자료융합을 이용한 누락치 추정 방법에 관한 연구
- 고객관계관리를 위한 새로운 방법론 -

김성호[1], 조성빈[2], 백승익[1]

[1] 한양대학교 경영학과, [2] 충북대학교 경영학과

요약

자료융합 (Data Fusion)은 두 개 이상의 표본집단으로부터의 서로 다른 설문지 조사 자료를 혼합하여 하나의 통합되어진 자료를 만들어 내는 과정이다. 본 연구에서는 고객관계관리를 위해서 효율적으로 고객정보를 수집하고 관리할 수 있는 하나의 방법론으로서 자료융합 (Data Fusion) 방법을 제시하고, 그 방법의 유용성을 평가하고자 한다. 좀 더 구체적으로 말하면, 상관계수를 이용하여 기증자를 찾는 자료융합 방법과 Correspondence Analysis를 이용하여 기증자를 찾는 자료융합 방법의 정확도를 비교하는데 그 주요 목적을 두고 있다.

I. 서론

오늘날 많은 기업들은 새로운 경영전략기법으로서 고객관계관리 (Customer Relationship Management: CRM)를 서둘러 도입하고 있다. 고객이 원하는 것이 무엇인지를 파악하고, 그 요구에 맞는 서비스와 제품을 적시에 제공하는 것이 오늘날 급변하는 시장에서 기업이 필요로 하는 가장 중요한 생존 전략일 것이다. 그러나, 많은 기업들은 고객의 욕구를 분석하고, 인터넷 상에서 다양한 마케팅 전략을 수행할 수 있는 정보시스템 도입에 막대한 노력과 투자를 하였음에도 불구하고, CRM을 위한 이러한 기업의 노력은 영업활동의 성과에는 그다지 크게 영향을 미치지 못하고 있는 것이 현실이다. 효율적인 CRM을 위해서는 고객과 관련된 모든 정보를 획득하고, 그것을 관리하고 활용할 수 있는 기업의 능력이 필수 요건일 것이다. 고객에 대한 전체적이고도 명확한 그림을 그리기 위해서 필요한 고객정보를 획득하고 관리하는 과정에는 상당한 비용과 시간이 소요된다. 만일 데이터가 고객에 대한 최신의 정보를 정확하게 반영하지 못한다면 이를 기반으로 한 의사결정은 오히려 역효과를 가져올 수도 있다. CRM을 서둘러 도입하고 있는 많은 기업들은 고객분석의 입력 자료가 되는 고객정보를 획득하고 관리하는 과정에 대해서는 상대적으로 적은 관심을 보이는 경향이 있다. CRM을 위한 정보시스템 도입에 앞서 기업 내·외에 산재해 있는 고객정보의 수집과 통합이 먼저 수행되어져야 할 것이다. 본 연구에서는 효율적으로 고객정보를 수집하고 관리할 수 있는 하나의 방법론으로서 자료융합 (Data Fusion) 방법을 제시하고, 그 방법의 유용성을 평가하는 데 그 주요 목적을 두고 있다.

