Using Neural Networks for Clustering-Based Market Segmentation

Harald HRUSCHKA
Martin NATTER

Forschungsbericht/
Research Memorandum No. 307
September 1992

Zusammenfassung

Die vorliegende Studie beschäftigt sich mit dem Einsatz künstlicher neuraler Netzwerke in der clusterbasierten Marktsegmentierung. Zur Lösung der dabei auftretenden Datenanalyseprobleme werden zwei Typen von Feedforward Netzwerken mit logistischen Aktivierungsfunktionen formuliert. Modelle des ersten Typs bestimmen Segmente auf der Grundlage von Segmentierungskriterien. Modelle des zweiten Typs bilden Segmente und differenzieren gleichzeitig zwischen diesen Segmenten auf der Grundlage zusätzlicher Segmentdeskriptoren. Die Parameter aller Modelle werden mit Hilfe einer erweiterten Version des Backpropagation-Verfahrens geschätzt.

Abstract

We study use of artificial neural networks in clustering-based market segmentation. To this end two types of feedforward neural networks with logistic activation functions are formulated. Models of the first type determine segments on the basis of segmentation criteria. Models of the second type simultaneously form segments and discriminate between these segments on the basis of additional segment descriptors. Parameters of all models are estimated by an extended version of backpropagation.

1 Introduction

Market segments represent more homogeneous divisions of a heterogeneous total market. In clustering-based segmentation segments are not known a priori. Instead segments have to be determined on the basis of a set of relevant variables (segmentation criteria) of respondents (Wind 1978).

Most segmentation studies proceed in two steps, determining segments in the first step and looking for discriminating characteristics (segment descriptors) in the second step. Data analytic methods applied in these steps are cluster analysis and discriminant analysis techniques, respectively. The large number of clustering algorithms available may be divided into hierarchical and partitioning (Jain and Dubes 1988) as well as overlapping (Arabie et al. 1981) and fuzzy methods (Hruschka 1986). Among discriminant analysis techniques linear and logistic methods are widespread in market segmentation. Contrary to the usual stepwise approach of segmentation studies, simultaneous classification and discrimination would be preferable, but there seems to be a shortage of appropriate techniques.

Artificial neural networks represent alternatives to better known statistical techniques. Certain types of artificial neural networks are closely related to well-known statistical methods. Linear discriminant analysis or principal components analysis, for example, are special cases of certain artificial neural network models. Therefore use of neural network models in market segmentation seems to be justified considering their greater generality. This article does not provide a general survey of neural networks, interested readers may consult the relevant literature (Rumelhart 1986a, Wasserman 1989, Hertz et al. 1991).

2 The Models

The models we consider are feedforward neural networks. In feedforward networks connections only run oneway, for example from input variables to hidden units and from hidden units to output variables. Hidden units differ from input and output variables by not being accessible from the outside world.

Neural networks with hidden units and nonlinear (transfer) functions are more powerful than linear models. They are able to form convex regions of the input space, whereas linear models can only separate the input space into hyperplanes.

Parameters (weights) of the neural network models discussed here indicate the strength of relationships between different variables (units). They are estimated by a variant of the so-called backpropagation algorithm.

2.1 Classification Models

The classification model shown in Figure 1 is a feedforward neural network using segmentation criteria both as input variables and output variables. Between input and output we put a layer of hidden units.

Theoretical values of segmentation criterion o for respondent p are calculated by a classification model in the following way

$$\hat{y}_{op} = f(\sum_h w_{oh} \quad g(\sum_i w_{hi} y_{ip})) \tag{1}$$

The weight w_{hi} measures the strength of the connection between a hidden unit h and segmentation criterion i as input variable. The weight w_{oh} measu-

