Expenditure Aggregation Techniques for Consumer Demand Systems: A Simulation Study[1]

by

Christopher J. Nicol

Department of Economics

University of Regina

Regina, Saskatchewan, S4S 0A2

CANADA

306-585-4182

Internet: nicolc@uregina.ca

FAX: 306-585-4815

March 19, 1997

[1]This research was supported in part by grants from the Social Sciences and Humanities Research Council of Canada, and the University of Regina President's Research Fund. Abble Chu, Sergio Gomez, Brent Langen and Jason Walters provided indispensable research assistance. Any errors are my responsibility.

I. Introduction

The availability of large cross-sections of microdata from household expenditure surveys around the world provides the opportunity to estimate more complicated models than was ever possible using aggregate, time-series data. Such models can include a wide range of household characteristics effects, many goods, effects on demand of labour force participation behaviour and nonlinearities in Engel function specifications.

Recently, the AI demand system of Deaton and Muellbauer (1980) has waned in popularity as a quadratic variant of this has been reported to be superior for some bodies of data. See, for example, Blundell, Pashardes and Weber (1993). A version of this model employed by Banks, Blundell and Lewbel (1994) is referred to as the IQAIDS by Lewbel (1996).

Although these large cross-sections of microdata present great opportunities for demand system estimation, there are associated difficulties. These difficulties were present with aggregate, time-series data, but of less pressing concern, owing to the many other problems which applied researchers faced in that earlier setting.

In principle, given extensive information on spending patterns contained in the microdata files, one might consider estimating as large a system of demand equations as possible. This would have the theoretical attraction that one could test many hypotheses in such a general framework, and model specification error would be less likely. However, although such large systems can certainly be specified, estimating them is a different matter. In particular, although there is now a lot of information on spending patterns, there are limitations on the quality and variability of price data, to facilitate estimation of these large systems.

The foregoing problem has always been around in demand estimation. One way in which it has been dealt with is to make assumptions that some goods are separable from others, included in the demand system. Alternatively, goods have been aggregated to yield more manageable systems for estimation. Unfortunately, where it has been possible to test separability assumptions, these are often rejected. See, for example, Derrick and Wolken (1985) and Diewert and Wales (1995).

An alternative to the assumption is to aggregate goods. Recently, Lewbel (1993, 1996) has proposed a generalisation of the Hicks (1936) - Leontief (1936) composite commodity theorem, which can be applied in situations where the Hicks-Leontief theorem does not hold exactly, as one would expect with empirical data. Lewbel (1996) shows that the assumptions required for the generalised composite commodity theorem to hold appear to be met for United States time-series price data, based on some *a priori* groupings of commodities. Lewbel (1996) also indicates, however, that if the assumptions for the GCCT do not hold for a particular set of data, then one could use methods proposed in Nicol (1991) to find appropriate groupings of goods.

In Nicol (1991), the requirements of a variation of Lewbel's (1996) GCCT, which appeared in an earlier paper by Lewbel (1993), were applied to Canadian microdata to obtain a grouping of goods which was less likely to influence the outcome of hypothesis tests.

The literature cited above is not the only time where these concerns have been analysed. Pudney (1981) proposed grouping goods on the basis of relationships amongst cross-price elasticities, obtained from demand system estimation. Unfortunately, this approach is sensitive to model specification error inducing incorrect groupings. A more general approach to detecting separable groups was proposed by Varian (1983), who bases the identification of these separable groups on revealed preference theory. This is a completely nonparametric approach, so is at the opposite extreme to Pudney (1981). However, in order to apply the Varian (1983) methodology, an initial grouping of goods must be chosen, as was the case in the Lewbel (1996) research. There is no guarantee that such an initial partition of the goods will have the desired properties, and therefore one might have to spend some time searching for a partition of the goods into groups which satisfies the requirements of the GCCT or GARP.

The issue of finding a suitable partition or grouping of goods thus has implications of a wide-ranging nature. If the suitable partition can be found, estimation of a demand system can proceed without the need to be concerned about separability issues. If one ignores the true separability structure, or aggregates goods incorrectly, there are many consequences. For one thing, hypothesis tests of almost any kind will be affected. (Nicol, 1991, provides an example of this). In addition, Derrick and Wolken (1985) show how even small errors in aggregation can lead to very large biases in estimated elasticities. Furthermore, since small errors in variables (of the type induced by incorrect aggregation) can lead to large differences in parameter estimates, any estimates of such magnitudes as equivalence scales, inequality measures and predictions can potentially also be seriously affected.

