Review of Income and Wealth
Series 48, Number 4, December 2002

COMBINING HOUSEHOLD INCOME AND EXPENDITURE DATA IN POLICY SIMULATIONS

BY HOLLY SUTHERLAND*

University of Cambridge and DIW, Berlin

REBECCA TAYLOR

National Centre for Social Research, London

AND

JOANNA GOMULKA

London School of Economics

The analysis of the distributional impact of fiscal policy proposals often requires information on household expenditures and incomes. It is unusual to have one data source with information on both and this problem is generally overcome with statistical matching of independent data sources. In this paper Grade Correspondence Analysis (GCA) is investigated as a tool to improve the matching process. GCA draws out the relationships between the common variables to enable the sample to be partitioned into more homogeneous groups, prior to matching. An evaluation is conducted using the UK Family Expenditure Survey, which is unusual in containing both income and expenditure at a detailed level of disaggregation. Imputed expenditures are compared with actual expenditures through the use of indirect tax simulations. The most successful methods are then employed to enhance data from the Family Resources Survey and the synthetic dataset is used as a microsimulation model database.

INTRODUCTION

Statistical matching and related dataset enhancement techniques (data "fusion") have been used when a single source of micro-data does not contain all the information necessary for a particular task. The use of such techniques may also form part of a wider strategy for improving the coherence of national data

Note: The research reported in this paper was supported by the ESRC *Analysis of Large and Complex Datasets* (ALCD) Programme (H519255052). Data from the *Family Expenditure Survey* are Crown Copyright. They have been made available by the Office for National Statistics (ONS) through the Data Archive and are used by permission. Data from the Family Resources Survey have been made available by the Department of Social Security (DSS) through the Data Archive. The ONS, the DSS and the Data Archive bear no responsibility for the analysis or interpretation of the data reported here. Thanks are due to Neela Dayal and Lavinia Mitton for considerable preliminary work. This is a substantially revised version of a paper presented under the title of "Creating Order out of Chaos? Identifying Homogeneous Groups of Households across Multiple Datasets"' to the 26th General Conference of the International Association for Research in Income and Wealth in Cracow. We are grateful for the comments received there and for the exceptionally helpful comments and suggestions from two referees. However, the usual disclaimers apply.

*Correspondence to: Holly Sutherland, Department of Applied Economics, Austin Robinson Building, Sidgwick Avenue, University of Cambridge, Cambridge CB3 9DE, UK (Holly.Sutherland @econ.cam.ac.uk).

collection and for the efficient and effective use of limited data resources. However, any review of practical applications of the methods, such as Cohen (1991) reveals that no major study has consisted simply of a straightforward application of a chosen algorithm. There is inevitably a substantial degree of *ad hoc* and problem-specific treatment. Despite over 25 years of history, statistical matching remains more of a craft than a science.

Our particular problem is to create a synthetic micro-dataset containing information on both household incomes and expenditures. From the policy perspective both types of information are necessary for the analysis of the effects of the combination of direct and indirect personal taxes (Salomäki, 1996; Office for National Statistics, 2001).

Most countries do not have single sources of micro-data including high-quality disaggregated information on both incomes and expenditures. The income data in most Household Budget Surveys is usually very limited and of relatively low quality. However, the UK Family Expenditure Survey (FES) is an exception. It contains both types of information and has been used for many years as the database of official analyses of both household income and expenditures (Office for National Statistics, 1996). The FES offers an opportunity to evaluate experiments with imputation methods that may be applied to other datasets. An additional, specific objective is to impute expenditure variables into a second British household survey dataset: the Family Resources Survey (FRS) (Department of Social Security, 1997). The potential advantages of using the FRS as database for policy analysis and simulation include a much larger sample size and detailed information necessary for the simulation of welfare benefit entitlement (such as information on savings).[1]

Section 1 provides an overview of statistical matching principles and introduces the method to be used in this study. As well as variables describing household characteristics, household income is a variable common to both datasets. Although one might imagine a straightforward relationship between household income and expenditure, two factors inhibit its identification. First, there are measurement problems: incomes and expenditures are measured over different reference periods. The period for most expenditures in FES is very short: two weeks (Dayal *et al.*, 2000). This means that while on average measured expenditures correspond to those in the population, in any one household we may observe atypical patterns such as zero expenditures on food or a very high proportion of fortnightly spending on a durable purchase (such as a car). Secondly, a linear relationship between household income and expenditure "can only be expected for a class of families homogeneous as regards tastes and needs and making their purchases on the same market" (Allen and Bowley, 1935, p. 37). Our method of identification of something like Allen and Bowley's "classes of families" makes use of Grade Correspondence Analysis (GCA) and is explained in some detail in Section 2.

Section 3 explains the matching process once the homogeneous groups have been identified. Households are ranked by income within groups and matched

[1]FRS is nearly four times the size of the FES: the 1995–96 Great Britain samples contain 26,435 and 6,690 households respectively. FRS does not cover N. Ireland so these observations were excluded from the FES data files.

across datasets according to group and rank. Section 4 presents an evaluation of alternative methods by comparing policy simulation results using both actual and imputed data from the FES. The impact of the policy changes is simulated using POLIMOD, the Microsimulation Unit's tax-benefit microsimulation model (Redmond, Sutherland, and Wilson, 1998). Section 5 presents POLIMOD results using FRS and imputed expenditures. We consider imputations to be sufficiently robust for a particular purpose if they lead us to the same policy conclusions as data taken from a single source. By their nature, there are countless policy simulations that *could* be carried out using different combinations of variables from the two datasets. Thus the conclusions we draw in Section 6 cannot be comprehensive; nor can they be applied mechanically as general rules. Instead, we focus our investigation on three practical issues that may be of relevance in other contexts where there is less scope for validation of results. These are (a) whether our chosen method of identifying groups of similar households shows sufficient promise to be implemented in other studies; (b) whether it matters if the ranking variable is not identical in the two data sources; and (c) whether it is necessary to include all variables of importance to the end analysis in the identification of the groups that are matched.

1. Statistical Matching of Income and Expenditure: The State of the Art

Statistical matching started in the early 1970s. To our knowledge Okner (1972) is the earliest generally available publication on the subject. Cohen (1991) and Baker, Harris, and O'Brien (1989) provide reviews of techniques and practical applications in the field. The problem in a nutshell is as follows. We have sample A with variables (X, Y) and sample B with variables (X, Z). We want to create data C with variables (X, Y, Z) by merging records from A and B with close values of X. This is legitimate if Y and Z are related to each other only through X, i.e. if, conditionally on X, Y and Z are independent (Sims, 1972, 1974), or if the relationship between Y and Z is known from other sources (for example, estimated from a different sample) and incorporated into the matching process (Paass, 1986). Under the assumption of conditional independence (which in practice can rarely be checked), a number of computational techniques for finding "good" matches are available. Usually the samples are divided into cells by values of X and matches allowed only within cells.

