응용통계연구 (2008)
21(5), 875–888

모자이크 플롯에서 변수와 범주의 순서화

이문주[1] · 허명회[2]

[1]고려대학교 통계학과; [2]고려대학교 통계학과

(2008년 5월 접수, 2008년 8월 채택)

요 약

Hartigan과 Kleiner (1981, 1984)에 의해 제안된 모자이크 플롯은 범주형 자료의 탐색에 매우 유용한 시각화 도구이다. 모자이크 플롯은 범주 셀의 빈도를 사각형의 크기에 비례하게 나타내므로 이해가 쉽고 데이터에 포함된 정보를 유지하지만 실제 모습은 변수 순서와 변수 내 범주의 순서에 따라 상당히 달라진다. 이에 우리는 본 연구에서 모자이크 플롯에서 크래머(Cramer)의 V 계수를 활용한 변수의 순서화 방법과 감마 계수를 활용한 범주의 순서화 방법을 제안하고 Titanic, Housing, PreSex 등 공개 자료에 적용한 결과를 제시한다.

주요용어: 모자이크 플롯, 변수의 순서화, 범주의 순서화, 크래머의 V, 감마 계수.

1. 연구배경과 목적

Hartigan과 Kleiner (1981, 1984)가 제안한 모자이크 플롯(mosaic plot)이 범주형 자료의 시각화 방법으로 매우 유용한 방법이라는 데 많은 자료 분석자들이 동의하고 있다. 모자이크 플롯은 1회 1 변수씩 개별 범주들의 빈도에 비례하게 주어진 사각형을 수평 또는 수직으로 나누어 표현하므로 이해가 쉽고 정보를 잃지 않는다. Friendly (1994)는 모자이크 플롯에 로그-선형 모형과의 편차를 추가해 넣는 방법을 개발하였고 Huh (2004)는 각 셀의 빈도를 면적으로 표현하는 대신 선으로 나타내는 모자이크 플롯의 한 변형을 제안한 바 있다. 그러나 모자이크 플롯에 향후 개선되어야 할 더 근본적인 문제들이 있다.

모자이크 플롯의 실제 모습은 변수들의 순서와 변수 내 범주들의 순서에 따라 다르다. 한 예로서, 그림 1.1에 제시된 2개의 모자이크 플롯을 보자. 두 플롯에서 열 변수($= A$)는 동일하며 행 변수($= B, C$)가 다른데 왼쪽 플롯에 비하여 오른쪽 플롯이 더 강한 연관성이 있다. 따라서 A와 B 간 모자이크 플롯보다는 A와 C 간 모자이크 플롯을 우선적으로 취할 필요가 있다. 또 한 예로서, 그림 1.2에서 3개 범주를 갖는 X와 2개 범주를 갖는 Y 변수 간 모자이크 플롯을 보자. 왼쪽 플롯은 X의 3개 범주의 순서를 있는 그대로 놓은 것이고 오른쪽 플롯은 X의 두 번째 범주와 세 번째 범주를 바꾼 것이다. 변수 X가 순서형이 아닌 경우라면 왼쪽 플롯보다는 오른쪽 플롯이 보기에 편하다.

수치형 자료를 위한 대표적 시각화 방법인 산점도 행렬이나 평행좌표 플롯도 변수들의 순서에 의존하나 최근 이들 그래픽 방법을 위한 변수 순서화 방법이 개발되었다 (Hurley, 2004). 범주형 자료의 그래픽 방법인 모자이크 플롯에서도 변수들을 순서화할 필요가 있고 뿐만 아니라 변수 내 범주들도 정렬할 필요가 있다(명목형 변수인 경우). 본 연구의 목적은 모자이크 플롯에서 변수와 범주의 순서화 방법을 제안하는 것이다.

[1](136-701) 서울특별시 성북구 안암동 5가, 고려대학교 통계학과, 석사과정 졸업. E-mail: elegize@korea.ac.kr
[2]교신저자: (136-701) 서울특별시 성북구 안암동 5가, 고려대학교 통계학과, 교수. E-mail: stat420@korea.ac.kr

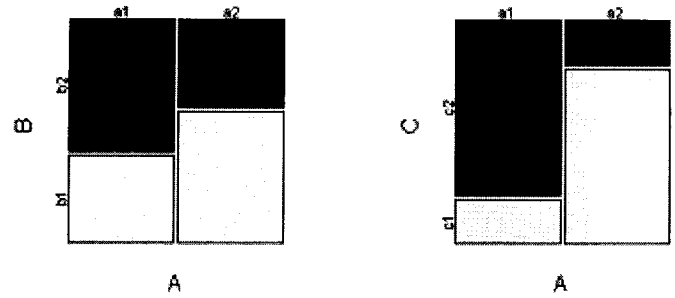

그림 **1.1.** 2 × 2 모자이크 플롯의 두 예

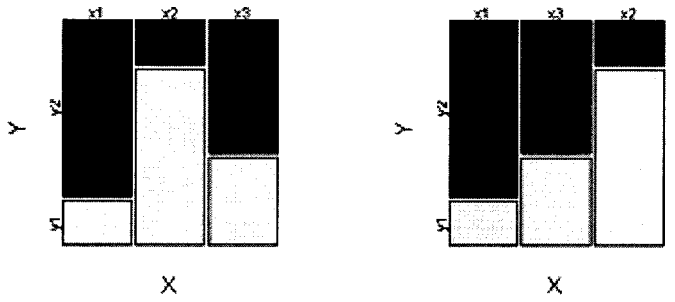

그림 **1.2.** 2 × 3 모자이크 플롯의 두 예

모자이크 플롯에서 범주의 순서화에 대한 기존 연구로는 Friendly (1994)가 있다. 그는 범주의 순서에 따라 모자이크 플롯의 모습이 달라지고 해석에 영향을 준다는 점에 주목하였고, 2개 변수의 모자이크 플롯을 위해서는 대응분석(correspondence analysis)의 1차원 결과를 활용할 것을 제안하였다. 그리고 3개 이상 변수에 대하여는 확장형 대응분석의 적용을 언급하였다 (Greenacre, 1984; van der Heijden과 de Leeuw, 1985). 그러나 대응분석이 차원축소적 방법이므로 제 1축에 국한된 범주 수량화는 정보 손실을 초래할 수밖에 없다. 더욱이, 자료가 변수를 3개 이상 포함하는 경우, 1회 1 변수씩 표현해가는 모자이크 플롯의 순차적 특성이 확장형 대응분석과 조화되지 않는다.

