Using Multivariate Adaptive Regression Splines (MARS) to Find Interactions of Socio-Demographics That Model Individual Differences in Australian Farmers Purchase Behavior

ABSTRACT

Socio-demographics play a major role in accounting for preference heterogeneity and market segmentation in discrete choice models. The use of demographic segments to account for heterogeneity in choice models has been proposed by Ben-Akiva & Lerman (1985) and complex models such as random coefficients logit have been used to account for unobserved differences in preferences. To enter demographics into a repeated choice stated preference model, they must be interacted but, due to the complexity of finding and modelling socio-demographic interactions (McLelland & Judd 1993), the interactions are often restricted to simple terms that act global over the data space. We use MARS to overcome the difficulties associated with detecting and integrating socio-demographic interactions in localized areas of the data space. In our study, heterogeneity that exists amongst farmers can be accounted for by localized interactions of the observable demographics with the experimentally designed choice attributes using basis function found by MARS. The MARS basis functions are hybrid into a conditional logit model that outperforms a hybrid of the MARS basis functions in a random coefficients logit.

1. INTRODUCTION

In 2006, the Australian Research Council (ARC) funded a study on the purchase behaviour of Australia's farming community. To ascertain farmers' retail purchasing preferences, the University of Technology, Sydney (UTS) administered a stated preference (SP) survey to a group of 2414 Australian farmers. The farmers were presented with binary choice sets, wherein each set had associated 15 attributes. They were shown 16 choice experiments and asked, based on the attributes presented, which of the two alternatives they would purchase. The primary aim of the study was to identify a valid basis for local segmentation and target marketing of Australia's farming community, based on their stated preference choice behaviour and individual characteristics. The study also sought to develop and test a staged approach for inferring market segments and interactions by utilising a hybrid discrete choice and data mining method to simultaneously account for interactions between individual characteristics and choice behaviour.

Due to geographic and specific farm type factors, we assumed the presence of localised effects but these effects were not known a priori. Given that the dependent variable in our farming data was binary, Rust & Donthu (1995) suggest combining a logistic regression with a kernel density estimation to account for localised differences. In order for kernel density estimation to work well, the choice of the kernel needs to be known a priori. Kernel density can be inflexible in that it does not allow for easy estimation of non-metric data (Breiman et al. 1984). Although we could look for other ways to detect local differences, an additional dimension of this study is to account for heterogeneity in individual farmer's preferences. A popular technique for the analysis of unobserved heterogeneity is a mixed logit model (McFadden & Train 2000). On the other hand, Allenby, Arora and Ginter (1998) argue that in a logit mixture model, heterogeneity is difficult to find because a major source of heterogeneity is in the consideration set.

While the consideration set as a source of heterogeneity may be dominating in a revealed preference (RP) study, this does not apply to a study based on stated preference (McFadden 2002). The reason why stated preference performs better is that RP data does not control for a respondent's choice set formation (Louviere, Hensher & Swait 2000). Before introducing a new product to market, SP is a marketer's only choice, since no revealed preference/purchase data exists. The use of a stated preference study makes it easier to segment and target customers when the structure of observed heterogeneity is known, especially in a pricing variable (Bell et al. 1998). In our study, the pricing variable was experimentally designed into the stated preference study. Previous studies (e.g., Kamakura & Russell 1989) showed that price elasticity can be an effective segmentation variable. There is no reason to believe that heterogeneity cannot exist in the non-intercept terms of a model (Allenby & Rossi 1998). Since the attributes used in the study were provided by farming focus groups of which we had no prior domain knowledge, we assumed that all experimentally designed attributes used in our study were potential sources of heterogeneity. Correctly integrating demographics into a choice model will better account for individual sources of heterogeneity (Ben-Akiva & Lerman 1985).

This study, as in the previous essay, follows the premise that omission of a relevant term from a utility function will yield heterogeneity (Swait & Louveire 1993; McFadden 1986). One relevant variable that was not initially used in the UTS rural choice study was "distance to retailer". Since this variable was observed and relevant to the model, the omission of the distance covariate could introduce heterogeneity into the model. Inclusion of this variable is important in a choice model as it can account for unobserved effects (Bronnenberg 2005).

We will show that by including the proper terms in the choice model, we can correct any heterogeneity that is present but unaccounted for. If there is large taste heterogeneity, taking into account interactions between demographics and attributes of the alternatives may be necessary (Brownstone, Bunch & Train 2000). However, one of the challenges in integrating interactions into the model is that the structure of these interactions may not be known a priori and are difficult to detect if they are localised and have a small sample size (McClelland & Judd 1993). Further complicating our choice model is the fact that we need to account for segment and interactive differences while simultaneously investigating ways to reduce heterogeneity in the model. In response to these difficulties, we estimated the farming data using a spline regression data mining technique called MARS. Although the MARS model can be used as a stand-alone model, in this essay, the data transformations and interactions found by MARS were subsequently estimated in a multinomial logit model (MNL). We have chosen this method because the hybrid of MARS-MNL is the best way to ascertain the strength of interactions, recover the standard errors for the parameters, and make accurate comparisons to other logistic regression methods (Zabaleta et al. 2008). Although MARS's predecessor, CART, is very good at interaction detection, it does not represent strong linear structure effectively. CART also lacks a smooth prediction surface, making it "rough" in a regression setting. Moreover, CART has an issue in modelling strongly additive and linear structure (Hastie, Tibshirani & Friedman 2001). The experimentally designed farming data used in this essay has strong linear structure, and the models used GIS data, for which MARS is better suited (Munoz & Felicismo 2004).

The purpose of the essay was not only to conduct a simple MARS model estimation and interpretation, but instead to look at the overall model performance and explanation of heterogeneity. MARS is used in this essay to investigate how a MARS model accounts for heterogeneity by correctly determining the covariate interactions within the data structure. An additional advantage of using MARS is that, if needed by the marketing manager, the MARS basis functions are transparent and easy to interpret. To demonstrate this, Appendix 7 shows the complete MARS basis function output of all the attributes used in the rural choice study.

This essay uses a hybrid MARS-MNL model to correctly identify interactivity between experimentally designed and demographic variables. As a comparison, it also models a mixed logit (MIXL), and shows that the MARS-MNL model hybrid accounts for all observed as well as unobserved heterogeneity. A shortcoming of mixed logit is that it can verify the existence of heterogeneity but it cannot denote the exact structure of the heterogeneity. Some recent attempts to perform a posterior analysis of mixed logit have been proposed (Hess 2007), but a model that simultaneously accounts for heterogeneity and other issues is optimal (Bhat & Guo 2004). By including the proper attributed interaction and transformations, we show that the MARS-MNL hybrid model will significantly help the marketing manager in that it increases the predictive strength of farming choice data, resulting in better targeting of rural customers.

The remainder of this essay is organised as follows:

Section 2 describes the dataset and looks at some previous studies that explored the correct form of socio-demographics in a choice model. Section 2 also describes the MARS algorithm;

Section 3 explains the set-up of the datasets as well as the use of MARS and mixed logit to investigate heterogeneity in the model;

Section 4 is a simulation of known spline basis function interactions and tests if MARS finds these basis functions in the presence of slightly noisy data;

Section 5 examines and analyses the results of MARS, Mixed Logit, and the hybrid of MARS with a choice model and mixed logit;

Section 6 summarises our findings.

2. DATA DESCRIPTION

2.1 Description of Data

In mid-2006, we pre-screened all postcodes in Australia using the Australian Bureau of Statistics (ABS) farming data to see which postcodes would be suited to a study on rural purchase behaviour. Since we were interested in specific types of farming projects with sufficient numbers of respondents, postcodes in the central area of Australia, as well as urban centres, were excluded. Initially, we planned to perform a web-based survey but the percentage of farmers in rural Australia with internet access was low at the time of the survey. After determining the postcodes to be used, potential respondents were recruited for the UTS study by a research agency specialising in rural respondents. Starting in late 2006, the research agency called the potential respondents and conducted a CATI interview to see if they were amenable to being part of a rural research study. During the CATI interview, potential respondents were asked about their awareness of specific farm products and rural retailers to determine their appropriateness to be included in the study. Suitable respondents were asked if they would like to complete a mail survey and receive a $50 voucher upon completion. In the end paper questionnaires were mailed to a total of 4,662 CATI respondents.

The mail questionnaire consisted of a best/worst preference task of farming and rural merchandising, ratings scale opinions of farming practices and products, and 16 stated preference choice questions. The stated preference questions used 14 attributes with binary and multiple levels. Initially, the retailer's brand name was collected as the choice (dependent) variable, but brand name was actually an attribute of the alternative, so it was included as a 15th attribute in the analysis. The resulting dataset was set up with the dependent variable being which of the two alternatives the respondent would choose to buy. There were 48 versions of the stated preference survey to control for version effects, if any. A 49th version of the questionnaire was sent to randomly selected respondents to ascertain preferences for farmers going directly to a supplier for products instead of through a rural retailer. At the end of each mail questionnaire demographic information was collected. Of the 4,662 questionnaires mailed out, 2,414 were returned with complete information before the deadline of mid-2007. A sample of one of the survey choice tasks can be found in Appendix 5.

To increase the precision of our model, we used an orthogonal OMEP designed stated preference data. Repeated observations were made on each responding farmer, as multiple observations are needed in a mixed logit to account for unobserved variation in the model (Bhat 2000). To find the variables that pertain to all rural choice scenarios is extremely difficult as some variables may be available in some localities but not in others. Linkages in farming purchase behaviour are difficult since there are many omitted variables in most farming models (Roe & Stockberger 2004). The stated preference questions used in the study came from focus groups conducted by UTS-CenSoc. The most prominent features that came up during the focus groups were used as attributes in the choice model and the nature of the stated preference study forces the respondents to trade-off only the attributes that are presented in their survey.

Distance from a supplier is an important covariate in a respondent's purchase decision. Until recently, geographic information systems (GIS) data has been difficult and costly to obtain. The availability of commercial software (e.g., MapInfo, Maptitude, Esri) has made calculation of distance from respondent to supplier easier to obtain. Using distance in an urban setting is valuable in retailer purchase decisions but only mildly predictive in a choice model since competing retailers are within similar distances. In a rural setting, distance from retailer carries much more significance and some respondents in our study travelled in excess of 150km to purchase their supplies. Distance was not asked in the survey and had to be calculated from the GIS coordinates of the respondent's farm to the rural retailer they told us they had purchased from. The 15 experimentally designed attributes of purchase and the distance from retailer attribute are listed in Table 16. The mixture of data scales makes analysis difficult and the presence of categorical data complicates the data space (Breiman 2001). To help alleviate any issues, any non-numeric data was effects coded with the highest value of the variable set as the effects code base level.

