Internet mobility survey sampling biases in measuring frequency of use of transport modes

Internet Mobility Survey Sampling Biases in Measuring Frequency of Use of Transport Modes

Marco Diana

28th February 2012

INTERNET MOBILITY SURVEY SAMPLING BIASES IN MEASURING FREQUENCY OF USE OF TRANSPORT MODES

Marco Diana*

Dipartimento di Ingegneria dell'Ambiente, del Territorio e delle Infrastrutture
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino – ITALY
Phone: +39 011 090 5638 - Fax: +39 011 090 5699
marco.diana@polito.it

Manuscript submitted for the 91st TRB annual meeting
Submission date: July 21, 2011 – Revised November 9, 2011 and February 27, 2012
Word count, excluding tables and captions: 6698
Number of tables and figures: 4 + 2

* *Corresponding author*

ABSTRACT

We develop a quantitative analysis of the biases that arise when measuring trip frequencies for a general population through an online survey instrument. Data from a national official survey in Italy, concerning both mobility behaviors and skills in using computers and internet, have been deployed to assess differences in mobility levels between those that can answer a computer/internet survey and those that cannot. Positive correlations were found between ability in using ICT tools and trip frequencies. These latter are about 15% to 150% higher for the "ICT literate", according to the travel means under consideration. A Heckman sample selection model showed us that these biases have different explanations. People knowing how to use internet are different from the others in they car driving behavior due to a range of self-related factors. Conversely, public transport patterns of use are more similar between the two groups: the observed bias is mainly due to the fact of using internet in itself, which could for example lead to a more active lifestyle. Such distinction is of practical interest because it can help defining a method to correct these biases. According to our results, the overestimation of public transport frequency of use of an internet survey could be corrected by looking at the internet diffusion in the population. On the contrary, corrections for car driving frequencies are more complex and should be based on differences in attitudinal and personal characteristics between internet survey respondents and the remainder of the population.

KEYWORDS

Internet surveys, Trip frequencies, Heckman model, sample self-selection

INTRODUCTION

Internet mobility surveys represent a data collection method that is attractive under several points of view. Low unit costs, flexibility in the administration of the questionnaire, real-time automatic data quality checks and continuous monitoring of the collection processes are only some of the advantages that are making this tool more and more popular among researchers. However, Information and Communication Technologies (ICT) are still not enough diffused in the population to draw a representative sample from those being able to answer a web survey. Related measurement biases make therefore difficult to infer mobility figures for a general population from an online survey, so that this tool can be practically used in limited ambits, i.e. to survey specific population segments (i.e. students) or as a complement to other surveying protocols (*1-3*). In view of the above mentioned large benefits of web surveys, there is thus an interest among researchers and practitioners to better understand the magnitude of such biases, how they arise and how they can be controlled for, in order to increase the use of this tool.

Many papers in the last decade deal with internet mobility surveys biases. Arentze et al. (*4*) developed secondary analyses on three different datasets from specific behavioral surveys to investigate sampling biases mainly concerning socioeconomic characteristics. Bricka and Zmud (*5*) studied if and how a web survey could somewhat compensate the fact that less active and mobile population segments tend to be over-represented by phone surveys. They compared mobility figures of those that answered to the same survey over the phone and through internet and found that the latter reported higher trip rates, even if the considered sample sizes for web surveys were rather small. On the other hand, Bayart and Bonnel (*6*) found that face-to-face survey respondents reported higher trip rates but less trips by car compared to internet survey respondents, even if these results could be affected by the fact that the web survey was an option only to those not reachable or that refused a personal interview. Bonsall and Shires (*7*, *8*) investigate among other things the discrepancies between self-administered paper and internet surveys concerning sampling biases and both stated and revealed preferences, although their study targeted a specific social group, namely Directors of Human Resources of firms with 10 or more employees.

These researches give some initial insight on the issue, but results are somewhat partial since they are mainly based on the comparison of two mobility-related surveys that are administered in different ways and that sometimes have a specific scope or are addressed to only part of the population. In the following, we try to give a wider picture by using a different approach. In particular, we consider a single national survey that sampled the whole population and that investigated key figures concerning both mobility behaviors and levels of use and skills related to ICT, through personal interviews and self-compiled questionnaires. Considering that some surveying modes can be used only with some more or less wide population subgroups (i.e. one needs to have a telephone for phone interviews, or needs to have a computer and be enough proficient in its use to answer a web survey), it becomes possible to use such dataset to study how mobility figures change among those different subgroups.

In a preliminary work following this approach, internet mobility survey biases were investigated by exploiting non-metric variables through correspondence analyses (*9*). It has therefore been possible to qualitatively delineate how vehicle ownership and the patterns of use of different travel means are affected when considering population subgroups characterized by different levels of access to land line phones, mobile phones, computers and internet. It was noted that the fraction of the population that is likely to answer an internet survey has more cars than the average. Concerning other travel means, an underestimation of

the intensity of use of feet and cars as a passenger and an overestimation of the use of public transport and of multimodality behaviors was observed for this subgroup.

These results allowed defining some guidelines to limit sampling biases due to the survey mode, at least for some specific mobility figures. However, a quantitative estimation of such errors was still missing. In this paper we show how we can achieve this by using ordinal-scaled data related to the frequency of use of different means in terms of number of trips. Measuring the bias induced by an internet survey has an obvious immediate interest to better understand to what extent data quality is affected. However, it also constitutes the first step to envisage correction procedures for such biases, once these latter are adequately modeled. This more ambitious research goal will be pursued in the final part of the paper, where a sample selection model will be introduced in order to understand the likely causes of the differences in mobility levels between potential internet survey respondents and the remainder of the population. On the basis of the estimation results of these models, different strategies to treat such sampling biases will be delineated.

