COMPOSITIONAL DATA ANALYSIS

AND ZEROS IN MICRO DATA

by

Jane M. FRY

Centre of Policy Studies, Monash University

Tim R.L. FRY and Keith R. MCLAREN

Department of Econometrics, Monash University

CENTRE of

POLICY

STUDIES and

the IMPACT

PROJECT

ABSTRACT

The application of compositional data analysis methods in economics has some attraction. In particular, this methodology ensures that the stochastic component of budget share models will satisfy the restriction of shares to the unit simplex. The methodology relies upon the use of "log-ratios" in the statistical analysis. Such an approach is not possible when the data to be analyzed includes observations where the observed budget share is zero. We, therefore, extend the methods of compositional data analysis to the situation where the data to be analyzed includes observations where the observed budget share is zero. The modified compositional data methods are discussed both in statistical terms and through potential economic interpretations of the method. Further, the modified methodology is applied to the 1988 Australian Household Expenditure Survey yielding estimates for a system of Engel curves.

TABLE OF CONTENTS

COMPOSITIONAL DATA ANALYSIS AND ZEROS IN MICRO DATA

JANE M. FRY

TIM R.L. FRY AND KEITH R. McLAREN*

1. INTRODUCTION.

The statistical literature has advocated the use of compositional data analysis methods when working with data on (budget) shares (for a comprehensive review of these methods see Aitchison (1986)). This methodology has the advantage that it ensures that the stochastic component of a model for budget shares will satisfy the restriction that shares should be constrained to the unit simplex. In earlier work (Fry *et al* (1996), McLaren *et al* (1995)) we have argued that compositional data methods should be used in the econometric specification of models of budget shares. Suitably modified to allow the mean to be parameterized on the basis of economic theory, the application of this methodology yields the following estimable form for our system of equations

$$y_i = \log\ \left(\frac{w_i}{w_N}\right)\ = \log\ \left(\frac{W_i}{W_N}\right)\ + v_i,\ \ i = 1, \ldots, N - 1 (= n) \qquad (1.1)$$

with w denoting the observed share, W denoting the "deterministic" component derived from economic theory and $\mathbf{v} = (v_1 \ldots v_n)'$ a multivariate normal disturbance term. This is simply a non-linear multivariate regression model and can be easily estimated with existing software for moderately large n.

If W is derived from an indirect utility function that satisfies certain "regularity" conditions, the attraction of this approach is that the shares are bounded to the unit simplex in terms of both the deterministic and the stochastic components. Unfortunately, a problem arises if any w is zero. Traditionally in demand analysis this is avoided by assuming that quantity demanded, q_i, is always positive and/or by the use of aggregate time series data. However, in microeconomic data, as collected in either expenditure surveys or in panel studies, the occurence of zero observations for shares is common. These zeros would seem to preclude the use of compositional data methods with such microeconomic data.

On the other hand, the model in (1.1) does have a certain intuitive appeal in the case of zero shares. If, for a certain range of values of the exogenous variables, there is a preponderance of zero observations for the share of a particular commodity, then it

1

would be desirable for an estimation method to select values of the model parameters such that the deterministic part of the model also gave zeros (corner solutions) when evaluated at those values of the exogenous variables. With a Normal specification of the v_i, the above model achieves this in that for any likely value for v_i, a $-\infty$ for the left hand side effectively forces a $-\infty$ for the deterministic component on the right hand side. The question that remains is how to handle the $-\infty$ in a numerical sense.

The problem of zero shares is not unique to microeconomic data and the statistical literature does include discussion of this modeling situation. This paper reviews and extends the existing statistical methodology for modeling compositions with zero components (shares)[1]. The modified version of the compositional data analysis method is then applied to Australian household expenditure data to illustrate the modeling procedure. The plan of the rest of the paper is as follows. Section two discusses compositional data analysis and the presence of zeros. It looks at the rationalizations used for the occurence of zeros and the strategies proposed to modify the methodology. The discussion is concerned both with the statistical details and the "economic" interpretation of the strategies. Section three describes the application of the modified zero replacement methodology to estimate a system of Engel curves using the 1988 Australian household expenditure survey. Finally section four contains some concluding remarks.