TABLE 1: SUMMARY STATISTICS OF RURAL CHOICE SURVEY ATTRIBUTES

Scale	Variable	Value	Label
Nominal	(ATTR1) Retailer	1	CRT
		2	Landmark
		3	Elders
Ordinal	(ATTR2) Staff product knowledge	1	No real product knowledge
		2	Limited product knowledge
		3	Moderate product knowledge
		4	Extensive product knowledge
Binary	(ATTR3) Staff professionalism	1	Not consistently professional in appearance & manner
		2	Consistently professional in appearance & manner
Binary	(ATTR4) Independence of advice	1	Unsure whether advice may be biased
		2	Trusted to provide unbiased advice
Ordinal	(ATTR5) Opening days	1	5 days only
		2	5.5 days (close at midday Sat)
		3	6 (closed Sun)
		4	7 days
Ordinal	(ATTR6) Opening hours	1	8am to 5pm
		2	7am to 5pm
		3	7am to 7pm
		4	7am to 9pm
Binary	(ATTR7) Store presentation	1	No investment in store presentation
		2	Significant investment in store presentation
Binary	(ATTR8) Store branding	1	No external store branding
		2	Easily recognized external store branding
Binary	(ATTR9) Product range	1	Limited range of brands
		2	Wide range of brands
Binary	(ATTR10) Stock availability	1	Often stock has to be ordered in
		2	Stock nearly always available
Binary	(ATTR11) On farm delivery	1	No free delivery
		2	Free delivery
Ordinal	(ATTR12) Professional advisory service	1	No on-farm professional advisory service
		2	On farm advice paid for in product margin
		3	On farm advice paid for as separate fee
		4	Free on-farm professional advisory service
Ordinal	(ATTR13) Payment terms	1	1.2% discount for < 30 days
		2	30 days
		3	60 days
		4	90 days
Binary	(ATTR14) Late payment fee	1	Late payment fee
		2	No late payment fee
Ratio	(ATTR15) Price	-0.15	15% less
		-0.10	10% less
		-0.05	5% less
		-0.02	2% less
		0.00	About Average
		0.02	2% more
		0.05	5% more

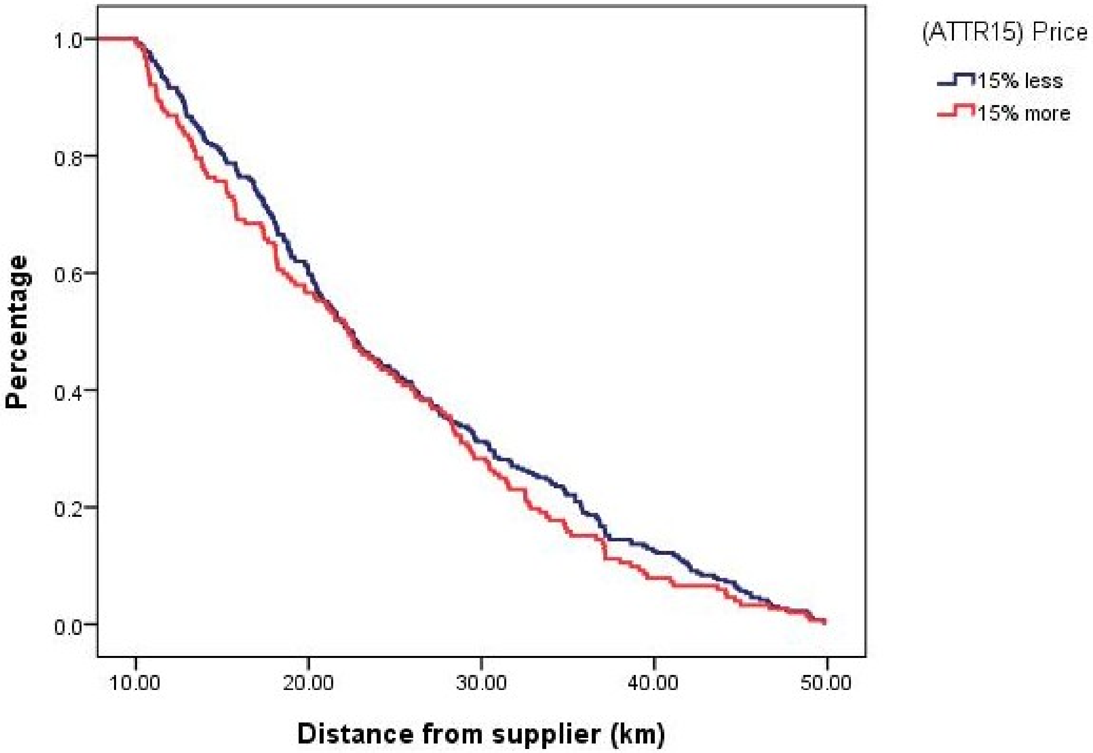
		0.10	10% more
		0.15	15% more
Ratio	(Distance) Distance from Retailer	-	-

2.2 Previous findings of geographic-demographics on choice

In retailer choice, where to shop can explain up to 70% of the variance in a model (Bell et al. 1998). Bayer and Timmins (2005) mention the fact that demographic interactions tend to act well in geographically localised regions. Previous models of farming choice behaviour have not investigated non-linearity nor localised interactive behaviour as it is difficult to ascertain which geographic regions result in different parameter estimates (Rust & Donthu 1995). Much of the literature using distance for retailer purchase behaviour is aggregated at the county or postcode level (Fisher & Hanemann 1998). For example, Archer & Lonsdale (1997) used a one-way ANOVA analysis of value per hectare (a proxy for income) of farmland by postal code and found significant differences between postcodes. They also investigated a two-way ANOVA, which showed a high degree of interactivity, but this analysis was not pursued further.

As the size of a farm gets larger, the likelihood of a farmer buying locally from a retailer diminishes (Roe & Stockberger 2004). Since there is likely to be localised utility maximisation when using spatial data (Xue & Brown 2002), using "distance to retailer" as one of our attributes allowed us to geographically disaggregate the data to increase the precision of our estimates. Figure 11 shows an empirical distribution of purchase probability against distance from rural retailer in our data. This graph can be thought of as demand for a retailer as a function of distance from the retailer. If standard gravity models are correct, creating different strata based on different price discounts should create a series of parallel lines (Sullivan 1990). As can be seen in Figure 11, the lines are not parallel, but cross at 23 km from the retailer. This graph clearly implies that other variables are moderating the effect of price over distance, or there is heterogeneity in the distance variables which needs to be accounted for. Although the difference in pricing discounts as a function of distance is never more than 8%, the point of Figure 11 is to graphically show that the pricing levels, as a function of distance, are not parallel, as they should be if there was no interaction. An extremely large random coefficients model might be needed to account for spatial heterogeneity; hence, a parsimonious way to model this heterogeneity is warranted (Bhat & Guo 2004).

FIGURE 1: EFFECTS OF PRICE DISCOUNT ON DISTANCE FROM RETAILER



2.3 Introduction to MARS

MARS is a non-parametric regression technique that builds flexible regression models using splines. MARS creates splines for all variables and data points, then finds the best group of splines based on variance-bias tradeoffs. The MARS algorithm creates a series of 'basis functions' (BFs), which are regression splines defined between points in the data called knots. At each knot point '*t*' in the data, MARS creates two basis functions (X_k -t) if X_k >t, 0, otherwise *and* one for (t- X_k) when X_k<t, 0 otherwise. MARS looks at every knot split point *t* and every variable X_k, so there are 2*X_k*t possible basis functions. MARS only has two scales: ordinal and categorical. All non-categorical data is treated as ordinal so any split point from a continuous variable will come from a point in the dataset or between any two points. MARS creates internal mutually exclusive dummy variables for categorical variables just as any standard regression technique. However, the basis functions for categorical variables in MARS are created differently than standard regression dummy variables in that the basis functions are a collection of the categorical levels. For instance, a categorical variable with levels A, B, C, D, and E could result in a MARS basis function that contains "A, B and D" or "B and E". In our analysis, we do not discuss how MARS handles categorical data since all categorical data is effects coded in the rural choice dataset to ensure accurate comparisons across modelling techniques. MARS is a forward stepping linear regression which has the functional form: $f(x)=\beta_0+\Sigma\beta_m h_m(x)$, where $h_m(x)$ is a function of one or more basis functions (Hastie, Tibshirani & Friedman 2001). The dependent variable (y) in the MARS model is usually continuous, but a binary dependent variable is still estimable as a linear probability model (LPM). Although the

β_m parameters are estimated by minimising the residual sum of squares (RSS) $\Sigma(y-f(x))^2$, estimation of the correct $H_m(x)$ is tricky since it can be a function of the variable itself or an interaction with previously existing basis functions. The first parameter estimated by MARS is always the constant in the form of $h_0(x)=1$. After the constant, terms that decrease the training error the most are added to the model . Two forms of the split at knot point *t* are created for variable X_k that are split into a pair of basis functions and parameterised as $B_1(X_k\text{-}t) + B_2(t\text{-}X_k)$. Whereas MARS's predecessor CART will continue to partition the data space until it exhausts the data, MARS will create pairs of basis functions up to the point that the number of basis function hits a predetermined limit. The final model at this limiting point is usually overfitted, so a back pruning mechanism that assesses the model on test error is necessary. A generalised adjusted penalty is used by MARS to determine this best set of basis functions to include in the model. Not only do generalised procedures perform well in model selection, but using generalised degrees of freedom performs even better when the data matrix used is orthogonal (Ye 1998). For computational ease, MARS uses the generalised cross validation (GCV) criteria $GCV(\lambda)=RSS/(1-\lambda/N)$. What makes MARS more robust is that λ, which is traditionally estimated as 1 degree of freedom per parameter in standard LPM where $\lambda=k$, is a function of the number of basis functions, split points and user's domain knowledge such that $\lambda=f$(BFs, split points, user specified penalty) in MARS. Friedman (1991) suggests a user specified penalty of 2-5 degrees of freedom per basis function but Steinberg et al. (2001) suggest using more than 5 based on empirical findings.

As a result of this methodology, the regression surface is built parsimoniously using non-zero basis functions locally and only where they are needed (Hastie, Tibshirani & Friedman 2001). For the inclusion of interactions in the model, MARS will only build a higher interaction if one of the lower order interaction variables already exists as a basis function. This assumption may not be always true and sometimes a significant interactive region may come from two or more insignificant main effects. However, Friedman (1991) considers the method of interactions coming from existing significant basis functions in MARS as a reasonable assumption.

3. METHODOLOGY

3.1 Data pre-processing and set up

The MARS or mixed logit algorithm will only estimate the choice model parameters correctly if the data is properly set up. We assume that the utility function is linear in parameters but not necessarily linear in attributes and that the respondents are using random utility theory (RUT) to make their retailer choice. RUT assumes that farmers choose an alternative based on attributes of the alternative and some random error (McFadden, 1986; Ben-Akiva & Lerman 1985). RUT assumes the utility U of a choice is a combination of systemic V and random error ε as follows: $U = V + \varepsilon$. We assume the systemic component V is a matrix of experimentally designed attributes of alternatives and observed demographics X with a parameter vector β and can be expressed as: $V = X\beta$. The choice model based on RUT is set up such that alternative one is chosen over alternative two is if $U_1 > U_2$ and we can express this as $U_1 - U_2 \geq 0 \geq V_1 + \varepsilon_1 - (V_2 + \varepsilon_2)$. Substituting the observed attributes X into the equation, we have $U_1 - U_2 \geq X_1\beta - X_2\beta + (\varepsilon_1 - \varepsilon_2)$, which can be reduced to $U_1 - U_2 \geq (x_1 - x_2)\beta + (\varepsilon_1 - \varepsilon_2)$. In a choice model, the term $X\beta$ is set up as the difference in the attributes of an alternative $(x_{1n} - x_{2n})\beta$ and the utility of choice is denoted as $U = X\beta + \varepsilon$.

In the farming choice survey, the respondent had to choose from two different retailers (see Appendix 5). The survey is a forced choice survey, in that it did not contain a "no choice" option. To reduce this model to a generic rather than an alternative specific model, we included the brand name as one of the attributes of the alternatives. The inclusion of distance from retailer as an attribute of an alternative was not as straightforward. Usually, the distance from a retailer is a geo-demographic variable that does not change over a respondent's dataset. Since this study used multiple choice sets per respondent, this would tend to make the effect of distance insignificant across choice sets. However, in the choice experiment UTS administered, different retailers were used in each of the respondent's choice sets, so it was determined that distance could be included as an experimentally designed variable rather than as a geo-demographic.

As part of the pre-processing of the data used in this essay, we estimated an LPM and MNL model on all the attributes from the farming study, shown in Table 16. The pre-processing models showed that all variables exhibited effects that were strongly, but not perfectly, linear. In the LPM and MNL pre-processing models, variables ATTR8 (store branding) and ATTR14 (late payment fee) were insignificant and thus were removed from all further analyses. Although the two removed variables could be potentially used as interactions, not including insignificant main effect variables is appropriate when higher order interactions are not going to be estimated (Taverna, Urban & McDonald 2004). The price and distance variables were numeric but all other experimentally designed variables were ordinal, and these were included in the models as effects coded variables. Although MARS can handle missing data, in order to make accurate comparisons between mixed logit and MARS models, we had to remove all missing data from the analysis (Howieson 1991). Each table of estimation results in Section 5 builds upon the model results from the previous table. For instance, Table 20 is an improvement to the model in Table 19. This culminates in our final model results in Table 22 of Section 5.