In Nicol (1991) a start was made in developing a methodology for identifying appropriate grouping patterns for goods based on a GCCT. However, this methodology was developed in the context of actual data. A more detailed analysis of these methods is required, in a controlled setting. This paper therefore uses simulation experiments to advance the understanding of methods which can be employed for grouping goods for aggregation, with a view to estimation of demand systems. This builds on the earlier research in Nicol (1991) by analysing how alternative grouping methods perform, when the nature of the underlying groupings in the data are known. The aim is to provide researchers with some simple yet powerful tools which can be used for preliminary data analysis, prior to estimation of demand systems.

The remainder of the paper is structured as follows. In Section II, the literature on the GCCT is reviewed, and related to the generation mechanism for the artificial data to be generated in the simulation experiments. Section III contains more detailed information on the experimental design and execution. Section IV provides some initial results, based on one grouping method. Section V summarises and concludes.

II. Stochastic Hicksian Aggregates

Suppose that a total of M goods are purchased by a household, but that it is desired to aggregate these goods into R groups, where $R < M$. Data on the "underlying" prices, $p_j, j = 1, \ldots, M$ are observed in the form of price indices, and the aggregated prices, $P_r, r = 1, \ldots R$, are constructed as geometric means of $p^r_{R_j}, j \in r$. Thus $\ln P_r = \sum_{j \in r, j=1}^{R_r} \lambda^r_{R_j} \ln p^r_{R_j}$, where R_r is the number of goods in the r'th group. Also, $\sum_{j \in r, j=1}^{R_r} \lambda^r_{R_j} = 1$, and $\lambda^r_{R_j}$ are weights reflecting the share of spending on good j in group r.

Lewbel (1988) defines SHA relative to the $\ln p^r_{R_j}$, as follows

$$(1) \qquad \ln p^r_{R_j} = \ln P_r + k^r_{R_j} + \mu^r_{R_j}, \qquad j \in r, \qquad r = 1, \ldots R$$

For notational simplicity, household subscripts are dropped from $\ln p^r_{R_j}, \ln P_r$ and $\mu^r_{R_j}$. It is assumed in (1) that $k^r_{R_j} = \mathrm{E}(\ln p^r_{R_j} - \ln P_r)$, and that $\mu^r = (\mu^r_{R_1}, \ldots, \mu^r_{R_r})^T \sim \mathrm{NI}(0, \Omega_r)$. Thus, $k^r_{R_j}$ are constant across a given price, $p^r_{R_j}$. This means that each household faces a constant component, $k^r_{r_j}$, as part of the price, $p^r_{R_j}$, the household faces. The $\mu^r_{R_j}$ are random components, and vary across households for every price.

To design the generation of SHA in a simulation experiment, use is made of (1) and the definition of $\ln P_r$. Substituting the latter in the former yields

$$
\begin{aligned}
\ln p^r_{R_j} &= \sum_{j\in r, j=1}^{R_r} \lambda^r_{R_j} \ln p^r_{R_j} + k^r_{R_j} + \mu^r_{R_j}, \\
\ln p^r_{R_j} - \sum_{j\in r, j=1}^{R_r} \lambda^r_{R_j} \ln p^r_{R_j} &= k^r_{R_j} + \mu^r_{R_j}, \\
(1-\lambda^r_{R_j}) \ln p^r_{R_j} - \sum_{j\neq i\in r, i=1}^{R_r} \lambda^r_{R_i} \ln p^r_{R_i} &= k^r_{R_j} + \mu^r_{R_j}, \qquad j\in r, \qquad r=1,\ldots R \qquad (2)
\end{aligned}
$$

The final equation above, (2), can be written in matrix form, in terms of all prices within the r'th SHA, as follows

$$\Lambda_r \ln p^r = k^r + \mu^r \qquad (3)$$

where $\ln p^r \equiv (\ln p^r_{R_1}, \ldots \ln p^r_{R_r})^T$, $k^r \equiv (k^r_{R_1}, \ldots k^r_{R_r})^T$ and $\mu^r \equiv (\mu^r_{R_1}, \ldots \mu^r_{R_r})^T$. The matrix, Λ_r is defined as follows

$$
\left\{
\begin{array}{ccccc}
(1-\lambda^r_{R_1}) & -\lambda^r_{R_2} & \cdots & \cdots & -\lambda^r_{R_r} \\
-\lambda^r_{R_1} & (1-\lambda^r_{R_2}) & -\lambda^r_{R_3} & \cdots & -\lambda^r_{R_r} \\
\vdots & -\lambda^r_{R_2} & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
-\lambda^r_{R_1} & -\lambda^r_{R_2} & \cdots & -\lambda^r_{R_{r-1}} & (1-\lambda^r_{R_r})
\end{array}
\right\} \qquad (4)
$$

