13강 다변량 자료의 시각화 이해 2

정보통계학과 이태림 교수

1. 평행좌표플롯을 작성하고 해석할 수 있다.
2. 주성분 분석으로 차원을 축소할 수 있다.
3. 공간자료를 표현하고 해석할 수 있다.
4. 다변량 시각화의 R에 의한 표현

학습개요 (다변량 자료의 시각화 2)

평행좌표 플롯
Parallell coordinate plot

- ▶ 평행좌표 플롯
- ▶ 평행좌표 조건부 플롯
- ▶ 끝잇기 알고리즘

주성분분석
(principal component analysis)

- ▶ 주성분점수
- ▶ 차원축소

공간자료 시각화
(scatter plot matrix)

- ▶ 등고선도
- ▶ 전망도

1. 평행좌표 플롯 시각화

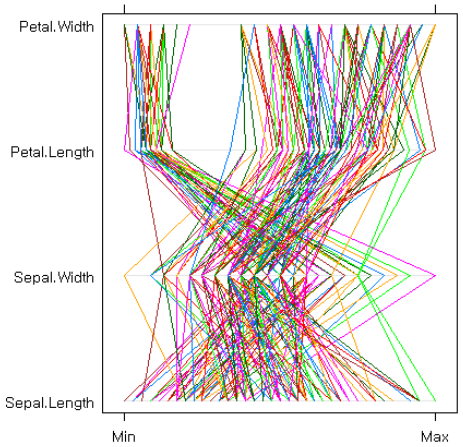
(Parallell coordinate plot)

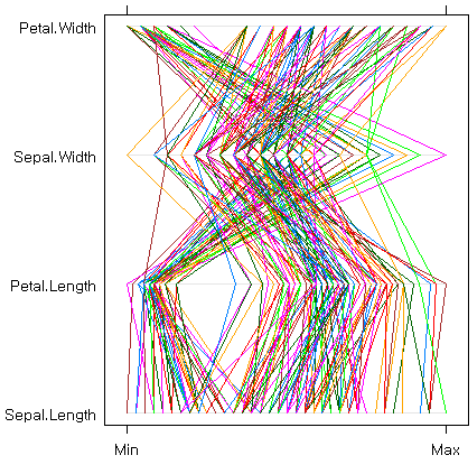
1 평행좌표 플롯 (parallel coordinate plot)

평행좌표 플롯 (parallel coordinate plot) :

PCP

변수들을 각기 평행선 상에 타점하여 점들을 연결하여 얻은 그래프





1 평행좌표 플롯 (parallel coordinate plot)

평행좌표 플롯 (parallel coordinate plot) R 코드:

PCP

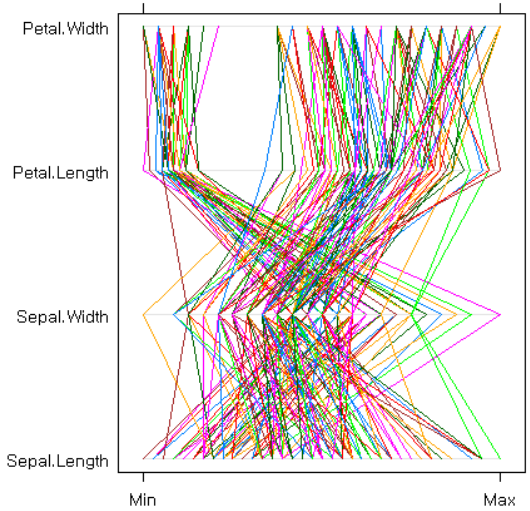
```
library(lattice)

    data(iris)
    parallel(~iris[,1:4])
    parallel(~iris[,c(1,3,2,4)])
    parallel(~iris[,1:4] | iris$Species)
    parallel(~iris[,c(1,3,2,4)] | iris$Species)
```

1 평행좌표 플롯 (parallel coordinate plot)

평행좌표 플롯 (parallel coordinate plot)