The key problem for matching and imputation is classifying the samples into homogeneous groups, with the definition of similarity between households depending on the variables which are going to be imputed. In this case, these would be households that can be expected to have similar patterns of expenditures. The most commonly used method defines groups as cells in a cross-tabulation of common variables. However, compromises have to be made in the definition of cells between the desire to match records with very close (ideally, identical) values of X, and not creating cells with small numbers of observations. Thus in general, there is a need to identify groups that are different from the "straight-edged" cells produced by a cross-tabulation.

One method is simply to regress expenditures on the common variables, using the whole sample. However, this involves an important limitation that does not apply to methods that explicitly identify homogeneous groups of households and then match individual records within the groups. The number of variables that may be estimated separately is limited. Where many variables are needed at a high level of disaggregation individual record matching is the only suitable approach.

2. Identifying Homogeneous Groups of FES Households

In order to identify groups of similar households within which to match across datasets we have used Grade Correspondence Analysis (GCA). The first step, as in any other method, is to select a group of common variables which will serve as a base for classification. Variables common to both FES and FRS are of the following five types (in addition to income):

(1) Basic information about the sample (month of interview and region).
(2) Demographic information: some variables are at the person level (age, sex and marital status) and some are at the household level (household composition).
(3) Household dwelling descriptions (category of dwelling; tenure type; total number of rooms).
(4) Variables describing ownership of durables.
(5) Individual labor market activity, at the person level (usual weekly hours, employment status; socio-economic group of the head of household).[2]

A priori, the relative effectiveness of any combination of variables, or transformations of them, is unknown. A sub-set of variables was chosen by regressing four categories of expenditure on different transformations of common variables and selecting those which were most consistently significant. The results were treated as indicative of possible suitable choices, not as an inviolable rule.

Once a group of variables has been selected, a table of their values can be constructed, with rows corresponding to households and columns to the selected variables. The next task is to classify rows into a number of groups (''clusters'') such that rows within each group are as similar to each other as possible. The method we have used for this purpose, Grade Correspondence Analysis (GCA) was initially developed to solve a classification problem for a different kind of table, namely for cross-tabulations. Suppose for example that we are interested in geographical pattern of housing tenure types in the UK, and on a sample of households we have cross-tabulated tenure (seven categories) by region (12 standard regions). To get a clearer picture we wish to aggregate regions into 3–4 groups and perhaps also tenure types into 2–3 groups, and we wish to do it in an optimal way. There may be different technical definitions of optimality (used by different varieties of so-called ''cluster analysis''), but the gist is that we want to see main geographical tendencies in housing types. It is not generally the case

[2]See Dayal *et al*. (2000) for details of how these variables were constructed. Stringent criteria for the sameness of variables were used and although the two surveys are similar in many respects it proved surprisingly difficult to define *identical* variables in each.

that adjacent categories form optimal groupings. A particular recipe for aggregation which is of interest here (Grade Correspondence Cluster Analysis, GCCA) consists of two steps. First, rows and columns of the cross-tabulation are permuted in such way that as many cells with large values as possible move close to the diagonal. Then the best aggregation is obtained by cutting the permuted table along straight lines, i.e. rows best put in the same group are now adjacent, and the same is true for columns. The whole procedure has been precisely defined, theoretically validated and made operational in Ciok *et al.* (1995). Its first step—purposeful reordering of the cross-tabulation—is GCA.

A preliminary step in GCA divides all entries in the table by the sample size. In the scaled table the sum total of all entries is 1 and they can be interpreted as probabilities. In our example it would be the probability of occurrence of a combination of tenure and region, say a proportion in the sample of owner-occupied households in the South East. The purpose of the algorithm is to find a permutation of the table (both rows and columns) which maximizes a measure of dependence between the two variables (so called Spearman's rho ρ^*). Two variables are statistically dependent if small values of one occur in general together with small values of the other, and similarly large values of one go together with large values of the other. In the context of Spearman's rho "smaller" means "preceding in the table," so the value of rho depends on the ordering of rows and columns. It can be seen that for strongly dependent variables most large entries in the cross-tabulation will be grouped close to the diagonal. GCA proceeds by permuting alternately rows and columns of the table according to a rule which assures that in each step rho is increased. The procedure stops when it cannot make further improvement. Unfortunately that does not mean that the optimal table has been found: it is possible that starting from a different initial ordering the algorithm would finish with a table with higher rho. Therefore GCA is usually repeated a number of times, from different starting points.

We have applied GCA not to cross-tabulations but to data tables, with household records as rows and variables as columns.[3] The sequence of calculations is exactly the same as described above, so cells with larger values are pushed towards the diagonal. Because the final, post-GCA table is approximately diagonal, adjacent rows are in general more similar to each other than to rows farther away, and the same is true for the columns. In particular, identical rows are placed next to each other. In other words, households which are identical with respect to the chosen variables appear in the table as homogeneous blocks of rows. Our data tables had about 6700 rows and 8 to 20 columns, so not surprisingly similarity between columns was less pronounced. However, it is still true that in the post-GCA ordering variables in general have stronger association to their neighbors than to variables farther from them. In our experience it was more visible for the variables at the extremes of the sequence, i.e. the first and the last few.

[3]This has been done before, see Ciok *et al.* (1997), Ciok (1998), and Szczesny and Pleszczynska (1997), and Szczesny *et al.* (1998), although not for large socio-economic data sets, and within a different context. A recent paper by Szczesny (2001) describes applications of GCA that are much more advanced than those used in our study.

In its original context of grade clustering, GCA would be followed by an algorithm splitting the table into optimal clusters. After one attempt we decided not to use it. The clusters formed on the basis of purely numerical information could not be described in terms of a relatively small number of variables, and had no discernible intuitive meaning. Also it was possible that, due to the differences between the FES and FRS, clusters formally corresponding to each other would not in fact be similar. We formed clusters by eye, using patterns evident in the GCA-transformed table. Typically, such tables would include a number of blocks of identical households. Some of them might be large enough to become clusters on their own. The blocks are surrounded by households which are not identical but have some values in common. We tried to identify groups of households, close to each other in the table although not necessarily adjacent, with several variables identical for the whole group. We based our clusters on such groups. Variables from the extremes of the final GCA sequence were usually more effective for classification than those from the middle.

Formation of clusters involved a number of more or less arbitrary decisions. The first, and likely the most important, was the choice of variables. If too many variables are included hardly any households are identical, and similarities between them become very difficult to detect. Therefore we experimented with a number of groups of 8 to 20 variables. Two groups were selected for further work. They were chosen because in both cases the final rho was very high, and the GCA was robust, in the sense that the results were not much altered by a change of starting point or change to a different year of FES data. For one of the groups clusters were defined in two different ways. First, clusters were created with simple definitions in terms of the variables used (e.g. two working adults, dwelling owner-occupied with mortgage, one car). These clusters followed the GCA order rather loosely, with many interleavings of different clusters. Next, clusters were created that followed the GCA ordering much more strictly, but as a result some clusters had extremely complex definitions.