Because $\lambda^r_{R_j}$ are weights which sum to one, the matrix Λ_r is singular by construction. Thus, (3) cannot be solved uniquely for the $\ln p^r$ vector by premultiplying by the inverse of Λ_r. However, if the last equation of (3) is dropped, we can rewrite the system as

$$\Lambda_{r-} \cdot \ln p^{r-} = k^{r-} + \mu^{r-} \qquad (5)$$

where Λ_{r-} has dimension $R_{r-1} \times R_{r-1}$, and the vectors $\ln p^{r-}, k^{r-}$ and μ^{r-} all have dimension $R_{r-1} \times 1$. The system, (5), can now be solved by pre-multiplying by Λ_{r-}^{-1}, to obtain the vector $\ln p^{r-} = \Lambda_{r-}^{-1}[k^{r-} + \mu^{r-}]$. Furthermore, the "last" element of $\ln p^r$ can be recovered by making use of (2) as follows

$$\begin{aligned} (1 - \lambda^r_{R_r}) \ln p^r_{R_r} - \sum_{i \in r, i=1}^{R_{r-1}} \lambda^r_{R_i} \ln p^r_{R_i} &= k^r_{R_r} + \mu^r_{R_r}, \\ \ln p^r_{R_r} &= (1 - \lambda^r_{R_r})^{-1} \{ \sum_{i \in r, i=1}^{R_{r-1}} \lambda^r_{R_i} \ln p^r_{R_i} + k^r_{R_r} + \mu^r_{R_r} \} \end{aligned} \tag{6}$$

The definition and solution method for SHA given above relates to price data in logarithmic form. However, it is equally likely that SHA, if they exist, are structured in terms of the price indices themselves. Thus, an alternative definition of SHA is as follows

$$p^r_{R_j} = P_r + \kappa^r_{R_j} + \tau^r_{R_j}, \qquad j \in r, \qquad r = 1, \ldots R \tag{7}$$

where $p^r_{R_j}$ and P_r are defined as before, $\kappa^r_{R_j}$ are the price specific constants such that $\mathrm{E}(p^r_{R_j} - P_r) = \kappa^r_{R_j}$, and $\tau^r_{R_j}$ are the stochastic components which vary across households. Clearly, an analogous solution method for the $p^r_{R_j}, j \in r$, is possible to that outlined above for the $\ln p^r_{R_j}, j \in r$.

Each of the $r = 1, \ldots R$ SHA can be defined with respect to the appropriate underlying $\ln p^r_{R_j}$ or $p^r_{R_j}$, constants, stochastic components and weights contained in the respective Λ_r matrices. Irrespective of which definition is used, it is clear that log prices or prices within an SHA are almost perfectly linearly related to each other, up to a random element. Therefore, any linear measure of association (such as simple correlation coefficients) would be able to identify which goods belonged to which SHA. In practice, things are a little more complicated than this. Economic data, by its nature, is highly correlated across time, due to business cycle considerations. Concurrently, however, relative prices do change over time, so there do exist goods which, while highly correlated, have relative prices diverge that over time. These are the sorts of goods which ought not to be grouped together to form composite goods. However, it is difficult in practice to disentangle the separate effect of across-group versus across-time price variation.

With the foregoing points in mind, log price data and price data are constructed firstly as "pure" SHA (SHAP). That is, the data are generated as indicated above, to yield R distinct SHA. These prices are then used to determine how well cluster analysis methods work in locating what are known to be the true SHA. These simulations constitute experiments 1 and 2, which will be described in detail in the next section. Experiments 1 and 2 thus serve as baseline measures against which the other two experiments (3 and 4) can be assessed.

