Analyzing Observed Heterogeneity in Preferences: A Semiparametric Estimation Approach

Paul Koster* and Hans Koster

Department of Spatial Economics, VU University Amsterdam

Abstract

The estimation of preference heterogeneity is of key importance in the choice modelling literature. We propose a semi-parametric estimation approach to analyze observed heterogeneity. Under the assumption that more similar people in terms of socio-economic characteristics have more similar preferences, we show how to estimate nonparametric distributions of the preferences. The model is applied to a stated choice experiment intended to measure the value of time and reliability.

This version: 25 February 2011

Keywords: nonparametric logit, semi-parametric logit, heterogeneity, local logit, willingness to pay

*Corresponding author: Paul Koster. Email: pkoster@feweb.vu.nl. Address: Faculty of Economics and Business Administration. De Boelelaan 1105, 1081HV Amsterdam, the Netherlands.
1.0 Introduction

The random utility model of McFadden (1974) is a very popular tool to measure preferences of individuals. It is widely applied in subjects as diverse as the demand for new products (e.g. Brownstone and Train, 1998), residential location choice (e.g. Bayer et al., 2007), brand choice (e.g. Swait and Erdem, 2007) and the value of travel time and reliability in transport economics (e.g. Small et al., 2005). Although the heterogeneity of individuals also should lead to heterogeneous preferences, economists are generally more interested in aggregate effects and sometimes regard heterogeneity as nuisance (Allenby and Rossi, 1998). The distribution of consumer preferences, however, plays a central role in many marketing activities, as heterogeneity in preferences leads to market segments, differentiated products and niches. Also welfare implications of policies may be different if heterogeneity of preferences is accounted for (Van den Berg and Verhoef, 2011).

Two types of taste heterogeneity may be considered. Observed heterogeneity is usually analysed by taking into account personal characteristics such as income, gender or age (see for example Bayer et al., 2007). The second type of heterogeneity is unobserved taste heterogeneity. Unobserved heterogeneity can be the result of lack of data about the household characteristics, the decision process of the individual, or the choice set that a respondent considers. A large literature has developed in order to estimate the amount of unobserved taste heterogeneity and usually mixed logit or latent class models are used to estimate unobserved heterogeneity (see for example Small et al., 2005; McFadden and Train, 2000; Train, 2010; Chesher and Santos Silva, 2002).

In this paper we focus on analysing observed heterogeneity using a semi-parametric estimation approach.\footnote{We will show that our procedure may be extended to capture unobserved heterogeneity as well.} We will investigate heterogeneity using local likelihood methods, proposed by Fan and Gijbels (1996) and Fan et al. (1995), which to our knowledge are not frequently been applied in the choice
modelling literature up to now.\textsuperscript{2} Our estimation process involves estimation of a weighted logit model for each individual. The kernel weight variables are dependent on the differences between socio-economic characteristics. That is, more similar individuals have more similar preferences. This estimation procedure results in nonparametric distributions of preferences. We summarise the results by regressing the estimated taste parameters on individual characteristics. A similar approach is used in the literature on hedonic pricing (Bajari and Kahn, 2005; Bajari and Benkard, 2005). We apply our methods to measure the value of travel time and travel time reliability based on a stated choice experiment held among participants of a rewarding experiment. It is shown that the \textit{average} willingness to pay (WTP) values for reductions in time (VOT), schedule delay early (VSDE) and late (VSDL) correspond to the coefficients estimated by an ordinary Logit model, which increases confidence in the estimation procedure. Furthermore, we demonstrate that there is substantial (observed) heterogeneity in the value of time and reliability.