2.   COMPOSITIONAL DATA ANALYSIS AND ZEROS.

In the statistical literature, two explanations for the occurence of zero observations are proposed. These are rounding (or trace elements) and "essential" (or true) zeros. The first of these rationalizes that the zero observation is an artefact of the measurement process. Basically, if we had a more accurate measurement instrument we would record a non-zero observation for the share. Thus the observed zero is a proxy for a very small number. The second rationalization argues that the observation should be zero as the true generating process leads to the occurence of zeros. The proposed modifications to deal with zero observations can then be derived by considering the cause of the zero (see Aitchison (1986) pp 266-274). In terms of economic modeling we might argue that of these two causes the second is analogous to the situation of a corner solution and so, if we believe that our zeros are the result of corner solutions, the modified zero replacement methodology may be useful. On the other hand, the existence of zeros through measurement errors is potentially less likely in economic data.

Although the "solutions" to the problem of zero observations are discussed by Aitchison (1986) according to the cause of the zero, we present them as competing techniques to deal with both types of zero and discuss the applicability in the case

---

[1]An excellent summary of the economic literature concerning zeros can be found in Pudney (1989).

of microeconomic data. The techniques to be considered are amalgamation, zero (or trace) replacement, modified Box-Cox and conditional modeling[2].

Amalgamation is the reduction of the number of components in the composition by the grouping together of certain components. That is, we decrease the number of commodities by aggregating into broader commodity groups. This is a simple approach and may have some attraction. However, if the emphasis of the modeling exercise is to study behavior at the disaggregate level then aggregation is *a priori* not an appropriate technique to deal with the zero observations. For much microeconometric work the emphasis is on disaggregate modeling and so we rule out the use of amalgamation.

Zero (trace) replacement is a technique usually discussed in the context of zeros caused by the measurement process. It is, however, possible to use it no matter how the zeros arose. The zero replacement technique and the other compositional data techniques for dealing with zeros typically assume that the number of zeros for each observation is fixed at M and with the same components being zero for each observation. In economic data that is unlikely to be true. However, there is no reason why the techniques cannot be applied, observation by observation, with M varying.

The zero replacement technique assumes that a composition has M zero and $N - M$ non-zero components. It is recommended that the zeros be replaced by "small" values. In particular, using arguments based upon a ternary representation of the data (Aitchison (1986) pp 266-267), it is suggested that we replace the zeros with $\tau_A = \delta(M + 1)(N - M)/N^2$ and then reduce the non-zeros by $\tau_S = \delta M(M + 1)/N^2$, where δ is the maximum rounding error. Bounds for δ can then be set within this ternary framework.

In terms of its application to (economic) data there is a problem with this technique as it stands. The compositional data methodology relies upon the use of share ratios. The zero replacement procedure is not ratio preserving. In particular, the ratios for the non-zero shares are distorted by the zero replacement procedure. For example, consider the case where $N = 5$, $M = 2$ and $\delta = 0.02\dot{7}$ (see Table 1). In this case the two zeros are each replaced by *0.01* and the non-zeros are each reduced by the appropriate amount. This does not preserve the share ratios.

An alternative procedure is to replace the zeros by the same number, τ_A, but to reduce each non-zero by $w_i \times \tau_S$. This both retains the share ratios for the non-zero components and makes an appropriate zero replacement. In the context of budget share

---

[2]It has also been proposed to replace the shares by ranks (Bacon-Shone (1992)). This discards a large amount of information and will not be considered here.

modeling this procedure has the appeal that the expenditure implicitly allocated as zero replacement is withdrawn from the non-zero commodities in an (approximately) optimal manner. That is, the amount taken from the non-zeros is proportional to the size of that non-zero value. We call this alternative zero replacement procedure modified Aitchison.

Table 1: Zero Replacement Procedures, $N = 5, M = 2, \delta = 0.02\dot{7}$.