Experiments 3 and 4 again relate to log prices and price data respectively. However, in these two experiments, recognition is made of a typical feature of economic data observed in practice: that all prices are highly correlated. With this in mind, the SHAP of experiments 1 and 2 are "mixed" with varying degrees of across-group price variability. Suppose the SHAP log price and price $(n \times M)$ matrices are defined as $\ln P(I)$ and $P(I)$ respectively. Also, suppose corresponding across-group-correlated log price and price matrices are defined as $\ln P(II)$ and $P(II)$ respectively. In experiment 3, linear combinations of $\ln P(I)$ and $\ln P(II)$ are formed as follows

$$\ln P(III.i) = \alpha_i \ln P(I) + (1 - \alpha_i) \ln P(II), \qquad \alpha_i \in (0,1) \tag{8}$$

Experiment 4 uses an analogous method to construct a sequence of prices, $P(III.i)$. Thus as α_i ranges from $1 \rightarrow 0$, an increased amount of "noise" or across-group price variability is introduced into the SHAP, $\ln P(I)$ and $P(I)$, to form "impure" SHA (SHAI), $\ln P(III.i)$ and $P(III.i)$. The potential for the cluster analysis methods to locate correctly the actual SHA contained in the data can be then observed.

III. Experimental Design and Execution

III.1 Generation of SHAP

The first step in the design of the experiments is to choose the number of disaggregated expenditure groups for which prices would typically be required. That is, M, in the notation of the previous section. Also required is the number of aggregates which the disaggregated prices would be grouped into, R in the notation of Section II.

From a practical view, in terms of price data reported by government statistical offices, the largest feasible system estimable is usually about $M = 20$. This would be the case, for example, if one were to make use of the Canadian Family Expenditure Survey public-use database, in conjunction with matching, available price data. A system as large as 20 equations is not usually feasible to estimate, for a variety of reasons, hence the need to aggregate the goods into a smaller number of groups. For this study, $R = 4$ is chosen. This is a convenient choice in light of some recent studies[2], and also since goods are sometimes classified as durables, semi-durables, nondurables and services– four groups.

Having selected M and R, the next step is to determine how to generate the required data. To simulate the type of price variability exhibited by SHAP, one suspects that the source of variability for these data would be similar to that shown by inter-regional price differentials. Using series of inter-regional price indices from *Consumer Prices and Price*

[2] See, for example, Browning and Meghir (1991), Nicol (1991b) and Buse (1992).

Indexes[3] for major Canadian cities and twenty goods, a sequence of four covariance matrices were constructed, one for each SHAP. To construct these covariance matrices, use was made of the raw price data and the definitions of SHAP. That is, goods were grouped into four categories on the basis of what would often be regarded as separable groups. Thus, group 1 contained eight categories; group 2, three; group 3, five; and group 4, four. The goods falling into each of the four broad categories are given in Table 1, along with the broad category definitions: A, "food and clothing"; B "shelter"; C "transportation"; and D "services".

While the raw data were used as a basis for the construction of appropriate covariance matrices, this was done in an indirect way, to guarantee that the generated prices would have the SHAP properties. For each of the four groups, "actual" $\ln P_r$[4] were constructed using as weights in the respective Λ_r matrices, the share of spending on the individual goods as proportions of total spending for the group. Given respective series of "actual" $\ln p^r_{R_j}$[5] and $\ln P_r$, a sequence of twenty series of differences, $\ln p^r_{R_j} - \ln P_r$. were calculated. Using these deviations, covariance matrices were then constructed, one for each group. These covariance matrices were the basis for generation of the stochastic components of the SHAP. That is, each covariance matrix was used in turn to generate 1000 "observations" for each of the deviation vectors[6]. This generation procedure therefore yielded the experimental counterparts, of $\mu^r_{R_j}$ and $\tau^r_{R_j}, j \in r, \forall r = 1, \ldots R$. To obtain the experimental counterparts of $k^r_{R_j}$ and $\kappa^r_{R_j}$, the means of the twenty series of differences, $\ln p^r_{R_j} - \ln P_r$. were calcu-

[3] Statistics Canada Catalogue No. 62–010.

[4] Or P_r, depending on the experiment.

[5] Or $p_{R^r_j}$, again, depending on the experiment.

[6] All random number generation used the IMSL, Version 1.0, April 1987 subroutine, DRNMVN.

lated. Thus, in conjunction with Λ_{r-}, $\ln p^{r-}$ and p^{r-} can be obtained using the respective relationships

$$\ln p^{r-} = \Lambda_{r-}^{-1} \cdot \{k^{r-} + \mu^{r-}\} \tag{9}$$

and

$$p^{r-} = \Lambda_{r-}^{-1} \cdot \{\kappa^{r-} + \tau^{r-}\} \tag{10}$$

Then, $\ln p_{R_r}^r$ and $p_{R_r}^r$ can be recovered by making use of (6).

