KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and Applied Economics

Department of Economics

Enriching income data with expenditure information: a semi-parametric imputation technique

by

André DECOSTER
Kris DE SWERDT

Public Economics

Center for Economic Studies
Discussions Paper Series (DPS) 08.11
http://www.econ.kuleuven.be/ces/discussionpapers/default.htm

**April 2008**

DISCUSSION PAPER

# Enriching income data with expenditure information: a semi-parametric imputation technique*

André Decoster
Kris De Swerdt

Center for Economic Studies - K.U.Leuven

April, 2008

**Abstract**

In this paper we describe a methodology for enriching an income dataset with information on expenditures using a semi-parametric imputation technique. Engel curves are first estimated semi-parametrically on household budget data. We then show how the technique can be used to impute expenditure information into a separate income dataset. As an example we show results from the imputation of expenditures in a separate income file using Belgian household budget data.

## 1 Introduction

For decades, Engel curves have been estimated parametrically. Many functional forms have been explored in the literature. One that is often used in applied work originates from the work of Working (1943) and Leser (1963) and therefore often referred to as the Working-Leser functional form. It relates budget shares in a linear way to the logarithm of total expenditures as follows:

$$w_i = \alpha_i + \beta_i log\left(x\right), \tag{1}$$

where $w_i$ represents the budget share on good $i$ , $x$ are total expenditures and $\alpha_i$ and $\beta_i$ are parameters to be estimated.

*The authors are grateful to Frederic Vermeulen, Laurens Cherchye, Guy Van Camp and especially Bart Capéau for helpful comments. Of course, none of these can be held responsible for any errors in the paper. A more elaborated version of this paper –especially of the empirical application– can be found in Decoster et al. (2004).

Though often used, a strict linear specification not necessarily provides the best description for every commodity. Indeed, the relation between expenditures and budget shares may very well be (highly) non-linear for certain commodities. In the last 15 years an new consensus seems to have emerged that, from the many possible non-linear functional forms, the quadratic specification –the so-called QUAIDS-demand system– is sufficiently flexible to capture the non-linearities of the Engel curves. The new *functional* form underlying the QUAIDS-demand system was derived in Banks et al. (1997) and was in fact inspired by a visual inspection of *nonparametrically* estimated Engel curves for different commodities (see also Blundell and Duncan, 1998; Blundell et al., 1998; Blundell et al., 2003).

Nonparametric *estimation* of Engel curves has since become common practice in the empirical literature on consumption behaviour and its (theoretical) advantages are widely described (for an overview see e.g. Yatchew, 1998). The main advantage of nonparametric estimation is of course the absence of the 'strait-jacket' of a functional form, leaving a maximum of flexibility in the estimation of the relation between income and expenditures. However, on imputation or even out-of-sample prediction using nonparametric techniques, the literature is rather silent. In this paper we try to fill in part of this gap by describing a methodology to impute expenditures in an income dataset within the theoretical framework of semi-parametric (or nonparametric) estimation.

Indeed, why should we give up the praised flexibility of nonparametric estimation when it comes to imputation? A nonparametric estimation does not have as a result a fixed set of coefficients that can then be used for imputation or prediction, but in this paper we demonstrate that it is perfectly possible to use nonparametric techniques for imputation as well. The paper is organized as follows. In section 2 we will justify the use of and describe the semi-parametric estimation technique. Section 3 then describes an algorithm to impute information using semi-parametric regression. Section 4 briefly describes the data. Results are shown in section 5 while section 6 concludes.

# 2 Semi-parametric estimation of Engel curves

## 2.1 nonparametric regression: brief overview

Estimating a regression function boils down to finding the conditional mean of the dependent variable given the independent variable(s), that is, given the (parametric) function $y = \beta' x + \epsilon$, we have $E(y|x) = \beta' x$. In general, one can of course write this conditional mean as:

$$E(y|x) = \int y f(y|x)\, dy, \tag{2}$$

where $f(y|x)$ is the conditional density of $y$ given $x$. Knowing that $f(y|x) =$

$\frac{f(x,y)}{f(x)}$ and $f(x) = \int f(x,y)\,dy$, where $f(x)$ is the marginal density of $x$ and $f(x,y)$ the joint density of $x$ and $y$, we can rewrite the conditional mean as:

$$E(y|x) = \frac{\int y f(x,y)\,dy}{\int f(x,y)\,dy}. \tag{3}$$

The objective of a nonparametric regression then, is to replace the numerator and denominator in (3) by nonparametric estimators. Introducing a kernel smoothing function $K(.)$, $H$ the number of observations and $b_x$ the kernel bandwidth associated with variable $x$, the expression in (3) can be written as:

$$\hat{E}(y|x) = \frac{\frac{1}{Hb_x}\sum_{h=1}^{H} K\left(\frac{x_h - x}{b_x}\right) y_h}{\hat{f}_b(x)}, \tag{4}$$

where $\hat{f}_b(x) = \frac{1}{Hb_x}\sum_{h=1}^{H} K\left(\frac{x_h - x}{b_x}\right)$, the estimated density of $x$. For a more formal derivation and necessary conditions we refer the reader to Pagan and Ullah (1999).[1] The estimator in (4) is known as the Nadaraya-Watson estimator after the work of Nadaraya (1964) and Watson (1964).