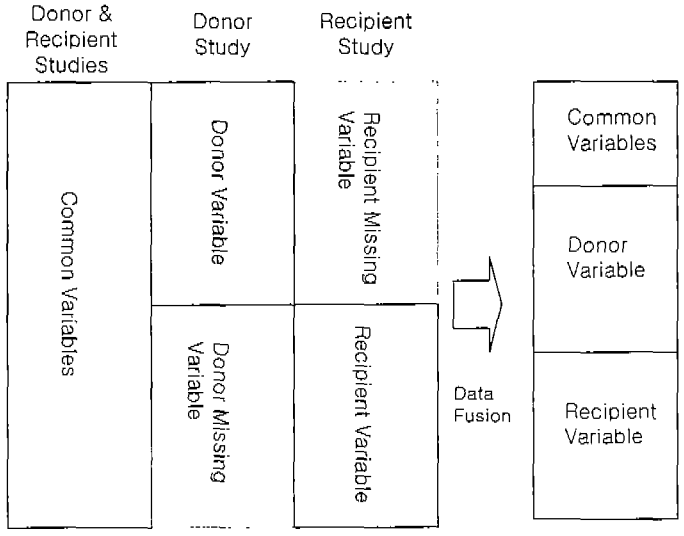
II. 자료융합

자료융합 (Data Fusion)은 두 개 이상의 표본집단으로부터의 서로 다른 설문지 조사 자료를 혼합하여 하나의 통합되어진 자료를 만들어 내는 과정이다. 이 방법은 커뮤니케이션 분야에서 누락되지 않은 정보를 기초로 누락된 정보(Missing Value)를 추정하는 하나의 방법으로서 (Baker et al., 1989), 다차원 척도법 (Multidimensional Scaling: Kruskal & Wish, 1978)의 특별한 경우인 Correspondence Analysis를 사용하여 누락치를 추정하는 방법이다. 자료융합은 커뮤니케이션 분야에서 뿐 만 아니라 설문지를 통한 마케팅 조사 등에서 빈번하게 발생하는 누락치를 추정하는 데에도 유용하게 사용될 수 있다.

자료융합에서는 응답자 표본을 기증자 (Donor)와 수혜자 (Recipient)로 구분한다. 기증자란 어떤 특정한 수혜자에게 그가 가지고 있는 누락치에 대한 대체값 (추정값)을 제공해 주

는 응답자를 말하며, 수혜자란 기증자로부터 자신이 가지고 있는 누락치에 대한 추정값을 제공받는 응답자를 말한다. 자료융합에서 각각의 수혜자는 우선 자기로부터 가장 가까운 거리에 있는 기증자를 찾는다. 기증자를 찾는 방법은 p개의 공통변수 (Common Variable)로 구성된 p 차원의 공간에서 특정한 수혜자로부터 다른 응답자들 (잠재적 기증자)과의 거리를 계산하여 그 수혜자로부터 가장 가까운 거리에 위치한 응답자를 기증자로 선택하게 된다. 일반적으로 공통자료로 응답자의 인구 통계적 정보 (Demographic Information)를 사용한다. 일단 특정 수혜자가 자신으로부터 가장 가까운 거리에 있는 기증자를 찾으면 그 기증자가 가지고 있는 변수들의 값이 수혜자의 누락치에 대한 추정치가 된다. 만일 이 기증자 역시 수혜자와 마찬가지로 특정한 변수에 대한 누락치를 지니고 있으면 그 다음으로 가까운 거리에 있는 응답자를 수혜자로 찾는다. 수혜자와 기증자와의 거리를 계산하기 위해서 일반적으로 상관계수 (Correlation Coefficient)와 Correspondence Analysis를 사용할 수 있다. 만약 수혜자나 기증자의 비누락치 (Non-Missing Value)로 구성된 공통자료가 응답항목별 자료 (Categorical Variables)라면 어떤 특정한 차원을 지닌 공간내에서 각각의 수혜자에 대한 기증자를 파악하기 위하여 응답자간의 거리 (예를 들면 Euclidean Distance나 Mahalanobis Distance), 혹은 응답자간의 상관계수를 직접적으로 계산하는 것은 불가능하다. 이러한 경우, 응답자간의 거리를 계산하기 위하여 사용되는 것이 Correspondence Analysis이다. [그림 1]은 자료융합 과정을 설명해 주고 있다. 본 연구에서는 상관계수를 이용하여 기증자를 찾는 자료융합 방법과 Correspondence Analysis를 이용하여 기증자를 찾는 자료융합 방법의 정확도를 비교하는데 그 주요 목적을 두고 있다.

[그림 1] 자료융합 과정



III. Correspondence Analysis

Correspondence Analysis (Hoffman and Franke 1986; Carroll et al., 1986, 1987)란 다차원 척도법(Multidimensional Scaling; Carroll & Arabie 1980)의 일종으로서 분석자료의 종류에 있어서 일반 다차원 척도법과 구별되는 분석기법이다. 일반적으로 다차원 척도법에 사용되는 입력자료의 종류는 다차원 척도법의 유형에 따라 등간척도 혹은 비율척도를 사용한 Rating

Data (Metric MDS의 경우)와 서열척도를 사용한 Ranking Data (Non-Metric MDS의 경우)로 나뉘어진다. 이에 비하여 Correspondence Analysis의 경우는 입력자료가 Dummy변수를 포함한 명목 척도이거나 혹은 응답 항목(Dichotomous이거나 Multichotomous)에 따른 응답의 빈도수(frequency)라는 점에서 일반적인 다차원 척도법과 구분되어진다. 따라서 Correspondence Analysis란 일반적으로 N-Way Contingency Table 혹은 Cross-Tabulation Table을 분석하는데 사용되는 기법이다.