Using the foregoing approach, a sequence of 5000 replications of 1000×20 matrices of $\ln p_j$ and p_j were generated. These were then used in experiments 1 and 2 to determine how well cluster analysis methods could perform in locating the four SHAP. Details of the cluster analysis methods used will be discussed in Section III.3 below.

III.2 Generation of SHAI

The generation of SHAI require the SHAP discussed in the previous sub-section, as well as the $\ln P(II)$ and $P(II)$ matrices discussed earlier. To generate the $\ln P(II)$ and $P(II)$, the covariance matrices of time-series price data were used. Time-series of price data have the property which economic data exhibits in general, in that all series are all highly correlated. Using expenditure categories which most closely corresponded to those on which the SHAP were based, twenty series of prices were extracted from *National Income and Expenditure Accounts, Annual Estimates, 1926–1986*[7]. The covariance matrices of $\ln p_j$ and p_j were used to generate the stochastic components of the vectors of $\ln P(II)$ and $P(II)$ respectively. To these were added the respective sample means of each of the price vectors, as calculated on

[7]Statistics Canada Catalogue No. 13–531 (occasional).

the basis of the actual price series. Again, 5000 replications of 1000×20 log price and price matrices were generated.

Given the SHAP, $\ln P(I)$ and $P(I)$, and the inter-group-correlated prices, $\ln P(II)$ and $P(II)$, the aggregates referred to in (8), $\ln P(III.i)$ and $P(III.i)$ can be generated, which are of course the SHAI. Five different sequences of these were generated, based on each of the log price and price data. These were calculated on the basis of $\alpha = (\alpha_1, \ldots \alpha_5) \equiv (0.9, 0.75, 0.5, 0.25, 0.1)$. Thus, an increasing amount of noise is introduced as one moves from $\ln P(III.1)$ (or $P(III.1)$), where $\alpha_1 = 0.9$, to $\ln P(III.5)$ (or $P(III.5)$), where $\alpha_5 = 0.1$.

III.3 Cluster Analysis Experiments

In all four experiments conducted in this simulation study, the "complete linkage" hierarchical agglomerative cluster analysis method is used. This method is the most appropriate for this sort of application, since the clusters (or aggregates) formed contain prices which are no more than a given "distance" apart. All prices outside a given aggregate are therefore "further" from each price within the aggregate than any price within the aggregate. A summary of the most widely used cluster analysis methods is contained in Aldenderfer and Blashfield (1984), where the complete linkage, hierarchical agglomerative method is also reviewed.

In empirical studies which use cluster analysis with actual data, it is necessary to specify a "stopping rule" where clustering ceases. This stopping rule is triggered in the cluster step immediately prior to a "large" change in the "fusion coefficient"[8]. In many empirical studies, stopping rules are chosen on an *ad hoc* basis. See Mojena (1977) for a brief evaluation. In

[8]Fusion coefficients measure the closeness of two "objects": in this case, prices. They are based on the

Nicol (1991a), however, the stopping rule had a statistical basis. That is, given the nature of the similarity measures used, a large change in the fusion coefficient could be interpreted as a statistically significant change. Thus, clustering ceased immediately before a cluster step which would require the merging of two statistically significantly different objects (prices/goods).

In these simulations, the concern is with whether cluster analysis can locate known SHA from within a larger array of price data. This is a different question from whether cluster analysis can, for data with the quality of prices, locate the *correct number* of clusters or aggregates. This latter subject is the topic of ongoing research, and the subject of another paper. Consequently, for present purposes, a stopping rule is not necessary, since it is already known at what point to cease clustering: when four aggregates or clusters have been identified.

One topic which must be considered further, however, is the choice of similarity or distance measure. Since SHAP are almost perfectly linearly related, they should be highly positively correlated. An obvious measure of similarity to use is therefore the matrix of correlation coefficients of prices. Thus, in the complete linkage cluster analysis method, prices are grouped according to high (positive), within-group correlation coefficients. In this case, the fusion coefficients are correlation coefficients.

While correlation coefficients might be expected to work well in locating goods which belong to the correct SHA when dealing with SHAP, they are less likely to work well when using SHAI. That is, the greater the noise or inter-group price correlation in the data, the

measure of similarity (or distance) used. The notions of similarity/distance and their relationship to fusion coefficients will be discussed more fully below.

more difficult it will be to locate the correct prices for each of the known SHA using the correlation matrix as a measure of similarity. In other words, as inter-group price correlation rises with $\alpha_i \to 0$, the more highly correlated are *all* prices, as well as the within SHA prices.