DATASET AND EXPERIMENTAL FRAMEWORK

We use the 2007 dataset from the "Aspects of everyday life" survey, a multipurpose data collection effort yearly administered by the Italian National Statistical Institute ISTAT to a stratified unequal probability sample of about 50,000 inhabitants that is representative of the whole population (*10*). The questionnaire is divided into three main parts: a face-to-face survey related to the whole household, a face-to-face survey administered to each individual member (or a proxy for children) and a paper and pencil individual questionnaire for each member of the household that can be either self-compiled or compiled with the help of the interviewee. The purpose of this annual survey is in broad terms to investigate habits and opinions of people on a variety of ambits, ranging from public services use to health conditions, quality of life or social inclusion. We do not report here any socioeconomic characterization of the sample, since it is representative of the whole Italian population for which official statistics are widely available.

In the present work, we essentially consider information from this survey that is related to the individual levels of mobility (in terms of trip frequency) through urban public transport, trains and car driving, matching it to the ability to perform different tasks with a personal computer and over the internet, that is represented by a set of later introduced binary variables. For the sake of briefness, the variables that we use in the remainder of the paper are presented more in detail along with the analyses that we perform. They are either directly contained in the publicly available distribution of the above dataset, or have been derived with some simple manipulation, such as the aggregation of some categories. Many of these variables are related to questions that were not asked to the entire sample; for example, the frequency of car driving is obviously surveyed only among people aged 18 or more, which is the minimum driving age in Italy. Therefore, the definition of the universe of individuals changes according to the variables being considered. In the following, we therefore define each time the universe to which we are referring any given analysis that we present.

QUANTITATIVE ANALYSIS OF THE BIASES

Correlation between ICT skills and modal usages

At the outset, we need a method to measure ICT skills of the Italian population. In our dataset, the proficiency of the respondents concerning computer and internet usage was measured

through eleven binary variables COMP_1, …, COMP_8 and WEB_1, …, WEB_3, each variable pertaining to a specific task that is specified in Table 1. Not all tasks have the same level of complexity. We assume that, considering any pair of variables from either the first or the second set, the variable with smaller proportion of people that declared being able to perform the related task identifies the most difficult task between the two. With such assumption, it is therefore possible to aggregate the responses of each individual into a single score, and to consider this score as a measure of the overall ability of the subject in using the computer (if COMP_1, …, COMP_8 responses are aggregated) or using internet (if we aggregate the remaining three). In the following we respectively call these two scores MU_COMP and MU_WEB. The method that we follow to perform such aggregation is detailed elsewhere (*9*) and is based on the computation of rather simple non-parametric multivariate statistics that were formerly introduced by Wittkowski et al. (*11*).

Table 1 about here

Having defined a measurement method for ICT skills, it is interesting to compute correlations between the two scores MU_COMP/MU_WEB and the frequencies of use of the three transport modes that are considered in the survey (Table 2). These frequencies were measured through ordinal variables whose categories are the following four: “Everyday”, “Sometimes a week”, “Sometimes a month”, “Sometimes a year” and “Never”. Therefore, such responses were then transformed into an annual trip rate per individual by respectively taking the following values: 300, 100, 20, 5 and 0. Of course this latter metric variable is only a rough estimation of the true number of trips that are yearly made with these modes. However, we believe that such approximation is not significantly affecting the results of this correlation analysis. When assigning different frequency values, we transform only one variable in each pair that is jointly considered to compute each correlation. Furthermore, such transformation is likely to be monotonic, unless we postulate a relationship between the way respondents translated their true trip frequency into a categorical response and their proficiency in using a computer or internet.

Table 2 about here

All these correlations are positive and significant, also given the sample size, and they become stronger when moving from urban public transport to train to car driving. We also notice that these values are slightly higher for MU_WEB than for MU_COMP when considering urban public transport and train modes, whereas the opposite is true concerning car driving. However, differences in values between rows are rather small.

Overall, these positive correlations are an indication of the fact that both the frequencies and the intensities of use of these modes are increasing when we consider more skilled individuals concerning computer and internet use. The following step is then to try to quantitatively assess the actual differences in the number of trips made through these three means between those that can answer to a computer and/or an internet self-administered survey and those that cannot. These differences could in fact be considered as a proxy of the biases induced by a survey implemented only through these specific means.

Identifying population subgroups that can answer a self-administered computer or online survey

In order to discriminate between those that could answer a computer and an internet surveys and those that are not able to do it, we set an “ability threshold” on the basis of the above

ability scores. In particular, we assume that an individual is able to answer a computer survey if the corresponding MU_COMP score is at least equal to the score of those survey respondents that answered "yes" only to COMP_1 and COMP_2, that are the two questions related to the two easiest tasks in computer use among those considered. We similarly consider that an individual is able to answer a survey over internet if the corresponding MU_WEB score is at least equal to the score of those survey respondents that answered "yes" to WEB_1, WEB_2 and WEB_3, since all these three represent the basic skills that are needed to proficiently use internet. Of course, a sufficient proficiency in using ICT is only a necessary condition to truly complete an online questionnaire, so that our approach can be viewed as a first attempt to define the target population subgroup for this survey instrument.