2절에서 특정변수를 목표로 하는 경우와 그렇지 않은 경우로 나누어 모자이크 플롯의 구성방법을 달리해야 할 필요성을 기술하고 변수의 순서화 알고리즘과 범주의 순서화 알고리즘을 제안한다. 그리고 3절에서는 Titanic 자료, Housing 자료, PreSex 자료 등에 2절의 변수 및 범주 순서화 알고리즘을 적용한 모자이크 플롯들을 제시한다. 마지막으로 4절에서는 관련 이슈를 토의하고 향후 연구 과제를 제시한다.

2. 변수와 범주의 순서화 알고리즘

대부분의 범주형 자료는 목표변수와 이와 관련된 다수의 설명변수로 구성되어 있다. 예컨대, 캘리포니아 대학교 버클리 입학자료(UCB Admissions) 사례에서는 합격여부(Admit)가 목표변수로 주어져 있고 관련 설명요인으로 지원자의 성(Gender)과 지원학과(Dept)가 고려된다 (Bickel 등, 1975). 성적

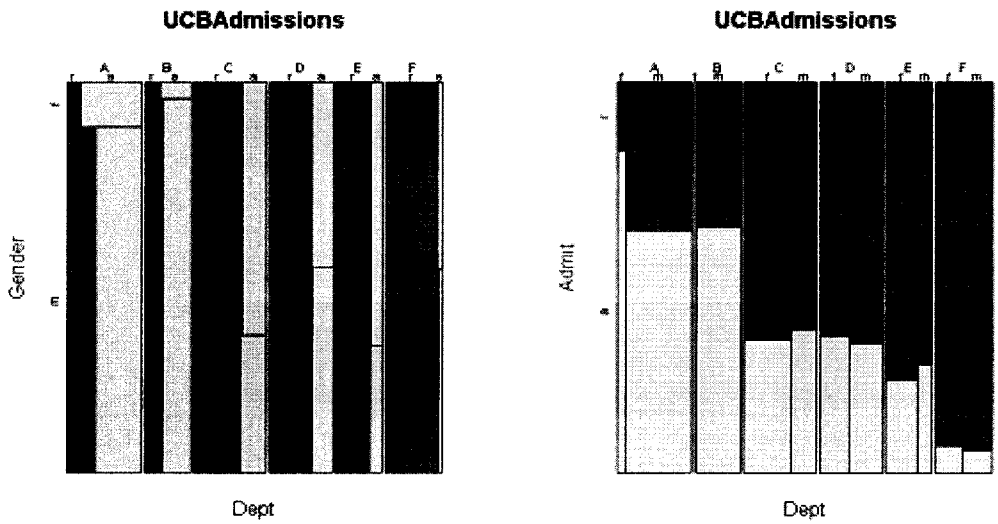

그림 **2.1.** UCBAdmissions 자료의 모자이크 플롯: Admit 목표

행태와 결혼에 관한 한 연구(PreSex) 사례에서는 현재의 결혼상태(Marital Status)가 성(Gender), 과거의 혼전 성관계(Pre-marital Sex), 혼외 성관계(Extra-marital Sex)로부터 받은 영향을 규명하고자 한다 (Thornes와 Collard, 1979). 이와 같은 상황에서 범주형 자료의 시각화를 다음과 같은 2개 국면(phase)으로 나누어 할 것을 제안한다.

국면 A. 목표변수와 설명변수 간 관련성을 시각화한다.

국면 B. 설명변수 간 상호 연관성을 시각화한다.

국면별로 자료 시각화의 포커스가 다르므로 각기 다른 방식으로 모자이크 플롯을 만들 필요가 있기 때문이다.

그림 2.1은 버클리 대학교 입학자료(UCBAdmissions)에 대한 2개의 모자이크 플롯을 보여주는데, 왼쪽 것은 Dept(A, B, C, D, E, F)를 수직 방향에, Gender(f, m)를 수평 방향에, Admit을 수직 방향에 놓은 것이다. 그러나 오른쪽 것은 Dept, Gender를 수직방향에 놓고 Admit(a, r)을 수평 방향에 놓은 것이다. 왼쪽 플롯에서는 Dept 간, 또는 Dept*Gender 조합 간 Admit 비율의 차이를 알아내기 어렵다. 그러나 오른쪽 플롯에서 Dept에 따라 Admit 비율에 현격한 차이가 있으나 Dept 내에서 Gender 간 차이는 미미함을 알 수 있다. 이와 같이 국면 A에서는 오른쪽 플롯과 같이 모든 설명변수를 한 방향에 두고, 목표변수를 다른 방향에 두는 것이 좋다.

한편, 그림 2.2의 두 플롯은 UCBAdmissions 자료에서 설명변수인 Dept와 Gender 간 연관성을 보여준다. 이 자료에서는 설명변수가 2개뿐이므로 변수 순서화는 필요 없지만 범주 순서화를 통하여 변수 간 연관성의 패턴을 쉽게 인지할 수 있다. 즉 그림 2.2의 오른쪽 플롯에서 Male 비율이 높은 학과가 *B-A-D-F-C-E* 순서임을 쉽게 볼 수 있다.

이와 같은 관찰을 토대로, 우리가 제안하는 모자이크 플롯에서의 변수 순서화 방법은 다음과 같다. 변수의 순서화 기준은 크래머의 V (Cramer, 1946)이다. 목표 변수가 있는 경우와 없는 경우에 따라 알고리즘이 다르다.