3.2 MIXL models for unobserved heterogeneity

A popular technique used to check if there is heterogeneity in a model is to use a mixed logit, which will allow for random preference variation. This model allows taste variation across individuals in the utility function, as well as parameters that are individual specific. In the previous section, we assumed the farmer's preferences were fixed and denoted the utility of choice as:

$$U= \beta X + \varepsilon.$$

To denote the fixed β parameter choice for alternative *i* and person *n*:

$$U_{in} = \beta_n X_n + \varepsilon_n$$

Following McFadden & Train (2000), in a mixed logit context, we assume that β_n is random, so the utility of alternative *i* is specific to person *n* as follows:

$$U_{in} = \beta_n X_{in} + \varepsilon_n$$

$\beta_n \sim f(\beta|\Theta,\Omega)$*;* Θ *is the mean of the* β_n*'s and* Ω *is the covariance matrix of the* β_n*'s*

In our mixed model of farming behaviour, we specified that the mixing distribution $f(\beta|\Theta,\Omega)$ is distributed normally. For choice analytics, the standard logistic specification still applies to the mixed logit model but the likelihood function is more complicated and is estimated numerically. If there is significant heterogeneity across respondents, we would expect the random parameters estimated by the mixed logit model (MIXL) to be significant. The mixed logit models were estimated in two stages using the farming data: first, estimation of a MIXL for the experimentally designed attributes only, and second, estimation of a MIXL using experimentally designed attributes with the demographic characteristics of the respondents included. In the first MIXL model all included parameters were estimated as random parameters except for the constant term. In the second MIXL model, we did not estimate random parameters for the constant nor for the demographics. In both MIXL models we accounted for panel data effects in the estimates.

Even though both MIXL are choice models, we included the constant to make consistent and accurate comparisons between the MIXL models and the MARS models, which must contain a constant since it utlises OLS. We estimated the MIXL models using 2500 random draws on 22 parameters in the R statistical software program. Many articles prefer Halton draws as they cover a multidimensional space more evenly (Bhat 2001). However, our 2500 random draw models converged in a reasonable time and the model performance should be about the same as 1000 Halton draws (Hess & Polak 2003). We estimated two LPMs using the same structure as the two MIXL models to directly compare our results to MARS models. To verify the impact of heterogeneity on the parameters in the model, we estimated MNL models with the same structure as the two MIXL models mentioned previously. We would expect the mixed logit estimates to be larger than MNL estimates, since MIXL incorporates unobserved attributes into the parameter estimates (Revelt & Train 1998).

3.3 MARS-MNL Model for observed heterogeneity

The 'basis functions' created by MARS assume that the parameters in the utility function can be multiple functions and transformations of the same independent variables, which include interactions that are functions of these independent variables. As MARS is a series of functions of the independent variables, it can be thought of as a generalised additive model (GAM) with the form of utility:

$$U=\beta_0 + \sum_1^k fk(X)\beta k$$

What differentiates the MARS model from GAM models is that it uses piecewise splines and, more importantly, the interactions are detected automatically in a MARS model and do not have to be specified a priori. The automatic detection of interactions in MARS is an important distinction because other stepwise logistic regression models eliminate variables that MARS finds to be jointly significant (Kuhnert, Do & McClure 2000). The MARS choice model is a series of basis function (BF) transformations of the alternative attribute differences **X** as well as respondents' demographics **Z** as follows:

$$Y = f(\mathbf{X},\mathbf{Z}) = BF_0 + BF_1 + BF_2 + BF_3 + BF_4 +\ldots BF_k$$

In a MARS model, the dependent variable (Y) can be continuous or discrete. In the rural choice study, the dependent variable was '0' if the respondent chooses not to buy or '1' if the respondent chooses to buy. Since the dependent variable is two-level discrete, the resultant MARS model is a Linear Probability Model (LPM). The BF data transformations need to be integrated into a MNL choice model to recover the standard errors of the transformed parameters and to ensure the predicted response stays in the [0,1] interval. The transformation of the BFs into the log-odds space makes the MARS-MNL model directly comparable to a mixed logit. Unlike a CART-MNL hybrid, a MARS-MNL hybrid can be easily re-parameterised into MNL as the probability space is the same for all the data (Friedman 2010). Using the binary farming choice data, the transformations found by MARS were integrated into a binary choice model as follows:

$$\text{Ln}\,\frac{P}{(1-P)} = \boldsymbol{BF}\theta$$

Where ***BF*** denotes the series of basis functions found by MARS, *P* represents the probability of the respondent choosing retailer 1, and Θ are the parameters of the basis functions. Three MARS models were estimated: a LPM MARS model, a MARS-MNL hybrid and a MARS-Mixed Logit hybrid. The MARS mixed logit model was estimated to verify to what extent the MARS BF transformations accounted for preference heterogeneity. The estimation of the MARS mixed logit accounted for the panel effects of multiple choice sets from the same individuals. The MARS-MIXL hybrid should show no significant random component parameters if MARS does in fact account for both observed and unobserved heterogeneity. The MARS models in this essay were estimated using Salford Systems' implementation of the MARS algorithm (Steinberg et al. 2001). The hybrid models of MARS with MIXL were estimated in the R statistical software program (version 2.13).

4. SIMULATION

4.1 Simulation Dataset up

For the known basis function simulation, we used continuous data and two multi-level categorical variables. The generated "X" data was experimentally designed using an orthogonal main effects plan (OMEP) by Dr. John Rose. The generated data tried to mimic the same data structure in the rural choice study, so the only variable that was interval scaled was price. Variables for distance and demographics were added later to the data matrix. Since many of the simulation variables had more than two levels, a fold over could not be used.

Two demographics covariates for MARS simulated respondents were generated from the distributions in Table 17. In all cases, the values were truncated to the nearest non-negative integer. Even though it was continuous, the simulated AGE demographic was binned into 6 levels (i.e. under 25, 25-35, 35-45, etc.) STATE was also binned into 6 levels. Both demographic variables were effects coded for the simulation.

TABLE 2:MARS SIMULATION DEMOGRAPHIC LEVEL DESIGN

Demographic Variable	Distribution
AGE	Normal $(45,5^2)$
STATE	Uniform(1,6)

Out of the 15 variables possible, we randomly selected four continuous variables and added one of the two demographic covariates. Of the four continuous variables selected, three were randomly picked to be two-way interacted with each other. The dependent variable Y is binary and set to 0 if the simulated consumer chooses not to buy and 1 if the consumer chooses to buy. The binary dependent variable makes the MARS model a LPM. Since we are testing the effects of different amounts of knots per variable, there are three MARS simulation models that are based on the following model:

$$Y = f(\mathbf{X}) = BF_0 + BF_1 + BF_2 + BF_3 + BF_4 + BF_5 = \boldsymbol{\beta}'\mathbf{X}$$

In the above "base" model BF_0 is the constant, BF_1 is a transformed continuous variable, BF_2, BF_3, and BF_4 are two-way interactions of continuous variables and BF_5 is a randomly selected effects coded categorical variable.

Since every MARS basis function is only a one knot transformation or an interaction of two basis functions, when we test 2 or 3 knots per variable in the simulation, we need to include extra basis functions. Even though we are using the same four variables, they are being transformed twice when using two knots, so we need three additional basis functions in the base model. If there are three knots per variable, we need 6 additional basis functions for the base model.

A uniform random variable (URV) with values between 0 and 1 was created with a choice heuristic to construct a dependent choice variable as Choice=1 if $\boldsymbol{\beta}'\mathbf{X}$ > URV; 0

otherwise. The purpose of the simulation was to see if MARS can see where the true knot locations are. There were 72 scenarios per simulated respondent and 1,000 simulated respondents. To test if MARS can detect the correct location of the knot points in a LPM model, the simulation was performed per the following steps:

MARS Simulation Procedure

The following simulation procedures were followed:

1. Using the simulated data, randomly pick a data point to insert a knot point and create a basis function before or after the knot point.
2. Create beta weights for all the basis functions (we will set the constant basis function to β_0=.5).
3. Using a uniform random error, create a dependent/choice variable in the following manner using all basis functions

   if (β'X>error) Choice=1, otherwise Choice=0 ;
4. Take the existing X variables and randomly add a normal error to them so they are away from the true basis function line
5. Estimate an MARS model
6. Record the position that MARS believe the knot point is located
7. Steps 1-6 should be run 100 times for 1 knot, 100 times for 2 knots, and 100 times for 3 knot simulations.
8. Run Steps 1-7 using 1%, 5%, 10% and 20% error for the error in Step 4.

4.2 Simulation Results

The results of the 1,200 simulations are shown below in Table 18. Given the large amount of data and two error structures within the simulation, we assumed a knot point was "correct" if MARS detected it within one standard error of its true location. Even though the models with more parameters fit better, and had a higher R² value, the purpose of the simulation was to see if MARS detected the correct knot location.

TABLE 3: MARS KNOWN BASIS FUNCTION SIMULATION RESULTS

Random		**Knots**	
Error	**1**	**2**	**3**
1%	95%	92%	88%
5%	95%	90%	85%
10%	93%	85%	81%
20%	88%	77%	65%

As expected, MARS performed worse as more error was entered into the data. Furthermore, finding a greater number of basis functions proved problematic for the MARS simulation. When there are multiple knots per variable, if the true knot locations are too close to each other, MARS may have a hard time seeing one knot after finding the other knot. This problem should be addressed in future simulation research. In the original MARS article, Friedman (1991) suggests that a minimum span should exist between data points for MARS to more effectively find knot points.

5. ANALYSIS AND RESULTS

5.1.1 Mixed Logit Model with ED variables

Table 19 shows the estimation results from a LPM, MNL and mixed logit models of the experimentally designed data only.

TABLE 4: MIXED LOGIT AND MNL MAIN EFFECTS ESTIMATED

Attribute	LPM Estimate	S.E.	MNL Estimate	S.E.	MIXL Estimate	S.E.	SD of MIXL	S.E.
Constant	0.490	0.007	-0.045	0.037	-0.106	0.085	-	-
ATTR1E1	0.021	0.005	0.083	0.025	0.223	0.072	0.779	0.233
ATTR2E1	-0.049	0.009	-0.245	0.044	-0.612	0.157	0.756	0.276
ATTR2E2	-0.041	0.009	-0.195	0.047	-0.420	0.131	0.456	0.315
ATTR2E3	0.031	0.009	0.151	0.043	0.311	0.118	-0.024	1.112
ATTR3E1	-0.015	0.004	-0.092	0.022	-0.209	0.066	0.541	0.206
ATTR4E1	-0.029	0.004	-0.145	0.022	-0.335	0.084	0.478	0.211
ATTR5E1	-0.025	0.008	-0.105	0.042	-0.183	0.099	-0.419	0.353
ATTR5E2	0.017	0.009	0.075	0.043	0.109	0.097	-0.487	0.332
ATTR5E3	-0.003	0.008	-0.024	0.042	-0.055	0.095	-0.423	0.382
ATTR6E1	-0.022	0.009	-0.124	0.043	-0.243	0.102	-0.050	1.099
ATTR6E2	0.002	0.009	0.008	0.044	0.047	0.101	0.299	0.485
ATTR6E3	0.020	0.008	0.117	0.043	0.202	0.100	0.358	0.420
ATTR7E1	-0.012	0.004	-0.066	0.022	-0.130	0.054	0.374	0.228
ATTR9E1	-0.034	0.004	-0.167	0.022	-0.336	0.085	0.456	0.205
ATTR10E1	-0.061	0.004	-0.289	0.022	-0.656	0.142	0.629	0.206
ATTR11E1	-0.022	0.004	-0.115	0.022	-0.282	0.076	-0.224	0.335
ATTR12E1	-0.013	0.009	-0.066	0.047	-0.240	0.120	-0.616	0.308
ATTR12E2	-0.022	0.010	-0.099	0.049	-0.162	0.116	-0.638	0.309
ATTR12E3	-0.020	0.009	-0.098	0.046	-0.151	0.105	0.520	0.322
ATTR15	-1.153	0.061	-5.460	0.322	-14.454	3.007	14.514	3.342
Distance	-0.002	0.001	-0.004	0.001	-0.010	0.003	-0.018	0.007
			LL	**-2346**	**LL**	**-2281**		

Highlighted coefficients significant at α=.05 level

In Table 19, 50% of the parameters estimated with random effects in the mixed logit model had significant heterogeneity. Price (ATTR15) has significant preference heterogeneity, which was to be expected (Bell et al. 1998). Some effects coded levels of the variable ATTR12 (Professional Advisory Service) were barely significant. We could not remove one level of the effects coded variables because they would form a grouping which meant all effects codes had to be included or all excluded (Hensher et al. 2005). In this essay, the mixed logit models assumed normal distributions for the random components. Since the normal distribution is symmetric about 0, it can have positive or negative values. In some cases, this will lead to the standard deviation of the random coefficient having a negative value (Glasgow et al 2001). This “wrong” sign can be corrected by using a distribution that has only positive values in the mixed logit estimation process. A log-normal distribution should be used if the response parameter must be specifically non-negative (Hensher & Greene, 2002), but we had no a priori reason to assume non-negativity.

