	2	4	5	2	4	5	2	4	5
No. of individuals		5000	1000	500	250	125	1000			500			250			125		
Max *t*-test	Ttime (mean)	−24.57	−10.75	−6.82	−4.48	−2.85	−15.21	−21.51	−24.05	−9.64	−13.63	−15.24	−6.34	−8.97	−10.03	−4.03	−5.70	−6.38
	Ttime (st.dev.)	15.77	7.98	5.13	3.78	3.12	11.28	15.94	17.83	7.24	10.25	11.46	5.36	7.58	8.47	4.42	6.24	6.99
	Cost (mean)	−27.28	−11.62	−7.61	−4.56	−2.96	−16.42	−23.23	−25.98	−10.76	−15.20	−17.00	−6.46	−9.13	−10.21	−4.16	−5.88	−6.58
Min *t*-test	Ttime (mean)	−26.60	−12.46	−9.21	−6.73	−4.20	−17.63	−24.93	−27.87	−13.04	−18.43	−20.61	−9.53	−13.47	−15.06	−5.94	−8.39	−9.39
	Ttime (st.dev.)	13.94	5.09	3.79	1.89	1.40	7.22	10.22	11.43	5.36	7.57	8.47	2.67	3.77	4.24	1.97	2.79	3.12
	Cost (mean)	−30.26	−14.16	−11.18	−7.44	−6.45	−20.02	−28.32	−31.67	−15.80	−22.34	−24.99	−10.53	−14.91	−16.67	−9.12	−12.88	−14.41

analyses computed on the Fisher information matrix. The D-error decreases with sample size, varying from 0.128 for the samples with 125 observations to 0.0052 for those with 5000 observations. The model capability to recover the true parameters also depends on the sample size. Table 6 shows the scale factors computed for each estimated parameter (i.e. the means and standard deviations over the 10 samples). As can be seen, while the parameters are recovered correctly on average (i.e. the ratios between the estimated and true parameters averaged over the 10 samples are similar to the true scale factor used to generate the EV1 error), the standard deviation of the scale factors over 10 samples clearly increase as sample size decreases, indicating that the number of cases where the true values are not correctly recovered increases. In fact, for a sample size of 5000 individuals, all estimated parameters in the 10 samples (i.e. 30 parameters) are between ±20% of the true values, but as the sample size decreases the number of estimated parameters falling outside this range increases: four for the sample sizes of 1,000 individuals, five for the sample sizes of 500 and 250 individuals, and 20 for the sample sizes of 125 individuals.

Models ML6 to ML17 report the results for the panel with identical observations. As can be seen, including identical observations yields better *t*-test against zero only when this increases the sample size (compare for example model ML2 with models ML6-ML8). However, if we compare models estimated with the same "sample size", but composed of different numbers of individuals and different numbers of identical observations (e.g. models ML2 vs. ML9-ML13), the *t*-tests of the panel data worsen, although the effect is not always clear. This result seems to depend on the estimation process, and maybe on the simulation involved in the estimation, because the D-error computed with the simulated data gives, as expected, the same result for a sample with 5000 observations and a sample of 1000 observations repeated 5 times each. Note that, although the data variability is the same in both types of samples (e.g. 1000 individuals with 1 observation each, and 500 individuals with 2 observations each) having identical observations does not add new information for estimation.

More importantly, including identical observations is not beneficial in terms of the model capability to recuperate the true parameters. In fact, the standard deviation of the scale factors increases as the number of repeated observations increases while keeping the sample size unchanged (compare models ML2, ML9 and ML13). This clearly means that having repeated observations does not allow one to reduce sample size.

We wish to emphasize that all the above results do not depend on the richness of the data (Cherchi and Ortúzar 2008), because all samples have the same average travel time and cost even when the repeated observations are accounted for.

The third test performed consisted in estimating a weighted utility function where each individual utility was multiplied by the number of identical trips (*R*) made during the panel period. For space reasons we do not report the results here, but will only comment them. We found that weighting the utility of each individual by *R* gives exactly the same statistical results as estimating a model with *N* individuals and *R* repeated identical observations. However, as expected, the capability to reproduce the true parameters improves, because weighting individual utilities is equivalent to increasing the weight of the systematic component of utility over the random part.