Correspondence Analysis 역시 다른 다차원 척도법과 마찬가지로 입력자료에 나타난 개체 (소비자, 제품, 기업, 제품의 사용상황 등)들을 몇 차원의, 일반적으로 p-차원의 공간에 점(point)으로 나타내는 기법이다. 이 과정에서 입력자료에 나타난 개체들간의 상대적인 관계를 Correspondence Analysis를 통하여 구성한 공간에서도 동일하게 유지하여 나타낸다는 것이 Correspondence Analysis (다른 다차원분석기법과 마찬가지로)의 특징이라고 할 수 있다. 또한 Correspondence Analysis에서는 입력자료의 가로와 세로에 나타난 개체들을 동시에 동일한 공간에 점으로 나타내는 Joint Space분석기법의 일종이라고 할 수 있다. 또한 Correspondence Analysis의 알고리즘에 따라서는 가로와 세로의 개체들간의 거리가 직접 비교가능 할 수도 있다 (Carroll, Green & Schaffer 1986, 1987참조).

Carroll et al., (1986, 1987)의 알고리즘에 따르면, 우선 분석의 대상이 되는 Contingency Table혹은 Cross Tabulation Table matrix F행렬 (F는 가로 I줄과 세로 J줄로 되어 있다고 하자)를 다음과 같이 H행렬로 정상화 (normalize)한다.

$$H = R^{1/2} F C^{1/2}$$

R행렬은 Ix Idiagonal matrix이며 C행렬는 Jx Jdiagonal matrix이다. 이들 R과 C는 각각 가로와 세로의 합계의 제곱근의 역수(Reciprocals of the Square Roots of Row and Column Marginal)로 구성되어 있다. 다음으로 H행렬은 다음의 식을 사용하여 chi-square 거리척도로 전환된다.

$$H = P \Delta Q'$$

여기에서 $PP = QQ = I$이며 Δ는 Diagonal Metric이다. 끝으로 p-차원상에서의 가로줄(X)과 세로줄(Y)에 나타난 개체의 좌표는 각각 다음과 같다.

$$X = R^{1/2} P(\Delta + I)^{1/2}$$

$$Y = X^{1/2} R(Q + I)^{1/2}$$

IV. 연구목적

본 연구의 구체적인 연구 목적은 다음과 같다.

1. 상관계수를 사용한 자료융합과 Correspondence Analysis를 사용한 자료융합을 통하여 얻어진 누락치에 대한 추정치는 얼마나 본래 값에 가까운가 (즉, 추정치의 신뢰도)?

2. 삭제된 속성의 수가 추정치의 신뢰도 (정확도)에 미치는 영향은 어느 정도인가?

3. 추정치를 사용하여 발견한 세분 시장의 안정도 (Rand Index로 측정)는 어느 정도인가?

V. 연구 방법

본 연구에서 사용한 자료는 600명의 응답자들로부터 수집된 자동차 딜러쉽에 대한 선호도 조사 자료이다. 구체적으로 본 연구에서 사용된 자료는 13개의 속성으로 구성되어 있으며 각 속성의 수준은 2개에서 8개이다 (표 1 참조).

[표 1] 자동차 딜러 쉽의 속성과 속성수준의 수

속성	속성변수	속성수준의 수	속성수준 변수
판매차종	X_1	6	X_{1-1}, X_{1-2}, X_{1-3}, X_{1-4}, $X_{1\ 5}$, X_{1-6}
전시장(매장)위치	X_2	3	X_{2-1}, $X_{2\ 2}$, X_{2-3}
전시장(매장)내부	X_3	4	X_{3-1}, X_{3-2}, $X_{3\ 3}$, X_{3-4}
차량구입 도우미	X_4	3	X_{4-1}, X_{4-2}, X_{4-3}
할부판매	X_5	3	X_{5-1}, X_{5-2}, X_{5-3}
구매거래	X_6	3	X_{6-1}, X_{6-2}, X_{6-3}
보상판매	X_7	4	$X_{7\ 1}$, X_{7-2}, X_{7-3}, X_{7-4}
전문영업사원	X_8	3	X_{8-1}, X_{8-2}, X_{8-3}
신차서비스	X_9	4	X_{9-1}, X_{9-2}, X_{9-3}, X_{9-4}
렌탈카	X_{10}	2	X_{10-1}, X_{10-2}
서비스플랜	X_{11}	3	$X_{11\ 1}$, X_{11-2}, X_{11-3}
부품	X_{12}	3	$X_{12\ 1}$, X_{12-2}, X_{12-3}
가격할인	X_{13}	8	X_{13-1}, X_{13-2}, X_{13-3}, X_{13-4}, $X_{13\ 5}$, X_{13-6}, X_{13-7}, X_{13-8}

