WORKING PAPER
ITS-WP-98-5

A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice

by

David A. Hensher
Tu Ton

February, 1998

ISSN 1440-3501

Established and supported under the Australian Research Council's Key Centre Program.

INSTITUTE OF TRANSPORT STUDIES
The Australian Key Centre in Transport Management

The University of Sydney and Monash University

NUMBER: Working Paper ITS-WP-98-5

TITLE: A Comparison of the Predictive Potential of Artificial Neural Networks and Nested Logit Models for Commuter Mode Choice

ABSTRACT:

Understanding and predicting traveller behaviour remains a complex activity. The set of tools in common use by practitioners and many of the tools used by researchers appear in many ways to exhibit complexity; yet often this richness of detail is in methods of estimation rather than in representation of how individuals actually evaluate alternatives and make decisions on a set of interrelated travel choices. Discrete choice methods championed by the multinomial logit model and its variants such as nested logit, heteroskedastic extreme value, and multinomial probit have added substantial behavioural richness into statistical specification and estimation (Hensher et al forthcoming), seeking to accommodate the role of both observed and unobserved influences on travel choices. The search for behavioural and analytical enhancement continues. Research in the field of artificial intelligence systems has been exploring the use of neural networks (eg Faghri and Hua 1991, Yang et al 1993) as a framework within which many traffic and transport problems can be studied. The main motivation for using neural networks could be due to some fascinating properties that neural networks possess. They are parallelism, the capacity to learn, allowing for the use of distributed memory and capacity for generalisation. Following these characteristics, one of the promises from neural networks is that they can tackle the problem of forecasting and modelling which is very common in travel demand modelling. The use of such tools in studying individual traveller behaviour opens up an opportunity to consider the extent to which there are representation frameworks which complement and/or replace existing analytical approaches.

This paper explores the merits of neural networks as part of a revised framework within which to explore the processes of traveller decision making, and how discrete choice methods might be integrated within such a framework to acknowledge the important role that the latter tools have played in the last 25 years in the development of better practice in travel demand modelling.

AUTHORS: David A. Hensher
Tu Ton

CONTACT: Institute of Transport Studies (Sydney & Monash)
The Australian Key Centre in Transport Management
C37, The University of Sydney NSW 2006
Australia

Telephone: +61 2 9351 0071
Facsimile: +61 2 9351 0088
E-mail: itsinfo@its.usyd.edu.au
Internet: http://www.its.usyd.edu.au

DATE: February, 1998

Introduction

Understanding and predicting traveller behaviour remains a complex activity. The set of tools in common use by practitioners and many of the tools used by researchers appear in many ways to exhibit complexity; yet often this richness of detail is in methods of estimation rather than in representation of how individuals actually evaluate alternatives and make decisions on a set of interrelated travel choices.

Discrete choice methods championed by the multinomial logit model and its variants such as nested logit, heteroskedastic extreme value, and multinomial probit have added substantial behavioural richness into statistical specification and estimation (Hensher et al 1996), seeking to accommodate the role of both observed and unobserved influences on travel choices. The search for behavioural and analytical enhancement continues.

Research in the field of intelligence systems has been exploring the use of artificial neural networks (ANN) (eg Faghri and Hua 1991, Yang et al 1993) as a framework within which many traffic and transport problems can be studied. Notable applications are in traffic control and scheduling of rail and air services. The use of such tools in studying individual traveller behaviour opens up an opportunity to consider the extent to which there are representation frameworks which complement and/or replace existing analytical approaches.

This paper explores the merits of neural networks as part of a revised framework within which to explore the processes of traveller decision making, and how discrete choice methods might be integrated within such a framework to acknowledge the important role that the latter tools have played in the last 25 years in the development of better practice in travel demand modelling.

The paper is structured around six sections. The next section is an overview of the empirical setting of the travel choice experiment followed by a description of common variables and data sets selected for contrasting the two modelling approaches: choice and ANN models. Section three describes the specific choice-based models (ie. nested logit models) in estimating commuter mode choice for selected studies. In section four, the basic ANN concepts are presented followed by a description of the specific structure of ANN for representing the same variables and data sets as the choice model. Section five presents the predictive performance comparison between the choice models and neural network models. The paper concludes with comments on the merits of neural networks and choice models in the prediction task of travel demand models.

Empirical Setting

Background of commuter choice studies

The case studies used in this research were extracted from a stated choice experiment. This experiment was part of a broader research effort examining potential impacts of transport policy instruments on reductions in greenhouse gas emissions in six Australian

capital cities: Sydney, Melbourne, Brisbane, Adelaide, Perth and Canberra (Hensher et.al 1995; Louviere et.al 1994). The universal choice set comprised the currently available modes plus the two 'new' modes of light rail and busway. Respondents evaluated scenarios describing ways to commute between their current residence and workplace locations using different combinations of policy-sensitive attributes and levels. The purpose of the exercise was to observe and model their observed coping strategies in each scenario.

Four alternatives appeared in each travel choice scenario: a) car (no toll), b) car (toll), c) bus or busway, and d) train or light rail. Twelve types of showcards described scenarios involving combinations of trip length (3) and public transport pairs (4): bus vs. light rail, bus vs. train (heavy rail), busway vs. light rail, and busway vs. train. Appearance of public transport pairs in each card shown to respondents was based on an experimental design. Attribute levels are summarised in Table 1 and an illustrative show card is displayed in Table 2.

Table 1: The Set of Attributes and Attribute Levels in the Travel Choice Experiment (all cost items are in Australian $'s, all time items are in minutes)

