인위적 데이터를 이용한 군집분석 프로그램간의 비교에 대한 연구

- A Research-In-Progress Paper -

김성호, 백승익, 최종연
한양대학교 경영학부

요약

인터넷 비즈니스나 전자상거래와 연관되어 고객관계관리 (Customer Relationship Management: CRM)가 널리 확산됨으로 해서 군집분석에 대한 관심이 한층 높아졌고, 다양한 군집분석 프로그램이 시장에 소개되어 지고 있다. 그러나, 군집분석 프로그램들은 다른 데이터 분석 기법과는 달리 그들의 정확성을 측정하기가 매우 힘들다. 본 논문에서는 이미 알려져 있는 군집구조를 지닌 인위적 데이터를 사용하여 반복적 군집분석 프로그램 (Convergent Cluster Analysis: CCA)과 보다 전통적인 단순군집 프로그램 (One-Shot Clustering Program: Howard-Harris 프로그램), 그리고 데이터 마이닝 기법 중의 하나인 데모그래픽 군집분석 프로그램의 정확성을 비교하기 위한 현재 진행 중인 연구의 방법론을 제시하는데 그 주요 목적을 두고 있다.

I. 서론

군집분석은 오랫동안 여러 분야에서 많은 연구가 (Joyce & Channon 1966; Green, Frank & Robinson 1967; Frank & Green 1968; Green & Krieger 1991) 되어졌을 뿐만 아니라, 때로는 수많은 논쟁을 불러 일으켰던 역사 (Morrison 1967; Neidell 1970)를 지니고 있다. 이런 과정을 거치면서 군집분석은 시장 조사와 시장 세분화 연구자들에 의해 구매자, 제품, 브랜드, 혹은 구매자의 구매상황 및 제품 사용상황 등을 기초로 몇 개의 동일한 집단으로 묶는 실증적 도구로 그 자리를 굳히고 있다. 특히 인터넷 비즈니스나 전자상거래와 연관되어 고객관계관리 (Customer Relationship Management: CRM)가 널리 확산됨으로 해서 군집분석에 대한 관심이 한층 높아졌다.

시장조사와 시장 세분화 연구자들은 구매자, 기업, 제품, 브랜드, 혹은 구매 상황 등의 객체를 기초로 몇 개의 소비자 집단으로 분류하기 위해서 여러 가지 군집분석 방법을 사용하여 왔다. 그 주된 방법으로서는 블록모형(Block Modeling; Arabie & Boorman 1982), 계층적 군집분석방법(Hierarchical Clustering Models; Blashfield 1976), 계층적 군집분석방법과 분할적 군집분석방법의 병용 (Combination of Hierarchical Clustering and Partitioning Methods; Milligan & Sokol 1980), 중복군집분석방법(Overlapping Clustering; Arabie, et al. 1981, Srivastava, et al., 1984), 그리고 혼합모형 (Mixture Models; Wolfe 1970, McLachlan & Basford

1988), 그리고 K-군집 중앙치을 이용한 중복군집 분석 (K-Cntroid Oerlapping Custering; Chaturvedi, et al., 1997) 등을 들 수 있다.

여러 가지 군집분석 방법 중 K-평균 분할적 군집분석 방법을 가장 널리 사용하고 있다. 현재 상업적으로 제공되는 K-평균 분할적 군집분석 프로그램의 예로서는 SAS의 FASTCLUS 프로그램, SPSS의 QUICK CLUSTER 프로그램, BMDP의 KM 프로그램, Howard-Harris(1966) 프로그램, 그리고 반복적으로 군집분석을 실행하여 (Replicated Clustering) 최적군집을 찾아내는 Sawtooth Software사의 PC용 군집분석 프로그램인 Convergent Cluster Analysis (CCA) 프로그램 (1988)과 메인 프레임용 프로그램인 CONCLUS (Helsen & Green 1991) 등이 있다.