		Commodity (i)				
		1	2	3	4	5
Observed						
Expenditure ($)		0	15	0	75	10
Share	w_i	0.00	0.15	0.00	0.75	0.10
	y_i	$-\infty$	$\log(1.5)$	$-\infty$	$\log(7.5)$	$\times$
Aitchison	w_i	0.01	0.14\dot{3}	0.01	0.74\dot{3}	0.09\dot{3}
	y_i	$\log(\frac{0.01}{0.093})$	$\log(\frac{0.143}{0.093})$	$\log(\frac{0.01}{0.093})$	$\log(\frac{0.743}{0.093})$	$\times$
Modified Aitchison	w_i	0.01	0.147	0.01	0.735	0.098
	y_i	$\log(\frac{0.01}{0.098})$	$\log(1.5)$	$\log(\frac{0.01}{0.098})$	$\log(7.5)$	$\times$

A related issue is then the setting of an appropriate range for the zero replacement value τ_A (and hence for δ and τ_S) in the modified Aitchison technique. In terms of expenditures the minimum value a zero should be replaced with is *0.01* (one cent). If we divide this by the maximum value of total expenditure in our data then we produce a sensible minimum value for our zero replacement τ_A. We can then find δ and τ_S. Analogously, if we divide *0.01* by the minimum value of total expenditure in our data we will produce a sensible maximum value for our zero replacement procedure. It is the combination of modified Aitchison and this setting of τ_A, δ and τ_S that we recommend in the use of the zero replacement technique[3].

Two points need to be considered in the use of zero replacement. Firstly the sensitivity to the replacement procedure. In this situation we need to consider different values of τ_A (equivalently δ) and see whether our parameter estimates change dramatically. This is essentially a robustness issue. If the estimates do not change dramatically then we can have some confidence in them. The second issue is whether our replacement observations are "outliers" in our statistical modeling. This issue can be assessed using the usual residual diagnostics.

---

[3]Another approach is to argue that zero should be replaced by the smallest possible number and this could be found by assessing the numerical accuracy (storage) capacity of the software to be used in the modeling. This would have the disadvantage of introducing a potential dependence of results on the hardware or software used.

As discussed in Fry *et al* (1996), in the case when W_i is zero (*e.g.* a corner solution) the compositional data methodology ensures that the modeled w_i will also be zero. In that paper we argue that the problem of zeros and hence $-\infty$ through the log transform is essentially a numerical one and we suggest using an appropriate replacement procedure. Such a procedure is the modified Aitchison procedure described in this paper. Thus zero replacement has this further advantage in econometric modeling.

The third technique of dealing with zero observations is the use of the modified Box-Cox transformation. That is, in place of the log-ratio transformation we use:

$$y_i = \frac{(w_i/w_N)^{\lambda_i} - 1}{\lambda_i}, \quad i = 1, ..., n \tag{2.1}$$

which can be applied so long as w_N is always non-zero. A similar transformation is applied to the deterministic part of the model, the W_i. Unfortunately, in microeconomic data it is often the case that we do not have a commodity for which w will always be non-zero. A further problem with this approach is that the transformed distributions which are produced and define the likelihood function do not have any "nice" properties (*e.g.* normality). Thus likelihood based estimation is extremely difficult[4]. The final nail in the coffin of this technique from an economist's perspective is that the technique is not "invariant" to the normalizing commodity, N.

Conditional modeling attempts to separate out the zero components and model them and the non-zeros using conditioning arguments. For example, if zeros are confined to the first M components, then it is possible to specify one model for the full composition with *no* zero (*i.e.N*) components and another model for the subcomposition with $N - M$ components. If these models occur with probabilities π and $1 - \pi$ respectively then the log-likelihood for the full model is a weighted sum of the log-likelihoods of the two models. This technique can be expanded and has some attraction in that it is similar the the "hurdle" models used in economics (see Pudney (1989)). However, the resultant modeling quickly becomes computationally difficult and thus we do not recommend this technique.