SHORT (< 30 mins.)	**Car no toll**	**Car toll rd**	**PUBLIC TRANSPORT**	**Bus**	**Train**	**Busway**	**Light Rail**
Travel time to work	15,20,25	10,12,15	Total time in the vehicle (one-way)	10,15,20	10,15,20	10,15,20	10,15,20
Pay toll if you leave at this time (otherwise free)	None	6-10, 6:30-8:30, 6:30-9	Frequency of service	Every 5,15,25	Every 5,15,25	Every 5,15,25	Every 5,15,25
Toll (one-way)	None	1,1.5,2	Time from home to closest stop	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8
Fuel cost (per day)	3,4,5	1,2,3	Time to destination from closest stop	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8
Parking cost (per day)	Free,$10,$20	Free,$10,$20	Return fare	1,3,5	1,3,5	1,3,5	1,3,5
Time variability	0, ±4,±6	0,±1,±2					
MEDIUM (30-45 mins.)							
Travel time to work	30,37,45	20,25,30	Total time in the vehicle (one-way)	20,25,30	20,25,30	20,25,30	20,25,30
Pay toll if you leave at this time (otherwise free)	None	6-10, 6:30-8:30, 6:30-9	Frequency of service	Every 5,15,25	Every 5,15,25	Every 5,15,25	Every 5,15,25
Toll (one-way)	None	2,3,4	Time from home to closest stop	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8
Fuel cost (per day)	6,8,10	2,4,6	Time to destination from closest stop	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8
Parking cost (per day)	Free,$10,$20	Free,$10,$20	Return fare	2,4,6	2,4,6	2,4,6	2,4,6
Time variability	0, ±7, ±11	0, ±2, ±4					
LONG (>45 mins.)							
Travel time to work	45,55,70	30,37,45	Total time in the vehicle (one-way)	30,35,40	30,35,40	30,35,40	30,35,40
Pay toll if you leave at this time (otherwise free)	None	6-10, 6:30-8:30, 6:30-9	Frequency of service	Every 5,15,25	Every 5,15,25	Every 5,15,25	Every 5,15,25
Toll (one-way)	None	3,4.5,6	Time from home to closest stop	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8	Walk 5,15,25 Car/Bus 4,6,8
Fuel cost (per day)	9,12,15	3,6,9	Time to destination from closest stop	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8	Walk 5,15,25 Bus 4,6,8
Parking cost (per day)	Free,$10,$20	Free,$10,$20	Return fare	3,5,7	3,5,7	3,5,7	3,5,7
Time variability	0, ±11, ±17	0, ±7, ±11					

Table 2: Example of the Format of a Travel Choice Experiment Showcard

SA101	1. CAR, TOLL ROAD	2. CAR, NON-TOLL ROAD
Travel time to work	**10 min.**	**15 min.**
Time variability	**None**	**None**
Toll (one way)	**$1.00**	**free**
Pay toll if you leave at this time (otherwise free)	**6-10 am**	**—**
Fuel cost (per day)	**$1.00**	**$3.00**
Parking cost (per day)	**Free**	**Free**
	3. BUS	**4. TRAIN**
Total time in the vehicle (one way)	**10 min.**	**10 min.**
Time from home to your closest stop	Walk **5 min.** / Car/Bus **4 min.**	Walk **5 min.** / Car/Bus **4 min.**
Time to your destination from the closest stop	Walk **5 min.** / Bus **4 min.**	Walk **5 min.** / Bus **4 min.**
Frequency of service	**Every 5 min.**	**Every 5 min.**
Return fare	**$1.00**	**$1.00**

Five three-level attributes were used to describe public transport alternatives: a) total in-vehicle time, b) frequency of service, c) closest stop to home, d) closest stop to destination, and e) fare. The attributes of the car alternatives were: a) travel times, b) fuel costs, c) parking costs, d) travel time variability, and for toll roads e) departure times and f) toll charges. The design allows orthogonal estimation of alternative-specific main effect models for each mode option: a) car no toll, b) car toll road, c) bus, d) busway, e) train, and f) light rail.

The master design for the travel choice task was a 27 x 3^{27} orthogonal fractional factorial, which produced 81 scenarios or choice sets. The 27 level factor was used to block the design into 27 versions of three choice sets containing two alternatives. Versions were balanced such that each respondent saw every level of each attribute exactly once. The 3^{27} portion of the master design is an orthogonal main effects design, which permits independent estimation of all effects of interest. Two 2-level attributes were used to describe bus/busway and train/light rail modes, such that bus/train options appear in 36 scenarios and busway/light rail in 45.

Description of common variables and data sets selected for contrasting the choice and ANN modelling approaches

Sydney, Melbourne and the pooled cities (combined Sydney and Melbourne) were selected for the comparative studies. Each data source was split into two sub-data sets: training and testing (see Table 3). Training data were used to feed into both choice and ANN models for estimation. The testing data were used to test both models to establish testing generalisation or predictive capability of models.

Table 3: List of Data Sources, Their Type and Sample Sizes Selected For Contrasting the Choice and ANN Models

Data set	Code Name	Data Sources	Type	Number of observations
1	Syd_Train	Sydney	Training	329
2	Syd_Test	Sydney	Testing	82
3	Mel_Train	Melbourne	Training	312
4	Mel_Test	Melbourne	Testing	78
5	SydMel_Train	Combined Sydney and Melbourne	Training	641
6	SydMel_Test	Combined Sydney and Melbourne	Testing	160

The arrangement of data sets for comparing choice and ANN models is shown in Table 4. Three choice models and three ANN models were estimated for Sydney, Melbourne and combined Sydney and Melbourne. Both choice and ANN models were trained/estimated with the same associated data sets. For example, the SydTrain data set was used by both choice and ANN models in modelling travel behaviour for Sydney.

Table 4: Matrix of Models and Associated Data Sets Used in Estimating and Testing of Both Choice and ANN Models

Model Training/Estimation			Model Testing			
			Data sets			
City	**Data set**	**Model**	**Self**	**Sydney**	**Melbourne**	**Pooled Cities**
SYDNEY	Syd_Train	CHOICE MODEL	Syd_Train	Syd_Test	Mel_Test	SydMel_Test
	Syd_Train	ANN Model	Syd_Train	Syd_Test	Mel_Test	SydMel_Test
MELBOURNE	Mel_Train	CHOICE MODEL	Mel_Train	Syd_Test	Mel_Test	SydMel_Test
	Mel_Train	ANN Model	Mel_Train	Syd_Test	Mel_Test	SydMel_Test
POOLED CITIES	SydMel_Train	CHOICE MODEL	SydMel_Train	Syd_Test	Mel_Test	SydMel_Test
	SydMel_Test	ANN Model	SydMel_Train	Syd_Test	Mel_Test	SydMel_Test

Table 5 provides a list of variables which were used as common variables by both the choice and ANN models. Six possible alternatives are drive alone (DA), ride sharing (RS), bus (BS), busway (BW), train (TN) and light rail (LR).

Table 5: Common variables and Alternatives Selected for Contrasting the Choice and ANN Models

Variable	Alternative
Cost ($)	All
Linehaul Time (mins)	All
Parking cost ($)	DA, RS
Access & Egress Time (mins)	BS, TN, LR, BW

Choice Modelling Approach To Commuter Choice

Nested logit models were estimated for Sydney, Melbourne and the pooled cities. The results are summarised in Table 6. All three models provide statistically significant effects for in-vehicle cost, parking cost, linehaul time and public transport access plus egress time.