An alternative measure of similarity could thus be based more directly on the original notion on which SHA were proposed. What is really desired is a grouping scheme which identifies those prices which are relatively constant over the sample, up to a random element. Such a similarity measure was proposed in Nicol (1991a) and has the convenient property that its elements have an asymptotically normal distribution. In particular, consider two prices, p_{jt} and $p_{kt}, j \neq k = 1, \ldots M, t = 1, \ldots, n$. Define the relative price at observation t as $\phi_{jkt} = p_{jt}/p_{kt}$. If ϕ_{jkt} is constant over all n observations, then it can be denoted merely as ϕ_{jk}. Then p_{jt} and p_{kt} are true Hicksian aggregates. However, since all prices contain a random element, ϕ_{jkt} are unlikely to be observed to be constant. On the other hand, if the random elements in prices are truly random and the two prices are SHA, then *on average* ϕ_{jkt} should be constant.

If two prices are, on average, constant, then estimates of their sample means constructed using different samples of prices should not be statistically significantly different. Using the generated data for experiments 2 and 4, each sample of 1000 observations on each price is split into two sub-samples, and alternative estimates of each sample mean, $\bar{\phi}_{jk,1}$ and $\bar{\phi}_{jk,2}$ are calculated. A constant relative price can be represented by the null hypothesis, $\mathrm{H}_0 : \phi_{jk,1} = \phi_{jk,2}$. When this null holds (for all $j \neq k$), then

$$Z_{jk} = \frac{(\bar{\phi}_{jk,1} - \bar{\phi}_{jk,2})}{(S_p^2/n_1 + S_p^2/n_2)^{0.5}} \overset{A}{\sim} N(0,1) \tag{11}$$

where $\bar{\phi}_{jk,\ell} = \frac{1}{n_\ell} \sum_t^{n_\ell} \phi_{jkt},\ \ \ell = 1,2;\ \ S_p^2 = \frac{(n_1-1)\cdot S_{jk,1}^2 + (n_2-1)\cdot S_{jk,2}^2}{(n_1+n_2-2)};\ \ S_{jk,\ell}^2 = \frac{1}{(n_\ell-1)} \sum_t^{n_\ell} (\phi_{jkt} -$

$\bar{\phi}_{jk,\ell})^2$; and $n_\ell, \ell = 1, 2$ are the two sub-sample sizes, both 500 in this application. The Z statistic in (11) is the standard test statistic for significant differences in population means (here, the difference is zero, under H_0) when the population variances, $\sigma^2_{jk,\ell}, \ell = 1, 2$, are identical. Large positive and negative deviations of Z_{jk} from zero are viewed symmetrically as reflecting statistically significant deviations from constancy of ϕ_{jkt}. A cluster analysis technique which uses the Z_{jk} as distance measures will thus group goods with relatively low Z_{jk}. A corresponding similarity measure can be constructed for experiments 1 and 3, with all the above notation redefined in terms of $\ln P(I)$ and $\ln P(III.i), i = 1, \ldots, 5$.

On the basis of the price data defined in sub-sections (III.1) and (III.2), and the similarity measures defined in this sub-section, the four cluster analysis experiments were conducted with each of the four types of price aggregates: with SHAP based on log prices and prices; and with SHAI based on log prices and prices[9]. Each of the two alternative similarity measures were employed in each of the four experiments. As mentioned previously, it is to be expected that the correlation matrices will work well for SHAP, but less well for SHAI as α_i declines. Some evidence for this *a priori* speculation is provided by Nicol (1991a), who only reported the results of clustering with the relative price distance measure. This was because that measure seemed to be superior at identifying aggregates which had least effect on the outcome of hypothesis tests, relative to the correlation matrices. Since Nicol (1991a) was using actual data with a high degree of across-group price variability, the chances of the correlation matrix correctly identifying SHA would be reduced. It will therefore be of some interest to analyse the results of this simulation experiment to see if a similar finding can be replicated in a controlled environment.

[9] The IMSL, Version 1.0, 1987 subroutines, DCLINK and CNUMB were used.