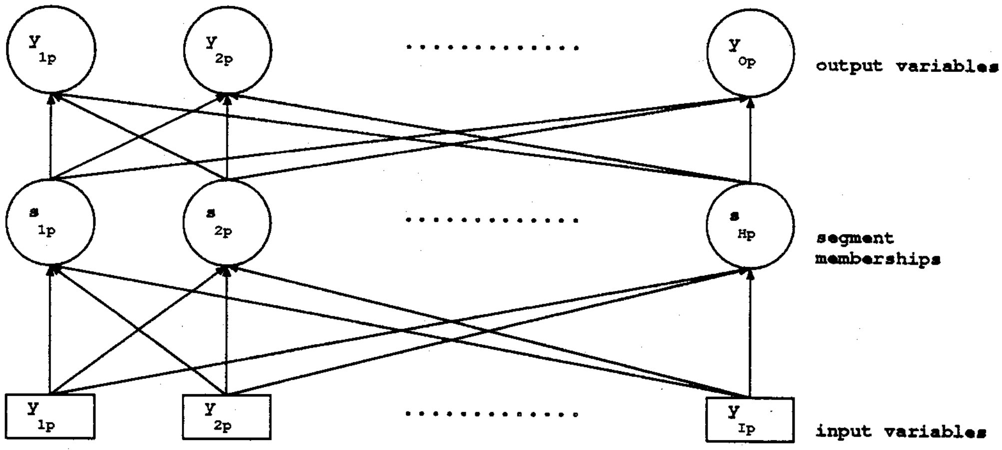

Figure 1: Classification Model

res the strength of the connection between segmentation criterion o as output variable and a hidden unit h. f and g are activation (transfer) functions.

As a rule, we use logistic activation functions. The logistic function (well known from econometrics or psychometrics) with argument z is

$$y = \frac{1}{1 + exp^{-z}} \tag{2}$$

The output values y of the logistic function lie in the unit interval $[0, 1]$. Its rate of change given a small change of z is

$$\frac{\delta y}{\delta z} = y(1 - y) \tag{3}$$

The logistic activation function leads to low values of this ratio if its output is near zero or one. It results in high values of this ratio if its output is near 0.5.

If we sum the products of all the inputs y_{ip} that have connections to h with w_{hi} and put this sum into the (logistic) transfer function g, we get the activation s_{hp} for hidden unit h. The output for criterion o, y_{op}, is calculated in a similar way: it is the sum of the products of the previously calculated hidden activations with w_{oh} put into the (logistic) transfer function f.

The full equation for computing the theoretical value $\hat{y}_{op}$ of criterion o for a respondent is

$$\hat{y}_{op} = \frac{1}{1 + exp^{\left(-\sum_h w_{oh} \left(\frac{1}{1+exp^{-\left(\sum_i w_{hi} y_{ip}\right)}} \right)\right)}} \tag{4}$$

The number of hidden units of the classification models is chosen to be smaller than the number of input (= output) variables. In this case the hidden units perform data reduction similar to principal component analysis. As a matter of fact, still using backpropagation to estimate weights but replacing the logistic by linear activation functions leads to results equivalent to principal components (Cottrell et al. 1988, Baldi and Hornik 1989). In other words, the classification models include principal components as a special case.

The values of each of the hidden units for a respondent show her(his) similarity to one of different classes of respondents. The values of these hidden units may be interpreted as measuring membership in different market segments. A respondent may be assigned to the m-th segment if for this respondent the m-th hidden unit has the largest value of all hidden units.

2.2 Classification and Discrimination Models

In the classification models only the segmentation criteria serve to determine segments. The task of the classification and discrimination models is more encompassing. These models simultaneously form segments (classes) of respondents and discriminate between these segments on the basis of additional segment descriptors.

Figure 2 demonstrates the structure of the classification and discrimination models. Each descriptor is connected to exactly one unit of a first hidden layer with K hidden units. Because of this restricted connectivity, these hidden units appear similar to latent variables occuring in linear structural equation models of the LISREL- or PLS-types (Bagozzi 1980, Fornell and Bookstein 1982). From now on we call these hidden units latent variables, following a definition of latent variables as being both causes or consequences of observable variables (James 1987).

The layer of latent variables inputs to each of H units of a second hidden layer. Moreover, this second hidden layer is also connected with the segmentation criteria in two ways, using them both as input variables and as output variables.

In these models there are two places where data compression is performed. The first place is the first hidden layer where we use the segment descriptors to form latent variables $l_{1p}, ..., l_{Kp}$. The second place is the second layer of hidden units which provide membership values $s_{1p}, ..., s_{Hp}$ for different market segments. Just like for the pure classification models, a respondent may be assigned to a segment on the basis of the maximum membership value. As a rule, we use logistic activation functions to compute values for hidden units (latent variables, segment memberships) as well as for output variables (segmentation criteria).