More generally we can write any nonparametric estimator as $\sum w_{b_h}(x)\, y_h$ where $w_{b_h}(x) = w_b(x_h, x)$ the weight assigned to observation $h$. For the Nadaraya Watson estimator (4), $w_{b_h}(x)$ is given by $\frac{\frac{1}{Hb_x} K\left(\frac{x_h - x}{b_x}\right)}{\frac{1}{Hb_x}\sum_{h=1}^{H} K\left(\frac{x_h - x}{b_x}\right)}$.

## 2.2 Engel curves: semi-parametric estimation

The general form for Engel curves can be written as follows:

$$y_i = g_i(x) + \varepsilon_i, \tag{5}$$

where the dependent variabel $y_i$ might be the expenditures on good $i$ or the budget share for good $i$ and $x$ is an explanatory variable such as total disposable income or total expenditures. The function $g_i(.)$ is an unknown function and $\varepsilon_i$ a random error term.

It is well known that expenditure patterns apart from total expenditures or disposable income also depend on other socio-economic and demographic characteristics. To include other characteristics in the specification, equation (5)

[1] One of the conditions needed for this result to hold is that the kernel function be symmetric. A Gaussian kernel based on the (symmetric) standard normal distribution is an example of a kernel smoothing function that satisfies this condition.

can be adapted to read as follows:

$$y_i = g_i(x, \mathbf{z}) + \varepsilon_i, \tag{6}$$

where $\mathbf{z}$ now represents a vector of characteristics.

Estimating equation (6) as stated would imply a fully nonparametric estimation of the relation between consumption, income and household characteristics. Theoretically there are no problems in estimating this relationship. It involves a straightforward generalization of the techniques described in the previous sections. For an $n$-dimensional problem we could use an $n$-variate standard normal distribution as the kernel function, rather than a univariate Gaussian one for example. The problem, however, is of a practical nature and has to do with data requirements. If the dimension of the vector is relatively large we would need immensely large datasets to estimate this relationship accurately. This problem is known in nonparametric analysis as the 'curse of dimensionality'. Intuitively, the higher the dimension of the vector of variables, the sparser will be the data in a neighbourhood (however defined) around the data point where we wish to estimate the relationship, and hence the fewer will be the observations over which to locally smooth or average, leading to inaccurate estimates with very slow rates of convergence to the true regression function.[2]

A proposed solution to this problem is the use of semi-parametric models, where part of the regression is entirely nonparametric and another part enters the equation in a parametric way. The model in (6) can then be rewritten as[3]:

$$y_i = \boldsymbol{\beta}_i' g(\mathbf{z}) + F_i(x) + \varepsilon_i, \tag{7}$$

where $\mathbf{z}$ is the vector of household characteristics, $g(.)$ a known function, $\boldsymbol{\beta}$ a vector of parameters to be estimated, $F_i(.)$ an unknown function to be determined as well and $\varepsilon_i$ a random error term with conditional mean equal to zero and variance $\sigma_\varepsilon^2$. The nonparametric part, the function $F(.)$, is now of lower

[2] A small numerical example will illustrate (Yatchew, 1998). Suppose we have a function $f$ defined on the unit interval, i.e. we are considering the one-dimensional case first. If we uniformly distribute $T$ data points over this interval, the typical distance between observations will be $1/T$, and hence the approximation error will reduce at rate $O(1/T)$ $O(.)$ stands for order of magnitude as we increase the density of points. Now suppose that $f$ is defined over the unit square and we again uniformly distribute $T$ data points over this square. The typical distance between data points will now be $1/\sqrt{T}$ (each data point occupies an 'area' of $1/T$). The approximation error will consequently reduce at rate $O(1/\sqrt{T})$ as the number of data points increases. More generally, for $n$-dimensional problems, the approximation error reduces at rate $O(1/\sqrt[n]{T})$. This implies, for example, that for a sample where $T = 100$ the approximation error will be 10 times as large in two dimensions as it would be in one dimension. Put differently, 10.000 observations would be needed in two dimensions to obtain the same accuracy as 100 observations in one dimension.

[3] Remark that is considered here as a single variable, i.e. total expenditures or disposable income. However, can also be a vector of variables, the dimension depending on the number of observations available for estimation (see text and footnote 2). In the results shown in section 5 both income and age were used in the nonparametric part. See also Decoster et al. (2004) and Schmalensee and Stoker (1999).

dimension as compared to expression (6). The estimation of this model and the parameters involved follows the method proposed by Robinson (1988). Take the expectation of expression (7) conditional on $x$ to get:

$$E\left(y_i|x\right) = \boldsymbol{\beta_i'} E\left(\mathbf{z}|x\right) + F_i\left(x\right). \tag{8}$$

Now subtract expression (8) from expression (7) to obtain:

$$y_i - E\left(y_i|x\right) = \boldsymbol{\beta_i'}\left[\mathbf{z} - E\left(\mathbf{z}|x\right)\right] + \varepsilon_i. \tag{9}$$

Equation (9) can now be estimated using ordinary least squares regression (OLS). Since we have no observations on the conditional means of $y$, these are replaced by their nonparametric estimates. That is, we replace $E\left(y_i|x\right)$ by its estimate $\dfrac{\sum_{h=1}^{H} K\left(\dfrac{x_h - x}{b_x}\right) y_{ih}}{\sum_{h=1}^{H} K\left(\dfrac{x_h - x}{b_x}\right)}$, and $E\left(z|x\right)$ by $\dfrac{\sum_{h=1}^{H} K\left(\dfrac{x_h - x}{b_x}\right) z_h}{\sum_{h=1}^{H} K\left(\dfrac{x_h - x}{b_x}\right)}$, for each element $z$ of vector $\mathbf{z}$.

