응용통계연구 제 15권 2호, 2002년, pp. 433-448

탐색적 자료분석시 그래프의 활용에 대한 연구

장대흥[1)]

요 약

탐색적 자료분석에서는 자료를 통계적 모형에 바로 적합시키기 보다 자료를 있는 그대로 보려는 데 중점을 두므로 현시성을 강조한다. 따라서, 다양한 그래프가 사용되는데. 본 논문에서는 이러한 그래프들을 이용하여 탐색적 자료분석의 몇 가지 유용한 사례들을 보이고자 한다.

주요용어: 줄기-잎-그림, 상자그림, 혼돈, 다반응값 최적화.

1. 서론

탐색적 자료분석에서 다루는 중요한 네 가지 주제로서 저항성의 강조, 잔차의 계산, 자료의 재표현을 통한 다각적 시도, 그래프를 통한 현시성 등을 들 수 있다. 탐색적 자료분석에서는 자료를 통계적 모형에 바로 적합시키기 보다 자료를 있는 그대로 보려는 데 중점을 두므로 현시성을 특히 강조한다. 따라서, 다양한 그래프가 사용되는데. 본 논문에서는 이러한 그래프들을 이용하여 탐색적 자료분석을 행한 몇 가지 유용한 사례들을 보이고자 한다. 첫째로, 줄기-잎-그림과 상자그림을 결합시키는 방법을 제시하고, 둘째로, 임의성(randomness)과 혼돈(chaos)을 구별할 수 있는 탐색적 방법을 보이고, 셋째로, 다반응값 최적화에 어떻게 탐색적 방법이 쓰일 수 있는 지를 언급하였다.

2. 그래프의 활용

2.1. 줄기-잎-그림과 상자 그림의 결합

상자 그림은 일변량 자료의 중심, 산포도, 비대칭성, 그리고 이상값이라는 4가지의 특징을 한 그림에 나타내는 그림이다. 상자 그림의 많은 장점에도 불구하고 일변량 자료의 밀도를 구체적으로 나타내지 못하는 단점이 있다. 상자 그림의 장점을 살리면서 상자 그림의 단점을 보완하기 위하여 여러 학자들이 상자 그림의 변형들을 제시하였다(McGill, Tukey와 Larsen(1978), Benjamini(1988), Frigge, Hoaglin과 Iglewicz(1989), Hintze와 Nelson(1998)). 이 중 Hintze와 Nelson(1998)이 제시한 바이올린 그림(violin plot)는 상자그림과 density trace를 하나로 결합한 그림이다. 그림의 결과가 보통 바이올린처럼 생겨서 붙인 이름이다. 그러나, 이 그림을 그리기 위하여서는 density trace를 구하여야 하는 데, 이 density trace의 개념이 초, 중, 고등학생은 물론이고, 대학생들(교양과목으로서의 기초통계학의 입장에서)

1) (608-737) 부산시 남구 대연3동 599-1, 부경대학교 자연과학대학 수리과학부 통계학전공, 교수
E-mail: dhjang@pknu.ac.kr

에게도 쉬운 개념이 아닐뿐더러 컴퓨터를 이용하지 않으면 이 바이올린 그림을 그리기가 쉽지 않다. 줄기-잎-그림과 상자 그림을 결합시키는 하나의 시도로서 종이와 펜만 가지고도 사용할 수 있는 방법을 우리는 모색하여 볼 수 있다.

줄기-잎-그림을 그릴 때 줄기와 잎을 구분하기 위하여 줄기와 잎 사이에 수직선을 그어준다. 줄기-잎-그림과 상자 그림을 결합시키는 아주 간단한 방법으로서 이 수직선 대신 상자 그림을 작게 그리는 방법을 생각할 수 있다. 상자 대신 Tufte(1983)이 제안한 점과 선분으로 대신하는 방법도 있다. 그러나, 그림인식상 수직선 옆에 상자 그림을 그리는 것이 수직선 대신 상자 그림을 그리는 것 보다 더 나아 보여 본 논문에서는 이 방법을 택하였다. 줄기-잎-그림과 상자 그림은 모두 개념 및 작도가 쉽고, 종이와 펜만 가지고도 작성할 수 있다. 따라서, 본 논문이 제안한 방법도 종이와 펜만 가지고도 작성할 수 있다. 줄기-잎-그림과 상자 그림을 결합시키는 방법의 순서는 다음과 같이 아주 간단하다.

1. 줄기-잎-그림을 그린다.
2. 상자 그림을 작성한 후 줄기와 잎 사이의 수직선 옆(수직선과 잎 사이)에 그린다.
3. 줄기-잎-그림에서 중앙값, 사분위수, 인접값, 이상값에 해당하는 자료들을 색깔, 밑줄, 또는 굵기 등을 이용하여 구별시킨다.

단계 3에서 중앙값, 사분위수, 이상값에 해당하는 자료들을 색깔로 구분하는 방법을 예로 들면 중앙값은 녹색, 사분위수는 파란색, 인접값은 노랑색, 이상값은 빨간색 등으로 선택하여 표시할 수 있다. 우리는 줄기와 잎 사이의 수직선 옆에 상자 그림을 그려 줄기-잎-그림과 상자 그림을 결합시킨 이 그림을 통하여 줄기-잎-그림과 상자 그림의 시너지효과를 맛볼 수 있다.