Table 6: Summary of Nested Logit Training Models

Variable	Alternative	Syd-Mel	Sydney	Melbourne
Cost ($)	All	-.59985 (-7.08)	-.52395 (-4.88)	-.78084 (-4.75)
Linehaul Time (mins)	All			-.05858 (-3.0)
Linehaul Time (mins)	DA, RS	-.07258 (-5.17)	-.06759 (-4.06)	
Parking cost ($)	DA, RS	-.10589 (-6.11)	-.08268 (-3.64)	-.11552 (-3.65)
Linehaul Time (mins)	BS, TN, LR, BW	-.06809 (-4.80)	-.08708 (-4.29)	
Access & Egress Time (mins)	BS, TN, LR, BW	-.04082 (-5.37)	-.03625 (-3.52)	-.03759 (-2.95)
Car Drive Alone Constant	DA	1.7512 (3.80)	0.6815 (1.34)	2.6909 (3.14)
Ride Share Constant	RS	0.8273 (1.85)	0.02613 (.05)	2.1230 (2.42)
Bus Constant	BS	-.11982 (-.55)	-.10222 (-.35)	-.11818 (-.33)
Train Constant	TN	0.24967 (1.16)	0.10207(0.35)	0.47955 (1.44)
Light Rail Constant	LR	0.38893 (2.23)	0.26275 (1.08)	0.55353 (2.12)
Inclusive Value	DA	0.58122 (5.29)	0.83055 (4.57)	
Inclusive Value	RS, BS, TN, LR, BW	0.39789 (2.13)	0.84010 (3.17)	
Inclusive Value	DA, RS			0.7958 (4.30)
Inclusive Value	BS, TN, LR, BW			0.5291 (2.02)
Sample Size		641	329	312
Log-likelihood at convergence		-1165.34	-367.81	-336.78
Adjusted Pseudo R^2		0.389	0.371	0.382

Note: Nested structures for Sydney and combined Sydney-Melbourne are DA vs the rest; nested structure for Melbourne in DA, RS versus the rest.

In searching for an appropriate model for each market, we found some variations in the specification of the taste weights; in particular the Melbourne model treats linehaul time as generic across all modes whereas the other two markets distinguish car and public transport. The nested structure is also different for Melbourne. We found that car drive alone is partitioned from the other modes for Sydney and the combined cities; suggesting that the unobserved influences on choice are more similar between all public transport modes including ride share; whereas for Melbourne the unobserved effects are similar within the drive alone and ride share alternatives. The taste weights for the inclusive value variables are all statistically significant and lie within the 0-1 range, the latter a requirement for the model form to be globally consistent with random utility maximisation. The overall goodness-of-fit of the models is impressive, with pseudo-r^2s of .371 to .389. The implied behavioural values of travel time savings (VTTS) for car travel are respectively for Syd-Mel, Sydney and Melbourne \$7.26/person hour, \$7.74/person hour and \$4.50. The latter is based on a generic taste weight across all modes, which tends to deflate the car-specific value. The public transport linehaul VTTSs for Syd-Mel and Sydney are respectively \$6.81 and \$9.97; the equivalent access plus egress VTTSs for public transport are \$4.08 and \$4.15. The Melbourne access plus egress VTTS is \$2.89/person hour.

Comparison of the taste weights is a meaningless exercise since each model has a different scale parameter. Our preferred basis for comparison is the marginal effects and elasticities. To demonstrate this, let us begin with the simple multinomial logit model with only the characteristics of each sampled individual in the utility expression, and taste weights not associated with any particular outcome. The notation P_j is used for Prob(y = j). By differentiation, we find that:

$$\begin{aligned} \partial \text{Prob}(y_q = j)/\partial \beta_k &= P_k(1 - P_k)\mathbf{x} \quad \text{if } j = k, \\ &= -P_0 P_k \mathbf{x} \quad \text{if } j \neq k. \end{aligned} \tag{1}$$

That is, every taste weight vector enters every probability. The taste weights in the model are not the marginal effects. Indeed these marginal effects need not even have the same sign as the taste weights. Hence the statistical significance of a taste weight does not imply the same significance for the marginal effect:

$$\partial \text{Prob}[y_q = j]/\partial \mathbf{x} = P_j(\beta_j - \bar{\beta}), \quad \bar{\beta} = \Sigma_j P_j \beta_j. \text{ (defined below as } \delta_j) \tag{2}$$

It follows that neither the sign nor the magnitude of δ_j need bear any relationship to those of β_j. The asymptotic covariance matrix for an estimator of δ_j would be computed using

$$\text{Asy.Var.}[\hat{\delta}_j] = \mathbf{G}_j \, \text{Asy.Var}[\hat{\beta}] \mathbf{G}_j' \tag{3}$$

where $\hat{\beta}$ is the full parameter vector. It can be shown that:

$$\text{Asy.Var.}[\hat{\delta}_j] = \Sigma_l \Sigma_m \, \mathbf{V}_{jl} \, \text{Asy.Cov.}[\hat{\beta}_l, \hat{\beta}_m'] \mathbf{V}_{jm}', \; j = 0,...,J, \tag{4}$$

where $\mathbf{V}_{jl} = [\mathbf{1}(j=l) - P_l]\{P_j\mathbf{I} - \delta\mathbf{x}\phi - P_j\delta\mathbf{x}\phi$

and $\mathbf{1}(j=l) = 1$ if j=l, and 0 otherwise.

Since $\beta = \partial\log(P_j/P_0)/\partial\mathbf{x}$, it has been suggested as an interpretation of the taste weights. "Logit" is not a natural unit of measurement, and is definitely not an elasticity. Thus the taste weights in the multinomial logit model are essentially uninformative. This is why marginal rates of substitution (eg value of travel time savings), marginal effects and elasticities are the preferred behavioural outputs for model comparison. For an MNL model in which attributes of alternatives are included as well as characteristics of sampled individuals, the marginal effects defined as derivatives of the probabilities are given as:

$$\delta_{jm} = \partial P_j/\partial\mathbf{x}_m = [\mathbf{1}(j=m) - P_jP_m]\,\beta \qquad (5)$$

The presence of the IIA property produces identical cross effects. The derivative above is one input into the more general elasticity formula:

$$\eta_{jm} = \partial\log P_j/\partial\log\mathbf{x}_m = (\mathbf{x}_m/P_j)[\mathbf{1}(j=m) - P_jP_m]\,\beta \qquad (6)$$

To obtain an unweighted elasticity for the sample, the derivatives and elasticities are computed by averaging sample values. The empirical estimate of the elasticity is

$$\hat{h}_{jm} = \left(\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\hat{P}_j(q)}\left[1(j=m) - \hat{P}_j(q)\,\hat{P}m(q)\right]\right)\beta = \left(\sum_{q=1}^{Q} w(q)\,\hat{q}_{jm}(q)\right)\beta \qquad (7)$$

where $P_j(q)$ indicates the probability estimate for the q^{th} observation and $w(q) = 1/Q$. A problem can arise if any single observation has a very small estimated probability, as it will blow up the estimated elasticity. There is no corresponding effect to offset this. Thus, a single outlying estimate of a probability can produces unreasonable estimates of elasticities. To deal with this common problem, one should compute "probability weighted" elasticities, by replacing the common weight $w(q) = 1/Q$ with

$$w_j(q) = \frac{\hat{P_j}(q)}{\sum_{q=1}^{Q} P_j(q)} \qquad (8)$$

With this construction, the observation that would cause the outlying value of the elasticity automatically receives a correspondingly small weight in the average.