Our estimation procedure has several advantages compared to other techniques. First, our method focuses on observed heterogeneity and adequately defines the sources of heterogeneity. Methods that account for unobserved heterogeneity have probably less predictive power because when the population changes it is unknown how the distribution of preferences changes since the source of heterogeneity is unknown. Furthermore, our procedure directly identifies preferences of individuals, whereas this is more difficult with simulation techniques. It is well known that the number of repetitions used in the simulation and the type of software may affect the estimation results, especially when sample are small (Chang and Lusk 2010). Third, our method does not make any distributional assumptions on the preferences, in contrast to other methods such as Mixed Logit.\textsuperscript{3} It may also be shown that local likelihood estimation techniques use far fewer degrees of freedom than fully nonparametric estimation (McMillen and Redfearn, \textsuperscript{3} One notable exception is Fosgerau (2007). His paper mainly focuses on the specification of the local utility function.\textsuperscript{3} Exceptions are for example Train (2010), Fosgerau (2006) and Fosgerau and Nielsen (2010) who estimate nonparametric distributions.}
2010). Fourth, although our procedure is very flexible, it is computationally light and can be estimated using standard statistical software packages.

The plan of the paper is as follows. Section 2 introduces the econometric setup. In Section 3 we discuss the choice experiment followed by the empirical results in Section 4. Section 5 concludes.

2.0 Econometric setup

2.1 Local Logit estimation

Fan et al. (1995) and Fan and Gijbels (1996) introduce local likelihood estimation. This procedure can be used to estimate the choice probabilities in a stated choice experiment, under the assumption that more similar choices in terms of the independent variables \(X\), result in more similar estimated choice probabilities (Fosgerau, 2007).\(^4\) It can also be used to analyze non-linearities in the taste parameters caused by individual characteristics. This approach is used by Frölich (2006) to analyze the effect of having children on the employment rate. He shows that Local Logit is much more efficient than parametric regression. Park et al. (2010) study the theoretical properties of the estimator and use simulation to show that Local Logit performs well, even for small samples.

First, we need to introduce some notation. The dependent variable \(y\) is the choice that is made. In our case it is binary; it is 1 if alternative A is chosen and 0 if alternative B is chosen. The dependent variable is a function of the independent variables \(X\), and the vector of taste parameters \(\beta\). Since we are interested in observed heterogeneity, we estimate how individual characteristics \(Z\) affect the taste parameters \(\beta\). The taste parameters depend in a non-linear way on \(Z\) via the nonparametric function \(g(\cdot)\). This means that all interactions of the different variables in \(Z\) are modeled implicitly (McMillen and Redfearn, 2010). The probability \(P[\cdot]\) that we want to estimate is given by the standard Logit formula.

\[
y = P[g(\beta; Z); X] + \varepsilon. \quad (1)
\]

\(^4\) This can be done by specifying the local utility as \(aX + b(X - X_0)\). The coefficients \(a\) and \(b\) are then locally estimated and depend on \(X_0\). \(P(a(X_0))\) is then the estimated probability at \(X_0\).
The goal is to estimate the probability of the chosen alternative as close as possible to 1. Suppose we observe \( N \) binary choices in total, made by \( I \) individuals. The vector of taste parameters of individual \( i = 1, \ldots, I \) can then be estimated by maximizing the local likelihood:

\[
\hat{\beta}_i = \arg \max \sum_{n=1}^{N} w_{h,i}[Z] \cdot \log[P^*(\beta_i; X_n)].
\]  

Equation (2) shows that the local log-likelihood is calculated by taking the log of the probability of the chosen alternative, multiplied by a vector of weights \( w_{h,i} \). The log of the probability of the chosen alternative depends on the independent variables and the taste parameters and needs to be maximized in order to arrive at the local estimate \( \hat{\beta}_i \). The weights depend on the characteristics of individual \( i \). In the next section we will show in more detail how these weights are calculated.

2.2 Kernel functions and bandwidth selection

The weights are assumed to be determined by the distance between individuals in socio-economic space using a Kernel function. When the differences in socio-economic space between two individuals \( \bar{i} \) and \( i \) become smaller, observations of \( i \) are weighted heavier in the local regression of \( \bar{i} \) (and vice versa). Important parameters of the Kernel functions are the bandwidths (denoted by \( h \)). When \( h \) is larger, more observations in the ‘neighborhood’ of \( \bar{i} \) are taken into account in Logit estimation of \( \bar{i} \). For a very large bandwidth, the resulting local estimates are equal to the estimates of a standard Logit model.