Table 6 Mean and standard deviation of the scale factor (λ) over 10 samples

		ML1	ML2	ML3	ML4	ML5	ML6	ML7	ML8	ML9	ML10	ML11	ML12	ML13	ML14	ML15	ML16	ML17
No. of repeated obs.		1	1	1	1	1	2	4	5	2	4	5	2	4	5	2	4	5
No. of individuals		5000	1000	500	250	125	1000			500			250			125		
Mean (λ)	Ttime (mean)	1.42	1.41	1.34	1.40	1.78	1.41	1.41	1.41	1.34	1.34	1.34	1.40	1.40	1.40	1.78	1.78	1.78
	Ttime (st.dev.)	1.54	1.57	1.45	1.53	2.31	1.57	1.57	1.57	1.44	1.44	1.44	1.55	1.56	1.56	2.26	2.26	2.26
	Cost (mean)	1.44	1.43	1.38	1.44	1.88	1.43	1.43	1.43	1.38	1.38	1.38	1.44	1.44	1.44	1.88	1.88	1.88
St.dev (λ)	Ttime (mean)	0.05	0.10	0.16	0.28	1.23	0.10	0.10	0.10	0.16	0.16	0.16	0.28	0.28	0.28	1.23	1.23	1.23
	Ttime (st.dev.)	0.08	0.24	0.30	0.59	1.18	0.23	0.23	0.23	0.31	0.31	0.31	0.61	0.61	0.61	1.10	1.10	1.10
	Cost (mean)	0.05	0.13	0.14	0.28	0.99	0.13	0.13	0.13	0.14	0.14	0.14	0.28	0.28	0.28	0.99	0.99	0.99

When simulated data are used, it is possible to test the capability of the ML to reproduce the true phenomenon using the marginal square error (MSE) computed for each estimated model:

$$MSE = \left(\mu\left(\widehat{\beta}\right) - \beta^{true}\right)^2 + \sigma^2\left(\widehat{\beta}\right) \tag{11}$$

where $\mu\left(\widehat{\beta}\right)$ and $\sigma^2\left(\widehat{\beta}\right)$ are the estimated mean and variance of each parameter.

The MSE tends to zero as the estimated mean becomes equal to the true mean and its standard deviation equals zero. This test allows verifying whether the estimated means and standard errors are close enough to the true parameters. However, as all parameters estimated in a discrete choice model are scaled by an unknown factor (λ) proportional to the inverse of the standard deviation, the MSE was also computed de-scaling the estimated parameters. Results for the de-scaled MSE are reported in Table 7. In line with the above comments, the MSE values (especially the de-scaled ones) are fairly close to zero for almost all models up to a sample size of 500 individuals (ML1-ML3), but increase drastically for smaller size of samples with repeated identical observations, no matter how many repeated observations are used for each individual.

5 Conclusions

We have analysed the problem of how to model panel data in the presence of repeated observations and to which extent repeated observations affect the efficiency of the estimated parameters and their capability to reproduce the true phenomenon. In particular, we tested the impact of sample size and repeated observations in a "short-survey-panel" context. The effect of the repeated observations was tested using both real data from the *Santiago Panel* and synthetic data generated especially with the aim of complementing the real case.

As in previous works, we were able to confirm that the largest improvement in overall model statistics is due to the panel correlation contribution. Considering the different ways to introduce heterogeneity discussed above, we believe that it is crucial to analyse empirically (i.e. for each application case) the most appropriate ways to accommodate heterogeneity.

In our particular case, and also in the case recently studied by Hess and Rose (2009), the effect of inter-respondent heterogeneity is dominant. In fact, this constitutes empirical evidence of the advantages of the panel approach, as it allows the inclusion of correlation among the answers from the same individual, which as we said in the previous paragraph, is normally responsible for a large increase in likelihood.

Our empirical results for real data show that the inclusion of intra-respondent heterogeneity demands more observations, which means that the repeated observations can affect the definition of model structure. Therefore, we could say that a potential benefit of considering a longer multi-day survey in a short panel context is the highest probability to capture the different kinds of heterogeneity among observations.