The parameter(s) of inclusive value(s) provides the basis for differences in cross-substitution elasticities as compared to the independently and identically distributed (IID) condition of the multinomial logit (MNL) model. The elasticity formulae for a nested logit model vary depending on whether an alternative (for a direct elasticity) or a pair of

alternatives (for a cross elasticity) are associated with the same branch of a nested partition. For the direct elasticity, it is identical to the MNL formula for alternative *m* which is *not* in a partitioned branch (eg it exists in a non-nested partition of tree). Where alternative *m* is in a partitioned part of the tree, the formula has to be modified to accommodate the correlation between alternatives within the branch. The NL direct elasticity for a partitioned alternative is:

$$[(1 - P_m) + \{\frac{1}{1- s_G}\}(1-P_{m|G})]\beta_k X_{mk} \quad (9)$$

The NL cross elasticity for alternatives m and m' in a partition of the nest is:

$$-[Pm + \{\frac{s_G}{1- s_G}\}P_{m|G}]\ \beta_k X_{mk} \quad (10)$$

The direct elasticities and marginal effects are summarised in Table 7. The marginal effects which define the partial derivative of the probability of mode choice with respect to an attribute of choice; suggest that price has a greater impact than travel time; however when an elasticity is calculated we found that linehaul travel time is slightly more elastic than cost for car for Sydney and the combined cities but less elastic for Melbourne. There appears to be no consistent trend in the ordering of direct elasticities between Sydney and Melbourne; for example, Melbourne commuters appear to be more sensitive to in-vehicle and parking costs compared to Sydney commuters, but the reverse applies for linehaul time except for drive alone.

Table 7: Summary of Nested Logit Training Models Marginal Effects and Direct Elasticities

Variable	Alternative	Syd-Mel	Sydney	Melbourne
Cost ($)	DA	-.93 (-6.3)	-1.14 (-7.99)	-1.87 (-12.11)
	RS	-1.73 (-5.68)	-1.59 (-5.77)	-2.20(-9.42)
	BS	-.43 (-3.61)	-.38 (-3.59)	-.57 (-4.18)
	TN	-.43 (-3.73)	-.42 (-3.42)	-.48 (-5.58)
	BW	-.54 (-4.55)	-.50 (-4.49)	-.65 (-5.42)
	LR	-.48 (-5.93)	-.46 (-4.67)	-.55 (-5.97)
LineHaul Time (mins)	DA	-1.01 (-.76)	-1.34 (-1.03)	-1.25 (-.91)
	RS	-1.89 (-.69)	-1.88 (-.74)	-1.48 (-.71)
	BS	-1.92 (-.44)	-1.85 (-.46)	-.35 (-.31)
	TN	-.45 (-.42)	-.65 (-.57)	-.32 (-.34)
	BW	-.52 (-.52)	-.72 (-.75)	-.41 (-.41)
	LR	-.52 (-.56)	-.74-.78)	-.39 (-.45)
Access & Egress Time	BS	-.74 (-.41)	-.97 (-.60)	-.62 (-.31)
	TN	-.33 (-.31)	-.30 (-.24)	-.27 (-.22)
	BW	-.37 (-.34)	-.33(-.31)	-.56 (-.41)
	LR	-.37 (-.25)	-.37 (-.32)	-.28 (-.29)
Parking Cost ($)	DA	-.42 (-1.11)	-.45 (-1.26)	-.73 (-1.79)

	RS	-.80 (-1.00)	-.63 (-.91)	-.86 (-1.39)

Note: Marginal effects are in brackets and multiplied by 100

Neural Network Approach To Commuter Choice

Basics of Neural Network Approach

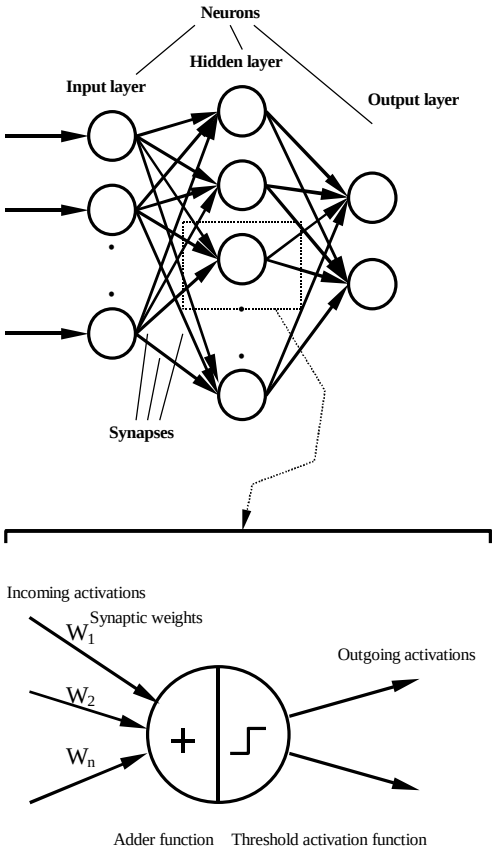
The term "neural networks" is used to describe a number of different models intended to imitate some of the functions of the human brain, using its basic structure. The history of neural network research started in 1943 when McCulloch and Pitts studied a collection of model neurons and showed that they were capable of calculating certain logical functions. Hebb, in a psychophysiological study published in 1949, pointed out the importance of the connection between synapses to the process of learning. In 1958, Rosenblatt described the first operational model of neural networks called perceptron. He put together the ideas of Hebb, McCulloch and Pitts. When two mathematicicans, Minsky and Papert, demonstrated the theoretical limits of the perceptron in 1969, the effect was dramatic: researchers lost interest in neural networks. The recent resurgence of interest in neural networks is largely due to individual contributions such as that of Hopfield, who showed the analogy between neural networks and certain physical systems in a 1982 study, bringing a rich and well understood formalism to bear on these networks. More recently, since 1985, new mathematical models have enabled the original limits of the perceptron to be greatly extended. Today, the first practical applications of neural networks are beginning to see the light of day, and the discipline is beginning to interest a larger and larger audience of students, researchers, engineers and industrialists.