Summarizing the above discussion, we can see that the modified Aitchison zero replacement technique is simple to implement, easy to work with and has the added attraction of having an economic rationale for its use. In the next section we, therefore, discuss its use in an empirical application.

3. EMPIRICAL APPLICATION.

The household expenditure survey (HES), conducted by the Australian Bureau of Statistics, collects expenditure, income and other information on a representative

---

[4]Least squares techniques are also difficult to use with Box-Cox procedures.

sample of Australian households. The sampling methodology is broadly based upon that of the family expenditure survey (FES) in the United Kingdom and its primary use is in determining the expenditure weights for use in the consumer price index. In this study we use the 1988 HES at the household level to estimate a system of Engel curves[5].

The HES collects expenditure information at the "fine-level" of 421 commodities. We aggregate these into 18 commodity groups, comparable to those in Rimmer and Powell (1994), for use in our analysis[6]. Table 2 contains the percentage of zero share observations for each of our broad groups in the HES.

Table 2: Zeros in Australian Micro Consumption Data.

		% Zeros in Sample	
	Commodity Group	**Full**	**Estimation**
1	Bread, cereal & grain products	1.0657	0.8205
2	Meat & fish	4.8858	3.0769
3	Dairy, eggs & oil products	1.2872	0.9231
4	Sugar, preserves and confectionery	6.9343	5.3333
5	Fruit & vegetables	3.0450	1.6410
6	Other food and non-alcoholic drinks	0.7889	0.1026
7	Alcohol	33.9380	25.8460
8	Cigarettes & tobacco	58.5880	62.1540
9	Clothing & footwear	27.0030	26.5640
10	Housing (current)	16.097	9.5385
11	Fuel (not including motor vehicle)	1.8408	2.0513
12	Furniture and other household goods	8.1522	4.2051
13	Private transport	8.1522	1.0256
14	Public transport	56.8030	64.9230
15	Leisure goods & services	24.3880	24.9230
16	Other goods	1.3287	0.1026
17	Other services	1.4118	0.6154
18	Privately purchased health	8.5813	2.8718
	Number of Households	7225	975

Source: 1988 Australian Household Expenditure Survey.

We can see that no group always has non-zero share observations and the range of the problem of zero observations is large. Rather than deal with the large sample

---

[5]The identification of relative price terms in a demand system using a cross section survey is not possible. Hence we estimate Engel curves.

[6]The exact mapping of "fine-level" commodities to our 18 groups is available on request from the authors.

size and attendant heterogeneity of the complete HES, we restrict attention to a relatively homogeneous sub-sample for our estimation. This sub-sample comprises two adult, single or dual income, no children households. The extent of the zero observation problem for this sample can also be found in Table 2. We see that although the pattern of percentages of zero shares is different from the full HES, the estimation sample does contain a large zero observation problem. As such we believe that applying the modified Aitichison zero replacement technique to this data will be a good test of the usefulness of the technique.

To implement the zero replacement we need to determine τ_A, δ and τ_S. Using our guidelines the minimum zero replacement value is given by 0.01/1929.85 and the maximum replacement value by 0.01/79.08. To find δ and τ_S we need to know N (=18) and M. Table 3 shows that for our data the maximum number of zeros for any observation is 8.

Table 3: Number of Zeros.

Zeros (M)	Count
0	48
1	242
2	281
3	229
4	96
5	51
6	19
7	8
8	1

We therefore set M to 8 and find the values for δ and then τ_S. We find that $\delta_{\min} = 0.00001865$ and $\delta_{\max} = 0.00045524$. To investigate robustness to zero replacement we also consider the median value $\delta_{med} = 0.00023695$. The procedure is then to work through the data, observation by observation, replacing zeros with τ_A and adjusting the non-zeros by subtracting $w_i \times \tau_S$. This produces a "zero-replaced" data file for analysis. A system of Engel curves is estimated on this "zero-replaced" data. The procedure is then repeated with different values for δ and hence τ_A and τ_S. The idea is to investigate the robustness issue.