Suppose we include \( Q \) variables in the Kernel function. Each of these variables has a corresponding Kernel function \( K_q(\cdot) \) and bandwidth \( h(q) \). Note that the Kernel function does not need to be the same for each variable. A general specification of the Kernel weights is then given by equation (3).

\[
w_{h,i,i} = \prod_{q=1}^{Q} K_q(h(q); z_{q,i} - z_{q,i}).
\]

Note that we ignore the panel structure in the estimation procedure, so we do not allow for correlation of the error term over a sequence of choices. This is a topic for further study.
An example of a Kernel function is the product of $Q$ standard normal distributions of the weight variables. Then the weights are given by:

$$K_q(h(q); z_{q,i} - z_{q,l}) = \frac{1}{h(q)} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{z_{q,i} - z_{q,l}}{h(q)} \right)^2 \right].$$  \hfill (4a)

For categorical variables, another Kernel function must be used, as argued by Racine et al. (2004). They show that for these variables one needs a kernel function that has the possibility to be an indicator function and that it must be possible to smooth out a categorical variable. If equation 4a is used for categorical variables, the second property will not hold. For a model with only ordered categorical variables or unordered categorical variables the Kernel weights are given by equations 4b and 4c respectively (Hall et al. 2007; Racine et al. 2004).

$$K_q(h(q); z_{q,i} - z_{q,l}) = \begin{cases} 1, & \text{if } z_{q,i} = z_{q,l} \\ h(q) |_{z_{q,i} - z_{q,l}}, & \text{if } z_{q,i} \neq z_{q,l} \end{cases}.$$  \hfill (4b)

$$K_q(h(q); z_{q,i} - z_{q,l}) = \begin{cases} 1, & \text{if } z_{q,i} = z_{q,l} \\ h(q), & \text{if } z_{q,i} \neq z_{q,l} \end{cases}.$$  \hfill (4c)

In equations 4b and 4c the bandwidth $h(q)$ must be between 0 and 1. If $h(q)$ equals 1, the weights are uniform and the variable is smoothed out. This means that the variable has no effect on the estimated taste parameters. If $h(q) = 0$, the function is an indicator function and the sample is divided in two parts, when the variable of interest is dichotomous. In our estimation, we use a mixed Kernel function with both continuous and categorical variables.

We employ a multivariate bandwidth, implying that for each socio-economic variable $q$ we determine a separate bandwidth $h(q)$. If the bandwidth is high ($h(q) \to 1$ for categorical variables, and $h(q) \to \infty$ for a continuous variable), or the model fit does not improve significantly, the variable can be excluded from the Kernel function. A drawback of using a multivariate bandwidth is that it is hardly possible to analyze the full grid of possible bandwidths if there are many variables in the weight matrix. A practical approach that can be employed is to add the variables one by one to and then optimize the bandwidth for each variable.
separately, holding the bandwidths of the other variables constant, or to group variables and assume that they have the same bandwidth (Frölich 2006). As a bandwidth selection criterion we use the leave-one-out cross validation statistic given by equation 5:

\[ CV = \sum_{n=1}^{n} (1 - P[\hat{\beta}_{-n}])^2. \] (5)

In this equation, \( P[\hat{\beta}_{-n}] \) is the probability of the chosen alternative evaluated at \( \hat{\beta}_{-n} \), which is the vector with locally estimated parameters. The \(-n\) subscript indicates that observation \( n \) is omitted in the local estimation of \( n \), meaning that the weight of \( n \) is set to 0 in the \( n_{th} \) local estimation (Ruppert et al. 2003).