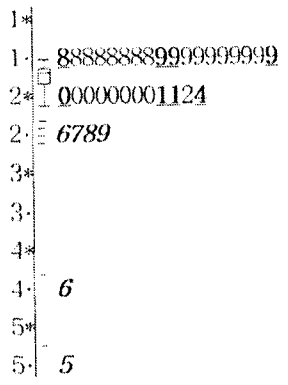


그림 2.1: 예 2.1에 대한 결합그림

예 2.1. Koopmans(1987) 책에 나와 있는 통계학수업 청강학생들의 나이에 대한 자료를 이용하여 다음 그림 2.1과 같이 줄기-잎-그림과 상자 그림을 결합시켜 하나의 그림으로

나타내어 보았다. 줄기-잎-그림에서 중앙값, 사분위수, 인접값에 해당하는 자료들은 숫자를 진하게 하고, 밑줄을 그었다, 이상값은 진하게 하여 구별하였다. 오른쪽으로 치우쳐있는 구조를 확인할 수 있다. 줄기-잎-그림과 상자 그림을 결합시켜 시너지효과를 맛볼 수 있다.

예 2.2. Milton(1992) 책에 나와 있는 흡연자와 비흡연자의 잠들기까지 걸린 시간에 대한 자료를 이용하여 다음 그림 2.2와 같이 줄기-잎-그림과 상자 그림을 결합시켜 하나의 그림으로 나타내어 보았다. 줄기-잎-그림에서 중앙값, 사분위수, 인접값에 해당하는 자료들은 숫자를 진하게 하고, 밑줄을 그었다, 이상값은 진하게 하여 구별하였다. 왼쪽이 비흡연자그룹이고 오른쪽이 흡연자그룹이다. 흡연자그룹의 특이한 패턴이 눈에 띈다. 줄기-잎-그림과 상자 그림을 결합시켜 시너지효과를 맛볼 수 있다.

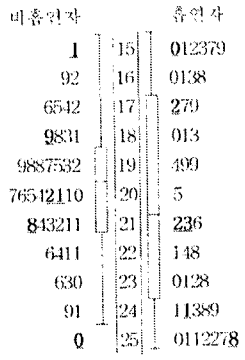


그림 2.2: 예 2.2에 대한 결합그림

예 2.3. 도부 양(渡部 洋)외 3인(1985) 책에 나와 있는 일본 두 지역의 년중 지진횟수에 대한 자료를 이용하여 다음 그림 2.3과 같이 줄기-잎-그림과 상자 그림을 결합시켜 하나의 그림으로 나타내어 보았다. 줄기-잎-그림에서 중앙값, 사분위수, 인접값에 해당하는 자료들은 숫자를 진하게 하고, 밑줄을 그었다, 이상값은 진하게 하여 구별하였다. 두 종류의 오른쪽으로 치우쳐있으나 형태가 아주 판이한 구조들을 확인할 수 있다. B 지역에서 하나의 아주 특이한 이상값이 눈에 띈다. 줄기-잎-그림과 상자 그림을 결합시켜 시너지효과를 맛볼 수 있다.

2.2. 임의성과 혼돈의 구별

혼돈에 대한 수리적인 연구는 프랑스 수학자였던 포앙카레(Henri Poincare)로 까지 거슬러 올라가는 데, 포앙카레는 1890년대에 태양계의 안정성에 대하여 연구를 하였다. 그러나, 비선형동역학(nonlinear dynamics)에 대한 연구는 Lorentz(1963) 이후에서야 비로소 활발히 연구되기 시작하였다. 언뜻 겉으로 보기에는 혼돈은 통계학에서 항상 언급하는 임의성과 매

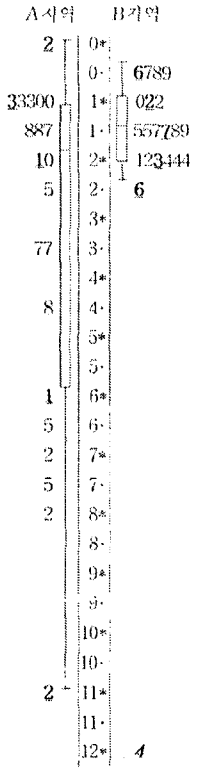


그림 2.3: 예 2.3에 대한 결합그림

우 흡사하다. 그러나, 혼돈은 결정론적(deterministic)이지만 임의성은 확률적(stochastic)이다. 이러한 혼돈을 임의성과 구별하는 것은 통계학자로서 중요한 작업이라 할 수 있다. 그러나, 일반적으로 혼돈과 임의성을 구별하기가 어렵다(Berliner(1992), Chatterjee와 Yilmaz(1992)). Lai와 Harrist(1997), 그리고 Lai와 Chen(2002)은 잡음이 들어간 혼돈을 구별하기 위한 통계방법들을 제안하였다. 이러한 혼돈을 임의성과 구별하는 탐색적 그래픽 방법이 있다면 혼돈을 임의성과 구별하는 데 유용한 수단이 될 것이다. 기존의 통계그래픽스 기법인 산포도행렬과 3차원 회전을 이용하면 혼돈과 임의성을 구별할 수 있기 때문에 이러한 통계그래픽스 방법들이 혼돈과 임의성의 구별을 가능하게 하는 탐색적 그래픽 방법이라는 것을 보일 수 있다.

