07 September 2005

A tale of two logits, compositional data analysis and zero observations*

Tim R. L. Fry and Derek Chong

School of Economics, Finance & Marketing
Royal Melbourne Institute of Technology
GPO Box 2476V, Melbourne, Victoria 3001
Australia

Abstract: The application of compositional data analysis through log ratio transformations corresponds to a multinomial logit model for the shares themselves. This model is characterized by the property of Independence of Irrelevant Alternatives (IIA). IIA states that the odds ratio – in this case the ratio of shares – is invariant to the addition or deletion of outcomes to the problem. It is exactly this invariance of the ratio that underlies the commonly used zero replacement procedure in compositional data analysis. In this paper we investigate using the nested logit model that does not embody IIA and an associated zero replacement procedure and compare its performance with that of the more usual approach of using the multinomial logit model. Our comparisons exploit a data set that combines voting data by electoral division with corresponding census data for each division for the 2001 Federal election in Australia.

J.E.L. Classification: C310, D720, M310
Keywords: compositional data, zero observations, vote shares, brand shares
Corresponding Author: tim.fry@rmit.edu.au

* The research reported in this paper was supported by Australian Research Council grant DP0449846: "Economic reform and Australian electoral decision making" and by the research program in "Market mix modeling" funded by Starcom Worldwide Pty Ltd. The views expressed, however, are those of the authors.

1. Introduction.

The application of compositional data analysis through log ratio transformations corresponds to a multinomial logit model for the shares themselves. This model is characterized by the property of Independence of Irrelevant Alternatives (IIA). IIA states that the odds ratio – in this case the ratio of shares – is invariant to the addition or deletion of outcomes to the problem. It is exactly this invariance of the ratio that underlies the commonly used zero replacement procedure in compositional data analysis. In this paper we investigate using the nested logit model that does not embody IIA and an associated zero replacement procedure and compare its performance with that of the more usual approach of using the multinomial logit model.

The plan of the rest of this paper is as follows. The next section describes the compositional data approach used by statisticians to model share data. Section two extends this approach to regression modeling of share data and discusses two key specifications that can used – the multinomial logit (MNL) and nested logit (NL). The issues that arise in modeling share data with zero observations are discussed in section three and zero replacement procedures for MNL and NL specifications are presented. Section four applies the MNL and NL specifications along with the associated zero replacement techniques to a data set that combines voting data by electoral division with corresponding census data for each division for the 2001 Federal election in Australia. Finally, section five contains some concluding remarks.

2. Compositional Data Analysis.

The restriction of shares to the unit simplex has been recognized by researchers in many fields (see *inter alia* Aitchison (1986), Barceló *et al* (1996), Fry *et al* (1996), Howel (1994) and McLaren *et al* (1995)). In particular, this restriction causes problems for traditional multivariate statistical methods which are based upon the Normal distribution. It is, however, possible to develop a framework for the statistical analysis of data on shares. Such techniques are termed compositional data analysis, hereafter CODA, (Aitchison (1986)). The advantage of CODA techniques is that they provide a unifying set of distributional assumptions which allow for the use of traditional multivariate statistical methods.

In the statistical literature a *composition* consists of M *parts*. The parts are labels which identify the components into which a total has been sub-divided (*e.g.* the parts are brands and the total is total market volume sales). The *components* are the numerical proportions in which the parts appear (*i.e.* the shares). A composition is defined by taking the elements of a *basis* (*e.g.* individual brand volume sales) and dividing them by the *size* of the basis (*e.g.* total market volume sales). This operation takes elements defined as non-negative and constrains them to lie between zero and one and to sum to one (*i.e.* to lie on the unit simplex, S^{M-1}). It should be noted that this unit sum constraint reduces the dimension of the space on which the vector of components (shares) is defined to $M-1$. The major obstacle to the statistical analysis of compositional data is that the restriction to the unit simplex necessarily leads to the lack of an interpretable covariance structure and, as a result, the multivariate Normal distribution is inappropriate.

In order to apply statistical analysis techniques based upon the Normal distribution a one-to-one transformation is required to map the data on shares to data suitable for analysis using multivariate Normal based techniques. That is we need to map from the unit simplex, S^{M-1}, to R^{M-1} and produce an interpretable covariance structure. One such transformation is the additive log-ratio (ALR) transform:

$$y_i = \ln\left(\frac{s_i}{s_M}\right),\ i = 1, \ldots,\ M-1$$

with an associated Jacobian given by $jac(\mathbf{y} \mid \mathbf{s}) = (s_1 \ldots s_M)^{-1}$.