The main motivations for using a neural network are parallelism, the capacity to learn, allowing for the use of distributed memory, capacity for generalisation and ease of computer simulation construction. Following these characteristics, one of the promises from neural networks is that they can tackle the problem of forecasting and modelling which is very common in travel demand modelling. The challenge is how to determine its suitability for traveller behaviour problems (in general and in particular to a specific problem and modeller's objectives and constraints), how to select and apply the relevant methods of neural networks, and the feasibility of an implemented system.

Structure and terminology

The basis of a neural network is to use artificial *neurons* to represent nerve cells. Neurons are the fundamental element of the human central nervous system. Neurons have five specialist functions: they receive signals coming from neighbouring neurons, they integrate these signals, they give rise to nerve pulses, they conduct these pulses, and they transmit them to other neurons which are capable of receiving them (Davalo and Naim, 1991). Figure 1 shows a typical neural network and a neuron's structure.

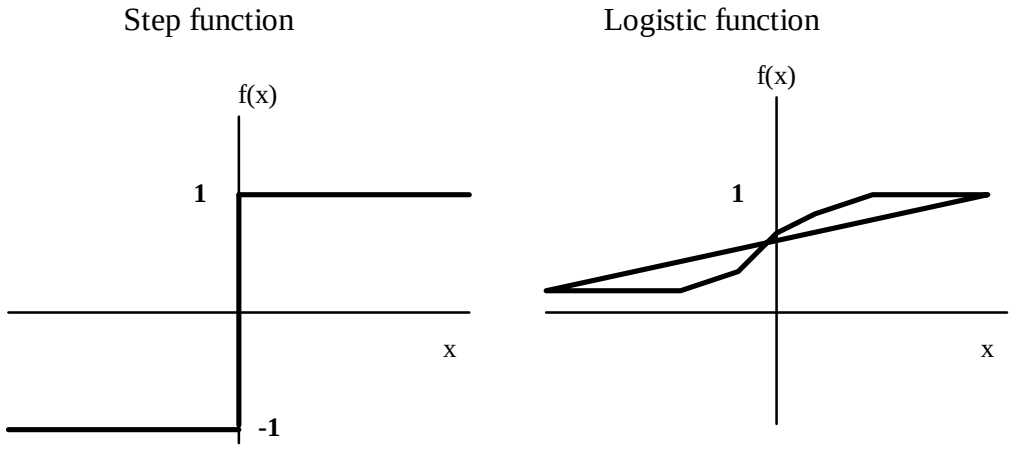
Figure 1: Structure of ANN and a Selected Neuron



Many different structures of neural networks may be used. However, the multi-layer network shown in Figure 1 represents the most popular structure. It represents the nervous system as successive layers of neurons. The two outermost layers correspond in one case to the layer which receives inputs from the external world and in the other to the layer which outputs the results of processing. The intermediate layers are called *hidden layers* and they may vary in number.

The neuron is the basic processor in neural networks. Each neuron has one output, which is generally related to the state of the neuron -its activation - and which may fan out to several other neurons. Each neuron receives several inputs over these connections, called synapses. The inputs are the activations of the incoming neurons multiplied by the weights of the synapses. The activation of the neuron is computed by applying a threshold function to this product. The threshold function is generally some form of non-linear function. Figure 2 describes two typical threshold functions: a step function (discrete) and a logistic function (continuous).

Figure 2: Two Typical Threshold Functions of Neural Networks



These functions can be represented mathematically as follows:

	Step function	Logistic function
f(x) = (11)	1 if x > 0 f'(x) if x = 0, where f'(x) refers to the previous value of f(x) (that is, the activation of the neuron will not change -1 if x < 0	$f(x) = 1/(1+e^{-x})$

where x is the summation (over all the incoming neurons) of the product of the incoming neuron's activation and the synaptic weight of the connection:

$$x = \sum_{i=0}^{n} Aiwi \qquad (12)$$

where n is the number of incoming neurons, A is the vector of incoming neurons, and w is the vector of synaptic weights connecting the incoming neurons to the neuron under study.

Key properties of neural networks

In the previous section, the basis of neural networks has been presented; it is now appropriate to discuss the key properties of neural networks.

i) *Parallelism*. Parallelism is fundamental in the architecture of neural networks when these are considered as sets of elementary units operating simultaneously. This parallelism in data processing is interesting because of the limitations of sequential methods of processing large problems needing an enormous quantity of data, sometimes giving rise to a combinatorial explosion in processing requirements. Through the use of parallel hardware, parallelism allows a greatly increased speed of calculation, but demands that problems to be resolved are stated, and thought of in a different, unconventional manner.

ii) *Capacity for Adaptation/ Learning.* This property of neural networks first manifests itself in their ability to learn, which allows networks to take account of new constraints or new data from the external world as they arise. Furthermore, it appears in certain networks by their capacity to self-organise, ensuring their stability as dynamic systems. This capacity for adaptation is particularly relevant for problems which evolve; these need to take account of situations which are not yet known in order to resolve problems. This may mean that the network is able to take account of a change in the problem that it is solving, or that it may learn to resolve the problem in a new manner.

iii) *Distributed memory.* In neural networks 'memory' corresponds to an activation map of the neurons; this map is in some ways a coding of facts that are stored. Memory is thus distributed over many units, giving a valuable property, resistance to noise. In the first place, the loss of one individual component does not necessarily cause the loss of a stored data item. This is different from the case of a traditional computer, in which individual data is stored in individual memory units, and in which the loss of one memory unit causes its data to be lost permanently. In a neural network the destruction of one memory unit only marginally changes the activation map of the neurons. Secondly, when one atomic piece of knowledge corresponds to one piece of data stored in a particular place, the problem of managing the full set of knowledge arises. In order to find or to use one particular fact, it is necessary to know precisely either its address or its contents. This technique cannot therefore take account of noisy data and preprocessing of data must therefore be used to eliminate the noise. This limitation is overcome in distributed memories such as neural networks, in which it is possible to start with noisy data and to make the correct data appear from the network's activation map without noise.

iv) *Capacity for Generalisation.* This capacity is crucial; its importance has been shown in recent years by the difficulty of acquiring rules for expert systems. Many problems are solved by experts in a more or less intuitive manner, making it very difficult to state explicitly the knowledge base and the rules which are necessary for its exploitation. It is therefore highly significant to consider a system which may learn the rules simply from a set of examples, or which may learn to mimic a behaviour (which is the case in travel choice modelling where there might be an

association between mode choice with associated attributes and the preferred mode selection), allowing the problem itself to be solved.