시장조사와 시장 세분화 연구자들 사이에서의 K-평균 군집분석의 높은 인기에 대한 이유는 다음과 같이 설명 할 수 있다. 우선 마케팅조사와 시장 세분화 연구자들은 많은 수의 응답자와 변수를 다루는 것이 사실이다. 예를 들면, 소비자 패널 데이터, 제품스캐너로부터 수집되는 가구(家口) 단위의 구매데이터, 혹은 동일한 조사대상에 대하여 장기적으로 수집되는 라이프 스타일 (Lifestyle)에 관한 자료 등이 그것들이다. 이러한 데이터들은 수백 명 심지어는 수천 명의 응답자들로부터 수집되는 방대한 자료들이다. 또한 많은 시장조사 연구에서는 데이터를 수집하는 과정에서 응답자의 구매상황, 제품사용, 제품 혹은 브랜드 선호도, 인구통계적자료, 라이프스타일 등을 포함하는 여러 가지 다른 종류의 자료들을 하나의 설문지를 통해 수집하는 것이 일반적이다. 따라서 시장세분화의 연구자들이 다루는 자료에 포함된 변수의 수와 종류는 상당히 많은 것이다. 이 때 K-평균 군집분석은 수많은 응답자와 변수를 지닌 대규모의 자료를 빠른 시간 내에 효율적으로 분석할 수 있다. 또한 마케팅 관리자들은 세분시장의 수를 결정하고 각 세분시장의 구성원을 파악하는 과정에 있어서 중복군집 (Overlapping Cluster) 혹은 계층적 군집(Hierarchical Cluster)에 대하여 매우 불편하게 생각하는 경향을 그 이유로 들 수 있다. 그들은 효과적이고 효율적인 마케팅전략의 수립을 위하여 한 사람의 구매자가 오직 하나의 군집 (세분시장)에만 소속되기를 원하는 것이다. 그들은 또한 군집분석을 통하여 나타난 세분 시장들이 몇 개의 의미 있고 요약된 통계(Summary Statistics, 예를 들면 군집평균, Cluster Centroid)에 의하여 선명하게 설명되기를 원한다. 예를 들면, 마케팅 관리자들은 어느 세분시장을 공략할 것인가에 따라 광고주제를 개발하고 광고매체를 선택 할 수 있다. K-평균 군집분석은 서로 비슷한 크기의 군집을 형성하고 마케팅 관리자들이 쉽게 이용할 수 있는 군집(세분시장)의 특성들을 보여줌으로써 마케팅 관리자들의 이러한 욕구를 충족시켜 주는 방법이라 할 수 있다.

2 연구목적

본 연구의 주요 목적은 시장 세분화에 가장 많이 사용되고 있는 두 개의 K-평균 군집분석 프로그램 (Howard-Harris와 Convergent Cluster)과 K-평균 군집분석 방법이 아닌 IBM 사의 Intelligent Miner에 제공하는 데모그래픽 군집분석 프로그램 (Demographic Cluster Analysis)들의 성과를 인위적인 데이터를 사용하여 비교하는데 있다. 구체적으로 본 연구의 목적은 시장 세분화 및 마케팅조사에서 가장 빈번하게 사용되고 있는 두 개의 K-평균 군집분석 프로그램과 Intelligent Miner의 군집분석 프로그램을 군집의 구조가 이미 알려져 있는 인위적자료

에 사용하여 그 가치를 비교 평가하는 탐색적 연구를 수행하는데 있다. 본 연구에 사용된 군집분석 프로그램은 Howard-Harris 프로그램(1966), Sawtooth Software사의 CCA(Convergent Cluster Analysis; 1989), 그리고 IBM의 Demographic Clustering 프로그램이다. 본 연구의 구체적인 연구사항은 다음과 같다.

1. 세 개의 군집분석 프로그램 (CCA, Howard-Harris, Demographic Clustering) 중에서 어느 프로그램이 이미 알려져 있는 군집의 구조를 가장 잘 도출하는가?

2. 반복적 군집분석 프로그램 (CCA)과 전통적인 K-평균 군집분석 프로그램 (Howard-Harris), 그리고 데모그래픽 군집분석 프로그램간에 알려져 있는 군집 구조의 도출에 있어서 유의적인 차이가 있는가?

3. 세 개의 군집분석 프로그램과 군집의 수, 표본의 크기, 군집을 구성하는 변수의 수, 그리고 소음형태 등의 실험 요인 중 어느 것이 유의적인가?

4. 본 연구에 사용된 실험 요인간에 유의적인 상호작용이 있는가?

본 연구 논문에서는 위의 연구를 수행하기 위한 방법론을 제시하는데 그 주요 목적을 두고 있다.