5.1.2 Mixed Logit Model with ED variables and demographics

Table 20 shows the results of an estimation of a LPM, MNL, and MIXL. In this table, the models included the experimentally designed data as well as the demographics. Since the functional form of the demographics was not known a priori, they were entered into the model as main effects.

TABLE 5: MIXED LOGIT AND MNL ESTIMATED WITH DEMOGRAPHICS

Attribute	LPM Estimate	S.E.	MNL Estimate	S.E.	MIXL Estimate	S.E.	SD of MIXL	S.E.
Constant	0.529	0.073	0.160	0.361	0.051	0.714	-	-
ATTR1E1	0.021	0.005	0.084	0.025	0.182	0.055	0.805	0.189
ATTR2E1	-0.048	0.009	-0.241	0.044	-0.467	0.112	0.587	0.202
ATTR2E2	-0.042	0.009	-0.199	0.047	-0.398	0.106	0.405	0.234
ATTR2E3	0.030	0.009	0.150	0.043	0.275	0.095	-0.104	0.437
ATTR3E1	-0.015	0.004	-0.091	0.022	-0.187	0.052	0.422	0.165
ATTR4E1	-0.029	0.004	-0.144	0.022	-0.282	0.061	0.389	0.168
ATTR5E1	-0.025	0.008	-0.106	0.042	-0.175	0.090	-0.017	0.393
ATTR5E2	0.017	0.009	0.077	0.043	0.073	0.089	0.428	0.273
ATTR5E3	-0.003	0.008	-0.024	0.042	-0.033	0.084	0.431	0.264
ATTR6E1	-0.022	0.009	-0.123	0.043	-0.214	0.088	0.015	0.451
ATTR6E2	0.002	0.009	0.008	0.045	0.055	0.091	0.257	0.317
ATTR6E3	0.020	0.008	0.117	0.043	0.173	0.089	-0.203	0.322
ATTR7E1	-0.012	0.004	-0.066	0.022	-0.087	0.045	0.276	0.185
ATTR9E1	-0.034	0.004	-0.168	0.022	-0.289	0.060	0.488	0.158
ATTR10E1	-0.061	0.004	-0.291	0.022	-0.618	0.103	0.585	0.163
ATTR11E1	-0.022	0.004	-0.115	0.022	-0.273	0.062	0.172	0.201
ATTR12E1	-0.013	0.009	-0.066	0.047	-0.215	0.100	0.639	0.251
ATTR12E2	-0.022	0.010	-0.099	0.049	-0.150	0.099	0.413	0.256
ATTR12E3	-0.019	0.009	-0.098	0.046	-0.110	0.091	0.373	0.256
ATTR15	-1.151	0.061	-5.464	0.323	-12.411	1.943	12.372	2.186
Distance	-0.002	0.001	-0.004	0.001	-0.009	0.002	0.009	0.005
Sec4Q2	-0.003	0.007	-0.020	0.036	0.008	0.071	-	-
Sec4Q5	-0.003	0.004	-0.012	0.018	-0.020	0.036	-	-
Sec4Q10	0.003	0.002	0.016	0.012	0.031	0.024	-	-
Sec4Q14	-0.001	0.004	-0.001	0.020	-0.004	0.037	-	-
Sec4Q17	-0.003	0.002	-0.016	0.010	-0.031	0.020	-	-
StateE3	-0.039	0.015	-0.191	0.075	-0.324	0.152	-	-
			LL	**-2341**	**LL**	**-2278**		

Highlighted coefficients significant at α=.05 level

Table 20 shows that the addition of all but one of the demographic variables was insignificant. Although many demographic variables were available for inclusion in the models (see Appendix 8), to ensure accurate comparison of model results, we only included the demographic variables that were determined to be significant in Table 22. The addition of demographics in the mixed logit model in Table 20 reduced the number of significant random parameters from ten in Table 19 to eight. This indicates that the addition of demographics in a model, albeit as main effects, accounts for some of the heterogeneity in the mixed logit (Gupta & Chintagunta 1994). The heterogeneity in distance was slightly insignificant at the 5% level when the demographics were included, so distance may have been a proxy for something in the demographics. This result leads us to suspect there may be interactive effects in the demographics.

The effects coded variable **StateE3** is the only significant demographic variable. Unlike other effects coded variables, State is categorical so we can use an individual effects code level. The variable **StateE3** is the effect of purchasing in South Australia (SA) relative to Western Australia (WA). The negative coefficient indicates that people in South Australia are less likely to purchase from a rural retailer relative to people in Western Australia. What makes this information useful for the marketing manager is that these states are contiguous and the farm size in these two states is large compared to the rest of Australia. Given these characteristics, one could purchase in WA as easy as SA, so the difference in retailer choice may not be due to geographic factors as much as macro environmental issues (e.g., if taxes in South Australia are higher than in Western Australia).

5.2.1 MARS ED data model

Table 21 contains the estimation results of three models using experimentally designed data only in MARS. The LPM model shows the results of a MARS model itself. The MNL model below is a hybrid of the MARS basis functions estimated in an MNL model (MARS-MNL). The final MIXL-MARS hybrid shows the results of the MARS basis functions estimated in a mixed logit with random parameters for all non-constant MARS basis functions.

Table 6: MARS Main Effects with transformations and interactions

MARS Attribute	**LPM**	**S.E.**	**MNL**	**S.E.**	**MIXL**	**S.E.**	**SD of**	**S.E.**
Function	**Estimate**		**Estimate**		**Estimate**		**MIXL**	
Constant	0.627	0.055	0.636	0.281	1.030	0.439	-	-
Basis Function 2	1.749	0.302	8.503	1.568	9.342	2.508	-3.519	1.148
Basis Function 3	-0.142	0.011	-0.705	0.061	-0.878	0.103	0.037	0.434
Basis Function 4	-0.108	0.010	-0.549	0.051	-0.700	0.081	0.087	0.414
Basis Function 5	-0.170	0.036	-0.918	0.187	-1.288	0.302	-0.341	1.676
Basis Function 6	-0.062	0.016	-0.297	0.081	-0.419	0.112	0.021	1.140
Basis Function 7	0.091	0.014	0.491	0.072	0.578	0.107	-0.207	0.371
Basis Function 11	0.029	0.006	0.140	0.032	0.185	0.045	0.009	0.987
Basis Function 12	-0.052	0.009	-0.276	0.048	-0.359	0.080	0.442	0.214
Basis Function 14	0.016	0.003	0.082	0.013	0.103	0.019	0.093	0.033
Basis Function 17	-0.004	0.001	-0.022	0.006	-0.030	0.010	-0.013	0.013
Basis Function 19	-0.195	0.062	-0.922	0.318	-1.561	0.572	1.791	1.814
Basis Function 20	-0.014	0.003	-0.077	0.017	-0.089	0.027	0.121	0.078
Basis Function 21	-0.138	0.034	-0.638	0.178	-0.769	0.269	0.342	1.770
Basis Function 22	-1.059	0.327	-4.853	1.721	-6.040	2.286	-0.383	86.079
Basis Function 24	0.029	0.009	0.133	0.046	0.165	0.062	0.223	0.216
Basis Function 26	-0.117	0.037	-0.641	0.197	-0.843	0.343	1.554	0.802
Basis Function 28	-0.055	0.020	-0.302	0.099	-0.356	0.172	-0.688	0.583
Basis Function 30	0.131	0.052	0.571	0.267	1.060	0.445	-0.004	9.562
Basis Function 33	0.319	0.061	1.678	0.320	2.255	0.576	-0.230	5.551
Basis Function 34	0.048	0.014	0.272	0.068	0.338	0.133	0.305	0.414
Basis Function 35	0.014	0.004	0.078	0.019	0.114	0.029	-0.003	0.613
Basis Function 36	0.465	0.083	2.405	0.432	0.996	1.196	-5.589	2.158
Basis Function 37	0.173	0.055	0.808	0.284	0.875	0.429	-0.104	6.203
Basis Function 42	-1.902	0.411	-10.220	2.115	-16.035	5.102	11.312	12.336
Basis Function 44	0.432	0.077	2.190	0.404	2.638	0.619	-0.119	8.269
Basis Function 46	-0.006	0.001	-0.028	0.007	-0.027	0.012	-0.017	0.029
Basis Function 48	-0.010	0.003	-0.048	0.016	-0.051	0.022	0.036	0.165
			LL	**-2285**			**LL**	**-2253**

Highlighted coefficients significant at α=.05 level

The 27 non-constant parameters in Table 21 are transformations and interactions of the experimentally designed data. Since no demographics were used in this MARS-MIXL model, the results in Table 21 are directly comparable to the MIXL in Table 19. The localised fitting of the MARS model makes all of the MARS-LPM coefficients significant at the 5% level. Even when the MARS basis functions are hybrid into a MARS-MNL model, all the parameters are significant at the 5% level. When the MARS functions are estimated in a mixed logit, only four of the random parameters have significant heterogeneity. The MARS-MIXL estimates for the basis functions are all significant except for basis function 36. Basis function 36 has the following form and is a function of the price of a product and the availability of stock:

BF36 = max(0, DIFF_ATTR15 - 0.05) * max(0, DIFF_ATTR10E1 + 2)

The positive coefficient indicates that people will pay more if the stock is available. Although farmers will get larger discounts purchasing from a nationwide retailer, the farmer is willing to pay more when the local retailer has the product in stock (Gustafson & Nganje 2006). There is significant heterogeneity in this parameter so the addition of demographics or another experimentally designed data interaction in the MARS models can account for this heterogeneity.

5.2.2 MARS ED data and Demographics model

Table 22 contains the estimation results for three models using experimentally designed data and demographics in MARS. The LPM model is the results of MARS itself and the MNL and MIXL are the results of MARS basis functions being hybrid into the multinomial logit and mixed logit (MARS-MNL and MARS-MIXL respectively).