The results from the synthetic data show that having repeated observations in a data panel increases the efficiency of the estimated parameters only because this

Table 7 Mean and standard deviation of the de-scaled MSE test over 10 samples

		ML1	ML2	ML3	ML4	ML5	ML6	ML7	ML8	ML9	ML10	ML11	ML12	ML13	ML14	ML15	ML16	ML17
No. of repeated obs.		1	1	1	1	1	2	4	5	2	4	5	2	4	5	2	4	5
No. of individuals		5000	1000	500	250	125	1000			500			250			125		
Mean *t*-test	Ttime (mean)	0.003	0.017	0.039	0.170	1.401	0.009	0.005	0.004	0.020	0.011	0.009	0.085	0.043	0.035	0.719	0.354	0.286
	Ttime (st.dev.)	0.004	0.025	0.055	0.240	0.751	0.013	0.006	0.005	0.028	0.014	0.011	0.120	0.060	0.048	0.391	0.189	0.153
	Cost (mean)	0.007	0.044	0.084	0.487	3.249	0.022	0.011	0.009	0.042	0.021	0.017	0.243	0.122	0.097	1.662	0.819	0.660
St.dev *t*-test	Ttime (mean)	0.002	0.008	0.013	0.023	0.036	0.004	0.002	0.002	0.007	0.003	0.000	0.011	0.006	0.005	0.018	0.009	0.008
	Ttime (st.dev.)	0.003	0.015	0.023	0.049	0.056	0.007	0.004	0.003	0.012	0.006	0.000	0.024	0.012	0.010	0.028	0.014	0.011
	Cost (mean)	0.005	0.021	0.031	0.057	0.079	0.011	0.005	0.004	0.016	0.008	0.000	0.028	0.014	0.011	0.040	0.020	0.016

increases the sample dimensions. Therefore, based on our results from real and synthetic data, we can say that there is a trade-off between the higher probability of capturing more effects (i.e. different types of heterocedasticity) in a longer multi-day-panel sample and the risk of a decreased capability of reproducing the true phenomenon (as this worsens in the presence of repeated observations).

Finally, our suggestion on the definition of a multi-day-panel survey length would be to consider not only the number of individuals, but also the level of routine expected. This last factor seems to be especially important in "short panel" surveys, as they commonly feature a large proportion of identical observations, which are actually harmful; we proved that they reduce the capability of reproducing the true phenomenon. Thus, even though having more observations per respondent requires smaller sample sizes in order to establish the statistical significance of the parameter estimates derived from choice data (Rose et al. 2009), our results show that this is effectively true if and only if the level of routine is not strong.

Acknowledgements We wish to thank the funding of the Chilean Fund for Scientific and Technological Research, FONDECYT (Project 1050672) and of the Millennium Institute on Complex Engineering Systems (Project P05-004F). The last version of this paper was written when the first author was visiting Ph.D. student at the German Institute for Economic Research (DIW-Berlin), financed by the Chilean Council for Scientific and Technological Research (CONICYT) and the German Academic Exchange Service (DAAD), the second author was visiting the Pontificia Universidad Católica de Chile funded by the post-doctoral visit programme of their School of Engineering, and the last author was Visiting Professor at the Centre for Transport Studies, University College London under a research grant from the UK Science and Engineering Research Council.

References

Abdel-Aty MA, Kitamura R, Jovanis PP (1997) Using stated preference data for studying the effect of advanced traffic information on driver's route choice. Transp Res 5C:39–50

Ampt ES, de Dios Ortúzar J (2004) On best practice in continuous large-scale mobility surveys. Transport Rev 24:337–363

Axhausen KW, Zimmermann A, Schönfelder S, Rindsfüser G, Haupt T (2002) Observing the rhythms of daily life: a six-week travel diary. Transportation 29:95–124

Axhausen KW, Löchl M, Schlich R, Buhl T, Widmer P (2007) Fatigue in long-duration travel diaries. Transportation 34:143–160

Bierlaire M (2003) BIOGEME: A free package for the estimation of discrete choice models. *3rd Swiss Transport Research Conference*, Monte Verit, Ascona, Switzerland.

Bliemer MCJ, Rose JM (2010) Constructing of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research B* (forthcoming).

Cantillo V, de Dios Ortúzar J, Williams HCWL (2007) Modelling discrete choices in the presence of inertia and serial correlation. Transportation Science 41:195–205

Cherchi E, Cirillo C (2008) A modal mixed logit choice model on panel data: accounting for systematic and random heterogeneity in preferences and tastes. *86th Seminar of the Transportation Research Board.* Washington D.C., USA.

Cherchi E, de Dios Ortúzar J (2007) On the efficiency of mixed logit parameters estimates: analysing the effect of data richness. *Proceedings XIII Congreso Chileno de Ingeniería de Transporte*. Santiago, Chile (*on CD*).