IV. Simulation Experiment Results

To assess the relative performance of clustering with different similarity measures and/or amounts of noise introduced into the SHAP, a statistic introduced by Rand (1971) is employed. This statistic can also be used to determine whether two alternative clustering procedures are statistically significantly different from each other, even when the true number of clusters and their "members" are unknown. In this study, however, the exact number of clusters is known, as are the memberships of the clusters. Consequently, the summary statistic can be used to indicate the number of "correct" versus "incorrect" linkages made in a given cluster procedure. In addition, the statistics so generated can be used to test for statistically significant differences in performance across methods and price definitions.

When clustering with M objects (prices) which belong to known groups, each object is either linked or not linked with every other object. If every object is assigned to the correct group, and no objects are linked with an object outside the true group, then the clustering algorithm will have been 100% successful. However, in practice, it is likely that some objects are linked with objects outside their true groups, and not linked with objects which are in their group. There are therefore a total of $\sum_{I=1}^{M-1} I = M(M-1)/2$ pairwise comparisons which can be made between each object and every other object, as to whether it has been grouped with the correct group members. Thus if a value of 1 is assigned to correct linkages/non-linkages and 0 to incorrect linkages/non-linkages, the Rand (1971) statistic defining the proportion correct can be defined as follows

$$\mathcal{C} \;=\; \sum_{i<j}^{M} \gamma_{ij}/[M(M-1)/2], \qquad \in \qquad [0,1] \tag{12}$$

where

$$\gamma_{ij} = \begin{cases} 1 \text{ if } X_i \text{ and } X_j \text{ are correctly grouped} \\ 1 \text{ if } X_i \text{ and } X_j \text{ are not grouped, correctly} \\ 0 \text{ otherwise} \end{cases}$$

As well as clustering with each type of generated data using the correlation matrices and relative price measures of similarity, clustering was also carried out using the similarity measure calculated: on the basis of *prices* for experiments 1 and 3; and on the basis of *log prices* for experiments 2 and 4. This facilitates a comparison of the relative performance of the clustering methods when implemented based on the "wrong" similarity measure metric for a given type of data. In practice, the true basis of the SHA definition (in terms of log prices or prices) is not known, so it is useful to obtain some guidance on how robust the clustering methods are to the metric of the similarity measure.

The statistic in (12) was calculated for each of the different types of prices generated, when clustered either with the correlation matrix measure of similarity, or the relative price measure. Summary statistics for each experiment are contained in Table 2 for experiments 1 and 3; and in Table 3 for experiments 2 and 4. Some notational shorthand is useful. Denote clustering with the correlation matrix based on log prices and prices by LPC and PC respectively. Denote clustering with the relative price measure of similarity based on log prices and prices by LPT and PT respectively.

Recall that experiments 1 and 3 have prices generated based on the log price definition of SHA. Table 2 shows that the correlation matrix can identify the SHA with 100% accuracy when the prices are SHAP. This is the case whether the correlation matrix is calculated using log prices or prices. All other statistics in Table 2 are statistically significantly different from 1. This indicates less than perfect performance when the relative price measure of similarity is used with SHAP, and when either similarity measure is used if noise is introduced into the SHA. This holds whether the similarity measures are calculated using log prices (correctly), or prices (incorrectly).

In experiments 2 and 4 prices are generated based on the price definition of SHA. Table 3 shows that the correlation matrix performs less than perfectly in identifying the SHA when they are SHAP. This is the case whether the correlation matrix is calculated using prices (correctly) or log prices (incorrectly). This contrasts with the results in Table 2. Therefore, true SHA cannot be identified as easily when they are defined in terms of prices, using the methods applied here. However, an approximately 96% success rate is still a major achievement. All other statistics in Table 3 are statistically significantly different from 1 by a much larger margin than that seen for the SHAP. However, the deterioration in performance of the correlation matrix as a similarity measure is not as rapid when noise is introduced into the SHAP as was the case in experiments 1 and 3. This observation is based on a cursory comparison of the relative values of comparative statistics in Tables 2 versus 3.

A more formal comparison of the performance of the different similarity measures, different metrics for calculation of the similarity measures and different amounts of noise in the SHAI is contained in Table 4 for experiments 1 and 3. For SHAP: PC and LPC are identical; both dominate PT; and PT dominates LPT. This last observation is interesting, because the SHAP are defined on the basis of log prices, yet LPT performs worst by a statistically significant margin.