Regularity of the deterministic component of demand systems is important. The Almost Ideal Demand system does not always satisfy this condition and thus we choose to estimate the Modified Almost Ideal Demand system (MAIDS) of Cooper and McLaren (1992). This system is attractive since it preserves regularity over a wide region. The parameterization of MAIDS we use is to set, in the original notation,

$\eta = \sum \beta_i = 1$, and to normalize on minimum total income by setting K to this value (*i.e.* in the original Cooper & McLaren notation: $\kappa = \log(K)$). This gives us regularity over all of the sample space. The MAIDS specification for the shares in our Engel curve analysis is therefore given by:

$$W_i = \frac{\alpha_i + \beta_i \log(Y_i/K)}{1 + \log(Y_i/K)} \tag{3.1}$$

where Y is total expenditure. A further advantage of our estimation sub-sample is that we do not need to worry about the incorporation of demographic variables into the system.

The results of zero replacement estimation are given in Table 4 below:

Table 4: MAIDS Engel Curve Estimation Results for Varying δ.

Group	$\delta = \delta_{\min}$		$\delta = \delta_{med}$		$\delta = \delta_{\max}$	
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
1	.062225	.017652	.062223	.017652	.062222	.017653
	(.011469)	(.002644)	(.011469)	(.002644)	(.011468)	(.002644)
2	.138843	.028883	.138840	0.028885	.138837	.028887
	(.026185)	(.006036)	(.026184)	(.006035)	(.026182)	(.006035)
3	.082192	.010650	.082189	.010651	.082185	.010652
	(.010458)	(.002411)	(.010458)	(.002411)	(.010457)	(.002410)
4	.029107	.009701	.029109	.009704	.029110	.009706
	(.007798)	(.001797)	(.007796)	(.001797)	(.007794)	(.001797)
5	.070686	.028468	.070686	.028468	.070686	.028469
	(.016803)	(.003873)	(.016802)	(.003873)	(.016801)	(.003873)
6	.059984	.105776	.059983	.1055776	.059981	.105775
	(.041704)	(.009613)	(.041704)	(.009613)	(.041703)	(.009613)
7	-.006562	.058975	-.006506	.058976	-.006450	.058997
	(.035835)	(.008260)	(.035829)	(.008259)	(.035823)	(.008257)
8	.089595	.002997	.089611	.003022	.089627	.003046
	(.022389)	(.005161)	(.022381)	(.005159)	(.022372)	(.005157)
9	-.007335	.075722	-.007290	.075725	-.007244	.075728
	(.052492)	(.012100)	(.052486)	(.012098)	(.052480)	(.012097)

Table 4 continues ...

Table 4 continued.

Group	$\delta = \delta_{\min}$		$\delta = \delta_{med}$		$\delta = \delta_{\max}$	
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
10	.038871	.065440	.038907	.065436	.038943	.065433
	(.043937)	(.010128)	(.043934)	(.010127)	(.043931)	(.010126)
11	.089692	.026183	.089694	.026184	.089696	.026184
	(.018869)	(.004349)	(.018868)	(.004349)	(.018867)	(.004349)
12	-.024668	.063897	-.024638	.063893	-.024609	.063888
	(.044827)	(.010333)	(.044825)	(.010333)	(.044824)	(.010332)
13	.187931	.124753	.187932	.124752	.187932	.124752
	(.070262)	(.016196)	(.070261)	(.016196)	(.070260)	(.016195)
14	-.011814	.016007	-.011740	.016021	-.011667	.016034
	(.028656)	(.006605)	(.028652)	(.006605)	(.028648)	(.006604)
15	-.070273	.106018	-.070218	.106018	-.070163	.106018
	(.069211)	(.015954)	(.069206)	(.015953)	(.069201)	(.015951)
16	.085192	.079205	.085189	.079205	.085186	.079205
	(.034718)	(.008003)	(.034717)	(.008003)	(.034717)	(.008003)
17	.118273	.111480	.118268	.111481	.118263	.111481
	(.059219)	(.013650)	(.059218)	(.013650)	(.059217)	(.013650)
18	.068058	.068192	.067761	.068151	.067464	.068110
	(.036206)	(.0083456)	(.036194)	(.008330)	(.036183)	(.008340)

Standard errors in parentheses.