Cherchi E, de Dios Ortúzar J (2008) Empirical identification in the mixed logit model: analysing the effect of data richness. Networks Spatial Econ 8:109–124

Cherchi E, Polak JW (2005) The assessment of user benefits using discrete choice models: implications of specification errors under random taste heterogeneity. Transp Res Rec 1926:61–69

Cirillo C, Axhausen KW (2006) Dynamic model of activity-type choice and scheduling. *Proceedings European Transport Conference*, Strasbourg France.

Cirillo C, Daly AJ, Lindveld K (2000) Eliminating bias due to the repeated measurements problem in SP data. In: de Dios Ortúzar J (ed), *Stated preference modelling techniques*. Perspectives 4, PTRC, London.

de Dios Ortúzar J, Willumsen LG (2001) Modelling transport, 3rd edn. Wiley, Chichester

de Dios Ortúzar J, DA Roncagliolo, UC Velarde (2000) Interactions and independence in stated preference modelling. In: de Dios Ortúzar J (ed), *Stated preference modelling techniques*. Perspectives 4, PTRC, London.

DICTUC (2003) *Encuesta de Movilidad 2001 de Santiago*. Informe final para el Ministerio de Planificación. Departamento de Ingeniería de Transporte y Logística, Pontificia Universidad Católica de Chile, Santiago.

Galbraith RA, Hensher DA (1982) Intra-metropolitan transferability of mode choice models. J Transport Econ Pol 16:7–29

Hess S, Rose JM (2009) Allowing for intra-respondent variations in coefficients estimated on repeated choice data. Transp Res 43B:708–719

Hess S, Train K (2009) Approximation issues in simulation-based estimation of random coefficient models. *Mimeo*, Institute for Transport Studies, University of Leeds.

McFadden D (1974) Conditional logit analysis of qualitative choice behaviour. In: Zarembka P (ed) Frontiers in econometrics. Academic, New York, pp 105–142

Mobidrive (2000) Mobidrive questionnaires. *Arbeitsberichte Verkehrs und Raumplanung*, 52, Institut für Ver-kehrsplanung, Transporttechnik, Strassen und Eisenbahnbau, ETH Zürich.

Revelt D, Train K (1998) Mixed logit with repeated choices: households' choices of appliance efficiency level. Rev Econ Stat 80:647–657

Rose JM, Bliemer MCJ (2008) Stated preference experimental design strategies. In: Hensher DA, Button KJ (eds) Handbook of transport modelling. Elsevier, Oxford

Rose JM, Bliemer MCJ (2009) Constructing efficient choice experimental designs. *Transport Rev* 29 (in press).

Rose JM, Hess S, Bliemer MCJ, Daly AJ (2009) The impact of varying the number of repeated choice observations on the mixed multinomial logit model. European Transport Conference, Leiden

Ruiz T (2006) Attrition in transport panels. In: Stopher P, Stecher C (eds) Travel survey methods: quality and future directions. Elsevier, Amsterdam

Sillano M, de Dios Ortúzar J (2005) Willingness-to-pay estimation with mixed logit models: some new evidence. Environ Plan 37A:525–550

Train KE (2003) Discrete choice methods with simulation. Cambridge University Press, Cambridge

Van Wisen LJG, Meurs HJ (1989) The Dutch mobility panel: experiences and evaluation. Transportation 16:99–119

Walker J (2001) Extended discrete choice models: Integrated framework, flexible error structures, and latent variables. Ph.D. Thesis, Center for Transport Studies, MIT.

Williams HCWL, de Dios Ortúzar J (1982) Behavioural theories of dispersion and the mis-specification of travel demand models. Transp Res 16B:167–219

Yáñez MF, Cherchi E, de Dios Ortúzar J, Heydecker B (2009) Inertia and shock effects over mode choice process: implications of the Transantiago implementation. *Transp Sci* (under revision).

Yáñez MF, de Dios Ortúzar J (2009) Modelling choice in a changing environment: assessing the shock effects of a new transport system. In: Hess S, Daly A (eds) Choice modelling: the state-of-the-art and the state-of-practice, Proceedings from the Inaugural International Choice Modelling Conference. Emerald, Bingley

Yáñez MF, Mansilla P, de Dios Ortúzar J (2010) The Santiago panel: measuring the effects of implementing Transantiago. *Transportation* 37. doi:10.1007/s11116-009-9223-y.

Yen J, Mahmassani H, Herman R (1998) A model of employee participation in telecommuting programs based on stated preference data. In: de Dios Ortúzar J, Hensher DA, Jara-Díaz SR (eds) Travel behaviour research: updating the state of play. Pergamon, Oxford