The inverse transformation, R^{M-1} to S^{M-1}, is the additive logistic transform and reconstructs the components as:

$$\begin{aligned} s_i &= \frac{\exp(y_i)}{1 + \exp(y_1) + \ldots + \exp(y_{M-1})},\quad i = 1, \ldots, M-1, \\ s_M &= \frac{1}{1 + \exp(y_1) + \ldots + \exp(y_{M-1})} \\ &= 1 - s_1 - \ldots - s_{M-1}. \end{aligned}$$

These transformations form the heart of CODA techniques. To model compositional data we apply the ALR transform to produce log-ratio data and then apply traditional multivariate statistical techniques (*e.g.* multivariate regression) to the transformed data. To return to the composition we simply apply the inverse transform, the additive logistic.

A major benefit of this approach is that it is straightforward to derive the associated distribution theory (Aitchison (1986) pp. 115-119) for the random variables. In particular, if the log-ratio vector $\mathbf{y}$ has an $M-1$ dimension Normal distribution, $N(\boldsymbol{\mu},\,\boldsymbol{\Sigma})$, then the composition, $\mathbf{s}$, (the vector of shares) will follow an additive logistic normal distribution, $L(\boldsymbol{\mu},\,\boldsymbol{\Sigma})$, defined on the unit simplex. The additive logistic normal distribution is particularly attractive in that, like the Normal distribution, it is capable of capturing the wide range of covariance structures encountered in observed data. Additionally within the CODA framework it can be shown that the basis $\mathbf{q}$ (*e.g.* the vector of brand volume sales) will follow a multivariate log-Normal distribution.

Before discussing the application of CODA techniques to regression models for share data some additional points need to be made. Firstly, the use of s_M as the denominator in the ALR transform is, at first, unusual in that the parts of the composition are treated asymmetrically. It is important, however, to note that reordering the parts and changing the component used as the denominator in the transform makes no difference to any statistical procedures. Thus all statistical procedures are invariant to the choice of the component used as the denominator. Secondly, the ALR is not the only transform that could be used. In particular, a centered log-ratio transform could be used and this centered version is related to the approach currently undertaken in the stochastic specification of attraction models (see *inter alia* Cooper (1993), Cooper and Nakanihi (1988) and Ghosh *et al* (1984)) and in the estimation of Logit and Addilog models in economics (see *inter alia* Bewley (1982a), (1982b), (1986) sand Chavas and Segerson (1986)). The centered log-ratio transform is $s_i^* = \ln(s_i/\tilde{s})$, where $\tilde{s}$ is the geometric mean of the M shares. Indeed, there is a one-to-one mapping between the centered log-ratio form and our preferred log-ratio approach (ALR) and so the two approaches are identical (see Aitchison (1986) and McLaren *et al* (1995)). The mapping is $\mathbf{s}^*= \mathbf{Gy} = \mathbf{F}'(\mathbf{FF}')^{-1}\mathbf{y}$ and $\mathbf{y} = \mathbf{Fs}^*$, where $\mathbf{y}$ is the vector of log-ratios and $\mathbf{s}^*$ is the vector of centered log-ratios and $\mathbf{F}$ has the general form:

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ 0 & 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

Since the parameter estimates obtained via maximum likelihood are invariant to the form of the transformation used the choice of one transformation over another is purely a matter of convenience. It is our opinion that on grounds of distributional assumptions and computational simplicity the additive log-ratio transform is preferable.

Finally, we note that often we are interested in modeling not just the shares but also the movements in the total. For example, we are interested in both the vote shares and the total turnout in an election or in both the brand shares and the total market sales. In other words, we wish to jointly model the composition and the size of the basis. A further advantage of CODA techniques is that they can be used to specify models for the joint modeling of data on shares and the size of the basis (for full details see Aitchison (1986) Chapter 9.2, 9.4). Essentially, the share data is transformed using the additive logistic transformation and the size (e.g. total sales) data is also transformed (to *log(total sales)*). The resultant transformed data is then modeled using an M dimensional multivariate regression model. The first $M - 1$ equations in the model concern the vector of log-ratios, $\mathbf{y}$, and the last equation concerns *log(total sales).*