Therefore, we define the two subgroups of the Italian population whose scores are above these thresholds and we study how mobility figures of those two subsamples are different from those of the whole sample. We will shortly refer to those subsamples as "computer subsample" (CS, N = 21,984,751 out of 57,029,046 individuals of the whole Italian population aged 3 or more) and "web subsample" (WS, N = 11,380,591 out of 55,398,010 individuals of the whole Italian population aged 6 or more). By looking at the different cardinality of CS and WS we can conclude that the latter subset is more selective concerning ICT proficiency levels in general terms, since we assume that abilities in using computers and internet are rather correlated albeit distinct concepts. Such difference will allow us to have a deeper understanding on how mobility figures are affected by an increasing familiarity with ICT technologies.

It is insightful to briefly look at differences in socioeconomic characteristics of CS and WS compared to the whole Italian population (N = 58,729,564 individuals). The following categories tend to be over-represented in both subgroups: males (55.6% in CS and 58.6% in WS, against 48.6% in the whole population), those aged between 10 and 45 (Figure 1), those having at least a high school diploma (Figure 2 top left), workers and students (Figure 2 top right), those living with their parents (Figure 2 bottom left) and in households where at least two cars are available (Figure 2 bottom right). We also looked at the proportions of individuals living in different built environments (metropolitan city centers, metropolitan suburbs, larger cities, smaller towns etc.) but these are quite well preserved in our subsamples. Therefore, one important preliminary finding is that studies primarily dealing with land use impacts on travel behaviors should not be affected when using web-based samples.

Figures 1 and 2 about here

These biases in CS and WS are somewhat expected, and can give us a sense of the risks of implementing mobility surveys only involving those subgroups. In particular, concerning employment status, we systematically over-represent more mobile population segments (workers and students) and under-represent less mobile ones (housewives, retired). Since WS is more "selective" than CS we would expect that the former is even more problematic than the latter. This is generally true for the considered socioeconomic variables, but remarkably not for car ownership. In the following, we will investigate if this same finding applies also to other mobility-related figures.

Survey biases concerning trip frequencies by mode

We look at biases related to the number of trips yearly made by train, by urban public transport and driving a car, since the dataset that we used included specific questions on those three transport modes. To run a quantitative analysis involving trip frequencies we have to overcome one practical difficulty, since trip rates can be estimated only approximately in our

dataset, as we made for the above correlation analysis. In order to control for such approximation in our data that could now bias our results, we run a sensitivity analysis by considering four different trip frequencies for each of the first four categories of these three variables. Therefore, we consider that the actual trip frequency for those that answered "Everyday" could be 700 (for those making almost two trips per day), 500, 365 or 250 trips/year (i.e. one trip per working day). Along the same lines, "Sometimes a week" translates into 200, 150, 100 or 60 trips/year; "Sometimes a month" becomes 60, 40, 30 0r 20 trips/year and "Sometimes a year" is coded 20, 15, 10 or 5 trips/year. The fifth category of these three variables is "Never" and the corresponding frequency is always taken equal to zero. Therefore, we generate 4 exp (4) = 256 different scenarios by considering all the combinations of levels for the different categories.

Coming to the presentation of the results, some key mobility figures concerning the intensity of use of different means are presented in Table 3. The nine columns of numbers report the results relative to the three considered travel modes for the whole sample, for CS and for WS. The three rows report the number of trips per year per person for the entire population (i.e. considering both those that use and those that do not use the travel means under consideration), the percentage of respondents that declared using the given mode at least sometimes per year, and the average trip rates per person of only these latter. The numbers that are shown in the first and in the third row are average values of the above mentioned 256 scenarios of the sensitivity analysis. The actual values for the different scenarios are rather dispersed around such means, given the wide variation in the considered trip frequencies within each category. However, we recall that the goal of the analysis is not to give an estimation of the absolute values of such mobility figures, but rather to estimate their sampling biases when considering our subsamples. Therefore, we also show in the table the minimum and maximum relative differences (in percent) between the mobility figures of the subsamples and the corresponding value of the general sample. These extreme values have been found by considering all the possible combinations of frequency values from our 256 scenarios, so that the true difference is surely in between this interval. Reporting such range of possible biases is more insightful than considering only the mean values, in order to see how our results could be affected by the actual trip frequencies that we judgmentally assigned.

Table 3 about here

In general terms, it can be seen that the two subsamples have mobility figures that are rather similar between them but always higher than those of the general sample. On the other hand, relative error ranges (shown in italics in Table 3) are surely appreciable according to our sensitivity analysis but not so wide, so that we can infer that our findings and the subsequent interpretation of the results are not seriously affected by the actual trip frequencies that we judgmentally assigned to the different categories.

Numerical results are consistent with the correlation values of Table 2: we observe that biases are higher for WS than for CS when considering urban public transport and train, whereas the opposite is true for the car driving mode. By comparing the results in the three lines of the table it is also apparent that the bias in terms of trips per year (first line) is largely due to the fact that subsamples overestimate the percent of users of the mode (second line). Therefore, we see that the measurement biases of trip rates of mode users are much smaller (third line). In particular, it seems that the frequency of car driving of those that drive at least sometimes a year can be rather accurately predicted only considering WS, a remarkable finding considering the fact that both the percentage of drivers and car ownership levels are higher for this group than for the general population. In other words, a higher share of car

drivers and a higher availability of cars in WS compared to the whole population must be cancelled out by the fact that drivers belonging to WS use less such transport means than the average driver, so that the number of car driving trips per year for car drivers is not significantly affected. These counter-balancing effects concerning mobility behaviors must be carefully considered, in order to avoid overlooking underlying biases that are not apparent if one only considers some specific mobility figures.

