MATCHING OF INCOME AND EXPENDITURE DATA BY MEANS OF NONPARAMETRIC ESTIMATION OF ENGEL CURVES

ANDRÉ **DECOSTER**

KRIS **DE SWERDT**

GUY **VAN CAMP**

CENTER FOR ECONOMIC STUDIES - KULEUVEN

MARCH 2004

REPORT OF THE D.W.T.C.-PROJECT AG/01/079

Abstract: In DWTC-project AG/01/30 use was made of statistical matching techniques to attribute fiscal information from an administrative fiscal file to households in the household budget survey. In the current project, we *impute* consumption expenditure data from the household budget survey in a fiscal file by means of semiparametric estimation of Engel curves on expenditure survey data (Robinson, 1988).

Acknowledgments: The authors are grateful to Bart Capéau, Frederic Vermeulen and Laurens Cherchye for helpful comments. Of course, none of these can be held responsible for any errors in the report.

1 INTRODUCTION

The research project AG/01/079 of which the present report is the result, is itself a "follow-up"-project of a similar project on statistical matching (DWTC-project AG/01/030, finished in 2002). In both the current and the past project the major aim is to construct a file with Belgian microdata on expenditures and (taxable) incomes. Indeed, in many relevant simulations of policy changes, such as simultaneous changes in personal income taxes or social security contributions on the one hand and indirect taxes on the other, the availability of microdata on both income and expenditures is highly recommended if one wants to assess the distributional consequences of these changes. However, no such single file, containing information on both expenditures and (gross) incomes currently exists in Belgium. There is a file with very detailed expenditures for a representative sample of Belgian households: the budget survey of the National Statistical Institute. And several files contain gross or net income information. In the previous matching project we did succeed in constructing one matched dataset.[1] We first briefly recall how we proceeded in this previous project, to identify the aim of the current project as an improved version of this first matching exercise. We then explain how the current project had to be redefined in the course of the project due to problems with the delivery of the necessary income data.

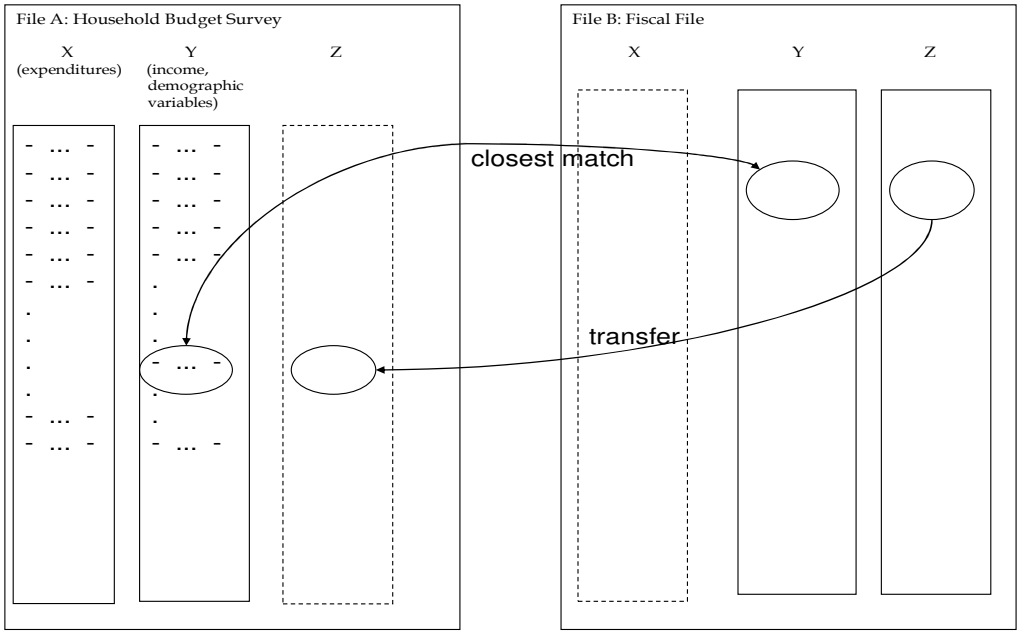
Since in both projects we have combined information from the household budget survey with that from an (administrative) fiscal file, we will refer from now on to these two files as the "expenditure survey" and the "fiscal file". If we would have a unique identifier for households appearing in the two files, i.e. the household budget survey and the fiscal file, there would be no problem in combining the two sources of information (provided we could retrace each household in the budget survey in the administrative fiscal file). For example, if in the two files individual household members would have been identified by their respective national registry numbers, we could easily find the matching number in the other file and complete the necessary data. However, due to obvious legal and privacy reasons we do not dispose of this kind of information. Other ways will have to be found to identify the most similar records in the two files.

In the DWTC-project AG/01/030 (see Decoster and Van Camp, 2002) *a statistical match* has been performed between the Belgian Household Budget Survey 1997-98 and a Fiscal File with income data of 1998. The essence of a statistical match is illustrated in figure 1.

---

[1] The usefulness of the matching technique has been proven by the use which has been made of it in another DWTC –project on equity in the financing of healthcare (Project S0/01/005). Use of the matching technique and the previously matched file made it possible to look at how the different components that contribute to the financing of health care, such as social security contributions, personal income taxes, indirect taxes and private payments, are distributed over the population. See De Graeve et al. (2003) for more details.

FIGURE 1 A STATISTICAL MATCH BETWEEN INFORMATION IN THE HOUSEHOLD BUDGET SURVEY AND THE FISCAL FILE



File A, the expenditure survey, contains information on expenditures, represented by matrix X (with n_A rows for the number of households, and m_A^X columns for the number of expenditure categories). In file B, the fiscal file with n_B observations, there is no expenditure information, represented in figure 1 by the dotted borderline for matrix X. Matrix Y contains the variables with information that may be considered as "common" to both datasets. Examples of the Y-variables are sociodemographic information (size of the household, age of the members of the household, place of residence, etc…). In the past DWTC-project also net income information (income after taxes) could be considered as such common information. Matrix Z contains the variables which are missing in the expenditure survey, e.g. income components before taxes. They are available in the fiscal file however. The purpose is to construct a file, call it C, that contains information on X, Y and Z.

One way to proceed, is to supplement the expenditure survey with information from the fiscal file in the following way. For each record in file A, search in file B for the record that most resembles the record under consideration in file A. Resemblance is defined in terms of the overlapping variable(s) in Y. In statistical matching, often a distance function is used to give meaning to the term 'resemblance'. As an example consider the following distance function between the record i in file A and record j in file B:

$$d\left(y_A^i, y_B^j\right) = \sum_l w_l \left| \frac{y_A^{il} - y_B^{jl}}{\sigma_l} \right| \qquad (1)$$

where y_A^{il} denotes the l-th element in the vector of overlapping variables for record y_A^i, y_B^{jl} the l-th element in the vector of overlapping variables for record y_B^j, σ_l the standard deviation of variable l taken over the two files, and w_l a weight attributed to component l of the overlapping variables in the distance function. The record in file B which minimizes the distance in (1) is chosen as the closest match. The values of the variables Z for this record in file B are then transferred to the corresponding record in file A. This procedure is repeated for every record i in file A.

Evidently, one is confronted with many more practical problems in a specific application. One of the obvious ones is the comparability of variable definitions, and a fortiori the comparability of the household definition. But also the choice of weights and other forms of distance functions have to be considered. We refer the reader to Decoster and Van Camp (2002), and Van Camp (2002) for an extensive treatment of many of these issues. One of these additional "problems" in the first matching project had to do with the fact that the fiscal dataset at our disposal was not representative for the population of sociological households, but only for the population of *fiscal* households who had filled in a tax form. A serious improvement could be reached by working with a *truly representative fiscal dataset*. **Replicating the previous matching exercise with a better fiscal dataset and a more recent budget survey was the core objective of this second matching project**.

The aim was to draw a sample of sociological households (contrary to the "fiscal" households in the previous project) from the National Registry, and to collect fiscal information for all household members of each sampled sociological household. We would then redo the statistical matching exercise with this fiscal dataset and the budget survey of 2000. We devoted a considerable amount of work to the design of the correct sampling procedure. Yet, the fiscal sample, was not yet at our disposal at the time of writing.[2] This forced us to conceive an alternative interpretation of the project description: *testing out an alternative methodology to match two datasets* than the statistical match of the previous project. The result is an enriched fiscal file, consisting not only of the income tax information, but also of expenditures.

The rest of this report is structured as follows. In section 2 we give an overview of the Household Budget Survey 2000, its characteristics and representativeness for the Belgian population. Section 3 will briefly describe the fiscal data that we used in this exercise. The theoretical details of the technique used in this text are discussed in section 4. Section 5 describes how these techniques were applied to data from the Household Budget Survey and data from the Fiscal File. Results will be shown in section 6. Finally section 7

[2] The document "Voorlopig rapport AG0179.doc", sent by email on Monday September 15, 2003 to Mr. Aziz Naji and Mr. Christian Valenduc, documents the efforts we undertook in order to obtain this file in due time, and is available from the authors upon request.

concludes. In Appendix 1 we describe the constructed data file and its contents. Appendix 2 briefly discusses what has been done in anticipation of a fiscal file that was requested but not yet at our disposal at the time of writing and which was part of the project proposal. That is, the exercise described in this text was performed using a different fiscal file than originally envisaged. The work done in anticipation of the file demanded is therefore not (entirely) relevant to the technique that we will describe in this text and is therefore deferred to an appendix.

2 THE HOUSEHOLD BUDGET SURVEY 2000

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.

1. Household expenditures. These are always reported at the household level. Hence, we cannot attribute consumption expenditures to individual household members.
2. 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.
3. Household characteristics. At the household level we find for example dwelling characteristics, number of children, …. 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*. We will now discuss the three broad categories of information in the expenditure survey in more detail.

2.1 HOUSEHOLD CHARACTERISTICS

First, we will show some distributional aspects concerning the population in the expenditure survey of 2000. We will verify whether or not the expenditure survey can be considered to be representative of the Belgian population. In Table 1 we show how the population in the expenditure survey is distributed with respect to age categories. In columns 4 and 5 we show the numbers and percentages reported by the National Statistical Institute.[3] As the table shows, according to the expenditure survey, extrapolated to the entire population, there were about 10,381,312 individuals in Belgium in the year 2000. Compared to the official statistics this entails a slight overestimation of the total number of inhabitants. Looking at the age categories separately we notice overestimations

[3] See URL http://www.statbel.fgov.be and the links therein.

as well as underestimations. The percentage differences between the expenditure survey and the official statistics are all less than 1 percent except for three categories. The age categories between 5 and 10 and between 10 and 15 are both overrepresented in the expenditure survey by more than 1.5 percentage points. The number of persons in the age group between 85 and 90 are underrepresented by slightly more than 1 percentage point in the expenditure survey relative to the official statistics. Overall we can conclude that , based on this criterion, the expenditure survey mirrors reasonably well the Belgian population as a whole.

TABLE 1 BELGIAN POPULATION: AGE CATEGORIES

Age Category	HBS2000[(a)]		Official statistics		Dif-tot (1)-(3)	Dif-perc (2)-(4)
	Subtotal (1)	Percent (2)	Subtotal (3)	Percent (4)		
<5	566,652	5.46	530,486	5.20	36,166	0.26
>=5 and <10	741,908	7.15	573,932	5.60	167,976	1.55
>=10 and <15	790,692	7.62	597,041	5.83	193,651	1.79
>=15 and <20	708,842	6.83	620,596	6.06	88,246	0.77
>=20 and <25	555,320	5.35	619,574	6.05	-64,254	-0.70
>=25 and <30	585,109	5.64	678,939	6.63	-93,830	-0.99
>=30 and <35	750,438	7.23	751,324	7.34	-886	-0.11
>=35 and <40	890,688	8.58	824,139	8.05	66,549	0.53
>=40 and <45	883,804	8.51	798,581	7.80	85,223	0.71
>=45 and <50	674,086	6.49	736,567	7.19	-62,481	-0.70
>=50 and <55	642,464	6.19	693,128	6.77	-50,664	-0.58
>=55 and <60	561,230	5.41	528,291	5.16	32,939	0.25
>=60 and <65	506,714	4.88	537,852	5.26	-31,138	-0.38
>=65 and <70	451,600	4.35	531,048	5.19	-79,448	-0.84
>=70 and <75	527,647	5.08	463,682	4.53	63,965	0.55
>=75 and <80	402,319	3.88	380,143	3.71	22,176	0.17
>=80 and <85	100,042	0.96	170,961	1.67	-70,919	-0.71
>=85 and <90	23,965	0.23	134,480	1.31	-110,515	-1.08
>=90 and <95	14,852	0.14	54,329	0.53	-39,477	-0.39
>=95	2,940	0.03	13,992	0.14	-11,052	-0.11
Total	10,381,312	100.00	10,239,085	100.00	142,227	0.00

(a)HBS2000 is short for Household Budget Survey of the year 2000.