Table 7: MARS Main effects and Demographics with transformations and interactions

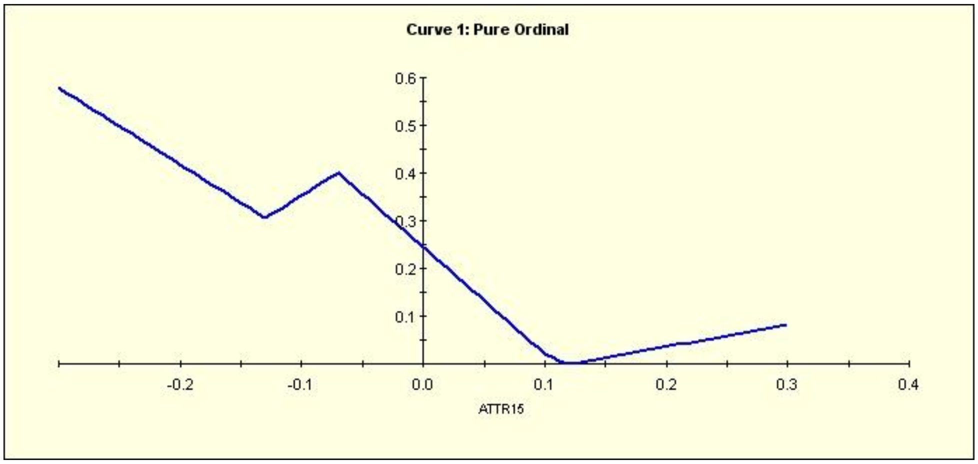
MARS Attribute Function	LPM Estimate	S.E.	MNL Estimate	S.E.	MIXL Estimate	S.E.	SD of MIXL	S.E.
Constant	0.695	0.038	0.867	0.206	1.031	0.274	-	-
Basis Function 2	1.469	0.174	7.646	0.978	9.054	1.411	2.845	1.003
Basis Function 3	-0.102	0.006	-0.484	0.034	-0.577	0.050	-0.007	1.074
Basis Function 4	-0.172	0.018	-0.846	0.094	-0.995	0.120	-0.011	1.628
Basis Function 6	-0.082	0.012	-0.396	0.061	-0.459	0.089	-0.005	1.816
Basis Function 8	4.979	0.605	24.795	3.227	28.646	4.494	-2.694	6.904
Basis Function 10	0.035	0.008	0.180	0.039	0.193	0.051	0.251	0.092
Basis Function 11	0.066	0.008	0.329	0.039	0.400	0.051	0.021	0.619
Basis Function 12	-0.039	0.008	-0.200	0.041	-0.225	0.055	-0.177	0.250
Basis Function 14	0.017	0.002	0.079	0.012	0.095	0.015	0.049	0.035
Basis Function 16	-0.908	0.174	-4.629	0.932	-5.284	1.704	-0.894	16.630
Basis Function 17	-0.007	0.001	-0.038	0.008	-0.047	0.013	0.006	0.043
Basis Function 18	-1.083	0.275	-4.502	1.400	-7.756	5.213	-12.927	12.839
Basis Function 19	-0.285	0.066	-1.332	0.345	-1.697	0.505	-0.126	10.365
Basis Function 20	-0.023	0.006	-0.121	0.031	-0.138	0.045	0.098	0.135
Basis Function 22	-0.011	0.003	-0.058	0.015	-0.071	0.020	-0.008	0.572
Basis Function 24	-0.019	0.005	-0.101	0.028	-0.122	0.034	0.086	0.105
Basis Function 25	-0.140	0.027	-0.738	0.148	-0.958	0.234	0.269	1.008
Basis Function 26	-0.132	0.038	-0.647	0.202	-0.737	0.281	0.546	1.212
Basis Function 29	0.047	0.010	0.233	0.053	0.268	0.082	-0.267	0.162
Basis Function 31	-1.390	0.229	-6.260	1.181	-6.569	1.433	0.118	13.823
Basis Function 32	0.044	0.012	0.200	0.061	0.232	0.074	-0.009	1.930
Basis Function 33	0.038	0.012	0.181	0.059	0.179	0.075	-0.258	0.160
Basis Function 36	-0.023	0.009	-0.113	0.047	-0.123	0.058	0.017	1.577
Basis Function 38	-0.093	0.034	-0.500	0.184	-0.654	0.297	0.345	2.181
Basis Function 40	0.002	0.001	0.010	0.003	0.130	0.077	-0.004	1.795
Basis Function 41	0.005	0.002	0.025	0.010	0.025	0.015	-0.035	0.038
Basis Function 45	-0.022	0.006	-0.119	0.034	-0.137	0.042	0.037	0.552
Basis Function 47	-1.440	0.234	-6.592	1.202	-7.345	1.544	1.941	1.559
Basis Function 53	-0.009	0.002	-0.045	0.011	-0.049	0.014	0.000	0.601
Basis Function 57	0.005	0.002	0.026	0.008	0.031	0.018	0.002	0.317
Basis Function 59	-0.018	0.005	-0.089	0.026	-0.095	0.033	0.075	0.102
			LL	**-2250**			**LL**	**-2228**

Highlighted coefficients significant at α=.05 level

The results in Table 22 represent the definitive models used in this essay. The MARS-MNL model results include demographics, experimentally designed variables, and the transformations and interactions of those demographics and experimentally designed variables. Although all demographics were included in the models (see Appendix 8), MARS used only six of the demographic variables available. The results of the MARS-MNL model show that all parameters are significant and that the model accounts for observed heterogeneity. To verify the heterogeneity reduction, the MARS basis functions were hybrid into a mixed logit to see if there was any residual unobserved heterogeneity in the model. No MARS-MIXL basis functions that included demographics contained any significant unobserved heterogeneity. In the MARS-MIXL model, only two basis functions have significant random parameter values: BF2, which is a transformation of the main effect of price only, and BF10, which is an interaction of the experimentally designed variables ATTR2 (Staff Product Knowledge) and ATTR4 (Independence of Advice). ATTR2 and ATTR4 were already shown to contain heterogeneity in Table 19 and Table 20. Since the MARS basis function two-way interaction of these variables did not eliminate their heterogeneity, this implies that a higher order interaction (i.e., three or four way) is needed to account for the heterogeneity.

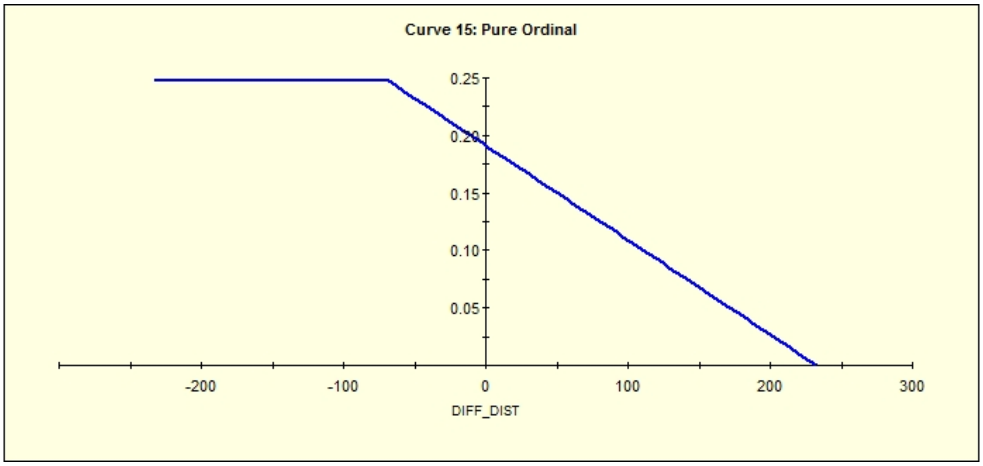
Economic theory says that the demand for a product has a negative relationship with price and that the price effect should be monotonically decreasing (Varian 1992). Figure 12 shows the relationship between percentage of price increase on the horizontal axis and the probability of purchase on the vertical axis. This plot is produced from MARS and shows the main effect of price conditioning on all variables that were interacted with price. There are clear moderation effects on price, which may be the reason why price has such a high degree of heterogeneity. The interactive effects with price are so strong that the response to price is actually positive when the price increase is 30% in Figure 12.

FIGURE 2: EFFECT OF PRICE VARIABLE WITH LINEARITY ASSUMED



In Figure 13, we see that the distance variable acts as we expect and is downwards sloping as distance from retailer increases (Sullivan 1990). One of the advantages of using MARS is that the basis function knot points can easily show the analyst where the variable's effect changes independently of it being interacted with other variables. In Figure 13, MARS finds that the negative sloping distance effect does not start until a retailer is 50km closer and has no effect when a retailer is more than 240km away from a respondent. We only plotted the quantitative variables because plots of categorical outcome variables are not helpful (Rutter & Elashoff 1994).

FIGURE 3: EFFECT OF DISTANCE VARIABLE WITH LINEARITY ASSUMED



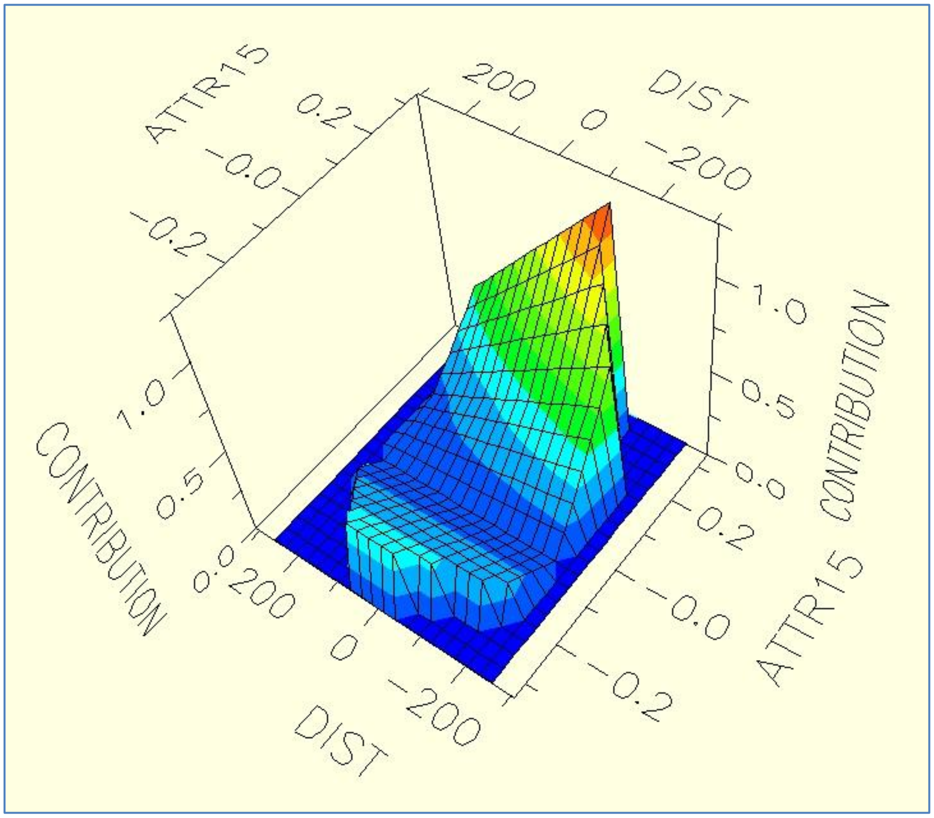
As shown in Table 18, the heterogeneity in price is quite significant and large compared to the MNL estimates. In Table 20, BF2 is a simple transformation of price (ATTR15): the linear effect of price has a change in slope when the difference in the retailers' prices is less than 13%. Although there is still heterogeneity in BF2, the ratio of the MIXL price parameter to the MNL price parameter of BF2 is much less in Table 20 than the ratio of MIXL price to MNL price in Table 18. This indicates that MARS transformations are improving the explanation of observed heterogeneity in the model.

The complete list of characteristics of the MARS basis functions used in Table 20 are in Table 22 of Appendix 7. Although MARS data mining will find structure in the data that is statistically correct, a check of all the basis functions should be made to ensure they fit in with social theory. Basis function 8 is a transformed function of the price attribute. In BF17, the price transformation from BF8 is interacted with distance as follows:

$$\mathbf{BF17 = max\ (0, DIST + 232.499) * max(\ 0, ATTR15 + 0.05)}$$

The resulting coefficient on BF17 is - 0.038, which indicates that people are less likely to purchase when they have to travel far and pay a higher price. Figure 14 graphically shows the interactions between distance to retailer and price discount. For instance, when the difference in retailer's price (ATTR15) is 20% higher, the probability of purchase increases only when the more expensive retailer is at least 100km closer than the alternative retailer. As the price difference becomes less, the effect of distance is not as pronounced. If higher order interactions were included in our MARS model, we would expect that the price-distance interaction would have a demographic interaction, since travelling long distances has socio-demographic components (Limtanakool et al. 2006).

Figure 4: Interaction of Product Price and Distance



Basis Function 26 in Table 22 is an interactive region of a 'fee for expert advice' and price.