3. Regression modeling in CODA

A direct application of the CODA approach would involve modeling the log-ratio transformed data, $\mathbf{y}$, in terms of $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. In particular, we may parameterize the mean, $\boldsymbol{\mu}$, to depend upon a set of variables, $\mathbf{Z}$ and a set of parameters, $\boldsymbol{\theta}$, according to a multivariate regression model:

$$y_i = \ln\left(\frac{s_i}{s_M}\right) = \mu_i(\mathbf{Z}, \boldsymbol{\theta}) + u_i,$$

where $\mathbf{u} = [u_i]$ is a stochastic term which is distributed as multivariate Normal $(\mathbf{0}, \boldsymbol{\Sigma})$. The advantage of this model is that, within this framework, the shares are distributed as additive logistic normal and the basis as multivariate log-Normal. The remaining issue is the specification of the functional form for the $\mu_i(\mathbf{Z}, \boldsymbol{\theta})$. By analogy with the arguments in Fry *et al* (1996), the parameterization chosen should retain any parameter interpretations from the underlying theory and, further, it should retain the logical consistency argument that shares from the model are restricted to the unit simplex. Such a parameterization is given by:

$$y_i = \ln\left(\frac{S_i(\mathbf{Z},\boldsymbol{\theta})}{S_M(\mathbf{Z},\boldsymbol{\theta})}\right) + u_i$$

where $S_i(\mathbf{Z},\boldsymbol{\theta})$ is the theoretical specification for the share of i which retains the logical consistency requirement.

We will consider two particular choices for the theoretical specification $S_i(\mathbf{Z},\boldsymbol{\theta})$. The multinomial logit (MNL) and the nested logit (NL). The first of these, the MNL, is very common in applied work. The MNL specification has:

$$S_i(\mathbf{Z},\boldsymbol{\theta}) = \frac{\exp(\mathbf{Z}^{'}\boldsymbol{\theta}_i)}{\sum_{j=1}^{M}\exp(\mathbf{Z}^{'}\boldsymbol{\theta}_j)}$$

A simple justification often used for this specification is that the share is a function of the relative "attractiveness"of i:

$$S_i(\mathbf{Z},\boldsymbol{\theta}) = \frac{A_i}{\sum_{j=1}^{M} A_j} = \frac{A_i(\mathbf{Z},\boldsymbol{\theta})}{\sum_{j=1}^{M} A_j(\mathbf{Z},\boldsymbol{\theta})},$$

with $A_i(\mathbf{Z},\boldsymbol{\theta}) = \exp(\mathbf{Z}^{'}\boldsymbol{\theta}_i)$. The estimating equations from this model specification are given by:

$$y_i = \ln(A_i(\mathbf{Z},\boldsymbol{\theta})) - \ln(A_M(\mathbf{Z},\boldsymbol{\theta})) + u_i = \mathbf{Z}^{'}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_M) + u_i.$$

The fact that there are identification issues involved in the use of share equations of the form:

$$S_i(\mathbf{Z}, \boldsymbol{\theta}) = \frac{A_i(\mathbf{Z}, \boldsymbol{\theta})}{\sum_{j=1}^{M} A_j(\mathbf{Z}, \boldsymbol{\theta})},$$

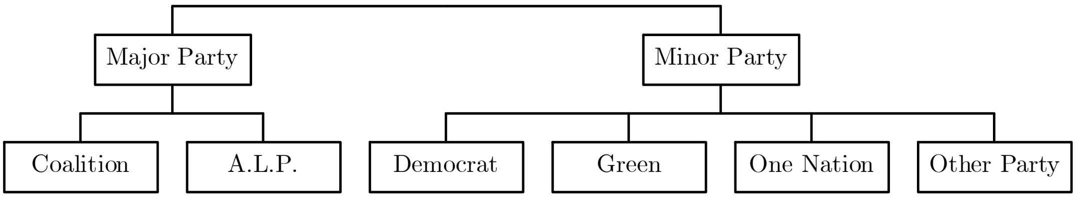
has long been recognized (Theil (1969)). In particular, if we re-scale by multiplying by an arbitrary, non-zero, constant, say, $\exp(a)$, we find:

$$S_i(\mathbf{Z}, \boldsymbol{\theta}) = \frac{\exp(a) A_i(\mathbf{Z}, \boldsymbol{\theta})}{\sum_{j=1}^{M} \exp(a) A_j(\mathbf{Z}, \boldsymbol{\theta})} = \frac{A_i(\mathbf{Z}, \boldsymbol{\theta})}{\sum_{j=1}^{M} A_j(\mathbf{Z}, \boldsymbol{\theta})}.$$

As a result, a normalizing restriction will be required to identify the model. For the MNL model the normalization used is to set $\boldsymbol{\theta}_M = \mathbf{0}$. This yields a simple multivariate linear regression specification for the y_i.