The function $F\left(.\right)$ can then be estimated from (8) as:

$$\hat{F}_i\left(x\right) = \hat{E}\left(y_i|x\right) - \hat{\boldsymbol{\beta}}_{\boldsymbol{i}}'\hat{E}\left(\mathbf{z}|x\right). \tag{10}$$

Alternatively, one can consider $F_i\left(x\right)$ in (7) as the conditional mean of $y_i - \boldsymbol{\beta_i'}\mathbf{z}$ given $x$. That is:

$$F_i\left(x\right) = E\left(y_i - \boldsymbol{\beta_i'}\mathbf{z}|x\right), \tag{11}$$

which can be estimated by nonparametrically regressing $y_i - \boldsymbol{\beta_i'}\mathbf{z}$ on $x$. To compute this, one can replace the vector of parameters $\boldsymbol{\beta}$by the vector of their estimates obtained from (9).[4] It can easily be verified that adding-up is automatically satisfied if the same bandwidth $b$ is used for all variables.

# 3 Semi-parametric imputation of expenditures

The techniques described in the previous section are well-known and can be found in any handbook on nonparametric econometrics (e.g. Härdle, 1990; Pagan and Ullah, 1999). Why then are these techniques not widely used in imputation? Yet, as we show in this section, it is perfectly possible and quite straightforward to do so.

[4] Note that both expression (10) and the procedure described in expression (11) yield consistent estimators for the function and should give similar results. For later imputation expression (11) is more interesting from a computational point of view. When using expression (10) imputation of the conditional mean of each of the demographic variables separately is needed, whereas using (11) only requires a single imputation step.

Intuitively, one might think of the imputation procedure as a missing values problem. That is, consider an expenditure survey with $H$ observations on expenditures, income and household characteristics $\mathbf{z}$ and an income survey with $J$ observations on income and (conceptually) the same household characteristics $\mathbf{z}$ as one file with a total of $H + J$ observations and $J$ missing values for the expenditures on each of the goods.[5] Each missing value will now be replaced by an imputed value that is calculated from the $H$ values that we do observe.

For the imputation of the parametric part we follow the normal procedure and use the estimated $\beta$-coefficients from (9) and apply them to the vector of household characteristics $\mathbf{z}$ in the income survey. It is the imputation of the nonparametric part in (7) that requires some more elaborate computations. The formal expression for the nonparametric part of the imputation of expenditures on commodity $i$ for a unit $j$ in the income survey with income $x_j$ is as follows:

$$\tilde{F}_i(x_j) = \left[ \frac{\sum_{h=1}^{H} K\left(\frac{x_h - x_j}{b_x}\right) \triangle y_{ih}}{\sum_{h=1}^{H} K\left(\frac{x_h - x_j}{b_x}\right)} \right], \tag{12}$$

where $H$ is the number of observations in the expenditure survey, $b_x$ the same bandwidth as in the estimation of (9). The expression $\triangle y_{ih}$ for each good $i$ is given by $\triangle y_{ih} = y_{ih} - \hat{\boldsymbol{\beta}}_{\boldsymbol{i}}' \mathbf{z}_h$. Applying (12) boils down to a nonparametric estimation of the function $F(.)$ where the points of estimation, $x_j$, are (the incomes of) the households in the income survey and the result are $J$ imputed values for the function $F(.)$ in the income survey. Expenditures can now be imputed by adding the parametric part of the imputation to the imputed values of the function $F(.)$. For each good $i$ and for every household $j$ in the income survey we then have:

$$\tilde{y}_{ij} = \tilde{F}_i(x_j) + \hat{\boldsymbol{\beta}}_{\boldsymbol{i}}' \mathbf{z}_j + \tilde{\varepsilon}_i, \tag{13}$$

where a tilde indicates an imputed value and a hat an estimated one. The vector $\mathbf{z}_j$ is again the vector of household characteristics, now for household $j$ in the income survey and $\tilde{\varepsilon}_i$ is an error term that we add to have the necessary variability in the imputed expenditures. Not adding this error term would boil down to an imputation of conditional means, with smaller standard errors and less variability. The choice whether or not to add an error term is entirely the researcher's to make and will probably depend on the type of analysis one wants to perform using the enriched data.

Adding error terms can be accomplished by taking random draws from the empirical distribution of residuals obtained from the estimation of regression (9) for each imputed value. That is, for each value that is imputed in the income survey a single value from the empirical distribution of residuals is randomly

[5] Obviously the income survey contains many more variables not registered in the expenditure survey. An imputation of expenditures would otherwise be useless.

drawn and added to the imputed conditional mean. Expenditures imputed with a negative value as a consequence can be replaced by zero and the remainder rescaled such that adding-up is not violated.