Using neural network models in representing commuter choice problems

Pattern association is the underlying mechanism of a multi-layer neural network. It enables the full set of human perceptions about a particular problem (such as travel choice preference) to be represented by neural networks. Pattern association can associate an input shape, pattern, representation of a concept or a situation, with other items, either of the same kind or totally different. All of the "knowledge" that a neural network possesses in the pattern association process is stored in the synapses, the weights of the connections between neurons. Once the knowledge is present in the synaptic weights of the network, presenting a pattern for input to the network will produce the correct output.

However, how does the network acquire that knowledge? This happens during "training". Pattern associations (between input and associated output) are presented to the network in sequence, and the weights are adjusted to capture this knowledge. The weight adjustment scheme is known as the learning law.

One of the learning methods formulated was Hebbian learning. Hebb formulated the concept of "correlation learning". This is the idea that the weight of a connection is adjusted based on the values of the neurons its connects:

$$\Delta w_{ij} = \alpha a_i a_j \tag{13}$$

where α is the learning rate, a_i is the activation of the ith neuron in one neuron layer, a_j is the activation of the jth neuron in another layer, and w_{ij} is the connection strength between the two neurons. A variant of this learning rule is the signal Hebbian law:

$$\Delta w_{ij} = - w_{ij} + S(a_i)S(a_j) \tag{14}$$

where S is a sigmoid or logistic function which has been presented above.

Since the learning method just described does not test the resultant weights to see if they yield acceptable output(s), this method is described as an *unsupervised learning method*. In general, an unsupervised learning method is one in which weight adjustments are not made based on comparison with some target output. There is no "teaching signal" feed into the weight adjustments. This property is known as *self-organisation*. Another form of training of neural networks which has gained in popularity is the *supervised learning*. Input-output patterns are presented one after the other to the neural network. The presentation of every input-output pattern to the network is called a *training cycle*. Each cycle might involve many iterations for the network to adjust its weights in an effort to match the desired output. This *error correction mechanism* can be expressed as follows.

$$\Delta w_{ij} = \alpha a_i [c_j - b_j] \quad (15)$$

where w_{ij} is the connection strength between the two neurons, α is the learning rate, a_i is the activation of the ith neuron, b_j is the activation of the jth neuron in the recalled pattern, and c_j is the desired activation of the jth neuron.

The selection of suitable learning methods and the error correction mechanism will identify the type of neural models to be used for a particular application. In general, back-propagation neural models demonstrate a multi-layer network with supervised learning. Back-propagation neural models are also the most popular networks in applications (Faghri and Hua 1991).

Training and testing neural networks represents the two major steps in the development of neural networks for any context. In training, the network is taught to produce the expected output for a given set of input patterns. The learning capacity of the network is built and evaluated during the training task.

Testing a neural network determines the ability of the network to generalise when presented with patterns on which it was not explicitly trained. In other words, for a given input, the network is tested to see if it can recall its knowledge of associative network towards the estimation of an accurate output within a specified tolerance. Capacities for learning (in training phase) and generalisation (in testing phase) represent the two key attractive properties for the study of travel behaviour. Patterns of travel attributes and associated travel choice(s) collected from an individual, a group of individuals and a whole sample can be used to train a number of different neural networks to mimic travel behaviour of individual, a group of individuals and a whole sample.

Specific neural network models and results

The ANN model building process for representing the commuter choice problem involves four major steps:

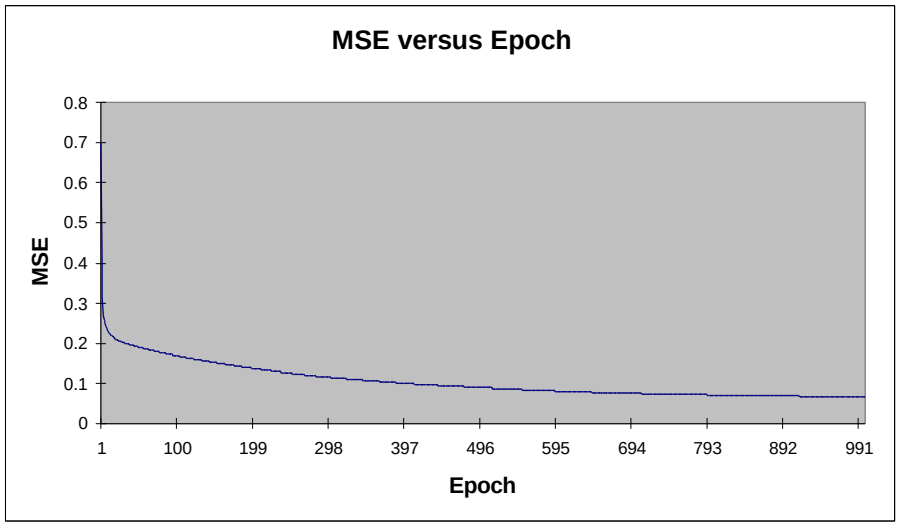
Step 1: This step specifies the structure of input and output layers. In other words, it focuses on the selection of the number of processing units (PEs) in input and output layers. This step is straight forward as it is constrained only by the number of input and output variables in the commuter choice problem. A total number of 12 PEs are used to represent the six mode choices and six associated attributes. These attributes are listed in Table 5 above. A total number of 6 PE s are used to represent the six mode choice decision vector in which at least one mode was chosen.

Step 2: This step focuses on the selection of the number of hidden layers, the associated number of processing units, and the interconnectivity between these layers and the input and output layers. The question of choosing the configuration for hidden layers is somewhat more difficult because the relationship between

neural network performance and the number and size of hidden layers is not well understood. Generalisation and convergence are two aspects of network behaviour which often work against one another. Generalisation is the ability of the network to produce reasonable results for novel or incomplete input data once the training process has been completed. Convergence is simply the ability of the network to learn the training data to within the error tolerance specified for the problem. In general, the more hidden neurons present, the greater the likelihood that the network will converge. However, if too many hidden units are used, the network will generalise poorly, "memorising" the training data rather than focusing on its significant features. The goal is to use as many hidden neurons as are needed to ensure convergence without using so many as to inhibit generalisation. In this research, a 30 PEs hidden layer was selected. This decision was taken after a number of test runs to test the sensitivity of the network performance due to the change in the size of the hidden layer (from 20 PEs to 40 PEs). The range from 20 PEs to 40 PEs was selected due to the size of input and output layers which have 12 PEs and 6 PEs, respectively. It was found that 30 PEs hidden layer provided the best result without taking so much computing time as with the 40 PEs hidden layer.