The corresponding comparisons for SHAI, experiment 3, are also contained in Table 4. As an increasing amount of noise is added to the SHAP, the performances of PC and

LPC deteriorate sharply. LPC dominates by a statistically significant margin only up to $\alpha_i = 0.75$. Thereafter, LPC is dominated everywhere by PT and LPT. The most surprising aspect of these results is the performance of PT. This becomes the best method at $\alpha_i = 0.5$ and beyond, yet the SHAP are based on the *log price* definition for experiment 3. In general, however, the *a priori* speculation that the relative price similarity measure would perform better the more noise there is in the SHA, is borne out.

Comparisons of the performances of experiments 2 and 4 are contained in Table 5. Here, although the performances of PC and LPC are not perfect when using SHAP (unlike experiment 1), these two methods appear to perform relatively better than for experiment 3, at higher levels of noise in the data[10]. A word of caution is necessary here, however. In generating the $P(II)$ prices for experiment 4, it was necessary to shrink the variances and covariances of the data design covariance matrix to avoid "negative" prices. This meant that the relative coefficient of variation between $P(I)$ and $P(II)$ was about 1.01. For experiment 3, however, this adjustment was not required, which meant that the relative coefficient of variation between $\ln P(I)$ and $\ln P(II)$ was about 0.27[11]. Consequently, a given value of α_i introduces more inter-group noise for experiment 3 than for 4. Taking this difference into account would mean that the performance of PC and LPC deteriorate at about the same rate in experiment 4 as in experiment 3.

One result which corresponds with experiment 3, therefore, is that the relative price similarity measure is better than the correlation similarity measure the more noise the data

---

[10] That is, for lower values of α_i.

[11] For the actual data used in the data generation design, the corresponding relative coefficients of variation were about 0.18 for log price data, and 0.14 for price data.

contain. However, while PC and LPC fall below both PT and LPT beyond $\alpha_i = 0.5$, LPT is always better than PT beyond $\alpha_i = 0.90$, and LPT is best with noisy data. This is curious since the SHAP are based on the *price* definition of SHA. In other words, clustering with PT is best with noisy data when SHA are defined in terms of *log prices*, whereas clustering with LPT is best with noisy data when SHA are defined in terms of *prices*.

To summarise, there are three main findings. Firstly, SHAP can be identified almost perfectly, irrespective of whether they are defined in terms of log prices or prices, when a correlation matrix is used as similarity measure. Secondly, when noise is introduced into the SHAP, the performance of the relative price measure of similarity increases dramatically when SHA are defined in terms of log prices, but calculating the measure based on *prices* is the best method of all. Finally, when SHAP are defined in terms of prices, the relative price measure of similarity based on *log prices* outperforms all other methods, when clustering relatively noisy SHAI.

V. Summary and Conclusions

Nicol (1991a) showed that the outcome of hypothesis tests and other estimation results could be affected by "inappropriate" aggregation of goods in demand systems. This finding highlights the need for a means of identifying appropriate groupings of goods: SHA, in the terminology of Lewbel (1988).

This paper has explored the relative performance of complete linkage, hierarchical agglomerative cluster analysis in identifying known SHA in sets of generated price data. Two measures of similarity were used: one based on different representations of correlation of the experimental data; the other based on different representations of relative price variation.

The results indicate that when SHA are contained in a set of price data, and when this is the only systematic source of price variability, correlation measures of similarity can locate the SHA with almost 100% precision. Unfortunately, economic data are not characterised by such a well-behaved structure. When inter-group price correlation is introduced into SHA, making the price data more "noisy", the ability of measures of correlation to identify the SHA rapidly deteriorates. However, an alternative measure of similarity based on the degree of relative price variability in the data performs statistically significantly better than the correlation measure. This performance does not seem to be grossly affected by increasing amounts of inter-group correlation of prices. Also, the relative price similarity measure should be constructed using *prices* when SHA are defined in terms of log prices; but the measure should be calculated using *log prices* when SHA are defined in terms of prices.

These results therefore indicate that if it is desired to group goods in a demand system on the basis of a SHA definition, the cluster analysis method used here with the relative price measure of similarity can be very effective at locating SHA structure in the data. However, some caveats are in order. First of all, this study does not look at the issue of whether the correct number of clusters can be easily identified. The study was conducted on the basis that a known number of clusters existed in the generated data, and cluster analysis was used to try to find them. Also, while Nicol (1991a) provides some evidence that hypothesis test results can be affected by "incorrect" grouping, this issue has yet to be examined in a controlled setting with the sort of generated price data used in this paper. These two subjects are the focus of ongoing research.