The second specification for $S_i(\mathbf{Z}, \boldsymbol{\theta})$ that we consider in this paper is the nested logit (NL) model. This model was introduced in the context of discrete choice modeling by McFadden (1978). However, Bechtel (1990) uses the NL model in the context of market shares and utilizes compositional data analysis techniques to facilitate estimation of the NL model with share data. The NL model recognizes the fact that often there is additional structure in a problem that can be exploited in the specification. For example, brands in fast moving consumer goods markets may belong to particular segments (e.g. standard, premium and economy) of the total market. Political parties can be categorized as major or minor parties (or left and right wing). Such a situation is represented in Figure 3.1 for Federal Elections in Australia. There are two major political parties – the Coalition formed by the Liberal Party and National Party and the Australian Labor Party (ALP). Four minor parties also contest in the election – Australian Democrats, Australian Greens, the One Nation Party and "Other Party" comprising of independents and other small groupings. The contest for a given electorate can be viewed as a contest between the major parties and the minor parties and within the major or minor groups between the political parties that comprise the group.

Figure 3.1: Example Nested Logit



A basic (two-level) NL model can be characterized as follows (see Train (2003)). First, the set of J outcomes is partitioned into K non-overlapping subsets (nests or branches), $B_k, k = 1, \ldots, K$. Within each subset there are J_k outcomes with $\sum_{k=1}^{K} J_k = M$. It is possible to write the theoretical specification for $S_i(\mathbf{Z}, \boldsymbol{\theta})$ from the NL directly (see p84 of Train (2003)). However, it is more informative to consider the decomposition $S_i(\mathbf{Z}, \boldsymbol{\theta}) = S_{i|B_k} S_{B_k}$. That is the share of i is given by the product of the share specification for i within subset B_k and the share specification for subset B_k.

We partition $\mathbf{Z}$ into two components $\mathbf{Z}_1$, $\mathbf{Z}_2$ where $\mathbf{Z}_1$ consists of variables that relate solely to the subset k and $\mathbf{Z}_2$ that consists of variables that relate solely to outcomes within the subset[1]. The NL specification comes from assuming that both of the probabilities in the decomposition follow a multinomial logit (MNL) form. That is,

$$S_{i|B_k} = \frac{\exp(\mathbf{Z}_1'(\boldsymbol{\theta}_{1i}/\lambda_k))}{\sum_{j\in B_k}\exp(\mathbf{Z}_1'(\boldsymbol{\theta}_{1j}/\lambda_k))}$$

and

$$S_{B_k} = \frac{\exp(\mathbf{Z}_2'\boldsymbol{\theta}_{2k} + \lambda_k I_k)}{\sum_{l=1}^{K}\exp(\mathbf{Z}_2'\boldsymbol{\theta}_{2l} + \lambda_k I_l)}$$

with $I_k = \ln\left(\sum_{j\in B_k}\exp(\mathbf{Z}_1'(\boldsymbol{\theta}_{1j}/\lambda_k)\right)$ is the "inclusive value" that summarizes the attractiveness of the particular subset k. Notice that when $\lambda_k = 1$, $k = 1,\ldots,K$. Thus, fitting the NL model also yields a convenient statistical test of the multinomial logit (MNL) specification.

Bechtel (1990) shows the the NL model can be estimated in a straightforward sequential manner using additive log-ratios and multivariate linear regression modeling. The first step is to consider each of the subsets (branches) individually. Noticing that within the branch the share specification is MNL and, normalizing on the last outcome in the subset, we can simply form estimate a linear multivariate regression for the y_i formed by the ALR for shares within the subsystem. From this estimation we can form an estimate of $\widehat{I_k}$ for the branch. Once we have a full set of $\widehat{I_k}$ then we can estimate another MNL model for the branch shares

[1] We also consistently partition $\boldsymbol{\theta}$.

that will depend upon the $\hat{I}_k$ and any variables that differ solely across branches, $\mathbf{Z}_1$. Again this is achieved by use of the ALR transformation and a multivariate linear regression. As with the discrete choice case this sequential estimation will yield unbiased and consistent estimators (see Train (2003))[2].