Before proceeding to the next stage we checked that our clusterings did relate to spending behavior. First, GCA was applied to disaggregated expenditure variables to classify FES households according to their expenditure patterns.[4] Then the two variables, expenditure class and cluster identifier (from one of the clusterings described above) were cross-tabulated. Next, GCA was applied to the cross-tabulation. For each of the three clusterings the final rho showed clearly that there was interdependence between clusters and expenditure patterns.

3. Matching

Having identified similar groups of households in each of the datasets the households within each group were ranked by household income and then matched sequentially, starting with the lowest income household from a donor cluster matched to the lowest income household in the equivalently defined and ranked cluster in the recipient file. We call this "rank by income" matching.

[4]The 27 variables are listed in Appendix 1 and their derivation is described in Mitton (1998). The method would have been equally appropriate if we had wished to impute directly from the FES database the full set of 400 disaggregated expenditure variables.

Households from any recipient cluster receive expenditure variables only from the corresponding donor cluster. Since the donor and recipient clusters are typically of different sizes, in general a recipient record would be assigned a combination of expenditures from two (rarely more) neighboring donor records. Intuitively, the whole donor cluster is treated as one piece of cake sliced into records. The slices have to be recut and combined to create the right number for the recipient cluster. To enable combining different donor records, the expenditures are in the form of shares of the total household expenditure.

4. Evaluation using Family Expenditure Survey

No formal statistical tests exist to distinguish between the alternative sets of imputed data. Our evaluation relies on comparison of microsimulation model results when imputed expenditure data are used, in relation to results based on actual expenditure data. Comparisons were made using FES data from a two-year sample (1994–95 and 1995–96). Expenditures were imputed from one (random) half of the combined sample into the other half. Then the roles of donor and recipient were reversed, providing an actual and an imputed set of expenditures for each household in the combined sample. Many alternative sets of imputed data were generated (Taylor, Sutherland, and Gomulka, 2001). Here we select five alternatives in order to focus on the three key questions raised in the introduction:

(1) *Un-clustered*, using rank-by-income matching across the whole of each sample. In this and imputations 2 to 4 below, the definition of income used as the ranking variable before matching is household disposable income after committed expenditures i.e. minus income tax, National Insurance contributions and housing costs.[5]
(2) *Clustered-A*, which uses pre-match cluster definitions set out in Appendix 2. A distinguishing feature is that the variables used include the presence of children. Cluster formation was based on simple combinations that made intuitive sense.
(3) *Clustered-B* uses somewhat different variables than A (not including presence of children—see Appendix 2) and also followed the intuitive approach to cluster formation.
(4) *Clustered-C* used the same variable and observation ranking as version B but followed the GCA ordering much more strictly in identifying clusters: the definitions have little intuitive meaning and are complex to reproduce.
(5) *Inc-exp ranked* uses the same clusters as variant B but instead of ranking both donor and recipient by the same income variable, two different variables are used: total household income in one and total household expenditure in the other. This artificially replicates the common situation where identical matching variables are not available in the two datasets.

Table 1 shows mean total expenditure and its shares across deciles of the income distribution for the five imputed expenditure datasets, compared with actual expenditure data.

[5]For a full definition see Dayal *et al*. (2000).

TABLE 1

SHARE (PERCENT) OF TOTAL HOUSEHOLD EXPENDITURE BY EQUIVALIZED INCOME DECILE GROUP AND IMPUTATION METHOD

			% Share of Total Expenditure									
			Equivalized Household Income Decile Groups									
Imputation	Mean	SD	1	2	3	4	5	6	7	8	9	10
Actual	**13812**	**12743**	**4.80**	**3.92**	**5.60**	**7.46**	**8.69**	**9.94**	**11.66**	**13.56**	**14.62**	**19.73**
Unclustered	13812	12225	3.34	3.57	4.94	6.66	7.99	9.61	11.07	13.25	15.63	23.94
Clustered-A	13789	11486	4.58	3.92	5.49	7.33	8.82	10.05	11.29	13.53	14.85	20.15
Clustered-B	13803	11521	4.56	3.96	5.60	7.22	8.50	9.97	11.58	13.38	15.20	20.04
Clustered-C	13802	11622	4.56	4.00	5.48	7.23	8.54	10.02	11.66	13.28	15.33	19.90
Inc-exp ranked	13803	12675	2.74	3.24	4.51	6.08	7.34	8.99	10.77	13.03	16.03	27.27

Notes: Amounts are £/year at 1994–96 prices. The modified OECD equivalence scale has been used to rank households.

We can see that total expenditure is very similar across all five variants (to be expected, given the imputation method), although the variation is somewhat less in the clustered datasets than in the un-clustered or the actual expenditure data. The share of total expenditure across the distribution of household income is much closer to the actual in the three clustered datasets than in either the un-clustered dataset or the imputation that uses different variables in the two datasets to rank within clusters. These both over-estimate the share of the high-income groups and under-estimate the share of expenditure among those on low incomes. There is little to choose between the three sets of clusters (A, B, C) in terms of the aggregate results shown in this table.

In order to assess the quality of the imputations in terms of the pattern of expenditure by income level we compare simulated indirect taxes for the existing tax system and some policy reforms, across the distribution of equivalized household income.[6] This approach allows us to summarize the effect of the imputations by combining information from the imputed expenditure datasets in a way that is both complex and is also relevant to questions of the type that policy simulation models are routinely called upon to address.

Table 2 shows the total amount of Value Added Tax (VAT) that is calculated to be due on the expenditures. This shows the extent to which the imputed total spending on goods and services that attract VAT reproduces actual expenditure on this group of goods and services. Appendix 1 lists expenditure variables by their tax treatment. Estimates of revenue from VAT are all close: within 1.5 percent of the actual estimates.

However, when we disaggreagte VAT payments by level of household income, we find more variation.[7] Figure 1 shows the distributional impact using the five alternative expenditure datasets compared with the real data. The line

[6]Using the modified OECD equivalence scale (1 for the first adult, 0.5 for additional adults (aged 14+) and 0.3 for additional children) and counting each household once in the ranking. Sensitivity to other equivalence scales was explored. Results were slightly different but conclusions were unaffected.

[7]Although the expenditure data varies between each comparison, a common micro-database for household income and characteristics variables is used. FES data are updated to 2000–01 prices and incomes, and use 2000–01 UK tax and social security policy as a starting point. All results use the data re-weighted to represent the national population.

TABLE 2
POLIMOD Output by Imputation Method, Using FES

Imputation	VAT £ million/year	Uniform Rate of VAT: Revenue-neutral Rate %	Uniform Rate of VAT: Households Losing %
Actual	**37253**	**9.98**	**51.40**
Unclustered	37479	9.96	52.17
Clustered-A	37354	10.06	50.72
Clustered-B	37393	10.04	51.22
Clustered-C	37432	10.04	51.28
Inc-exp ranked	37429	10.05	52.93

Notes: Results are expressed in terms of 2000–01 prices and incomes. "Losing"' is defined as being worse off by £0.10 or more per week.