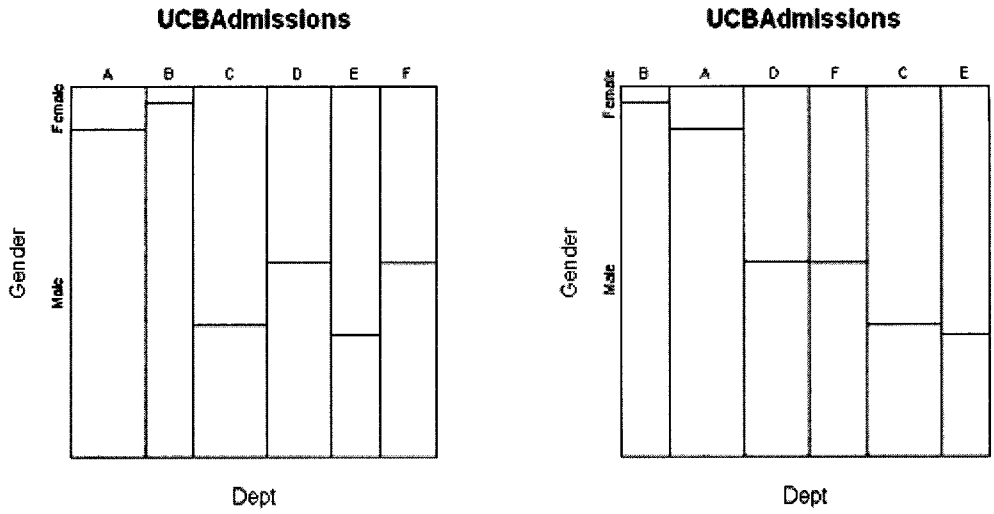

그림 **2.2.** UCBAdmissions 자료의 모자이크 플롯: Dept와 Gender 간 연관성

A. 목표변수(Y)에 대한 설명요인 $X_1, \ldots, X_{p-1}$의 영향을 보고자 하는 경우.

1) 첫 설명변수 선택: $j_1^* \leftarrow \arg\max_{j=1,\ldots,p-1} V(Y, X_j)$.
   X_1과 $X_{j_1^*}$를 바꾼다.
   단계 번호 s를 2로 놓는다.
   여기서 $V(Y, X_j)$는 Y와 X_j 간 크래머의 V이다. 즉, 각각 r개와 c개의 범주를 갖는 두 변수 R과 C간 카이제곱 통계량을 $\chi^2(R, C)$라고 하고 n을 총 빈도라고 할 때 두 변수 간 크래머의 V는

$$V(R, C) = \sqrt{\frac{\chi^2(R,\, C)}{n \cdot \min(r-1,\, c-1)}}$$

   로 정의되며 항상 0과 1 사이의 값을 취한다.

2) 다음 설명변수 선택: $j_s^* \leftarrow \arg\max_{j=s,\ldots,p-1} V^{X_1,\ldots,X_{s-1}}(Y, X_j)$.
   X_s와 $X_{j_s^*}$를 바꾼다.
   단계 번호 s를 1만큼 올린다.
   여기서
   $V^{X_1,\ldots,X_{s-1}}(Y, X_j) = \sum_{x_1} \cdots \sum_{x_{s-1}} p(x_1, \ldots, x_{s-1})\, V(Y,\, X_j | x_1, \ldots, x_{s-1})$,
   $p(x_1, \ldots, x_{s-1})$는 $X_1 = x_1, \ldots, X_{s-1} = x_{s-1}$에 대한 상대적 빈도,
   $V(Y,\, X_j | x_1, \ldots, x_{s-1})$는 $X_1 = x_1, \ldots, X_{s-1} = x_{s-1}$에 조건화된 부자료에서 Y와 X_j간 크래머의 V. 해당 교차표가 주변 합이 0인 행 또는 열을 갖는 경우 0으로 처리.

3) s가 p이면 변수 순서화 작업을 종료한다.
   그렇지 않으면 단계 2)로 돌아간다.
   작업을 종료하면 $X_1, \ldots, X_{p-1}$을 순서화한 결과로 $X_{j_1^*}, \ldots, X_{j_{p-1}^*}$을 얻는다.

B. 설명변수 $X_1, \ldots, X_{p-1}$ 간 상호 연관성을 보고자 하는 경우.

1) 처음 2개의 설명변수 선택: $(j_1^*, j_2^*) \leftarrow \arg\max_{j_1, j_2} V(X_{j_1}, X_{j_2})$.
X_1과 $X_{j_1^*}$를 바꾼다. X_2과 $X_{j_2^*}$를 바꾼다.
B.1: $k_1^* \leftarrow \arg\max_{k=3,\ldots,p-1} V^{X_1}(X_2, X_k)$,
B.2: $k_2^* \leftarrow \arg\max_{k=3,\ldots,p-1} V^{X_2}(X_1, X_k)$.
$V^{X_1}(X_2, X_{k_1^*}) \geq V^{X_2}(X_1, X_{k_2^*})$이면 $j_3^* \leftarrow k_1^*$로 놓는다. 그렇지 않으면 X_1과 X_2, j_1^*과 j_2^*를 바꾸고 $j_3^* \leftarrow k_2^*$로 놓는다. X_3와 $X_{j_3^*}$를 바꾼다. 단계 번호 s를 4로 놓는다.

2) 다음 설명변수 선택: $j_s^* \leftarrow \arg\max_{j=s,\ldots,p-1} V^{X_1,\ldots,X_{s-1}}(X_{s-1}, X_j)$.
X_s와 $X_{j_s^*}$를 바꾼다. 단계 번호 s를 1만큼 올린다.

3) s가 p이면 변수 순서화 작업을 종료한다.
그렇지 않으면 단계 2)로 돌아간다.
알고리즘을 마치면 $X_1, \ldots, X_{p-1}$을 순서화한 결과로 $X_{j_1^*}, \ldots, X_{j_{p-1}^*}$을 얻는다.

앞의 알고리즘에서 크래머의 V 대신 다른 연관성 측도로 대치해 볼 수도 있다. 그렇지만 기본적으로 모자이크 플롯은 카이제곱 검정 통계량 χ^2의 시각적 표현이기 때문에 χ^2에 기반한 연관성 측도를 변수 순서화에 기준으로 활용하는 것이 자연스럽다. 이 부류의 연관성 측도로는 크래머의 V 외에 ϕ(phi), 우발계수(contingency coefficient) C, Tschuprow's T 등이 있지만 이 중에서 최근에는 0과 1사이의 값을 취하는 크래머의 V가 가장 자주 활용되고 있다 (Garson, 2008).

각 변수의 범주들을 순서화하는 방법으로 다음 알고리즘을 제안한다. 두 변수 간 감마(gamma) 계수를 최대화하는 범주 순서를 찾아 모자이크 플롯에 넣는 방법이다. 목표변수의 유무에 관계없이 투입되는 변수들이 차례로 $X_1, \ldots, X_{p-1}$임을 가정한다.

