Procedimento em três passos de análise de dados ecológicos: estudo de um caso

Sergio Camiz
Sapienza Università di Roma
www.camiz.net sergio@camiz.net

Conteúdo

O problema da investigação

Camiz, Altieri and Manes (in press), «Pollution Bioindicators: Statistical Analysis of a Case Study», *Water, Air and Soil Pollution*.

- utilizar as plantas como bio-indicadores de poluição, empregando alguns parâmetros funcionais que se podem considerar como indicadores de stress.

- estabelecer o impacto que têm as condições de poluição sobre as folhas em ambiente natural;
- estabelecer quais parâmetros se podem medir e utilizar como indicadores indiretos de poluição ambiental.

Dados

- amostras de folhas de *Pinus pinea* L., coletadas em *árvores* em *três sítios*:
  - Roma — muito poluído;
  - Civitavecchia — potencialmente poluído;
  - Castelporziano — não poluído;
- en ***nove ocasiões***:
  - Fevereiro 1988
  - Março 1988
  - Maio 1988
  - Junho 1988
  - Julho 1988
  - Outubro 1988
  - Dezembro 1988
  - Janeiro 1989
  - Fevereiro 1989
- folhas de *três gerações*:
  - 1986
  - 1987
  - 1988

no total series temporais de 1 a 26 *meses de idade*.

Indicadores

- alteração das ceras (*Sta1 ... Sta5*), estimada en cinco níveis de dano crescente;
- freqüência relativa de Fungos e esporos (*Fusp*) e de Partículas (*Part*);
- Peroxidase (*POD*), segundo três frações: solúvel (*F1*), iônica (*F2*) e ligada à parede celular (*F3*);
- conteúdo de íons (*Sulf*atos, *Fosf*atos, *Cloret*os).
- dados de sítios (quando disponível)
  - . Dióxido de Enxofre (µg/m3) (SO_2);
  - . Dias de chuva nos 30 antes das coletas (*Gpio*);
  - . Quantidade de chuva em mm. (*Piog*);
  - . Temperatura mínima nos 30 dias antes (*Tmin*);
  - . Temperatura média dos 30 dias antes (*Tmed*);
  - . pH das chuvas (*pH*).

Física, Ecologia e Ciências Humanas

- Em física se empregaram modelos matemáticos desde Galilei:
  - . é possível construir experimentos:
  - . os modelos são funcionais;
  - . a matemática que se emprega pode ser muito complexa, para descrever fenômenos muito difíceis de compreender.
- Em ecologia e ciências humanas o emprego de modelos matemáticos é muito recente:
  - . é mais difícil construir experimentos, mas se podem obter dados de campo ou fazer entrevistas;
  - . os modelos descrevem tendências, causadas às vezes por muitos fatores;
  - . a matemática empregada deve ter em conta muita incerteza.

A quantidade de dados que se obtêm numa investigação em ecologia e em ciências humanas é muito grande e os dados podem ser heterogêneos.

Os objetivos podem ser também heterogêneos.

Assim é necessário um enfoque específico: —> *Análise de Dados*.

Para seu emprego, há que estruturar em três passos as análises e o estudo mesmo, porque em cada passo há diferentes objetivos, de forma que os métodos que se empregam em um passo não podem ser empregados nos outros.

As três etapas de uma investigação

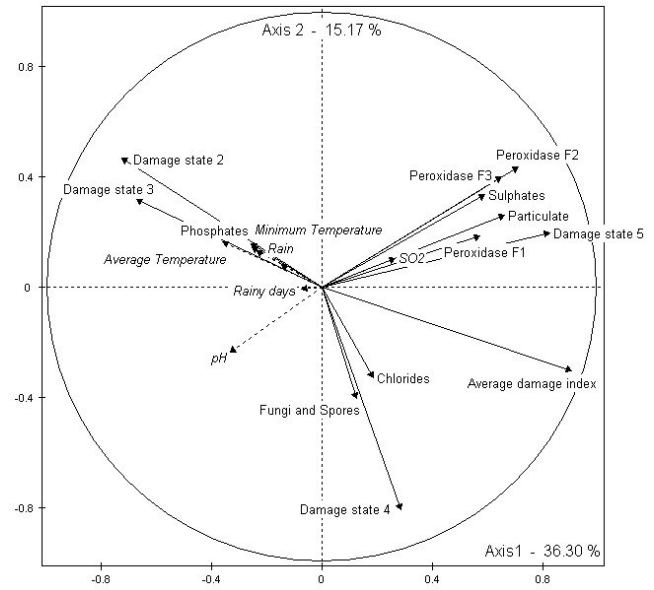
- *Exploratória*:
  - . Quadro de referência e objetivos do estudo, coleta de dados;
  - . busca de estruturas e relações;
  - . formulação de hipóteses.
- *Confirmatória*:
  - . projeto experimental;
  - . construção de relações;
  - . teste de hipóteses, inferência estatística.
- *Modelos*:
  - . formulação de um modelo matemático;
  - . implementação, calibração;
  - . simulação, predições.