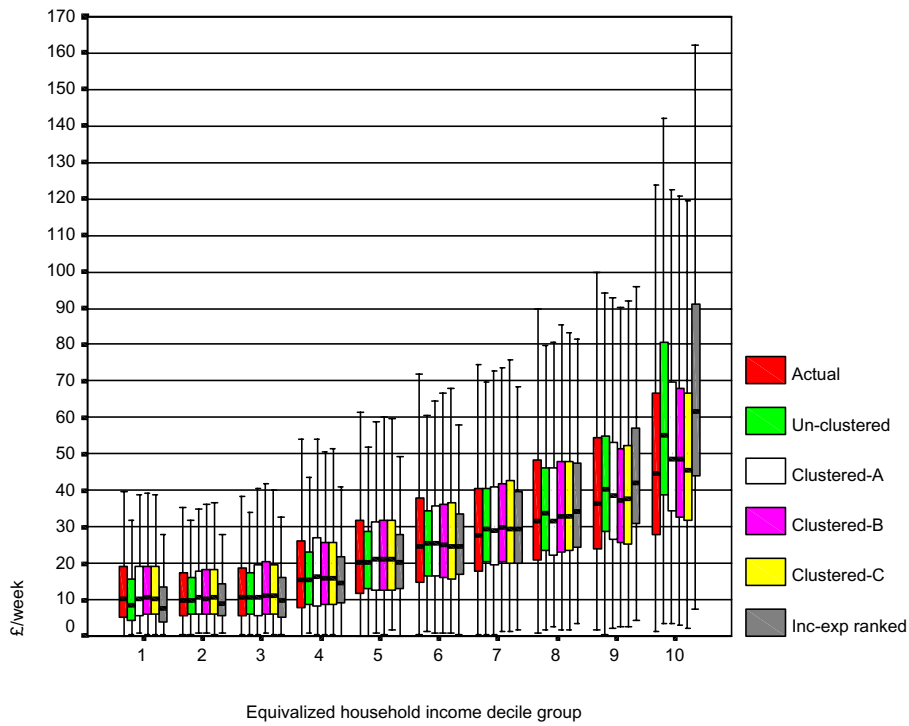

Figure 1. Incidence of VAT by Household Income: Comparison Using Actual Expenditure Data and Five Alternative Imputations

through the centre of the boxes in the plots represents the median, the box represents the inter-quartile range, and the whiskers reach out to the lowest and highest values.[8] There is a noticeable difference in the median and inter-quartile range of un-clustered estimates and the actual values in the bottom and top decile

[8]Excluding outliers, which are defined as more than 1.5 times the third quartile above the box and 1.5 times the first quartile below the box.

groups. There is little to choose between the clustered estimates, but those using non-identical ranking variables prior to matching perform relatively poorly.

The simulation of policy *changes* can provide additional information about the distribution of expenditures of different types. We replace the current variable rate structure of VAT—shown in Appendix 1—by a uniform rate on all goods and services, set at a revenue-neutral level.[9] This results in some households gaining and some losing, depending on their spending patterns. The proportion in each group, as well as the estimated revenue-neutral VAT rate are examples of estimates that are commonly used by policy makers, and which at the same time might be expected to be sensitive to differences in patterns of expenditure. However, Table 2 shows that this is not the case for aggregate results: the value of the revenue-neutral uniform rate (around 10 percent) is not very sensitive to the imputation method and the proportion of households losing from the reform is also fairly stable. Figure 2 shows the value of the median change in VAT for each income decile group. There is some divergence in the imputation which uses non-identical ranking variables at high and low income levels, but otherwise the distributions are remarkably similar.

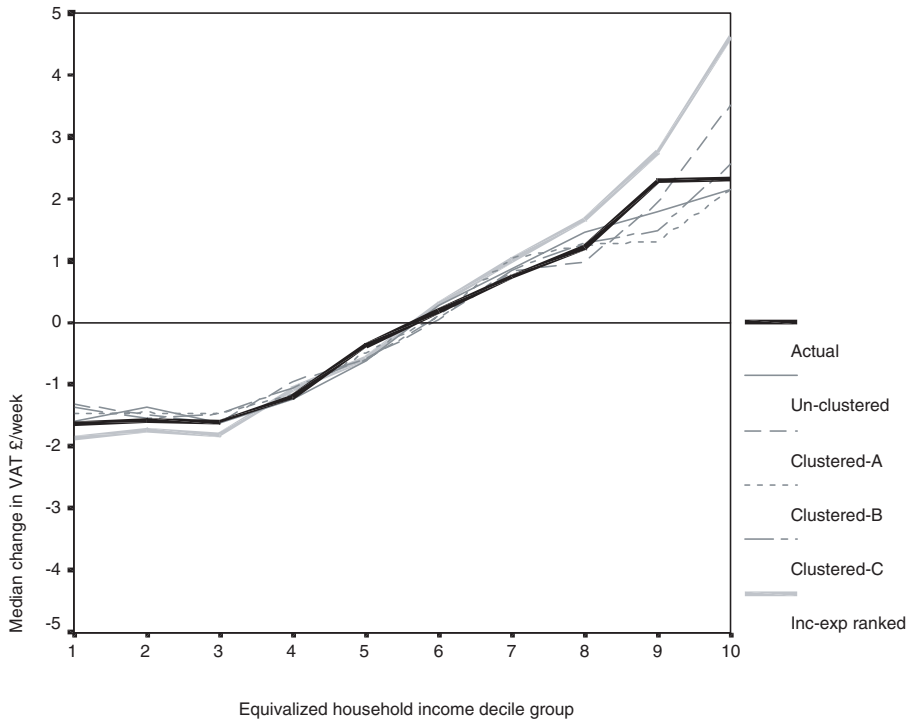

Figure 2. A Uniform Rate of VAT: Median Changes in VAT Payments by Household Income Using Actual Expenditure Data and Five Alternative Imputations (all households)

[9]In calculating the impact of tax changes, it is assumed that households do not change the quantity of goods bought.

Sampling error is one influence explaining differential effects across datasets. The size of this effect can be used as a benchmark to provide some indication of the significance of the differences between imputations, and between results based on imputed and actual data. The actual data are split randomly into two equally-sized samples and used to generate two equivalent sets of POLIMOD results. The increase in VAT payments for households losing due to a uniform rate of VAT is examined over quantiles of the distribution of household income in Figure 3. Kernel regression is used to smooth the curves to reduce the impact of extreme values. The figure shows that there is a substantial amount of variation between the two samples of actual data, and that this is comparable in size to the differences between results using the actual and imputed estimates.[10] This suggests that the relationship between household income and size of the VAT increase is estimated "well enough" using imputed expenditures.