PCP

Sepal.Length와 Sepal.Width의 연결선들은 서로 엇갈리는 방향으로 놓이는 경향을 보이는데 이런 패턴은 두 변수 간 음의 상관



- Sepal.Width와 Petal.Length의 연결선들도 서로 엇갈리는 음의 상관

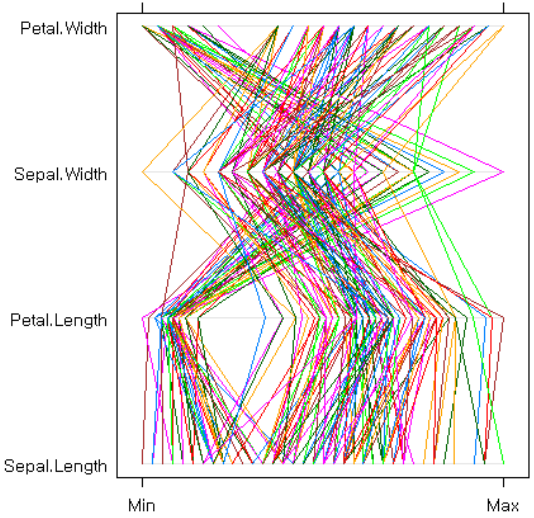
- Petal.Length와 Petal.Width의 연결선들은 나란한 경향을 보이는데 이런 패턴은 두 변수간 양의 상관

1 평행좌표 플롯 (parallel coordinate plot)

평행좌표 플롯 (parallel coordinate plot)

PCP

Sepal.Length와 Petal.Length의 연결선들은 서로 평행한 방향으로 놓이는 경향을 보이는데 이런 패턴은 두 변수 간 양의 상관



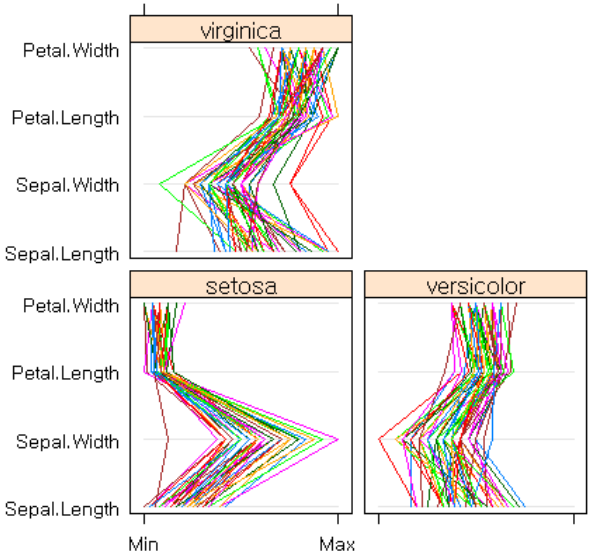
- Sepal.Width와 Petal.Width의 연결선들은 서로 엇갈리는 음의 상관
- Petal.Length와 Sepal.Width의 연결선들은 엇갈린 경향을 보이는데 이런 패턴은 두 변수간 음의 상관

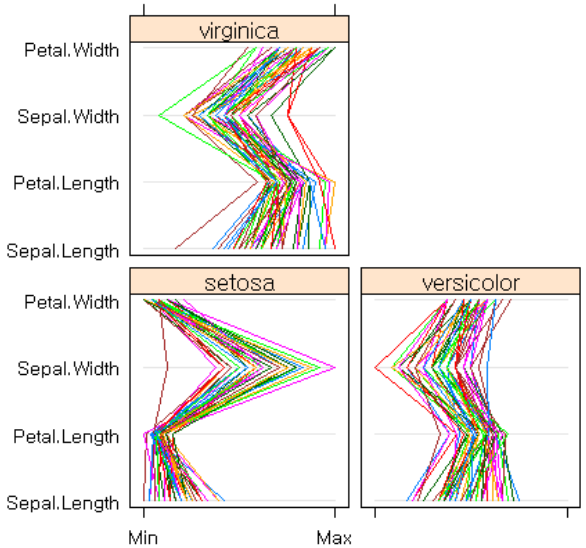
1 평행좌표 플롯 (parallel coordinate plot)

조건부 평행좌표 플롯 (conditioning parallel coordinate plot)

PCP

품종군별 Sepal.Width, Sepal.Width, Petal.Width, Petal.Length의 평행좌표 플롯





1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

[0] 개 변수 간 상관계수 행렬을 산출하여 상관도가 큰 순서대로 변수 쌍을 나열한다. 예를 들어 6개( ) 변수 자료에서 다음 순서로 변수 쌍이 나열되었다고 하자.
1–5, 4–6, 3–6, 2–6, 1–4, 1–2, 3–4, 4–5, 2–4, 1–6, 5–6, 1–3, 3–5, 2–5, 2–3,

1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

[1] 가장 상관도가 높은 두 변수를 연결하여 묶는다. 나머지 변수들은 독자적 묶음으로 간주된다. 예에서, 변수 묶음은 1–5, 2, 3, 4, 6이 된다 (변수 묶음의 수는 5개).

[2] 변수 묶음이 1개가 될 때까지, 변수 묶음

양쪽 끝 간 상관도가 가장 큰 변수 묶음을 찾아

해당 끝을 연결한다. 예에서, 최우선 연결은 4–6

1–5, 4–6, 2, 3 이 된다.

1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

다음의 우선 연결은 3–6 (or 6–3)이다. 따라서 변수 묶음은

이 된다.

1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

3] 다음의 우선 연결은 3–6 (or 6–3)이다. 따라서 변수 묶음은

1–5, 4–6–3, 2

다음 우선 연결은 2–6이지만 6은 변수 묶음의 끝에 있지 않다.
따라서 부적합하다. 다음 우선 연결은 1–4이다. 따라서 변수 묶음은

5–1–4–6–3, 2

다음 우선 연결은 1–2이지만 1이 변수 묶음의 끝에 있지 않으므로 부적합.
이어지는 우선 연결 3–4, 4–5, 2–4, 1–6, 5–6, 1–3, 3–5도 마찬가지로
적합하지 않다.

1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

다음 우선 연결 2–5는 적합하다.
이에 따라 변수 묶음은
2–5–1–4–6–3
이 된다 (or 3–6–4–1–5–2).
모든 변수들이 연결되었으므로 끝.

1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

	Sepal. Length	Sepal. Width	Petal. Length	Petal. Width
Sepal.Length	1.00	-0.12	0.87	0.82
Sepal.Width	-0.12	1.00	-0.43	-0.37
Petal.Length	0.87	-0.43	1.00	0.96
Petal.Width	0.82	-0.37	0.96	1.00

Petal.Length – Petal.Width

Sepal.Length – Petal.Length – Petal.Width

Sepal.Width – Sepal.Length – Petal.Length – Petal.Width

1 평행좌표 플롯 (parallel coordinate plot)

끝잇기 알고리즘(endlink) R 코드: Hurley(2004)

PCP

```
library(gclus)
    round(cor(iris[,1:4]),2)
    order <- order.endlink(cor(iris[,1:4]))
    order
    round(cor(iris[,order]),2)
    parallel(~iris[,c(order)])
```

2. 차원축소에 의한 시각화

2 차원축소에 의한 시각화(주성분분석)

주성분분석(principal component analysis)

다변량 자료 분석의 근본 문제는 차원 수가 크다는 데서 비롯되므로
차원 수를 줄이기 위한 방법

사영(projection)

: p-차원 공간에 놓이는 표준화 변환 관측개체들을 1차원 공간으로 축소시키는 방법

: p-차원 공간의 단위 벡터에 내려진 x_1 , x_2 …. x_n 의 그림자는

$$(x_1^t u)\, u \ , \cdots \ , (x_n^t u)\, u$$

2 차원축소에 의한 시각화(주성분분석)

차원축소(dimension Reduction)

제곱크기(squared norm)의 보정 평균을 최대화함으로써 사영으로 인한 손실을 최소화

maximize $\sum_{i=1}^{n} (x_i^t u)^2/(n-1)$ with respect to (w.r.t.) u

subject to $u^t u = 1.$

$\sum_{i=1}^{n} x_i x_i^t/(n-1)$ $(= R)$, 여기서 R은 상관행렬

$R u_1 = \lambda_1 u_1, \quad u_1^t u_1 = 1.$

$x_1^t u_1, \cdots, x_n^t u_1$

$(x_1^t u_1, x_1^t u_2), \cdots, (x_n^t u_1, x_n^t u_2)$

2 차원축소에 의한 시각화(주성분분석)

주성분 부하(principal component loadings)

변수 1…..변수p 의 위치점

1차원 공간 $u_{11}, \cdots, u_{1p}$

2차원 공간 $(u_{11}, u_{21}), \cdots, (u_{1p}, u_{2p})$

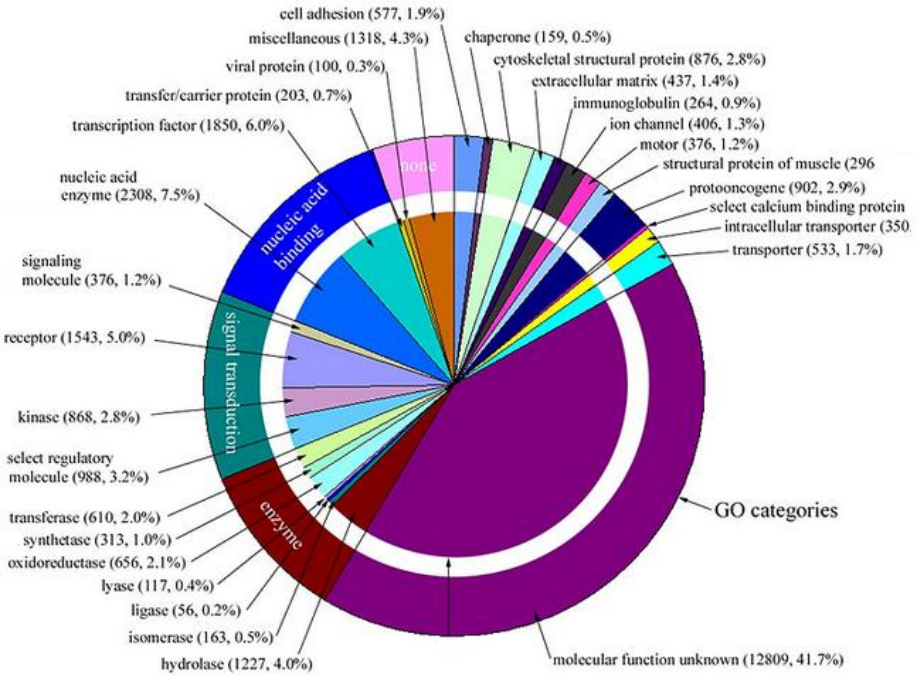
주성분분석 R 함수

```
princomp( x, cor=T )
```

- 여기서 x는 다변량 자료행렬이고 cor=T이면 상관 행렬 분해(표준화 변환)를, cor=F이면 공분산행렬 분해(중심화 변환)를 지시
- 출력 객체(object)에 주성분 점수는 scores에, 주 성분 부하는 loadings에 남는다.

2 차원축소에 의한 시각화(주성분분석)

주성분 분석 예제(PCA) : 유럽 국가들의 종류별 단백질 섭취 자료를 분석



- Albania, Austria, …, Yugoslavia 등 25개(= n) 나라
- 변수 : Beef, Chicken, Egg, Milk, Fish, Cereal, Potato, Bean, Fruit 등 9개(= p) 단백질 종류

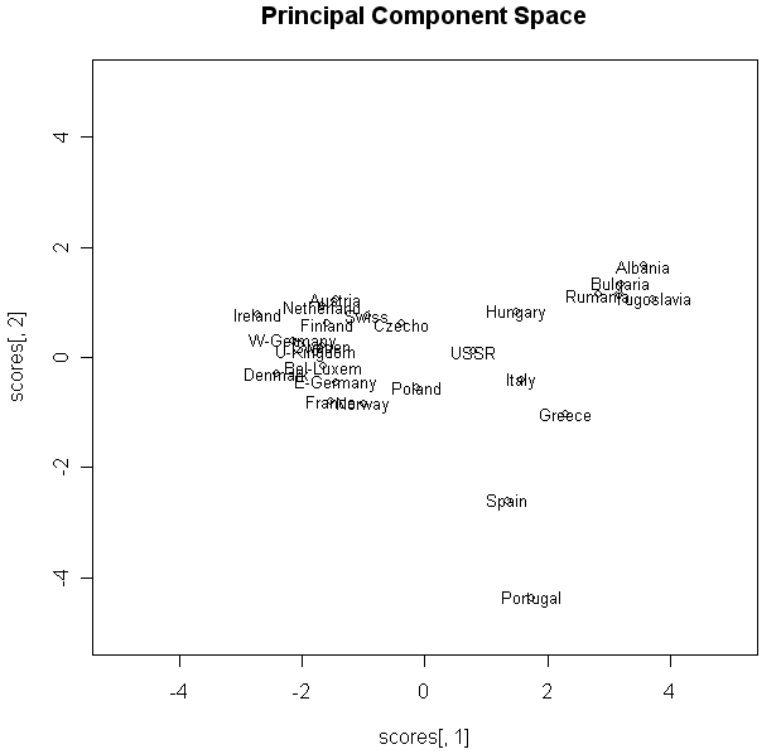
2 차원축소에 의한 시각화(주성분분석)

주성분 분석 예제(PCA) : R 프로그램

```
protein <- read.table("protein.txt", header=T)
                protein
pca <- princomp(protein[,2:10], cor=T)
pca
names(pca)
pca$loadings[,1:2]
pca$scores[,1:2]
attach(pca)
plot(scores[,2] ~ scores[,1], main = "Principal Component Space",
xlim=c(-5,5), ylim=c(-5,5))
text(y=scores[,2], x=scores[,1], label=protein$Country, cex=0.8)
x11()
plot(loadings[,2] ~ loadings[,1], main = "Principal Component
Loadings", xlim=c(-1,1), ylim=c(-1,1))
text(y=loadings[,2], x=loadings[,1], label=colnames(protein[,2:10]),
                              cex=0.8)
for (i in 1:9) {arrows(0,0,0.8*loadings[i,1],0.8*loadings[i,2], length = 0.1)}
```

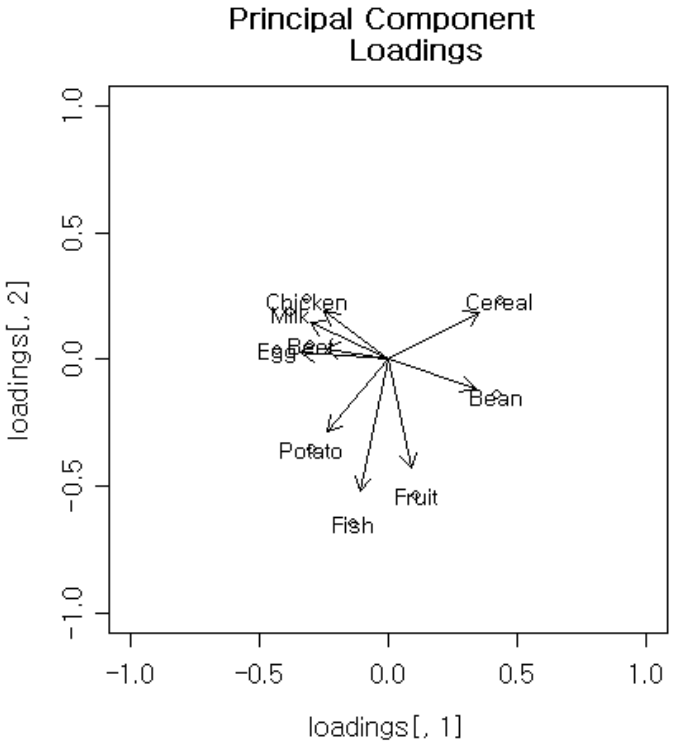
2 차원축소에 의한 시각화(주성분분석)

주성분 분석 출력결과 (principal component space)



2 차원축소에 의한 시각화(주성분분석)

주성분 분석 출력결과 (principal component loadingss)



- 아래 약간 오른쪽에 포르투갈과 스페인이 있는데 이들은 생선(fish), 콩(bean), 그리고 과일(fruit) 섭취로 특성화된다.
- 그리스와 이탈리아는 콩(bean) 섭취와 관련이 깊다.

3. 공간자료의 표현

3 공간자료의 표현

공간자료(special data)

$$\{(x_1, y_1, z_1), \cdots, (x_n, y_n, z_n)\}$$

- (x_i, y_i)는 2차원 평면상의 위치 좌표를 나타내고 z_i 는 그 위치점에서의 특성값

- 위치점 집합이 그리드(grid)로 표현되는 경우

$$\{(x_{jk}, y_{jk}) : x_{jk} = a_1 + h_1 j, y_{jk} = a_2 + h_2 k\}_{j=0,1,\cdots,n_1, k=0,1,\cdots,n_2}$$

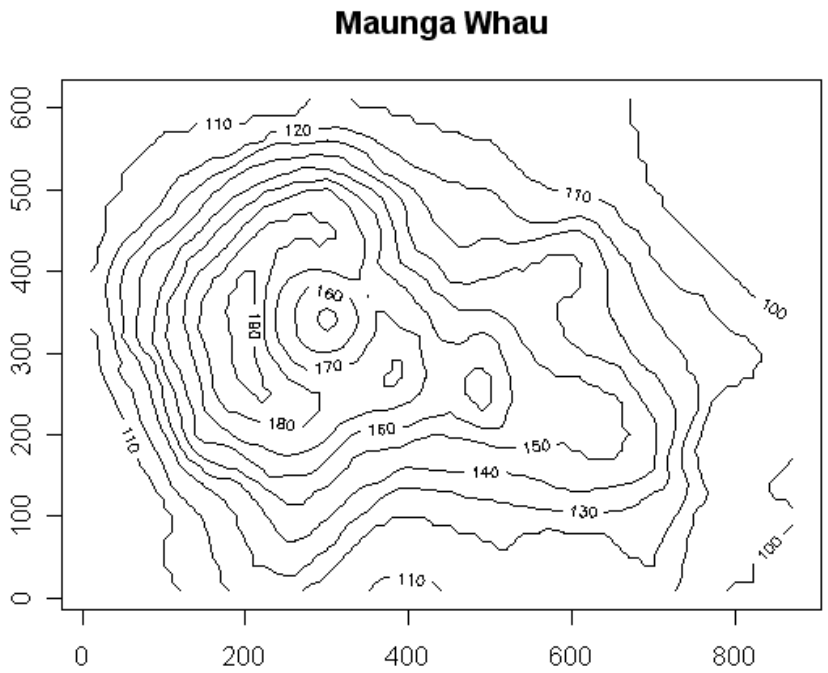
3 공간자료의 표현

공간자료의 시각화 예제

Auckland의 Maunga Whau 화산의 87x61 그리드 상에서의 고도

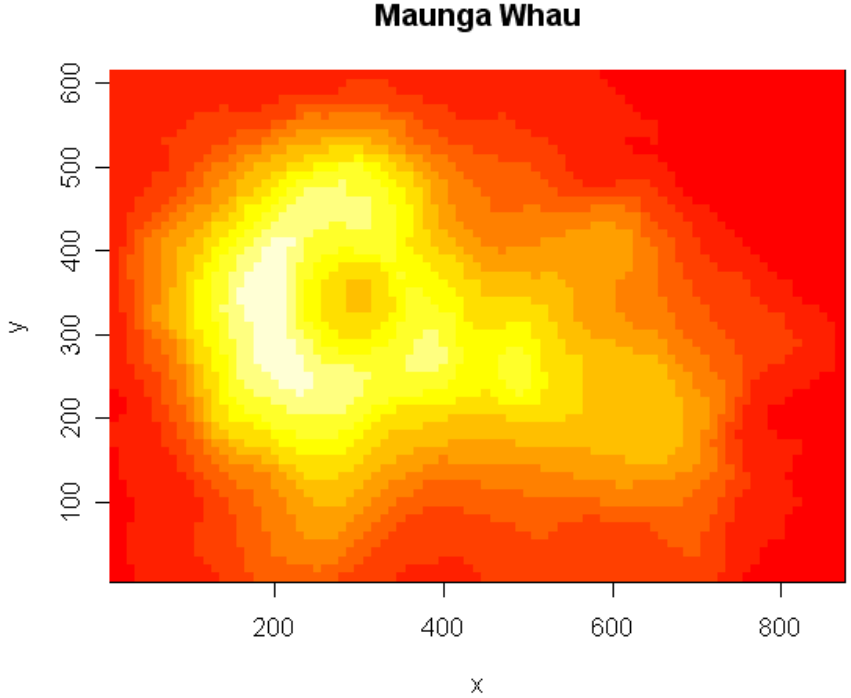
```
x <- 10*(1:87); y <- 10*(1:61)
  contour(x,y,volcano,main="Maunga Whau")
  image(x,y,volcano,main="Maunga Whau")
filled.contour(x,y,volcano,main="Maunga Whau")
 persp(x, y, volcano,phi=30,theta=30,scale=F,main="Maunga Whau")
```

공간자료의 시각화 예제 (출력결과)



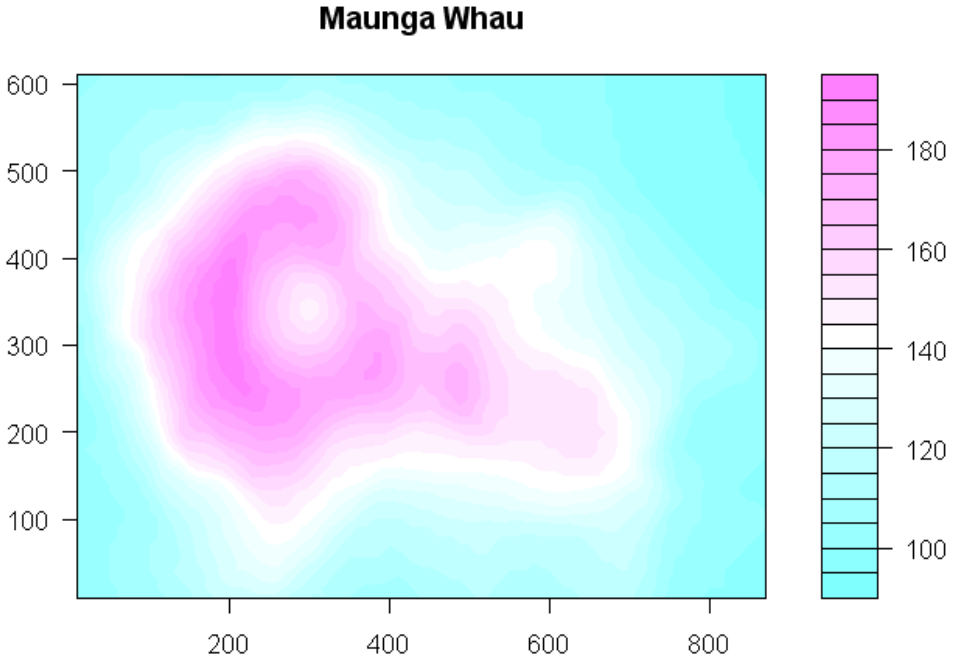
volcano 자료에 대한 등고선도

공간자료의 시각화 예제 (출력결과)



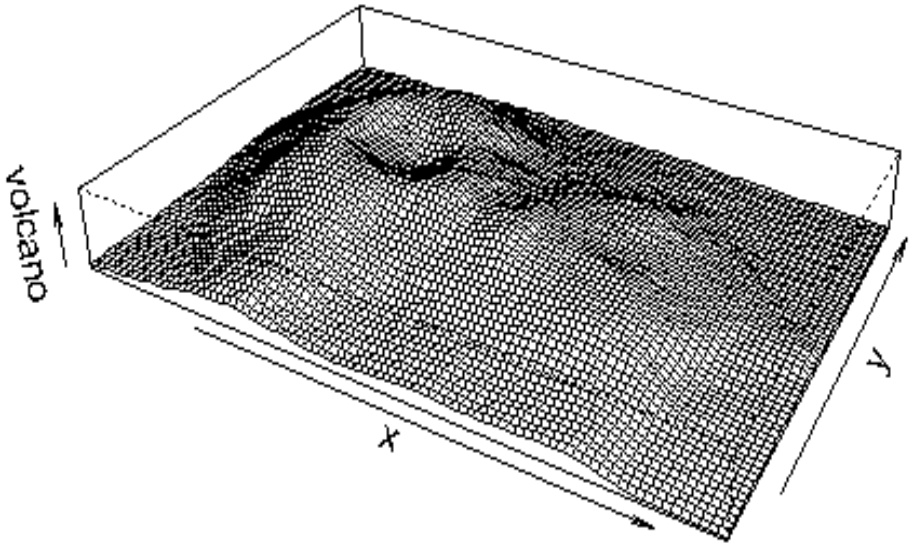
volcano 자료에 대한 이미지 그래프

공간자료의 시각화 예제 (출력결과)



volcano 자료에 대한 칼라 등고선도

공간자료의 시각화 예제 (출력결과)



volcano 자료에 대한 전망도(perspective plot)

다음시간안내

웹을 이용한 동적 • 대화형 데이터 시각화 1