비선형동역학 모형으로서 가장 간단한 모형이 다음과 같은 로지스틱방정식(logistic equation)이다.

$$x_{k+1} = f_\alpha x_k = \alpha x_k(1 - x_k), \quad k = 0, 1, 2, \ldots \tag{2.1}$$

이러한 로지스틱방정식에서 $\alpha = 4$일 때 혼돈현상이 일어난다. 즉, 초기값을 0과 1 사이

의 임의의 값을 잡고, (2.1)식을 반복하여 돌리면 어느 정도의 횟수가 지난 후(초기값의 영향이 사라진 후)에는 나오는 값들이 겉으로 보기에는 [0,1]에서 무작위로 나오는 것처럼 보여 혼돈이 일어난다. 그러나, 이러한 혼돈은 통계학에서 항상 언급하는 임의성과는 본질상 다르다. 그러므로, 혼돈과 임의성을 구별하는 것은 중요하다.

$x_0, x_1, ..., x_n$을 관측된 시계열자료라 하자. 그러면, 다음과 같은 차 시차(lag)를 갖는 행렬을 다음과 같이 만들 수 있다.

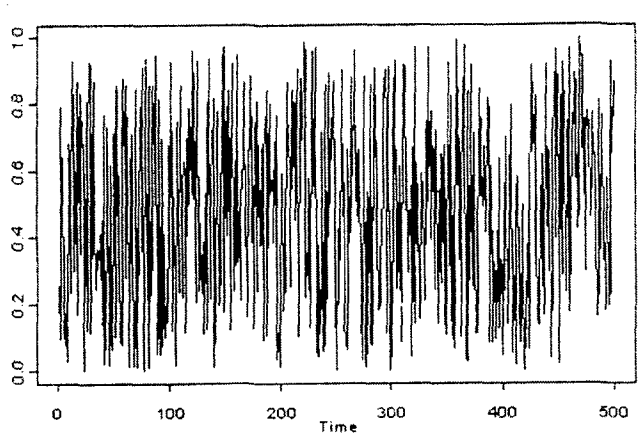
$$D = \begin{pmatrix} x_0 & x_1 & x_2 & ... & x_{n-l} \\ x_1 & x_2 & x_3 & ... & x_{n-l+1} \\ x_2 & x_3 & x_4 & ... & x_{n-l+2} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_l & x_{l+1} & x_{l+2} & ... & x_n \end{pmatrix} \tag{2.2}$$

이 행렬을 시차행렬(lag matrix)라 하자. $d_i' = (x_i, x_{i+1}, x_{i+2}, ..., x_{n-l+i}),\ i = 0, 1, 2, \ldots, l$ 이라 하면 D는 $D = (d_0', d_1', ..., d_l')'$로 표기할 수 있다. Peitgen, Jurgens와 Saupe(1992)는 $l = 1$인 1차 시차그림(lag plot)를 이용하여 의사난수발생기(pseudo-random number generator)에서 추출한 시계열자료를 로지스틱방정식과 Henon끌개(Henon attractor)에서 발생한 시계열자료들과 비교하였다. 우리는 ℓ차로 더 일반화시키기 위하여 산포도행렬(scatterplot matrix)을 이용할 수 있다. d_i'과 d_j' $(j > i,\ i = 0, 1, 2, \ldots, l-1)$를 각각의 산포도에 그리면 이 산포도행렬은 행렬 D의 행벡터 사이의 관계를 한 그림으로 나타낼 수 있다. 이렇게 그린 산포도행렬을 시차그림행렬(lag plot matrix)이라 하자. 이러한 시차그림행렬은 혼돈과 임의성의 구별을 가능하게 하는 탐색적 그래픽 방법으로 쓰일 수 있다.

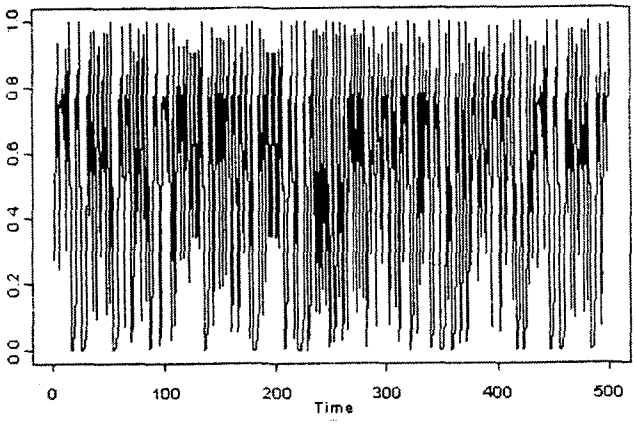
혼돈과 임의성의 구별을 가능하게 하는 또 다른 탐색적 그래픽 방법으로서 3차원 회전(rotation)이 있다. d_i', d_j', d_k' $(i < j < k,\ i = 0, 1, 2, ..., l-2)$를 3차원 산포도에 나타낸 후 회전을 시키면 우리 눈의 착시 현상에 의하여 입체구조를 확인할 수 있다. 이러한 3차원 산포도를 3차원 시차그림(three-dimensional lag plot)라 하자. 이러한 3차원 시차그림도 혼돈과 임의성의 구별을 가능하게 하는 탐색적 그래픽 방법으로 쓰일 수 있다.