Given H segments and K latent variables and denoting each segment criterion in its role as input variable by y_{ip}, each descriptor by x_{jp}, the full equation for computing the theoretical value $\hat{y}_{op}$ of criterion o for respondent

p becomes

$$\hat{y}_{op} = \frac{1}{1 + exp\left(-\sum_h w_{oh} \left(\frac{1}{1+exp\left(-\left(\sum_i w_{hi} y_{ip} + \sum_k w_{hk} \frac{1}{1+exp^{-\sum_j w_{kj} x_{jp}}} \right) \right)} \right) \right)} \tag{5}$$

The weights of this model type are denoted by $w_{oh}, w_{hi}, w_{hk}, w_{kj}$. w_{oh} measures the strength of the connection between segment memberships and segmentation criterion o as output variable, w_{hi} between segmentation criteria as input variables and segment memberships, w_{hk} between latent variables and segment memberships, w_{kj} between segment descriptors and latent variables.

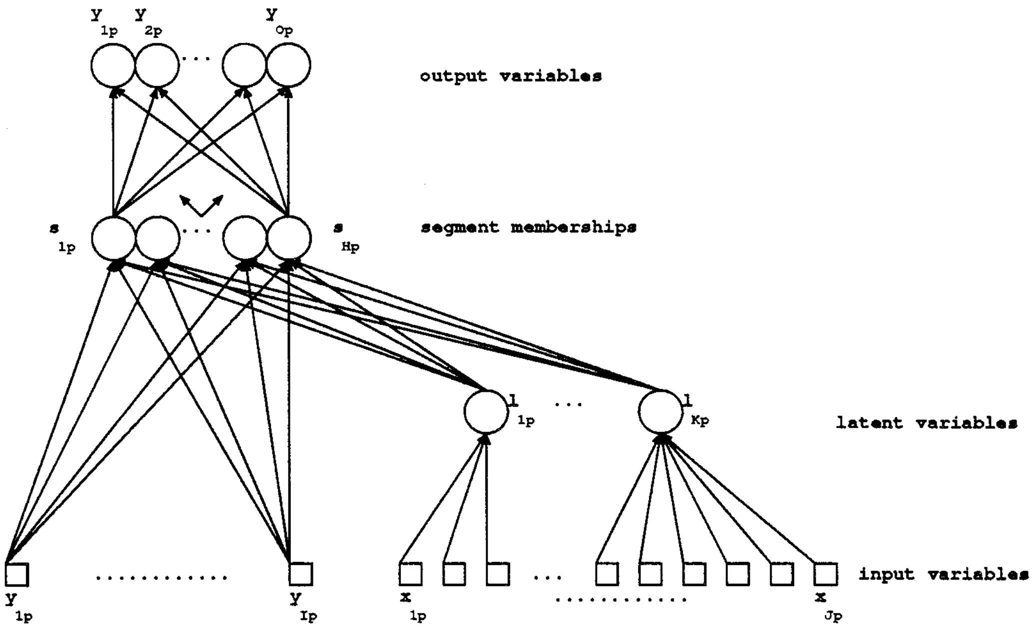

Figure 2: Classification and Discrimination Model

2.3 Estimation

Backpropagation is the most popular method to determine the parameters (weights) in feedforward networks (Rumelhart 1986b). In each of several iterations, adjustment of weights starts with the output units. Errors between actual and estimated state values are propagated layerwise backwards.

Backpropagation usually tries to minimize the error measure E which is half the sum of the squared differences between actual and computed outputs over all observations. In this case backpropagation is a least squares procedure.

$$E = \frac{1}{2}\Sigma_{p=1}^{P}(y_p - \hat{y}_p)^2 \tag{6}$$

Before starting the learning process, weights are initialized to small random numbers. The backpropagation algorithm proper runs for a number of iterations each with a forward and a backward pass. During the forward pass values of hidden units or output variables are determined layer after layer starting with the input units on the basis of the weighted summation and the activation functions.

In this study weights are initialized randomly in the interval $[-0.1, +0.1]$. The basic backpropagation method is extended by including a momentum term. Details on backpropagation and some of its extensions may be found in (Hertz et al. 1991).

The literature is controversial on importance and frequency of local minima using backpropagation. At least for the data analyzed here local minima did not occur.

Given y_{op} as segmentation criterion o for respondent p and its theoretical value $\hat{y}_{op}$ computed by one of the models, badness of fit is measured by the sum of squared errors(SSE)

$$\sum_p \sum_o (\hat{y}_{op} - y_{op})^2 \tag{7}$$

The decision on the number of segments is based on differences of SSE for consecutive numbers of segments. In other words, the number of segments is not increased, if decrease in SSE becomes small.