δ	Maximized Log-L
$\delta_{\min}$	29680.1
δ_{med}	29681.7
$\delta_{\max}$	29683.3

The estimates of both the parameters and the standard errors are remarkably similar for all values of δ. In addition, interval estimates for both the α's and the β's based on the extreme values of δ would overlap. The parameters where the differences are largest are the estimates of α_{18} and β_{18}. As these estimates are obtained from the adding up constraint, and so may suffer from "rounding errors" it is not surprising that they exhibit the largest differences. However, even these estimates are very close to each other. Thus we can have reasonable confidence in the robustness of the procedure. Additionally, the maximized log-likelihood values are also very close, confirming the robustness of the procedure. If a single set of estimates is to be used for interpretation, then perhaps those for the minimum-δ replacement data should be used.

Given the normalization of MAIDS achieved by setting K equal to the minimum expenditure in the sample, sufficient conditions for the regularity of the underlying demand system are that $\alpha_i, \beta_i \geq 0$, $i = 1, \ldots, 18$. From Table 4, with $\delta = \delta_{\min}$, all of the estimated β_i satisfy this condition, but for five commodities (Alcohol, Clothing & footwear, Furniture and household goods, Public transport and Leisure goods & services) the estimated α_i are (insignificantly) negative. One approach would be to set these α_i to zero, with the interpretation that a consumer with total expenditure equal to the minimum total expenditure would be at a corner solution for these five commodities, and would only consume positive quantities at higher levels of total expenditure.

An alternative, and more appealing interpretation, would be to maintain these point estimates as the parameter values and to write[7]:

$$\alpha_i + \beta_i \log(Y_i/K) = 0 \tag{3.2}$$

and solve for the implied values of Y_i. For each commodity where the estimate of α_i is negative, the above equation determines that level of expenditure, Y_i, at which the estimated share (and hence expenditure) for commodity i is zero. The appropriate interpretation is then that for total expenditure less than Y_i expenditure on commodity i will be zero (*i.e.* a corner solution). For the five commodity groups with negative α_i the expenditure levels which act as the thresholds are given in Table 5. The ordering and magnitudes of these total expenditure thresholds triggering positive consumption on the commodity group seems to make sense.

Table 5: Total Expenditure Thresholds for Commodity Groups with $\alpha_i < 0$.

Commodity Group	Threshold Value
Alcohol	$88.39
Clothing & footwear	$87.12
Furniture and other household goods	$116.34
Public transport	$165.42
Leisure goods & services	$153.44

Our normalization of MAIDS also allows the following interpretation of the α_i and β_i parameters (see Cooper & McLaren (1992), p658). In the MAIDS model the shares, w_i, move monotonically from α_i for the "poor" to β_i for the "rich". Therefore, if $\alpha_i > \beta_i$, then commodity i is a necessity and if $\beta_i > \alpha_i$ then commodity i is a luxury. From Table 4, with $\delta = \delta_{\min}$, we see that our estimates tell us that commodity groups 1, 2, 3, 4, 5, 8, 11, 13, 16 and 17 are necessities. The other commodity groups are, according to the MAIDS model, luxuries. Overall, this classification of necessities

---

[7]Note that in the equation the Y_i and the α_i are not separately identified. The Y_i are introduced merely to interpret the α_i.

and luxuries makes sense. However, the classification of Cigarettes and tobacco as a necessity is, perhaps, a little surprising until we take account of the possible addiction factor for this commodity group.

We can also derive the estimated expenditure elasticities for the MAIDS model (using the estimates when $\delta = \delta_{\min}$). This elasticity is given by:

$$\frac{\partial w_i}{\partial \log(Y)} = \frac{\beta_i - w_i}{1 + \log(Y/K)}.$$
(3.3)

The values of the estimated elasticities, evaluated at the sample means of the w_i, and for the minimum, mean and maximum values of $\log(Y/K)$ are given in Table 6. These are all of a reasonable magnitude and are consistent with the discussion of the parameter estimates above.