A SAMPLE SELECTION MODEL TO STUDY INTERNET SURVEY BIASES

Theoretical framework and model definition

Beyond quantifying through descriptive statistics the measurement error of self-administered internet surveys, we believe it is important to give an interpretation of these results that can support both researchers and practitioners in finding appropriate corrective action, in order to make such survey instrument practically useful. In this section we therefore try to explain the increase in trip frequencies that we observed among the most skilled internet users through an appropriate model. We restrict our analysis to the WS subgroup, given the more practical interest of implementing self-administered mobility surveys on the web rather than merely using a computer.

Our theoretical modeling framework is as follows. We would like to understand to what extent the observed increase in trip frequencies in the WS subsample is given by the fact that the ability in using internet increases the demand for trips (e.g. because it enlarges social relationships, increases the stimulus in visiting different places etc.), and to what extent such increase is due to a sample self-selection mechanism, since individuals belonging to WS are different from the others. Self-selection is due to the fact that individuals are not randomly assigned to the WS subsample, so that we do not truly have a treatment and a control group with homogeneous individuals, like in classic scientific experiments. For example, people belonging to WS could be better educated and have higher revenues, thus traveling more, independently on the fact that they can also use internet.

Studying how these two intertwined effects come into play is useful to understand the true nature of the bias in internet mobility surveys, and therefore to propose appropriate corrective actions. One possibility is that the bias that we observed could be mostly due to the "treatment", i.e. the ability in using internet, with a limited influence given by the self-selection mechanism of the WS sample. In this case, there would be a stronger relationship between ICT skills and trip frequencies, so that the bias in trip frequencies induced by an internet survey can be corrected by considering the ICT diffusion in the population. Conversely, if the self-selection mechanism is prevalent and the "treatment" has a limited influence on mobility levels, then the gap between the WS subsample and the remainder of the population concerning mobility levels should be explained by investigating the causes of such mechanism. Therefore, internet mobility survey biases should probably be corrected by looking at the differences between WS subsample and population, that is probably a more challenging and articulated task.

An appropriate method to study this problem is the definition of a sample selection model (*12*) that is also called Heckman or Heckit model, switching regression model or Tobit-5 model in the literature. Applying such model in our case, we have two regression equations, one for the WS subgroup (superscript "1") and the other for the remainder of the population (superscript "0"), whose endogenous variable Y is the trip frequency for the mode under investigation. Both equations have the same specification, i.e. the same set of explanatory

variables *X*. Denoting with β the regression coefficients and with *U* the error terms, we define these two *outcome equations* as follows:

$$Y^0 = X\beta^0 + U^0, \qquad Y^1 = X\beta^1 + U^1$$

For each individual we observe either Y^0 or Y^1, but individuals are not randomly assigned to the WS subgroup. Therefore, we define a binary indicator variable *D* that is set to one if the subject belongs to WS and zero otherwise. In turn, *D* is set to one when a latent (unobserved) variable *D** is greater than or equal to zero:

$$D^* = Z\theta + U^D$$
$$D = 1 \text{ if } D^* \geq 0, \qquad D = 0 \text{ otherwise}$$

where *Z* is a set of explanatory variables of the *selection equation*, θ are the corresponding coefficients and U^D the error terms. Assuming that the error terms in the above three equations have a 3-dimensional normal distribution, the above selection equation is a binary probit model and it can be jointly estimated with the two outcome equations through either maximum likelihood or a two-step procedure. We refer the interested reader to Heckman (*12*) for details on the estimation procedure; here we only report that at least one exogenous variable in *Z* should not belong also to *X* in order to avoid identification problems. Such exclusion restriction is hopefully easy to fulfill in our case, as it will be pointed out when presenting the model specification.

The possibility of computing some treatment parameters on the basis of the model estimation results is one of the most attractive features of this tool for practitioners. In particular, Heckman et al. (*13*) introduce the *average treatment effect* (ATE) and the *effect of treatment on the treated* (TT). In our framework, ATE can be defined as the average variation in trip frequencies when a randomly selected individual learns to use internet (which is our "treatment"), whereas TT is the average variation in trip frequencies when an individual belonging to WS learned to use internet. Therefore, if ATE = TT then we are in the case shown in Table 3, since no self-selection is observable and WS is a random subsample. Conversely, if TT tends to zero with ATE being different from zero, then the ability to browse through the web has a limited effect on trip frequencies, and the observed bias is mostly due to self-selection. In the more general case, the effect of self-selection is given by the difference between TT and ATE, that can either reinforce the treatment effect when TT>ATE (*14*) or attenuate it when TT<ATE (e.g. in the application discussed in *13*). For the sake of briefness we do not report here the computational procedure to derive ATE and TT, which is fully spelt out elsewhere (*13*).

Model specification and results

Concerning the model specification, we consider six different pairs of outcome equations, thus originating a set of models numbered from 1 to 6, since our aim is to study trip frequencies for both the general population (models 1-2-3) and for the mode users (models 4-5-6) through urban public transport (models 1 and 4), train (models 2 and 5) and car driving (models 3 and 6).