BF26 = max(0, ATTR12E2 + 1) * max(0, 0.13 - ATTR15)

The negative coefficient on BF26 in the MARS-MNL model indicates there is a threshold on how much people will pay for this for advice. Bennet, Bratton & Robson (2000) found that people will travel farther and pay more for expert advice in a B2B, but they did not verify that there is a threshold to this moderation, which was found by the MARS model. Some demographic variables that are theoretically assumed to exist were made available to MARS but not used in the final analysis. For instance, it is suggested that farm ownership and tenure of renters has an effect on farm purchasing and probability (Barry et al. 2001), but the demographic covariate for own/rent was not found to have an effect on purchase behaviour substantial enough to be included in the final MARS model.

In order to be of benefit to the retailers, the attributes that are important can be determined by using their contribution to the explained sum or squares, regardless if the attribute is interacted or not. Once the explained sum of squares is computed for each variable, the variable with the highest explained sum of squares is set to 100 and all the other variables are scaled accordingly. The variable importance graphic included in Table 23 below shows the results of important attributes for the MARS Basis Functions in Table 21.

TABLE 8: VARIABLE IMPORTANCE OF MARS MAIN EFFECTS

Variable	Score	
DIFF_ATTR15	100.00	||||||||||||||||||||||||||||||||||||||||||
DIFF_ATTR10E1	62.94	||||||||||||||||||||||||||
DIFF_ATTR2E1	42.04	|||||||||||||||||
DIFF_ATTR9E1	36.91	|||||||||||||||
DIFF_DIST	27.72	|||||||||||
DIFF_ATTR1E1	24.67	||||||||||
DIFF_ATTR12E3	23.23	|||||||||
DIFF_ATTR11E1	17.33	||||||
DIFF_ATTR4E1	16.35	||||||
DIFF_ATTR6E1	14.74	|||||
DIFF_ATTR12E2	13.49	|||||
DIFF_ATTR5E3	8.30	|||
DIFF_ATTR7E1	8.19	|||
DIFF_ATTR5E2	6.61	||

The variable importance listing in Table 24 shows how the importance ranking of variables changes with additional covariates, using the results from Table 22.

TABLE 9: VARIABLE IMPORTANCE OF MARS MAIN EFFECTS AND DEMOGRAPHICS

Variable	Score	
DIFF_ATTR15	100.00	||||||||||||||||||||||||||||||||||||||||||
DIFF_ATTR10E1	65.85	|||||||||||||||||||||||||||
DIFF_ATTR9E1	42.07	|||||||||||||||||
DIFF_ATTR2E1	41.52	|||||||||||||||||
DIFF_ATTR4E1	33.43	|||||||||||||
DIFF_ATTR1E1	27.39	|||||||||||
STATEE3	19.59	|||||||
DIFF_ATTR2E2	18.97	|||||||
SEC4Q5	18.19	|||||||
DIFF_ATTR12E3	16.71	||||||
DIFF_ATTR5E2	16.54	||||||
DIFF_DIST	15.97	||||||
SEC4Q14	15.95	||||||
DIFF_ATTR5E3	15.09	|||||
DIFF_ATTR11E1	12.42	||||
DIFF_ATTR12E2	10.07	|||
SEC4Q2	9.41	|||
SEC4Q17	6.99	||
DIFF_ATTR7E1	4.83	|
SEC4Q10	3.52	|
DIFF_ATTR6E1	2.73	

The above tables can be used by the retailers to see which variables are best to concentrate on. The non-linearity in the importance ranking indicates that the top 4 variables are especially important. For instance, in Table 23, the variables for ‘price’, ‘having product in stock’, ‘more staff knowledge’, and ‘a wide range of products’ are shown, respectively, to be the most important features. The insight given to marketing managers is enhanced when demographic covariates are entered into the model, and the resulting variable importance is shown in Table 24. When demographics are entered into the model, ‘wide range of products’ becomes more important than ‘more staff knowledge’. This indicates that respondents’ characteristics interact more with ‘wide range of products’ than ‘staff knowledge’, and this can help the marketing manager to alter their marketing mix.

5.3 Comparison of all models

Table 25 summarises the log-likelihood values of the MNL and MIXL models in Table 20 and Table 22. A series of chi-square tests were conducted to see which of the four models performed the best. The results of the tests show that the MARS-MXL model outperformed all but the MARS-MNL model (p-value=.054). This indicates that the MARS-MNL model in Table 22 sufficiently accounted for all of the observed heterogeneity. Since only two of the random parameters were significant in the MARS-MIXL model in Table 22, this indicates that any unobserved heterogeneity in the model has been better captured by the MARS-MNL model. Additionally, the MARS-MNL model has 31 less parameters and, all things considered, the more parsimonious model is preferred even if the log-likelihood values were identical (Munoz & Felicisimo 2004).

TABLE 10: ALL MODELS LOG LIKELIHOOD VALUE AND LIKELIHOOD RATIO TEST STATISTICS

	MNL	**Mars-MNL**	**MIXL**	**Mars-MIXL**
LL	-2346.05	-2250.18	-2280.50	-2227.90
npar	22	32	43	63

		LL ratio test p-values		
	MNL	**Mars-MNL**	**MIXL**	**Mars-MIXL**
MNL	-	0.000	0.000	0.000
MNL Mars		-	n/a	0.054
MXL			-	0.000

Given that the MARS-MNL model in Table 22 explains all the observed heterogeneity, it should be the model used as the final explanatory model. There is no reason to use the MARS-MIXL model estimates other than as a benchmark against the MARS-MNL. To quantify the amount of residual unobserved heterogeneity in the farming choice model, the MARS basis functions from Table 22 were estimated in a mixed logit *without* accounting for panel data effects. The MARS-MIXL model without panel effects was 4 log likelihood points (-2231.3) different from the MARS-MIXL model with panel data effects (-2227.9). The results of the estimation of the MARS-MIXL model without panel data effects (see Appendix 9) shows that the MARS-MNL accounts for virtually all observed and unobserved heterogeneity.

Although the MARS-MNL performs better than other models, we want to ensure that the model results are stable. MARS uses cross validation internally to ascertain the best degrees of freedom penalty when building basis functions. However, to ensure predictive accuracy, we tested our model on hold out samples. We did this using k-fold cross validation (CV) since it is less likely to lead to overconfidence in parameter estimation (Armstrong 2012). The rural choice study had multiple scenarios per respondent. To make it a true “hold out” sample, we were careful to not include the same respondents in the learn model and the test validation set.

We used the two most popular values for k: 2 and 10. In 2-fold CV, half of the data was used to learn/build the model and the other half of the data was used to test if the model predicted correctly or not. This process was then reversed and the data used to learn was then used to test predictive accuracy. The results for the two fold CV hold out modelling are in Table 26.

TABLE 11: MARS TWO FOLD CROSS VALIDATION HOLD OUT SAMPLE

CV Set	Percent Correct
1	67.28%
2	65.61%
overall	***66.45%***

The hit rate accuracy of the MARS model in Table 22 is 69%, so the two fold CV does well at 66.5%. The second hold out sample is performed via 10-fold cross validation and the results are shown in Table 27. The 10-fold cross validation performed better than the 2-fold and are extremely close to the Table 22 hit rate. This improved performance of the 10-fold CV is expected, since in 10 fold cross-validation, more data is being used in the learn (9/10th of the data) than on the test data (1/10th of the data).

TABLE 12: MARS TEN FOLD CROSS VALIDATION HOLD OUT SAMPLE

CV Set	Percent Correct
1	70.82%
2	66.51%
3	63.83%
4	70.96%
5	67.92%
6	62.29%
7	70.07%
8	71.09%
9	70.08%
10	69.37%
overall	***68.29%***

Breiman, et. al. (1984) mention in their book that cross validation folds of more than 10 have negligible improvements in predictive accuracy, so no higher level CV levels were used in this essay.

6. ANALYSIS AND CONCLUSION

In interaction detection, which variables should be used to determine interactions and the interactive structure are not known a priori; furthermore, interactions of small sample size are particularly difficult to detect (McClelland & Judd 1993). This essay used a hybrid MARS-MNL model to correctly identify interactions between experimentally designed and demographic variables for retailer choices of Australian farmers. The motivation behind our study is that it is easier to segment and target customers when you know the structure of observed heterogeneity (Bell, Ho & Tang 1998). Much of the current literature uses a mixed logit model to detect the presence of unobserved heterogeneity. However, the detection of unobserved heterogeneity does not fully describe the exact structure of heterogeneity. Few recent attempts to perform a posterior analysis of mixed logit have been proposed as a way to look for the structure of heterogeneity (Hess 2007). Observed heterogeneity can be difficult to detect in logit mixture models because a major source of heterogeneity is in the consideration set (Allenby, Arora & Ginter 1998). In this essay, we controlled for observed heterogeneity in the consideration set formation by using stated preference instead of revealed preference data (McFadden 2002; Louviere, Hensher & Swait 2000). For instance, customers usually have a notional price range in mind for a product. To eliminate any heterogeneity present in notional price range, we have set up price as a percentage.

We used a data mining tool called MARS to simultaneously detect non-linearity in the main effects of a choice model as well as determine the structure of interactions in the choice data. The MARS model is a spline regression technique that creates an LPM model, which is a series of transformations known as “basis functions”. To make the MARS model directly comparable to other mixed techniques, we used the basis functions from MARS to transform the choice data so that it can be estimated by a conditional logit. Parameters for the basis functions were estimated from a MNL choice model and these were compared to the results of a mixed logit estimation (which we call the MARS-MNL hybrid). Mixed logit models are based on the assumption that heterogonous differences in respondents can be accounted for in the use of random parameters. Our MARS-MNL model shows that heterogeneous differences can be accounted for in the proper interactions and function formation of the observed attributes.

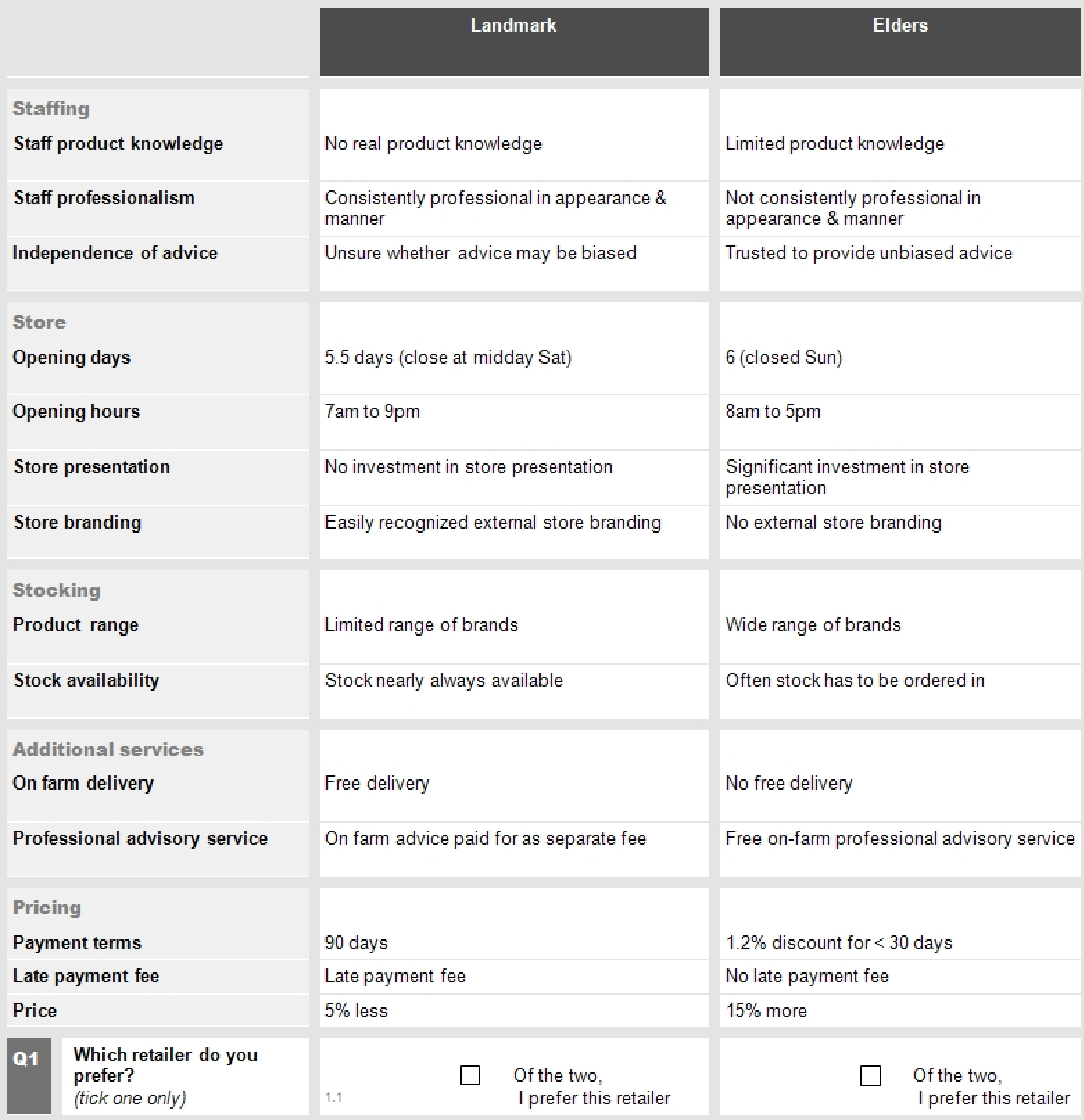
Domain knowledge of which respondent’s demographics are important and can help explain heterogeneity in a model. However, in a choice model, demographic covariates must be entered interactively, and it is highly unlikely that an analyst, even with domain knowledge, will know the precise structure of the interaction. In this essay, ‘domain knowledge’ was incorporated in the first step of the hybrid model. Initially, all attributes available from the rural choice survey were used, but two attributes were removed from any further analysis since they were statistically insignificant in a LPM model. In the data mining hybrid step, no domain knowledge was included, because we felt that to do so would weaken the presumption that the MARS-MNL technique is generalisable. Although the MARS-MNL hybrid works without any a priori knowledge, we assume that marketing managers will use their tacit knowledge to enhance the data mining models.

