IDENTIFICAÇÃO DE OBSERVAÇÕES INFLUENTES NA CLASSIFICAÇÃO DE AMOSTRAS DE PLANTAS DO GÊNERO *MINTHOSTACHYS* VIA ANÁLISE DISCRIMINANTE

Daniel Cañari CASAÑO[1]
Doris Gómez TICERÁN [2]
Olga Lidia Solano DÁVILA[1]
Yakov Quinteros GÓMEZ[3]
Joaquina Albán CASTILLLO[3]

- RESUMO: No presente estudo são explorados métodos de identificação de observações influentes no contexto de uma análise discriminante conduzida para classificar as medições feitas em 100 espécimes do gênero *Minthostachys* com pubescência abundante e em *Minthostachys* com pubescência escassa, recolhidos na província andina de Cajatambo do Departamento de Lima, Perú. Os dados usados no presente trabalho vêm de um inventário florístico realizado no ano de 2005. As variáveis morfológicas estudadas no ramo principal de cada *Minthostachys* foram: comprimento do peciolo, comprimento da folha e largura da folha. Estudos taxonômicos e sistemáticos das amostras foram realizados utilizando o sistema de classificação de Cronquist, que classificou 51 plantas de *Minthostachys* como de pubescência abundante e 49 plantas de *Minthostachys* com pubescência escassa. Através da análise foram discriminadas corretamente 92 das plantas de *Minthostachys* que sobre o total representa o 92%, um valor suficientemente grande para afirmar a eficácia da função discriminante. Para as 100 plantas de *Minthostachys* , eliminando uma informação por vez , foi calculado o valor da Distância de Mahalanobis, a probabilidade de erro de classificação e os escores da função discriminante de Fisher (Campbell, 1978; Fung, 1992, 1995). A análise discriminou corretamente 92 plantas de *Minthostachys* ou seja 92% de un total de 100, um valor suficientemente grande para evidenciar a eficácia da função discriminante. Das comparações dos valores da Distância de Mahalanobis, a probabilidade de erro de classificação, os escores da função discriminante de Fisher com e sem a observação em avaliação, as maiores mudanças nos valores dessas medidas envolvidas na análise discriminante, cada vez que se elimina uma observação, ocorreram quando foram retiradas as observações, 64, 90 e 100, portanto há evidencia significativa que essas observações são influentes.

- PALAVRAS CHAVES: Medida de influência, observação influente, análise discriminante linear, gênero *Minthostachys.*

1 Ministerio de Vivienda, Construcción y Saneamiento, Oficina de Estadística, CEP: 31, Lima, Perú. E-mail: *dcanari@vivienda.gob.pe*.
[2] UNMSM, Facultad de Ciencias Matemáticas, Departamento de Estadística, CEP: 31, Lima, Perú. E-mail: *dorisgomezt@gmail.com / solano_2010@gmail.com*
[3] UNMSM, Museo de Historia Natural, Departamento de Etnobotánica y Botánica Económica, CEP: 31, Lima, Perú. Yakov281 @hotmail.com. E-mail: *yakov281@hotmail.com / jalbanc@gmail.com*

1 Introdução

Desde 1985, um grupo de pesquisadores do Departamento de Etnobotánica e Botânica Econômica do Museu de História Natural de Lima pertencente a Universidad Nacional Mayor de San Marcos (UNMSM), tem interesse em estudos de populações de plantas medicinais andinas, particularmente do gênero *Minthostachys*, considerada uma das plantas medicinais mais relevante dos Andes do Peru. Trata-se de uma planta perene. Quando jovem é herbácea e na fase adulta é arbustiva, podendo atingir de 1 a 1,5 metros de altura. Suas folhas são verdes, pecioladas, lanceolada-elípticas e aromáticas. Geograficamente está distribuída ao longo da cordilheira dos Andes, desde a Venezuela, Colômbia, até a Argentina, crescendo entre 500 e 4000 metros acima do nível do mar . Geralmente crescem nas margens das lavouras ou em zonas úmidas e é utilizada pelos habitantes dos Andes do Peru desde tempos imemoriais, para fins medicinais, alimentício e, nos últimos anos o óleo extraído da planta tem sido comercializado, por exemplo, como repelente de insetos.