We resort on prior knowledge from the published literature to specify the outcome equations. There are a lot of studies on the determinants of car driving frequency, mainly in connection with car ownership (*15*), but relatively less works concerning the use of urban public transport and trains, since the standard approach in those cases is to estimate the demand for such means through a disaggregated mode choice model. However, we notice that

trip frequencies through different modes are somewhat correlated and influenced by the same factors, although the magnitude and even the direction of such influence could be different (e.g. when considering car ownership). For this reason and in order to simplify a comparative interpretation of the results, we will keep the same specification of the outcome equation across the six models. We therefore consider as exogenous variables X gender, age, occupational status, number of household vehicles, presence of children and kind of urban environment, using dummy coding for categorical variables.

Concerning the selection equation, that is obviously the same for all six models, the latent variable D^* is MU_WEB minus the threshold value that was previously introduced to define the WS subsample. The explanatory variables Z are age, educational level, the availability of an internet connection in the household and the frequency of use of internet. Among these latter, we postulate that the availability of an internet connection has no effect on the trip frequency with a given means, so that the above mentioned exclusion restriction is fulfilled. In this case, such exogenous variable is said to be a valid instrument.

We notice that income is not among the exogenous variables because it was not available in the dataset, so that we considered the occupational status as a proxy. This can be seen as a limitation of these models. However, we recall that our goal is not to understand the determinants of the levels of use of different travel means, but rather to see how these figures are affected when considering a sample affected by the selectivity bias induced by the survey instrument. Another limitation is given by the fact that the software that we use for the estimation, namely the *sampleSelection* package of R, does not offer a straightforward way to treat unequal probability samples as this one. Therefore, the following results have been obtained not considering observation weights.

Estimation results for these six models, together with their ATE and TT values, are presented in Table 4. We preliminarily note that the parameters for the selection equation are all significant and quite similar across different models, although not identical since each selection equation is jointly estimated with the corresponding outcome equations through maximum likelihood. All signs of the selection equations are those expected, so that comments on these results are not essential.

Table 4 about here

Concerning outcome equations, model estimation results are insightful on a number of issues. Model 3 coefficients are consistent with the findings of many studies on the determinants of car use (*15*, *16*). Age effect is significant and is changing sign when considering the two outcome equations. According to Figure 1, few older and probably less mobile persons are in the WS subgroup, so that the influence of the highly mobile middle-aged population segments that browse the web is probably prevailing in this case. The corresponding model 6, studying the intensity of car driving only among car users, shows similar patterns.

The use of urban public transport seems well captured by model 1, and it is instructive to notice the changes of sign of coefficients compared to model 3, that are not detailed here for the sake of briefness. No substantial differences can be pointed out between the coefficients of the two outcome equations in model 1, so that we can anticipate that the sample self-selection effect is probably not so relevant in this case. Unlike car driving models, we can see here more marked differences between the public transport frequency of use model for the general population (model 1) and the same model only for transit riders (model 4). In particular, sex, occupational status and the fact of living in a metropolitan city suburb become insignificant for those that are not able to answer an internet survey. The level of demand of transit customers is therefore influenced by different factors that are not considered

in this study. Finally, models for train use (i.e. models 2 and 5) show a greater proportion of less significant variables, and also more marked differences between the coefficients of the two outcome equations.

Beyond estimation results, for the purpose of the present study it is particularly important to focus on the treatment parameters. Model 3 has an ATE value that is much greater than TT. According to the discussion on the previous subsection, we can conclude that the WS subgroup upward bias in the frequency of car driving that is shown in Table 3 is largely due to a self-selection effect in the sample (that counts for 40.4-13.5=26.9 trips/year), and to a lesser extent to the effect of having learned to use Internet. Yet this latter effect, measured by TT, is appreciable, which can incidentally be seen as a confirmation of previous research unraveling the complementarity effect between ICT and transport demand. Therefore, an online mobility survey will give biases on the frequency of car driving primarily because the related subsample behaves differently from the population, whereas the fact of using internet in itself has a limited influence. Correcting for such biases could therefore be rather challenging.

The results of the other models can be interpreted along the same lines. As opposed to model 3, in it very interesting to see that models 1 and 2 show similar values for ATE and TT. The related bias has therefore a different origin in this case: if people that cannot answer an online survey learned to use a computer, then they would increase trip rates through public transport roughly as much as people in the WS subsample already did. Unlike the above car use model, it seems therefore possible to correct for such bias by looking at statistics on internet diffusion. On the other hand, treatment effects for model 4, compared to those of model 1, tell us that the additional increase in the frequency of use of urban public transport among its customers that are able to use internet could again be due to self-selection.

Finally, ATE of model 6 is the one nearest to zero, an expected results since the observed bias in car driving frequency from Table 3 is the smallest one (363-355=8 trips/year). The signs of both ATE and TT seem not correct, probably due to the fact that the treatment effect is not significant in this case. Concerning this point, we nevertheless observe that our treatment parameters are related but not directly comparable to the biases shown in Table 3, even if the measurement unit is the same (trips per year). Beyond the different underlying definitions, the exogenous variables of the outcome equations can in fact only partly explain the observed variance of trip frequencies. On the other hand, we did not perform a sensitivity analysis for the Heckman model based on trips frequencies, as previously done. However, the resulting approximation in the values of the endogenous variable in outcome equations should not affect our main conclusions concerning the selection mechanism.