In Table 2 the number of households in the expenditure survey is shown for different household sizes, again extrapolated to the entire Belgian population, and the corresponding figures that can be found in the official statistics of the National Statistical Institute.[4] We notice a slight discrepancy between the expenditure survey and the official statistics. According to the expenditure survey there were 4,260,111 households in Belgium whereas the official statistics indicate 4,237,773 households in the year 2000.

[4] Households in both sources of information are private households. In the comparison of the figures no adjustments have to be made for collective households.

More in particular, we observe an underrepresentation in the expenditure survey of the number of 'small' households and an overrepresentation of the 'larger' households, except for the largest families with eight or more members. Percent-wise the two distributions do not seem to be that different.

TABLE 2 BELGIAN POPULATION: HOUSEHOLD COMPOSITION, NUMBER OF MEMBERS

Number of members	HBS2000[a]		Official Statistics		Dif-tot (1)-(3)	Dif-per (2)-(4)
	Subtotal (1)	Percent (2)	Subtotal (3)	Percent (4)		
1	1,308,685	30.7	1,321,597	31.2	-12,912	-0.46
2	1,292,446	30.3	1,317,807	31.1	-25,361	-0.76
3	679,125	15.9	705,376	16.6	-26,251	-0.70
4	612,716	14.4	585,067	13.8	27,649	0.58
5	244,182	5.7	214,339	5.1	29,843	0.67
6	92,773	2.2	63,057	1.5	29,716	0.69
7	21,694	0.5	18,013	0.4	3,681	0.08
8 or more	8,490	0.2	12,517	0.3	-4,027	-0.10
Total	4,260,111	100.0	4,237,773	100.0	22,338	0.0

(a)HBS2000 is short for Household Budget Survey of the year 2000.

2.2 EXPENDITURES[5]

The expenditure survey provides a rich source of consumption expenditure information. In Table 3 and Table 4 the number of different expenditure codes that are found in the expenditure survey are shown. Each code represents some consumption good or category of goods. There are eight broad expenditure categories that can be identified in the expenditure survey and those are listed in the first column of Table 3. The total number of codes we find in each category and its percentage of the total number of codes are shown in columns 2 and 3 respectively. As can be seen, the expenditure survey collects information on consumption expenditures for 966 different (categories of) goods. By far the largest broad category is that of food, beverages and tobacco which accounts for over a third of the total number of codes available. The housing and health categories are those for which the least number of different codes are available, representing respectively 2.4 and 3.3 percent of the total.

In Table 4 we report the number of households and the percentage, in terms of the total number of households, which report expenditures for at least one item in the respective broad categories. From this table it can be seen that all households spend part of their budget on at least one of the items in the category food, beverages and tobacco. Also all households report house-related expenditures. Of all the households 80% spend part of their budget on clothing or shoes. Given the broadness of this category -it is the second

[5] Note that all amounts reported in this and the next section are in Belgian francs and not in Euro.

largest as can be seen from Table 3- we would have expected a higher occurrence of spending in this category. However, since households participating in the expenditure survey are only questioned during one month, this figure may not be so surprising after all. Households' expenditures on clothing might be concentrated around sales periods.

TABLE 3 NUMBER AND PERCENTAGE OF EXPENDITURE CODES FOR 8 DIFFERENT CATEGORIES IN THE HOUSEHOLD BUDGET SURVEY 2000

Category Description	Number	Percent
1 Food, Beverages and Tobacco	350	36.2
2 Clothing and Footwear	140	14.5
3 Housing	32	3.3
4 Furniture and Appliances	109	11.3
5 Health	23	2.4
6 Transport and Communication	89	9.2
7 Culture, Free Time and Education	120	12.4
8 Other Goods and Services	103	10.7
Total	966	100.0

TABLE 4 NUMBER AND PERCENTAGE OF HOUSEHOLDS REGISTERING AT LEAST ONE EXPENDITURE IN CATEGORY *I* IN THE HOUSEHOLD BUDGET SURVEY

Category Description	Average expenditures	Average budget share (in %)	Percent of households reporting expenditures
1 Food, Beverages and Tobacco	178,346.30	17.49	100.0
2 Clothing and Footwear	69,553.81	5.53	80.0
3 Housing	298,657.70	30.93	100.0
4 Furniture and Appliances	76,492.38	5.70	98.6
5 Health	55,233.58	5.13	86.8
6 Transport and Communication	193,961.90	12.90	95.8
7 Culture, Free Time and Education	101,491.30	8.54	99.6
8 Other Goods and Services	201,404.60	16.31	99.5
Total	1,143,389.00		

2.3 INCOME

The income amounts in the expenditure survey are reported for each member of the household and are net of taxes. We distinguish four broad income categories in the expenditure survey: income from an economic activity, income from social security, capital income and income from other sources.

In Table 5 and Table 6 we give an overview of the distribution of income from economic activity and income from social security respectively for individual household members. The percentage in the last column represents the percent of the population of individuals reporting a strictly positive amount for the corresponding income source.

Firstly, the low figure for "vacation allowance" is rather striking. For this figure we would expect a percentage that is of the same order of magnitude as that of "income from labour". One would expect that all those who work, or at least the majority of them, also receive a vacation allowance. We have no explanation for this.[6] Almost one third of all the individuals report an income from labour (salaried income) whereas individuals engaging in independent activities seem to form only a small minority in the expenditure survey. The mean income of the latter category, on the other hand, is much higher than that of salaried workers: 1,346,297 Belgian francs for independent workers versus 664,576 for salaried workers. Table 6 shows that about 17.5% of the individuals receive a pension income and that 5.7% receive some sort of unemployment benefit. Notice also that an unemployed person gets, on average, less than half of what a working person earns on average.

TABLE 5 DISTRIBUTION OF INCOME FROM ECONOMIC ACTIVITY: INDIVIDUAL HOUSEHOLD MEMBERS

Type of Income	Minimum	Maximum	Mean	Percent
Salary	4,500	3,669,768	664,576	30.1
Vacation Allowance	8,628	1,780,488	441,424	2.5
Other Income from Labour	780	3,120,000	178,214	6.6
Contributions Individual Household Members	12	720,000	115,811	2.3
Independent Activity	30,000	10,992,936	1,346,297	5.0

TABLE 6 DISTRIBUTION OF INCOME FROM SOCIAL SECURITY: INDIVIDUAL HOUSEHOLD MEMBERS

Type of Income	Minimum	Maximum	Mean	Percent
Pension	1,476	2,249,640	459,844	17.5
Unemployment Benefits	13,500	871,104	298,983	5.7
Disability Allowance	14,124	902,616	365,899	2.3

Table 7 and Table 8 are similar to Table 5 and Table 6 but the unit of analysis here is the household rather than its members. Table 7 reveals that almost three quarters of the Belgian households get their income from salaried labour whereas 15.2% get their main income from an independent activity. The low percentage of households reporting a vacation allowance only confirms what we found in Table 5. Almost ten percent of Belgian households report income from own produce. This can apply to farmers, keeping part of their produce for own consumption or entrepreneurs who keep part of production for own and household consumption. But it can equally well apply to households growing their own vegetables for example. Only 6.1% of all households report to receive some kind of extra-legal benefits and almost 7% earn extra income from side activities. Table 8 shows that nearly 45% of Belgian households have at least one member that receives a pension income and nearly 14% have at least one member that is unemployed. Far over a third of Belgian households receive family benefits (child allowances).

[6] Most probably people report their vacation allowance as part of their salary.

TABLE 7 DISTRIBUTION OF INCOMES FROM ECONOMIC ACTIVITY: HOUSEHOLD LEVEL

Type of Income	Minimum	Maximum	Mean	Percent
Salary	4,500	3,669,768	664,576	73.3
Vacation Allowance	8,628	1,780,488	441,424	6.0
Other Income from Labour	780	3,120,000	157,349	18.3
Contributions Individual Household Members	12	720,000	115,811	5.7
Independent Activity	30,000	11,000,000	1,387,314	15.2
Extra Legal Benefits (employees)	864	1,785,600	178,539	6.1
Extra Legal Benefits (independent activity)	2,640	833,460	184,363	2.6
Side Activities : Own Produce	192	255,480	17,409	9.6
Side Activities : Other Income	-868,212	2,984,400	114,839	6.8

TABLE 8 DISTRIBUTION OF INCOMES FROM SOCIAL SECURITY

Type of Income	Minimum	Maximum	Mean	Percent
Pension	1,476	2,249,640	438,430	44.9
Unemployment Benefits	13,500	871,104	298,354	13.8
Disability Allowance	6,816	902,616	346,583	6.2
Child and Other Allowances	9,216	696,300	127,534	38.5
Payments by Medical Funds	588	435,000	39,659	38.1
Other Social Benefits	456	868,908	159,702	6.1

In Table 9 we report income from capital, such as real estate, equity etc…. We also include fictitious income for households that own their own home (imputed rental value) and also cadastral income is taken up. The table shows 71% of Belgian households to be homeowner and this figure corresponds well with the percentage of households that have reported a value for the cadastral income, even though some of the minimum amounts reported might seem implausible. In previous editions of the household budget survey the percentage of households reporting a value for cadastral income did not always correspond to the percentage of homeowners (Decoster and Van Camp, 2002).

TABLE 9 INCOME FROM CAPITAL

Type of Income	Minimum	Maximum	Mean	Percent
Gross Revenues from Capital	8,040	8,512,788	371,879	8.7
Fictitious Income Owner Occupied Dwelling	90,000	600,000	229,458	71.0
Property Value Taxes	-1,617,300	-12	-231,269	5.2
Expenses for Property Let	-1,522,152	-2,136	-202,903	1.5
Net Income from Equity	24	3,507,060	136,817	7.4
CI (Cadastral Income)	99	361,700	37,613	72.0

Finally, in Table 10 we show the figures for other income items that are also reported at the household level and that are not classified under either of the following categories:

income from economic activity, income from social security and income from capital. It concerns alimony received and paid, amounts received through life insurances, other insurance payments, other amounts received and amounts to be subtracted for money lost or received in excess of what was to be received. Only a minor percentage of households seem to pay or receive any form of alimony. Similarly, only a small percentage of households report to have received money stemming from insurance contracts. Overall, this category of income seems to be represented by only a small part of the total Belgian household population (except then for "other amounts received").

TABLE 10 OTHER INCOME: HOUSEHOLD LEVEL

Type of Income	Minimum	Maximum	Mean	Percent
Alimony Received	7,813	639,324	139,193	4.7
Alimony Paid	-494,904	-12,000	-122,849	3.7
Life Insurance : amounts received	7,392	3,232,116	420,647	0.6
Other Insurance : amounts received	1,104	3,907,500	243,146	2.6
Other Amounts Received	600	3,660,000	49,672	11.1
Lost Money/Money Wrongfully Received	-2,459,412	-240	-82,017	1.1

3 FISCAL FILE

The fiscal file 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 fiscal file might actually belong to the same sociological household. The reconstruction of fiscal households into sociological ones is impossible from the information available in the fiscal file at our disposal.[7] 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.[8] Both facts make it useless to check the representativeness of the fiscal file with respect to external sources of information on sociological households (as we did with the expenditure survey).

In the previous matching exercise the budget survey was supplemented with income information from the fiscal file. To find the most comparable units in the fiscal file the sociological households in the budget survey were deconstructed into fiscal units. Since the matching exercise of this report rests on an imputation procedure of expenditures into the fiscal file and because it was both theoretically and practically unfeasible to assign "household" expenditures to individual household members, let alone to tax units into a sociological household, 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 rest of the report.[9] 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 following 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 fiscal file, 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 children at charge, number of other persons at charge, 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 .

---

[7] This reconstruction was one of the major aims of the fiscal file that had to be delivered by the FOD Financiën to carry out the original research proposal.

[8] We estimate this to be about 11% of the population (see Decoster and Van Camp, 2002). Also this weakness of the fiscal file used here, would have been cured, had the promised file been delivered.

[9] Although we work with sociological households in the budget survey, we did only retain the households for which at least one of the members was eligible for take up in the fiscal file. For more details see Appendix A.2.