Esse conjunto de propriedades da planta se manifesta como recurso valioso que poderia ser melhor explorado em forma sustentável e contribuir para melhorar a saúde dos moradores dos Andes do Peru. Nesse contexto é relevante investigar o seu potencial, em especial em Cajatambo, dado que até o ano 2004 a planta não era encontrada nesse lugar. A Provincia de Cajatambo é uma comunidade andina localizada na parte ocidental dos Andes do Departamento de Lima, a uma altitude de 3.376 metros acima do nível do mar, com uma população de aproximadamente 9.618 habitantes, dos quais 56% pertencem a população indígena (INEI, 2005).

Em 2005, uma equipe de pesquisadores do Laboratório de Etnobotânica do Museu de História Natural da UNMSM fez um inventário florístico em Cajatambo e através das determinações taxonómicas, a maioria das *Minthostachys* foi identificada como da espécie *tomentosa*.

A análise estatística dos dados das variaveis morfológicas: o comprimento do pedúnculo, a largura do pedúnculo, o número de veias do cálice, o comprimento da corola, a largura da corola usando análise de componentes principais mostrou duas possíveis espécies de *Minthostachys* (Gomez et al, 2008), a *Minthostachys* com pubescência abundante e a *Minthostachys* com escassa pubescência. Após muitos anos de confusões taxonômicas e virtual indeterminabilidade de seus espécimes, Schmidt (2008), fez um resumo geral do estado do conhecimento sobre *Minthostachys,* com ênfase na etnobotánica e no conteúdo farmacológico do óleo.

No contexto descrito, o objetivo deste trabalho é identificar observações influentes aplicando as medidas desenvolvidas por Campbell (1978), Fung (1992) e Fung (1995) em dados de amostras do gênero *Minthostachys tomentosa*, com abundante pubescência, e com escassa pubescência, coletadas na Província de Cajatambo, do Departamento de Lima.

Trata-se de uma aplicação de técnicas de estatística multivariada conhecidos na literatura por análise discriminante ou discriminação e classificação que são frequentemente utilizados para simplificar o tamanho do problema estatístico (Anderson, 1984; Manly, 2005), onde os resultados, podem ser afetados pela presença de algumas observações que têm um comportamento diferente da maioria dos dados, que muitos pesquisadores têm chamado de observações discordantes, *outliers*, influentes (Beckman e Cook,1983). Têm-se desenvolvido muitos estudos com métodos ou medidas estatisticas

para detectar dados influentes (Muñoz et al., 2001). Uma observação influente ao ser omitida da análise, dá origem a alterações nas estimativas de alguns ou de todos os parâmetros envolvidos no estudo. Pode ser considerado como um caso especial de observação discordantes. Já um dado é discordante, quando na opinião do pesquisador, está localizado longe das outras observações que compõem o conjunto de dados para análise. Também é chamado aberrante ou dissidente, para citar alguns termos que tem-se atribuído ao longo dos anos (Beckman e Cook,1983).

É importante mencionar que haverão observações discordantes que não são influentes, desde que as estimativas dos parâmetros permaneçãm essencialmente inalteradas quando essas observações são omitidas (Beckman e Cook,1983).

A Análise de Influência, para Belsley et al. (1982) tem sido amplamente estudada e divulgada em várias aplicações de análise de regressão e no contexto da análise discriminante, foi abordada inicialmente por Campbell (1978), que propôs medidas de influência com base na função de influência dada por Hampel (1974).

Anos mais tarde, Fung (1992, 1995) com base na relação entre os coeficientes da função discriminante linear de Fisher e os coeficientes do modelo de regressão linear múltipla, propôs algumas medidas seguindo a metodologia utilizada na análise de regressão. Apresenta-se a seguir a teoria mais relevante para identificar observações influentes no contexto da análise discriminante.