3. 군집분석 프로그램

3.1 Howard-Harris 프로그램

Howard-Harris 프로그램은 P개의 변수에 의하여 정의된 N명의 응답자행렬(N x P matrix)를 Euclidean 거리척도를 사용하여 K개의 집단으로 분할하는 K-평균 군집분석 프로그램으로 현재 시장세분화에 가장 많이 쓰이고 있는 프로그램 가운데 하나이다. 응답자간의 Euclidean거리는 다음과 같이 구한다.

$$d_{ij}^2 = \sum_{t=1}^{P} (x_{it} - x_{jy})^2$$

식 (1)

d_{ij} = 응답자 i와 응답자 j 간의 Euclidean 거리척도
x_{it} = 응답자 i의 속성 t의 평가

이들 응답자간의 Euclidean거리의 합은 집단의 평균으로부터의 편차의 합에 직접적으로 비례한다.

$$\sum_{i=1}^{n_k} (x_i - \mu_k)^2 = \frac{1}{2} n_k \left[\sum_{i=1}^{n_k} \sum_{j=1}^{n_k} (x_i - x_j)^2 \right]$$

식 (2)

μ_k = 집단 k의 평균(group centroid)
n_k = 집단 k의 크기

모든 응답자로부터 구한 총분산은 다음과 같이 나타난다.

$$V_t = \frac{1}{2} N \left[\sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2 \right]$$

식 (3)

총분산은 집단 내 분산(V_w)과 집단간 분산(V_b)으로 나누어지며 각각의 집단 내 분산은 식 (2)에 나타나 있다.

$$V_t = V_b + V_w$$

식 (4)

따라서 집단 내 분산의 합은 다음과 같다.

$$V_w = \sum_{k=1}^{K} V_k$$

식 (5)

*N*명의 응답자를 *K*개의 집단으로 나누기 위한 기준은 단순히 집단내 분산(V_w)을 최소화하는 분할 *P(N,K)*를 구하면 된다. 본 연구에서는 PC용 Howard-Harris 프로그램 (Smith 1990)을 사용하였다.

3.2 Convergent Cluster Analysis(CCA) 프로그램

CCA는 PC용 군집분석 프로그램으로 사용자가 지정하는 회수(*M*회)의 군집분석을 실행한 후 각각의 군집분석 결과와 나머지 (*M*-1)개의 군집분석 결과를 군집의 재생성 (再生性, reproducability)을 사용하여 비교한 후, 가장 높은 평균재생성을 지닌 군집분석결과를 최적 (그리고 최종) 군집분석 결과로 선정하는 K-평균 군집분석 프로그램이다. 재생성이란 두 개의 군집분석 결과를 *K* x *K* 표(cross-tabulation table; *K*는 군집의 수를 나타냄)로 나타내어 대각선상에 위치하는 응답자들의 수의 합을 말한다. CCA와 같은 반복적 군집분석방법은 대부분의 K-평균 군집분석 프로그램들이 초기치에 민감하고 따라서 최종적인 군집분석의 결과가 동일한 자료에 대해서도 상이할 수 있다는 단점을 해결할 수 있다. CCA는 다음과 같은 다섯 가지의 초기치 설정 옵션을 가지고 있다.

1. Distance-Based: 이 방법은 비교적 멀리 떨어져 있는 응답자들을 초기치로 설정하는 방법으로 가장 빠르게 군집분석을 수행하는 옵션이다. 그러나 이 방법을 사용했을 경우 항상 동일한 군집분석의 결과가 나타나기 때문에 반복적으로 군집분석을 수행할 필요가 없다.

2. Hierarchical-Based: 데이터에 있는 응답자의 수가 50명 이상일 경우에는 50명의 응답자를 무작위 추출하여 계층적 군집분석 (complete linkage)을 수행하여 각 군집의 군집평균 (cluster centroid)을 초기치로 택하는 방식이다. 데이터의 응답자의 수가 50명 미만일 경우는 전체 데이터를 초기치 설정에 사용한다.

3. Density-Based: Hierarchical-Based 방법과 마찬가지로 초기치 설정을 위하여 50명의 응답자를 무작위 추출한다. 이 방법에서는 응답자들이 비교적 밀집된 지역의 중심에 가까이 위치한 응답자를 무작위 하부자료로 추출한다.

4. Mixed Strategy: 첫번째 군집분석은 Distance-Based 옵션을 사용하여 실행되고 홀수번째의 반복적 군집분석은 Density-Based 방법을 사용하여, 그리고 짝수 번째의 반복적 군집분석은 Hierarchical-Based 옵션을 사용하여 실행된다. Sawtooth사는 대부분의 K-평균 군집분석에 있어 이 방법을 사용할 것을 권하고 있다.