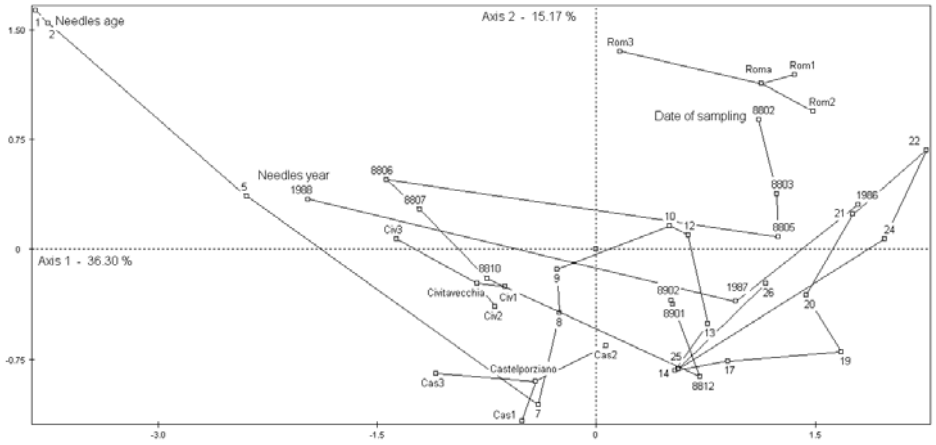
Análise exploratória

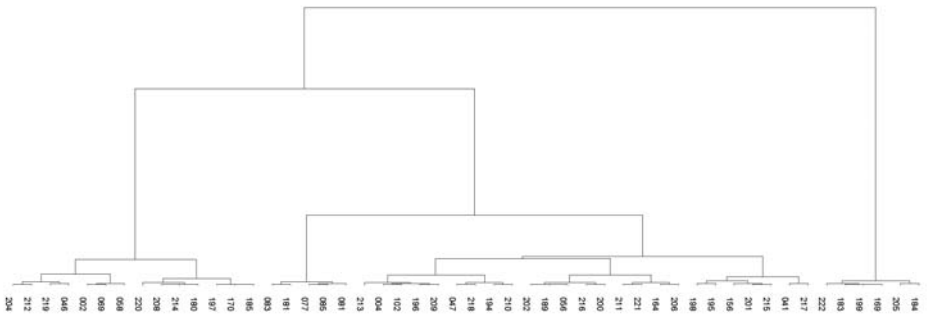
Neste passo é necessário *estruturar* os dados de acordo com *ordenação* e *classificação*, ou seja organizá-los de maneira a identificar as fontes da diversidade e os grupos de dados homogêneos.

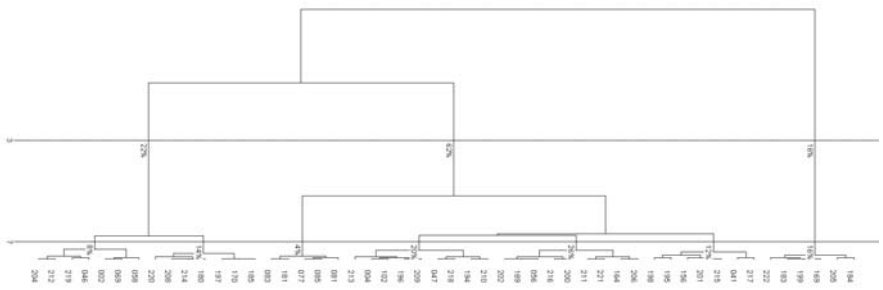
- **Controle e primeiro estudo dos dados:**
  - **. Estatísticas descritivas**
- **Busca de relações e fatores:**
  - **. Análise de componentes principais ou de correspondência**
- **Busca de estruturas:**
  - **. Classificação (hierárquica)**