4 NONPARAMETRIC ESTIMATION OF ENGEL CURVES

The imputation of expenditures in the fiscal file will be based on a relationship between expenditures and explanatory variables, such as income, demographic variables (age, household size, etc….) available in both the fiscal file and the expenditure survey. In fact this is nothing else than estimating Engel curves and using them for the imputation. The novelty of the approach explored here, however, lies in the fact that we rely on *nonparametric* techniques for both the estimation and the imputation, contrary to the usual parametric techniques based on specific functional forms. Therefore in this section we first shortly describe Engel curves, and then explain how we have estimated and used them nonparametrically.

4.1 ENGEL CURVES

The most straightforward way to describe Engel curves is as a *graphical representation of the relationship between consumption of an item and income or total expenditures*. In general Engel curves will be positively sloped, indicating that more income leads to more consumption. The only exception to this finding is when goods are inferior in which case the Engel curves will have a negative slope.

Often, however, also the relation between the *budget share* of a commodity and (the logarithm of) total expenditures is called an Engel curve. In this case the slope will be positive only for luxury goods, where consumers will spend more of their budget, as a percentage, on those kinds of goods as they get richer. Budget shares of necessities and inferior goods will be declining in total expenditures.

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

$$w_i = \alpha_i + \beta_i \log(x) \tag{2}$$

where w_i represents the budget share on good i, x are total expenditures and α and β are parameters to be estimated.

This linear specification has been questioned by others, claiming it might be plausible for some goods but does not necessarily apply to all goods. Banks et al. (1997) explored this issue further using quadratic terms in expenditure in the specification. Visual analysis of Engel curves led them to conclude that for some (categories of) goods a quadratic specification might be more reasonable. This has formed the basis of their QUAIDS demand system (see also Blundell et al., 2000 and Blundell et al., 2003).

The preliminary evidence in Banks et al. (1997), to advocate a revision of the functional form of the Working-Leser Engel curves was itself based on a nonparametric estimation of

an Engel curve. But the question then arises why at all, we would still rely on parametric estimations to impute expenditures (or budget shares). Indeed, the major advantage of nonparametric analysis being the lack of a straitjacket of a functional form, it is unclear why we would give up this advantage if it comes to imputation.

Therefore, in the next section we will review the nonparametric regression technique in more detail, and explain how we have used it to impute expenditures into the fiscal file.

4.2 NONPARAMETRIC ESTIMATION OF ENGEL CURVES

In parametric analysis one imposes a structure on the Engel curve and estimates the parameters that are part of the specific functional form used. For example, in the above Working-Leser specification (2), one would estimate the parameters α and β, under the assumption that the function specified is the correct one. In nonparametric analysis no functional form is imposed on the data at hand, or the relationship(s) between them. Rather, if one wants to estimate some (unknown) function, $f(x)$, this is done *directly* from the data, without imposing any structure on the function f.

4.2.1 Kernel density estimation with adaptive kernels

We will start with explaining the principles of nonparametric estimation in the case X is a discrete random variable. The reason is that this will make some concepts intuitively clear.

Let X be a discrete random variable that generates data $x_h, h = 1,...,H$. We wish to estimate the density $f(x)$ at some point x. One way to proceed would be to create an interval of length b around x and count the percentage of observations x_h, lying in this interval. This would give us an idea of the density at the point x. More formally one could write:

$$\hat{f}(x) = \frac{1}{Hb}\sum_{h=1}^{H} I\left(x - \frac{b}{2} \le x_h \le x + \frac{b}{2}\right) \tag{3}$$

where b is the length of the interval we are considering and $I(.)$ is an indicator function that takes the value one if its argument evaluates to true and zero otherwise. From (3) it is clear that we are averaging or *smoothing* the points in the neighbourhood of x, where the locality is determined by the width of the interval b. It is the latter that determines by what amount the data are averaged. The estimator in (3), however, is discontinuous at the points $x \pm \frac{b}{2}$ and in general it will be rather rough as an approximation of the density. One therefore has sought for a procedure that leads to a smooth and continuous approximation of the true density.

Rosenblatt (1956) was the first to address this issue and proposed the following alternative

$$\hat{f}(x) = \frac{1}{Hb}\sum_{h=1}^{H} K\left(\frac{x_h - x}{b}\right) \tag{4}$$

where $K(.)$ is some kernel smoothing function that is continuous and integrates to one. Several functional forms have been proposed in the literature for the smoothing function $K(.)$ and we refer the reader to Härdle (1990) for an overview. In this text we will work with the so called Gaussian kernel, where the function $K(.)$ takes the form of a standard normal distribution, that is:

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right) \tag{5}$$

It has been pointed out in the literature that the choice of smoothing function is of less importance, i.e. different smoothing functions will yield similar results. What is of crucial importance though in nonparametric estimation is the choice of the window- or bandwidth b.

One can intuitively feel that large bandwidths will produce estimates of the density that will be too smooth while the opposite holds for bandwidths that are too narrow.

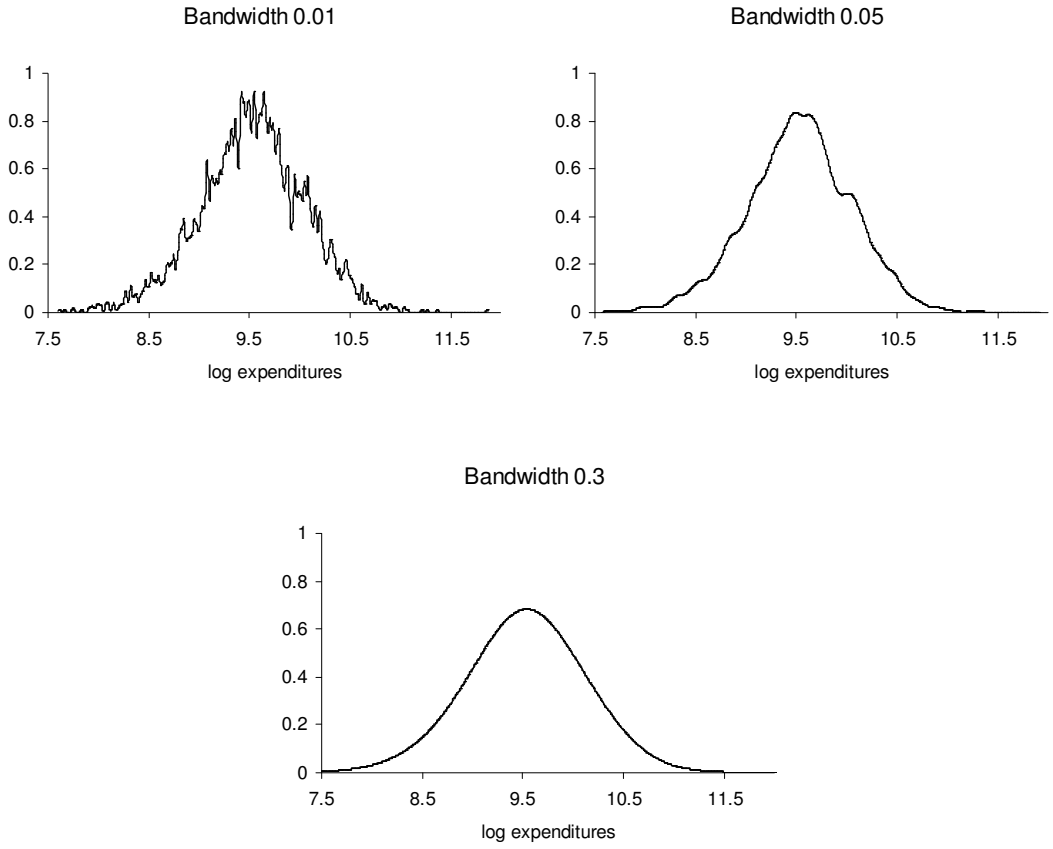
To give some idea of the influence of the bandwidth on the density estimation we present in Figure 2 an estimation of the density of the logarithm of total expenditures where we used a different bandwidth for each estimation. As is clear from the figures, a very small bandwidth shows too much detail in the function. In fact, a bandwidth equal to zero will merely replicate the observed data. A large bandwidth on the other hand will lead to a very smooth function as can be seen in the last graph labeled "Bandwidth 0.3" in Figure 2, but might suffer from serious bias.

Hence, the choice of bandwidth has to face the trade-off between bias and variance. That is, a too large bandwidth will lead to a function that is too smooth with very little variability while a (too) small bandwidth will lead to under-smoothing and hence too much variability and noise. There are several ways to determine the 'optimal' bandwidth, one of which we will describe below after we have dealt with nonparametric regression. Here we present as a rule of thumb a formula which is easy to implement and which will be used as a starting point in the procedure to be described later. The formula reads as follows (e.g. Deaton, 1997):

$$b^* = 1.06 * \min\left(\sigma, 0.75 * IQR\right) H^{-\frac{1}{5}} \tag{6}$$

where b^* denotes the optimal bandwidth, σ the standard deviation, H the number of observations and IQR the interquartile range, i.e. the ratio between the quantile values for the 75th and 25th percentiles.

FIGURE 2 DENSITIES WITH DIFFERENT BANDWIDTHS



The bandwidth thus defined is the same for all points at which we wish to estimate the density. However, typically, the data are sparser in the tails of the distribution than they are near the mode for example. Using the same bandwidth for observations in the tail as for observations in the middle of the distribution, will lead to less observations in the neighbourhood of the estimation point x in the tail. Thus it seems desirable to have a different bandwidth for estimation points where the density of observations is rather sparse. A kernel estimation procedure taking these considerations into account might look like this:

$$\hat{f}(x) = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{b_x}K\left(\frac{x_h - x}{b_x}\right) \tag{7}$$

where b_x can be thought of as the bandwidth corresponding to the point of estimation x. As such the bandwidth is made to vary with each point at which the density is being

estimated. This estimator is called an adaptive kernel estimator and it will be used in our application. One way to implement this is by specifying b_x as follows:

$$b_x = b * \delta_x, \quad \delta_x = \left[\frac{\tilde{f}(x)}{G} \right]^{-\lambda} \tag{8}$$

where b is the bandwidth used to get a preliminary estimate of the density, say $\tilde{f}(x)$, G is the geometric mean of these preliminary estimates at all x and λ $(0 < \lambda \leq 1)$ is a sensitivity parameter determining the degree of adaptation introduced. In the literature it is suggested that in general $\lambda = 0.5$ gives good results (for more details, see Pagan and Ullah, 1999).

4.2.2 Nonparametric regression

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 $y = \beta' x + \varepsilon$, we have $E(y|x) = \beta' x$. In general, one can of course write this conditional mean as:

$$E(y|x) = \int y f(y|x)\, dy \tag{9}$$

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:

$$E(y|x) = \frac{\int y f(x,y)\, dy}{\int f(x,y)\, dy} \tag{10}$$

The objective of a nonparametric regression then, is to replace the numerator and denominator in (10) by nonparametric estimators. Using the kernel specification of the previous section for the denominator, and extending this in an intuitive way towards the expression in the numerator, one obtains:

$$\hat{E}(y|x) = \frac{\frac{1}{Hb} \sum_{h=1}^{H} K\left(\frac{x_h - x}{b} \right) y_h}{\hat{f}_b(x)} \tag{11}$$

where $\hat{f}_b(x) = \frac{1}{Hb} \sum_{h=1}^{H} K\left(\frac{x_h - x}{b} \right)$, the estimated density of x. For a more formal derivation and necessary conditions we refer the reader to the text of Pagan and Ullah

(1999).[10] The estimator in (11) 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 $\sum_{h=1}^{H} w_{b_h}(x)\, y_h$ where $w_{b_h}(x) = w_b(x_h, x)$ is the weight assigned to the h^{th} observation. For the Nadaraya-Watson estimator (11), the one used in our application, $w_{b_h}(x)$ is given by

$$\frac{\frac{1}{Hb} K\left(\frac{x_h - x}{b}\right)}{\frac{1}{Hb}\sum_{h=1}^{H} K\left(\frac{x_h - x}{b}\right)}.$$

The problem of determining an appropriate bandwidth is the same in (11), the regression application, as it was for the density estimation. We have opted in our application for the "cross validation" method described in Härdle (1990) and Pagan and Ullah (1999).

As do many other methods, the method of cross validation is based on minimizing some loss function. The technique consists of leaving out the observation x_k when estimating the regression at point k:

$$\hat{E}(y|x_k) = \frac{\sum_{\substack{h=1 \\ h\neq k}}^{H} K\left(\frac{x_h - x_k}{b}\right) y_h}{\sum_{\substack{h=1 \\ h\neq k}}^{H} K\left(\frac{x_h - x_k}{b}\right)} \tag{12}$$

For a given bandwidth, b, one then calculates the following measure of squared deviations between y_k and its estimate:

$$CV(b) = \frac{1}{H}\sum_{k=1}^{H}\left(y_k - \hat{E}(y|x_k)\right)^2 w(x_k) \tag{13}$$

where $w(x_k)$ is a weight function applied to the k^{th} observation.[11] The value for b which minimizes the cross validation function (13) is selected as the (global) optimal bandwidth.

In the empirical application we started with the 'rule-of-thumb-bandwidth' (6). We then constructed an interval around this bandwidth and, in a number of steps, selected bandwidths from this interval. Each of the selected bandwidths is in turn used to estimate the conditional means (12) which subsequently act as input to compute values for

---

[10] One of the conditions needed for this result to hold is that the kernel function be symmetric. A condition that is fulfilled here, since we are using the Gaussian kernel based on the (symmetric) standard normal distribution.

[11] We used a standard normal function to assign weights to the observations.

expression (13). So we obtain as many values for expression (13) as there are bandwidths selected from the constructed interval. The bandwidth for which expression (13) is minimized, is then chosen as input in expression (8). That is, the optimal bandwidth, b, which minimizes (13) was used to determine the preliminary estimates of the density function(s) in expression (8).

4.2.3 Nonparametric estimation of Engel curves

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

$$y_i = g_i(x) + \varepsilon_i \tag{14}$$

Where the dependent variable 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 ε_i a random error term.

As an example and using the techniques explained in the previous section, we show in Figure 3 the estimation of expression (14) for the budget shares of the sixteen commodities for which we will impute expenditures in the fiscal file and which are listed in Table 11.[12] The logarithm of total household expenditures was used as explanatory variable x. In this exercise we used a grid of estimation points with a fixed interval between each point. The expression for determining the estimation points is as follows:

$$x_h = x_{h-1} + \left(\frac{x_M - x_m}{H - 1} \right), \quad h = 2, \ ..., \ H \ \textit{and} \ x_1 = x_m \tag{15}$$

where x_M is the maximum and x_m the minimum value of the explanatory variable x. The variable H is the total number of estimation points. The horizontal axes in Figure 3 have been relabeled, however, to show the percentiles of the logarithm of total expenditures. After having estimated the Engel curves at all the estimation points we only show the results for the estimation points that correspond to the 2nd through 99th percentiles. That is, estimation results for points corresponding with respectively the 1st and 100th percentiles are not shown on the graphs. The horizontal axes thus give an indication of (the place in) the distribution of total expenditures without being influenced by possible outliers in the bottom and the top of the distribution.

---

[12] In the figure we thus show $E(y|x) = g(x)$, that is the conditional mean of y given x.

TABLE 11 CATEGORIES OF GOODS ANALYZED

CATEGORY
Food
Beverages (non-alcoholic)
Alcohol
Tobacco
Clothing
Rent
Private Transportation
Public Transportation
Health
Energy
Maintenance
Leisure
Fuel (heating)
Car Fuel (diesel)
Car Fuel (leaded, unleaded)
Other Goods

FIGURE 3 ENGEL CURVES FOR THE DIFFERENT GOODS ANALYZED

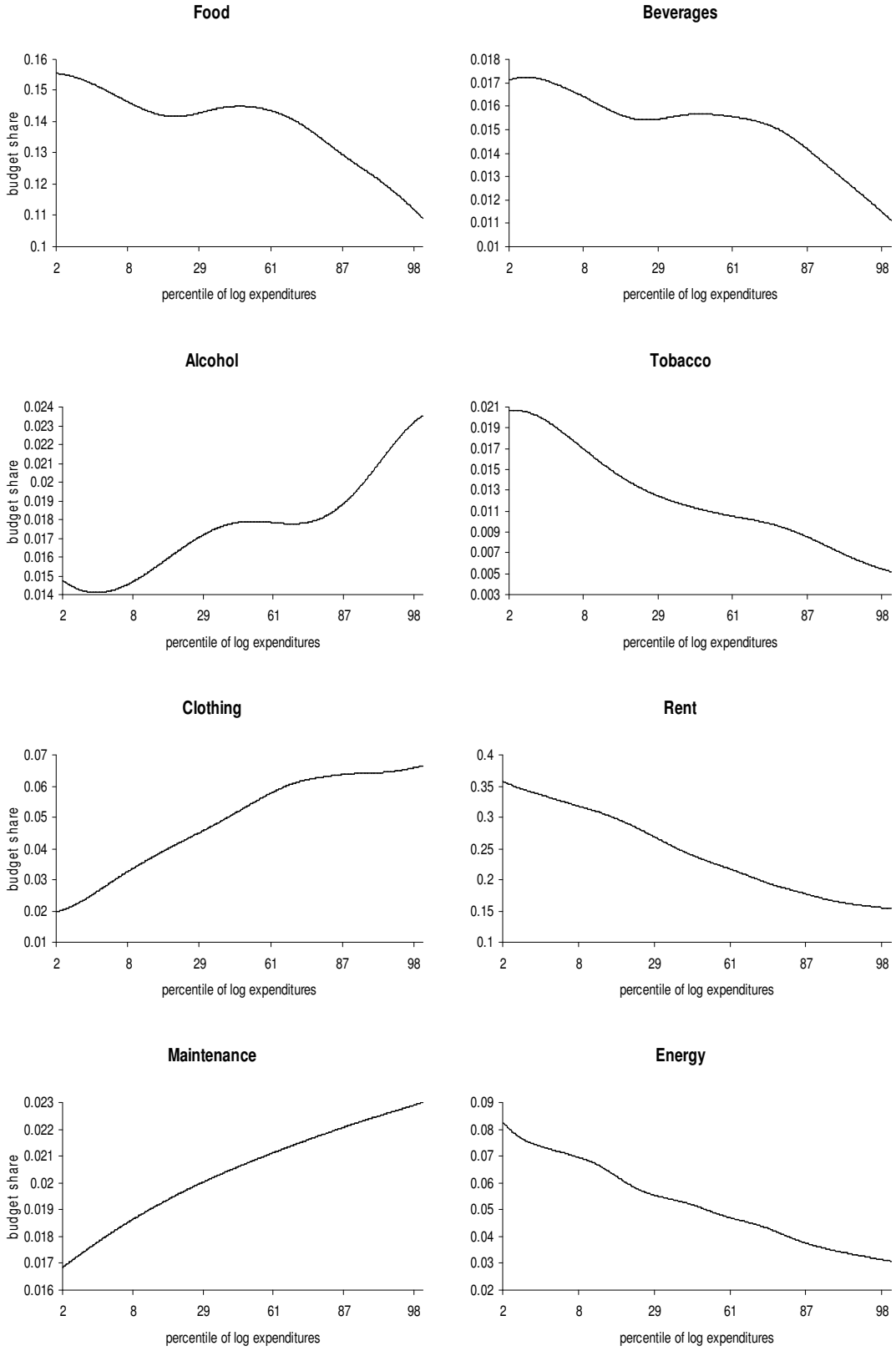
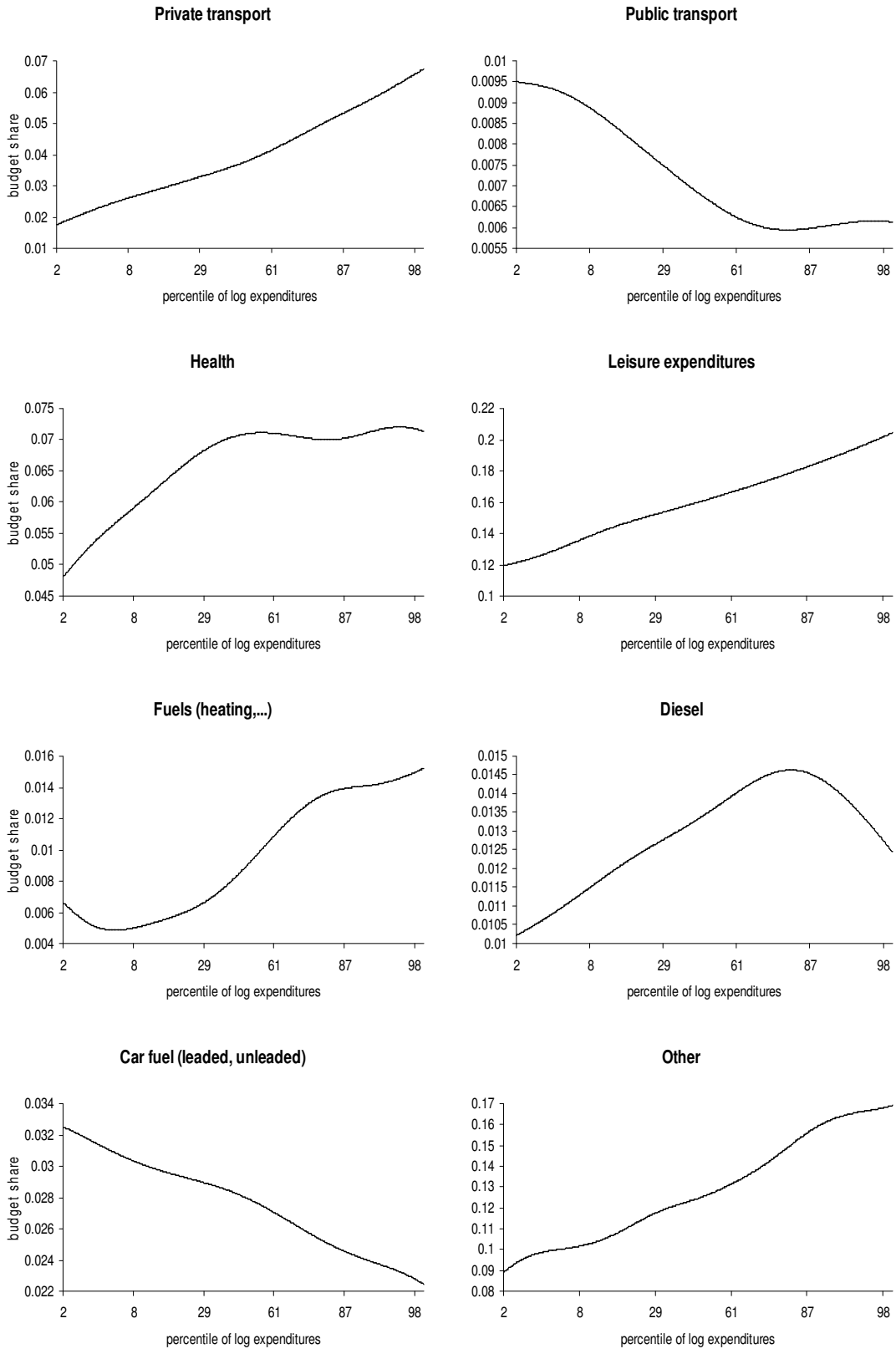
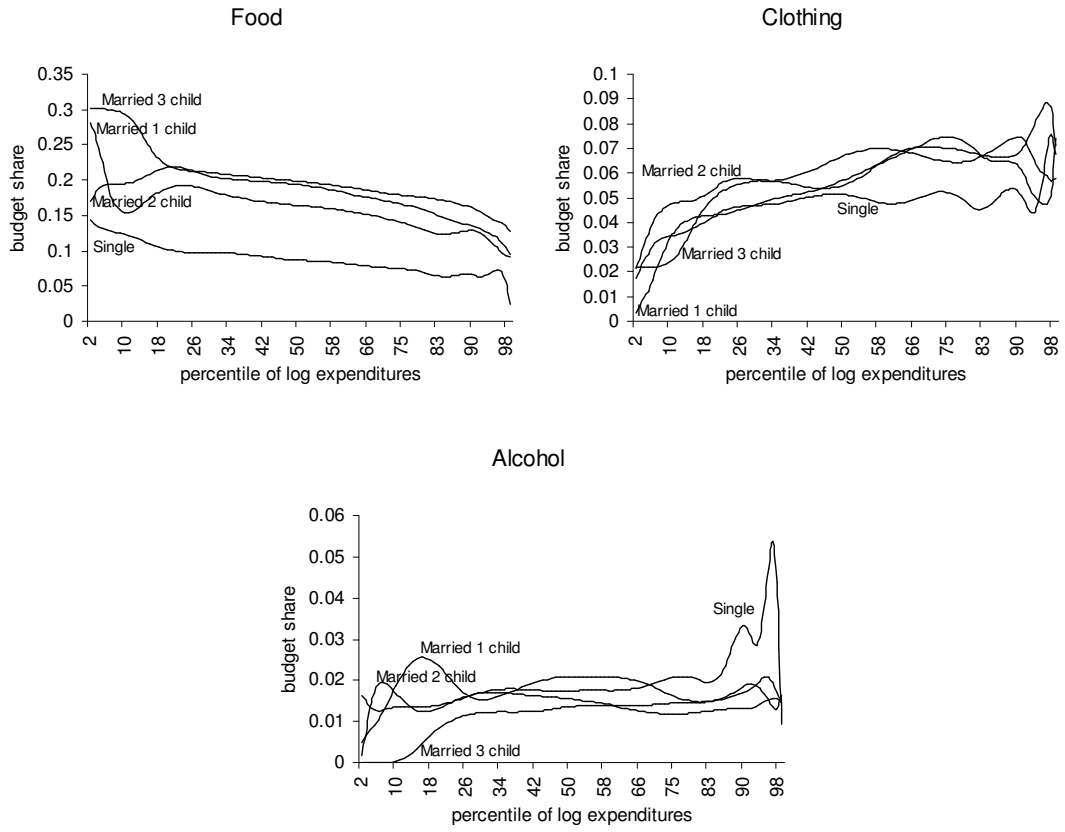


FIGURE 3 CONTINUED



It is well known that expenditure patterns depend on other characteristics than total expenditures (or disposable income) alone. This is illustrated in Figure 4, where we show Engel curves for three different commodities, food, clothing and alcohol, for households with different demographic characteristics. The Engel curves were estimated for the following subpopulations of households: single persons, married couples with one child, married couples with two children and finally married couples with three or more children.