# 4 The data

The expenditure information we use comes from the 2001 Belgian Household Budget Survey, while the income dataset we use is an administrative file of tax returns for incomes earned in 2001. In what follows we will briefly describe the two datasets and point out their differences and how we dealt with them. The main difference of course is that one is survey data while the other is an administrative dataset. We will first describe the budget survey, followed by a brief description of the income data.

## 4.1 Household Budget Survey

The expenditure survey of 2000 is a sample of Belgian private sociological households. In this context a household is defined as all people that live together and who jointly make decisions concerning, for example, the household budget. Collective households such as convent communities, hospitals or prisons are not included in the expenditure survey. In total 3,816 households participated in the expenditure survey 2000 representing 8,892 individuals. We can further distinguish three broad categories of information in the survey.

- Household expenditures. These are always reported at the household level. Hence, we cannot attribute consumption expenditures to individual household members.
- Income. Amounts are reported by the household members personally. It are mostly net incomes that we observe in the budget survey. Some incomes that are not attributable to individual household members are reported for the household as a whole.
- Household characteristics. At the household level we find for example dwelling characteristics, number of children, etc. At the individual level it will mostly be relationship characteristics, such as the relation of a household member to the head of the household. The latter is considered the one who defends the household's interests and takes care of most of the administrative duties. Typically it is the person that has the highest income and contributes most to household income.

This information is collected by effectively contacting the respondents who fill out most of the requested information. As of 1999 a random sample of about

300 households is drawn each month. Those households then record all expenditures and income during that month. Additional questionnaires provide information concerning the dwelling and other socio-economic and demographic characteristics. This way we can think of the budget survey as a *continuous* survey.

## 4.2 Income dataset

The income dataset we use is an administrative file of tax forms filled in by *fiscal units* in 2001 (income earned in 2000). The file contains detailed income and tax information on 24,881 fiscal units, drawn at random from the administrative file which covers the whole population of fiscal units. The distinction between 'fiscal households' and sociological households is of course a crucial one. The file consists of tax units who actually received and returned a tax form. This has two important implications. First, we do not observe sociological households, since different records in the file with tax return data might actually belong to the same sociological household. The reconstruction of fiscal households into sociological ones is impossible from the information available in the income file at our disposal. Secondly, even if sociological households could have been constructed, we miss a considerable part of the population which does not receive and/or return a tax form.[6]

A way to proceed is to 'construct' fiscal units in the budget survey, i.e. to break down sociological households into fiscal households. However, since expenditures are recorded at the household level it was both theoretically and practically unfeasible to assign 'household' expenditures to individual household members, let alone to possibly different tax units in a sociological household. As a result we did not work with constructed fiscal units in the budget survey. We therefore neglect this lack of comparability in the household definition for the remainder of the paper. Needless to say that this lack of uniformity in the household definition may compromise the results reported here. Yet, we feel that the procedure described in the preceding sections, i.e. imputation via nonparametric estimation of Engel curves, is a viable and promising one and it would give better results when applied to datasets for which the observation units are comparable.

To apply the nonparametric Engel curve regressions on the income data, we do still need common explanatory variables in the two datasets. The following variables are common and can therefore be used: net disposable income, number of dependent children, number of other dependent persons, number of children younger than three years of age, civil status (married or not), region, age of the reference person and sex of the reference person.

[6] We estimate this to be about 11% of the population (see Decoster and Van Camp, 2002).

# 5 Results

In both the expenditure survey and the income data file observations with missing values for income and/or age of the reference person were left out of the analysis. Also observations with a reference person younger than 20 years have been discarded in the income data file. There were instances in the income dataset where the age of the head of the household was implausible, e.g. 1 or 2. The cut-off point was taken as the minimum age of the head of household observed in the expenditure survey, i.e. 20 years. This resulted in 3,207 observations for the expenditure survey and 23,820 observations for the income dataset.

Expenditures were aggregated into sixteen different commodity groups and imputed as such in the income file. Table 1 shows some summary statistics for the different aggregate commodities in the expenditure survey and the income file respectively. The results for the income file are calculated using the imputed values.[7] The last column of the table shows the percentage difference in mean observed (expenditure survey) and imputed values (income file) for each of the sixteen commodities as a percentage of the observed (mean) values. Clearly there are considerable differences between the observed and the imputed values.

One of the explanations for the considerable difference in observed and imputed mean expenditures might be the difference in units of observation (see section 4.2). As stated before, we were unable to obtain a sample of sociological households with fiscal information. Hence, we had no choice but to treat the fiscal households in the income file as sociological ones in the imputation procedure since in the expenditure survey expenditures are at the *sociological* household level. Disposable income and household size in the fiscal units on average being smaller than in more comprehensive sociological units, the lower average for imputed values does not come as a surprise.

A two-sample t-test, the results of which we do not report here, indicated a rejection of the null hypothesis of equal means in eleven out of the sixteen cases. Only for 'tobacco', 'maintenance', 'private transport', 'fuels (heating,...)' and 'car fuel (leaded, unleaded)' could the hypothesis of equal means not be rejected at a 95% confidence level or more.