1) X_1과 X_2의 범주 순서화: $(\pi_1^*, \pi_2^*) \leftarrow \arg\max_{\pi_1, \pi_2} \gamma(\pi_1(X_1), \pi_2(X_2))$.
$X_1^* \leftarrow \pi_1^*(X_1)$, $X_2^* \leftarrow \pi_2^*(X_2)$로 놓는다.
s를 3으로 놓는다.
여기서 $\pi_k(X_k)$는 X_k 범주들의 순서를 재지정하는 순열로서($k = 1, 2, \cdots$) X_k가 m_k개의 명목범주를 갖는 경우 $m_k!$개 경우를 고려하게 된다. 그러나 X_k가 순서 범주형 변수인 경우는 2개 경우만 고려한다. 그리고 $\gamma(X_1, X_2)$는 X_1과 X_2간 감마 계수이다 (Goodman과 Kruskal, 1979). 즉,

$$\gamma(X_1, X_2) = \frac{\Pi_c\left(F^{[12]}\right) - \Pi_d\left(F^{[12]}\right)}{\Pi_c\left(F^{[12]}\right) + \Pi_d\left(F^{[12]}\right)},$$

여기서 $F^{[12]} = \{f_{ij}^{[12]}\}$는 X_1과 X_2간 교차표의 표기이고

$$\Pi_c\left(F^{[12]}\right) = 2\sum_i \sum_j f_{ij}^{[12]} \left(\sum_{h>i} \sum_{k>j} f_{hk}^{[12]}\right),$$

$$\Pi_d\left(F^{[12]}\right) = 2\sum_i \sum_j f_{ij}^{[12]} \left(\sum_{h>i} \sum_{k<j} f_{hk}^{[12]}\right)$$

는 $F^{[12]}$에서 일치·비일치(concordant/discordant pairs) 쌍의 수이다. 다만, 해당 교차표가 주변합이 0인 행 또는 열을 갖는 경우 0으로 처리.

2) X_s의 범주 순서화: $\pi_s^* \leftarrow \arg\max_{\pi_s} \gamma^{X_1^*,\ldots,X_{s-1}^*}(X_{s-1}^*, \pi_s(X_s))$.
$X_s^* \leftarrow \pi_s^*(X_s)$로 놓는다.
s를 1만큼 늘린다.
여기서

$$\gamma^{X_1,\ldots,X_{s-1}}(X_{s-1}, X_j) = \sum_{x_1} \cdots \sum_{x_{s-1}} p(x_1,\ldots,x_{s-1})\gamma(X_{s-1}, X_j | x_1,\ldots,x_{s-1}).$$

즉, $\gamma^{X_1,\ldots,X_{s-1}}(X_{s-1}, X_j)$는 $X_1 = x_1,\ldots,X_{s-1} = x_{s-1}$에 조건화된 부자료에서 얻는 Y와 X_j간 감마 $\gamma(X_{s-1}, X_j | x_1,\ldots,x_{s-1})$를 조건화 칸의 구성비율 $p(x_1,\ldots,x_{s-1})$을 가중치로 평균낸 것이다.

3) s가 p이면 범주 순서화 작업을 종료한다.
그렇지 않으면 단계 2)로 돌아간다.
알고리즘을 마치면 $X_1,\ldots,X_{p-1}$는 범주들의 순서가 재지정된 $X_1^*,\ldots,X_{p-1}^*$로 바뀐다.

두 변수 중 어느 한 변수가 2개의 범주를 갖는 경우 감마 기준은 그 변수의 이항 비율 크기 순서로 다른 한 변수의 범주들을 정렬한다 (그림 1.2와 그림 2.2 참조).

3. 사례

3.1. Titanic 자료

Titanic은 타이타닉 호에 승선한 2,201명에 대한 기록 자료이다. 자료에 포함된 4개 변수는 Class(1^{st}, 2^{nd}, 3^{rd}, Crew: 등급), Sex(Male, Female: 성), Age(Child, Adult: 나이), Survived(No, Yes: 생존여부) 등이다. Class를 순서형 범주로 간주하고 앞 절의 제안 알고리즘에 따라 변수와 범주의 순서를 재배열하고 모자이크 플롯을 그려보도록 한다.

> **국면 A**: Survived와 관련된 첫째 요인으로 Sex가 선택되었고(Cramer's V = 0.46) Survived의 (No, Yes)와 Sex의 (Male, Female)이 대응하였다($\gamma = 0.82$). Survived와 관련된 둘째 요인은 Class이며 (Cramer's V = 0.22) Class의 4개 순서형 범주는 (Crew, 3^{rd}, 2^{nd}, 1^{st})로 정렬되었다 ($\gamma = 0.16$). Age(Adult, Child)가 마지막 변수였다(Cramer's V = 0.10, $\gamma = 0.35$). 그림 3.1 참조.

따라서 국면 A의 모자이크 플롯으로부터 다음과 같은 것들을 알아낼 수 있다. Survived에 관련 있는 첫째 요인은 Sex이다. Male의 생존률이 작았고 Female이 높았다. 다음 요인은 Class로 전체적으로는 (Crew, 3^{rd}, 2^{nd}, 1^{st})의 순서로 커졌지만 Male에서는 2^{nd} Class의 생존률이 가장 작았고 Female에서는 3^{rd} Class의 생존율이 가장 작았다. 마지막 요인은 Age로 모든 Sex*Class의 셀 그룹에서 Child의 생존률이 Adult의 생존률보다 높았다.

> **국면 B**: Class와 Sex가 먼저 선택되었다(Cramer's V=0.40). Class와 Sex의 순서는 각각 (1^{st}, 2^{nd}, 3^{rd}, Crew)와 (Female, Male)였다($\gamma = 0.66$). Class에 붙어 Age가 진입하였고(Cramer's V = 0.22) (Child, Adult)의 순서가 되었다($\gamma = 0.33$). 그림 3.2 참조.

따라서 국면 B의 모자이크 플롯으로부터 다음과 같은 것들을 알아낼 수 있다. Sex와 Class의 연관성이 크게 가장 나왔다. 즉, Female은 Male에 비해 상대적으로 1^{st} Class와 2^{nd} Class의 비율이 현저하