Using farming choice data, we have shown that the MARS-MNL model accounted for observed sources of heterogeneity. No significant residual unobserved heterogeneity could be detected when the MARS basis function transformations were included in a mixed logit model (Appendix 9). Since the omission of a relevant term from a utility function will yield heterogeneity (Swait & Louveire 1993; McFadden, 1986) and omitted variable bias (Greene 1993), the inclusion of proper interactive structure detected by the use of our MARS-MNL hybrid will correct for these issues. Although MARS can handle multi-way interactions, the MARS models in this essay were confined to two-way interactions. This was done to make the MARS method more easily interpretable, as higher order interactions of a basis function make explanation of the effect difficult (King, Tomz & Wittenberg 2000). By including the proper attribute interaction and transformations, we have shown that the MARS-MNL hybrid model can have significant positive impacts for the marketing manager who wants to accurately target a marketing mix to Australian farmers. According to Landmark Farm Services (Annual Report 2008), investigation of ways to segment the farming retailer market is one of their priorities.

Although there have existed hybrid MARS-Logit models for interaction detection, there have not been any attempts at a MARS discrete choice model hybrid. To check the stability and external validity of the MARS-MNL hybrid, we would like to investigate the hybrid technique on non-farming datasets. Since the MARS basis functions in this essay can be thought of as functions of errors, observed heterogeneity and unobserved heterogeneity, we can formally investigate endogeneity in a choice model similar to a control function approach (Petrin & Train 2010). Also, since Allenby, Arora and Ginter (1998) propose that a major source of heterogeneity is in choice set formation, we would like to use the MARS technique on revealed preference data. Some talks with Sydney research agencies are already underway to address this issue. Finally, there is some evidence of confounding between taste preference heterogeneity and the error component (Swait & Bernardino 2000; Hess 2007). This may indicate that the mixed logit model is mis-specified, and a GMNL would perhaps be a more suitable model. We would like to run our MARS basis functions in a GMNL to see if the MARS transformations work better or worse in a GMNL setting.

APPENDIX 1 – FARMING CHOICE SURVEY QUESTIONNAIRE

FIGURE 5: FARMING RETAILER CHOICE SURVEY QUESTION



	Landmark	Elders
Staffing		
Staff product knowledge	No real product knowledge	Limited product knowledge
Staff professionalism	Consistently professional in appearance & manner	Not consistently professional in appearance & manner
Independence of advice	Unsure whether advice may be biased	Trusted to provide unbiased advice
Store		
Opening days	5.5 days (close at midday Sat)	6 (closed Sun)
Opening hours	7am to 9pm	8am to 5pm
Store presentation	No investment in store presentation	Significant investment in store presentation
Store branding	Easily recognized external store branding	No external store branding
Stocking		
Product range	Limited range of brands	Wide range of brands
Stock availability	Stock nearly always available	Often stock has to be ordered in
Additional services		
On farm delivery	Free delivery	No free delivery
Professional advisory service	On farm advice paid for as separate fee	Free on-farm professional advisory service
Pricing		
Payment terms	90 days	1.2% discount for < 30 days
Late payment fee	Late payment fee	No late payment fee
Price	5% less	15% more
Q1 **Which retailer do you prefer?** *(tick one only)*	1.1 □ Of the two, I prefer this retailer	□ Of the two, I prefer this retailer

APPENDIX 2 – MAP OF AUSTRALIAN FARM RETAILER LOCATIONS

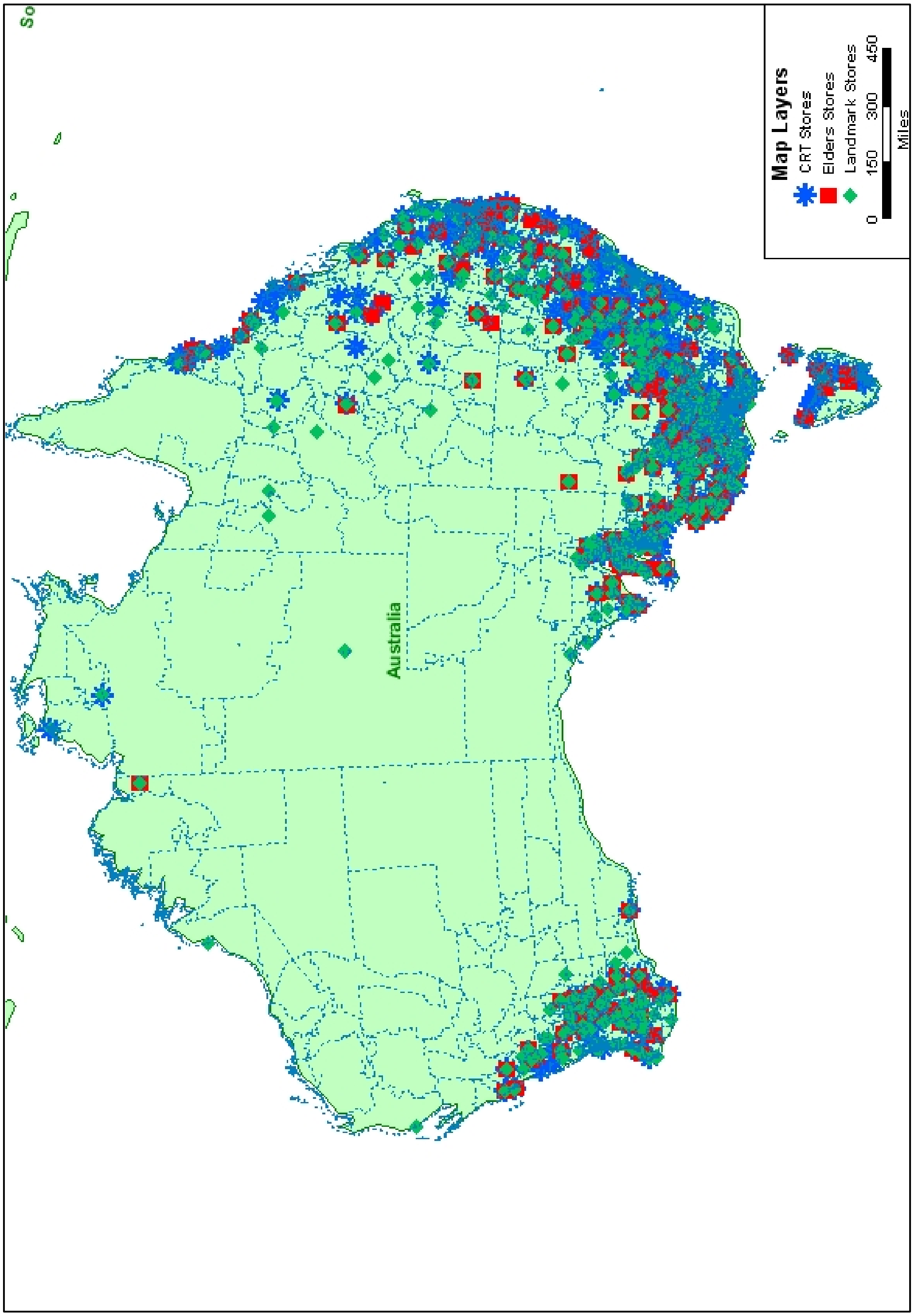


FIGURE 6: LOCATIONS OF THE THREE MAJOR FARM RETAILER OUTLETS IN AUSTRALIA

APPENDIX 3 – DESCRIPTION OF MARS BASIS FUNCTIONS

TABLE 13: MARS BASIS FUNCTIONS USED IN FINAL MODEL

Basis Function	Attribute Transformation
BF2	max(0, 0.13 - ATTR15)
BF3	max(0, ATTR10E1 + 2)
BF4	max(0, ATTR2E1 + 1)
BF6	max(0, ATTR9E1 - 1.39972e-009)
BF7	max(0, 1.39972e-009 - ATTR9E1)
BF8	max(0, ATTR15 + 0.05)
BF9	max(0, -0.05 - ATTR15)
BF10	max(0, ATTR4E1 - 1.94556e-009) * BF4
BF11	max(0, 1.94556e-009 - ATTR4E1) * BF4
BF12	max(0, ATTR12E3 + 1.05438e-009) * BF7
BF14	max(0, ATTR1E1 + 2) * BF3
BF16	max(0, -1 - ATTR2E2) * BF8
BF17	max(0, DIST + 232.499) * BF8
BF18	max(0, ATTR5E3 - 1) * BF8
BF19	max(0, 1 - ATTR5E3) * BF8
BF20	max(0, SEC4Q14 - 32) * BF9
BF22	max(0, ATTR11E1 + 2) * BF7
BF24	max(0, 12 - SEC4Q5) * BF6
BF25	max(0, SEC4Q5 - 8) * BF8
BF26	max(0, ATTR12E2 + 1) * BF2
BF29	max(0, 37 - SEC4Q14) * BF8
BF31	max(0, 1 - ATTR5E2) * BF8
BF32	max(0, ATTR5E2 - 5.78517e-011) * BF4
BF33	max(0, 5.78517e-011 - ATTR5E2) * BF4
BF36	max(0, ATTR6E1 + 1.38267e-009) * BF7
BF38	max(0, ATTR7E1 + 2) * BF8
BF40	max(0, 36250 - SEC4Q17) * BF8
BF41	max(0, SEC4Q10 + 0.00234259) * BF8
BF44	max(0, STATEE3 + 1)
BF45	max(0, ATTR2E2 + 1) * BF44
BF47	max(0, ATTR5E2 + 1) * BF8
BF53	max(0, 27 - SEC4Q14) * BF44
BF57	max(0, SEC4Q2 - 59.5) * BF7
BF59	max(0, SEC4Q2 - 67.5) * BF44