자료융합을 이용한 누락치에 대한 추정치의 정확도를 평가하기 위하여 본 연구에서는 13개의 속성과 그에 속한 총 49개의 속성수준 자료 (컨조인트 부분가치 자료)를 이용하였다. 처음의 6가지 속성(속성변수 X_1에서 X_6까지에 해당됨)은 600명의 응답자 자료를 자료융합을 위한 공통변수로 사용하고, 나머지 7가지 속성 (속성변수 X_7에서 X_{13}까지에 해당됨)은 각각 200명으로 구성된 세 개의 집단으로 나누고, 각 집단마다 누락한 속성의 수를 달리하였다. 즉,

- 집단 A: 7가지 속성 중 무작위로 하나의 속성을 선택, 그에 해당되는 속성수준모두를 누락시킨다.
- 집단 B: 7가지 속성 중 무작위로 두 개의 속성을 선택, 그에 해당되는 속성수준모두를 누락시킨다.
- 집단 C: 7가지 속성 중 무작위로 세 개의 속성을 선택, 그에 해당되는 속성수준모두를 누락시킨다.

세 개의 집단에서 누락된 값을 추정하기 위해서 상관계수와 Correspondence Analysis를 각각 사용하여 누락된 값을 추정하였다.

5.1 방법 I (상관계수를 이용한 자료융합)

공통 속성인 X_1에서 X_6까지의 속성변수를 계량변수(Metric Variable)로 취급하였다. 이들 계량형 변수를 기초하여 응답자끼리의 상관계수를 구하였다. 각 집단에 대하여 구해야 할 상관계수의 수는 $n(n-1)/2$ 개, 즉 19,900개이다.

응답자 i와 j의 상관계수 ($i \neq j$):

$$r_{ij} = \frac{\sum_{k=1}^{7}(\bar{Y}_{ik} - Y_{ik})(\bar{Y}_{jk} - Y_{jk})}{\sqrt{\sum_{k=1}^{7}(\bar{Y}_{ik} - Y_{ik})^2}\sqrt{\sum_{k=1}^{7}(\bar{Y}_{jk} - Y_{jk})^2}}$$

Where,

Y_{ik}: 응답자 i의 k 속성변수

$i = 1, \ldots, 200$: 응답자

$k = 1, \ldots, 7$: 공통속성변수

X_7에서 X_{13}까지의 속성 중 누락된 속성의 속성수준은 누락치가 발견된 응답자 I(수혜자)와 가장 상관계수가 높은 응답자 j(기증자)의 속성수준으로 대체하여 누락치를 추정하였다. 만약 가

장 상관계수가 높은 응답자 j도 같은 속성수준이 누락되어 있으면 그 다음 상관계수가 높은 응답자 k의 속성수준으로 대체한다.

5.2 방법 II (Correspondence Analysis를 이용한 자료융합)

공통 속성인 Y_1에서 Y_6까지의 속성변수에 대하여, 각 속성수준 중 가장 큰 값을 그 속성을 나타내는 이상점으로 인식하고 범주변수(Categorical Variable)로 취급한다. 범주형 이상점 변수에 기초하여 SAS의 Correspondence Analysis를 적용하여 각 응답자의 5차원 좌표(공통속성의 수보다 하나 적은 5차원이 사용됨)를 구하고 각 응답자간의 거리를 계산하였다.