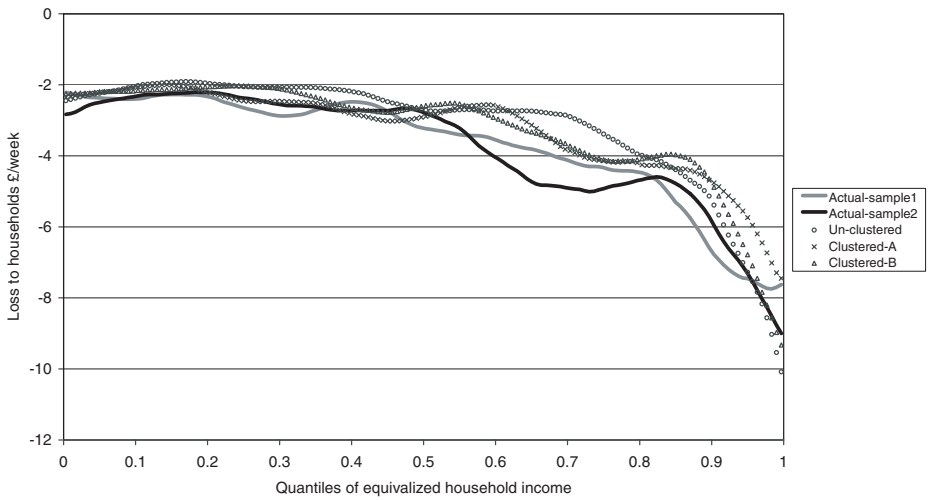

Figure 3. Losses from the Introduction of a Uniform Rate of VAT: Nonparametric Regression of the Size of the Loss on Household Income

The wide dispersion of VAT payments *within* income groups shown in Figure 1 suggests that while the imputations may perform reasonably well for the whole sample, the same may not always be the case for sub-groups. An example of this is shown in Figure 4 which plots the median change in VAT within each decile group focusing only on households with children. Using actual expenditure data shows that on average households at both the bottom and the top of the income distribution pay less tax under the uniform system. There is little change on average among middle-income households with children. However, none of the imputations capture the tax reduction for households with children in the top decile group. The un-clustered data do not successfully replicate the results using actual

[10]The imputed data were also split randomly in two so that the sample sizes of all the datasets remained comparable.

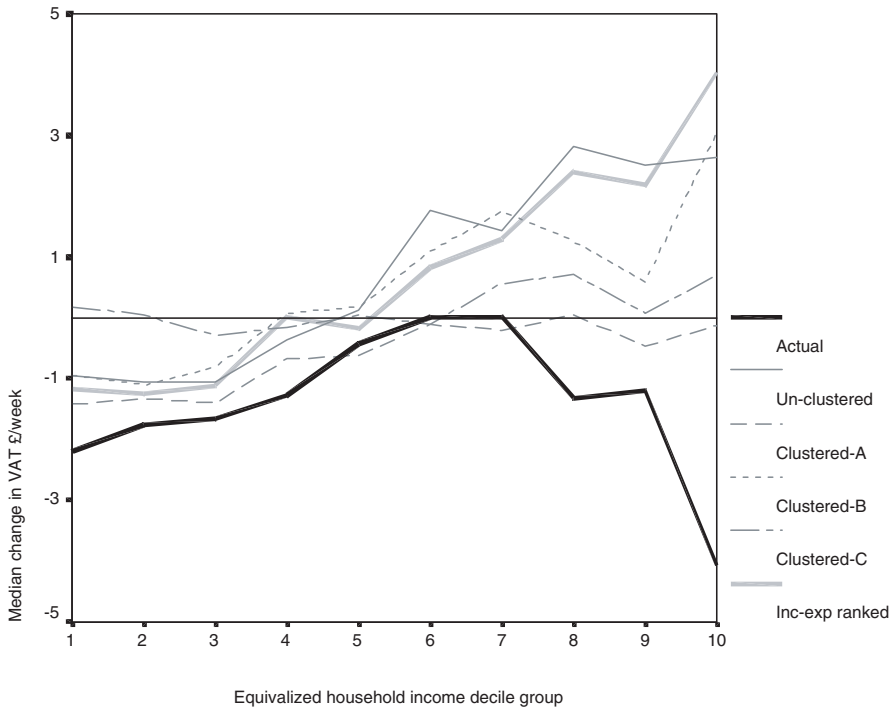

Figure 4. A Uniform Rate of VAT: Median Changes in VAT Payments by Household Income Using Actual Expenditure Data and Five Alternative Imputations (households with children)

expenditure data. And the different cluster definitions result in different estimates. In particular, the best-performing dataset is Clustered-A which is based on clusters that include presence of children in their definition. The imputation using non-identical ranking variables fails to capture the distributional impact of the VAT change on households with children: it simply reproduces the distributional profile for all households (see Figure 2).

The lessons learned from this evaluation exercise can be summarized as:

(1) Clustering into similar groups prior to ranking and matching does improve the quality of the imputations.
(2) The variable that is used for ranking should be the same in both datasets.[11]
(3) The dimensions of importance in the subsequent analysis should be included in the cluster definitions.

5. Imputation of Expenditures into Family Resources Survey Data

Two imputations were selected to be implemented with FRS data, both based on matching within clusters identified using the same variables as in the more

[11]In experiments not reported here we found that the precise definition of the income variable used to rank did not matter, so long as it was the same in both datasets (Taylor *et al.*, 2001).

successful experiments with FES: variant A (clusters including child information—see Appendix 2) and variant B (clusters not including children in their definitions).

In comparing results using FES with those using FRS and imputed expenditures it is important to remember that the expenditure data are not the only source of difference. We have found that, although the surveys are similar in many respects, the distributions of household income are significantly different (Dayal *et al*., 2000). FRS incomes are lower than FES incomes on average, but for some sub-groups the opposite is the case.[12] We would therefore not expect the composition of income decile groups in the two datasets to be identical. For this reason we would not expect the expenditures of (say) the bottom income decile group in the FRS sample to be the same as the expenditures in the bottom decile group of the FES. In addition, routine data adjustments (e.g. re-weighting to correct for differential non-response) and the process of policy simulation (e.g. the modeling of non-take-up of some social security benefits) may either exacerbate or mitigate these underlying differences. Thus, in the comparisons that follow we would not expect to find identical results using the alternative data sources even if we were able to exactly replicate the spending behavior of FES households in FRS. Rather, if the two sets of results lead us to the same policy conclusions then we can conclude that the imputations are sufficiently robust to be considered as adequate for the particular purpose.

TABLE 3
POLIMOD Estimates of VAT Under Three Policy Scenarios Using FES and Enhanced FRS

		Uniform Rate of VAT		VAT on Children's Clothing			
		Revenue-neutral			% affected		% Falling on Households
Data/Imputation	VAT £million/year	Rate %	Households losing %	Increase in VAT £million/year	All	With Children	Without Children
FES 94/5 + 95/6	37253	9.98	51.4	625	22.6	58.2	15.1
FRS + imputation A	37009	9.85	50.0	694	27.1	58.5	29.1
FRS + imputation B	37021	9.85	49.9	720	28.7	41.7	55.7

Notes: Results are expressed in terms of 2000–01 prices and incomes.

Table 3 shows the aggregate results using FES and enhanced FRS data under three policy scenarios. Estimates for the total amount of VAT under the existing tax system show that the imputed datasets contain less expenditure that attracts VAT than the FES data. This is consistent with the somewhat lower revenue-neutral uniform VAT rates obtained when using imputed data and a slightly lower proportion of households losing when the revenue-neutral rate is implemented. However, all these differences are small and there is little to choose between the two imputations. Figure 5 shows the distribution of VAT across income decile groups, which confirms the similarity of the datasets at this general level.

A third policy scenario shows the effect of the imposition of standard rate VAT (17.5 percent) on children's clothing, which is currently zero-rated. The differences are larger for this relatively small component of expenditure, affecting

[12]See also Frosztega *et al*. (2000).

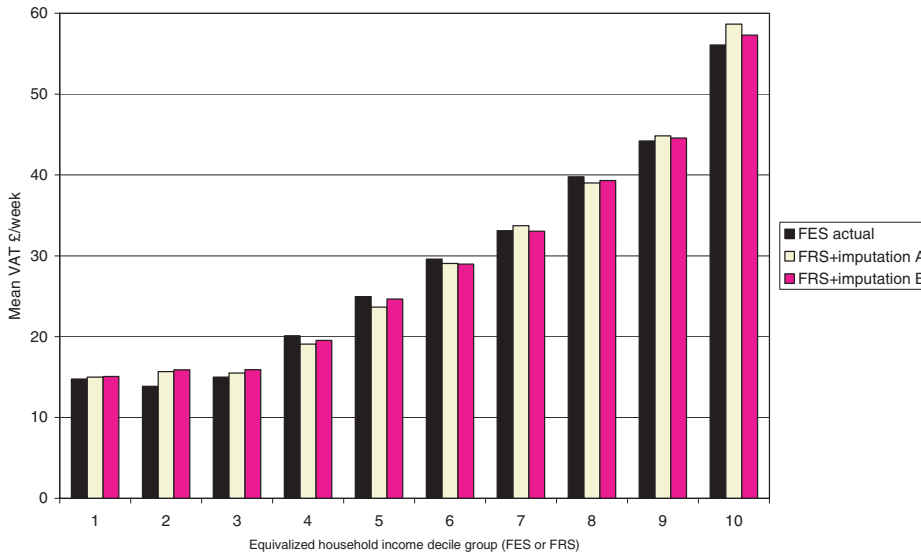

Figure 5. Mean VAT by Household Income Decile: Comparing FES Actual with FRS Imputed

only a quarter of households. The revenue estimate is between 11 percent and 15 percent larger for the imputed datasets compared with the real data. The imputed data also show a larger proportion of households being affected than the actual data. Although the FES data indicate that 15 percent of spending on children's clothing is carried out by households without children, the estimates using imputed data suggest that this proportion is even higher. This is particularly the case for variant B, which at no stage controls for the presence of children in the matching process. Without this control we find that 56 percent of the change falls on households without children and that only 42 percent of households with children are affected (compared with 58 percent in the real data). When the presence of children is controlled for, the results using imputations are closer to the actuals. However, they do not tell exactly the same story. On the one hand, variant A closely matches the actual proportion of households with children affected (59 percent compared with 58 percent). On the other hand, the share of the impact on households without children is nearly double what it is using FES (29 percent compared with 15 percent).

Figure 6 shows the average VAT paid on children's clothes across the household income distribution. The actual relationship is quite flat, and all the imputations fail to capture fully the flattening at the top of the distribution. Variant B, which does not control for children behaves particularly badly in this respect and predicts twice as much extra tax paid by the top decile group than estimated using FES.

This is even clearer in Figure 7, which plots the same information for households with children only. Variant B consistently underestimates the effect, regardless of income level. Variant A follows the actual more closely, and overall the policy analyst would probably draw similar conclusions from simulations using FRS data combined with this imputation and those using FES data.

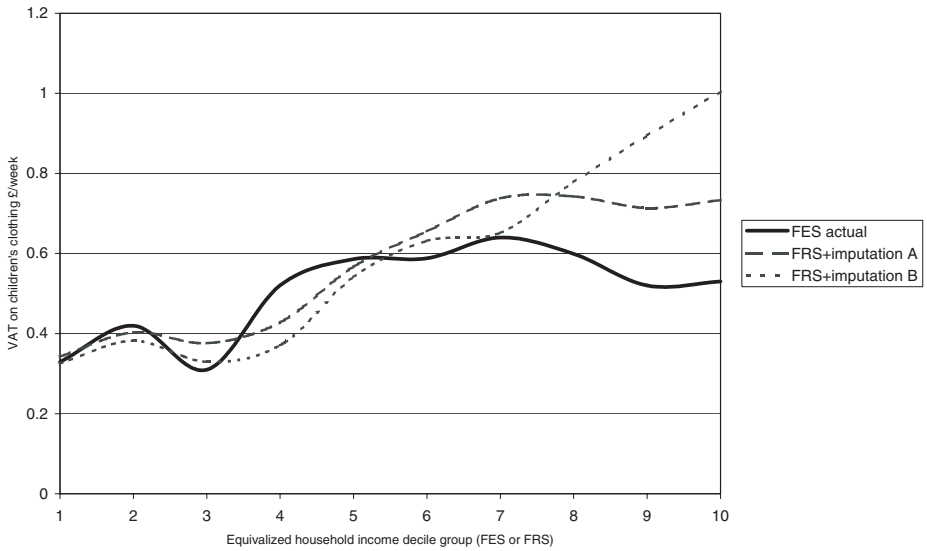

Figure 6. VAT on Children's Clothing by Household Income: Comparing FES Actual with FRS Imputed (all households)

Clearly, if changes affecting a specific group of households are to be modeled, the imputation method must take account of the characteristics of that group. In the case of tax on children's clothing, presence of children must be explicitly controlled for in the imputation process. However, there are many groups of

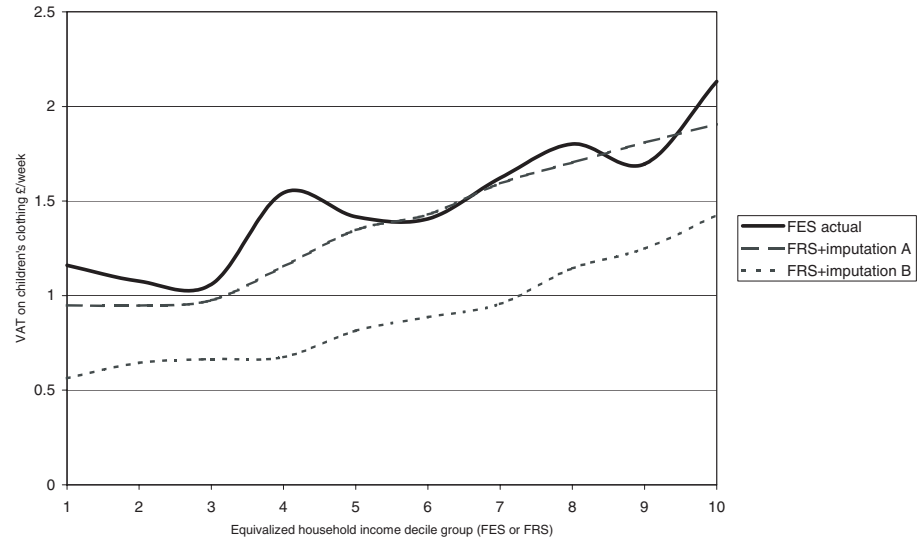

Figure 7. VAT on Children's Clothing by Household Income: Comparing FES Actual with FRS Imputed (households with children)

interest from a policy point of view and it is unlikely that a general imputation method could be found that simultaneously anticipated all such groups and was able to incorporate sufficient information. (The methods used here involved finding a balance between the number of variables and categories taken into account in the cluster definitions, and in the number of observations in each group to be matched. If the samples are divided into too many groups before matching, the chances are increased that a relatively high-income household is matched with a relatively low-income household.)