CONCLUSIONS

This paper offered a quantitative analysis of the biases in measuring trip frequencies for a general population when using an online survey instrument. By using a dataset from personal interviews that contained information both on trip frequencies and on informatics-related skills, it has been possible to split the Italian population in two groups: those that can answer a computer (or internet) survey and those that are not able to do that. The former group has a higher trip rate for all the three considered transport modes. Positive correlations were in fact found between ability in using ICT tools and trip frequencies. Concerning car driving, upward biases are due to an overestimation of those driving cars that belong to the CS and WS subsamples, since drivers in these groups do not drive significantly more than the others.

Our sample selection model showed us that the observed biases have different origin, according to the mode under investigation. People belonging to the WS subsample are actually different from the others in they car driving behavior, beyond the fact of using internet. Conversely, public transport patterns of use of the WS subsample and of the general population are more similar, and the observed bias is mainly due to the fact of using internet, according to a complementarity effect between ICT and travel levels of use that is well documented in the literature. In this latter case, the overestimation of trip frequencies of an internet survey could more simply be corrected by looking at the internet diffusion in the population, whereas this might not be sufficient when studying car driving.

One desirable extension of the present study is to more precisely individuate those individuals that are well suited to answer internet surveys. A sufficient level of ICT proficiency is in fact only a necessary condition, that does not guarantee that those individuals would effectively complete a self-administered online questionnaire. Beyond this, the analyses that we presented could also be expanded to further explore the intertwined relationships between ICT uses or skills on one hand, and mobility patterns on the other. In particular, the ISTAT dataset contains a lot of relevant nominal and binary variables, which could not all be used in the present work, given the need of having a manageable set of variables with the selected analysis technique. It is anticipated that also other kinds of analyses could be useful in this situation, such as data mining techniques. In fact, several different non-metric variables could be jointly considered through association analysis and association rules. Finally, recalling that we use data from a continuous survey, it could be possible to add a longitudinal perspective to our study by considering several different survey waves rather than just one, in order to understand how the relationship between mobility patterns and ICT uses is evolving in more recent years.

REFERENCES

1. Alsnih, R. Characteristics of web-based surveys and applications in travel research. In: P. Stopher, C. Stecher (eds), *Travel Survey Methods: Quality and Future Directions*. Elsevier, Amsterdam, Netherlands, 2006, 569-592.
2. Bonnel, P., J.-L. Madre. New technology: Web-based, In: P. Stopher, C. Stecher (eds), *Travel Survey Methods: Quality and Future Directions*. Elsevier, Amsterdam, Netherlands, 2006, 593-603.
3. Dillman, D.A., J.D. Smyth, L.M. Christian. *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method*. Wiley, Hoboken (NJ), 2009.
4. Arentze, T., I. Bos, E. Molin, H. Timmermans. Internet-based travel surveys: selected evidence on response rates, sampling bias and reliability, *Transportmetrica*, 1(3), 2005, 193-207.
5. Bricka, S., J. Zmud. Impact of internet retrieval for reducing nonresponse in a household travel survey. Presented at the 82nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2003.
6. Bayart, C., P. Bonnel. The mixing of survey modes: application to Lyon web and face-to-face household travel survey, Proceedings of the 10th World Conference on Transport Research, Lisbon, 11-15 July 2010.
7. Bonsall, P., J. Shires. Employer expectations for commuting and business-related travel in an environment rich in information and communication technologies, *Transportation Research Record 1977*, 2006, 268-276.

8. Bonsall, P., J. Shires. Estimating the robustness of questionnaire results: lessons from a mixed-mode survey of expectations for tele-working and road-based business travel, *Transportation*, 36(1), 2009, 47-64.
9. Diana, M. Relationships among household communication tools ownership and use, ICT skills and mobility patterns: evidence from Italy, *9th International Conference on Survey Methods in Transport*, Puyehue, Chile, 14-18 November 2011.
10. ISTAT. *Indagine multiscopo sulle famiglie – Aspetti della vita quotidiana – Anno 2007 – Manuale utente e tracciato record*. Available at http://www.istat.it/it/files/2011/01/20074.zip?title=Aspetti+della+vita+quotidiana+-+30%2Fnov%2F2010+-+2007.zip – Accessed February 27, 2012 (in Italian).
11. Wittkowski, K.M., E. Lee, M. Nussbaum, F.N. Chamian, J.G. Krueger, Combining several ordinal measures in clinical studies. *Statistics in Medicine*, 23(10), 2004, 1579-1592.
12. Heckman, J.J. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4), 1976, 475-492.
13. Heckman, J.J, J.L. Tobias, E. Vytlacil. Four parameters of interest in the evaluation of social programs. *Southern Economic Journal*, 68(2), 2001, 210-223.
14. Cao, X. Disentangling the influence of neighborhood type and self-selection on driving behavior: an application of sample selection model. *Transportation*, 36(2), 2009, 207-222.
15. De Jong, G., J. Fox, M. Pieters, A.J. Daly, R. Smith. A comparison of car ownership models. *Transport Reviews*, 24(4), 2004, 379-408.
16. Vance, C., R. Iovanna. Gender and the automobile: analysis of nonwork service trips. *Transportation Research Record 2013*, 2007, 54-61.