4. Zero observations in CODA

The final area in which CODA techniques may need to be modified to deal with a real life problem is the situation in which there are observed shares in our data set which are zero (see Adolph (2004), Aitchison (1986), Bacon-Shone (1992) and Fry *et al* (2000)). The statistical literature identifies two key explanations for the occurrence of zeros in compositional data. 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. Thus the observed zero is a proxy for a very small number. The second explanation argues that the observation should be zero as the data generating process leads to the occurrence of zeros. The proposed modifications to the CODA methodology to deal with the problem of zero observations are then derived by considering the cause of the zero observations (see Aitchison (1986) pp266-274). It is, however, possible to use any of the modifications regardless of how the zero observations arose (see Fry *et al* (2000)).

The modifications proposed are amalgamation, zero (trace) replacement, mod-

[2] However, the sequential estimator is not efficient.

ified Box-Cox, the use of ranks, and conditional modeling. Amalgamation is the reduction of the number of components in the composition by grouping together certain components. This may lead to certain “aggregate” brands (*e.g.* “Private Label”) which might be fairly heterogenous in character and will complicate the interpretation of the resultant estimated model. Zero replacement simply replaces the observed zeros with, appropriately chosen, small values and adjusts the non-zero components in an analogous manner. Modified Box-Cox uses a Box-Cox transformation in place of the log-ratio transform. This approach can be used in situations where one of the brands always has a share which is non-zero. Unfortunately, this approach seriously complicates the distribution theory and is, therefore, not as attractive as it appears.

Bacon-Shone (1992) proposes to replace the share data by ranks. Although this eliminates the problem of zero observations, it discards a large amount of information. The final modification to the CODA approach is to separate out the zero and non-zero components and model them using conditioning arguments. This is the preferred approach of Aitchison and Kay (2003). Adolph (2004) has implemented such an approach to yield a “zeros–inflated” compostional data model and applied it to the selection of central bankers. However, without an assumption of conditional independence between the data generating process for zero observations and the compositional data process the resultant modeling quickly becomes computationally difficult.

The choice of which modification to use is one that has received far less attention. Fry *et al* (2000) argue strongly that a zero replacement technique which is ratio preserving, is simple to implement, easy to work with, has a simple rationale and gives sensible results should be used. Recently, Aichison (2003a, 2003b) has suggested that a good starting point for analysis is the ratio preserving zero replacement procedure "modified Aitchison" suggested independently by Fry *et al* (2000) and Martin-Fernández *et al* (2000). 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. 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. This is the "modified Aitchison" procedure and in its application τ_A is often chosen within the context of the data at hand (e.g replacing zero budget shares with sensible values consistent with the dataset – see (Fry *et al* 2000, 2001).

The requirement that a zero replacement procedure for CODA retains the share ratios for the non-zero components means that those share ratios are, by construction invariant to the addition of components to the composition. This is exactly the property of Independence of Irrelevant Alternatives (IIA) that is inherent in the multinomial logit (MNL) model for discrete choice. Namely, that the odds ratio is invariant to additions or deletions to the choice set. Thus, the "modified Aitchison" zero replacement procedure that has become the default procedure is consistent with the MNL theoretical specification for $S_i(\mathbf{Z}, \boldsymbol{\theta})$. IIA in the discrete choice literature is viewed as an extremely restrictive property for models to have. It is argued that when outcomes are competing within a choice set we would expect that as we expand or contract the choice set the odds ratios might not be invariant. This has led to the development of a range of alternative model specifications that do not embody IIA but which can allow for tests for IIA. That is, specifications that include as a special case the (restricted) MNL model. Such models would of course yield alternative specifications for the theoretical specification $S_i(\mathbf{Z}, \boldsymbol{\theta})$. Interestingly, one such choice of non-IIA model is the nested logit (NL) model!

If the "modified Aitchison" zero replacement procedure is consistent with the MNL theoretical specification for $S_i(\mathbf{Z}, \boldsymbol{\theta})$ then we need to ask what zero replacement procedure would be consistent with a NL specification for $S_i(\mathbf{Z}, \boldsymbol{\theta})$? Fortunately, the decomposition of $S_i(\mathbf{Z}, \boldsymbol{\theta})$ in the NL specification gives us the answer. In the NL both the model for branches and the model for outcomes (components) within branches are of the MNL form. Thus, we can simply apply the "modified Aitchison" procedure within the branch(s) that the zero observation(s) appear. That is, the "modified Aitchison" procedure is used within the subset (branch) composition. If the the conditions ($\lambda_k = 1$, $k = 1, \ldots, K$) are met for the NL to collapse to the MNL then this is the same as the more usual MNL share ratio preserving "modified Aitchison" procedure. Thus the same statistical test for NL against MNL can also tell us about zero replacement procedures.