In addition, a full analysis of VAT on children's clothing might require us to look at sub-groups—say lone parent families, or those with pre-school children. In such cases it is most unlikely that imputation methods that simply controlled for children as a whole would provide results that were close to those from the original data.

6. Conclusions

It is clear that spending patterns, as captured by FES data, vary considerably from sample to sample. This can be explained by the short reference period and the dominating influence of atypical expenditures. It means that it is particularly difficult to predict expenditure patterns that capture micro-level diversity, as exhibited in the actual data. Imputed expenditure variables are only ever an adequate second-best substitute for actual data when the variables are used at a sufficient level of aggregation to mask differences that are not controlled by the imputation procedure. Since we cannot say *a priori* what this level of aggregation should be, we can only be confident in the imputations when the dimensions that are important to the end analysis have been controlled for.

Our results comparing clustered and un-clustered rank-by-income matching suggest that matching within similar groups adds to the quality of the imputation and that the identification of the groups using GCA is a fruitful approach. However, our experience with the craft of statistical matching has taught us that the identification of *optimal* groups is a goal not worth pursuing if the enhanced dataset is to be used for multiple tasks that cannot be anticipated. Many cluster definitions were "good enough" in specific contexts but none could be relied on to perform well in any context. Two firm conclusions can be drawn:

(1) It is important for the ranking variable to be the same in both donor and recipient datasets. In many practical "expenditure-to-income" imputation exercises, this is not possible and it is likely that the imputations are of lower quality than those produced in our experiments.
(2) The dimensions of importance to the subsequent policy analysis should be included in the cluster definitions. In practice, this places clear limits on the uses of any datasets that have been enhanced through matching.

Appendix 1: Expenditure Variables using FES 1995–96 by Tax Treatment

Expenditure Category	Mean £/week	SD	Percentage of Households with +ve Expenditure	Current Tax Treatment (2000–01)
1 Housing expenditure + household services + other household expenditure	23.90	54.23	95.0	VAT (17.5%)
2 Motoring expenditure	12.58	65.47	58.9	VAT (17.5%)
3 Food (which attracts VAT)	14.45	16.89	94.3	VAT (17.5%)
4 Leisure goods and services	17.82	39.07	86.2	VAT (17.5%)
5 Adult clothing and footwear	13.89	26.84	62.5	VAT (17.5%)
6 Household goods + personal goods and services	32.99	48.70	98.3	VAT (17.5%)
7 VAT-exempt goods*	45.76	113.19	98.5	No VAT
8 Food (zero-rated for VAT)	39.19	24.08	99.8	No VAT
9 Books and newspapers	4.09	5.23	91.8	No VAT
10 Domestic fuel and power	12.88	7.78	97.7	Reduced rate VAT (5%)
11 Other zero-rated goods (includes transport and drugs and medicines)	8.44	85.19	79.1	No VAT
12 Children's clothing	3.55	10.36	26.6	No VAT
13 Insurance premia	10.31	10.77	87.2	Insurance premium tax only
14 Beer	6.96	13.01	57.9	VAT + excise duty
15 Cider	0.30	1.57	20.5	VAT + excise duty
16 Fortified wine	0.37	1.30	19.2	VAT + excise duty
17 Wine	2.11	5.56	40.8	VAT + excise duty
18 Champagne	0.11	1.21	15.1	VAT + excise duty
19 Spirits	1.63	4.84	30.3	VAT + excise duty
20 Cigarettes	5.23	10.04	35.0	VAT + excise duty
21 Cigars	0.14	1.36	2.2	VAT + excise duty
22 Pipe tobacco	0.40	1.89	6.6	VAT + excise duty
23 Motor fuel	9.90	12.41	60.7	VAT + excise duty
24 Motor fuel (diesel)	0.79	4.25	5.2	VAT + excise duty
25 Pools stakes	0.38	1.29	18.7	Excise duty
26 Other betting stakes	1.18	4.67	27.2	Excise duty
27 Lottery stakes	2.30	3.18	69.7	Excise duty

*Includes postal services, life insurance, financial services, education, health, burial and cremation and trade union and professional subscriptions.

Appendix 2: Cluster Definitions

Variables used in Cluster Definition A

Demographic characteristics

children	= 1 if children in household (defined as under 16 or under 19 if in full time secondary education), 0 otherwise
male	number of males
female	number of females

Employment status

retired	number of retired adults in household
worker	number of employees and self-employed adults in household
other	number of adults in household who are neither retired nor in work (includes unoccupied, unemployed, sick)

Housing tenure

ownall	= 1 if housing tenure is owner occupier (no mortgage) or living rent free, 0 otherwise
ownsome	= 1 if housing tenure is owner occupier with mortgage, 0 otherwise
rent	= 1 if housing tenure is rented, 0 otherwise

Car ownership

car0	= 1 if no cars in household, 0 otherwise
car1	= 1 if 1 car in household, 0 otherwise
car2	= 1 if 2 or more cars in household, 0 otherwise

*Regional groups**

high	= 1 if region is in high spending group, 0 otherwise
mid	= 1 if region is in medium spending group, 0 otherwise
low	= 1 if region is in low spending group, 0 otherwise

*Standard regions of Great Britain are grouped by mean levels of household expenditure into three categories (high contains Greater London, South East and East Midlands; mid contains Scotland, South West, North West, West Midlands and East Anglia; low contains North and Wales).

Cluster Definitions B and C

These make use of the same variables as A except:

(1) They do not use the demographic variables.
(2) They divide housing tenure into more disaggregated categories (own outright, own on mortgage, rent free, rent furnished, rent from Local Authority or Housing Association, other rent unfurnished).
(3) They use more aggregated car ownership categories.

Table A2 shows the precise definition of the clusters under A.