Step 3: Once the network configuration was established, each individual ANN model was trained with the data sets shown in Table 4 above. Each record in the selected training data set consists of 12 input variables and the associated vector of six values presenting the desired output. This desired output was used to correct the error that the network predicts during the training cycle. The error was then back-propagated from the output PEs to the input PEs via the process of adjusting the weight values which connects the output and input PEs. A sensitivity analysis of network performance in terms of mean square error (MSE) due to changes in the number of training cycle (called *epoch*) from 100 to 10000 epochs was carried out (see Figure 3). It was found that the MSE tends to stabilise at 1000 epoch. In terms of computing time a 1000 epoch training of a 641 records for the combined Sydney and Melbourne model took about two minutes.

Figure 3: Typical Training Curve for Melbourne Neural Network



Step 4: Once the network was trained, the testing of the trained network can then be proceeded. A number of models to be tested and associated data sets was presented in Table 4 above. The next section contrasts the results of this step with that of the nested logit models.

The direct elasticities are summarised in Table 8. Six attributes were selected for calculating direct elasticties: fuel cost, public transport fare, travel time by car, travel time by public transport, access and egress time for public transport, and parking cost.

Table 8: Summary of ANN Models – Direct Elasticities

Variable	Alternative	Syd-Mel	Sydney	Melbourne
Fuel cost	DA	-0.42	-0.87	-1.82
	RS	0.00	-2.22	0.00
PT fare	BS	0.80	2	2.50
	BW	-0.59	-1.18	2.50
	TN	0.00	0	0.00
	LR	-0.83	-1.43	0.00
Car time	DA	-0.83	0.00	-1.82
	RS	-1.11	-2.22	0.00
PT time	BS	0.00	0.00	2.50
	BW	0.59	0.00	5.00
	TN	0.00	0.00	-2.50
	LR	-1.67	-1.43	2.00
PT Access & Egress Time	BS	-0.80	2.00	0.00
	BW	0.59	0.00	-2.50
	TN	0.00	-2.22	2.50
	LR	0.00	-2.86	2.00
Parking cost	DA	-0.42	0.00	-0.61
	RS	0.00	-2.22	0.00

An inspection on the overall direct elasticities patterns reveal that:

- In terms of model transferability between different cities, all three (Sydney, Melbourne and combined Sydney and Melbourne) have quite different levels and distribution patterns of responsiveness to changes in attribute levels across all six attributes.

- Ignoring the sign (for the moment), the results are broadly consistent with the choice models in respect of the relative magnitudes between attributes.

In terms of the distributional profile of the impact from selected attributes, the results confirm the findings from the choice models. There is no consistent trend in the ordering of direct elasticities between Sydney and Melbourne; for example, Melbourne commuters appear to be more sensitive to in-vehicle cost and parking costs compared to Sydney commuters, but the reverse applies for line haul time except for drive alone. Melbourne ride-sharers (RS) seem to be insensitive to any changes from the six selected attributes. Detail inspection of the impact of each attribute on the six modes across the 3 contexts reveals a number of counter-intuitive results from ANN models. Specific comments are:

i) All three contexts (Sydney, Melbourne and combined Sydney and Melbourne) have responded as expected to the increase in fuel cost with a drop in DA (drive alone) and RS (ride share) demand (see Table 9). However, a counter-intuitive result arose in which the demand for bus use decreased in all 3 city contexts.

ii) In terms of the impact of public transport fares, there is no consistent trend in the sensitivity among the six modes of transport. There is a counter-intuitive result from ANN models with respect to bus use.

iii) In terms of the impact of travel time by car, there is a consistent pattern among the three cases with the exception of Sydney where DA was not sensitive to an increase in travel time, and Melbourne for ride share.

iv) In terms of public transport linehaul time, the dominating number of counter-intuitive positive elasticities is worrying and further raises concern about the behavioural validity of ANN models.

These findings suggest a behavioural weakness of the ANN approach. There appears to be a lack of an underlying model structure consistent with economic theory for representing behavioural responses to changes in the levels of attributes influencing travel choices. Previous studies have emphasised the classification power of ANN approach, and not the behavioural potential associated with change. The comparison between the predictive power of ANN and choice models is presented in the next section.

Comparison Of The Predictive Potential Of Neural Networks And Nested Logit Models For Commuter Mode Choice

A prediction success table is used as a format for comparing the prediction capability of both choice and ANN models. A detailed format of this table is shown in Appendix A. The prediction success tables for both choice and ANN models are shown as follows for Sydney, Melbourne and the pooled cities with full details on actual predictions available from the authors on request.

Table 9: Comparison between Sydney choice and Sydney ANN models

Case	Model	Predicted share less observed share						Weighted Percent correct	Weighted Success index
		DA	RS	BS	BW	TN	LR		
1. Sydney model on Combined Syd-Mel testing data	Choice	Better		Better	Better	Better	Better	Better	
	ANN		Better						Better
2. Sydney model on Syd testing data	Choice	Better	Better	Better	Same	Better	Better	Better	Better
	ANN				Same				
3. Sydney model on Mel testing data	Choice	Better	Better	Better	Better	Better	Better	Better	
	ANN								Better
4. Sydney model on Syd training data	Choice	Better	Better	Better	Better	Better	Same		
	ANN						Same	Better	Better

As shown in Table 9, choice models outperform ANN models in terms of the predicted share less observed share and weighted percent correct measures. In terms of the weighted success index ANN models perform reasonably well except case 2 with the Sydney model and Sydney testing data. The classification power of ANN models are reflected by their weighted success indices. An interesting finding is in Case 3 with the Sydney model and Melbourne testing data. In this case, even the choice model provides better weighted percent correct but it does not mean a better weighted success index is associated with it. In fact, the significant gain for ANN model in terms of the weighted success index comes from the contribution of the high percent correct for BW (busway) and LR (light rail), the two 'new modes'.

Table 10: Comparison between Melbourne choice and Melbourne ANN models

Case	Model	Predicted share less observed share						Weighted Percent correct	Weighted Success index
		DA	RS	BS	BW	TN	LR		
5. Mel model on Combined Syd-Mel testing data	choice	Same		Better		Better		Better	Better
	ANN	Same	Better		Better		Better		
6. Mel model on Syd testing data	choice		Better	Better	Better	Better		Better	Better
	ANN	Better					Better		
7. Mel model on Mel testing data	choice			Better	Better	Better	Better		
	ANN	Better	Better					Better	Better
8. Mel model on Mel training data	choice	Better	Better	Better		Better	Better		
	ANN				Better			Better	Better

In comparing the Melbourne choice and Melbourne ANN models, the trend continues for choice models in terms of getting better predicted share less observed share and weighted percent correct measures (see Table 10). Both choice and ANN models are equal in the measures of the percent correct and weighted success index.