3 Pilot Application

The main objective of the study presented here is to gain a general understanding of usage patterns of household cleaners in Austria. Therefore usages of brands in different situations are chosen as segmentation criteria. The respondents are a representative random sample of 1007 housewives. After deletion of incorrect data and limitation to the more frequent brands and situations, the data base consists of 854 respondents.

7 different brands A, B, C, D, E, F, G of cleaners and 5 different usage situations $1, \cdots, 5$ (Table 1) are finally distinguished. This leads to 35 different usages $A1, A2, A3, A4, A5, B1, \cdots, G1, G2, G3, G4, G5$ that serve as segmentation criteria. $A1$ up to $G5$ are all binary variables, where $A1 = 1$ means that the respondent uses cleaner A in situation 1, $A1 = 0$ that the housewife does not use cleaner A in situation 1 etc.

Table 1: Usage Situations

1	Synthetic Surfaces
2	Lacquered Surfaces
3	Tiles
4	Ceramics, Enamel
5	Floors, Stairs

Descriptors that may explain possible segment membership of respondents are both psychographic variables (items measuring attitude towards housework or attitude towards cleaners) and sociodemographic variables (age, household size, number of children, income, housewife's education and occupation, second residence, location size, number of household members with income, household income). Table 2 gives an overview of the descriptors used, table 3 provides details on the psychographic items.

Table 2: Descriptors

Attitude Towards Housework
Attitude Towards Cleaners
Age
Household Size
Number of Children
Housewife's Education
Housewife's Occupation
Second Residence
Size of Household Location
Household Members with Income
Household Income

Table 3: Psychographic Items

(I1)	Cleaning the household is cumbersome
(I2)	It is better to buy products that save work even if they are a bit more expensive
(I3)	I appreciate it if my family helps with the housework
(I4)	If you do not see to it that the household is absolutely clean infections are probable
(I5)	Most of the cleaners are too sharp
(I6)	For specific chores in the household you need special cleaners
(I7)	I like to try new cleaners

4 Results

4.1 Classification Models

The classification models have the following structure:

35 input variables $A1, \cdots, G5$

H hidden units (segment memberships) ($H : 1, \cdots, 10$)

35 output variables $A1, \cdots, G5$

Table 4: Classification Models: Badness of Fit

Number of Segments	Number of Parameters	SSE
1	106	1988.9
2	177	1640.0
3	248	1383.2
4	319	1079.8
5	390	769.8
6	461	641.3
10	745	301.1

Table 4 demonstrates the change of SSE for different numbers of segments. The classification model for 5 segments is selected as SSE decreases become small for higher numbers of segments.

A comparable model for 5 segments but linear activation functions has the same number of parameters. This model may be rated as simpler because of its linearity, but it leads to a worse fit. The SSE of the linear model amounts to 1435.3 (the nonlinear model results in a SSE of 769.8). These results allow us to restrict interpretation to the five segment model with logistic activation functions whose nonlinearity is advantageous for the data at hand.

Table 5 shows the sizes of the five segments formed, i.e. the numbers of respondents in the different market segments. Each respondent is assigned to the segment in which she has the highest membership value.

Table 5: Classification Model: Segment Sizes

1	2	3	4	5
160	160	140	243	151

Table 6 characterizes each segment by those usages that are at least 75% more frequent in the segment compared to the whole sample and are indicated by at least 5% of the respondents.

Table 6: Characteristic Usages of the Segments

Segment 1	D1 F1
Segment 2	A1 A2 A3 A5 D1 D2 D3
Segment 3	A1 B1 B2 E3 F3 F4
Segment 4	G3 G4
Segment 5	B1 B2 B3 B5 C3 G2

4.2 Classification and Discrimination Models

Psychographic and sociodemographic variables are forward-connected with a first layer of 5 hidden units. The hidden units of the first layer are interpreted as latent variables ($l_1, .., l_5$) measuring the attitude of the housewife towards housework (l_1), the attitude of the housewife towards cleaners (l_2), the size of the location the housewife lives in (l_3), the family context (l_4) and the social status (l_5) of the housewife.

The 5 latent variables are forward-connected with a second layer of H hidden units. The other inputs of the second layer of hidden units are the usages $A1$ up to $G5$ (fully connected). This second layer of hidden units is also (fully) connected with usages $A1$ up to $G5$ representing the final or output layer.

Table 7: Classification and Discrimination Models: Badness of Fit

Number of Segments	Number of Parameters	SSE
4	364	1081.3
5	440	719.4
6	528	614.2

Table 7 demonstrates the change of SSE for various classification and discrimination models that differ by the number of segments. The model for 5 segments is selected because SSE decreases become small for higher numbers of segments.