TABLE A2
CLUSTERS FOR VARIANT A

		% FES	% FRS
1	car2 = 1, rent = 0, all adults workers, children = 0	6.2	5.8
2	car2 = 1, rent = 0, all adults workers, children = 1	6.0	5.9
3	car1 = 1, ownsome = 1, all adults workers, children = 1, high = 1	3.6	3.4
4	car1 = 1, ownsome = 1, all adults workers, children = 1, high = 1	2.4	2.3
5	car1 = 1, ownsome = 1, all adults workers, children = 0, high = 0	5.1	4.6
6	car2 = 1, ownsome = 1, some adults workers, some not	4.0	3.5
7	car1 = 1, ownsome = 1, all adults workers, children = 1, high = 0	4.8	4.2
8	car1 = 1, ownall = 1, all adults workers	2.8	2.9
9	car2 = 1, (rent = 1, some or all adults workers) or (ownall = 1, some adults workers, some not)	3.4	3.3
10	car1 = 1, ownsome = 1, other = 1, remaining adults workers	4.3	4.0
11	car2 = 1, worker = 0	1.3	1.3
12	car1 = 1, children = 1, (rent = 1, all adults workers) or (ownsome = 1, some adults workers, some not)	1.9	1.7
13	car1 = 1, ownsome = 1, worker = 0	2.0	2.1
14	car1 = 1, rent = 1, all adults workers, children = 0	2.4	2.2
15	car0 = 1, ownsome = 1, some or all adults workers	3.3	3.4
16	car1 = 1, ownsome = 0, some adults workers, some not, children = 1	1.9	1.6
17	car1 = 1, ownsome = 0, worker = 0, children = 1	1.4	1.6
18	car1 = 1, some adults workers, some not, children = 0	4.1	4.2
19	car0 = 1, ownsome = 0, some or all adults workers, children = 1	2.0	2.0
20	car0 = 1, ownsome = 0, some or all adults workers, children = 0	3.8	3.7
21	car0 = 1, worker = 0, children = 1	4.2	4.7
22	(car0 = 1, ownsome = 1, worker = 0) or (car1 = 1, ownsome = 0, all adults other), children = 0	2.4	2.5
23	car1 = 1, ownall = 1, all adults retired or some retired and some other, children = 0	3.9	4.0
24	car1 = 1, rent = 1, all adults retired or some retired and some other, children = 0	1.7	1.8
25	car1 = 1, ownall = 1, all adults retired, children = 0, (female = 1, high = 0) or (female = 2, mid = 1) or (female = 1, male = 1, high = 0)	3.7	3.9
26	car0 = 1, ownsome = 0, all adults other,children = 0	3.8	4.1
27	car0 = 1, rent = 1, all adults retired or some retired and some other, children = 0, (but not in clusters 29, 30)	2.8	3.2
28	car0 = 1, ownall = 1, all adults retired or some retired and some other, children = 0 (but not in cluster 31)	3.5	4.2
29	car0 = 1, rent = 1, all adults retired, children = 0, (male = 1, high = 0) or (female = 2, mid = 1) or (female = 1, male = 1, high = 0)	1.7	1.9
30	car0 = 1, rent = 1, retired = 1, female = 1, children = 0, high = 0	3.1	3.4
31	car0 = 1, ownall = 1, retired = 1, female = 1, children = 0, high = 0	2.3	2.5

Notes: Percentages shown are for combined 1994–95 and 1995–96 FES datasets and for 1995–96 FRS.

REFERENCES

Allen, R. G. D. and A. L. Bowley, *Family Expenditure. A Study of Its Variation*, P.S. King & Son Ltd, 1935.

Baker, K., P. Harris and J. O'Brien, "Data Fusion: An Appraisal and Experimental Evaluation," *Journal of the Market Research Society*, 31(2), 153–212, 1989.

Ciok, A., "A Comparative Study of Exploratory Methods Applied to Car Switching Data," *Bulletin of the Polish Academy of Sciences—Technical Sciences*, 46, 133–48, 1998.

Ciok, A., W. Bułhak and R. Skoczylas, "Exploration of Control-Experimental Data by Means of Grade Correspondence Analysis," *Biocybernetics and Biomedical Engineering*, 17, 101–13, 1997.

Ciok, A., T. Kowalczyk, E. Pleszczynska and W. Szczesny, "Algorithms of Grade Correspondence—Cluster Analysis," *Archiwum Informatyki Teoretycznej i Stosowanej* (The Collected Papers on Theoretical and Applied Computer Science), 7, 5–22, 1995.

Cohen, M. L., "Statistical Matching and Microsimulation Models," in C. F. Citro and E. A. Hanushek (eds), *Improving Information for Social Policy Decisions. The Uses of Microsimulation Modeling*, *Vol. II*, National Academy Press, Washington D.C., 1991.

Dayal, N., J. Gomulka, L. Mitton, H. Sutherland and R. Taylor, "Enhancing Family Resources Survey Income Data with Expenditure Data from the Family Expenditure Survey: Data Comparisons," Microsimulation Unit Research Note MU/RN/40, Department of Applied Economics, University of Cambridge, 2000.

Department of Social Security, *Family Resources Survey: Great Britain 1995–96*, The Stationery Office, London, 1997.

Frosztega, M. and the Households Below Average Income team, "Comparisons of Income Data between the Family Expenditure Survey and the Family Resources Survey," *GSS Methodology Series*, No. 18, Office for National Statistics, London, 2000.

Mitton, L., "POLIMOD: The Calculation of VAT and Excise Duties on Household Expenditure," Microsimulation Unit Research Note MU/RN/29, Department of Applied Economics, University of Cambridge, 1998.

Okner, B., "Constructing a New Microdata Base from Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement*, 1, 325–62, 1972.

Office for National Statistics, *Family Spending*: *A Report on the* 1995–96 *Family Expenditure Survey*, The Stationery Office, London, 1996.

———, "The Effects of Taxes and Benefits on Household Income 1999–2000," *Economic Trends*, 569, The Stationery Office, London, 2001.

Paass, G., "Statistical Match: Evaluation of Existing Procedures and Improvement by Using Additional Information," in G. H. Orcutt, J. Merz and H. Quinke (eds), *Microanalytic Simulation Models to Support Social and Financial Policy*, Elsevier Science Publishers B.V., 1986.

Redmond, G., H. Sutherland and M. Wilson, *The Arithmetic of Tax and Social Security Reform*: *A User's Guide to Microsimulation Methods and Analysis*, Cambridge University Press, Cambridge, 1998.

Salomäki, A., "Including Consumption Expenditure and Welfare Services in a Microsimulation Model," in A. Harding (ed.), *Microsimulation and Public Policy*, Elsevier, Amsterdam, 1996.

Sims, C. A., "Comments and Rejoinder," *Annals of Economic and Social Measurement*, 1, 343–5, 355–7, 1972.

———, "Comment," *Annals of Economic and Social Measurement*, 3, 395–7, 1974.

Szczesny, W., "Grade Correspondence Analysis Applied to Contingency Tables and Questionnaire Data," *Intelligent Data Analysis*, 5, 1–35, 2001.

Szczesny, W., A. Ciok and E. Pleszczynska, "Clustering Land Districts According to their Farm Magnitude Repartition," *Statistics in Transition*, 3, 757–68, 1998.

Szczesny, W. and E. Pleszczynska, "A Grade Statistics Approach to Exploratory Analysis of the HSV Data," *Biocybernetics and Biomedical Engineering*, 17, 235–45,1997.

Taylor, R., H. Sutherland and J. Gomulka, "Using POLIMOD to Evaluate Alternative Methods of Expenditure Imputation," Microsimulation Unit Research Note MU/RN/38, Department of Applied Economics, University of Cambridge, 2001.