Table 11: Comparison between Pooled Cities choice and Pooled Cities ANN models

Case	Models	Predicted share less observed share						Weighted Percent correct	Weighted Success index
		DA	RS	BS	BW	TN	LR		
9. Syd-Mel model on Combined Syd-Mel testing data	choice	Better	Same	Better	Better		Better		
	ANN		Same			Better		Better	Better
10. Syd-Mel model on Syd testing data	choice		Better		Better				
	ANN	Better		Better		Better	Better	Better	Better
11. Syd-Mel model on Mel testing data	choice	Better	Better	Better	Same	Same	Better	Better	
	ANN				Same	Same			Better
12. Syd-Mel model on Syd-Mel training data	choice	Better		Better	Better	Better	Better		
	ANN		Better					Better	Better

Table 11 confirms the finding from Tables 9 and 10. The strength of the choice model is clearly in the area of matching the predicted share and observed share whereas the ANN models are good at matching individual share.

Conclusions

The research reported in this paper is still in its preliminary stage. Further research will be carried out in terms of testing the performance of both approaches with more segmented data sets (eg different income groups, etc.), and the methodology in assessing the predictive capability of the two approach. There are a number of issues relating to the development of ANN. They are data, model validation and model structure issues. In terms of data issues, the back-propagation mechanism used in this research is certainly the supervised training algorithm. Supervised learning implies that the network requires a set of "good" pattern associations to train with. A good set of teaching facts are required. More research is required in terms of determining the impact of data quality on the performance of the specific choice neural network. Dia and Rose (1996) did look at this type of problem on the neural network for an incident detection system. On the model validation issue, the multi-layered ANN model used in this research might not be a good representation of the commuter choice problem. Even though, this kind of model is applicable to a wide class of problems. However, other ANN models should also be implemented in searching for the best representative model for the travel behaviour problem. In terms of model structure issue, more research is required to determine the appropriate structure for the network or its topology. In other words, it is about the input and output patterns and the number and size of hidden layers that are appropriate to the problem.

One important finding from this research is the confirmation of the predictive power of the choice modelling approach in matching the overall market share whereas the ANN models offer their contribution in matching individual market share. There is no clear indication as to which approach is better. However, we still consider the choice modelling approach continues to be used as a policy-based quantitative tool, but the ANN approach might have a role in the process of capturing existing mode choice preferences.

Acknowledgment

The research and development reported in this paper is supported under the Australian Research Council Research Centres program.

References

Davalo, E. and Naim, P. (1991) Neural Networks, translated by A. Rawsthorne, (Macmillan Computer Science Series: London).

Dia, H. and Rose, G. (1996) Impact of Data Quality on the Performance of Neural Network Incident Detection Models, Working Paper ITS-WP-96-17, Institute of Transport Studies, The University of Sydney and Monash University.

Faghri, A. and Hua, J. (1991) Evaluation of artificial neural network applications in transportation engineering, Transportation Research Record 1358, 71-80.

Hensher, D.A. and Louviere, J.J. (1998) A Comparison of Elasticities Derived from Multinomial Logit, Nested Logit and Heteroscedastic Extreme Value SP-RP Discrete Choice Models, paper presented at the 8th WCTR, Antwerp, Belgium.

Hensher, D.A., Louviere, J.J. and Swait, J. (forthcoming) Combining sources of preference data, Journal of Econometrics.

Hensher, D.A., Milthorpe, F.W. and Lowe, M.(1995) Greenhouse Gas Emissions and the Demand for Urban Passenger Transport: Final Report: Summary of Approach and Selective Results from Application of the ITS/BTCE Simulator, Report 8, Institute of Transport Studies, The University of Sydney, November.

Louviere, J.J., Hensher, D.A., Anderson, D.A., Raimond, T. and Battellino, H. (1994) Greenhouse Gas Emissions and the Demand for Urban Passenger Transport: Design of the Stated Preference Experiments, Report 3, Institute of Transport Studies, The University of Sydney, March.

Yang, H., Kitamura, R., Jovanis, P. P., Vaugh, K. M. and Abdel-Aty, M.A. (1993) "Exploration of route choice Behaviour with advanced traveler information using neural network concepts", Transportation 20, No.2, 199-223.

Appendix A
Prediction Success Table Format

Observed Choice	**Predicted Choice**			**Total Observed Count**	**Observed Share**
	1	2 …	J		
1	N_{11}	N_{12}	N_{1J}	$N_{1.}$	$N_{1.}/N_{..}$
2	N_{21}	N_{22}	N_{2J}	$N_{2.}$	$N_{2.}/N_{..}$
.					
.					
.					
J	N_{J1}	N_{J2}	N_{JJ}	$N_{J.}$	$N_{J.}/N_{..}$
Predicted Count	$N_{.1}$	$N_{.2}$	$N_{.J}$	$N_{..}$	1
Predicted Share	$\frac{N_{.1}}{N_{..}}$	$\frac{N_{.2}}{N_{..}}$	$\frac{N_{.J}}{N_{..}}$	1	
Proportion Successfully Predicted	$\frac{N_{11}}{N_{.1}}$	$\frac{N_{22}}{N_{.2}}$	$\frac{N_{JJ}}{N_{.J}}$	$\frac{N_{11}+\ldots+N_{JJ}}{N_{..}}$	
Success Index	$\frac{N_{11}}{N_{.1}}-\frac{N_{.1}}{N_{..}}$	$\frac{N_{22}}{N_{.2}}-\frac{N_{.2}}{N_{..}}$	$\frac{N_{JJ}}{N_{.J}}-\frac{N_{.J}}{N_{..}}$	$\sum_{i=1}^{J}\left[\frac{N_{ii}}{N_{..}}-\left(\frac{N_{.i}}{N_{..}}\right)^2\right]$	
Proportional Error in Predicted Share (Predicted Share Less Observed Share)	$\frac{N_{.1}-N_{1.}}{N_{..}}$	$\frac{N_{.2}-N_{2.}}{N_{..}}$	$\frac{N_{.J}-N_{J.}}{N_{..}}$		

(Source: Hensher and Johnson, 1981, Table 3.1, p.54)

INSTITUTE OF TRANSPORT STUDIES

The Australian Key Centre
in Transport Management

The University of Sydney
and Monash University