To give a first impression of the *distribution* of the observed and imputed values, Table 1 also shows the 25th, 50th and 75th percentiles in each of the two files respectively. This immediately reveals an additional problem in the imputation and a second possible explanation for the lower average of the imputed values: the commodities for which we find a large percentage of households in the expenditure survey reporting zero expenditures. For tobacco, public transport

[7] In the imputation we also include age of the head of the household in the nonparametric part of (7) (see footnote 3). Instead of using a bivariate Gaussian kernel, we used the product of two univariate kernels. This boils down to assuming that age and income are independent, an assumption corroborated by the statistically insignificant correlation between the two in the budget survey.

and diesel, at least half of the households have zero expenditures. For heating this amounts to even more than 75%. We do not explore the explanations for these zero expenditures here: infrequency of purchase (e.g. heating) during the recall period, or corner solutions (e.g. tobacco).[8] But clearly a two-stage estimation process would be recommended here: firstly, a discrete choice model, secondly, for the positive expenditures a semiparametric model as the one set out above. We leave this as further research for the moment as it is not as straightforward as it seems. Furthermore, also the discrete choice part of the two-stage estimation could be estimated nonparametrically.

To have a detailed picture of how well the distribution of the imputed values mirrors the distribution of the observed ones, Figure 1 shows the density functions for the different commodities. We have restricted the estimation for the graphs of Figure 1 to the strictly positive amounts only. The dotted line in each of the graphs in the figure shows the density of the observed values (expenditure survey) whereas the solid line is the density function of the imputed values (income file). On the horizontal axis the expenditures in Euros are shown. The density functions were obtained by kernel density estimation.

[8] See Decoster and Vermeulen (1998) and Vermeulen (2003).

Table 1: Observed versus imputed values: Household Budget Survey versus Income dataset

| | Household Budget Survey | | | | | Income dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| good | mean (1) | sd | p25 | p50 | p75 | mean (2) | sd | p25 | p50 | p75 | % diff. [(2)-(1)]/(1) |
| Food | 3,486.64 | 2,015.02 | 1,917.80 | 3,152.91 | 4,590.29 | 2,767.79 | 1,777.39 | 1,449.03 | 2,561.64 | 3,854.36 | -20.62% |
| Beverages | 376.01 | 290.02 | 168.67 | 306.69 | 511.06 | 320.16 | 269.07 | 120.16 | 272.25 | 460.69 | -14.85% |
| Alcohol | 486.14 | 1,064.80 | 28.85 | 208.23 | 572.63 | 415.45 | 964.67 | 0.00 | 168.43 | 498.87 | -14.54% |
| Tobacco | 256.06 | 546.09 | 0.00 | 0.00 | 257.61 | 270.17 | 530.83 | 0.00 | 40.33 | 255.85 | 5.51% |
| Clothing | 1,478.95 | 2,004.14 | 118.10 | 772.54 | 2,019.24 | 1,314.67 | 1,768.98 | 0.07 | 742.91 | 1,854.38 | -11.11% |
| Rent | 5,556.89 | 2,469.48 | 4,125.34 | 5,205.76 | 6,351.03 | 4,808.16 | 2,363.60 | 3,297.14 | 4,503.37 | 5,887.23 | -13.47% |
| Maintenance | 591.57 | 2,274.56 | 0.00 | 41.65 | 511.65 | 610.27 | 2,392.25 | 0.00 | 141.00 | 490.03 | 3.16% |
| Energy | 1,182.45 | 732.14 | 687.46 | 1,014.98 | 1,511.46 | 1,034.52 | 682.76 | 576.23 | 926.86 | 1,354.66 | -12.51% |
| Private transport | 1,208.35 | 2,386.90 | 0.00 | 121.96 | 1,437.09 | 1,227.24 | 2,150.48 | 0.00 | 401.98 | 1,553.11 | 1.56% |
| Public transport | 170.34 | 439.21 | 0.00 | 0.00 | 130.89 | 188.85 | 412.45 | 0.00 | 55.58 | 166.39 | 10.87% |
| Health | 1,801.18 | 2,226.24 | 556.87 | 1,249.09 | 2,318.80 | 1,427.24 | 1,965.48 | 198.07 | 955.19 | 1,911.40 | -20.76% |
| Leisure | 4,548.92 | 4,706.49 | 1,720.28 | 3,313.25 | 5,895.30 | 3,755.18 | 4,041.77 | 1,084.82 | 2,873.35 | 5,183.90 | -17.45% |
| Fuels (heating, ...) | 297.20 | 1,349.42 | 0.00 | 0.00 | 0.00 | 333.91 | 1,284.41 | 0.00 | 0.00 | 158.07 | 12.35% |
| Diesel | 346.78 | 669.88 | 0.00 | 0.00 | 496.78 | 390.01 | 601.29 | 0.00 | 148.05 | 521.77 | 12.47% |
| Car fuel (leaded, unleaded) | 672.35 | 818.46 | 0.00 | 437.28 | 1,078.04 | 653.14 | 741.47 | 20.82 | 422.16 | 1,023.48 | -2.86% |
| Other | 3,739.62 | 5,087.10 | 1,248.19 | 2,416.07 | 4,758.96 | 2,876.58 | 4,627.51 | 314.86 | 1,849.01 | 3,952.28 | -23.08% |

sd is the standard deviation ; p25, p50 and p75 are the quantile values of the 25th, 50th and 75th percentiles respectively