APPENDIX 4 – DEMOGRAPHIC PORTION OF FARMING SURVEY

TABLE 14: DEMOGRAPHIC QUESTIONS ASKED IN FARMING CHOICE SURVEY

Scale	Variable	Demographic Question
Ratio	Sec4Q2	Please tell us your age in years
Ratio	Sec4Q5	Please indicate the highest level of education you have completed in years
Ratio	Sec4Q7	Please indicate the number of bedrooms in your house
Ratio	Sec4Q8	Please indicate the number of persons in your household
Ratio	Sec4Q10	Please indicate your average annual household income (including all household members)
Binary	Sec4Q12	Please indicate which of the following applies to your farming property [Fully owned] / [Being leased]
Ratio	Sec4Q14	Please indicate the average number of hours worked per week on farming related activities
Ratio	Sec4Q15	Please indicate the number of persons engaged in full time farming activities at the farm you work at
Ratio	Sec4Q16	Excluding tractors, please indicate the total number of vehicles owned
Ratio	Sec4Q17	Please indicate the revenue generated by your total farming activity in the last 12 months
Binary	Sec4Q19	Please indicate if you have access the internet
Nominal	State	State farm is located in (** *Effects code base case was Western Australia***)
Nominal	FarmType	Farming Type
Ratio	Hectarage	Farm Hectarage

APPENDIX 5 – MARS-MIXL without Panel Data Effects

Table 15: MARS-MIXL Model with no panel effects

MARS Attribute Function	MXL Estimate	S.E.	SD of MXL	S.E.
Constant	0.769	0.497	-	-
Basis Function 2	21.713	6.461	-11.694	4.206
Basis Function 3	-1.060	0.268	0.515	0.282
Basis Function 4	-1.562	0.430	0.279	1.225
Basis Function 6	-0.765	0.272	-0.016	6.393
Basis Function 8	49.097	12.527	0.008	62.783
Basis Function 10	0.242	0.112	0.422	0.271
Basis Function 11	0.625	0.160	0.067	1.397
Basis Function 12	-0.416	0.149	-0.023	5.426
Basis Function 14	0.162	0.047	0.107	0.126
Basis Function 16	-10.663	3.896	-11.074	7.493
Basis Function 17	-0.077	0.025	0.000	0.281
Basis Function 18	-5.812	4.029	-0.601	111.720
Basis Function 19	-2.970	1.027	-0.092	31.604
Basis Function 20	-0.295	0.126	-0.007	3.658
Basis Function 22	-0.135	0.050	0.004	1.726
Basis Function 24	-0.282	0.114	0.817	0.346
Basis Function 25	-1.308	0.454	0.025	10.981
Basis Function 26	-1.504	0.661	0.031	32.284
Basis Function 29	0.320	0.286	-1.741	0.842
Basis Function 31	-9.065	3.174	-0.057	31.748
Basis Function 32	0.396	0.192	0.111	1.294
Basis Function 33	0.273	0.156	-0.601	0.348
Basis Function 36	-0.168	0.099	-0.225	0.633
Basis Function 38	-1.081	0.464	-0.095	15.044
Basis Function 40	0.163	0.075	-0.025	1.580
Basis Function 41	0.041	0.025	-0.085	0.058
Basis Function 45	-0.280	0.103	0.060	1.333
Basis Function 47	-10.283	3.082	0.266	19.650
Basis Function 53	-0.096	0.062	0.008	1.197
Basis Function 57	0.059	0.029	0.001	1.189
Basis Function 59	-0.220	0.102	0.006	2.846
	LL	**-2231**		

Highlighted coefficients significant at α=.05 level

REFERENCES

Allenby, GM, Arora, N & Ginter, JL 1998, ‘On the heterogeneity of demand’, *Journal of Marketing Research*, vol. 35, no. 3, pp. 384-389.

Allenby, GM & Rossi, PE 1998, ‘Marketing models of consumer heterogeneity’, *Journal of Econometrics*, vol. 89, no. 1-2, pp. 57-78.

Archer, JC & Lonsdale, RE 1997, ‘Geographical Aspects of US Farmland Values and Changes During the 1978-1992 Period’, *Journal of Rural Studies*, vol. 13, no. 4, pp. 399-413.

Armstrong, JS, 2012. ‘Illusions in Regression Analysis’, Forthcoming in *International Journal of Forecasting*, 2012.

Barry, PJ, Moss, LM, Sotomayor, NL & Escalante, CL 2000, ‘Lease Pricing for Farm Real Estate’, *Review of Agricultural Economics*, vol. 22, no. 1, pp. 2–16.

Bayer, P & Timmins, C 2005, ‘On the equilibrium properties of locational sorting models’, *Journal of Urban Economics*, vol. 57, no. 3, pp. 462-477.

Bell, D, Ho, TH & Tang CS 1998, ‘ Determining where to shop: fixed and variable costs of shopping’, *Journal of Marketing Research*, vol. 35, no. 3, pp. 352-369.

Ben-Akiva, ME & Lerman, SR 1985, *Discrete choice analysis: theory and application to travel demand*, The MIT Press.

Bennett, RJ, Bratton, WA & Robson, PJ 2000, ‘Business advice: the influence of distance’, *Regional Studies*, vol. 34, no. 9, pp. 813-828.

Bhat, CR 2000, ‘Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling’, *Transportation Science*, vol. 34, no. 2, pp. 228-238.

Bhat, CR 2001. "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model." Transportation Research Part B: Methodological **35**(7): 677-693.

Bhat, CR & Guo, J 2004, ‘A mixed spatially correlated logit model: formulation and application to residential choice modeling’, *Transportation Research Part B: Methodological*, vol. 38, no. 2, pp. 147-168.

Breiman, L 2001, ‘Random forests’, *Machine learning*, vol. 45, no. 1, pp. 5-32.

Breiman, L, Friedman, J, Stone, CJ & Olshen, RA 1984, *Classification and Regression Trees*, New York, New York, Chapman and Hall.

Bronnenberg, BJ 2005, ‘Spatial models in marketing research and practice’, *Applied Stochastic Models in Business and Industry*, vol. 21, pp. 335–343.

Brownstone, D, Bunch, DS & Train, K 2000, ‘Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles, *Transportation Research Part B: Methodological*, vol. 34, no. 5, pp. 315-338.

Fisher, AC & Hanemann, WM 1998, ‘The Impact of Global Warming on Agriculture Rethinking the Ricardian Approach’, *CUDARE Working Papers*, University of California, Berkeley.

Friedman, J 2010, Wald Lectures. University of Technology, Sydney.

Friedman, JH 1991, ‘Multivariate adaptive regression spline, *The annals of statistics*, pp. 1-67.

Glasgow, G et al. 2001, ‘*Mixed Logit Models in Political Science*’, *Eighteenth Annual Political Methodology Conference*, Emory University.

Greene, WH 1993, *Econometric analysis*, New York, Macmillan.

Gupta, S. & Chintagunta, PK 1994, ‘On using demographic variables to determine segment membership in logit mixture models’, *Journal of Marketing Research*, vol. 31, no. 1, pp. 128-136.

Gustafson, CR. & Nganje, WE 2006, ‘Value of Social Capital to Mid-Sized Northern Plains Farms’, *Canadian Journal of Agricultural Economics*, vol. 54, pp 421–438.

Hastie, T, Tibshirani, R & Friedman, J 2001, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, Springer-Verlag.

Hensher, DA, Rose, JM & Greene, WH 2005, *Applied choice analysis: a primer*, Cambridge Univ Press.

Hensher, DA & Greene, WH 2002, ‘The Mixed Logit Model: the State of Practice’, *Institute of Transportation Studies*, University of Sydney.

Hess, S 2007, ‘Posterior analysis of random taste coefficients in air travel behaviour modelling’, *Journal of Air Transport Management*, vol. 13, no. 4, pp. 203-212.

Hess, S & Polak, JW 2003, ‘On the performance of shuffled Halton sequences in the estimation of discrete choice models’, *European Transport Conference*, Strasbourg.

Howieson, B 1991, ‘A Security Analyst’s Action Recommendations: An Application of Recursive Partitioning to Modelling Judgement’, *Australian Journal of Management*, vol. 16, no. 2, pp. 165-185.

Kamakura, WA & Russell, GJ 1989, ‘A Probabilistic Choice Model for Market Segmentation and Elasticity Structure’, *Journal of Marketing Research*, vol. 26, no. 4, pp. 379-390.

King, GM, Tomz, M & Wittenberg, J 2000, ‘Making the Most of Statistical Analyses: Improving Interpretation and Presentation’, *American Journal of Political Science*, vol. 44, no. 2, pp. 341-355.

Kuhnert, P, Do, K-A & McClure, R 2000, ‘Combining non-parametric models with logistic regression: an application to motor vehicle injury data’, *Computational Statistics & Data Analysis*, vol. 34, no. 3, pp. 371–386.

Limtanakool, N, Dijst, M & Scwanen, T 2006, ‘The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium-and longer-distance trips’, *Journal of Transport Geography*, vol. 14, no. 5, pp. 327-341.

Louviere, JJ, Hensher, DA & Swait JD 2000, *Stated choice methods: analysis and applications*, Cambridge Univ Pr.

McClelland, G & Judd, C 1993, ‘Statistical difficulties of detecting interactions and moderator effects’, *Psychological Bulletin*, vol. 114, no. 2, pp. 376-390.

McFadden, D 1986, ‘The Choice Theory Approach to Market Research’, *Marketing Science*, vol. 5, no. 4, pp. 275-297.

McFadden, D 2002, ‘Disaggregate Behavioral Travel Demand’s RUM Side: A 30-Year Retrospective’, in D Hensher and J King (eds.), *The Leading Edge in Travel Behavior Research*, Oxford, Pergamon Press.

McFadden, D & Train, K 2000, ‘Mixed MNL Models for Discrete Response’, *Journal of Applied Econometrics*, vol. 15, no. 5, pp. 447-470.

Muñoz, J & Felicísimo, Á. 2004, ‘Comparison of statistical methods commonly used in predictive modelling’, *Journal of Vegetation Science*, vol. 15, pp. 285-292.

Petrin, A & Train^, K 2010, ‘A control function approach to endogeneity in consumer choice models’, *Journal of Marketing Research*, vol. 47, no. 1, pp. 3-13.

Revelt, D & Train, K 1998, ‘Mixed logit with repeated choices: households' choices of appliance efficiency level’, *Review of Economics and Statistics*, vol. 80, no. 4, pp. 647-657.

Roe, B & Stockberger, A 2004, ‘Explaining Economic Linkages Between Farms And Local Communities Looking Beyond Farm Size’, *2004 AAEA Annual Meetings*, Denver, Co.

Rust, R & Donthu, N 1995, ‘Capturing Geographically Localized Misspecification Error in Retail Store Choice Models’, *Journal of Marketing Research*, vol. 22, pp. 103–110.

Rutter, C & Elashoff, R 1994, ‘Analysis of longitudinal data: Random coefficient regression modelling’, *Statistics in Medicine*, vol. 13, no. 12, pp. 1211-1231.

Steinberg, D, Bernstein, B, P. L. Colla & Martin, K 2001, *MARS user guide*, San Diego, CA, Salford Systems.

Sullivan, A 1990, *Urban Economics*, Boston, Mass, Irwin.

Swait, J & Bernardino, A 2000, ‘Distinguishing taste variation from error structure in discrete choice data’, *Transportation Research Part B: Methodological*, vol. 34, no. 1, pp. 1-15.

Swait, J & Louviere, JJ 1993, ‘The role of the scale parameter in the estimation and comparison of multinomial logit models’, *Journal of Marketing Research*, vol. 30, no. 3, pp. 305-314.

Taverna, K, Urban, DL & McDonald, RI 2005, ‘Modeling landscape vegetation pattern in response to historic land-use: a hypothesis-driven approach for the North Carolina Piedmont, USA’, *Landscape Ecology*, vol. 20, no. 6, pp. 689-702.

Varian, HR 1992, *Microeconomic Analysis*. New York, WW Norton New York.

Xue, Y & Brown, D 2002, ‘Decision Based Spatial Pattern Analysis’, *Systems and Information Engineering Technical Papers*, University of Virginia.

Ye, J. 1998. "On measuring and correcting the effects of data mining and model selection." Journal of the American Statistical Association **93**(441): 120-131.

Zabaleta, J et al. 2008, ‘Interactions of cytokine gene polymorphisms in prostate cancer risk’, *Carcinogenesis*, vol. 29, no. 3, pp. 573.