5. Application

The analysis in this paper is based upon data obtained from the 2001 Australian Federal Election and the 2001 Australian Census. The 2001 Australian Census, conducted by the Australian Bureau of Statistics, took place on August 7^{th} 2001 and provides a snapshot of the Australian community at that point in time. The Australian Census occurs once every five years and aims to collect accurate information on the number and characteristics of people living in Australia on census night. The census data can be mapped to varying geographical areas such as state level and postal area. Of particular relevance to us is that we are able to map census data to Commonwealth Electoral Division (CED), which allows us to obtain an accurate profile of each electorate. We are then able to combine this data with the vote count data available by electoral division from The Australian Electoral Commission. The close proximity of the 2001 Australian Federal Election (10^{th} November 2001) and the 2001 Australian Census (7^{th} August 2001) provides a unique opportunity for analysis to be undertaken.

We use the results of the voting in the House of Representatives to calculate the vote shares of the political parties in each of the 150 CEDs in Australia. The parties that contest on a national scale are the Australian Labor Party (ALP), the Coalition made up of both the Liberal Party and the National Party, The Australian Democrats and The Australian Greens. The One Nation Party and the other minor parties and independents (combined to form the group "Others") do not contest all CEDs.

The census data contain information regarding income, age, education, occupation, and dwelling types. We use variables computed from the census data to form a profile of each electorate. The majority of the census data is in the form of proportions, with the exception of median age and median weekly income and the Gini coefficient variables. In addition to the census variables there are four dummy indicator variables included. Three of the dummy variables indicate whether or not the candidate contesting the electoral division for the ALP, Coalition, or Other Parties is the current incumbent member of the parliament. The fourth dummy indicates if both the Liberal and National Parties contest the given electorate. A full description of all variables is provided in Chong *et al* (2005).

We use this data to fit a nested logit (NL) model. The structure of the NL model is that given in Figure 3.1 above. Namely, that there exist two branches – Major and Minor Parties – the Major Party branch consists of two parties (ALP and Coalition) and the Minor Party branch consists of four parties (Democrats, greens, One Nation and Others). We also fit an MNL model to this data that ignores the additional structure of the NL specification. Four political parties (ALP, Coalition, Democrats and Greens) contest all 150 electoral divisions. There are 108 electoral divisions in which all six political parties contest the division. A further 42 divisions have either one or two of the One Nation and Other Party not contesting the division. Thus we have a zero observations problem.

As argued in Chong et al (2005), we could replace the zero vote for the party not contesting with the value one[3]. Consistent with the arguments above when using the MNL specification we make the zero replacement using "modified Aitchison" to ensure that the share ratios in the full choice set remains unchanged. However, for the NL specification the zero replacement procedure is implemented as "modified Aitchison" within the branch that the zero observation(s) appear. In this case the zero observation(s) appear in the Minor Party branch and zero replacement is carried out preserving share ratios *within* that branch but not across the two branches.

Fitting the NL model also yields a convenient statistical test of the multinomial logit (MNL) specification and hence of the IIA assumption embodied within the MNL specification. In our application[4] the value of the log–likelihood test for the NL model against the (restricted) null MNL model is 161.293. This is a strong rejection of the MNL specification and, it is our opinion, that this strongly suggests that zero replacement should be carried out consistent with the NL model. That is, "modified Aitchison", or ratio preserving, *within branches.*

[3] The idea is that had the party contested the division then the candidate would have voted for themself!

[4] Full results available upon request from the authors.

6. Conclusions.

The traditional application of regression models for compositional data analysis through log ratio transformations corresponds to a multinomial logit model for the shares themselves. This model is characterized by the property of Independence of Irrelevant Alternatives. Independence of Irrelevant Alternatives states that the odds ratio – in this case the ratio of shares – is invariant to the addition or deletion of outcomes to the problem. It is exactly this invariance of the ratio that underlies the commonly used zero replacement procedure "modified Aitchison" in compositional data analysis. We present a model for shares, the nested logit model, that does not embody this property. We then discuss a zero replacement procedure that is consistent with this more general model. A simple statistical test can be used to determine which model is consistent with a data set. We apply these models and associated zero replacement procedures to a dataset that combines voting data by electoral division with corresponding census data for each division for the 2001 Federal election in Australia. We find evidence that the nested logit model and procedures are preferred.

7. References.

Adolph, C. (2004), "Succession in the Temple: Central Banker Careers and the Politics of Appointment", mimeo, August 25, 2004. http://chris.adolph.name/

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

Aitchison, J. (2003a), "Compositioanl data analysis: Where are we and where shoul we be heading?", Compositional Data Analysis Workshop (CoDAWork03), October 15–17, Girona, Spain. http://ima.udg.es/Activitats/CoDaWork03/

Aitchison, J. (2003b), *The Statistical Analysis of Compositional Data*, pub: Blackburn Press, New Jersey. Reprint of 1986 edition with additional material.

Aitchison, J. and J.W. ay (2003), "Possible solutions of some essential zero problems in compositional data analysis" Compositional Data Analysis Workshop (CoDAWork03), October 15–17, Girona, Spain. http://ima.udg.es/Activitats/CoDaWork03/

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

Barceló, C., Pawlowsky, V. and E. Grunsky (1996), "Some Aspects of Transformations of Compositional Data and the Identification of Outliers", *Mathematical Geology*, **28**, 501-518.

Bewley, R.A. (1982a), "On the Functional Form of Engel Curves: The Australian Household Expenditure Survey 1975-76", *Economic Record*, **58**, 82-91.

Bechtel, G.G. (1990), "Share-Ratio Estimation of the Nested Multinomial Logit Model", *Journal of Marketing Research*, **27**, 232-237.

Bewley, R.A. (1982b), "The Generalised Addilog Demand System Applied to Australian Time Series and Cross Section Data", *Australian Economic Papers*, **21**, 177-192.

Bewley, R.A. (1986), "Allocation Models: Specification, Estimation and Applications", pub: Ballinger, Cambridge, MA.

Chavas, J-P. and K. Segerson (1986), "Singularity and Autoregressive Disturbances in Linear Logit Models", *Journal of Business & Economic Statistics*, **4**, 161-169.

Chong, D., Fry, T.R.L., Davidson, S. and L Farrell (2005), "Compositional data analysis of vote shares in the 2001 Australian Federal Election", Compositional Data Analysis Workshop (CoDAWork05), October 19-21, Girona, Spain.
http://ima.udg.es/Activitats/CoDaWork05/

Cooper, L. (1993), “Market-Share Models”, in J. Eliashberg and G.L. Lilien *eds.*, *Handbooks in Operations Research and Management Science, Volume 5: Marketing*, pub: North Holland, Amsterdam.

Cooper, L.G. and M. Nakanishi (1988), *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness*, pub: Kluwer, Boston.

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”, *Journal of Econometrics*, **73**, 377-385.

Fry, J.M., Fry, T.R.L. and K.R. McLaren (2000), “Compositional Data Analysis and Zeros in Micro Data”, *Applied Economics*, **32**, 953-959.

Fry, J.M., Fry, T.R.L., McLaren, K.R. and T.N. Smith (2001), “Modelling Zeroes in Microdata”, *Applied Economics*, **33**, 383-392.

Ghosh, A., Neslin, S. and R. Shoemaker (1984), “A Comparison of Market Share Models and Estimation Procedures”, *Journal of Marketing Research*, **21**, 202-210.

Howel, D. (1994), “Statistical Analysis of Compositional Data in Anatomy”, *The Anatomical Record*, **240**, 625-631.

Martin-Fernández, J.A., Barceló, C. and V. Pawlowsky-Glahn (2000), “Zero replacement in compositional data sets” in H. Kiers, J. Rasson, P. Groenen and M. Shader, eds., *Studies in Classification, Data Analysis and Knowledge Organisation, Proceedings of 7th Conference of The International Classification Society*, pub: Springer-Verlag, Berlin, pp 155-160.

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.

McFadden, D. (1978), “Modeling the choice of residential location” in A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull, eds., *Spatial Interaction Theory and Planning Models*, pub: North-Holland, Amsterdam, pp 75-96.

Theil, H. (1969), “A Multinomial Extension of the Linear Logit Model”, *International Economic Review*, **10**, 251-259.

Train, K.E. (2003), *Discrete Choice Methods with Simulation*, pub: Cambridge University Press, Cambridge.