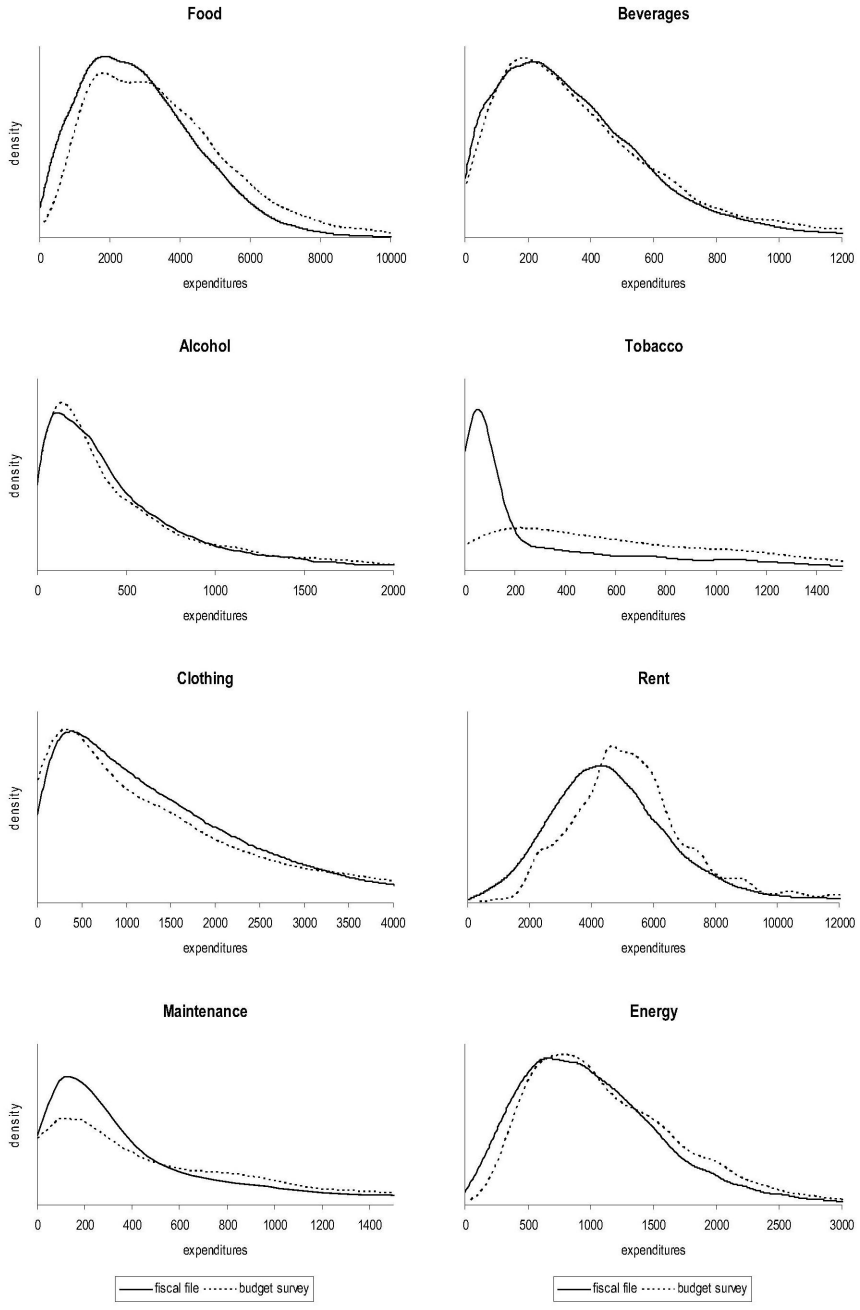


Figure 1: Density functions for the different commodity groups: observed (dotted) versus imputed (solid)