Table 6: Estimated Expenditure Elasticities for Minimum, Mean and Maximum Values of $\log(Y/K)$.

	Commodity Group	Expenditure Elasticity		
		Minimum	Mean	Maximum
1	Bread, cereal & grain products	-.002971	-.001123	-.000936
2	Meat & fish	-.007330	-.002772	-.002309
3	Dairy, eggs & oil products	-.004769	-.001803	-.001502
4	Sugar, preserves and confectionery	-.001294	-.000489	-.000408
5	Fruit & vegetables	-.002814	-.001064	-.000887
6	Other food and non-alcoholic drinks	.003053	.001154	.000962
7	Alcohol	.004369	.001652	.001376
8	Cigarettes & tobacco	-.005773	-.002183	-.001819
9	Clothing & footwear	.005537	.002094	.001744
10	Housing (current)	.001771	.000670	.000558
11	Fuel (not including motor vehicle)	-.004234	-.001601	-.001334
12	Furniture and other household goods	.005904	.002232	.001860
13	Private transport	-.004212	-.001593	-.001327
14	Public transport	.001855	.000701	.000584
15	Leisure goods & services	.011752	.004444	.003703
16	Other goods	-.000399	-.000151	-.000126
17	Other services	-.000453	-.000171	-.000143
18	Privately purchased health	.0000006	.0000002	.0000002

These elasticities are estimated at the sample means for w_i.

One final issue to address in this section is the question of whether the zero replacement procedure has produced discernible patterns or outliers in the residuals from the estimation. Residual diagnostics in the form of Durbin-Watson statistics do

not indicate any patterns in the residuals[8]. However, a full residual diagnostic analysis looking for potential outliers has not been attempted at this juncture. It should be noted that, given that only 48 observations have no zero shares, the overwhelming majority of observations have been subject to the zero replacement procedure. As a result, it is our opinion that residual analysis to ascertain whether zero replaced observations are outliers is unlikely to be productive. We, therefore, rely on the robustness of the estimates to indicate the acceptability of the procedure.

4. Conclusions.

This paper has discussed and modified techniques for dealing with the problem of zero observations in budget share models when the stochastic specification is made using compositional data analysis methodology. Since the compositional data approach is attractive in economics, the new techniques of dealing with the problem of zeros are of interest. When the methodology is applied to Australian household expenditure data we find that the results indicate that the procedures are simple to use and yield results that are robust to the zero replacement used. Further, these results are consistent with *a priori* expectations. Given the success of the application of the methodology it will be interesting to see how it works with pooled time-series cross-section data.

---

[8]These are available on request.

5. REFERENCES.

Aitchison, J.A., (1986), *The Statistical Analysis of Compositional Data*, London: Chapman Hall.

Bacon-Shone, J., (1992), "Ranking Methods for Compositional Data", *Applied Statistics*, **41**, 533-537.

Cooper, R. J and K.R. McLaren, (1992), "An Empirically Oriented Demand System with Improved Regularity Properties", *Canadian Journal of Economics*, **25**, 652-667.

Fry, J.M., Fry, T.R.L. and K.R. McLaren, (1996), "The Stochastic Specification of Demand Share Equations: Restricting Budget Shares to the Unit Simplex", in press *Journal of Econometrics*.

McLaren, K.R., Fry, J.M. and T.R.L. Fry, (1995), "A Simple Nested Test of the Almost Ideal Demand System", *Empirical Economics*, **20**, 149-161.

Pudney, S.E., (1989), *Modelling Individual Choice: The Econometrics of Corners, Kinks and Holes*, Oxford: Basil Blackwell.

Rimmer, M. and A.A. Powell, (1994), "Engel Flexibility in Household Budget Studies: Non-parametric Evidence versus Standard Functional Forms", CoPS/IMPACT Paper OP-80, Monash University.