5. 사용자 정의 옵션: 이 방법에서는 프로그램 사용자가 제공하는 군집소속에 따라 계산된 군집평균을 초기치로

설정한다. 이 방법을 사용하기 위해서는 CCA의 실행 이전에 실행한 군집분석의 결과 (군집소속 데이터)가 필요하다.

3.3 데모그래픽 군집분석 마이닝 프로그램

Intelligent Miner는 모든 데이터 베이스에 있는 레코드들을 한 번에 두 개씩 비교하여 두 데이터 항목의 유사성을 계산하고 그 값을 기초로 하여 군집을 형성하게 된다 (Cabena, et al., 1997). 레코드들 사이의 유사성은 그 레코드의 필드 값을 비교하여 결정하게 된다. 그런 다음 동일한 군집내의 모든 레코드 유사성 쌍의 합에서 다른 군집내의 모든 유사성의 합을 뺀 값을 최대화하도록 연속적으로 군집을 만들게 된다.

4. 연구방법

군집분석의 연구자들은 Monte Carlo 방법을 이용하여 이미 알려져 있는 군집구조 (cluster structure)를 가진 인위적인 데이터 (synthetic data)를 분석함으로써 주어진 군집분석의 알고리즘 (프로그램)을 평가해 볼 수 있다. 그리고 일단 군집이 형성되면 다양한 형태와 양의 소음 (noise)을 추가함으로써 데이터의 질을 떨어뜨릴 수 있다 (degraded). 이 질이 떨어진 데이터를 이용하여 다양한 상황 하에서의 군집분석 프로그램의 정확도을 비교 분석해 볼 수 있다.

4.1 군집조성 프로그램

본 연구에서는 이미 알려져 있는 군집구조를 지닌 인위적인 데이터를 만들기 위해 Milligan(1985)에 의해 개발된 군집조성 프로그램 (cluster generation program)을 사용하였다. Milligan의 프로그램은 다음과 같은 특성을 지니고 있다.

- 군집조성 프로그램은 각 분석마다 2개에서 5개 사이의 군집을 조성한다.
- 군집들은 4개, 6개, 혹은 8개의 차원에 의하여 나타내어 진다.
- 각각의 군집은 50개, 100개, 150개, 혹은 200개의 개체(points)로 구성된다(군집의 크기).
- 소음(noise)을 추가함으로써 원래 군집을 구성하고 있는 변수의 좌표에 오차(error)를 첨가할 수 있다.
- 한 개, 두 개, 혹은 세 개의 소음차원(변수; noise dimension)을 추가함으로써 전체 데이터에 오차(error)를 첨가할 수 있다.
- 데이터크기의 20% 혹은 40%에 해당되는 outlier를 추가로 데이터에 포함시킬 수 있다.
- 군집의 구조가 전혀 없는 데이터 (무작위자료)도 형성할 수 있다.

이 프로그램은 다변량 정상분포 (truncated multivariate normal distribution)로부터 표본을 추출함으로써 군집을 구성하는 점들의 좌표를 도출해 낸다. 이 과정에서 각 군집의 군집평균(cluster centroid)의 ±1.5 표준편차 내에 있는 모든 점으로부터 특정한 군집을 구성하는 점들을 도출해 낸다. 그리고 군집을 나누는 기준에 의해 군집간에 서로 겹치는 부분은 없도록 군집을 구성한다.(non-overlapping clusters). Milligan의 군집조성 프로그램은 기본적으로 구형(spherical)의 군집을 도출해 낸다. 본 연구에서는 서로 같은 크기의 군집을 도출해 내는 옵션을 이용하였다.

군집을 수성하는 변수에 오차(error)가 첨가된 좌표값(E_{ij})은 다음과 같다.

$$E_{ij} = T_{ij} + \lambda \varepsilon_{ij}$$

T_{ij} = 군집을 구성하는 점 i의 j번째 오차없는 기본좌표 (error-free coordiante)
ε_{ij} = 무작위 추출된 오차(random error)
λ = 소음의 강도

4.2 실험설계

본 연구에서는 인위적 데이터의 Monte Carlo실험을 위하여 다음과 같은 실험 변수들을 사용하여 실험을 설계하였다.