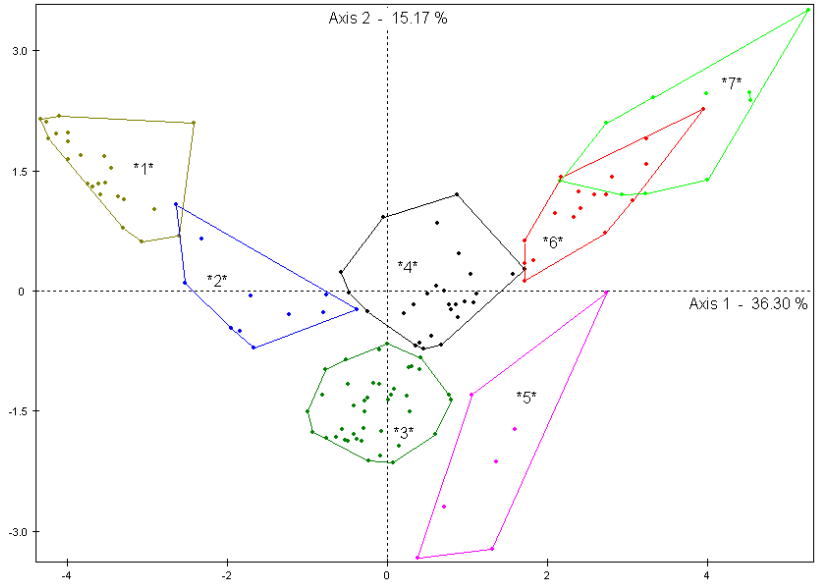
Ao final se interpretam os resultados integrando fatores e classes com o que se conhece dos caracteres e das unidades.



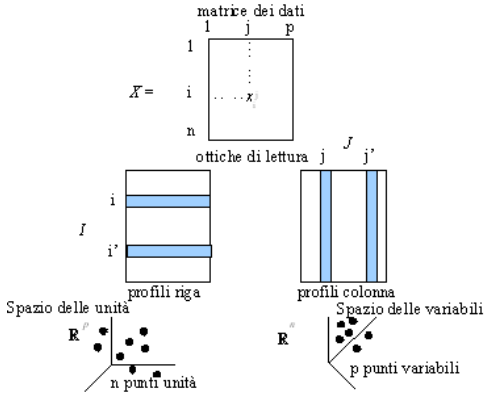








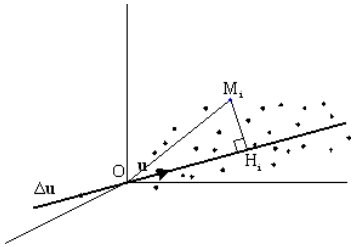
Análise de componentes principais



As unidades são imaginadas representadas em um espaço R^p gerado para os caracteres, que formam uma base ortogonal e se busca uma nova base ortogonal que maximiza a inércia dos dados projetada sobre os "primeiros" vetores.

Como a inércia em relação à origem vale

$$\sum_{i=1}^{n} p_i (OM_i)^2 = \sum_{i=1}^{n} p_i (M_i H_i)^2 + \sum_{i=1}^{n} p_i (OH_i)^2$$



como direção mais importante tem que buscar a direção que minimiza a primeira quantidade e maximiza a segunda.

Em geral, se X é a tabela de dados centralizados (média $= 0$), N é a matriz diagonal dos pesos das unidades ($\sum_{i=1}^{n} p_i = 1$) e M a matriz diagonal da métrica do espaço (em *ACP*, $\boldsymbol{M} = \mathbf{diag}(1/\sigma_i^2)$), a inércia em relação a uma reta vetor de comprimento unitário $\boldsymbol{u}$ é:

$$I_{r^\perp} = \|c\|_N^2 = c'Nc = (XMu)'N(XMu) = u'MXNXMu$$

Para encontrar a direção de inércia máxima, tem que resolver o problema de maximização

$$\begin{cases} u'MX'NXMu = Max_u \\ u'Mu = 1 \end{cases}$$

e como sua Lagrangiana é

$$\mathcal{L} = u'MX'NXMu - \lambda\,(u'Mu - 1)$$

sua solução é

$$X'NXMu = \lambda u$$

com λ sendo o máximo autovalor de $X'NXM$, e $\boldsymbol{u}$ seu autovetor correspondente. λ é a variância dos dados sobre a direção $\boldsymbol{u}$.