Just like for the pure classification model a comparison to a linear model with the same structure shows significantly better results for the nonlinear model. The linear model leads to a SSE of 1411.32, the model with logistic activation functions to a SSE of 719.4. Moreover, model weights for connections with the latent variable have very low absolute values. Therefore the linear model does not give evidence to an influence of latent variables on segment memberships. That is why we restrict further interpretation to the nonlinear classification and discrimination model.

Table 8 shows the sizes of the five segments formed, i.e. the numbers of respondents in the different market segments.

Latent variable l_1 is formed by items I1, I2, I3 and I4. l_1 is postulated to measure the attitude of the housewife towards housework. The network inputs produce 15 different values for l_1. These 15 values can be aggregated to 3 value ranges (Table 9).

l_1 takes low values, if I3 and I4 are answered with yes. l_1 produces high values if these items are answered with no. Therefore l_1 measures how negative the attitude of a housewife towards housework is.

If we look at the segment memberships we can see that the impact of l_1 on segments 1 and 5 is not remarkable. On the other hand, high values of l_1 lead to higher membership values for segments 2 and 3. This means that housewives belonging to segment 2 (3) have a rather negative attitude towards housework. High values of l_1 go with low values for segment 4. So housewives belonging to segment 4 tend to have a positive attitude towards housework.

l_2 is determined by the values of items I5, I6 and I7. l_2 is formed to measure attitude towards cleaners. l_2 can take 8 different values which can be reduced to 3 value ranges (Table 10).

l_2 has low values if I5 is answered with yes and I6 is answered with no. l_2 has high values if I6 and I7 are answered with no. If I5 is answered with yes and only one of I6, I7 is 'no', l_2 has values belonging to the second level. l_2 measures how negative the attitude towards cleaners is.

High values of l_2 increase the memberships in segments 1 and 2. This can be interpreted as a tendency of housewives belonging to these segments for having a negative attitude towards cleaners. l_2 has virtually no impact on

Table 8: Classification and Discrimination Model: Segment Sizes

1	2	3	4	5
140	220	186	133	175

segment 3. If l_2 has values in the middle range, membership for segment 4 is likely to take high values. High values of l_2 result in low values of memberships for segment 5. Housewives belonging to segment 5 have a positive attitude towards cleaners.

l_3 is based on the size of the location (number of inhabitants) the respondent lives in. l_3 takes 4 values and increases with the size of the location. Table 11 shows the relationship between l_3 and segment memberships (e.g. members of segments 2 and 3 tend to live in bigger cities etc.).

l_4 should give some idea of the family context a housewife lives in. The following descriptors were measured

1. household size (one person, two persons, three persons, more than three persons)
2. age of the housewife (20-29, 30-39, 40-49, 50-59 years)
3. number of children (no child, one child, more than one child)

l_4 assumes 18 different values. High values of household size and age result in high values for l_4. Medium age results in medium values for l_4.

There is no impact of family context on segments 1 and 2. Segment memberships for segments 3 and 5 are high if l_4 is low. Housewives belonging

Table 9: Latent Variable l_1

value range	l_1	I1	I2	I3	I4
1	0.0 – 0.1			yes	yes
2	0.1 – 0.4				
3	0.4 – 0.8			no	no

to segment 3 are usually in the lowest age category, live alone and have no children. On the other hand, housewives of segment 5 typically are not in the highest age category. If l_4 is high, membership values for segment 4 are also high.

l_5 is postulated to measure the social class a housewife belongs to, using the following indicators

1. education (primary, vocational, secondary school)
2. occupation (full time, part time, no)
3. second residence (no, yes)
4. number of household members with an income (one person, more than one person)
5. income (3 income classes, 1 category for missing)

The indicators second residence and number of household members with an income are taken out of the model because eliminating them did not increase badness of fit to an important extent. The remaining indicators (school, occupation and income) produced 14 different levels for l_5. Values of l_5 near zero have the highest frequency. l_5 can be divided roughly into 3 value ranges.

Table 10: Latent Variable l_2

level	l_1	I5	I6	I7
1	0 – 0.15	no	yes	
2	0.15 – 0.25	yes		
3	0.5 – 0.7		no	no

If a housewife was in secondary school and has a part time job, the value of l_5 is high. If a housewife has no occupation and average household income, l_5 assumes a low value.