예 2.4. 다음 그림 2.4 (a)와 (b)는 각각 S-Plus의 의사난수발생기에서 뽑은 500개의 난수들과 로지스틱방정식에서 추출한 500개의 시계열자료를 나타낸 그림들이다. 언뜻 보기에 로지스틱방정식에서 추출한 500개의 시계열자료들이 임의성을 띄고 있는 것처럼 보인다. 이 것을 확인하기 위하여 각각의 500개 수들을 이용하여 $l = 5$인 시차그림행렬을 그리면 다음 그림 2.5와 그림 2.6과 같다. S-Plus의 의사난수발생기에서 뽑은 500개의 난수들은 그림 2.5의 시차그림행렬에서 보는 바처럼 임의성을 띄고 있다. 어떤 시차그림에서도 임의성을 잃지 않고 있다. 그러나, 로지스틱방정식에서 추출한 500개의 시계열자료는 그림 2.6의 시차그림행렬에서 보는 바처럼 시차를 두고 특정한 함수 형태를 갖는다. 즉, 확률적이지 못하고 결정론적이다. 예로, d_i'와 d_{i+1}'사이에는 2차 곡선의 형태($f_4(x)$)를 유지하고 있다. 대응되는 그림 2.5의 시차그림과 비교하면 그 차이를 곧 알 수 있다. 그림

2.6의 시차그림행렬에서 d'_i와 d'_{i+1}, d'_i와 d'_{i+2}, d'_i와 d'_{i+3}, d'_i와 d'_{i+4}에 대한 시차그림은 각각 $f_4(x)$, $f_4(f_4(x))$, $f_4(f_4(f_4(x)))$, $f_4(f_4(f_4(f_4(x))))$라는 함수 형태를 나타내고 있다. 부차적으로 히스토그램을 작성하여 보면 그림 2.6의 하단에서 보는 것처럼 로지스틱방정식에서 추출한 500개의 시계열자료는 균일(uniform)하지 못 하고 [0, 0.1]과 [0.9, 1] 사이에 숫자들이 많이 몰려있고, 욕조(bath-tube) 모양의 U-자형의 모습을 하고 있다.



(a)



(b)

그림 2.4: (a) S-Plus의 의사난수발생기에서 뽑은 500개의 난수들 (b) 로지스틱방정식에서 추출한 500개의 시계열자료들

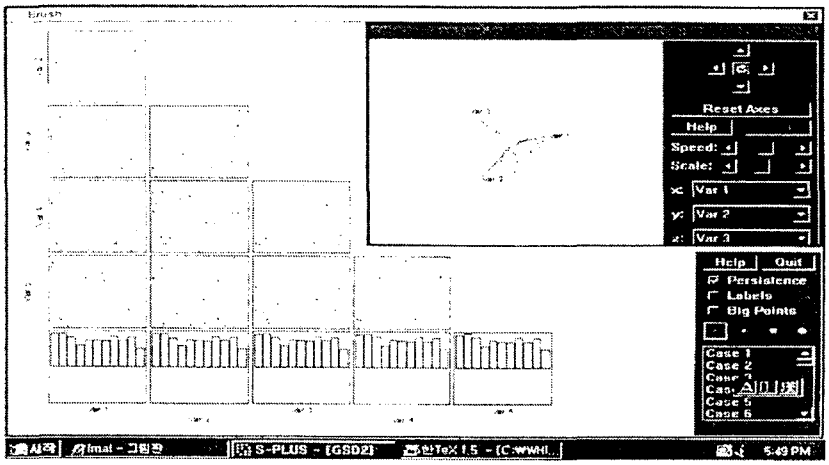


그림 2.5: S-Plus의 의사난수발생기에서 뽑은 500개의 난수들을 이용하여 그린 시차그림행렬과 3차원 시차그림(그림에서 $var1$, $var2$, $var3$, $var4$, $var5$는 각각 d_0, d_1, d_2, d_3, d_4를 나타냄.)

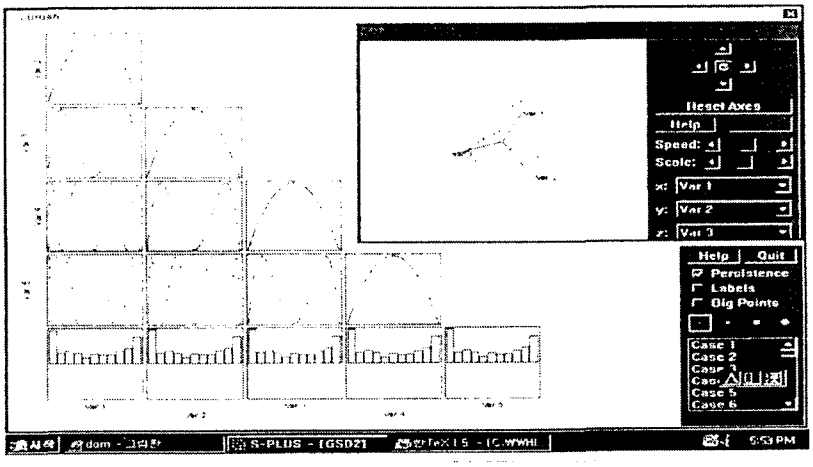


그림 2.6: 로지스틱방정식에서 추출한 500개의 시계열자료들을 이용하여 그린 시차그림행렬과 3차원 시차그림(그림에서 $var1$, $var2$, $var3$, $var4$, $var5$는 각각 d_0, d_1, d_2, d_3, d_4를 나타냄.)