응답자 i와 j의 거리 ($i \neq j$):

$$d_{ij} = \sqrt{\sum_{k=1}^{5} (d_{ik} - d_{jk})^2}$$

Where,

d_{ik} : 응답자 i의 k 차원의 좌표.
$i = 1, \ldots, 200$: 응답자
$k = 1, \ldots, 5$: 차원

Y_7에서 Y_{13}까지의 속성 중 누락된 속성의 속성수준은 누락치가 발견된 응답자 i(수혜자)와 가장 거리가 가까운 응답자 j(기증자)의 속성수준으로 대체하여 예측한다. 만약 가장 거리가 가까운 응답자 j도 같은 속성수준이 누락되어 있으면 그 다음 거리가 가까운 응답자 k의 속성수준으로 대체하고, 만약 응답자 k도 속성수준이 누락되어 있으면, 이 과정은 속성수준이 누락되지 않고 거리가 다음 순서로 가까운 응답자 i에게까지 확장한다.

5.3 자료혼합방법의 평가기준

상관계수와 Correspondence Analysis를 이용한 자료융합방법을 사용하여 기증자로부터 구한 추정치가 삭제하기 전의 수혜자의 값(Original Value)에 얼마나 가까운가를 평가하기 위한 기준으로 다음의 두 가지 척도를 사용하였다.

5.3.1 상관계수 (Correlation Coefficient)

수혜자가 지니고 있던 본래의 값 (Original Value)과 자료혼합을 통하여 구한 추정치(Recovered Value)간의 상관계수를 각 응답자의 수준에서 구하였다.

5.3.2 Rand Index

수혜자의 본래의 값으로 구성된 자료와 자료융합을 통해 예측되어진 추정치로 구성되어진 자료를 각각 군집분석을 한 후, 각각의 자료를 사용하여 구한 군집 소속간의 일치도(i.e., 세분시장의 안정도)를 자료혼합의 평가척도로 사용하였다. 군집간의 일치도를 구하는 척도로서 많이 사용되는 것은 Rand Index이다. Rand Index란 군집에 속해 있는 응답자 쌍들의 빈도수의 비율이다. Rand Index의 분자는 동일한 두 명의 응답자 쌍이 동일한 군집에 소속되어 있는가 혹은 상이한 군집에 소속되어 있는가의 빈도 수이며 분모는 전체 응답자 쌍의 수이다. 만약 두 개의 군집이 정확하게 일치한다면 Rand Index는 1.0이다. 만일 군집의 구성원들간에 일치가 전혀 이루어지지 않는다면 Rand Index는 0이다. 그러나 본래의 Rand Index에는 상향적 오차가 존재하므로 그 오차를 수정하여 많이 쓰이고 있는 것이 Adjusted Rand Index (ARI)이다. Rand Index가 높을수록 자료융합을 통해 본래의 값을 정확하게 추정한 것이다. [표 2]는 Rand Index와 ARI를 구하는 식을 보여 주고 있다.

[표 2] Rand Index와 Adjusted Rand Index (ARI)

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

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

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

VI. 연구결과

Y_1에서 Y_3까지의 속성에 해당하는 27개의 속성수준(즉, Y_{1-1}에서 Y_{13-8}까지의 속성수준)의 원래자료와 위의 두 가지 방법에 의하여 누락치를 예측하여 보완함으로써 누락치가 없어진 예측된 자료를 이용하여 두 자료의 일치도를 다음과 같이 측정하였다.

6.1 상관계수에 의한 평가 결과

각 응답자별로 두 자료간의 상관계수를 구한다. 즉, 각 집단별로 응답자 200명에 대한 200개의 상관계수를 구한 후 이에 대한 평균상관계수를 계산한다.

집단의 평균상관계수:

$$\overline{R} = \frac{\sum_{i=1}^{200} R_i}{200}$$

Where,

R_i: 응답자 i의 원래자료와 예측된 자료와의 상관계수

i = 1, 2, . . . , 200: 각 집단내의 응답자

연구결과의 안정성을 시험하기 위하여 Monte Carlo Simulation을 실시한다. 100회의 시뮬레이션을 실시한 결과 평균상관계수의 평균은 [그림 2]에서 보는 바와 같이 일정한 값에 수렴하고 있음을 알 수 있다.