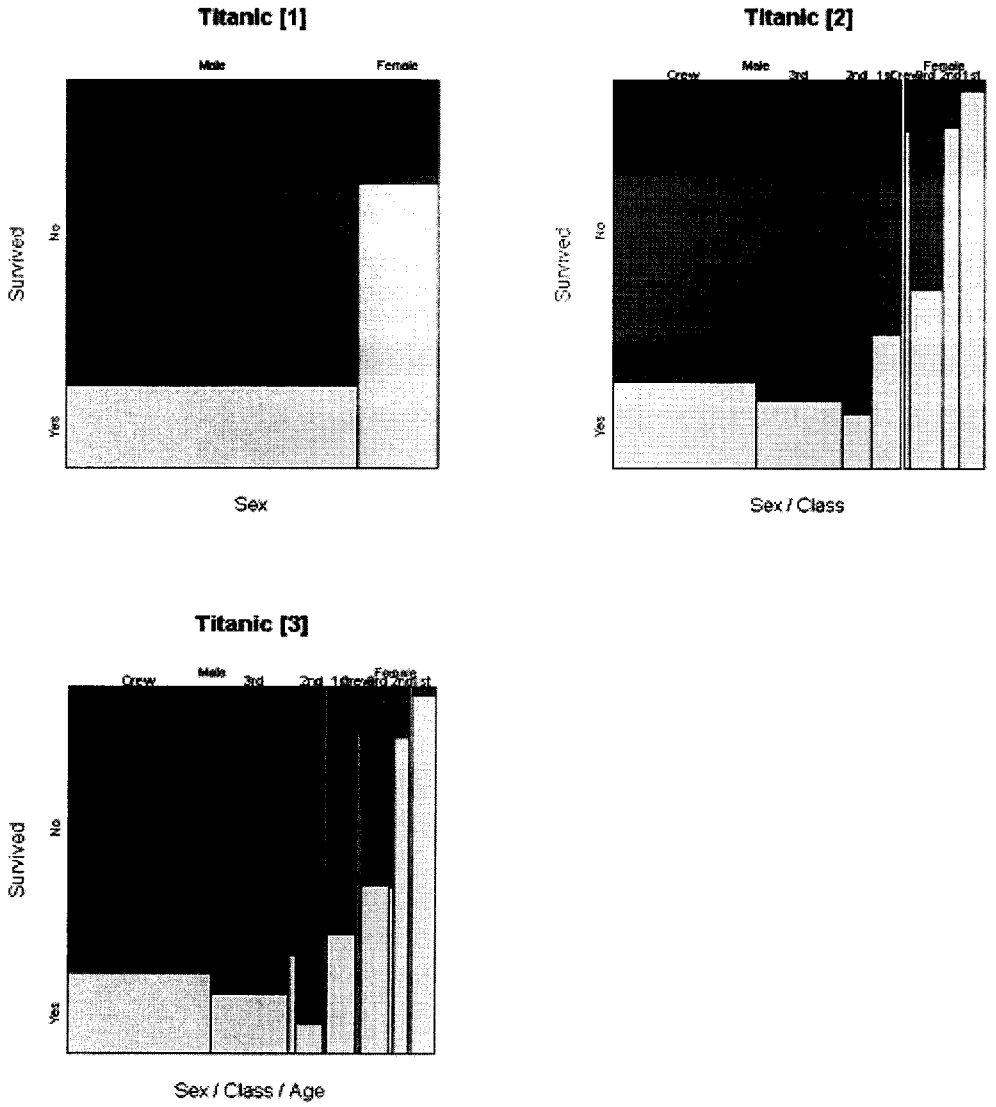

그림 **3.1.** TITANIC 자료에서 목표 SURVIVED와 3개 설명변수 간 모자이크 플롯

게 큰 반면 Crew 비율이 현저하게 작다. Sex*Class 셀 중에서 Child가 극히 적거나 없는 셀이 2개 있다(Female Crew와 Male Crew).

Titanic 자료를 잘 적합하는 로그선형모형(log-linear model)은 Class*Sex*Age, Class*Sex*Survived, Class*Age*Survived 등의 3-요인 상호작용을 포함하는 것으로 나타났다(G^2 = 1.69, 자유도 4, p-값 0.79, AIC = 57.7).

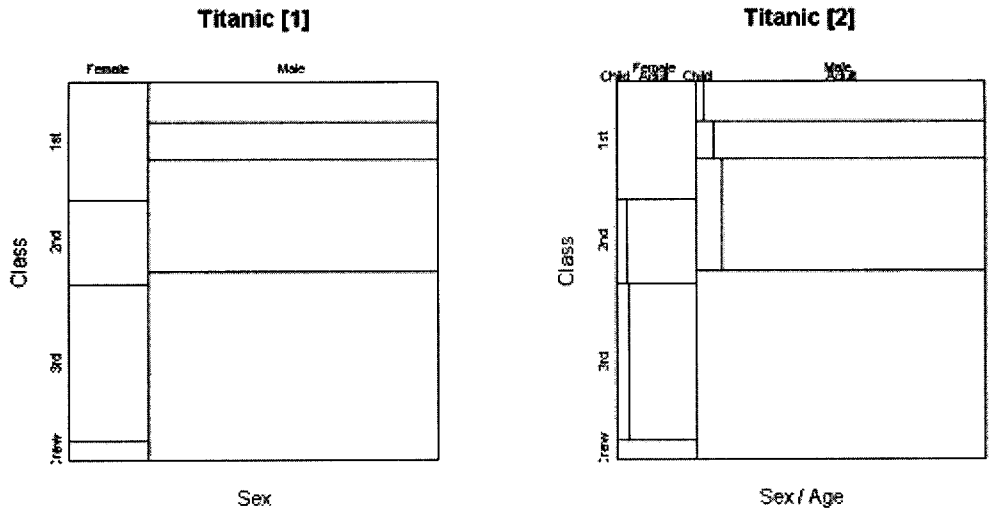

그림 **3.2.** TITANIC 자료에서 3개 설명변수 간 모자이크 플롯

3.2. Housing 자료

Housing 자료는 72가구 거주자로부터 Sat(Satisfaction, 만족도), Infl(Influence, 영향력), Type(주거형태), Cont(Contact, 접촉성)을 조사한 자료이다 (Madsen, 1976). 이 중에서 목표변수인 Sat는 Low, Medium, High 등 3개 범주로 구성되어 있다. 이를 위한 설명변수로 Infl, Type, Cont를 고려하는데 Infl은 Low, Medium, High로, Type는 To(Tower), Ap(Apartment), At(Atrium), Te(Terrace)로, Cont는 Low, High로 구성되어있다.

> **국면 A**: Sat와 가장 관련 있는 첫째 요인으로 Infl가 선택되었고(Cramer'sV = 0.18) Infl와 Sat의 3개 범주는 모두 Low, Medium, High로 순서화된다($\gamma = 0.33$). 다음으로 Type가 선택되며(Cramer's V = 0.14) Type의 4개 범주는 Te, Ap, At, To로 정렬된다($\gamma = 0.22$). 마지막으로 Cont가 Low, High의 순서로 들어온다(Cramer's V = 0.18, $\gamma = 0.24$). 그림 3.3 참조.