그림 2.5의 오른쪽 상단의 창에서 보는 것처럼 S-Plus의 의사난수발생기에서 뽑은 500개의 난수들을 이용하여 만든 시차행렬 D에서 d_0', d_1', d_2'를 선택한 후 이를 사용한 3차원 시차그림을 보면 대체적으로 입방체 모양을 띄고 입방체 내부를 무작위로 채우고 있다. 반면, 그림 2.6의 의 오른쪽 상단의 창에서 보는 것처럼 로지스틱방정식에서 추출한 500개의 시계열자료들을 이용하여 만든 시차행렬 D에서 d_0', d_1', d_2'를 선택한 후 이를 사용한 3차원 시차그림을 보면 매우 특이한 입체구조를 갖는다. 즉, 특수한 함수 형태를 갖는다. 즉, 확률적이지 못하고 결정론적임을 알 수 있다.

혼돈의 예로서 우리는 로지스틱방정식 외에 각종 끌개, 즉, Henon 끌개, Lorez 끌개, Rossler 끌개 등의 시계열자료들을 이용할 수도 있다.

2.3. 다반응값 최적화

단반응값 최적화 문제에서는 최대경사법, 정준분석, 능선분석 등을 이용하여 최적화를 시행하나 다반응값 최적화 문제는 단반응값 최적화 문제보다 복잡하기 때문에 이러한 기법들을 사용할 수 없다. 많은 연구자들이 이러한 다반응값 최적화 및 다반응값 문제에 대하여 연구를 하였고, 최근까지도 연구가 활발히 진행되고 있다(Vining과 Myers (1990), Del Castillo와 Montgomery (1993), Derringer (1994), Luner (1994), Lin과 Tu (1995), Copeland와 Nelson (1996), Del Castillo (1996), Del Castillo, Montgomery와 McCarville (1996), Vining (1998), Box (1999), Myers (1999), Kim과 Lin (1998, 2000), 그리고, Carlyle, Montgomery와 Runger (2000)).

다반응값 최적화 문제에서 평균반응값들의 분포를 아는 것이 중요하다. 하나의 방법으로 모든 평균반응값들의 등고선도(contour plot)를 그릴 수 있으나, 이 방법은 설명변수의 수나 반응변수의 수가 많아지면 그리기가 불가능하다. 산포도행렬과 평행좌표그림을 이용하면 평균반응값들의 (조건부)분포를 알 수 있고 경험적으로 최적조건을 찾을 수 있다. 다반응값 최적화를 위한 해석적인 기법들(직접탐색기법, 수학적 최적화기법, 소망함수(desirability function) 등)을 사용하기 전이나 후에 산포도행렬과 평행좌표그림을 이용하여 평균반응값들의 (조건부)분포를 알아 보고 경험적으로 최적조건을 찾아 볼 수 있다.

y_1, y_2,...,y_r을 r개의 반응변수들이라 하고 x_1, x_2,...,x_k를 k개의 설명변수(입력변수)들이라 하고, 반응변수들이 흥미영역 R에서 입력변수들의 다항식으로 나타내어 진다고 하면 n개의 실험점을 가지고 최소제곱법을 이용하여 반응변수들을 다항식(주로, 1차식 또는 2차식)으로 나타낼 수 있다. $x = (x_1, x_2, \ldots, x_k)$에서의 i번째 추정반응값을 $\hat{y}_i(x)$라 하고 오차분산으로 나눈 추정반응값분산을 $V(x)$라 하자. $V(x)$를 고려하는 이유는 최적조건에서의 $\hat{y}_i(x)$에 대한 정밀도(precision)를 고려하기 위해서다. 그러면, 다음과 같은 순서로 산포도행렬과 평행좌표그림을 그려 평균반응값들의 (조건부)분포를 알아 보고 경험적으로 최적조건을 찾아 볼 수 있다.

순서 1. 추정반응함수식 $\hat{y}_1(x)$, $\hat{y}_2(x)$,...,$\hat{y}_r(x)$들을 구한다.
순서 2. 몬테칼로시뮬레이션을 이용하여 $(x_1, x_2, ..., x_k, \hat{y}_1(x), \hat{y}_2(x), ..., \hat{y}_r(x), V(x))$의 조합을 원하는 수 만큼 구한다.

순서 3. 순서 2에서 구한 점들을 이용하여 산포도행렬이나 평행좌표그림을 그린다.
순서 4. 브러싱기법을 이용하여 경험적으로 최적조건을 찾는다.

예 2.5. Derringer와 Suich(1980)는 타이어의 지상접촉면에 대한 연구에서 소망함수를 이용하여 다반응값 최적조건을 찾았다. 이 실험에서는 3개의 입력변수(x_1:hydrated silica level, x_2:silane coupling agent level, x_3:sulfur)를 갖는 중심합성계획을 이용하였고, 4개의 반응변수(y_1:PICO abrasion index, y_2:200% modulus, y_3:elongation at break, y_4:hardness)가 있고, 반응변수들에 대한 조건은 다음과 같다.