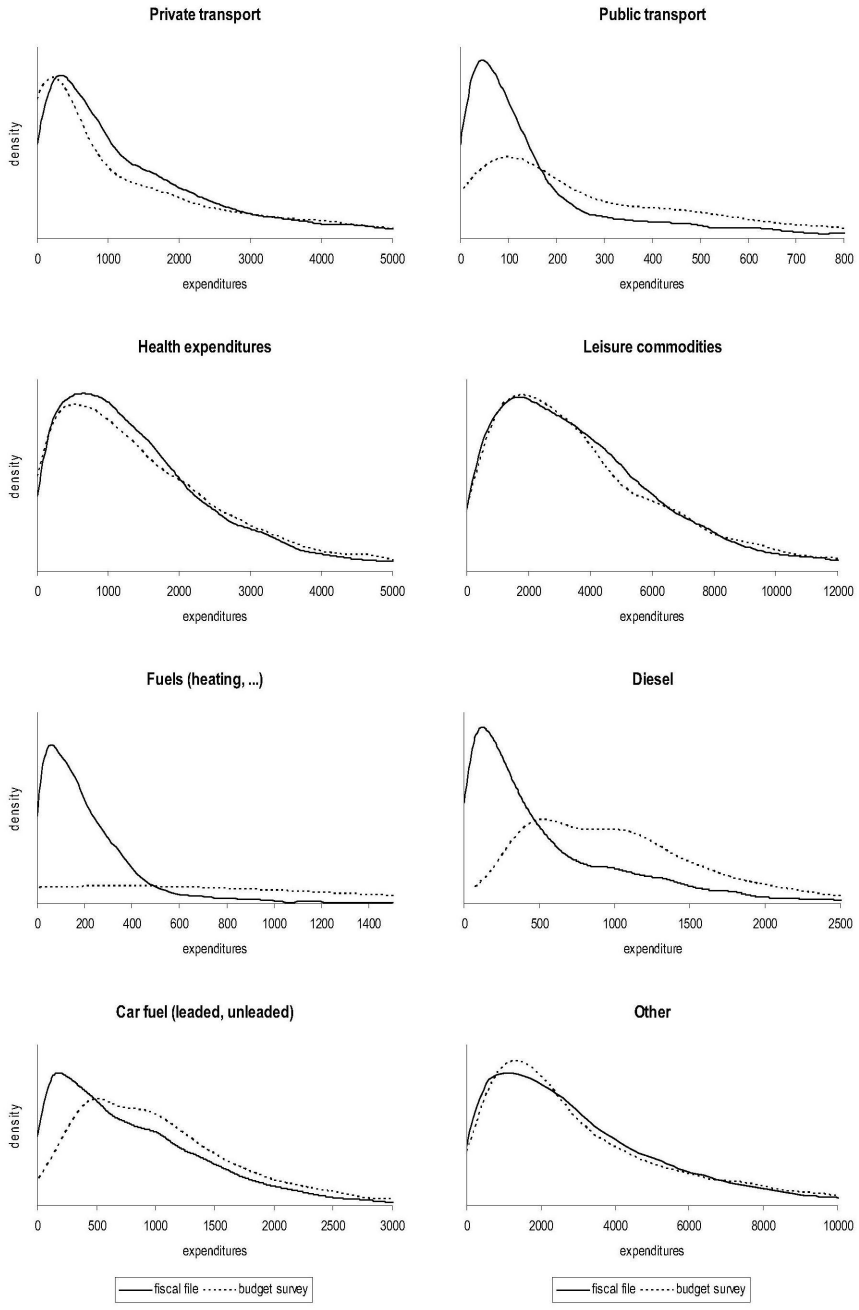


Figure 1: continued

In general, the density functions in both files are quite similar and close together for the majority of the goods analyzed. For food, beverages, alcohol, clothing, rent, energy, leisure commodities, health expenditures and private transport, we feel quite confident that our approach delivers satisfying results. This seriously mitigates the conclusion of Table 1 on the difference in the averages. Moreover, for the other commodities (tobacco, public transport, heating, diesel, and car fuel) we already predicted that the lack of a model which deals with zero expenditures, seriously hampers a trustworthy estimation. Table 2 shows the number and proportion of zero expenditures in the budget survey and after the imputation in the income file. On the one hand the commodities for which we find a diverging distribution in Figure 1 perfectly correspond with the ones who have a large proportion of zeroes in the budget survey. On the other hand the still larger proportion of zeroes in the income file for the non-problematic commodities follows from negative imputed values that were put equal to zero[9] (problematic commodities are those with a large proportion of zeroes in the expenditure survey). Since this probably follows from the addition of the error term in expression (13), more research is needed here, to explore specifications which preclude these error terms to turn negative ( a lognormal distribution seems to be a natural candidate for this).

Overall, the results show that this kind of imputation allows enriching an income file or survey with a distribution of detailed expenditures on which (policy) simulations can be run.

Table 2: Zero observations in the budget survey and after imputation in the income file

| good | Nonzero (budget survey) | Zero (budget survey) | Percentage (budget survey) | Percentage (income file) |
|---|---|---|---|---|
| Food | 3207 | 0 | 0.0% | 4.1% |
| Beverages | 3164 | 43 | 1.3% | 10.0% |
| Alcohol | 2486 | 721 | 22.5% | 32.4% |
| Tobacco | 1082 | 2125 | 66.3% | 34.5% |
| Clothing | 2669 | 538 | 16.8% | 25.0% |
| Rent | 3207 | 0 | 0.0% | 0.3% |
| Maintenance | 1737 | 1470 | 45.8% | 34.0% |
| Energy | 3207 | 0 | 0.0% | 1.6% |
| Private transport | 2089 | 1118 | 34.9% | 35.0% |
| Public transport | 1215 | 1992 | 62.1% | 31.0% |
| Health expenditures | 3128 | 79 | 2.5% | 19.4% |
| Leisure commodities | 3204 | 3 | 0.1% | 12.6% |
| Fuels (heating, ...) | 427 | 2780 | 86.7% | 54.5% |
| Diesel | 1008 | 2199 | 68.6% | 34.1% |
| Car fuel (leaded, unleaded) | 1982 | 1225 | 38.2% | 23.7% |
| Other | 3199 | 8 | 0.3% | 20.4% |

[9]Remember that in the graphs in Figure 1 we show the density functions for the strictly positive amounts only.