따라서 국면 A의 모자이크 플롯에서 다음 사항들을 알 수 있다. 만족도와 관련 있는 변수는 첫째 영향력(Infl)이며 다음이 주거형태(Type)이고 마지막으로 접촉성(Cont)이다. 영향력이 클수록, 주거형태는 Te, Ap, At, To의 순서로, 접촉성이 클수록 만족도가 크다.

> **국면 B**: Type과 Cont가 가장 큰 상호 연관성을 나타났다(Cramer's V = 0.15). Type는 Tower, Apartment, Atrium, Terrace의 순서일 때 Cont는 Low, High의 순서일 때 연관성이 가장 컸다($\gamma = 0.23$). Cont에 붙어 Infl가 진입하였고(Cramer's V = 0.10) High, Med, Low의 순서가 되었다($\gamma = 0.16$). 그림 3.4 참조.

따라서 국면 B의 모자이크 플롯에서 다음 사항들을 알 수 있다. 주거형태와 접촉성이 상호 연관도가 크며, 주거형태별로 접촉성과 영향력이 연관되어 있다. 즉 Tower, Apartment, Atrium, Terrace의 순서로 접촉성이 크며 접촉성이 클수록 영향력이 작은 경향이 있다.

Housing 자료를 잘 적합하는 로그선형모형을 찾은 결과 Sat*Inf, Sat*Type, Sat*Cont, Infl*Type, Infl*Cont, Type*Cont 등 모든 2-요인 상호작용이 선택되었고 또한 3-요인 상호작용 Type*Cont*Sat가

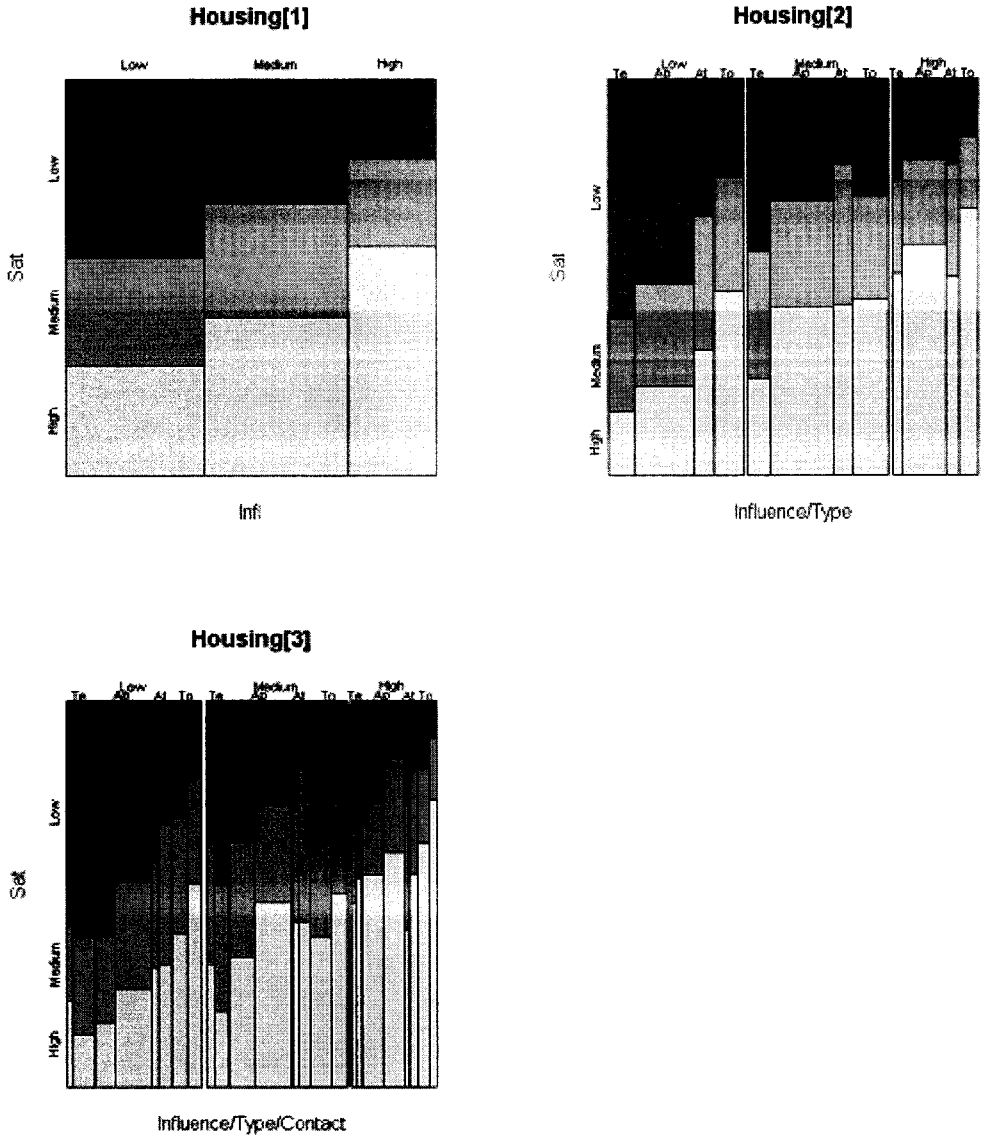

그림 **3.3.** Housing 자료에서 목표 Sat과 3개 설명변수 간 모자이크 플롯

선택되었다($G^2 = 31.8$, 자유도 34, p-값 0.58, AIC = 107.8).