FIGURE 4 ENGEL CURVES FOR DIFFERENT DEMOGRAPHIC SUBGROUPS



Given the differences in the location and, to a lesser extent, shape of the Engel curves for different demographic groups, the message of Figure 4 is to include household demographic characteristics in the estimation of the Engel curves and the subsequent imputation of expenditures.

To include demographic characteristics in the specification, equation (14) can be adapted to read as follows:

$$y_i = g_i(x, z) + \varepsilon_i \tag{16}$$

where z now represents a vector of demographic characteristics.

If we were to estimate equation (16) as it is stated here, this would imply a fully nonparametric estimation of the relation between consumption, income and household characteristics. If the dimension of the vector z is relatively large, however, nonparametric estimation will lead to some practical difficulties. 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 the univariate Gaussian one. The problem, however, is of a practical nature and has to do with data requirements. 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.[13]

A proposed solution to this problem is the use of semiparametric models, where part of the regression is entirely nonparametric and another part enters the equation in a parametric way. When using demographic variables in an Engel curve analysis the partially linear model seems the preferred way to follow.

In a partially linear setting we write (16) as:

$$y_i = \beta_i' z + F_i(x) + \varepsilon_i \tag{17}$$

where z is the vector of household characteristics, β a vector of parameters to be estimated, $F(.)$ an unknown function to be determined as well and ε_i a random error term with conditional mean equal to zero and variance σ_ε^2. The nonparametric part, the function $F(.)$, is now of lower dimension compared to expression (16). The estimation of

---

[13] A small numerical example will illustrate (Yatchew, 1998). Suppose we have a function f defined on the unit interval, that is, we are considering first the one-dimensional case. 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 a rate $O(1/T)$, where $O(\ )$ 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 point 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 a rate $O\left(1/\sqrt{T}\right)$ as the number of data points increases. More generally, for n-dimensional problems, the approximation error reduces at a rate of $O\left(1/\sqrt[n]{T}\right)$. This means, 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.

this model and the parameters involved follows the method proposed by Robinson (1988). Take the expected value of expression (17) conditional on x to become:

$$E\left(y_i|x\right)=\beta_i'E\left(z|x\right)+F_i\left(x\right) \tag{18}$$

Now subtract expression (18) from expression (17) to obtain:

$$y_i-E\left(y_i|x\right)=\beta_i'\left[z-E\left(z|x\right)\right]+\varepsilon_i \tag{19}$$

Equation (19) 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 $\frac{\sum_{h=1}^{H}K\left(\frac{x-x_h}{b}\right)y_{ih}}{\sum_{h=1}^{H}K\left(\frac{x-x_h}{b}\right)}$, and $E\left(z|x\right)$ by $\frac{\sum_{h=1}^{H}K\left(\frac{x-x_h}{b}\right)z_h}{\sum_{h=1}^{H}K\left(\frac{x-x_h}{b}\right)}$.

The function $F\left(.\right)$ can then be estimated from (18) as:

$$\hat{F}_i\left(x\right)=\hat{E}\left(y_i|x\right)-\hat{\beta}_i'\hat{E}\left(z|x\right) \tag{20}$$

Alternatively, one can consider $F_i\left(x\right)$ in (17) as the conditional mean of $y_i-\beta_i'z$ given x. That is:

$$F_i\left(x\right)=E\left(y_i-\beta_i'z|x\right) \tag{21}$$

which can be estimated by nonparametrically regressing $y_i-\beta_i'z$ on x. To compute this, one can replace parameters β by their estimates obtained from regression (19). It is this approach that we have followed in this text.[14]

[14] Note that both expression (20) and the procedure described in expression (21) yield consistent estimators for the function $F\left(.\right)$ and should give similar results. It was computationally more interesting to use expression (21). When using expression (20) we would have to impute the conditional mean of each of the demographic variables separately, whereas when using (21) we only have to do one imputation.

5 NONPARAMETRIC IMPUTATION OF CONSUMPTION IN THE FISCAL FILE

5.1 ESTIMATION IN THE EXPENDITURE SURVEY

In this subsection all the expressions used and estimations done are with respect to the expenditure survey. That is, in this section we only use data from the expenditure survey whenever estimations or computations are done.

In the application we have selected two explanatory variables for the nonparametric part of the regression, i.e. in the function $F(.)$. The first one is disposable income of the household. We used income, and not the usual total expenditure variable because it is disposable income which is available in the fiscal file. The second variable in the nonparametric part is "age". In this we follow Schmalensee and Stoker (1999). We feel that age of the household head may enter nonlinearly in the equation. Moreover, it is not implausible for age to interact with income. The regression equation thus reads as follows:

$$y_i = \beta_i' z + F_i(x, age) + \varepsilon_i \tag{22}$$

where y_i are the expenditures on good i (not the budget shares), z is a vector of household characteristics, $F_i(.,.)$ an unknown function, x is the income of the household, age is the age of the reference person or head of the household, and ε_i a random error term. The same estimation procedure applies as in section 4.2.3 above. The conditional means of the respective goods *for each household* are determined as follows:

$$\hat{E}(y_i | x, age) = \frac{\sum_{h=1}^{H} K\left(\frac{x - x_h}{b_x}\right) K\left(\frac{age - age_h}{b_{age}}\right) y_{ih}}{\sum_{h=1}^{H} K\left(\frac{x - x_h}{b_x}\right) K\left(\frac{age - age_h}{b_{age}}\right)} \tag{23}$$

where b_x and b_{age} are the respective bandwidths for the variables x and age of the household for which (23) is estimated and H is the total number of households in the expenditure survey. We write both the denominator and numerator of (23) as a product of two univariate standard normal distributions instead of using a bivariate or joint standard normal distribution for x and age. This is because the correlation between x and age in the expenditure survey was not statistically significantly different from zero, so that we can approximate the bivariate standard normal function as the product of two univariate standard normal functions (see e.g. Greene, 2000). A similar expression applies for all the demographic variables in the vector z.

The bandwidth selection of both income and age is based on expressions (12) and (13), the result of which is used as initial input in the determination of the adaptive kernel

bandwidths determined by expression (8). To find the bandwidth that minimizes expression (13) we considered twenty different values for the bandwidth in a range around the 'rule-of-thumb' bandwidth obtained in formula (6). Bandwidths were calculated this way for both income and age separately.

To describe the determination of the adaptive bandwidths we repeat expression (8) here:

$$b_x = b * \delta_x \ , \ \ \delta_x = \left[\frac{\tilde{f}(x)}{G} \right]^{-\lambda} \tag{24}$$

Remember that G is the geometric mean of some preliminary estimate of the density $\tilde{f}(x)$ at all points of estimation x. That is:

$$G = \left(\prod_{h=1}^{H} \tilde{f}(x_h) \right)^{\frac{1}{H}} \tag{25}$$

Taking logarithms of both sides of (25) we get:

$$\ln G = \frac{1}{H} \sum_{h=1}^{H} \ln \tilde{f}(x_H) \tag{26}$$

so that G is determined as:

$$G = \exp\left(\frac{1}{H} \sum_{h=1}^{H} \ln \tilde{f}(x_h) \right) \tag{27}$$

First we estimate the density $\tilde{f}(x)$ at a grid of estimation points.[15] We then apply (27) and (24) to determine δ_x. For λ we have taken a value of 0.5, which, as stated in section 4.2.1, is said to give good results. The result for δ_x is then multiplied by the global optimal bandwidth found by minimizing expression (13) and given by b in (24). We thus obtain for each estimation point x a different bandwidth. The bandwidths are then multiplied by a factor 1.3.[16]

The bandwidths that are obtained are then used to estimate the adaptive kernel regression (22) for each category of goods listed in Table 11. The procedures followed are

---

[15] The procedure which is described is performed twice, once for income and once for age. The grid of points at which we estimate the density is chosen to coincide with the actual observations in the expenditure survey.

[16] The reason we multiply by 1.3 is to have δ_x equal or close to 1 in the region of the mode of the distribution. The number 1.3 holds for a normal density distribution and will vary with the true underlying density distribution. Nevertheless, 1.3 can be used as a rough guide (see Pagan and Ullah, 1999 for more details).

the ones described in section 4.2.3. That is, for each household we estimate the conditional mean(s): [17]

$$\hat{E}\left(y_i \middle| x, age\right) \quad i = 1, \ldots, n \tag{28}$$

where y_i are the expenditures on good i by the household, n is the number of goods listed in Table 11, x is net disposable income of the household and age is the age of the household head. For each household we also estimate:

$$\hat{E}\left(z_k \middle| x, age\right) \quad k = 1, \ldots, K \tag{29}$$

where z_k is the k^{th} element in the $k-$dimensional vector z of demographic characteristics of the household. The demographic variables are listed in Table 12. Evidently, since the ultimate objective is to use the regression for imputation in the fiscal file, the choice was confined by the availability of the corresponding variables in the fiscal file.

TABLE 12 DEMOGRAPHIC CHARACTERISTICS

Married (yes/no)
Number of Children at Charge
Number of Other Persons at Charge
Number of Children less than Three Years of Age
Region (Flanders, Brussels, Wallonia)
Age of the Household Head
Sex of the Household Head

Once we have estimated the conditional means in (28) and (29) using expression (23) where the bandwidths are replaced by adaptive bandwidths for both income and age of the reference person, we construct the following differences for every household and for every good i and demographic characteristic k:

$$y_i - \hat{E}\left(y_i \middle| x, age\right) \quad i = 1, \ldots, n \tag{30}$$

and

$$z_k - \hat{E}\left(z_k \middle| x, age\right) \quad k = 1, \ldots, K \tag{31}$$

The differences thus calculated are then used to run an OLS regression as in (19) which provides estimates for the coefficients β in the partially linear model (22). That is, the following regression was estimated:

[17] As in Schmalensee and Stoker (1999) we standardized all variables that acted as arguments to the kernel function, i.e. income and age of the reference person.

$$y_i - \hat{E}\left(y_i \middle| x, age\right) = \beta_{i1}\tilde{z}_1 + \beta_{i2}\tilde{z}_2 + \beta_{i3}\tilde{z}_3 + \beta_{i4}\tilde{z}_4 + \beta_{i5}\tilde{z}_5 + \beta_{i6}\tilde{z}_6 + \beta_{i7}\tilde{z}_7 + \varepsilon_i \tag{32}$$

for each good i. The variables $\tilde{z}_k = z_k - \hat{E}\left(z_k \middle| x, age\right)$ $k = 1$, ... , 7 are the constructed differences for each of the household characteristics listed in Table 12 and ε_i an error term. The error terms are saved and will be used in the imputation step described below.

In the estimation of the function $F_i\left(x, age\right)$ we used expression (21). The following differences are constructed for each good i and for each household h:

$$y_{ih} - \hat{\beta}_i' z_h \tag{33}$$

with z_h the vector of k household characteristics for household h and $\hat{\beta}_i'$ the vector of estimated coefficients for good i obtained from regression (32). Expression (33) is then used as the dependent variable in a nonparametric adaptive kernel regression with income and age as explanatory variables which gives an estimate of the function $F_i(.)$ in the expenditure survey. That is, in expression (23) we replace y_{ih} by $y_{ih} - \hat{\beta}_i' z_h$.[18] Bandwidths were calculated, both optimal (global) and adaptive, for the constructed variable in (33).