1. 자료의 크기 (2): 100; 200
2. 군집의 수 (3): 2개; 3개; 4개
3. 군집을 구성하는 변수의 수 (3): 4; 6; 8
3. 오차의 형태 (6)
   a. 기본좌표의 혼란: λ=0(error free situation); λ=1; λ=2
   b. 추가적인 소음차원(additional noise demension): 1차원; 2차원; 3차원
4, 군집분석 프로그램 (3): Howard-Harris; CCA; Demographic Clustering

위의 실험 변수들을 사용한 교차 디자인 (full factorial design)으로부터 Monte Carlo 실험을 위한 108개의 데이터를 구성하였다. 각각의 데이터는 3개의 K-평균 군집분석 프로그램에 의하여 분석되었다.

4.3 군집구조의 추출(recovery) 측정척도

본 연구에서는 두 개의 군집 (알려져 있는 군집구조와 군집분석 프로그램을 사용하여 도출한 군집구조)간의 일치도를 알아보기 위하여 Adjusted Rand Index (Hubert and Arabie 1985; 이하 ARI)를 이용하였다. Rand Index란 두 개의 빈도간의 비율로서 분자(分子)는 동일한 두 명의 응답자 쌍이 동일한 군집에 소속되어 있는가 혹은 상이한 군집에 소속되어 있는가의 빈도수이며 분모(分母)는 전체 응답자 쌍의 수이다. 예를 들면, 전체 응답자의 수가 N일 경우 분모는 $N(N-1)/2$이다. 만일 두 개의 군집이 서로 정확하게 일치한다면 Rand index는 1.0이다. 만일 군집의 구성원들간에 일치가 전혀 이루어지지 않는다면 Rand index는 0이다. 그러나 원래 Rand index에는 상향적 오차(upward bias)가 존재하므로 보통 그 오차를 수정하여 사용되고 있는 것이 ARI(Hubert & Arabie 1985)이다. Milligan & Cooper(1986)는 그들의 폭 넓은 군집분석 시뮬레이션 연구를 통하여 원래의 Rand Index를 더 이상 군집분석 연구에서 사용하지 않을 것을 권하고 있다. [표 1]은 Rand와 ARI에 관한 아이디어와 그를 구하는 수식을 보여주고 있다.

[표 1] Rand와 Adjusted Rand Index

True Structure (알려져 있는 군집구조)	**Test Structure** (군집분석을 통하여 발견한 군집구조)		
	동일군집에 속한 개체의 쌍	상이한 군집에 속한 개체의 쌍	합 .
동일군집에 속한 개체의 쌍	A	B	A+B
상이한 군집에 속한 개체의 쌍	C	D	C+D
합	A+C	B+D	R

Original Rand Index:

$$\frac{(A+D)}{R} = \frac{(A+D)}{\frac{1}{2}N(N-1)}$$

Adjusted Rand Index

$$\frac{R(A+D)-[(A+B)(A+C)+(C+D)(B+D)]}{N^2-[(A+B)(A+C)+(C+D)(B+D)]}$$

$$R = \frac{1}{2}N(N-1)$$

5. 요약

본 논문에서는 이미 알려져 있는 군집구조를 지닌 인위적 데이터를 사용하여 반복적 군집분석 프로그램 (Convergent Cluster Analysis: CCA)과 보다 전통적인 단순군집 프로그램 (One-Shot Clustering Program: Howard-Harris 프로그램), 그리고 데이터 마이닝 기법 중의 하나인 데모그래픽 군집분석 프로그램의 정확성을 비교하기 위한 현재 진행 중인 연구의 방법론을 제시하는데 그 주요 목적을 두고 있다. 본 연구 논문 발표에서는 제시된 연구 방법론을 사용하여 평가되어진 결과를 발표할 예정이다.

참고문헌

김성호 (1998), "컨조인트 최적제품 포지셔닝모형을 이용한 시장세분화에 관한 연구," 마케팅연구, 제13권 2호, 103-28.

Arabie, Phipps, J. Douglas Carroll, Wayne S. DeSarbo and Yoram Wind (1981), "Overlapping Clusterng: A New Methodology for Product Positioning," *Journal of Marketing Research*, 18, 310-7.

Berl, Janet, Gordon Lewis, and Rebecca Sue Morrison (1976), Applying Models of Choice to the Problem of College Selection, in Carroll, J. S. and Payne, J. W. (eds.), *Cognition and Social Behavior*, Hilsdale, NJ: Lawrence Elbaum Associates, 203-19.