Em geral, para a descomposição em autovalores

$$X'NXM = U\Lambda U'$$

e se ordenam autovalores e autovetores em ordem decrescente de λ.

Resulta também que

$$NXMX' = XU\Lambda U'X' = V\Lambda V'$$

de forma que os mesmos autovalores servem para a análise no espaço $\mathbb{R}^n$.

Pode ser demonstrado que se pode reconstruir a matriz dos dados com

$$\underset{(n,p)}{X} = \sum_{\alpha=1}^{p} \sqrt{\lambda_\alpha}\ \underset{(n,1)}{v_\alpha}\ \underset{(1,p)}{u_\alpha'}$$

Assim se se ordenam autovalores e autovetores em ordem decrescente, resulta que:

Teorema de Eckart e Young: a reconstrução da matriz utilizando os primeiros $r < p$ autovalores e autovetores

$$\underset{(n,r)}{X} = \sum_{\alpha=1}^{r} \sqrt{\lambda_\alpha}\ \underset{(n,1)}{v_\alpha}\ \underset{(1,r)}{u_\alpha'}$$

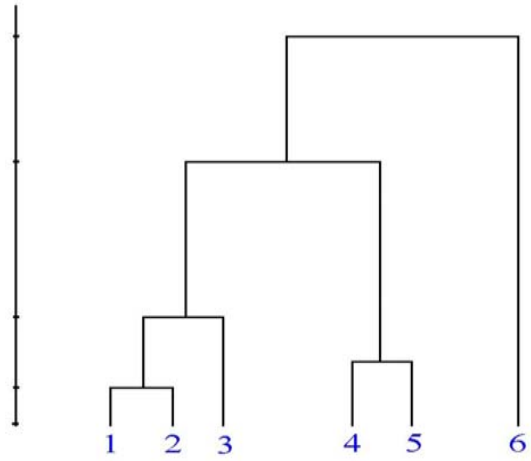
é a melhor de posto r.

Classificação hierárquica

A construção de um *dendrograma* permite conhecer o conjunto de relações entre os objetos que se quer classificar através de uma taxonomia.

Cortando o dendrograma se obtêm as partições necessárias.

Dendrograma



Clasificação hierárquica ascendente

Para construir um dendrograma, é necessário

- decidir um critério para associar os objetos;
- eleger um índice de associação;
- eleger um critério para associar objetos e grupos (função objetivo) a cada passo;
- buscar os objetos e os grupos que agregando-se otimizam a função objetivo;
- eleger um critério para voltar a calcular o índice entre objetos e grupos já criados.

Algoritmo iterativo

- No início cada objeto é um grupo (singleton) e se constroem as relações de associação entre objetos.

- A cada passo:
  - se elege o par de grupos que unindo-se otimizam a função objetivo;
  - os dois grupos combinam-se em um novo grupo;
  - se calcula a associação do novo grupo com os outros;
- se repete essa operação n-1 vezes.

- O processo termina quando não há apenas um grupo.

Classificação hierárquica ascendente

Funções objetivo:

- ***Ligação completa***: a associação entre dois grupos é a pior entre unidades de ambos grupos;
- ***Ligação simples***: a associação entre dois grupos é a melhor entre unidades de ambos grupos;
- ***Ligação promedia***: a associação entre dois grupos é a média entre unidades de ambos grupos;
- *Ward:* a associação entre dois grupos é o incremento na variância dentro dos grupos.

Resultados sintéticos

- distinção entre dano natural, dano antrópico, condições ambientais, distinto comportamento dos níveis de dano;
- diferentes condições para os três sítios;
- variação por anos das folhas e por meses de amostragem;
- distinção entre folhas de diferentes sítios dependente da idade.

Análise confirmatória

- Teste de hipóteses
- Diferença entre
  - . sítios
  - . idade das folhas
  - . estações
- *ANOVA*

		Total	*Roma*	*Civitavecchia*	*Castelporziano*	*1986*	*1987*	*1988*
Sulphate	*N*	131	47	40	44	18	63	50
	Mean	986.305	1389.659 A	1141.225 B	414.614 C	1647.833 A	1089.968 B	617.54 C
	St.dev	728.816	805.976	599.51	195.915	1152.36	602.916	385.39
Phosphates	*N*	131	47	40	44			
	Mean	291.389	230.234 CC	302.675 AB	346.454 AA			
	St.dev	210.84	206.191	191.117	215.845			
Chlorides	*N*	131	47	40	44	18	63	50
	Mean	1266.145	924.149 B	981.15 B	1890.545 A	1699.667 A	1420.968 A	915 B
	St.dev	1277.783	591.999	270.49	1957.438	713.658	1677.167	573.582
Peroxidase F1	*N*	129	47	38	44	18	62	49
	Mean	0.343	0.594 A	0.057 C	0.321 B	0.972 A	0.408 B	0.029 C
	St.dev	0.678	0.969	0.136	0.434	0.822	0.748	0.059
Peroxidase F2	*N*	130	46	40	44	18	62	50
	Mean	0.049	0.109 A	0.01 B	0.024 B	0.08 A	0.064 A	0.02 B
	St.dev	0.072	0.091	0.015	0.024	0.085	0.082	0.032
Peroxidase F3	*N*	131	47	40	44	18	63	50
	Mean	0.107	0.241 A	0.031 B	0.032 B	0.109 B	0.159 A	0.04 C
	St.dev	0.156	0.193	0.041	0.034	0.132	0.19	0.065

Resultados sintéticos

- diferenças significativas entre sítios, anos e meses de amostragem;
- interação significativa entre sítios e idade;
- o comportamento dos indicadores pode ser diferente para os sítios, com a variação da idade, as vezes com uma componente estacional;
- os cloretos dependem de uma tempestade en Castelporziano em Outubro, assim não é interessante seguir seu estudo.

Modelos

Para os íons e as peroxidases:

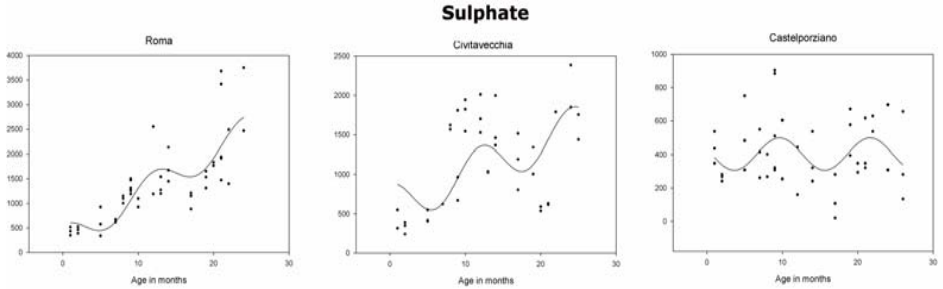
$$\textbf{linear} \quad f(x) = I + Lx + SF(x) + \varepsilon ,$$

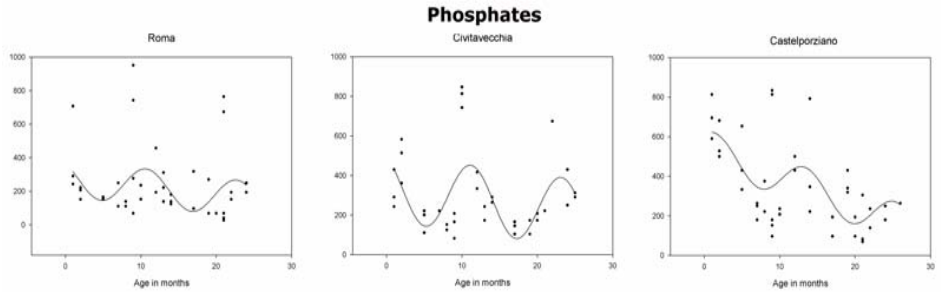
$$\textbf{quadrático} \quad f(x) = I + Lx + Qx^2 + SF(x) + \varepsilon ,$$

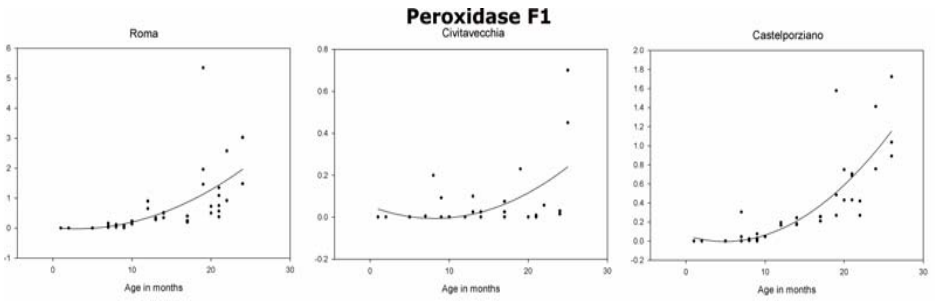
$$\textbf{exponencial} \quad f(x) = I + Se^{Ex} + SF(x) + \varepsilon ,$$

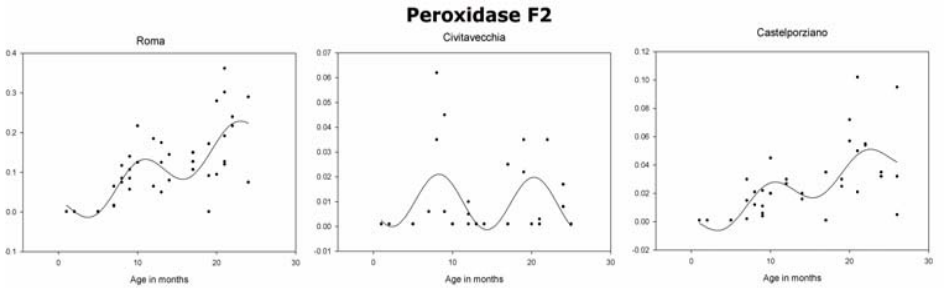
mais a componente estacional

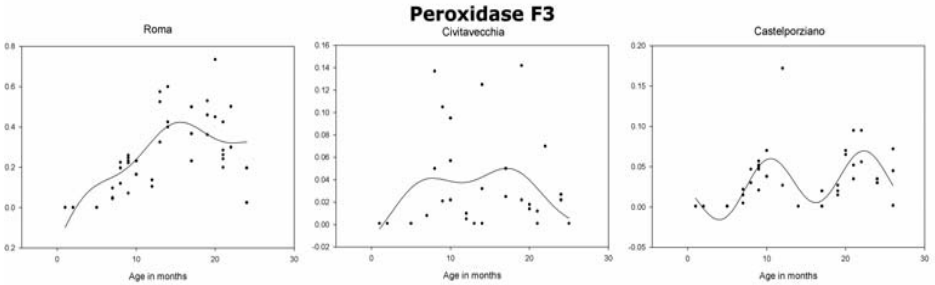
$$SF(x) = A\cos(\pi x + P)$$











Conclusões

- **Os indicadores são sensíveis aos diferentes estados de poluição;**
- **A poluição é indicada pelas diferentes tendências dos indicadores, e não pelos valores medidos em um dado instante;**
- **Alguns indicadores têm capacidade de recuperação estacional;**
- **O uso de análise exploratória permitiu orientar-se na direção de modelos separados por sítio e considerar variações estacionais: algumas hipóteses foram confirmadas pela *ANOVA*;**
- **Os modelos permitem verificar também o que já se conhece do comportamento dos indicadores sob stress.**

Bibliografia

Benzécri, J.P., 1973. *L'analyse des données*. 2 vols., Dunod, Paris.

Camiz, S., 1993b. «Computer assisted procedures for structuring community data». *Coenoses*, 8(2): 97-104.

Camiz, S., 2001. «Exploratory 2- and 3-way Data Analysis and Applications». *Lecture Notes of TICMI*, http://www.emis.de/journals/TICMI/lnt/vol2/lecture.htm. Tbilisi University Press, vol. 2.

Camiz, S., A. Altieri , F. Manes, 1993. «Effetti degli inquinanti atmosferici su aghi di *Pinus pinea* L. in ambiente naturale». *Statchem93 - Atti del Convegno su Statistica e Chemiometria per lo studio dell'ambiente*, Università di Venezia, Società Italiana di Statistica.

Lebart, L., A. Morineau, and K.M. Warwick, 1984. *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for large Matrices*. John Wiley & Sons, New York.

Miller, R.G.Jr., 1981. *Simultaneous Statistical Inference*. Springer Verlag, New York.

Mood, A.M., F.A. Graybill and D.C. Boes, 1974. *Introduction to the theory of Statistics*. McGraw-Hill, New York.