$$y_1 > 120,\ y_2 > 1000,\ 400 < y_3 < 600,\ 60 < y_4 < 75$$

20개의 실험점들을 이용하여 2차 다항식 $\hat{y}_1(x)$, $\hat{y}_2(x)$, $\hat{y}_3(x)$, $\hat{y}_4(x)$ 를 구한 후 순서 1과2를 통하여 그림 2.7처럼 250개의 $(x_1, x_2, x_3, \hat{y}_1(x), \hat{y}_2(x), \hat{y}_3(x), \hat{y}_4(x), V(x))$ 조합을 찾을 수 있다. 그림 2.7의 산포도행렬을 통하여 평균반응값들의 조건부분포를 알 수 있다. 브러싱기법을 이용하여 반응변수들에 대한 조건을 다음과 같이 더 강하게 제약하면,

$$y_1 > 130,\ y_2 > 1300,\ 400 < y_3 < 600,\ 60 < y_4 < 75$$

그림 2.8와 같은 산포도행렬을 구할 수 있다. 그림 2.8로부터y_1과 y_4사이에 양의 상관이 있고, y_1과 y_3사이에 음의 상관, y_1과 y_2사이에 약한 양의 상관이 있음을 알 수 있다. 브러싱기법을 이용한 최적조건을 찾기 위하여 반응변수들에 대한 조건을 다음과 같이 좀 더 강하게 제약하면,

$$y_1 > 135,\ y_2 > 1400,\ 400 < y_3 < 600,\ 60 < y_4 < 75$$

그림 2.9과 같이 3개의 점이 남는다. 이 점들은 표 2.1과 같다.

표 2.1: 최적조건

번호	x_1	x_2	x_3	$\hat{y}_1(x)$	$\hat{y}_2(x)$	$\hat{y}_3(x)$	$\hat{y}_4(x)$	$V(x)$
1	0.769	1.282	−1.216	137.19	1471.37	408.01	70.59	0.992
2	0.494	0.888	−1.084	136.60	1456.10	418.84	69.79	0.424
3	0.589	0.740	−0.960	139.47	1456.66	401.41	69.39	0.326

Derringer와 Suich(1980)는 소망함수를 이용하여 다반응값 최적조건 $(x_1, x_2, x_3, \hat{y}_1(x), \hat{y}_2(x), \hat{y}_3(x), \hat{y}_4(x)) = (-0.050,\ 0.145,\ -0.868,\ 129.5,\ 1300,\ 465.7,\ 68.00)$을 찾았다. 그들은 $V(x)$를 고려하지 않았다. 표2.1에서 번호 2와 3번의 조합은 Derringer와 Suich(1980)가 찾은 최적조건보다도 $\hat{y}_1(x)$와 $\hat{y}_2(x)$의 값이 더 크다. 물론, 소망함수를 어떻게 정의하느냐에 따라 우리가 찾은 다반응값 최적조건과 비슷한 결과를 얻을 수 있다.

위와 같은 방법으로 평행좌표그림을 그리면 그림 2.10, 2.11, 그리고 2.12와 같다. 이 평행좌표그림을 이용하면 평균반응값들의 분포를 알아 볼 수 있고, 산포도행렬처럼 경험적으로 최적조건을 찾아 볼 수 있다.

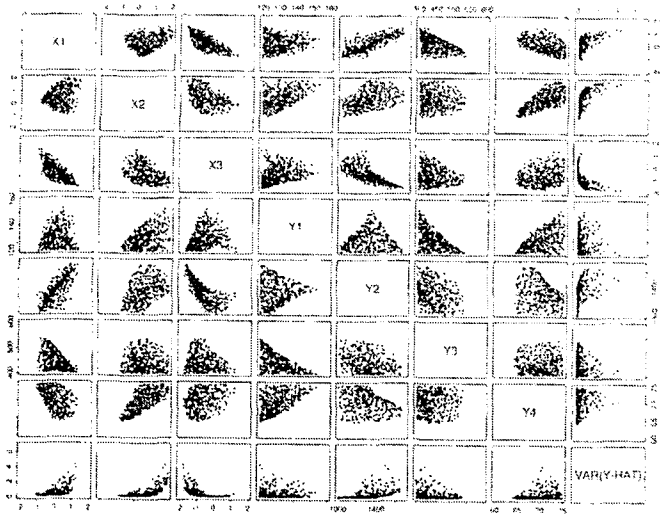


그림 2.7: 예제 2.5를 위한 산포도행렬 (1)

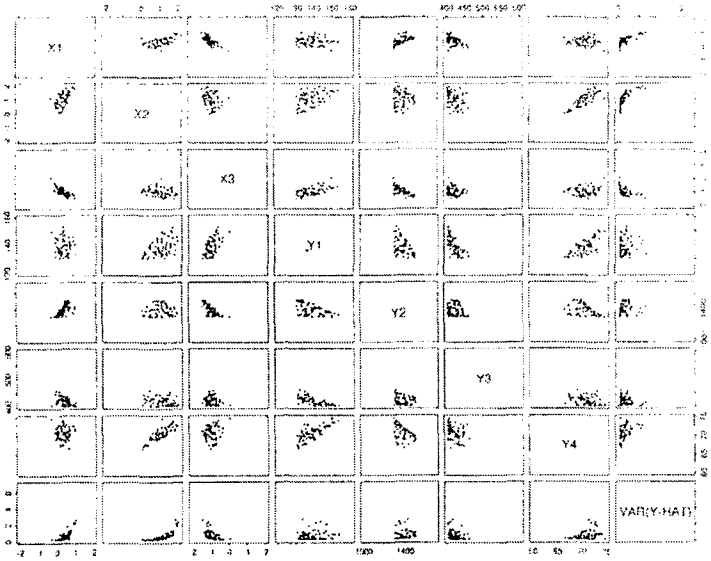


그림 2.8: 예제 2.5를 위한 산포도행렬 (2)