5.2 IMPUTATION IN THE FISCAL FILE

All the steps performed up until now use data from the expenditure survey only. The variable constructed in (33) and the bandwidths calculated are now used to impute the function $F_i\left(x, age\right)$ in the fiscal file for each good i. The imputation steps will be described in what follows.

Intuitively, one might think of the imputation procedure as a missing values problem. That is, consider the expenditure survey and the fiscal file as one file with J missing values for the expenditures on each of the goods.[19] Each missing value will now be replaced by an imputed value that is calculated from the H values that we do observe.

More formally, the expression for the nonparametric part of the imputation of expenditures on commodity i for unit j in the fiscal file with income x_j and age age_j is as follows:

---

[18] We thus estimate for each household the following conditional mean: $\hat{E}\left(y_i - \beta_i' z \middle| x, age\right)$.

[19] Note that we do not actually merge the two files into one. We take an observation in the fiscal file and compare it with each observation in the expenditure survey using expression (34).

$$\tilde{F}_i\left(x_j, age_j\right) = \left[\frac{\sum_{h=1}^{H} K\left(\frac{x_h - x_j}{b_{x_h}}\right) K\left(\frac{age_h - age_j}{b_{age_h}}\right) \Delta y_{ih}}{\sum_{h=1}^{H} K\left(\frac{x_h - x_j}{b_{x_h}}\right) K\left(\frac{age_h - age_j}{b_{age_h}}\right)} \right] \tag{34}$$

where H is the number of observations in the expenditure survey. The variables b_{x_h} and b_{age_h} are the bandwidths corresponding to each x_h and age_h respectively. The variables Δy_{ih} are given by $\Delta y_{ih} = y_{ih} - \hat{\beta}_i' z_h$ for each good i. The result of (34) is J imputed values for the function $F(.,.)$ in the fiscal file. Expenditures can now be imputed using the estimated β coefficients previously obtained from the estimation of (32). For each good i in Table 11 and for every household j in the fiscal file we then have:

$$\tilde{y}_{ij} = \tilde{F}_i\left(x_j, age_j\right) + \hat{\beta}_i' z_j + \tilde{\varepsilon}_i \tag{35}$$

where a tilde indicates an imputed value and a hat an estimated one. The vector z_j is again the vector of household characteristics listed in Table 12, now for household j in the fiscal file 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. In order to perform analyses we want the imputed values not to be too different from the ones we used in calculating them. That is, we want the distribution of imputed values to be as close as possible to that of the observed values (Paulin and Ferraro, 1996; Little and Rubin, 1987).

We add error terms by taking random draws from the empirical distribution of residuals obtained from the estimation of regression (32) for each imputed value. That is, for each value that we impute in the fiscal file we randomly draw a single value from the distribution of residuals and add this to the imputed conditional mean. Imputed expenditures with negative values are replaced by zero.

6 RESULTS[20]

In both the expenditure survey and the fiscal 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 fiscal file. There were instances in the fiscal file 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.[21] This resulted in 3,207 observations for the expenditure survey[22] and 23,820 observations for the fiscal file.

Table 13 shows some summary statistics for the different categories of commodities in the expenditure survey and the fiscal file respectively. The results for the fiscal file are calculated using the imputed values. The last column of the table shows the percentage difference in mean observed (expenditure survey) and imputed values (fiscal 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 3). We refer to Appendix 2 for an extensive discussion of this difference and its central place in the attempt to improve the previous matching exercise. [23] Unfortunately we were unable to obtain a sample of sociological households with fiscal information (see also section 1 and footnote 2). Hence, we had no choice but to treat the fiscal households in the fiscal 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 13 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

---

20 Here all the amounts shown are expressed in euros.

21 Since in the fiscal file 99 is a code rather than the actual age, reference persons with an "age" of 99 were also deleted.

22 We used a file constructed from the household budget survey containing only sociological households for which at least one member had a positive probability of being sampled in the fiscal file. The way this was done is explained in section A.2 of Appendix 2. See also footnote 9.

23 In Appendix 2 we describe how sociological households can be split into fiscal units. From the 3,816 sociological households 4,260 *potential* fiscal units were distilled.

expenditures. For tobacco, public transport 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).[24] 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. Due to time constraints, we did not choose this track, but simply applied the one shot semiparametric model on all sixteen commodities. We are fully aware that this will bias the imputation downwards for the commodities with a high number of zeroes (both through the coefficients of the parametric part and through the nonparametric regression).

To have a detailed picture of how well the distribution of the imputed values mirrors the distribution of the observed ones, Figure 5 shows the density functions for the different commodities. We have restricted the estimation for the graphs of Figure 5 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 (fiscal file). On the horizontal axis the expenditures in euros are shown. The density functions were estimated using the kernel density technique described in section 4.2.1, but without adapting the bandwidth at each estimation point. Formula (6) was used as an approximation of the optimal bandwidth.

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

TABLE 13 OBSERVED VERSUS IMPUTED VALUES: BUDGET SURVEY VERSUS FISCAL FILE

Good	Household Budget Survey					Fiscal File					Percentage difference [(2)-(1)]/(1)
	mean (1)	sd	p25	p50	p75	mean (2)	sd	p25	p50	p75	
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 5 DENSITY FUNCTIONS FOR THE DIFFERENT GOODS ANALYZED: OBSERVED VERSUS IMPUTED

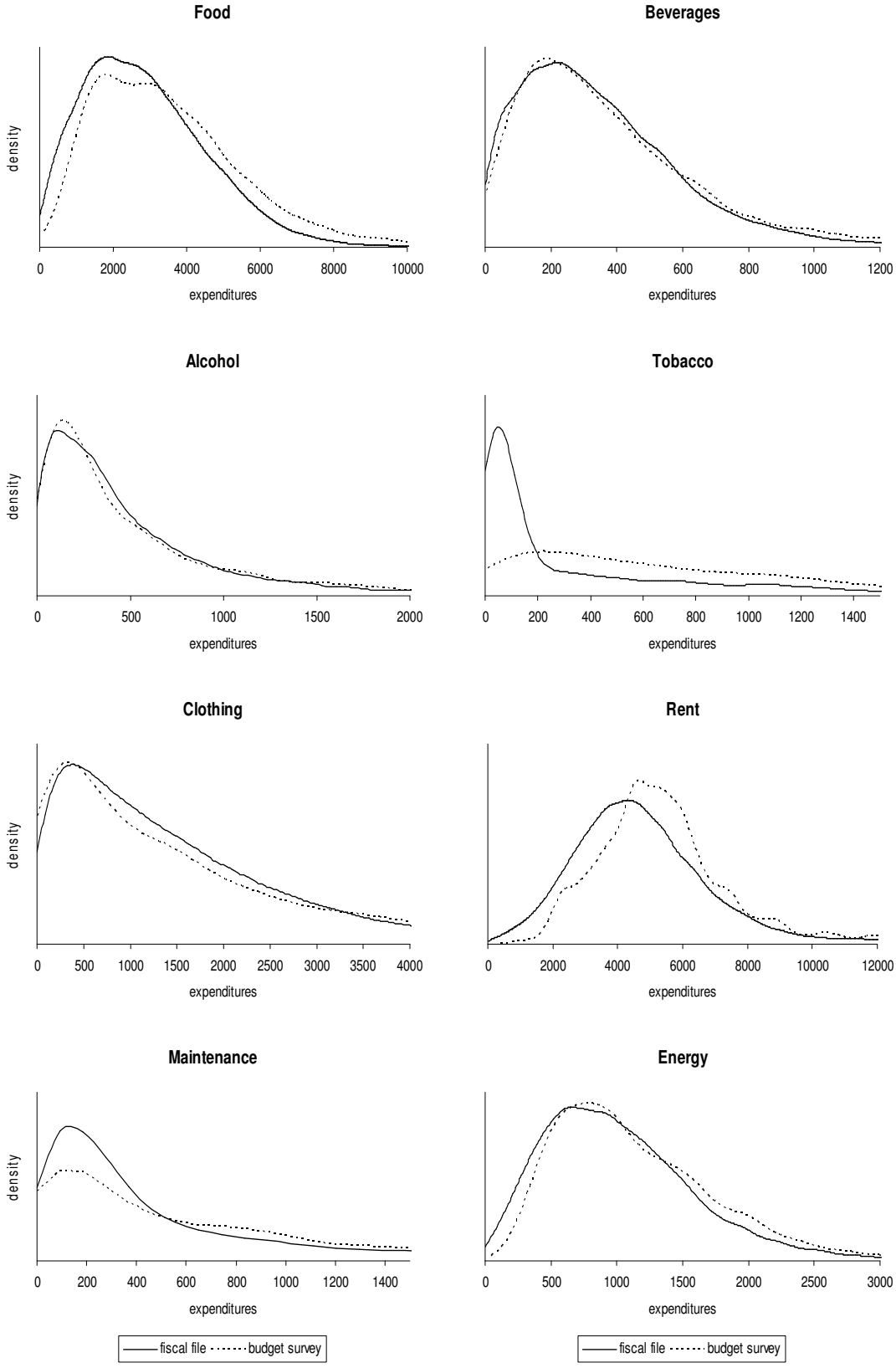
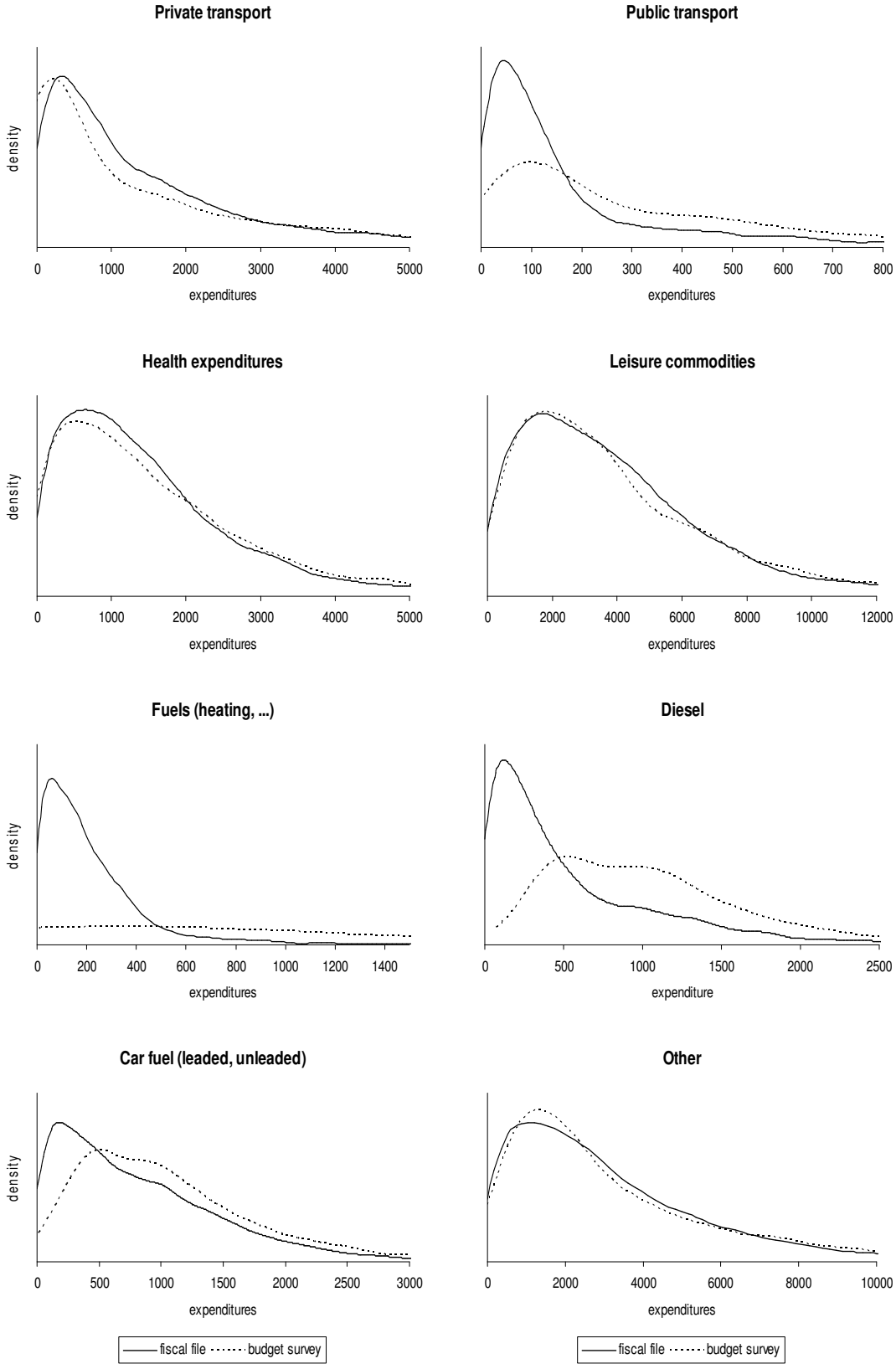


FIGURE 5 CONTINUED



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 13 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 14 shows the number and proportion of zero expenditures in the budget survey and after the imputation in the fiscal file. On the one hand the commodities for which we find a diverging distribution in Figure 5 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 fiscal file for the non-problematic commodities follows from the negative imputed values which were put to zero[25] (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 (35), 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 confirm our intuition that this kind of imputation allows to enrich the fiscal file with a distribution of detailed expenditures on which policy simulations can be run. A full description of the data file can be found in Appendix 1.