Blashfield, R. K. (1976), "Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods," *Psychological Bulletin*, 83, 377-88.

Blashfield, R. K. and M. S. Aldenderfer (1978), "Computer Programs for Performing Iterative Partitioning Cluster Analysis," *Applied Psychological Measurement*, 2, 533-541.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1997), Discovering Data Mining, Prentice Hall.

Chaturvedi, Anil, J. Douglas Carroll, Paul E. Green, and John A. Rotondo (1997), "A Feature-Based Approach to Market Segmentation via Overlapping K-Centroids Clustering," *Journal of Marketing Research*, 34 (August), 370-7.

DeSarbo, Wayne S., J. Douglas Carroll, Linda Clark, and Paul E. Green 91984), "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables," *Psychometrika*, 49, 59-78.

Dickenson, J. R. (1986), *Bibliography of Marketing Research Methods*, Lexington, MA: Lexington, 580-97.

Forgy, E. W. (1965), "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications," abstract in *Biometrics*, 21, 768.

Frank, R. E. and Paul E. Green (1968), "Numerical Taxonomy in Marketing Analysis: A Review Article," *Journal of Marketing Research*, 5, 83-98.

Green, Paul E., Frank J. Carmone, and Jonathan Kim (1990), "A Preliminary Study of Optimal Variable Weighting in K-Means Clustering," *Journal of Classification*, 7, 271-85.

Green, Paul E., R. E. Frank, and P. J. Robinson (1967), "Cluster Analysis in Test Market Selection," *Management Science*, 13, B-387-400.

Green, Paul E. and Abba Krieger (1991), "Segmenting Markets with Conjoint Analysis," *Journal of Marketing* 55, 20-31.

Hartigan, J. A. and M. A. Wong (1979), "Algorithm AS136: A K-Means Clustering Program," *Applied Statistics*, 28, 100-28.

Helsen, Kristiaan and Paul E. Green (1991), "A Computational Study of Replicated Clustering with an Application to Market Segmentation," *Decision Science* 22, 1124-41.

Howard N. and B. Harris (1966), "A Hierarchical Grouping Routine, IBM 360/65 FORTRAN IV Program," Philadelphia: University of Pennsylvania Computer Center.

Hubert, Laawrence and Phipps Arabie (1985), "Comparing Partitions," *Journal of Classification*, 2, 193-218.

Jancey, R. C. (1966), "Multidimensional Group Analysis," *Australian Journal of Botany*, 14, 127-30.

Joycce, T. and C. Channon (1966), "Classifying Market Survey Respondents," *Applied Statistics*, 15, 191-215.

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley*

Symposium on Mathematical Statistics and Probability, Vol. 1, 231-297.

McIntyre, R. M. and R. K. Blashfield (1980), "A Nearest Centroid Technique for Evaluating the Minimum-Variance Clustering Procedure," *Multivariate Behavioral Research*, 15, 225-38.

McLachlan, G. J. and K. E. Basford (1988), *Mixture Models: Inferences and Applications to Clustering*, New York: Marcel Dekker.

Milligan, Glenn W. (1985), "An Algorithm for Generating Artificial Test Clusters," *Psychometrika*, 50, 123-7.

Milligan, Glenn W. and Lisa M. Sokol (1980), "A Two-Stage Clustering Algorithm with Robust Recovery Characteristics," *Educational and Psychological Measurement*, 40, 755-9.

Milligan, Glenn W. and Martha C. Cooper (1986), "A Study of Comparability of External Criteria for Hierarchical Cluster Analysis," *Multivariate Behavioral Research*, 21, 441-58.

Neidell, L. A. (1970), Procedures and Pitfalls in Cluster Analysis, Proceedings, Fall Conference, Chicago: American Marketing Association.

Sawtooth Software (1990), *CCA System for Convergent Cluster Analysis*, Ketchum, ID: Sawtooth Software.

Sawtooth Software (1986), *ACA System for Adaptive Conjoint Analysis,* Ketchum ID: Sawtooth Software.

Srivastava, R. K., Mark I. Alpert, and Allan P. Shocker (1984), "A Customer-Oriented Approach for Determining Market Structures," *Journal of Marketing Research*, 48, 32-48.

Ward, J. H. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58, 236-44.

Wind, Yoram (1978), Issues and Advances in Segmentation Research, *Journal of Marketing Research*, 25 (August), 317-37.

Wolfe, J. H. (1970), "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research*, 5, 329-50.