3.3 Model Selection

A variable that is included in the Kernel function only potentially affects the locally estimated preference parameters. In order to investigate if adding a weight variable significantly improves the estimation we use the \( AIC_c \) statistic as proposed by Hurvich et al. (1998). This statistic trades off the number of parameters used in the model versus the model fit. The number of parameters is approximated using the projection or hat matrix.

The Binary Logit model is a special case of the group of Generalized Linear Models (GLMs), which encompasses models that are member of the exponential family. A variance function is needed at the local estimate \( \hat{\beta}_n \) where the variance is multiplied by the kernel weights:

\[
M_n[X\hat{\beta}_n] = \frac{e^{x\hat{\beta}_n}}{(1 + e^{x\hat{\beta}_n})^2} \cdot W_{h,n}.
\] (6)

This vector has size \( N \cdot 1 \). Note that the variance function is locally defined and evaluated at the local estimate \( \hat{\beta}_n \). The locally defined weight matrix is as follows:

\[
\hat{W}_n = diag(M[X\hat{\beta}_n]).
\] (7)
In the case of ordinary least squares (OLS) the diagonal of $W_n$ is simply given by the kernel weights because the variance function is equal to one. The $n^{th}$ row of the hat matrix -denoted by $\hat{H}(n,:)$- is then:

$$\hat{H}(n,:) = X(n,:)\widehat{\Omega}_nX\widehat{W}_n.$$  \hspace{1cm} (8)

The middle part of the equation is the locally estimated covariance matrix. If local standard errors are high, this will result in high values of $H(n,:)$. If the bandwidth goes to infinity (or to 1 for categorical and dummy variables), the trace of $\hat{H}$ will converge to the number of parameters to be estimated in the model.

For small datasets, equation (8) can be calculated easily. However, for larger datasets the matrix $W_n$ becomes too big and therefore a less computational intensive routine is needed (Greene, 2003). As we only use the trace of the hat matrix, it is sufficient to calculate only the diagonal elements $\hat{H}(n,n)$. We do so by calculating the $n^{th}$ column of $W_n$, which is a $N \times 1$ vector with on the $n^{th}$ row the variance element. This results in:

$$\hat{H}(n,n) = X(n,:)\widehat{\Omega}_nX\widehat{W}_n(:,n).$$ \hspace{1cm} (9)

The trace of the hat matrix is then the sum of these diagonal elements. Suppose we have $N$ choices and estimate $K$ parameters. Let $\hat{\beta}$ the $N \times K$ vector with all the locally estimated parameters of the individuals, so a vector $[\hat{\beta}_1, ..., \hat{\beta}_N]$. Furthermore, let $LL[\hat{\beta}]$ the global likelihood of the estimated model. To evaluate the performance of a chosen bandwidth, we use a likelihood criterion proposed by Hurvich et al. (1998):

$$AIC_c(h) = \frac{-2 \cdot LL[\hat{\beta}]}{N} + \frac{2 \cdot tr(\hat{\beta}) + 1}{N - tr(\hat{\beta}) - 2}. \hspace{1cm} (10)$$

The $AIC_c$ gives a good indication for model fit, while, at the same time, it accounts for the number of parameters in the model. Adding a weight variable might improve the global likelihood, but also may increase the trace of the hat matrix. Suppose that for model $f = 1 \ldots F$ the vector of optimal bandwidths is given by $h^*_f$ and the corresponding $AIC_c$ is given by $AIC_c(h^*_f)$. Then the best model can be found by maximizing $AIC_c(h^*_f)$ for $f = 1 \ldots F$. As a rule of thumb models are considered as significantly different
if the difference between model criteria is larger than $4/N$ (Charlton 2009), meaning that the $AIC_c$ must decrease with at least $4/N$ points when a socio-economic variable is added to the weight matrix.\textsuperscript{6}

### 3.0 Experimental setup and data

#### 3.1 Setup of the choice experiment

A stated choice experiment is developed to collect data about the preferences of morning-commuters participating in a peak-avoidance project. In order to reduce congestion travelers are \textit{rewarded} if they do not travel during the morning peak. A typical example of a choice question is given by figure 1.