List of figures

List of tables

TABLE 1 Binary (Yes/No) variables to test ICT skills

Label	Description
COMP_1	Ability to copy or move a file or a directory
COMP_2	Ability to copy and paste information in a document
COMP_3	Ability to use basic functions in a spreadsheet
COMP_4	Ability to compress files
COMP_5	Ability to connect and install external devices
COMP_6	Ability to write code for computer programs
COMP_7	Ability to connect a computer with a LAN
COMP_8	Ability to check and fix computer problems
WEB_1	Ability to use a search engine
WEB_2	Ability to send e-mails with attachments
WEB_3	Ability to send messages to chat, newsgroups, forums

TABLE 2 Correlations between ICT ability and frequency of use of travel means

	Urban PT	Train	Car driver
MU_COMP	0.08	0.16	0.33
MU_WEB	0.10	0.17	0.30

NB: Data under the column "Urban PT" are referred to people aged 14 or more that declared that some public transport service is available where they live (N = 40,272,424 after weighting); data under "Train" are referred to people aged 14 or more (N = 48,733,825 after weighting) and data under "Car driver" are referred to people aged 18 or more (N = 45,907,111 after weighting).

TABLE 3 Individual trip rates for the general population and for CS and WS subsamples (*)

	Urban PT			Train			Car driver		
	All	CS	WS	All	CS	WS	All	CS	WS
Trips per year, general mean	44	56	64	14	25	32	256	340	333
(relative differences range)		*+15-36%*	*+30-56%*		*+68-90%*	*+103-143%*		*+29-34%*	*+27-31%*
Percent of users of the mode	30.3%	34.4%	37.4%	30.7%	45.8%	51.2%	72.2%	92.2%	91.6%
Trips per year, mode users	147	164	172	46	56	62	355	369	363
(relative differences range)		*+0.8-19%*	*+4.9-26%*		*+13-27%*	*+22-45%*		*+1.2-5.1%*	*+0.2-3%*

(*) Data under the column "Urban PT" – "All" are referred to people aged 14 or more that declared that some public transport service is available where they live (N = 40,272,424 after weighting); data under "Train" – "All" are referred to people aged 14 or more (N = 48,733,825 after weighting) and data under "Car driver" – "All" are referred to people aged 18 or more (N = 45,907,111 after weighting). Data under "CS" and "WS" columns are referred to the fraction of the corresponding "All" sets belonging to the two subsamples defined in the preceding section.

TABLE 4 Heckman models estimation results and related treatment parameters

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Log-likelihood	-174427	-192072	-222920	-52540	-65244	-161809
Selection equation						
Intercept	-1.402***	-1.429***	-1.329***	-1.302***	-0.958***	-1.236***
Age	-0.029***	-0.028***	-0.030***	-0.029***	-0.032***	-0.031***
At least high school diploma or student	0.415***	0.450***	0.462***	0.358***	0.349***	0.452***
Internet use at least sometimes a week	1.928***	1.900***	1.911***	1.880***	1.722***	1.863***
Household has access to Internet	0.434***	0.426***	0.387***	0.473***	0.487***	0.397***
N	29701	37010	35270	8239	10973	25877
Outcome equation outside WS sample						
Intercept	51.566***	10.917***	141.834***	162.726***	44.053***	210.28***
Female	4.294***	-0.076	-70.656***	-0.635	-2.622^{+}	-26.414***
Age	-0.543***	-0.126***	-0.733***	-1.307***	-0.466***	-0.252***
Worker	-8.797***	0.907*	70.432***	-2.072	1.180	44.092***
Number of cars in the household	-7.515***	-0.363	52.393***	-16.605***	-1.817^{+}	19.701***
Living with children	-6.529***	-1.397***	8.504***	-18.581***	-1.484	9.556***
Living in a metropolitan city center	56.905***	2.715***	-40.528***	39.498***	4.900*	-53.675***
Living in a metropolitan city suburb	4.008**	3.844***	-7.236**	1.241	11.647***	-9.201***
Living in another city above 50.000	19.176***	-0.098	-10.757***	15.594***	-3.400^{+}	-16.772***
N	22788	28767	27800	5843	6796	19029
Outcome equation for WS sample						
Intercept	102.158***	40.300***	81.113***	201.152***	73.928***	145.864***
Female	14.732***	2.108	-27.688***	22.284***	0.089	-17.712***
Age	-0.951***	-0.196**	1.364***	-1.278***	-0.543***	0.586***
Worker	-34.612***	-10.444***	76.282***	-39.762***	-12.573***	60.652***
Number of cars in the household	-11.931***	-2.220*	34.752***	-18.049***	-4.081*	21.147***
Living with children	-1.943	-3.767*	11.222***	-9.490	-1.037	10.105***
Living in a metropolitan city center	59.986***	-1.568	-59.307***	23.479***	-4.167	-57.432***
Living in a metropolitan city suburb	1.508	10.815***	-3.967	-14.868	21.387***	-0.413
Living in another city above 50.000	6.127*	-5.416**	-18.917***	-24.763***	-10.148**	-16.691***
N	6913	8243	7470	2396	4177	6848
Treatment parameters						
ATE (trips/year)	17.6	17.5	40.4	23.3	16.1	-9.9
TT (trips/year)	17.3	15.2	13.5	7.3	16.7	-11.4

NB: '***' => p<0.001; '**' => p<0.01; '*' => p<0.05; '$^{+}$' => p<0.1; otherwise not significant at the 0.1 level.