# 6 Conclusion

In this paper we described an alternative approach to impute expenditure information into an income dataset that lacks such information. In the literature on consumption theory nonparametric estimation of Engel curves has become common parlance. However, on imputation or even out-of-sample prediction using nonparametric techniques, the literature is rather silent. In this paper we have tried to fill in part of this gap by describing a methodology to impute expenditures in an income dataset within the framework of semi-parametric estimation. We illustrated the technique by imputing expenditure information from the 2001 Belgian Household Survey into a dataset with income information coming from tax returns. As this technique can be used to impute expenditure into any income dataset, it can certainly serve to enrich typical income data files that are used as input to tax-benefit microsimulation models. The enriched dataset then allows incorporating both indirect as well as direct taxes and benefits.

Indeed, often tax-benefit microsimulation models are developed based on representative income surveys. One reason is the focus of interest on changes in income as a consequence of changes in the tax-benefit legislation. Income surveys typically contain all the necessary information required for such analyses. Another reason, however, is that, while interested in the effects of both indirect and direct taxation, information on both is often not available in a single dataset. Having a technique available that helps 'combine' this information in a single dataset and that is founded in economic theory is appealing. Moreover, with the speed at which processing power of personal computers increases the computational burden of nonparametric techniques no longer can be – and will be even less so in the future – the main reason to abstain from using them.

There still is the issue of what is called the 'curse of dimensionality' in the literature on nonparametric regression that might be an argument against using nonparametric techniques. However, the same argument applies to the estimation itself. To circumvent this problem use is made of semi-parametric techniques, both in estimation as well as in the imputation technique described in this paper. Ideally one would want to estimate a fully nonparametric regression, i.e. imposing no structure whatsoever on the functional form, leaving the relation between expenditures and explanatory variables completely free. However, the 'curse of dimensionality', entails that such fully nonparametric estimation often requires (unrealistically) vast amounts of data, increasing with the dimension of the vector of explanatory variables. However, the techniques for imputation described in this paper can be easily extended to a fully nonparametric setting were such huge datasets available (or only a limited amount of explanatory variables sufficient).

The methodology presented here is but one of many that can be used to enrich income data with information on expenditures. It is therefore important to compare results obtained by using different techniques in a meaningful way, a topic that we are currently exploring. Other caveats that arise, and that may

be specific to the estimation of Engel curves and the use of budget surveys, is the occurrence of zero expenditures. Some will be the result of infrequency of purchase while others are the result of either corner solutions or non-existing preferences for the good in question. A case in point is expenditure on tobacco (e.g. Vermeulen, 2003). An obvious avenue for further research would then be to first estimate a discrete choice model, parametrically or nonparametrically, and to take account of the resulting probabilities in the imputation step. Also the replacement of negative imputed values by zeroes as briefly touched upon in the text is an ad-hoc way to deal with negative imputed expenditures.

# References


[1] Banks, J., Blundell, R., and Lewbel, A., 1997, "Quadratic Engel curves and consumer demand", The Review of Economics and Statistics, 79 (4), pp. 527-539

[2] Blundell, R. and Duncan, A., 1998, "Kernel Regression in Empirical Microeconomics", Journal of Human Resources, 33, 62-87

[3] Blundell, R., Browning, M. and Crawford, I.A., 2003, "Nonparametric Engel Curves and Revealed Preference", Econometrica, 71, 205-240

[4] Blundell, R., Duncan, A. and Pendakur, K., 1998, "Semiparametric Estimation and Consumer Demand", Journal of Applied Econometrics, 13, 435-461

[5] Decoster, A. De Swerdt, K., and Van Camp, G., 2004, "Matching of income and expenditure data by means of nonparametric estimation of Engel curves", Report of the D.W.T.C. project AG/01/79, Center for Economic Studies, Leuven

[6] Decoster, A. and Van Camp, G., 2002, "De Constructie van één Samengesteld Bestand op Basis van Twee Bestaande Bestanden: Koppeling van de Budgetenquête 1997-98 en het Fiscaal Bestand 1999 (inkomsten 1998)", DWTC-Project AG/01/030 Eindrapport Deel 2

[7] Decoster, A. and Vermeulen, F., 1998, "Modelling household consumption on micro data with a focus on the source of the zeroes", Discussion Paper Series, Center for Economic Studies, Leuven

[8] Härdle, W., 1990, Applied Nonparametric Regression, Cambridge University Press

[9] Leser, C., 1963, "Forms of Engel Curves", Econometrica, 31, 694-703

[10] Nadaraya, E.A., 1964, "On Estimating Regression", Theory of Probability and its Applications, 9, 141-142

[11] Pagan, A. and Ullah, A., 1999, Nonparametric Econometrics, Cambridge University Press

[12] Robinson, P.M., 1988, "Root-N Consistent Semiparametric Regression", Econometrica, 56, 931-954

[13] Schmalensee, R. and Stoker, T.M., 1999, "Household Gasoline Demand in the United States", Econometrica, 67, 645-662

[14] Vermeulen, F., 2003, "Do Smokers Behave Differently? A Tale of Zero Expenditures and Separability Concepts", Economics Bulletin, Vol.4, No.6, 1-7

[15] Watson, G., 1964, "Smooth Regression Analysis", Sankhya, 26, 359-372

[16] Working, H., 1943, "Statistical Laws of Family Expenditure", Journal of the American Statistical Association, 38, 43-56

[17] Yatchew, A., 1998, "Nonparametric Regression Techniques in Economics", Journal of Economic Literature, Vol.36, 669-721