<table>
<thead>
<tr>
<th>Departure time from home</th>
<th><strong>Alternative A</strong></th>
<th><strong>Alternative B</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>6:05</td>
<td>6:50</td>
</tr>
<tr>
<td>Total travel time</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>Travel time from home to camera A</td>
<td>30 min, 15 min</td>
<td>20 min, 10 min</td>
</tr>
<tr>
<td>Travel time from camera A to camera B</td>
<td>5 min, 10 min</td>
<td>5 min, 10 min</td>
</tr>
<tr>
<td>Travel time from camera B to work</td>
<td>10 min, 15 min</td>
<td>5 min, 10 min</td>
</tr>
<tr>
<td>Arrival time at work</td>
<td>6:35, 6:45</td>
<td>7:10, 7:25</td>
</tr>
<tr>
<td>Reward</td>
<td>4 euro</td>
<td>0 euro</td>
</tr>
</tbody>
</table>

*Figure 1: Example of a choice question.*

Respondents were asked to choose between two departure times. Each departure time has two travel times with a corresponding probability and reward. Above the choice question the preferred arrival time (PAT) of the traveler is given, which is based on earlier answers in the questionnaire. It is the time a traveler wants to arrive if there is no rewarding and there is no travel time unreliability. More details about the design and the process of data cleaning are given in Knockaert et al. (2011). We use a slightly smaller sample than in their study, since we excluded some respondents for which the socio-economic characteristics were not available.

\textsuperscript{6} Some authors suggest using the $AIC_c$ statistic also as a selection criterion for bandwidth selection (Hurvich et al. 1998). In our case this will result in under smoothing, since the data has a panel structure. In the local estimation the panel structure is ignored, leading to an underestimation of local standard errors and therefore leading to a too low bandwidth (undersmoothing). It is therefore better to use cross-validation for the selection of the optimal bandwidth and the $AIC_c$ for the selection of the model.
3.2 Utility specification

In the basic model we estimate, we assume that the deterministic part of the utility \((V)\) of an departure time from home \((t_h)\) is explained by three types of variables: expected reward \(E(R)\), expected travel time \(E(T)\) and expected schedule delay \(E(SD)\). Schedule delay is defined as the deviation of the arrival time from the preferred arrival time \(PAT\) (Small 1982). The expected utility of choice \(n\) made by individual \(i\) choosing alternative \(j\) is defined as follows:

\[
U_{nj} = V_{nj}[\beta_i; E(R)_{nj}, E(T)_{nj}, E(SD)_{nj}] + \varepsilon_{nj}.
\]  

Equation (11) shows that utility is an additive function of the deterministic part and the random component.\(^7\) A common approach is to specify the schedule delay as a piecewise linear function where the marginal disutility of being early and late are possibly valued differently by travelers. Equations 12a and 12b show this specification based on the standard scheduling model introduced by Vickrey (1969) and Small (1982). This model was extended to a model with stochastic travel times by Noland and Small (1995). The schedule delay of masspoint \(z\) of alternative \(j\) of choice \(n\) is given by:

\[
SDE_{njz} = \text{Max}(0; PAT - t_{nj} - T_{njz})
\]  

\[
SDL_{njz} = \text{Max}(0; t_{nj} + T_{njz} - PAT)
\]  

These equations show that the schedule delay disutility is a piecewise linear function of arrival time, meaning that besides the disutility of travel time, there is additional disutility of arriving early or late. As figure 1 shows, every departure time has 2 realizations of travel time (so \(z=2\), with a corresponding probability \(p_{ni}\) and \(1 - p_{ni}\) respectively. The model variables can then be calculated as:

\[
E(R)_{nj} = p_{nj} \cdot R_{nj1} + (1 - p_{nj}) \cdot R_{nj2}
\]  

\[
E(T)_{nj} = p_{nj} \cdot T_{nj1} + (1 - p_{nj}) \cdot T_{nj2}
\]  

\[
E(SDE)_{nj} = p_{nj} \cdot SDE_{nj1} + (1 - p_{nj}) \cdot SDE_{nj2}
\]  

\(^7\) This is not necessary for a random utility model. For example, Fosgerau and Bierlaire (2010) propose a multiplicative specification.
\[ E(SDL)_{nj} = p_{nj} \cdot SDL_{n1} + (1 - p_{nj}) \cdot SDL_{n2} \]  

(13d)

The deterministic part of the utility (V) is then:

\[ V_{nij} = \beta_{R,i} \cdot E(R)_{nj} + \beta_{T,i} \cdot E(T)_{nj} + \beta_{SDE,i} \cdot E(SDE_{nj}) + \beta_{SDL,i} \cdot E(SDL_{nj}) \]  

(14)

Note that the formulation is not as restrictive as it seems, as the preference parameters \( \beta \) are individual-specific. In order to compare the local estimates, the WTP values are used. These are defined in equations (15a)-(15c):

\[ VOT = -\frac{\beta_{T,i}}{\beta_{R,i}} \]  

(15a)

\[ VSDE = -\frac{\beta_{SDE,i}}{\beta_{R,i}} \]  

(15b)

\[ VSDL = -\frac{\beta_{SDL,i}}{\beta_{R,i}} \]  

(15c)

The WTPs are used because the locally scale parameter may vary over the different local estimations. In other words, the ratio of the parameter and the scale of the error (\( \beta_{x,i}/\lambda_i \)) is estimated. Therefore the absolute values of the locally estimated taste parameters for different individuals cannot be compared directly. When we calculate the WTP values for attribute \( x \), we take ratios \(-\beta_{x,i}/\beta_{R,i}\). Therefore the scale \( \lambda_i \) drops out and the estimates of the different individuals can be compared.

**4.0 Estimation results**

**4.1 Baseline results**

In this section we present the estimation results. First, we will start with a simple model where we estimate a model using a univariate weight matrix. So, we assume that observed heterogeneity is caused by only one individual characteristic. In section 4.3, we will present a model with all variables included. In that section we will regress the estimated taste parameters on individual characteristics. Table 1 summarises the main estimation results.
Table 1: The value of time, schedule delay early and late for different estimation procedures

<table>
<thead>
<tr>
<th></th>
<th>Ordinary Logit</th>
<th>Semiparametric Logit with univariate kernel</th>
<th>Semip. Logit with multivariate kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age</td>
<td>Income</td>
<td>Education</td>
</tr>
<tr>
<td>VOT</td>
<td>32.62</td>
<td>44.60</td>
<td>(30.58)</td>
</tr>
<tr>
<td>VSDE</td>
<td>22.97</td>
<td>27.84</td>
<td>(11.47)</td>
</tr>
<tr>
<td>VSDL</td>
<td>16.20</td>
<td>18.67</td>
<td>(6.23)</td>
</tr>
<tr>
<td>Number of Individuals</td>
<td>396</td>
<td>396</td>
<td>396</td>
</tr>
<tr>
<td>AIC&lt;sub&gt;C&lt;/sub&gt;</td>
<td>1.1060</td>
<td>1.0846</td>
<td>1.0987</td>
</tr>
</tbody>
</table>

NOTE: The standard deviations of the coefficients are between parentheses.

Table 1 shows rather high WTP values for the ordinary Logit estimation compared to earlier estimations for Dutch commuters (Tseng, 2008). This is because of two reasons. First, the average income in our sample is higher than the Dutch average. This leads to higher WTP values. Second, our experiment offers a reward instead of a cost. It is well known that travellers are more sensitive to cost components than to reward components. A second observation is that the VSDE is higher than the VSDL, where usually VSDE<VSDL is found. This is because most travellers that participate in the rewarding experiment have a preferred arrival time early in the morning. The average WTP estimates follow a similar pattern for the semiparametric estimations. The VOT is higher and the VSDE and VSDL are close to the mean values found in the ordinary Logit estimations.

4.1 Univariate kernel weights

For the univariate kernel weights estimation we used the cross-validation statistic to optimize the bandwidth. In figure 1 and 2 the results are plotted for age and income. The age variables show a bimodal pattern, where younger and older travellers have a higher WTP. The fact that younger travellers have higher WTP values is caused by the fact that we omit the other household characteristics. As shown by figure 5, having young children leads to a substantial higher WTP. The age variable might pick up the effect, as younger people generally have younger children. The curves of VOT, VSDE and VSDL follow a
similar pattern so it is likely that the differences in WTP are mainly caused by differences in the sensitivity to rewards. As expected, a higher income results in higher WTPs, but the VOT is decreasing for the last income class.

**Figure 2: The effect of age**

**Figure 3: The effect of income**

There is a minor effect of education, where a higher education increases WTP values. Finally, having young kids (<5 year) in the household increases WTPs strongly. In all cases most heterogeneity seems to be in the reward coefficient since the pattern of VSDL<VSDE<VOT remains the same.

**Figure 4: The effect of education**

**Figure 5: The effect of having young kids**
4.2 Multivariate kernel weights

For the multivariate kernel weights we tried to use the cross-validation statistic to determine the optimal bandwidth. This time it performed poorly and leads to severe undersmoothing. It may be that outliers may result in probabilities of the chosen alternative that are close to 1 resulting in a low CV value.\textsuperscript{8} This again is an argument to account for the local standard errors using for example the $AIC_C$ statistic, but as argued before this will also lead to undersmoothing. Therefore we used eye-ball ing to determine the smallest possible bandwidths that lead to plausible estimation results. As a starting point we used the optimal univariate bandwidths. Since in the univariate case variables may be correlated with omitted variables, we know that it is not plausible that these bandwidths are lower in the multivariate case. The distributions of VOT, VSDE and VSDL are given in figure 6. The heterogeneity in the VOT is larger than the heterogeneity in VSDE and VSDL. The distributions do not follow a typical distribution, although a shifted lognormal distribution might come close.

\textsuperscript{8} This is a typical difference with OLS where outliers have a large effect on the mean squared error.
To obtain results presented in Table 4, we regressed the logs of the WTP-values on socio-economic characteristics of individuals. When we had estimated this using OLS, the estimates are consistent but may be inefficient because the error term is likely correlated over the same individual. A gain in efficiency may be obtained by estimating a Seemingly Unrelated Regression (SUR) model, where we account for the correlation of the error terms of the three equations (Zellner 1962). First, females have 18% higher WTP values than males. The latter is consistent with earlier findings of Lam and Small (2001) and Amador et al. (2005). Furthermore, the income effect is visible: travellers with a higher income have higher WTP values because they have a lower marginal utility of income but as in the univariate case the effect is decreasing for the last income class. Singles have more scheduling flexibility and have 12% lower values of schedule delay than non singles. Travellers with very young kids have high WTP values. Their VOT is 56% higher than for travellers without young kids and the values of schedule delay are more than 30% higher. The WTP values are also decreasing in age. So, even if we control for other observed characteristics there is a

---

9 One may also use flexible estimation techniques (e.g. local linear regression). For presentation purposes however, we assume a linear relationship between the individual characteristics and the WTP-values. As is observed in Table 4, the $R$-squared of the regressions are quite high. We also investigate whether taking logs of WTP-values changes the results. It appears that the results are then very similar.

10 More details about the estimation of SUR models can be found in standard econometric textbooks. We estimated the model in STATA. The standard errors of the OLS estimation are rather similar.
decreasing effect of age. All covariates that are significant, affect the WTP values in the same direction except for the effect of income on the VSDL.