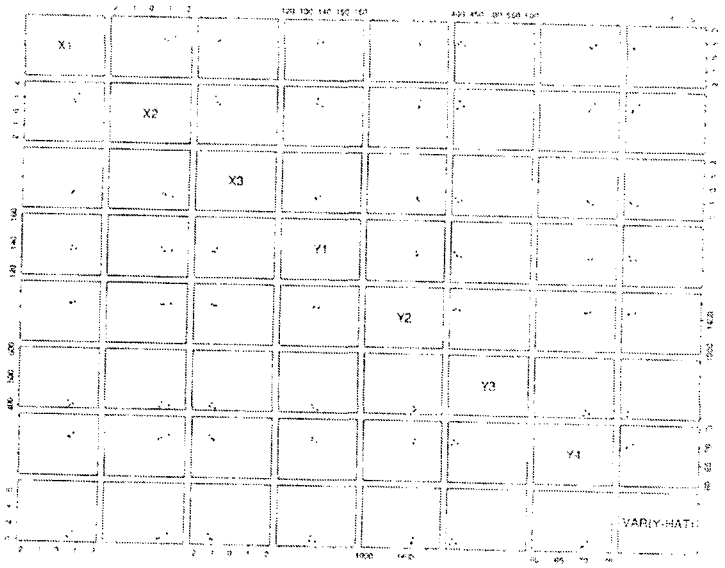


그림 2.9: 예제 2.5를 위한 산포도행렬 (3)

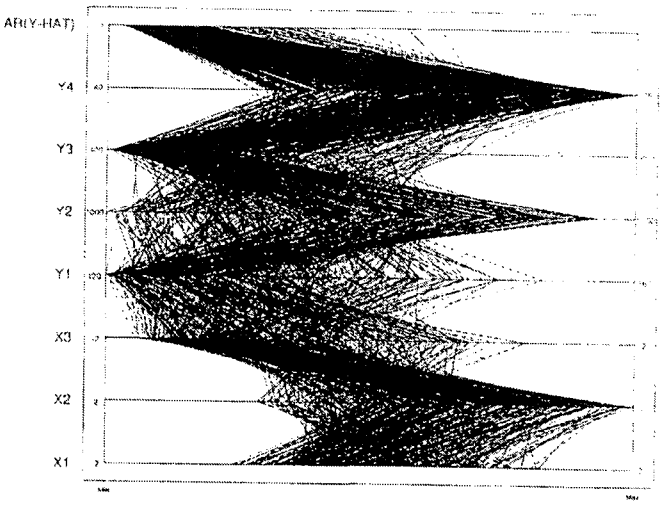


그림 2.10: 예제 2.5를 위한 평행좌표그림 (1)

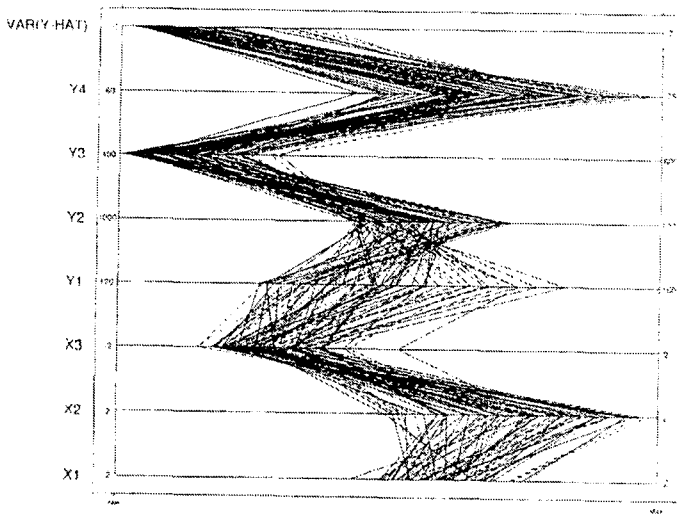


그림 2.11: 예제 2.5를 위한 평행좌표그림 (2)

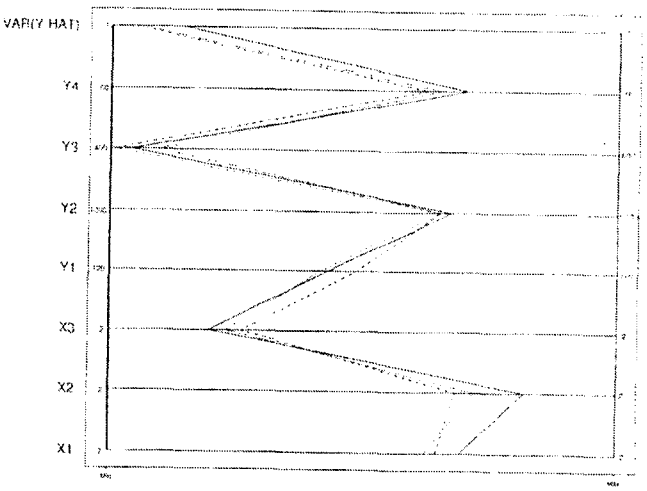


그림 2.12: 예제 2.5를 위한 평행좌표그림 (3)

3. 결론

본 논문에서 우리는 세 가지 예들을 통하여 탐색적 방법, 특히 그래프들을 이용하는 방법이 자료분석에 유용함을 보았다. 앞으로도 그래프들을 이용하여 탐색적 자료분석을 행하는 적용 예들이 더욱 다양하게 연구되어야 할 것이다.