High values for l_5 increase membership values for segment 1 but reduce membership values for segments 3 and 5.

Table 12 summarizes the relationships between membership values of the five segments and all latent variables considered.

Table 13 describes each segment by those usages that are at least 75% more frequent in the segment compared to the whole sample and indicated by at least 5% of the respondents.

5 Conclusions

Neural network models include some of the better known data analytic methods used in marketing research as special cases. Selection of more traditional models or their nonlinear generalizations may be based on estimating the parameters of different, but related appropriate neural network models.

This article illustrates the structure of neural network models developed to

Table 11: l_3 and Segment Memberships

l_3	Segment
2	1
4	2
4	3
1	4
3	5

Table 12: Segment Memberships and Latent Variables

	1	2	3	4	5
l_1		neg.attitude tow.housework	rather neg.att. tow. housework	pos. att. tow. housework	
l_2	neg.att. tow. cleaners	neg.att.tow. cleaners	rather neg.att. tow. cleaners		pos.att.tow. cleaners
l_3	medium	high	high	low	medium
l_4	high		low	high	high
l_5	high	medium		medium	

Table 13: Characteristic Usages of the Segments

Segment 1	A1 A2 A3 A5 D1 D2 D3
Segment 2	G2 G3 G4
Segment 3	F1 G1
Segment 4	E3 F3 F4
Segment 5	B1 B2 B3 C1 C3

solve problems of clustering-based market segmentation. Of course, marketing research offers a lot of other possible applications. In a priori segmentation customer segments are given. Neural networks may then be used to discriminate segments on the basis of customer attributes (Mazanec 1992). As certain neural network models are generalizations of linear factor analytic methods, other possible applications are positioning studies intending to measure customer perceptions of brand attributes. Being alternatives to some of the usual conjoint analysis estimation methods, neural networks may also be used to test components of the marketing-mix.

Estimating the parameters of neural network models may be made difficult by local minima and high computing times. Alleviating these problems constitutes an area of active research. Our experience demonstrates that given a similar model specification computing times are comparable to those of better-known nonlinear estimation methods.

The capability to approximate nonlinear functions constitutes an important strength of neural networks. Models with nonlinear activation functions and hidden units may become useful additions to the marketing researcher's toolkit.

References

Arabie, P., J.D. Carroll, W. De Sarbo and J. Wind (1981), "Overlapping Clustering. A New Method for Product Positioning," *Journal of Marketing Research*, 18, 310–317.

Baldi, P. and K. Hornik (1989), "Neural Networks and Principal Component Analysis. Learning from Examples without Local Minima," *Neural Networks*, 2, 53-58.

Bagozzi, R.P. (1980), *Causal Models in Marketing.* New York: Wiley.

Cottrel, G.W., P. Munro, and D. Zipser (1988), "Image Compression by Back Propagation. An Example of Extensional Programming," in *Advances in Cognitive Science*, Vol. 3. N.E. Sharkey ed. Norwood, NJ: Ablex, 546–569.

Fornell, C. and F.L. Bookstein, "Two Structural Equation Models: LISREL and PLS Applied to Consumer-Voice Exit Theory," *Journal of Marketing Research*, 19, 440–452.

Hertz, J., A. Krogh and R.G. Palmer (1991), *Introduction to the Theory of Neural Computation.* Redwood City, CA: Addison-Wesley.

Hruschka, H. (1986), "Market Definition and Segmentation Using Fuzzy Clustering Methods," *International Journal of Research in Marketing*, 3, 117–134.

Jain, A.K. and R.C. Dubes (1988), *Algorithms for Clustering Data.* Englewood Cliffs, NJ: Prentice Hall.

James, L.R., S.A. Mulaik and J.M. Brett (1987), *Causal Analysis. Assumptions, Models and Data.* 4th Printing. Beverly Hills, CA: Sage.

Mazanec, J.A. (1992), "Market Segmentation with a Neural Network Model. Preliminary Findings," *Journal of Travel and Tourism Marketing*, 1, 39–59.

Rumelhart, D.E., G.E. Hinton and J.L. McClelland (1986), "A General Framework for Parallel Distributed Processing," in *Parallel Distributed Processing. Explorations in the Microstructure of Cognition*, Volume 1. Rumelhart, D.E. and J.L. McClelland eds. Cambridge, MA: MIT Press, 45–76.

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

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

Wind, Y. (1978), "Issues and Advances in Segmentation Research," *Journal of Marketing Research*, 15, 317–337.