평균상관계수의 평균:

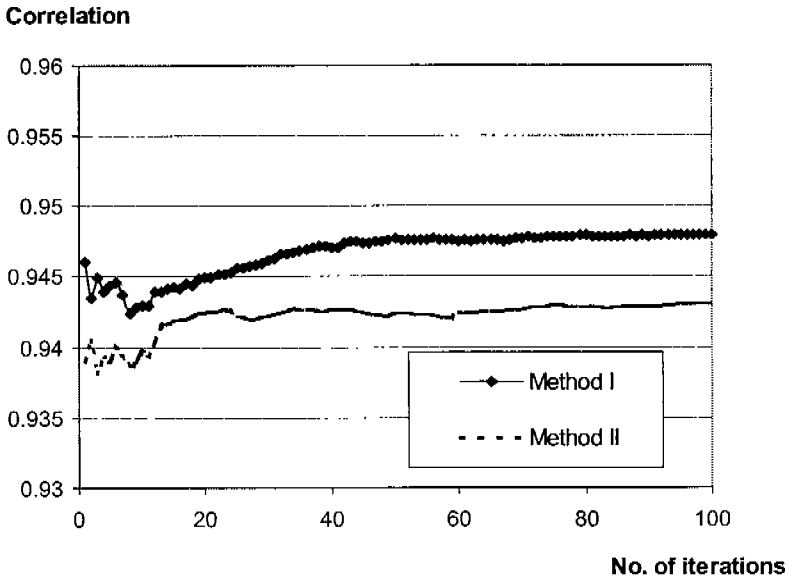
$$\overline{\overline{R}} = \frac{\sum_{i=1}^{100} \overline{R}^{(i)}}{100}$$

Where.

$\overline{R}^{(i)}$: i번째 시뮬레이션의 평균상관계수

i= 1, 2, . . . , 100: 시뮬레이션 횟수

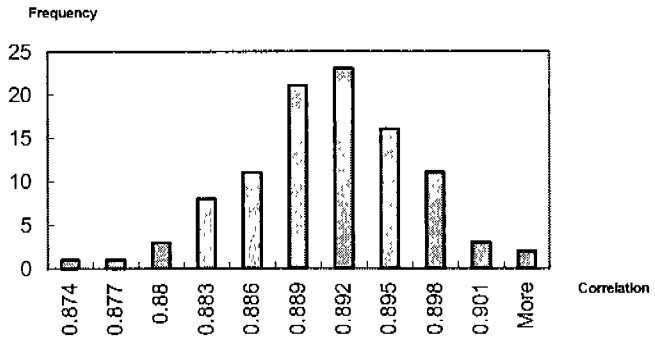
[그림 2] 평균상관계수의 수렴여부의 예



(집단 A에 대하여 방법 I과 방법 II를 적용한 결과)

또한 각 집단에 대한 평균상관계수의 분포도 [그림 3]에서 보는 봐와 같이 대략 정규분포의 모양을 따르고 있어 평균상관계수의 평균값을 대표 값으로 사용할 수 있었다.

[그림 3] 시뮬레이션결과에 따른 평균상관계수의 히스토그램의 예



(집단 B에 대하여 방법 I을 적용한 결과)

[표 3]는 Monte Carlo Simulation에 의한 평균상관계수의 분포특성을 요약하고 있다. 누락된 속성이 하나인 경우 평균상관계수의 평균은 두 방법 모두 0.94를 상회하여 매우 높았다. 누락된 속성이 두 개인 집단 B에 대한 예측은 상관계수가 0.89에 근접하여 역시 매우 정확한 예측이 이루어지고 있음을 보여주고 있다. 마지막으로 누락된 속성이 3개인 경우에도 (즉, 누락율 약 43%인 경우) 평균상관계수가 0.84 이상으로

써 두 방법에 의한 예측이 상당히 정확도가 높음을 보여준다. 예상했던 바와 같이 삭제된 속성의 수가 증가할수록 본래의 값과 추정 값간의 상관계수는 낮아진다는 것을 보여주고 있다. 또한 표준편차는 모두 매우 낮아 예측이 상당히 안정적임을 보여주고 있다. 누락된 속성의 수가 증가함에 따라 표준편차가 점차 증가함을 확인할 수 있다. 누락된 속성의 수가 1 개일 때는 상관계수에 의한 방법이 좋은 예측을 보여주며, 누락된 속성의 수가 증가하여 두 개 이상일 경우에는 Correspondence Analysis를 사용하여 거리를 구하고 예측한 방법 II가 더 좋은 예측을 보여주고 있다. 방법 II가 방법 I에 비하여 평균 상관계수의 분포에 대한 편차가 큼을 또한 주목할 수 있을 것이다.

[표 3] *Monte Carlo Simulation*에 의한 평균상관계수의 분포특성

	집단 A 삭제된 속성의 수 = 1		집단 B 삭제된 속성의 수 = 2		집단 C 삭제된 속성의 수 = 3	
	방법 I	방법 II	방법 I	방법 II	방법 I	방법 II
평 균	0.9479	0.9431	0.8899	0.8909	0.8422	0.8484
표준편차	0.0040	0.0045	0.0055	0.0058	0.0060	0.0072

6.2 Rand Index에 의한 평가 결과

본 연구에서는 자료융합 기법의 또 하나의 평가기준으로서 세분시장의 안정도를 채택하였다. 세분시장을 도출하기 위하여 다음과 같은 네 가지의 군집분석 방법을 사용하였으며 선행연구(김성호 1999)의 결과에 따라 군집의 수를 2개로 설정하였다.

- K-평균 군집분석 기법: Howard-Harris Program (1966), CCA (Sawtoo고 *Software* 1988)
- Data Mining Technique: Enterprise Miner (SAS), Clementine (SPSS)

[표 4]는 누락치를 추정하기 위한 방법 I (상관계수)과 방법 II (Correspondence Analysis)의 분산분석 결과를 보여주고 있다. 본 연구의 주된 초점은 방법 I과 방법 II의 비교에 있으므로 세분시장의 안정도를 도출하기 위하여 사용된 네 가지 군집분석 프로그램의 성과를 비교한 구체적인 결과는 생략하였다. (각 군집분석의 프로그램의 비교평가에 대한 연구는 현재 수행 중이다). 본 연구에서는 전체표본(N = 600)을 세 개의 하부자료(각각의 표본의 크기는 200)로 분할하여 각 20회의 시뮬레이션을 실행하였다. 따라서 총 480회의 시뮬레이션이 실행되었다 (삭제된 속성의 수[3] X 누락치 추정방법[2] X 군집분석 프로그램[4] X 시뮬레이션[20회]).

[표 4]에 나타난 바와 같이 분산분석의 전체모형은 통계적으로 유의하였다 ($p<0.0001$). 또한 누락치의 추정방법, 삭제된 속성의 수, 그리고 이들 실험변수 간의 상호작용 모두 통계적으로 유의한 것으로 나타났다 ($p = 0.001$). 또한 [표 5]는 실험변수의 평균 및 상호작용에 따른 실험요인의 조합의 평균을 보여 주고 있다. 누락치의 추정방법에 있어서는 상관계수에 의한 누락치의 추정 (방법 I)이 Correspondence Analysis에 의한 방법 (방법 II)보다 우수한 것으로 나타났으며 (0.3422 vs. 0.2534) 통계적으로 유의적이었다 ($p<0.05$). 삭제된 속서의 수에 있어서는 예상했던 바와 같이 삭제된 속성의 수가 증가할수록 누락치의 추정은 정확도가 감소하는 것으로 나타났다 ((0.4409 vs. 0.2554 vs. 0.1971). 이들 역시 0.05 수준에서 통계적으로 유의하였다.

[표 4] Rand Index의 분산분석 결과

Source of Variation	Sum of Squares	D.F	Mean Squares	F	p-value
모형	6.4887	5	1.2977	38.44	<0.0001
방법	0.9461	1	0.9461	28.02	<0.0001
삭제 속성의 수	5.1854	2	2.5927	76.79	<0.0001
방법*삭제속성 수	0.3572	2	0.1786	5.29	0.0053
오차	16.0042	474	0.0338		
합계	22.4929	479			

[표 5] 실험요인의 Rand Index 평균

	삭제 속성 1개	삭제 속성 2개	삭제 속성 3개	평균
방법 I (상관계수)	0.4513	0.3326	0.2425	0.3422
방법 II (Correspondence Analysis)	0.4304	0.1781	0.1517	0.2534
평균	0.4409	0.2554	0.1971	

VII 결론 및 향후 연구방향

본 연구에서는 자료융합 방법을 누락치의 추정에 적용하여 그 성과를 실제자료의 시뮬레이션을 통하여 탐색적으로 평가하였다. 두 가지의 누락치 추정방법과 두 가지의 평가기준을 채택한 연구결과는 전반적으로 고무적이라고 할 수 있다. 본 연구의 결과 중 흥미로운 사실은 상관계수를 평가기준으로 사용한 경우에는 방법 II(Correspondence Analysis에 의한 추정방법)의 성과가 다소 우수하게 나타난 반면 군집의 안정도를 평가기준으로 사용한 경우에는 방법 II (상관계수에 의한 추정방법)이 현저하게 우수한 것으로 나타났다. 이러한 상반된 결과 (특히 군집의 안정도가 낮게 나타난 결과)에 관해서는 향후 연구가 뒤따라야 하겠지만 우선은 다음과 같은 요인에 의하여 설명될 수 있을 것이다.