참고문헌

[1] 渡部 洋, 鈴木規夫, 山田文康, 大塚雄作 (1985). 〈 探索的データ解析入門 〉, 朝倉書店, 東京.

[2] Benjamini, Y. (1988). Opening the box of the box, *The American Statistician*, vol. 42, 257-262.

[3] Berliner, L. M. (1992). Statistics, probability and chaos, with discussion and a rejoinder by the author, *Statistical Science*, vol. 7, 69-122.

[4] Box, G. E. P. (1999). Statistics as catalyst to learning by scientific method: Part II -a discussion, *Journal of Quality Technology*, vol. 31, 16-29.

[5] Carlyle, W. M., Montgomery, D. C. and Runger, G. C. (2000). Optimization problems and methods in quality control and improvement, *Journal of Quality Technology*, vol. 32, 1-17.

[6] Chatterjee, S. and Yilmaz, M. R. (1992). Chaos, fractals and statistics, *Statistical Science*, vol. 7, 49-68.

[7] Copeland, K. A. F. and Nelson, P. R. (1996). Dual response surface optimization via direct function minimization, *Journal of Quality Technology*, vol. 28, 61-70.

[8] Del Castello, E. (1996). Multiresponse process optimization via constrained confidence regions, *Journal of Quality Technology*, vol. 28, 61-70.

[9] Del Castello, E. and Montgomery, D. C. (1993). A Nonlinear programming solution to the dual response problem, *Journal of Quality Technology*, vol. 25, 199-204.

[10] Del Castello, E. Montgomery, D. C. and McCarville, D. R. (1996). Modified desirability functions for multiple response optimization, *Journal of Quality Technology*, vol. 28, 337-345.

[11] Derringer, G. (1994). A balancing act: Optimizing a product's properties, *Quality Process*, vol. 27, 51-58.

[12] Derringer, G. and Suich, R. (1980). Simultaneous optimization of several response variables, *Journal of Quality Technology*, vol. 12, 214-219.

[13] Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989). Some implementations of the box plots, *The American Statistician*, vol. 43, 50-54.

[14] Hintze, J. L. and Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism, *The American Statistician*, vol. 52, 181-184.

[15] Kim, K. and Lin, D. K. J. (1998). Dual response surface optimization: A fuzzy modeling approach, *Journal of Quality Technology*, vol. 30, 1-10.

[16] Kim, K. and Lin, D. K. J. (2000). Simultaneous optimization of mechanical properties of steel by maximizing exponential desirability functions, *Applied Statistics* vol. 49, 311-325.

[17] Koopmans, L. H. (1987). *Introduction to Contemporary Statistical Methods*, 2nd ed., Duxbury Press, Boston.

[18] Lai, D. and Chen, G. (2002). Testing chaos based on empirical distribution fuction: a simulation study, *Journal of Statistical Computation and Simulation*, vol. 72, 77-85.

[19] Lai, D. and Harrist, R. B. (1997). A nonparametric statistical approach in noisy chaos identification, *Communications in Statistics - Simulation and Computation*, vol. 26, 291-300.

[20] Lin, D. K. J. and Tu, W. (1995). Dual response surface optimization, *Journal of Quality Technology*, vol. 27, 34-39.

[21] Lorenz, E. N.(1963). Deterministic non-periodic flow, *Journal of Atmospheric Science*, vol. 20, 130-141.

[22] Luner, J. (1994). Achieving continuous improvement with the dual response approach: A demonstration of the Roman catapult, *Quality Engineering*, vol. 6, 691-705.

[23] McGill, R., Tukey, J. W. and Larsen, W. A. (1978). Variations of box plots, *The American Statistician*, vol. 43, 50-54.

[24] Milton, J. S.(1992). *Statistical Methods in the Biological and Health Sciences*, McGraw-Hill, Inc., New York.

[25] Myers, R. H. (1999). Response surface methodolgy - Current status and future directions, *Journal of Quality Technology*, vol. 31, 30-44.

[26] Peitgen, H-O, Jurgen, H., and Saupe, D. (1992). *Chaos and Fractal - New Frontiers of Science*, Springer-Verlag, New York.

[27] Tufte, E. R.(1983). *The Visual Display Quantitative Information*, Graphics Press, Cheshire

[28] Vining, G. G. (1998). Compromise approach to multiresponse optimization, *Journal of Quality Technology*, vol. 30, 309-313.

[29] Vining, G. G. and Myers, R. H. (1990). Combining Taguchi and response surface philosophies: A dual response approach. *Journal of Quality Technology*, vol. 22, 38-45.

[2001년 12월 접수, 2002년 8월 채택]

A Study for the Application of Graphs in Exploratory Data Analysis

Dae-Heung Jang[1)]

ABSTRACT

Revelation is emphasized in exploratory data analysis. Hence, many graphical techniques are used in exploratory data analysis. Three applications of graphs in exploratory data analysis are presented.

Keywords: Stem-and-leaf plot; Box plot; Chaos; Multiresponse optimization.

1) Professor, Division of Mathematical Sciences, Pukyong National University.
E-mail: dhjang@pknu.ac.kr