3.3. PreSex 자료

PreSex 자료는 1,036명에서 조사한 4개 변수 Marital Status(결혼상태), Extra-marital Sex(혼외성관계), Pre-marital Sex(혼전성관계), Gender(성)로 구성되어 있다. Marital Status의 범주는 Di-

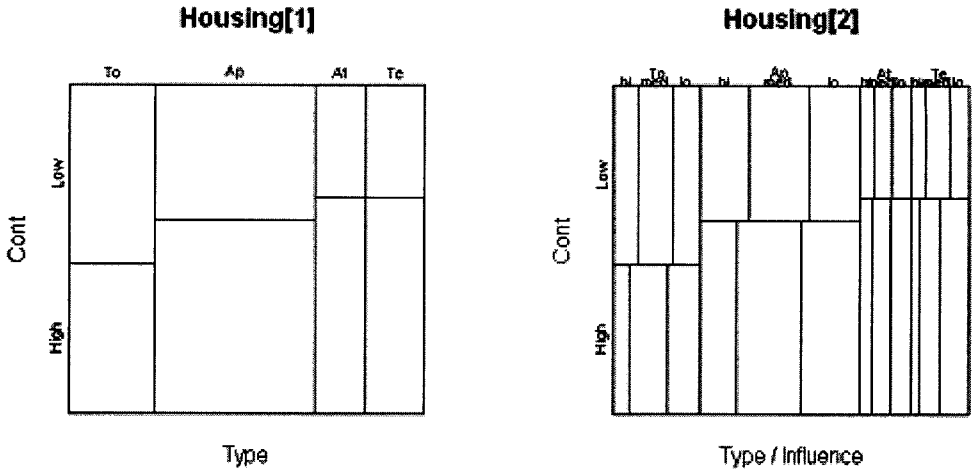

그림 **3.4.** HOUSING 자료에서 3개 설명변수 간 모자이크 플롯

vorced와 Married, Extra-marital Sex와 Pre-marital Sex의 범주는 Yes와 No, Gender의 범주는 Women과 Men이다.

> **국면 A**: 목적 변수 Marital Status와 가장 크게 관련된 요인은 Extra-marital Sex이다(Cramer's V = 0.24). 두 변수의 범주들은 (Divorced, Married)와 (Yes, No)가 대응할 때 감마 값이 가장 크다($\gamma = 0.70$). 이어서 선택되는 변수는 Pre-marital Sex이고(Cramer's V = 0.20) 범주 순서는 (Yes, No)이다 ($\gamma = 0.46$). 마지막 변수인 Gender는 (Women, Men)의 순서로 잡힌다(Cramer's V = 0.07, $\gamma = 0.12$). 그림 3.5 참조.

따라서 국면 A의 모자이크 플롯에서 다음 사항들을 알 수 있다. Extra-marital Sex 유경험 그룹에서 Divorced 비율이 높고 Extra-marital Sex 무경험 그룹에서는 Divorced 비율이 낮다 (Married 비율이 높다). 세부적으로는 Extra-marital Sex = Yes와 Pre-marital Sex = No 그룹에서 Divorced 비율이 가장 높고 Extra-marital Sex = No와 Pre-marital Sex = No 그룹에서 Divorced 비율이 가장 작다 (Married 비율이 가장 크다). Extra-marital Sex와 Pre-marital의 각 조합에서 Women의 Divorced 비율이 Men의 Divorced 비율보다 크다.

> **국면 B**: Gender와 Pre-marital Sex의 상호 연관성이 가장 크게 나타났다(Cramer's V = 0.27). Gender와 Pre-marital Sex의 순서가 각각 (Men, Women)과 (Yes, No)일 때 감마 값이 가장 컸다($\gamma = 0.58$). Pre-marital Sex에 붙어 Extra-marital Sex가 (Yes, No)의 순서로 진입하였다(Cramer's V = 0.24, $\gamma = 0.56$). 그림 3.6 참조.

따라서 국면 B의 모자이크 플롯에서 다음 사항들을 알 수 있다. Gender와 Pre-marital Sex의 연관성이 두드러진다(Men 그룹에서 Pre-marital Sex 유경험 비율이 크고 Women 그룹은 Pre-marital Sex 유경험 비율이 작다). Gender 그룹에서 Pre-marital Sex 경험과 Extra-marital Sex 경험이 병존한다.

PreSex 자료에 로그선형모형을 적합한 결과 $M*E*P$, $M*G$, $E*G$, $P*G$ 등 한 개의 3-요인 상호작용과 세 개의 2-요인 상호작용을 포함하는 모형이 최적으로 나타났다($G^2 = 0.76$, 자유도 4, p-값