- Correspondence Analysis를 사용하여 군집분석을 수행한 경우, 원래의 27개의 변수를 5차원에 공간에 투사함으로써 정보의 손실이 있었다는 점을 들 수 있다. 따라서 본래 27개의 변수에 의한 군집의 구조가 5차원 공간으로 축소되는 과정에서 뒤틀어질 수 있다.
- 군집분석을 효과적으로 수행하기 위해서는 군집 당 응답자의 수를 150 이상으로 하여야 하는 반면(Wind 1978) 본 연구에서는 하부표본의 크기를 200으로 설정하였다는 점이 Rand Index에 영향을 미쳤을 것으로 보여진다. 다라서 향후 연구에서는 하부표본의 크기를 확대하여 시뮬레이션 연구를 수행하는 것이 필요하다.

본 연구는 탐색적인 연구의 성격에도 불구하고 향후 연구를 위한 다음과 같은 문제를 제기하고 있다.

- 본 연구에서는 13개의 속성 중 임의로 처음의 6개를 공통변수로 선정하였다. 공통변수

의 선택은 각각의 수혜자에 대한 기증자를 선택하는 근거가 된다는 점에서 매우 중요하다. 향후 연구는 공통자료의 선택에 관하여 진행되어야 할 것이다. 가능한 공통변수로서는 응답자의 이상점(Ideal Point), 속성의 중요도(Attribute Importance Weight), 혹은 응답자의 배경변수(인구통계적 변수, 사이코 그래픽 변수, 라이프스타일 등을 포함) 등을 사용할 수 있다. 또한 이들 공통변수의 척도에 따라 상관계수(product moment correlation) 혹은 동급상관계수(rank correlation)를 누락치의 추정에 적용할 수 있을 것이다. 향후 연구에서는 이러한 공통변수의 선택이 추정치의 정확도에 미치는 영향을 실펴 보아야 할 것이다.

- 향후연구에서는 또한 군집을 도출하는데 사용한 프로그램의 비교를 다루어야 할 것이다. 특히 K-평균 군집분석은 그 초기치의 선정에 매우 민감한 결과를 나타내므로 군집의 안정도를 평가기준으로 채택하는 경우 군집분석 프로그램의 선정에 주의하여야 한다.
- 본 연구에서는 컨조인트 부분가치자료를 이용하였으나 향후 연구에서는 다른 종류(응답자의 매체행동, 태도, 브랜드선호도 등)의 실제자료를 사용하는 시도가 필요할 것이다.

참고문헌

Baker, K., Harris, P. and O' Brien, J. (1989), "Data Fusion: AnAppraisal and Experimental Evaluation," *Journal of the Market Research Evaluation,* 31(2), 153-212

Carroll, J.D. and Arabie, P. (1980), "Multidimensional Scaling," in M. R. Rosenzweig and L. W. Porter (eds.), *Annual Review of Psychology,* Volume 31, Palo Alto, CA: Annual Review, 607-49.

Carroll, J.D., Green, P.E. and Schaffer, C.M. (1987), "Comparing Interpoint Distances in Correspondence Analysis," *Journal of Marketing Research,* 24 (November), 455-50.

Carroll, J.D., Green, P.E., and Schaffer, C.M. (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Marketing Research,* 23 (August), 271-80.

Hoffman, D.L. and Franks, G.R. (1986), "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research," *Journal of Marketing Research,* 23 (August), 213-27.

Kruskal, J. and Wish, M. (1978), *Multidimentional Scaling,* Newbury Park, CA.

김성호 (1999), "컨조인트 최적제품 포지셔닝모형을 이용한 시장세분화에 관한 연구," **마케팅硏究**, 103-118