TABLE 14 ZERO OBSERVATIONS IN THE BUDGET SURVEY AND AFTER IMPUTATION IN THE FISCAL FILE

good	Nonzero (budget survey)	Zero (budget survey)	Percentage (budget survey	Percentage (fiscal 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%
Total	3207			

[25] In the graphs in Figure 5 we show the density functions for the strictly positive amounts only.

7 CONCLUSION

In this report we explored an alternative approach to combine data on gross incomes and data on consumption. Such a combined file is useful when one wants to assess, for example, changes in personal income taxes or social security contributions and/or indirect taxes simultaneously. A previous DWTC-project (project AG/01/30), of which the current one was intended to be a follow-up project, made use of statistical matching techniques to combine expenditure survey data with fiscal data.[26] We briefly touched upon the vast literature on statistical matching techniques in the first part of the report However, due to data delivery problems in the current project, we looked at an alternative procedure and actually worked the other way around, that is, we imputed data on consumption in a file containing fiscal data, rather than complementing the expenditure survey with data on gross incomes. For this we used semiparametric estimation and imputation of Engel curves.

In the first part of the report a descriptive analysis was given of the expenditure survey for the year 2000. In general the survey can be considered representative of the Belgian private household population. The same, however, can not be said of the fiscal file we had at our disposal. As mentioned in the text, the original project proposal envisaged a fiscal file based on a (representative) sample of the Belgian population. This file did not reach us in time however. The file we did work with consisted of a sample of fiscal units, that is, people having filled out and returned a tax form. We stressed that the observational units in this file, i.e. fiscal units, are not comparable to the observational units found in the expenditure survey, i.e. sociological households. The latter can consist of several fiscal units. We noted beforehand that this could possibly bias the results of the empirical exercise.

The second part of the report described in more detail what is meant by nonparametric estimation and laid out some theoretical concepts underlying the techniques. However, due to the so called 'curse of dimensionality', it is more practical to restrict oneself to semiparametric estimation techniques. More specifically, in this report we explored the partially linear model and applied it to the estimation of Engel curves. The results of the estimation were then used to impute consumption expenditures in the fiscal file.

Given the differences in observational units between the two files, and hence their incomparability, the results of the empirical exercise are promising and plead for the use of the techniques described in the report. For many commodities the density functions of both observed and imputed values were close enough together to allow distributional analysis of, e.g., indirect taxes in the fiscal file. When the difference between the density functions was large, we suggested this to be the result of a preponderance of zero observations for (most of) these goods in the household budget survey. An obvious avenue for further research then would be to first estimate a discrete choice model, parametrically or nonparametrically, and to take account of the resulting probabilities in

[26] See Decoster and Van Camp (2002).

the imputation step. Also the replacement of negative imputed values by zeroes needs further refinement.

Needless to say that a comparison between the results of a statistical matching technique (as used in project AG/01/030) and the technique used here, forces itself as a natural extension to the current report.

8 REFERENCES

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

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

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

[4] Czajka, J. L., Hirabayashi, S. M., Little, R. J.A. and Rubin, D. B., 1992, "Projecting from Advance Data using Propensity Modelling: An Application to income and Tax Statistics", *Journal of business and economic statistics*, Vol. 10, No.2, 117-131

[5] Decoster, A. and Van Camp, G., 2001, "Koppeling van de Budgetenquête 1997-98 en het Fiscaal Bestand 1999 (inkomsten 1998). Nota voor de gebruikers van het gekoppelde bestand.", mimeo, Centre for Economic Studies, K.U.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] De Graeve, D., Cantillon, B., Schokkaert, E., Kerstens, B., Van Camp, G. and Van Ourti, T. (2003), Billijkheid in de financiering van gezondheidszorg: Eindrapport, DWTC-Project SO/01/005, Leuven: Katholieke Universiteit Leuven, Centrum voor Economische Studiën (URL:http://www.belspo.be/belspo/fedra/proj.asp?l=nl&COD=SO/01/005)

[9] Gong, X., Van Soest, A. and Zhang, P., 2000, "Sexual Bias and Household Consumption: A Semiparametric Analysis of Engel Curves in Rural China", IZA Discussion paper No. 212

[10] Greene, W.H., 2000, *Econometric Analysis*, Prentice Hall

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

[12] Horton, N. J. and Lipsitz, S. R., 2001, "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables", *The American Statistician*, vol.55, No.3, 244-254

[13] Kadane, J. B., 1978, "Some Statistical Problems in Merging Data Files", *1978 Compendium of Tax Research*, U.S. Department of the Treasury, 159-171

[14] Kadane, J. B., 2001, "Some Statistical Problems in Merging Data Files", *Journal of Official Statistics*, Vol. 17, No.3, 423-433

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

[16] Little, R.J.A and Rubin, D.B, 1987, *Statistical Analysis with Missing Data*, John Wiley & Sons

[17] Moriarity, C. and Scheuren, F., 2001, "Statistical Matching: A paradigm for Assessing the Uncertainty in the Procedure", *Journal of Official Statistics*, Vol. 17, No.3, 407-422

[18] Moriarity, C., 2001, *Statistical Properties of Statistical Matching*, PhD dissertation, Columbian School of Arts and Sciences, George Washington University

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

[20] Nordbotton, S., 1996, "Neural Network Imputation Applied to the Norwegian 1990 Population Census Data", *Journal of Official Statistics*, Vol.12, No.4, 385-401

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

[22] Paulin, G. D. and Ferraro, D. L., 1996, "Do Expenditures Explain Income? A Study of Variables for Income Imputation", *Journal of Economic and Social Measurement*, 22, 103-128

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

[24] Rodgers, W. L., 1984, "An Evaluation of Statistical Matching", *Journal of Business and Economic Statistics*, Vol.2, No.1, 91-102

[25] Rosenblatt, M., 1956, "Remarks on Some Nonparametric Estimates of a Density Function", *Annals of Mathematical Statistics*, 27, 642-669

[26] Rubin, D.B., 1986, "Statistical Matching using File Concatenation with Adjusted Weights and Multiple Imputations", *Journal of Business and Economic Statistics*, 4, 87-94

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

[28] *Standaard Belastingalmanak 2001*, Standaard Uitgeverij

[29] Stern, S., 2001, "Valuing Housing Subsidies in a New Measure of Poverty: A Statistical Match of the American Housing Survey to the Current Population Survey", Housing and Household Economics Statistics Division, Working Paper, U.S. Census Bureau

[30] Sutherland, H., Taylor R. and Gomulka, J., 2001, "Combining Household Income and Expenditure Data in Policy Simulations", Microsimulation Unit Discussion Paper MU0101

[31] Taylor, R., 2000, "Guidelines for Identifying Clusters using Grade Correspondence Analysis: Practical and Technical Issues", Microsimulation Unit Research Note MU/RN/39

[32] Taylor, R., Gomulka, J. and Sutherland H., 2000, "Creating Order out of Chaos? Identifying homogeneous groups of households across multiple datasets", Paper Prepared for the 26th General Conference of the Interantional Association for Research in Income and Wealth, Cracow, Poland, 27 August to 2 September 2000

[33] Van Camp, G., 2002, "Statistische Koppelingstechnieken: Twee Mogelijke Benaderingen om Informatie uit Meerdere Bestanden te Combineren", DWTC-Project AG/01/030 Eindrapport Deel 1

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

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

[36] Wilke, R., 2003, "Semiparametric Estimation of Regression Functions under Shape Invariance Restrictions", ZEW Discussion paper, No. 03-64, Mannheim

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

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

Appendix 1 STRUCTURE OF THE DATA FILE WITH IMPUTED EXPENDITURES

In this appendix we describe the structure of the data file with the imputed expenditures. The fiscal file we had at our disposal was an Excel file, Euro2000.xls. This file contains 4 worksheets, 'base', 'Euro2000a', 'Euro2000b', 'Feuil3' and 'Feuil4'. To keep the same structure and not to lose any information, i.e. information that was not at all used in the imputation, we have made a copy of this file, Euro2000adj.xls, where we added an extra worksheet 'Imputed Consumption'.

In the file Euro2000adj.xls all observations that were not used in the imputation and hence that have no imputed consumption values, were deleted. This left us with a file of 23,820 observations. The extra worksheet that was added contains an identification number, which is the same has in the other sheets, and the imputed expenditures, for all the goods listed in Table 11. The variables and their description are listed in Table 15.

TABLE 15 VARIABLE NAMES AND DESCRIPTION OF IMPUTED EXPENDITURES IN THE SHEET 'IMPUTED CONSUMPTION' OF EURO2000ADJ.XLS

Variable Name	Description
food	Expenditures on food
beverage	Expenditures on non-alcoholic beverages
alcohol	Expenditures on alcohol
tobacco	Expenditures on tobacco
clothing	Expenditures on clothing and shoes
rent	Expenditures on rent (also imputed rents for homeowners)
energy	Energy related expenditures
maint	Maintenance costs
private	Expenditures on private transport
public	Expenditures on public transport
health	Health and hygiene related expenditures
leisure	Leisure related expenditures
fuel	Expenditures on fuel for heating etc.
diesel	Expenditures on car fuel (diesel)
gasoline	Expenditures on car fuel (leaded, unleaded)
other	Expenditures on other goods
exp	Total expenditures

For a description of the other, mostly income related, variables in the fiscal file we refer the reader to the Ministry of Finance (FOD Financiën, Studie- en Documentatiedienst).

In the worksheet Euro2000a we recoded the variables 'Sexe' and 'Region'. 'Sexe' takes on the values 1 and 2, where 1 stands for male and 2 for female. The variable 'Region' can take on one of three different values, 1, 2 or 3. A 1 indicates that the fiscal unit resides in the Brussels Capital Region, 2 indicates the Flemish Region while a 3 indicates that the fiscal unit resides in the Walloon Region.

Appendix 2

In this appendix we describe the work that has been done in anticipation of a fiscal file that was part of the initial project proposal. One of the aims of the original project proposal was to draw a sample from the Belgian population and to have this sample supplemented with fiscal information by the Ministry of Finance. This way a *representative* sample would have been obtained of sociological households and not only *fiscal units*. The procedure used to draw this sample is also discussed in this appendix. Concerning the efforts we undertook in order to obtain this file in due time, we refer to the document "Voorlopig rapport AG0179.doc".[27]

A.1 Sampling procedure

Our procedure is based on 'systematic sampling'. Basically, in systematic sampling, a first number is drawn randomly, after which each record with a fixed distance relative to the previously drawn record is selected. For a population of size N and a sample of size n one determines x as $\frac{n}{N}=\frac{1}{x}$. The first record, then, will be selected by drawing a random number between 1 and x. From then on each record with a distance of x to the previously selected record will be taken up in the sample.[28]

A.1.1 Sample frame

First we have to determine the sample frame which is the set of all units that have a positive probability of being selected. In our case this would be the set of all household heads of a private household observed in the national registry. Since the aim was to match with a fiscal file we wanted the household situation to be that on January 1st 2001, since this is the date that people used to fill out their tax forms for the income they earned in the year 2000. For each household head we have its age, nationality and the number of household members. The total number of household heads determines the size of the sample frame, N.

A.1.2 Determining the size of a representative sample

The size of a representative sample from a sample frame of size N can be determined as follows:

$$n=\frac{(2.576)^2*N*0.5*0.5}{(N-1)*(0.005)^2+(2.576)^2*0.5*0.5} \tag{36}$$

---

[27] This document was sent by email on Monday September 15, 2003 to Mr. Aziz Naji and Mr. Christian Valenduc. See also footnote 2.