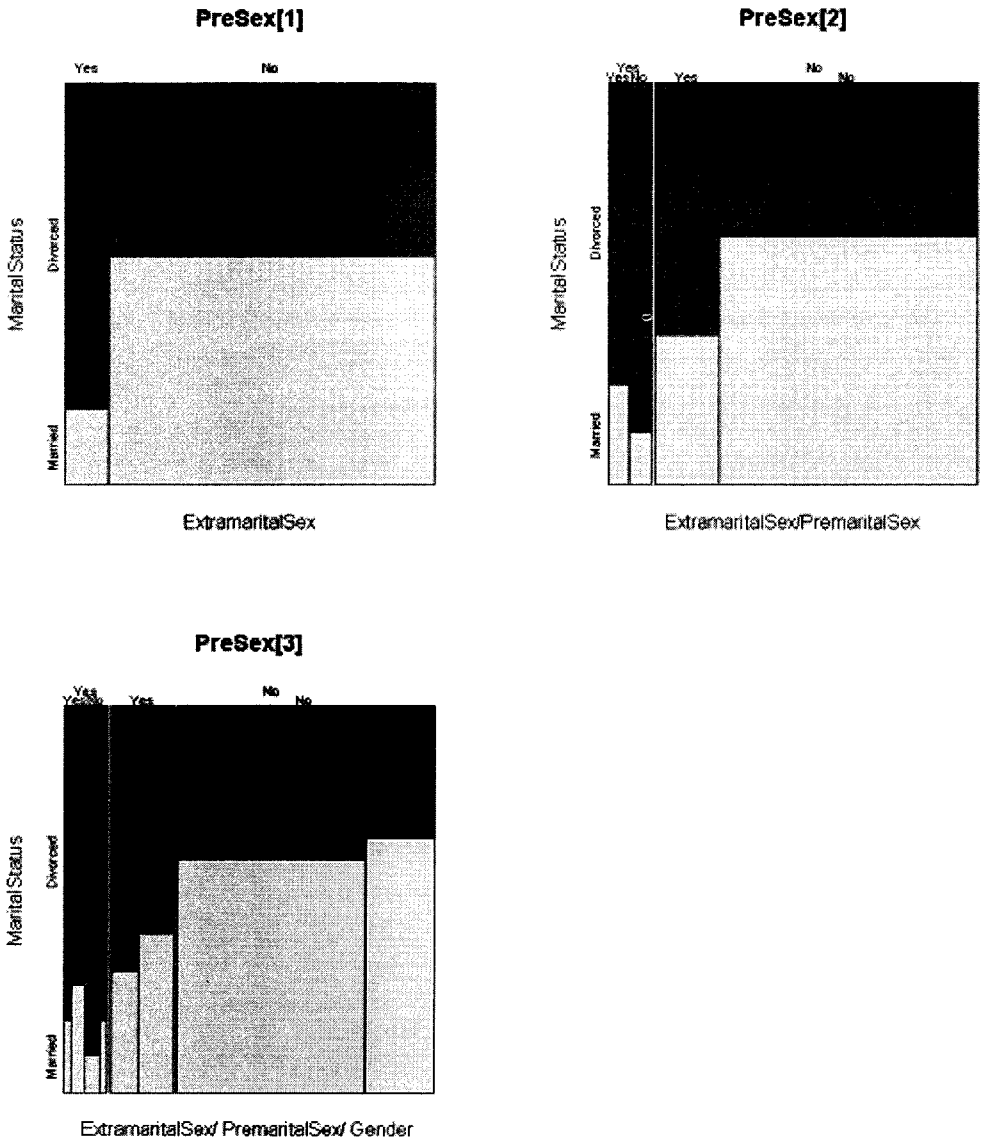

그림 **3.5.** PreSex 자료에서 MaritalStatus과 3개 설명요인 간 모자이크 플롯

0.94, AIC = 24.8; 여기서 M은 Marital Status, E는 Extra-marital Sex, P는 Pre-marital Sex, G는 Gender임).

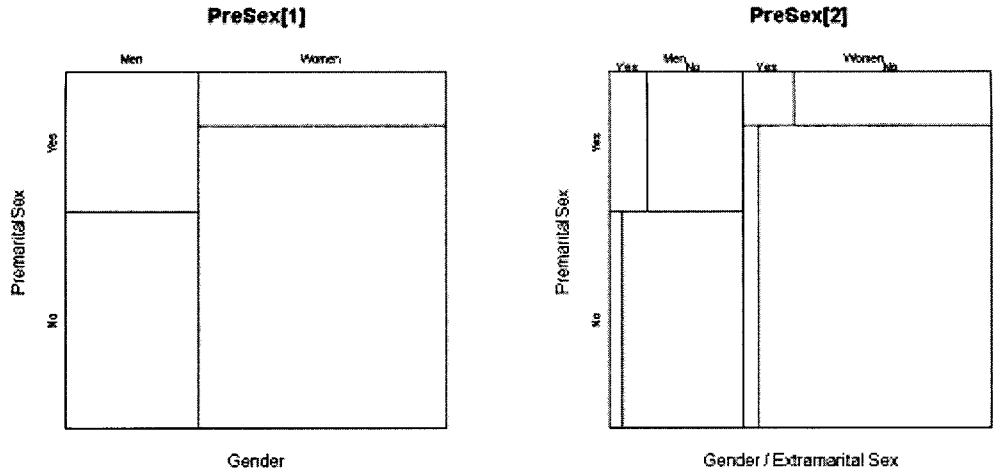

그림 **3.6.** PRESEX 자료의 3개 설명요인 간 모자이크 플롯

4. 맺음말

2절에서 변수와 범주의 순서화 알고리즘을 제안하면서 변수 순서화에서는 Cramer의 V에, 범주 순서화에서는 감마 계수 γ에 의존한 바 있다. 그러나 Cramer의 V는 명목 연관성(nominal association) 측도 중 하나이고 감마 계수 γ도 순서 연관성(ordinal association) 측도 중 하나이다. 이 연구에서는 V와 γ에 기반한 변수 순서화 방법과 범주 순서화 방법을 연구하였으나 향후 좀 더 포괄적인 연구가 필요하다.

이 연구에서는 변수 순서를 정하면서 마지막 변수까지 플롯에 넣었다. 그러나 PreSex 사례에서 국면 A의 경우 마지막 변수로 Gender가 들어오면서 Cramer의 V가 0.07에 그치는 등 특정 설명변수는 모자이크 플롯에서 추가적인 기여를 못하는 수가 있다. 따라서 일정 수준에 못 미치는 변수는 진입되지 않도록 하는 것이 좋을 것이다. 향후 이에 대한 연구가 필요하다.

참고문헌

Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley, *Science,* **187**, 398–403.

Cramer, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association,* **89**, 190–200.

Garson, G. D. (2008). Nominal association: Phi, contingency coefficient, Tschuprow's T, Cramer's V, lambda, uncertainty coefficient, *Statnotes: Topics in Multivariate Analysis,* Retrieved from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm 06/25/2008.

Goodman, L. A. and Kruskal, W. H. (1979). *Measures of Association for Cross Classifications,* Springer-Verlag, New York.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis,* Academic Press, London.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables, In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface,* (ed. by W.F. Eddy). New York: Springer-Verlag, 268–273.

Hartigan, J. A. and Kleiner, B. (1984). A mosaic of television ratings, *The American Statistician,* **38**, 32–35.

Huh, M. Y. (2004). Line mosaic plot: Algorithm and implementation, *COMPSTAT, 2004 Symposium,* Physica-Verlag/Springer.

Hurley, C. B. (2004). Clustering visualizations of multidimensional data, *Journal of Computational & Graphical Statistics,* **13**, 788–806.

Madsen, M. (1976). Statistical analysis of multiple contingency tables: Two examples, *Scandinavian Journal of Statistics,* **3**, 97–106.

Thornes, B. and Collard, J. (1979). *Who Divorces?*, Routledge and Kegan, London.

van der Heijden, P. G. M. and de Leeuw, J. (1985). Correspondence analysis used complementary to log-linear analysis, *Psychometrika,* **50**, 429–447.

Ordering Variables and Categories on the Mosaic Plot

Moon-Joo Lee[1] · Myung-Hoe Huh[2]

[1]Dept. of Statistics, Korea University; [2]Dept. of Statistics, Korea University

(Received May 2008; accepted August 2008)

Abstract

Mosaic plots, proposed by Hartigan and Kleiner (1981, 1984), are very useful in visualizing categorical data. In mosaic plot, multi-way classified cell frequencies are represented by rectangles with proportional area. The plot is easy to understand while preserving the information contained in the data. Plot's appearance, however, does change substantially depending on the order of variables and the orders of categories with variable put into the plot. In this study, we propose the algorithms for ordering variables and categories of the categorical data to be explored via mosaic plots. We demonstrate our methods to three well-known datasets: Titanic, Housing and PreSex.

Keywords: Mosaic plot, ordering variables, ordering categories, Cramer's V, Gamma coefficient.

[1]Graduate Student, Dept. of Statistics, Korea University, Seoul 136-701, Korea. E-mail: elegize@korea.ac.kr
[2]Corresponding author: Professor, Dept. of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr