Market Share Forecasting: An Empirical Comparison of Artificial Neural Networks and Multinomial Logit Model

DEEPAK AGRAWAL and CHRISTOPHER SCHORLING
Purdue University

We empirically compare the forecasting ability of artificial neural network (ANN) with multinomial logit model (MNL) in the context of frequently purchased grocery products for a retailer. Using scanner data on three grocery product categories, we find that performance of ANN compares favorably to MNL in forecasting brand shares. We test the sensitivity of the forecasting error in the two approaches to the length of the estimation period and the clustering of households which is used to define homogenous segments of households. We find the results to be robust to these variations. We also derive a few empirical propositions regarding performance of ANN and MNL from our analysis. The results are consistent with those in Kumar, Rao and Soni (1995) and suggest that although neural networks suffer from interpretability problem, they are a useful method to forecast brand shares in grocery product categories where large amounts of scanner data are readily available.

Artificial neural network (ANN) models are increasingly being used as a decision aid in a number of areas such as manufacturing, marketing, and retailing. Some applications in manufacturing have been in systems design, capacity planning, and product quality control (see Huang and Zhang, 1994 for a review of ANN applications in manufacturing). In marketing, applications have been reported in the areas of customer service, segmentation, and prospect identification (Shepard and Ratner, 1994; Venugopal and Baets, 1994). Some applications have also been reported in retailing particularly in fashion forecasting, retail assortment planning, and retail inventory management (Belt, 1993; Dragstedt, 1991). Retailers are also exploring neural networks to forecast retail demand in the grocery products industry (Thall, 1992). The use of neural networks in grocery retailing is especially appealing because large amounts of UPC (universal product code) scanner data routinely become available from the electronic check-out systems installed in the retail stores. Accurate demand forecasting is crucial for profitable retail operations because without a good

Deepak Agrawal, Purdue University, Krannert Graduate School of Management, West Lafayette, IN 47907-1310. E-mail: <agrawal@mgmt.purdue.edu.>. Christopher Schorling, Purdue University, School of Industrial Engineering and Technische Universität Berlin, Fachbereich Wirtschaft und Management, Strasse des 17. Juni 135, 10623 Berlin, Germany.

Journal of Retailing, Volume 72(4), pp. 383–407, ISSN: 0022-4359

forecast, either too-much or too-little stocks would result, directly affecting revenue and competitive position.

In the traditional econometric modeling arena, one technique which has emerged quite robust is the multinomial logit model (MNL) for the polychotomous choice problem found in grocery scanner panel data. The MNL model has been shown to be more appropriate for modeling consumer's probability of choice as a function of a mix of continuous and discrete predictor variables found in the panel data as compared to its rivals such as multiple regression, log linear, multiple discriminant and multinomial probit models (Gensch and Recker, 1979; Green, Carmone and Wachspress, 1977; Maddala, 1983; Malhotra, 1984). One widely recognized advantage of MNL is its ability to provide closed form solutions for the choice probabilities in a competitive setting where marketing activities of all players are taken into consideration. The choice probabilities can be aggregated to yield estimates of brand shares for a particular marketing mix environment.

An alternative to MNL is ANN which can also be used to forecast brand shares. Although it is a feasible alternative to MNL, its forecasting ability compared to MNL is not clear. In general there is a continuing debate about the comparative performance of the ANN and the traditional econometric approaches in several different contexts (Federowicz, 1994; Hruschka, 1993; Kumar, Rao, and Soni, 1995; Shepard and Ratner, 1994).

Recently Kumar et al. (1995) compared ANN with logistic regression model for a binary choice problem and found ANN to be quite comparable in performance. Similarly we focus here on the polychotomous choice problem of a grocery retailer and compare the forecasting ability of ANN with that of well established multinomial logit model. The main objective here is to assess whether ANN is an acceptable alternative to MNL for forecasting brand shares of grocery products.[1] We use scanner data from three frequently purchased categories, namely, peanut butter, dishwashing liquid, and catsup, to do the comparison.

Our results indicate that ANN forecasts brand shares better than MNL in peanut butter and dishwashing liquid categories, and moderately better in the catsup category. Although these results may be specific to the categories we use, a few generalizations seem to emerge from our analysis. For example, the results show that ANN performs significantly better than MNL when the number of brands in the category are numerous (as in dishwashing liquid and peanut butter categories). This is consistent with the previous findings in the literature that neural networks outperform other methods when complex and non-linear data patterns are present (Hruschka, 1993; Kumar et al., 1995; Venugopal and Baets, 1994). Our findings are also consistent with Kumar, Rao, and Soni (1995) who state:

> The neural network approach is parsimonious, produces better classification, handles complex underlying relationships better, and is stronger at interpolation. On the other hand, the logistic regression technique has a superior solution methodology (closed form versus heuristic) and better interpretability (p. 261-262).

We also analyze sensitivity of the forecasting error to the length of the estimation (training) period for MNL (ANN) model, and to the different schemes for classifying households into homogenous segments (the segmentation is used to circumvent the need for household

level estimation as we discuss later). We find the above results to be reasonably robust to the different clustering criteria and the length of estimation period.

The main contribution of this paper is in demonstrating that neural networks can be usefully employed in demand forecasting for the grocery product retailers, and that their performance is comparable and sometimes even better than that of more traditional econometric approaches such as multinomial logit model. This finding is important especially because the two approaches differ to a significant extent in terms of interpretability, software requirements, data preparation, computational ease, computing time, and analytical effort. We found neural networks to be relatively easier in terms of analytical and computational effort but the logit model to be much better in terms of interpretability.

We organize the rest of the paper as follows. In the next section we describe logit model estimation. Then we present neural network estimation in section 3. Next we present and discuss forecasting results. We conclude with a summary of findings and directions for future research.

MULTINOMIAL LOGIT MODELL APPLICATION

The multinomial logit model of brand choice assumes that the utility derived from a brand is a function of brand specific characteristics and its marketing mix activities such as price, advertising, and merchandising, and an error term. A household chooses a brand from among the brands available in the category which provides highest utility. If the error term is assumed to be distributed double exponentially, then Thiel (1969) has shown that a brand's choice probability can be expressed as a ratio of the exponentiated utility of the brand to the exponentiated sum of utilities of all brands. Specifically,

$$p_{it}^h = \exp(u_{it}^h) / \sum_j \exp u_{jt}^h \tag{1}$$

$$u_{it}^h = \alpha_i + \gamma\, propen_i^h + \sum_k \beta_k X_{it}^k + \xi_{it}^h \tag{2}$$

where

p_{it}^h = the probability that household h purchases brand i at purchase occasion t,
u_{it}^h = the utility to household h of purchasing brand i at purchase occasion t,
$propen_i^h$ = a measure of loyalty (or heterogeneity) across households in the panel, specifically household h's propensity towards brand i based on purchase behavior before time t,
X_{it}^k = the level of marketing mix variable k for brand i at purchase occasion t,
α_i = brand specific constant, specifically the effect of characteristics unique to a brand,
γ = effect of household loyalty,
β_k = effect of marketing mix activity k,and
ξ_{it}^h = error term distributed double exponentially.

Choice of a Store

The households in the scanner panel can shop in more than one store. The store environment in the scanner panel, namely brand price and merchandising activities, may be similar across stores yet a household may purchase different brands in different stores. This may occur if something other than price and merchandising activities affects brand choice. For example, a store's overall positioning on the price-quality dimension may systematically affect brand choice. To circumvent this confounding effect of store positioning, we focus on only one store for estimating both the logit model and the neural network.

In each category we choose that store which has maximum number of purchases made by the households. Note that restricting analysis to only one store is also consistent with the retail level nature of this demand forecasting study.

Model Estimation With Brand Loyalty

To estimate the MNL specified in Equations 1 and 2, it is necessary to have an estimate of loyalty, the $propen_i^h$ term. This was estimated using the procedure described in Agrawal (1996) (first proposed in Srinivasan and Kibarian, 1989) which is arguably advantageous over previous methods in that it filters out the effect of marketing mix activities in estimating brand loyalty. Also it provides an estimate of brand loyalty without needing semi-parametric estimation procedures which are computationally much difficult.

The procedure used for estimating MNL with brand loyalty is as follows:

STEP 1

The data are divided into three separate parts, caliberation, estimation, and the prediction periods.

STEP 2

A logit model is estimated with the following utility formulation on the caliberation period data using maximum likelihood procedure:

$$u_{it}^h = \alpha_i + \sum_k \beta_k X_{it}^k + \xi_{it}^h. \tag{3}$$

STEP 3

The estimated parameters are used to compute predicted probability of choice, $\hat{q}_{it}^h$ over the calibration period. The loyalty measure is then computed as:

$$propen_{it}^h = \delta propen_{it-1}^h + (1-\delta)(y_{it-1}^h - \hat{q}_{it-1}^h) \tag{4}$$

where,

$y^h_{it-1} = 1$ if i is chosen at $t-1$, else 0.

Here $\sum_i propen^h_{it} = 0$ and $0 < \delta <, 1$ is a smoothing parameter. The $propen^h_{it}$ measure is initialized as 0.0 for all brands. The $propen^h_{it}$ estimates were found *not* to be significantly sensitive to different values of δ. A value of 0.75 for δ was used in all estimations. The $propen^h_i$ vector for household h on the last purchase occasion T in the calibration period is used as a measure of brand loyalty as it represents the most stable vector in the calibration period of household propensities towards different brands.

STEP 4

Next a logit model is estimated with the following utility formulation on the estimation period data using maximum likelihood procedure:

$$u^h_{it} = \alpha_i + \gamma\, propen^h_i + \sum_k \beta_k X^k_{it} + \xi^h_{it} \tag{5}$$

where the $propen^h_i$ is a measure of loyalty across households in the panel. The estimation provides the estimates of model parameters $\hat{\alpha}_i$'s, $\hat{\gamma}$, and $\hat{\beta}_k$.

STEP 5

Next the household level predicted probability of choice of each brand on each occasion on the prediction period data is computed as follows:

$$\hat{p}^h_{it} = \frac{\exp\left(\hat{\alpha}_i + \hat{\gamma}\, propen^h_i + \sum_k \hat{\beta}_k X^k_{it}\right)}{\exp\left(\hat{\alpha}_j + \hat{\gamma}\, propen^h_j + \sum_k \hat{\beta}_k X^k_{jt}\right)} \tag{6}$$

where

$\hat{p}^h_{it}$ = predicted probability of choice of brand i for household h at purchase occasion t.

STEP 6

Lastly the predicted brand share is computed as the average predicted probability of choice of a brand in a particular week as follows:

$$\hat{p}_{it} = \frac{\sum_{h=1}^{n} \hat{p}^h_{it}}{n} \tag{7}$$

where

$\hat{p}_{it}$ = predicted share of brand i in week t, n = the number of purchase occasions in week t.

These predicted probabilities are then compared with actual brand shares to compute forecasting error, MAE (mean absolute error), as described later.

Model Estimation at the Segment Level

One limitation of the above MNL estimation at the household level is that it is necessary to have a large number of purchase observations made by that household. Typically however a scanner panel does not contain many purchases for a household. For example, the average number of purchases made by a household (in the store with most purchases) is, 15.3, 16.7, and 11.3 over a 161 week period in the dishwashing liquid, peanut butter, and catsup categories respectively. There are only 7, 8, and 3 households respectively in these categories who made more than 50 purchases over the 3 year period. With so few observations it is difficult to estimate the logit model coefficients reliably at the household level.

One solution to this problem is to estimate the MNL at a segment level where all households are homogenous with in the segment. This clustering of households provides us a sufficient number of observations to estimate the logit model reliably. Also by treating all households alike within a cluster we do away with the need to estimate a household specific loyalty measure ($propen_l^h$) in the brand utility function (Equation 2). This implies one less parameter to estimate thus increasing degrees of freedom in estimation. Furthermore, in segment level analysis there is no need to use any purchase observations to calibrate the loyalty measure thus more data become available for parameter estimation.[2]

A desirable clustering criteria for grouping households into homogenous segments needs to be independent of the purchase behavior and the associated store environment used in estimating the logit model. There are several possibilities for this type of clustering.

One is to use a portion of the purchase data (for example, first 25% of the number of weeks) to derive homogenous clusters. This approach reduces the number of observations available for model estimation. Another approach could be to use demographic data such as household size, income, and education for clustering the households. This approach does not guarantee homogenous clusters in terms of purchase behavior with respect to product categories of interest. That is, households may be similar on demographics yet exhibit dissimilar purchase patterns in the frequently purchased low price product categories available in a typical grocery store. A third approach is to use all available purchase data and employ an independent criterion, largely unrelated to the associated store environment, to classify households into homogenous clusters. We use such an approach here.

Method of Clustering Households

We classify households into different segments based on the number of different brands they purchase. This criterion essentially uses the variety seeking dimension to cluster the households (see McAlister and Pessemier, 1982 for discussion on role of variety seeking in choice behavior) and is also consistent with the notion of consideration sets being indica-

tors of household preferences (Horowitz and Louviere, 1995). In order to test the sensitivity of forecasting ability of the MNL and ANN, we define 4 different clusters:

1. *Cluster 0*: Group of all households who purchase in the store of interest. This cluster is an all inclusive one and does not differentiate among households. This serves as a benchmark.
2. *Cluster 1*: Group of households who purchase only one brand from among the ones available in the category. This cluster contains households who do not desire variety because they purchase the same brand regardless of the store environment.
3. *Cluster 2*: Group of households who purchase two or three different brands from among the ones available in the category. This cluster contains brand switching households.
4. *Cluster 3*: Group of households who purchase four or more different brands from among the ones available in the category. This cluster contains relatively more variety seeking households.

This clustering method is appealing because it uses all the information available in the purchase panel and yet saves purchase observations for model estimation. Also it does not require any additional data (for example, demographic data) for clustering.

The difference between this and the previous MNL estimation is that the data are now divided into two parts, estimation and prediction periods. The MNL is estimated on the estimation period data (without brand loyalty) using the utility formulation in Equation 3 and brand choice probabilities are estimated on the prediction period data. The predicted probability of choice of each brand on each occasion on the prediction period data is computed as follows:

$$\hat{p}_{it} = \exp\left(\hat{\alpha}_i + \sum_k \hat{\beta}_k X_{it}^k\right) \Big/ \sum_j \exp\left(\hat{\alpha}_j + \sum_k \hat{\beta}_k X_{jt}^k\right) \tag{8}$$

where

$\hat{p}_{it}$ = estimated value of choice probability,
$\hat{\alpha}_i$ and $\hat{\beta}_k$ = estimated values of parameters.

These predicted choice probabilities are then averaged over all purchase occasions during a week to yield an estimate of weekly brand share.

Effect of the Length of Estimation Period

It is well documented in literature that number of observations used in the MNL estimation affects the quality of estimates. In other words how well the model is able to fit the data and subsequently predict brand choice probabilities depends on the number of data points

used to estimate the logit coefficients. Same is believed to be true of ANN (Wasserman, 1989). To test whether the comparison of forecasting ability of the logit model and neural network is affected by the number of observations in the estimation period, we use three different data partitionings.

At the outset we would expect the forecasting ability of the MNL (ANN) to be directly related to the number of observations in the estimation (training) period. However the relative forecasting ability of ANN versus MNL is of particular interest to us. The three data partitions we use to test the sensitivity of the MNL and ANN to the number of purchase observations in the estimation (training) period are:

1. *Data Partition 80-20*: The first 80% of the weeks are used to estimate (train) the MNL coefficients (ANN) and the rest 20% are used to test the forecasting ability of the MNL (ANN) model.
2. *Data Partition 65-35*: The first 65% of the weeks are used to estimate (train) the MNL coefficients (ANN) and the rest 35% are used to test the forecasting ability of the MNL (ANN) model.
3. *Data Partition 50-50*: The first 50% of the weeks are used to estimate (train) the MNL coefficients (ANN) and the rest 50% are used to test the forecasting ability of the MNL (ANN) model.

Both the MNL (without brand loyalty) and ANN models are estimated for all 12 cluster-data partition combinations. The MNL model with brand loyalty is estimated only for cluster 0, because for other clusters (1, 2, and 3), household heterogeneity is conceptually not an issue.

Calculation of Forecasting Error

We interpret the predicted choice probabilities (averaged over all purchase occasions during a week) from MNL as predicted brand shares given a particular store environment (i.e. given prices and merchandising activities of all brands in that week in the store). These predicted brand shares are then compared to the actual brand shares observed in prediction period data. The actual brand shares are simply the brand shares each week in the store. For example, if brands, 1, 2, and 3 were purchased 20, 15, 10 times during the week then the brand shares are 0.44, 0.33, and 0.22 respectively.

We calculate the forecasting error, a measure of the forecasting ability of the MNL model as follows:

$$MAE_{MNL} = \left(\frac{100}{T \times I}\right) \sum_{t} \sum_{i} \left| \hat{p}_{it} - s_{it} \right| \tag{9}$$

where

MAE_{MNL} = mean absolute error obtained from the MNL model,
$\hat{p}_{it}$ = predicted brand share of brand i in week t given the store environment,
s_{it} = actual brand share of brand i in week t given the store environment,
i = index for brand i (i = 1 to I, where I = number of brands in the category), and

t = index for week t (t = 1 to T, where T = number of weeks in the prediction period of the category).

This measure of forecasting error called the mean absolute error is well established in literature (see Makridakis, Wheelwright and McGee, 1983, p. 44).

Description of the Data Sets and the Specific MNL Model

The three categories we use in this paper to compare the forecasting ability of MNL with ANN were obtained from Information Resources Inc. (I.R.I.). We obtained two files from I.R.I. One is the purchase data file which contains information on the brand purchased, the week and the store in which the purchase was made. The other is the store environment file which contains brand price (expressed in dollars per unit weight or volume), feature (F_{it} = 1 if featured by the retailer in its advertising, 0 otherwise), and display (D_{it} = 1 if displayed by the retailer in the store, 0 otherwise) indicators for each brand in the category for each week in each store. These three variables form the predictor variable set X_{it}^k in the model estimations.[3] A brief description of the three datasets is provided in Table 1.

Each data set was divided into two or three partitions as described above. Logit estimates were obtained using maximum likelihood procedure. Same observations were used to train the ANN. The logit estimates were then used to forecast brand shares in the prediction period and similarly ANN was used to forecast the brand shares over the same prediction period given store environment. It is noteworthy that we used identical clustering and data partitions for estimation and prediction in the two approaches.

A sample outcome of MNL parameter estimation is given in Table 2. We next describe the neural network model and its implementation.

TABLE 1

Data Description

	Catsup	*Peanut Butter*	*Dishwashing Liquid*
Number of Brands	4	6	11
Number of Purchases	2301	3927	2493
Number of Households	204	235	163
Number of Weeks	161	161	161

TABLE 2

MNL Estimates for Peanut Butter, Cluster 3, 80-20 Data Partition

Brand	*Constant*	*Price*	*Feature*	*Display*	*Price*(F or D)*
1	−0.00350	−0.11707	0.00447	0.00644	0.07779
2	0.00981	−0.11707	0.00447	0.00644	0.07779
3	−0.01126	−0.11707	0.00447	0.00644	0.07779
4	−0.00363	−0.11707	0.00447	0.00644	0.07779
5	−0.00587	−0.11707	0.00447	0.00644	0.07779
6	0.00000	−0.11707	0.00447	0.00644	0.07779

Notes: Sample size = 1,009
Log likelihood function value at the optimum = −1,666.85

ARTIFICIAL NEURAL NETWORK APPLICATION

We provide a short overview of the artificial neural networks (ANNs) and the backpropagation training algorithm. It is not our intention to explain more than the basic concepts here. Furthermore we do not discuss network topologies other than the layered feedforward. The reader may refer to Masson and Wang (1990) and Rumelhart, Hinton, and Williams (1986) for an introduction to ANN, and to Chauvin and Rumelhart (1995), Wasserman (1989), and White et al. (1992) for a more detailed description of ANN learning algorithms and topologies.

An ANN consists of a number of connected nodes (in the literature nodes are also referred to as neurons, units, or cells) each of which is capable of responding to input signals with an output signal in a predefined way. These nodes are ordered in layers. A network consists of one input layer, one output layer, and an arbitrary number of hidden layers in between. This number can be chosen by the user such that the network performs as desired. Typically one or sometimes two hidden layers are used. One reason for this is that one hidden layer is sufficient to approximate any continuous function to an arbitrary precision (Hornik, Stinchcombe and White, 1989).

For an illustration consider the three-layer ANN in Figure 1. This ANN consists of three layers, the input layer (the leftmost), one hidden layer (in the middle), and the output layer (the rightmost). The nodes are connected such that each node is connected to all nodes of the previous and the successive layer (if such layers exist). The input layer is only connected forward to the first hidden layer and the output layer only backward to the last hidden layer. All connections are assigned a weight (a real number). Often an ANN also contains biases (denoted by node b in Figure 1). These are dummy nodes which always provide an output of +1. They are useful in translating the [0, 1] output from the logistic function.

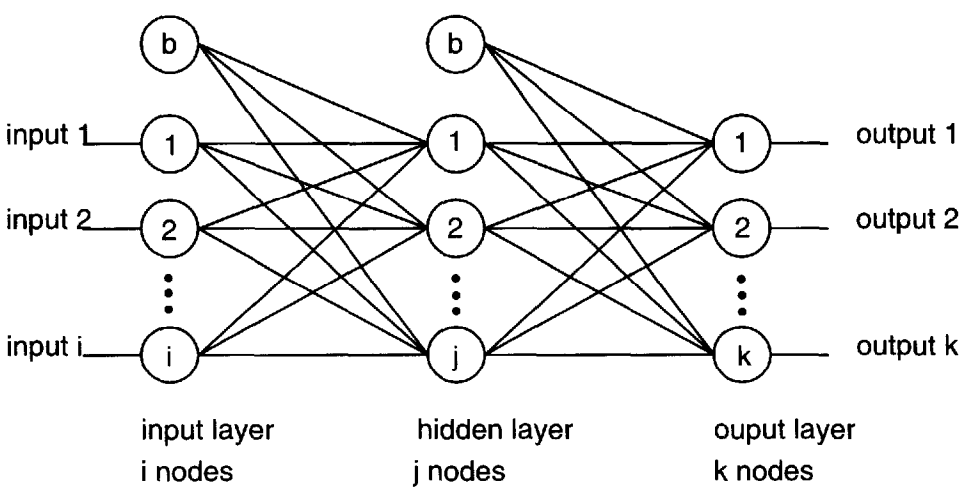

Figure 1. **A Three Layer Artificial Neural Network with Biases**

Similar to estimation of logit model over an estimation period data, the ANN gets trained on a set of training data. ANN starts out by an initial set of weights chosen randomly, typically between (−1, 1). It then adapts the weights in such a way that given the input signals, the ANN's output signal(s) match the desired output signal(s) as closely as possible (the convergence limit is specified by the user).

We use a particularly popular algorithm called the backpropagation algorithm in this study. The basic algorithm works as follows. The input to a node is computed as the sum of the outputs of the preceding nodes multiplied by the weight of the connection. This is expressed as

$$NET = \sum_{i=1}^{n} OUT_i w_i \tag{10}$$

where

OUT_i = the output of node j in the previous layer,
w_i = the corresponding connection weight.

For the input layer OUT_i is simply the vector of input values. This sum is then transformed to a value between 0 and, 1 using the so called logistic or sigmoid function

$$OUT = \frac{1}{(1 + e^{-NET})}. \tag{11}$$

Starting with the first hidden layer this calculation is done from left to right until the output layer is reached. All training pairs are presented to the ANN and the sum of squared errors over the whole training set is computed. If the sum of squared error exceeds the specified error goal, the ANN adjusts the connection weights. This is called a training epoch. The ANN then begins another training epoch until either the maximum number of training epochs is reached or the sum of squared errors reaches the specified error goal. The training is said to be complete when either of this happens. One can think of this as moving on the (often multidimensional) error surface in the direction of the steepest descent. How well a network is trained is measured by the mean sum-squared error over the complete training dataset.

The connection weights are adjusted as follows. Starting with the weights connecting output layer and the last hidden layer the weight adjustments are propagated backwards using

$$\delta_{p,\,output} = OUT(1 - OUT)(TARGET - OUT) \tag{12}$$

where $\delta_{p,\,output}$ is the delta value of node p in the output layer.

Based on this the weight change is calculated:

$$\Delta w_{pq,\,k} = \eta \delta_{qk} OUT_{pj} \tag{13}$$

where

$\Delta w_{pq,k}$ = weight change of connection from node p in layer $k-1$ to node q in layer k,
η = learning rate (which can be set by the user),
δ_{qk} = delta value for the node q in layer k, and
OUT_{pj} = output of node p in layer j (same as $k-1$).

The new weight assigned to this connection is computed as

$$w_{pq,k}(n+1) = W_{pq,k}(n) + \Delta w_{pq,k} \tag{14}$$

where n denotes the current iteration (before weight adjustment) and $n+1$ the next iteration (after weight adjustment). This procedure is repeated for all nodes in the output layer. Afterwards the incoming connections of the previous layer are updated.

For layers other than the output layer

$$\delta_{p,j} = OUT_{p,j}(1 - OUT_{p,j})\left(\sum_q \delta_{q,k} w_{pq,k}\right) \tag{15}$$

is used, where

$\delta_{p,j}$ = delta value of node p in layer j,
OUT_{pj} = output of node p in layer j,
δ_{qk} = delta value for the node q in layer k, and
$w_{pq,k}$ = weight of connection from node p in layer $k-1$ (same as j) to node q in layer k.

The other steps remain the same. This procedure continues until a specified error is reached or a specified number of training epochs are over.

Data Preparation for Neural Network Implementation

In this study we use a fully connected ANN. A fully connected ANN is the default ANN unless there is specific information suggesting a partially connected network. For implementation we used the neural network toolbox in the software package MATLAB which also offers improvements (namely, momentum and an adaptable learning rate) of the back-propagation algorithm to increase the convergence speed (see Demuth and Beale, 1992; Vogl, Mangis, Rigler, Zink, and Alkon, 1988).

To speed up the learning process of the ANN in the scanner data context we employed a few data transformations. First we replaced all feature and display indicator variables of 0-1 by .1 and .9 respectively and replaced a brand share of, 1.0 with 0.99 and a brand share of 0.0 with 0.01 wherever they occurred. This was done because the log-sigmoid function in the ANN is particularly slow in learning with values close to 0 or 1 which require very large or even infinite weights. Next we normalized the prices of all brands to the interval .1

to .9.[4] Again the reason is to increase the learning speed because ANN tends to converge slowly if it has to handle relatively large numbers (for example, 18 cents per ounce versus a .6 on the .1 to .9 scale).

Configuration of the Neural Network

The number of nodes in the output layer equals the number of brands, and the number of nodes in the input layer equals three times the number of brands since there are three store environment variables (price, feature, and display) for each brand.[5]

In order to determine the number of hidden layers and the number of nodes to use in these hidden layers it was necessary to conduct preliminary experiments. This was done by examining different possible configurations and choosing that combination which resulted in lowest mean forecasting error defined as,

$$MAE_{ANN} = \left(\frac{100}{T \times 1}\right) \sum_{t} \sum_{i} \left| \hat{p}_{it} - s_{it} \right| \tag{16}$$

where

MAE_{ANN} = mean absolute error obtained from ANN,
$\hat{p}_{it}$ = predicted brand share of brand i in week t given the store environment,[6]
s_{it} = actual brand share of brand i in week t given the store environment,
i = index for brand i (i = 1 to I, where I = number of brands in the category), and
t = index for week t (t = 1 to T, where T = number of weeks in prediction period of the category).

We used the peanut butter data (cluster 3) for this experimentation. The ANN was first trained with 5, 10, 15, 20, and 25 nodes respectively in one hidden layer. After training was considered complete (also see discussion on choosing number of epochs below) the testing partition of the data was used to compute MAE_{ANN} which became the first criterion to evaluate the capability of the ANN as a forecasting tool.

Furthermore since the weights and biases of the ANN are randomly initialized to values between −1 and, 1, we can treat the MAE_{ANN}, after a specified number of training epochs, as a random variable. The standard error of the computed MAE_{ANN} over a set of random starting points was used as a measure of stability in the results and chosen as the second criterion for selecting the ANN configuration. We implemented the ANN 20 times for each cluster-data partition combination to compute the mean and the standard error of MAE_{ANN} (reported in Tables 4(a), 4(b), 4(c)).

Choosing Number of Nodes in the Hidden Layer

The results of the investigation indicated that optimal number of nodes is somewhere between 1 and 15. Further, a plot of MAE_{ANN} against number of nodes (with 100 training epochs) indicated that five nodes in the hidden layer should be a good choice (see Figure 2).

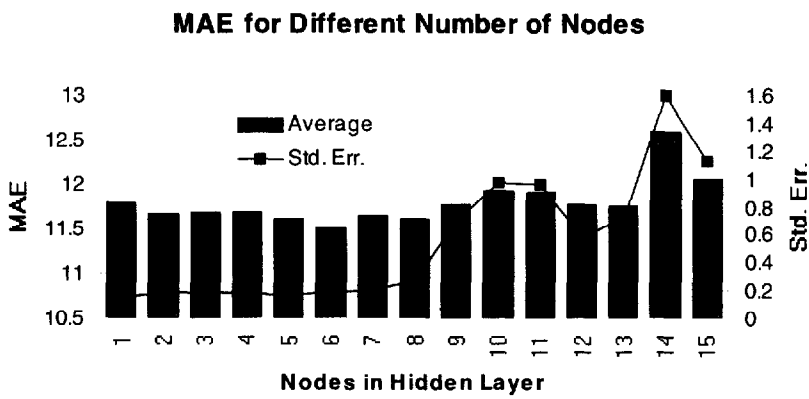

Figure 2. **MAE_{ANN} for Different Number of Nodes (100 Traning Epochs)**

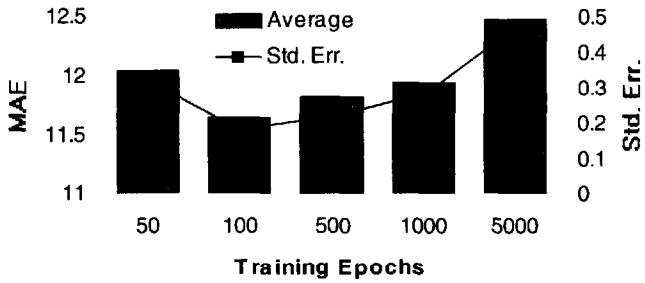

Figure 3. **MAE_{ANN} for Different Traning Epochs (5 Nodes)**

Choosing Number of Epochs

Along with number of nodes in the hidden layer, we simultaneously experimented with different number of training epochs. The initial experiments showed that given a particular number of nodes, longer training of the ANN did not always improve its forecasting capability. Therefore rather than specifying a target error we searched for an "optimal" number of training epochs after which to stop the training. The investigations indicated that, 100 training epochs (with 5 nodes in the hidden layer) would provide a good performance of the network (see Figure 3).

Furthermore we found that adding a second hidden layer did not yield better results. Thus we decided upon using a backpropagation ANN with one hidden layer containing 5 nodes. We trained the ANN with 100 epochs. For all implementations of the ANN in the 12 cluster-data partition combinations we used this same configuration. The idea behind using the

TABLE 3

ANN Weights for Peanut Butter, Cluster 3, 80-20 Data Partition

Weight Matrix from Input Layer to Hidden Layer

From Input	*Node 1*	*Node 2*	*Node 3*	*Node 4*	*Node 5*
Price 1	−0.8074	−1.0452	0.1942	0.1794	0.9745
Feature 1	0.9224	0.3823	0.6856	0.0228	−0.9633
Display 1	0.7714	0.4622	−0.8877	0.9732	−0.6205
Price 2	−0.2871	0.4185	−0.2733	0.9831	0.5051
Feature 2	0.2141	−1.1423	1.1842	0.1161	1.4139
Display 2	0.5462	0.1132	−0.0528	−1.1181	0.5454
Price 3	−0.8441	0.4200	−0.0696	1.2445	−0.4598
Feature 3	−0.4902	0.9270	0.4466	0.4483	0.3492
Display 3	−0.8390	0.2252	0.7707	−0.5130	−0.0810
Price 4	0.6571	−0.0975	−0.9276	−0.7167	−0.4713
Feature 4	−1.7767	0.8133	0.8371	2.1876	−1.4159
Display 4	0.8250	−0.0291	0.0338	−0.4203	1.0195
Price 5	0.3552	−0.7074	−0.7674	0.6943	−0.8705
Feature 5	0.0171	−0.2698	−0.4446	0.7693	0.1057
Display 5	−0.0554	0.8439	−0.8986	0.4646	0.5866
Price 6	0.5880	−0.5307	−1.9352	0.5278	−0.3217
Feature 6	0.0523	0.4961	1.2040	0.4623	0.8618
Display 6	−0.7215	−0.8178	−0.4276	0.1516	0.5243
From	*Node 1*	*Node 2*	*Node 3*	*Node 4*	*Node 5*
Bias Node 1	0.8487	−0.5568	0.6955	−1.0452	0.2929

Weight Matrix from Hidden Layer to Output Layer

Hidden Layer	*Brand 1*	*Brand 2*	*Brand 3*	*Brand 4*	*Brand 5*	*Brand 6*
Node 1	0.8874	1.0899	−1.2412	−1.7445	−1.1004	−1.5929
Node 2	0.7269	−0.9130	−0.3286	−0.5914	0.6763	0.7588
Node 3	−1.8173	1.1529	−0.4368	−0.2228	−0.9054	1.3262
Node 4	−0.3487	−2.2821	−1.6283	−0.4907	−1.2490	1.4386
Node 5	−2.3702	0.6261	0.1160	0.1272	−0.6555	−0.7317
From	*Brand 1*	*Brand 2*	*Brand 3*	*Brand 4*	*Brand 5*	*Brand 6*
Bias Node 2	−0.0912	−0.1621	−1.9573	−0.5771	−0.7915	−1.1851

Note: Training duration 100 epochs, 1 replication.

same configuration is simply that in a real world setting it is not practical to always search for the optimal configuration. Rather a configuration may be selected only once, after preliminary investigations. We did, however, also search for an optimal number of epochs in each case (with 5 nodes in the hidden layer). We report the best performance of the ANN with the optimal number of epochs in each case alongside the results with 100 epochs in Tables 4(a), 4(b), and 4(c).

A sample of the weights of a trained ANN is given in Table 3. We next discuss the results of estimation.

DISCUSSION OF RESULTS

The forecasting errors obtained from the MNL and the ANN for the 12 cluster-data partition combinations for each of the three categories are summarized in Tables 4(a), 4(b), and 4(c). As an illustration of the brand share forecasts from the two approaches, consider brand 1 in cluster 0 of the peanut butter category. For this brand the actual price and merchandising activities in weeks 621 and 622 are as follows:

Week	Actual price	Feature	Display
621	14.125	0	0
622	14.312	0	1

The price is rescaled on a 0.1-0.9 scale and feature and display variables are modified for ANN as follows:

Week	Normalized price	Feature	Display
621	0.483	0.1	0.1
622	0.503	0.1	0.9

TABLE 4a

MAE Comparisons Between ANN and MNL for Catsup

Data	MNL		ANN			Comparison		Error
Split	MAE	N	MAE	Std. Error	Best Result	t-value	p-value	Improvement
Cluster 1								
80:20	25.75	412	5.47	0.073	5.03	277.808	4.7E-36	79%
65:35	44.31	298	5.77	0.119	5.15	323.866	2.55E-37	87%
50:50	47.76	208	6.09	0.147	5.56	283.469	3.2E-36	87%
Cluster 2								
80:20	13.17	1186	13.66	0.130	12.83	−3.769	0.000649	−4%
65:35	14.03	911	14.23	0.063	14.23	−3.175	0.002495	−1%
50:50	14.10	712	14.43	0.062	14.43	−5.323	1.94E-05	−2%
Cluster 3								
80:20	23.66	176	25.96	0.201	25.86	−11.443	2.88E-10	−10%
65:35	21.85	125	25.82	0.286	25.82	−13.881	1.07E-11	−18%
50:50	23.55	91	25.33	0.046	25.33	−38.696	7.74E-20	−8%

Data	MNL		MNL with Brand Loyalty		ANN			Comparison		Error
Split	MAE	N	MAE	N	MAE	Std. Error	Best	t-value	p-value	Improvement
Cluster 0										
80:20	10.32	1774	12.97	1462	10.63	0.158	9.74	−1.962	0.0323	−3%
65:35	10.92	1334	11.32	1048	10.98	0.068	10.98	−0.882	0.1943	−1%
50:50	16.48	1011	11.29	758	11.63	0.073	11.63	66.438	3E-24	29%

Note: The reported MAE and standard error for ANN are based on 20 replications of the experiment with 100 training epochs each The value reported under best results is the mean MAE of 20 replications with an optimal number of training epochs.

TABLE 4b

MAE Comparisons Between ANN and MNL for Peanut Butter

Data	MNL		ANN			Comparison		Error
Split	MAE	N	MAE	Std. Error	Best Result	t-value	p-value	Improvement
Cluster 1								
80:20	11.12	341	10.87	0.047	10.72	5.319	1.96E-05	2%
65:35	11.21	319	11.48	0.039	11.46	−6.923	6.69E-07	−2%
50:50	11.39	224	12.35	0.116	11.47	−8.276	5.05E-08	−8%
Cluster 2								
80:20	9.12	1554	8.11	0.035	8.04	28.857	1.86E-17	11%
65:35	9.07	1161	7.84	0.036	7.74	34.167	7.98E-19	14%
50:50	10.86	830	7.99	0.037	7.95	77.568	1.54E-25	26%
Cluster 3								
80:20	14.09	1009	11.4	0.094	11.21	28.617	2.18E-17	19%
65:35	13.55	787	11.25	0.045	11.25	51.111	4.11E-22	17%
50:50	11.32	587	12.63	0.156	11.5	−8.397	4.05E-08	−12%

Data	MNL		MNL with Brand Loyalty		ANN			Comparison		Error
Split	MAE	N	MAE	N	MAE	Std. Error	Best	t-value	p-value	Improvement
Cluster 0										
80:20	7.02	2994	6.48	2534	6.02	0.027	6	37.037	2E-19	14%
65:35	9.10	2267	7.70	1848	5.96	0.044	5.96	71.364	7E-25	35%
50:50	11.54	1641	11.89	1325	6.33	0.040	6.2	130.25	8E-24	45%

Note: The reported MAE and standard error for ANN are based on 20 replications of the experiment with 100 training epochs each The value reported under best results is the mean MAE of 20 replications with an optimal number of training epochs.

The MNL and ANN forecasts in weeks 621 and 622 turn out to be:

Week	Actual Share	MNL Forecast	ANN Forecast
621	0.26	0.259	0.265
622	0.30	0.270	0.297

Clearly in this example both approaches give a comparable and rather good brand share forecast. In this section we organize our discussion of results into three parts. First we discuss forecasting ability across the three categories, next forecasting ability across clusters within a category, and lastly forecasting ability across data partitions within a cluster and a category.

Forecasting Errors Across Categories

The results indicate that the ANN performs significantly better than the MNL in dishwashing liquid and peanut butter categories, and moderately better in the catsup category.

TABLE 4c

MAE Comparisons Between ANN and MNL for Dishwashing Liquid

Data	MNL		ANN			Comparison		Error
Split	MAE	N	MAE	Std. Error	Best Result	t-value	p-value	Improvement
Cluster 1								
80:20	18.18	174	10.05	0.208	9.99	39.087	6.41E-20	45%
65:35	19.39	139	10.36	0.190	9.70	47.526	1.62E-21	47%
50:50	18.35	106	9.67	0.232	9.56	37.414	1.46E-19	47%
Cluster 2								
80:20	13.06	616	10.89	0.052	10.89	41.731	1.87E-20	17%
65:35	13.16	511	10.57	0.041	10.48	63.171	7.51E-24	20%
50:50	13.04	394	10.03	0.113	9.98	26.637	8.24E-17	23%
Cluster 3								
80:20	12.43	1216	9.19	0.051	9.11	63.529	6.75E-24	26%
65:35	12.51	907	8.69	0.082	8.69	46.585	2.36E-21	31%
50:50	13.49	658	8.86	0.068	8.86	68.088	1.82E-24	34%

Data	MNL		MNL with Brand Loyalty		ANN			Comparison		Error
Split	MAE	N	MAE	N	MAE	Std. Error	Best	t-value	p-value	Improvement
Cluster 0										
80:20	10.57	2006	10.87	1893	7.58	0.098	7.53	30.510	7E-18	28%
65:35	10.87	1557	10.98	1436	7.5	0.123	7.26	27.398	5E-17	31%
50:50	11.29	1158	12.66	1047	7.34	0.112	7.22	35.268	4E-19	35%

Note: The reported MAE and standard error for ANN are based on 20 replications of the experiment with 100 training epochs each The value reported under best results is the mean MAE of 20 replications with an optimal number of training epochs.

We calculate how much better is ANN than the MNL in forecasting ability by the following measure which we call as an error improvement measure:

$$\text{Error improvement by ANN} = \left(\frac{MAE_{MNL} - MAE_{ANN}}{MAE_{MNL}}\right) \times 100 \tag{17}$$

where,

MAE_{MNL} = mean absolute error obtained from the MNL, and
MAE_{ANN} = mean absolute error obtained from the ANN.

A positive error improvement means that the ANN performed better than the MNL and a negative error improvement means it performed worse than the MNL. We further conduct a *t*-test to test the statistical significance of the two MAE estimates. The null and alternate hypotheses can be stated as:

$$\mathbf{H_0}: \quad MAE_{ANN} \geq MAE^*_{MNL}$$

$$\mathbf{H_a}: \quad MAE_{ANN} < MAE^*_{MNL}.$$

where MAE_{ANN} represents a random variable and MAE^*_{MNL} represents a constant for the purpose of the test. Define $t = MAE_{ANN} - MAE^*_{MNL} / S_{MAE_{ANN}}$, where $S_{MAE_{ANN}}$ is the standard error of MAE_{ANN}. If $t > t_{critical}$ (one-tailed) then we can reject $\mathbf{H_0}$.

The results in Table 4(a)-4(c) indicate that MAE does not decrease in six out of the nine cases with the inclusion of loyalty variable in MNL model across the three categories. The MAE decreases only for 80:20 and 65:35 data partitions in Peanut butter category and in 50:50 partition in Catsup category. In rest 6 cases, the MAE in fact increases. This increase may be due to reduction in number of observations in the estimation of MNL model with loyalty. Even in the three cases where MAE decreases with inclusion of loyalty variable, the MAE does not decrease below the MAE obtained from ANN. Thus the gains obtained in the estimates by inclusion of loyalty variable are seemingly offset by the loss in forecasting ability due to fewer degrees of freedom. Since MNL without loyalty does better than MNL with loyalty, we compare it with ANN in all subsequent discussion.

As can be seen from Tables 4(a), 4(b), and 4(c) the forecasting performance of ANN is particularly stronger in the dishwashing liquid category where ANN outperforms MNL in all 12 cases (statistically significant at the .01 level) and the forecasting error improvement ranges from 17% to 47%. In the peanut butter category ANN outperforms MNL in all but three cases. The error improvement by ANN ranges from 11% to 45% in 9 cases, whereas MNL improves the forecast in the range of 2-12% in 3 cases. The results are weaker in catsup category for ANN where although it does significantly better than the MNL in cluster 1, it gives similar and sometimes even worse forecasting error than the MNL in other clusters. In this category MNL outperforms ANN in 8 out of 12 cases (again at significance level of .01), however, the error improvement by MNL is in the range of 1% to 18% only, compared to improvement of 29% to 87% by ANN in 4 cases (ANN and MNL perform equally in 2 cases at the .01 significance level).

The differences in the performance of the two approaches across the three categories can perhaps be explained by the complexity of the choice problem reflected in the number of available brands. Dishwashing liquid has 11 brands compared to 6 in the peanut butter category, and 4 in the catsup category. The choice problem is thus more complex to model in dishwashing liquid case, followed by peanut butter, and then catsup. This leads to the following proposition:

P1: *Artificial neural network performs better than multinomial logit model in forecasting brand shares when the choice problem is complex for the forecasting model, such as when the number of brands in the category are numerous.*

Forecasting Errors Across Clusters

Cluster 1 contains households which buy only one brand, i.e., strongly brand loyal households who buy only one brand regardless of store environment. In this cluster the store envi-

ronment does not have good predictive power because the households continue to buy their favorite brand regardless of store environment. In this situation we find that the ANN does better than the MNL (results are particularly strong in catsup category), indicating that ANN is better able to recognize the pattern in brand shares and discount the effect of input store environment in this case. This leads to our second proposition:

> **P2**: *Artificial neural network is better able to recognize the statistically negligible effect of input variables on the choice outcome for the strongly loyal customers, and is thus better able to discern the overall data pattern for them.*

Cluster 2 contains households who purchase two or three different brands. In this cluster the store environment should have some impact on brand choice. The MNL forecasting errors are lower for this cluster than cluster 1 in all three categories indicating better predictive power of the store environment variables. In contrast, the forecasting ability of ANN diminishes slightly in dishwashing liquid case, dramatically in catsup case, but improves in peanut butter case. In cluster 2, ANN does better than MNL in both dishwashing liquid and peanut butter and slightly worse in catsup.

Cluster 3 contains households who buy 4 or more brands. Depending on the number of available brands this type of brand switching behavior can be interpreted in different ways. If the number of available brands is small (such as in catsup category), it is possible to interpret it more as a variety seeking behavior, because in this cluster store environment alone does not explain the choice behavior of this group of households very well. On the other hand, if the number of available brands is large then it can be interpreted more as a reaction to store environment, because switching among 4 or more brands from a large set of available brands may be driven by price and merchandising. Thus in catsup category, with only four brands, this type of switching may be interpreted as variety seeking behavior, whereas in dishwashing liquid, with eleven brands, it can be more of a store environment explanation.

Under extreme switching behavior (such as buying 4 out of 4 available brands in catsup category) we would not expect MNL to forecast better in *cluster 3* than *cluster 2*. Also, unlike *cluster 1*, data patterns are less recognizable in this cluster which should result in deterioration of ANN's performance compared to that in *cluster 2*. These effects should also be moderately present in peanut butter category. The results in Tables 4(a) and 4(b) support this reasoning.

However in *cluster 3* of dishwashing liquid, with 11 brands, switching behavior is not necessarily extreme and the effect of store environment on brand switching is expected to be stronger. Accordingly we would expect better performance from MNL, and since data patterns are not driven by extreme switching behavior, better performance from ANN also, compared to *cluster 2*. The results in Table 4(c) support this reasoning.

The results with respect to *clusters 2 and 3* combined indicate that under extreme switching behavior (*clusters 2 and 3* in catsup), although performance of both MNL and ANN declines, the MNL performs relatively better than the ANN, whereas under normal switching (*clusters 2 and 3* in peanut butter and dishwashing liquid categories), the ANN outper-

forms the MNL (with the exception of peanut butter, *cluster 3*, 50-50 data partition). This leads to our third proposition:

> **P3**: *Under normal switching conditions, the artificial neural network outperforms the multinomial logit model but under extreme brand switching (due to variety seeking, for example) multinomial logit model outperforms artificial neural network in forecasting ability.*

One reason for relatively better performance of MNL over ANN under extreme brand switching could be that brand specific constants in the utility function in the MNL are better able to capture the variety seeking behavior than what ANN is able to do with the underlying data patterns.

As a base case if no clustering was imposed on the households and all were treated alike, we would get *cluster 0* estimates. We find that MNL performs relatively better in this cluster than it does in other clusters. One reason for this could simply be that number of observations used in estimation are much greater in this cluster.

Not surprisingly, with the exception of *cluster 0* versus *cluster 1* of catsup, ANN performs significantly better in this cluster than it does in other clusters. This could also be driven by availability of more training data which yields better connection weights which forecast with greater accuracy. What is interesting however is that ANN outperforms MNL in almost all cases (except 80:20 partition of catsup) across the three categories in this cluster. Thus ANN forecasts brand shares better than MNL when household heterogeneity is not explicitly considered leading to our fourth proposition:

> **P4**: *Artificial neural network forecasts better than multinomial logit when household heterogeneity is not explicitly considered.*

For the three categories ANN, in fact, performs best (gives lowest) when household heterogeneity is not explicitly considered and all households are considered alike in training and estimation. Thus one tentative conclusion is that while using ANN it is not important to cluster households in to homogenous segments. Since the number of observations is largest for *cluster 0*, this finding seems to be driven by the availability of more data for ANN training.

Forecasting Errors Across Data Partitions

The effect of having more data to estimate the MNL model is apparent in the results. Forecasting error declines across the data partitions in all clusters and all three categories as more observations are used to estimate the MNL (there are three exceptions though, *cluster 3* of peanut butter and catsup, and *cluster 1* of dishwashing liquid). The pattern is less clear in the case of ANN. For peanut butter case, there is no apparent pattern. In catsup the forecasting error decreases as more data are used to train the network but in dishwashing

liquid case the error actually increases as more data are used to train the network. This leads us to our final proposition:

> **P5**: *Artificial neural network is less sensitive to the number of observations used to train the network compared to the multinomial logit model which provides better forecasts if more data are used to estimate the model.*

In summary our comparison of forecasting ability of ANN and MNL indicates at least five patterns which we state as propositions above. These empirical patterns need to be further tested and validated in subsequent research studies.

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

The results in this paper indicate usefulness of neural networks in forecasting brand shares for a retailer in frequently purchased consumer goods categories. We used three categories, namely, peanut butter, dishwashing liquid, and catsup, and found that neural network performs relatively better than multinomial logit model approach in majority of cases.

The overall better performance of ANN could be related to the assertion that ANN is better able to handle non-linearities in the data (for example, Hruschka, 1993). A scanner dataset, especially with a large number of available brands, may contain significant non-linearities which ANN perhaps picks up well. Figure 4 summarizes the error improvements by ANN over MNL in the cases considered here. We compared a total of 36 cases (3 categories x 4 clusters x 3 data partitions). Out of 36, ANN did better in 25 cases, almost as well in 2 cases, and worse in 9 cases.[7] However the worse performance of ANN ranged between 1-18% with a mean of 7% compared to an improvement in the range of 2-87% with a mean of 34%.

Despite the seemingly better performance of ANN, it is noteworthy that MNL coefficients are directly interpretable as response elasticities whereas ANN coefficients (connec-

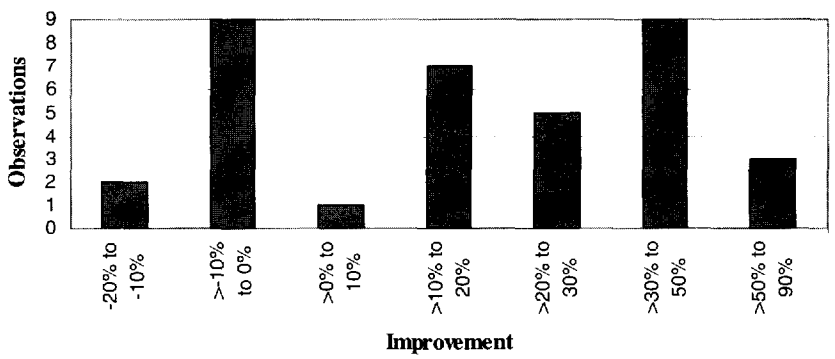

Figure 4. **Error Improvement by ANN over MNL**

tion weights and biases) are not directly interpretable. For example, the MNL estimates in Table 2 are response coefficients of price, feature, and display, whereas ANN connection weights and biases in Table 3 can not be meaningfully interpreted. Thus, unlike MNL, ANN provides no insight into the causes of the outcomes.

It is theoretically possible to use ANN to obtain a rough estimate of the response elasticities. A user can input several store environment scenarios and get a forecast for each. These forecasts can provide a rough estimate of the response elasticity. For example, feature and display can be kept constant at 0.01 and price can be varied.[8] The predicted brand shares can then be used to estimate the price elasticity. However the resulting elasticities are not as reliable as those obtained from MNL.

Another important factor to recognize is that neural network approach is useful only at an aggregate level, given the amount of data per household typically available in scanner data panels. ANN requires a sufficiently long purchase history (running into hundreds of purchases per household) to properly train the network for an individual household. This limitation also applies to MNL but MNL can model desegregate choice behavior quite satisfactorily by incorporating household heterogeneity measure. Thus ANN is useful only when we are interested in forecasting market level brand shares. The MNL, on the other hand, can be useful for forecasting both market level brand shares and disaggregate household choice probabilities.

It is also important to recognize that there is some amount of data preparation required for both ANN and MNL. For example, brand shares need to be computed for both approaches. There is some effort involved in configuring the ANN, particularly in selecting an optimal number of nodes and training epochs. This also applies to MNL where the utility model has to be chosen. Beyond these efforts ANN is much easier to work with and takes less effort and time to produce a forecast.

Thus one advantage of ANN is that it requires less effort than MNL in terms of data preparation and analytical effort. We spent on an average four times as many hours on MNL as on ANN in this study.[9] Both models were estimated on a mainframe computer. ANN training took 2 minutes on average and forecasting only a few seconds per case. In contrast, MNL estimation and forecasting took approximately 30 minutes per case because of the several steps involved. For ANN estimation we used default procedures and options in the Neural Networks Toolbox in MATLAB software, and for MNL estimation used FORTRAN software and GQOPT optimization subroutines.

Our findings are similar to Kumar et al. (1995) who find ANN to produce a better prediction rate but the traditional econometric method (logistic regression) to be a superior method in terms of interpretability. Similarly the tradeoff between ANN and MNL is a tradeoff between need for a quick, accurate brand share forecast provided by ANN and need for causal insights provided by response coefficients of MNL model.

A retailer armed with an accurate brand share forecast can better manage inventory, better plan merchandising activities, and even better negotiate with the manufacturers. For the retailer, whose chief concerns include category volume and category profits, accurate brand share forecasts are clearly equally (if not more) important than disaggregate household level choice probabilities.

We hope that future studies would test the five propositions stated in the paper, and similar comparisons between ANN and other econometric methods would be undertaken for retailing and other decision areas.

NOTES

1. Other polychotomous models such as multinomial probit become difficult to implement in presence of more than four choice alternatives (Maddala, 1983).
2. Also note that a retailer may not be much interested in household specific choice behavior. Instead a segment level analysis such as the one proposed here may be more relevant from a retailer's viewpoint.
3. An interaction term between price and feature or display was also included in the MNL model as it increased the log likelihood ratio indicating better fit.
4. The normalization was done using the formula [{(Pr$_{...}$-MinPrice)/(MaxPrice-MinPrice)} $\times$ 0.8 + 0.1] where MinPrice is the least price and MaxPrice is the highest price charged in the category dataset.
5. We included lagged store environment also as input variables but this did not increase the forecasting ability of the ANN. Therefore we omitted these lagged variables from further analysis.
6. The ANN forecast was rescaled such that brand shares summed to 1.
7. This applies to ANN in which 100 epochs are used in training. However when optimal number of training epochs are used, the ANN performs equivalent to MNL in three out of these nine "worse" cases.
8. The response elasticity should be estimated only in the vicinity where ANN is trained. For example, if the average price is 0.5 in training period then we can vary price from 0.4 to 0.6 to estimate price elasticity.
9. This was after steady state, i.e. after the initial learning during which all programs and the sequences were set up. The initial learning and setting up of MNL forecasts consumed in excess of 60 hours including time spent in fixing bugs. In contrast, ANN preparation and initial learning took less than a quarter of this time, i.e. less than, 15 hours.

REFERENCES

Agrawal, Deepak. (1996). "Effect of Brand Loyalty on Advertising and Trade Promotions: A Game Theoretic Analysis with Empirical Evidence," *Marketing Science*, **15**(Winter): 86-108.

Belt, Debbie. (1993). "Neural Networks: Practical Retail Applications," *Discount Merchandiser*, (October): 9-11.

Chauvin, Y. and D.E. Rumelhart (eds.). (1995). *Backpropagation: Theory, Architectures, and Applications*. Hillsdale, NJ: Erlbaum .

Demuth, Howard and Mark Beale. (1992). *Neural Network Toolbox User's Guide*. South Natick, MA: The MathWorks Inc.

Dragstedt, Carl. (1991). "Shopping in the Year 2000: Neural Net Technology is the Brain of Retail's Future," *Discount Merchandiser Technology*, Supplement, (September): 37-40.

Federowicz, Alex. (1994). "An Alternate View on Neural Networks," *Direct Marketing News*, July 25, **16**(28): 23, 51.

Gensch, Dennis H. and Wilfred W. Recker. (1979). "The Multinomial, Multiattribute Logit Choice Model," *Journal of Marketing Research*, **16**(February): 124-132.

Green, Paul E., Frank J. Carmone and David P. Wachspress. (1977). "On the Analysis of Qualitative Data in Marketing Research," *Journal of Marketing Research*, **14**(February): 52-59.

Hornik, K., M. Stinchcombe and H. White. (1989). "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks*, **2**: 359-366.

Horowitz, Joel L. and Jordan J. Louviere. (1995). "What is the Role of Consideration Sets in Choice Modeling," *International Journal of Research in Marketing*, *12*: 39-54.

Hruschka, Harald. (1993). "Determining Market Response Functions by Neural Network Modeling: A Comparison to Econometric Techniques," *European Journal of Operational Research*, **66**: 27-35.

Huang, S.H. and H.C. Zhang. (1994). "Artificial Neural Networks in Manufacturing: Concepts, Applications, and Perspectives," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, Part A, **17**(2): 212-228.

Kumar, Akhil, Vithala R. Rao and Harsh Soni. (1995). "An Empirical Comparison of Neural Network and Logistic Regression Models," *Marketing Letters*, **6**(4): 251-263.

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge, MA: Cambridge University Press.

Makridakis, Spyros, Steven C. Wheelwright and Victor E. McGee. (1983). *Forecasting: Methods and Applications*. New York: John Wiley and Sons.

Malhotra, Naresh K. (1984). "The Use of Linear Logit Models in Marketing Research," *Journal of Marketing Research*, **21**(February): 20-31.

Masson, Egill and Yih-Jeou Wang. (1990). "Introduction to Computation and Learning in Artificial Neural Networks," *European Journal of Operational Research*, **47**: 1-28.

McAlister, Leigh and Edgar A. Pessemier. (1982). "Varied Consumer behavior: An Interdisciplinary Review," *Journal of Consumer Research*, **9**(December): 311-322.

Rumelhart, D.E., G.E. Hinton and R.J.Williams. (1986). "Learning Internal Representations by Error Propagation." Pp. 318-362 in *Parallel Distributed Processing, Explorations into the Microstructure of Cognition*,Vol. 1, D.E. Rumelhart and J.L.McClelland (eds.). Cambridge, MA: MIT Press.

Shepard, David and Bruce Ratner. (1994). "Using Neural Nets with EDA and Regression Models," *Direct Marketing News*, May 23, **16**(20): 27, 79.

Srinivasan V. and Thomas Kibarian. (1989). *Purchase Event Feedback: Fact or Fiction.* Unpublished paper, Graduate School of Business, Stanford University, February.

Thall, Neil. (1992). "Neural Forecasts: A Retail Sales Booster," *Discount Merchandiser*, **32**(10): 41-42.

Thiel, Henri. (1969). "A Multinomial Extension of Linear Logit Model," *International Economic Review*, **10**(October): 251-259.

Venugopal, V. and W. Baets. (1994). "Neural Networks and their Applications in Marketing Management," *Journal of Systems Management*, (September): 16-21.

Vogl, T.P., J.K. Mangis, A.K. Rigler, W.T. Zink and D.L. Alkon. (1988). "Accelerating the Convergence of the Back-Propagation Method," *Biological Cybernetics*, **59**: 257-263.

Wasserman, P.D. (1989). *Neural Computing: Theory and Practice*. New York: Van Nostrand Reinhold.

White, Halbert et. al. (1992). *Artificial Neural Networks: Approximation and Learning Theory*. Cambridge, MA: Blackwell Publishers.