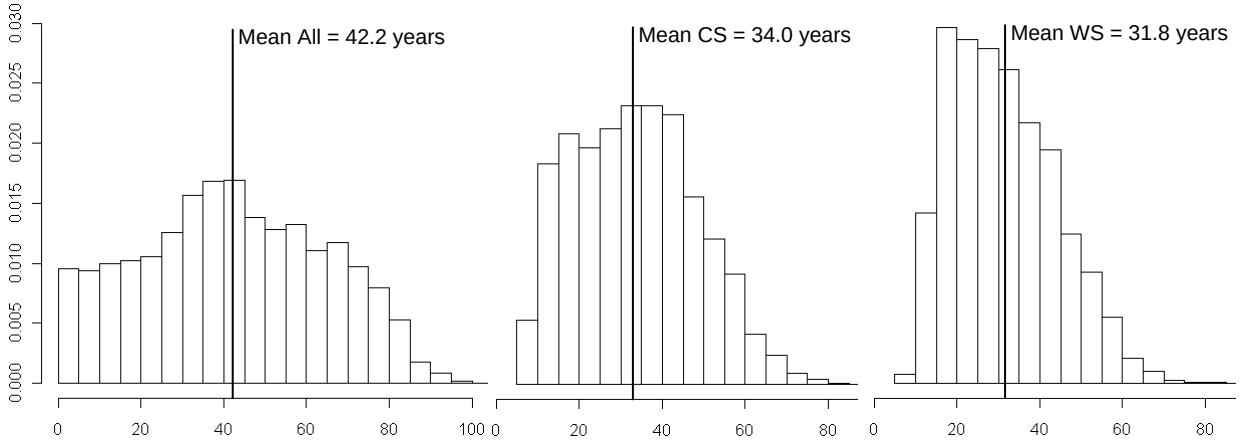

FIGURE 1 Age histograms for the Italian population and for CS and WS subgroups

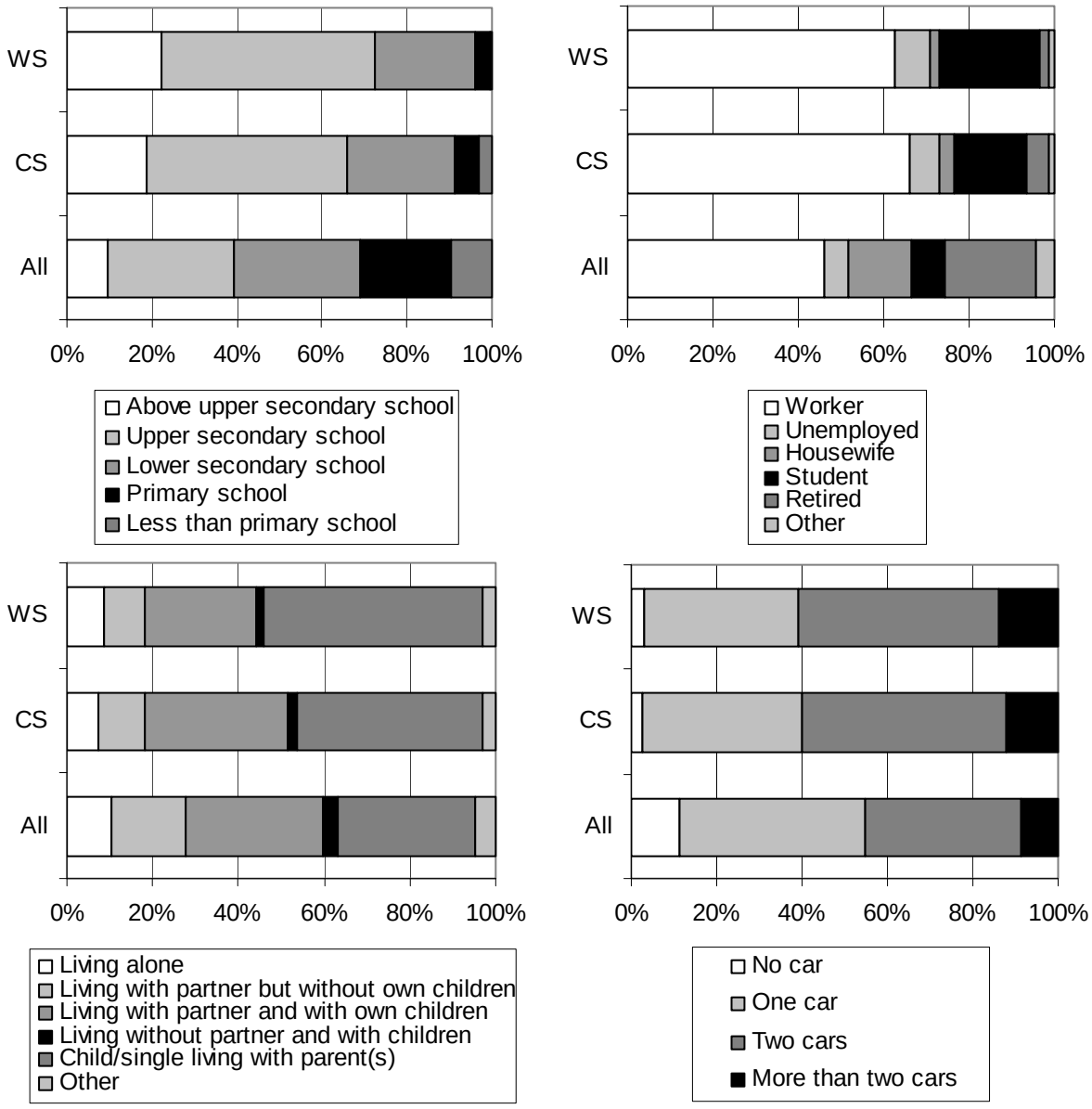

FIGURE 2 Educational levels, employment status, position in the household and car ownership for the entire Italian population and for CS and WS subgroups