[28] Note that complete random sampling is preferred, but can be time consuming when working with large data files.

where N is, as before, the size of the population or sample frame, 0.5, in this case, is the proportion of some population characteristic (which gives a variance when squared), 2.576 is a statistical significance level and 0.005 is a precision level to be chosen by the researcher. [29] In our case it is the amount that the sample proportion of a certain variable is allowed to deviate from the (true) population proportion of that same variable. The next step then is to draw a sample of size n from the population using systematic sampling as described in the introduction to this section.

In the representative sample certain groups, that is households with certain characteristics, will be underrepresented while others will be overrepresented.[30] This implies that for the underrepresented groups we will not have many candidates to choose from for matching with similar groups in the Household Budget Survey. This will influence the performance of the match, at least for those groups. To overcome this problem we perform additional draws for groups that are underrepresented in the representative sample. Note that this is done deliberately in order to enhance the performance of the matching procedure.

A.1.3 Drawing additional units

Based on the variables in the sample frame the population can be divided in mutually exclusive groups. For each group so defined it is possible to determine the 'ideal' size based on a similar formula as described in the previous section. The formula reads:

$$n_g^* = \frac{(2.576)^2 * N_g * 0.5 * 0.5}{(N_g - 1) * (0.01)^2 + (2.576)^2 * 0.5 * 0.5} \tag{37}$$

where g is an index indicating the group that we are considering, the other values are as in (36).[31] We then calculate for each group g the difference between the size determined in (37) and the number of units we already have selected via the representative sample,

---

[29] What is shown here is a specific example of a more general form which reads: $n \geq \frac{z_\alpha^2 N \operatorname{var}(Y)}{N d\{\bar{Y}^2\} + z_\alpha^2 \operatorname{var}(Y)}$, where Y is the variable of interest, $\bar{Y}$ its mean, z_α is the value of a standard normal distribution corresponding to a significance level of α and $d\{.\}$ is a precision parameter indicating the amount the sample mean is allowed to deviate from the (true) population mean.

[30] Note that it is not over- or underrepresentation at the population level, since we are working with a representative sample. The representativeness should be thought of in terms of matching candidates. Small groups might not be well represented in the matched file if there are only a few of them to choose from.

[31] Note that we have used 0.01 as precision level in this case. This is solely a consequence of practical considerations concerning the sample sizes that were obtained and that would be practically feasible. Note also that, although the first sample might be representative of the population, in order to have sample sizes that permit reliable statistical inference, additional units have to be drawn for certain groups.

i.e. n_g. If this difference is positive we will resample from the units not yet selected of that group. So, for each group that has a positive difference we have a sample frame of $N_g - n_g$ units. Since the numbers in these subsamples are typically small relative to the entire population we used pure random sampling in this case and thus no longer resorted to systematic sampling.

The file we had at our disposal to draw the sample from represented 10,162,945 individuals and 4,306,311 households. For each household a household number, the number of members in the household, the age of the household head and its nationality were available. We will limit the results shown to the sample that was ultimately selected. We experimented using several grouping procedures and values for the parameters in (36) and (37). We used a different value for the precision parameter in (37) than we did in (36) (0.01 and 0.005 respectively). The reason is that the sample sizes that resulted when using the same value in both cases were very large.

The population was divided according to four age categories, younger than 25, between 25 and 54, between 55 and 64 and 65 and older. [32] We also distinguished seven categories based on the number of household members: 1, 2, 3, 4, 5, 6 and 7 or more person households. In terms of resampling this resulted in 28 groups for which to determine whether or not additional units had to be drawn. Some results are shown in Table 16. As one can see the first sampling procedure, based on systematic sampling, resulted in a sample of size 66,251. Then an additional 202,659 households were drawn based on random sampling resulting in a total sample size of 268,910 households. Based on the results for the representative sample, that is, the first, systematic, drawing, we can feel confident about the procedure followed. Looking at the percentage distribution of groups in both the population and the representative sample, we see that they are very similar. As Table 16 makes clear the largest groups can be found in the age category between 25 and 55 up to four household members. Singles in this age category form the largest group amongst Belgian households. Also elderly persons, both singles and couples, form a considerable group. Together they represent almost a quarter of the Belgian household population.

[32] The age categories were chosen so as to represent four major 'cycles' in a person's life, namely youth, working age, pre-retirement age and retirement age.

TABLE 16 RESULTS FROM SAMPLING PROCEDURE

Category		Population (number)	Percent in population	1st sample (number)	Percent in 1st sample	Final sample (number)	Ideal Size
Age	Members (number)						
<25	1	66,268	1.54	1,010	1.52	13,268	13,268
<25	2	22,116	0.51	344	0.52	9,479	9,479
<25	3	6,505	0.15	95	0.14	4,672	4,672
<25	4	1,447	0.03	28	0.04	1,330	1,330
<25	5	272	0.01	5	0.01	267	267
<25	6	55	0.00	2	0.00	54	54
<25	7	23	0.00	0	0.00	22	22
>=25 and <55	1	609,110	14.14	9,338	14.09	16,149	16,149
>=25 and <55	2	482,083	11.19	7,535	11.37	16,037	16,037
>=25 and <55	3	491,340	11.41	7,626	11.51	16,047	16,047
>=25 and <55	4	514,016	11.94	7,728	11.66	16,070	16,070
>=25 and <55	5	191,466	4.45	2,975	4.49	15,266	15,266
>=25 and <55	6	54,450	1.26	861	1.30	12,715	12,715
>=25 and <55	7	24,450	0.57	351	0.53	9,883	9,883
>=55 and <65	1	171,792	3.99	2,677	4.04	15,128	15,128
>=55 and <65	2	284,863	6.62	4,321	6.52	15,676	15,676
>=55 and <65	3	116,566	2.71	1,830	2.76	14,522	14,522
>=55 and <65	4	46,710	1.08	697	1.05	12,241	12,241
>=55 and <65	5	14,908	0.35	223	0.34	7,852	7,852
>=55 and <65	6	5,330	0.12	99	0.15	4,034	4,034
>=55 and <65	7	4,399	0.10	69	0.10	3,477	3,477
>=65	1	540,665	12.56	8,337	12.58	16,095	16,095
>=65	2	538,220	12.50	8,293	12.52	16,093	16,093
>=65	3	86,954	2.02	1,303	1.97	13,931	13,931
>=65	4	19,814	0.46	317	0.48	9,029	9,029
>=65	5	7,163	0.17	108	0.16	5,003	5,003
>=65	6	3,152	0.07	45	0.07	2,648	2,648
>=65	7	2,174	0.05	34	0.05	1,922	1,922
Total		4,306,311	100.00	66,251	100.00	268,910	268,910

A.2 CONSTRUCTING FISCAL UNITS

In order to construct comparable variables for the two files to facilitate the matching, the sociological households in the Household Budget Survey have been first split into fiscal units. In this section we briefly discuss how households in the Household Budget survey of 2000 have been split up into fiscal units. We mainly considered family conditions and wealth conditions for constructing fiscal units. Finally we regrouped the fiscal units into their respective sociological households and constructed some extra variables. In first splitting off fiscal units we try to construct (fiscal) variables that also would have been included in the fiscal sample. A brief description of each step that was performed is given in each of the following subsections.

A.2.1 Family conditions

A first split is made based solely on relationship codes. Put differently, we look at the (family) tie each of the household members has with respect to the head of the household. As such, couples that live together but that are not married will be split up, since both of them will receive a separate fiscal form to fill out and are thus treated as two different fiscal units. If the cohabiting partner has children then the latter are taken as dependents of that cohabiting partner. That is, he or she and his or her children make up a fiscal unit that is different form the head and his or her children, if any. Other examples where a split was performed based on this criterion, was when two parents or parents-in-law were part of the household. They are considered married couples and were split off as (a) distinct fiscal unit(s). In this step we have not yet considered any income variables. The sole purpose was to split off those persons or couples that, on the basis of their relation to the head of the household, will most likely be treated as separate fiscal units by the tax authorities.

A.2.2 Wealth conditions

Next we construct four broad income categories from the income codes that we identified in Table 5 to Table 10 of the main text. The categories thus formed are: professional income, income accruing from (invested) property, income from equity and other income. These categories, taken together, make up a person's or household's net wealth and will be used to determine whether or not taxes have to be paid. Income codes that are reported at the household level are attributed to the head of the household, since it is practically impossible to know to whom they really belong.

For certain income concepts some manipulation was done in order to make the concepts comparable to what is usually observed in fiscal data. For pension amounts received, for example, we had to determine whether or not it concerned periodic payments. If not, the capital amount was converted to a periodic payment. Note that no household members are split off in this step. It is an intermediate step to determine a person's net wealth position. It is net wealth that will be used in the next step to decide whether or not a person has to be treated as a separate fiscal unit.

A.2.3 Splitting of fiscal units based on income and wealth

We considered next the net wealth of potential persons at charge of the household head. Children of a married couple and older than sixteen, for example, whose net income exceeds 76,000 Belgian francs, are no longer considered at charge of the household head. Similar rules apply for other persons at charge. Persons potentially dependent of the head of the household but that have net incomes that exceed the legal limits are split off as potential fiscal units. This way we constructed a file of 4,260 potential fiscal units from the original 3,816 households in the Household Budget Survey.

These data now permit us to simulate which fiscal units will actually receive a tax form to fill out and return. People that do not have any professional income and whose net subsistence amounts or assets are less than the tax free amount will not be sent a tax form or their form will not be processed and will not turn up in the fiscal file. Persons receiving income only from pension or disability benefits are also exempt, under certain conditions, of filling out and returning a fiscal form. The amounts we use to decide whether or not a person will be sent a tax form are the broad income categories that we identified in subsection A.2.2. This elimination process resulted in a file with 3,490 fiscal units retained and 770 that were removed from the 4,260 potential fiscal units.[33]

A.2.4 Construction of Additional Variables

Per fiscal unit we complement the file with some characteristics: fiscal couple (yes/no), number of children at charge, number of other persons at charge and number of children younger than three years of age. Then we re-refine the broad income categories that we used earlier to construct the fiscal units. The result of this is a file with thirteen income concepts.[34] We apply Belgian fiscal legislation rules to these amounts to obtain comparable amounts for the two files (Household Budget Survey and Fiscal File).

Finally, we constructed some additional variables per sociological household of which at least one member was (probably) sent a tax form and for whom the form would also be processed. To do this we reconstructed the sociological households, that is, we regrouped fiscal units belonging to one sociological household. The variables that can then be constructed are: total number of household members, total number of fiscal forms in the household, total number of married couples in the household, and total number of persons younger than 14 years of age. As such we obtained a file containing 3,207 sociological households of which at least one member is considered to be a fiscal unit. The variables that were ultimately retained in this file can be found in Table 17. The variables that are used in the empirical exercise of the main text are in bold. Net disposable income is constructed from the income variables. It is net disposable income that has been used in the application in section 5 of the main text.

---

[33] If the originally envisaged fiscal file would have been at our disposal no such elimination would have been necessary. Here it was convenient for the empirical exercise that has been carried out as described in section 5 of the main text (see also footnote 22)

[34] The thirteen concepts constructed are: Income from labour, Income from an independent activity, Income from pension, Unemployment benefits, Disability benefits, Income from (invested) property, Other income from property, Alimony received, Alimony paid, Property taxes, and two additional variables indicating whether a fiscal processing occurred the previous year and the results thereof. One variable will be the amount to be retrieved, hence taxes paid in excess of what was needed and an amount to be paid after processing, hence if a fiscal unit has paid too little taxes.

TABLE 17 VARIABLES RETAINED FOR 3,207 SOCIOLOGICAL HOUSEHOLDS

Sequence number
Household Number (as in Household Budget Survey)
Civil status of the household head
Age of the household head
Sex of the household head
Region
Fiscal Couple (yes/no)
Number of children at charge (of household head)
Number of other persons at charge (of household head)
Number of children younger than 3 years of age
Number of persons in the household
Number of children younger than 14 years of age
Number of fiscal forms in the household
Total number of dependent persons
Number of married couples in the household
Total number of dependent children (household as a whole)
Extrapolation coefficient
Salaried Income (man/woman)
Income from independent activity (man/woman)
Income from pension (man/woman)
Capital from pension (man/woman)
Unemployment benefits (man/woman)
Disability benefits (man/woman)
Income from real estate
Other income from property (immobile)
Alimony paid
Alimony received
Taxes paid on income from property
Taxes to be recovered from previous fiscal year
Extra taxes from previous fiscal year to be paid
Net disposable income