Table 4: SUR regressions of individual characteristics and VOT, VSDE and VSDL

<table>
<thead>
<tr>
<th></th>
<th>log(VOT)</th>
<th>log(VSDE)</th>
<th>log(VSDL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 36-40</td>
<td>-0.316 (0.027)***</td>
<td>-0.196 (0.026)***</td>
<td>-0.165 (0.025)***</td>
</tr>
<tr>
<td>Age 41-55</td>
<td>-0.462 (0.025)***</td>
<td>-0.231 (0.024)***</td>
<td>-0.112 (0.023)***</td>
</tr>
<tr>
<td>Age &gt;55</td>
<td>-0.659 (0.032)***</td>
<td>-0.212 (0.031)***</td>
<td>0.001 (0.029)</td>
</tr>
<tr>
<td>Income €2500-€3500</td>
<td>0.398 (0.039)***</td>
<td>0.493 (0.037)***</td>
<td>0.524 (0.035)***</td>
</tr>
<tr>
<td>Income €3500-€5000</td>
<td>0.640 (0.040)***</td>
<td>0.554 (0.039)***</td>
<td>0.516 (0.037)***</td>
</tr>
<tr>
<td>Income &gt;€5000</td>
<td>0.450 (0.043)***</td>
<td>0.501 (0.041)***</td>
<td>0.511 (0.040)***</td>
</tr>
<tr>
<td>Education – Bachelor Degree</td>
<td>-0.017 (0.025)</td>
<td>-0.043 (0.024)*</td>
<td>-0.027 (0.023)</td>
</tr>
<tr>
<td>Female</td>
<td>0.186 (0.021)***</td>
<td>0.189 (0.020)***</td>
<td>0.184 (0.019)***</td>
</tr>
<tr>
<td>Single</td>
<td>-0.046 (0.035)</td>
<td>-0.122 (0.033)***</td>
<td>-0.115 (0.032)***</td>
</tr>
<tr>
<td>Dinky</td>
<td>0.012 (0.029)</td>
<td>-0.024 (0.027)</td>
<td>0.021 (0.026)</td>
</tr>
<tr>
<td>Young Kids</td>
<td>0.561 (0.031)***</td>
<td>0.394 (0.030)***</td>
<td>0.320 (0.029)***</td>
</tr>
<tr>
<td>Kids at Primary School</td>
<td>0.077 (0.029)***</td>
<td>0.009 (0.027)</td>
<td>-0.009 (0.026)</td>
</tr>
<tr>
<td>Other Household Type</td>
<td>0.101 (0.098)</td>
<td>0.024 (0.094)</td>
<td>0.011 (0.089)</td>
</tr>
<tr>
<td>Constant</td>
<td>3.207 (0.046)***</td>
<td>2.768 (0.043)***</td>
<td>2.318 (0.042)***</td>
</tr>
</tbody>
</table>

Number of observations   396  396  396

R-squared                0.859  0.745  0.681

NOTE: Standard errors are between parentheses. Significance at *0.10, **0.05 and ***0.01 levels.

5. Conclusions and discussion

In this paper we analyzed observed taste heterogeneity using a semi-parametric estimation approach. We show that the approach is a powerful tool to analyze observed heterogeneity and to obtain the full distribution of the WTPs under the assumption that more similar people have more similar WTP values. In future research the specification of the local choice probability can be changed in order to capture other correlation structures. The method may be extended to any random utility model. For example, a similar strategy can be used to estimate a local Nested Logit or local Mixed Logit model to capture unobserved heterogeneity. In that case the unobserved part of heterogeneity can be estimated in a non-parametric way.
Acknowledgements

Part of the work was done while the first author was visiting the Danish Technical University, Department of Transport. We like to thank Mogens Fosgerau, Katrine Hjorth, Jan Rouwendal and Vincent van den Berg for their valuable comments.

References