2 Metodologia

Na análise discriminante o interesse principal é alocar um individuo $x=(x_1,...,x_p)'$ com p medidas, em um dos k grupos ou populações pré determinadas.

2.1 Análise discriminante linear em dois grupos

Sejam $G_1 e\ G_2$ as duas populações ou classes de objetos e $X^{(k)}=(X_1^{(k)},...,X_p^{(k)})'$, com $k=1,2$ um vetor aleatório de valores em R^p que contém as medições dos indivíduos de cada uma das populações, com os parâmetros $\mu^{(k)}=(\mu_1^{(k)},...,\mu_p^{(k)})'$ e Σ_k, e se valores observados do vetor aleatório $X^{(k)}$, diferem de um grupo para outro através de suas medidas, então cria-se uma regra para classificar o novo indivíduo, $x=(x_1,...,x_p)'$ de R^p em uma das duas populações G_1, G_2.

Dadas essas considerações, tomam-se amostras aleatórias de cada uma das populações, para estimar os parâmetros de interesse, onde $\bar{x}^{(1)}$, $\bar{x}^{(2)}$ e S são as estimativas dos vetores de médias e da matriz de covariância comum $\sum=\sum_k$, respectivamente.

Fisher (1936), partiu em busca de uma combinação linear do vetor x, $Y=\hat{\alpha}'x$, em cada população, de modo que seja o máximo da relação do quadrado da diferença de médias com à sua variância; ou seja, que fornece o máximo para a proporção:

$$\lambda = \frac{\left(\hat{\alpha}'\bar{x}^{(1)} - \hat{\alpha}'\bar{x}^{(2)}\right)}{\hat{\alpha}'S\hat{\alpha}}.$$

Demonstra-se que o vetor $\hat{\alpha}$ é proporcional a forma $S^{-1}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)$ e a combinação linear,

$$Y = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' S^{-1}x, \tag{1}$$

é conhecida como a função discriminante linear de Fisher.

Fazendo, $\hat{\alpha} = S^{-1}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)$, define-se a regra de classificação a seguir:

$$\begin{aligned} &\textit{Alocar } x \textit{ ao grupo } G_1 \textit{ se } \hat{\alpha}'x - \frac{1}{2}\hat{\alpha}'\left(\bar{x}^{(1)} + \bar{x}^{(2)}\right) \geq 0 \\ &\textit{caso contrario, alocar } x \textit{ ao grupo } G_2 \end{aligned} \tag{2}$$

Alguns aspectos importantes relacionados com a questão da discriminação em dois grupos são:

a. A Distância de Mahalanobis na população, $\Delta^2 = \left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$, estimada pela expressão;

$$\hat{\Delta}^2 = D^2 = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' S^{-1}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) \tag{3}$$

b) A probabilidade de classificar erroneamente uma observação do grupo G_j no grupo G_i seguindo a regra de classificação, R , dada por:

$$\hat{\Delta}^2 = P(i/j;R) = \phi\left(-\frac{1}{2}D^2\right) \tag{4}$$

Onde, ϕ é a função distribuição cumulativa normal no ponto $\left(-\frac{1}{2}D^2\right)$, $i, j = 1,2 \quad i \neq j.$

c) A função discriminante linear de Fisher dada por:

$$Y = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' S^{-1}x \tag{5}$$

d) Os escores da função discriminante linear de Fisher, dados como

$$\hat{\alpha}' x - \frac{1}{2}\hat{\alpha}'\left(\bar{x}^{(1)} + \bar{x}^{(2)}\right) \quad (6)$$

Um problema que muitas vezes aparece quando se faz análise discriminante é a presença de observações que alteram os valores das medidas: Distância de Mahalanobis, a probabilidade de erro de classificacão, a função discriminante linear de Fisher e os escores da função discriminante, envolvidas em nesta questão. Confrontado com este problema, tem se proposto na literatura um conjunto de técnicas para detectá-los denominadas análise de influência. A idéia básica por trás da análise de influência é comparar os valores das estimativas das medidas: Distância de Mahalanobis, a probabilidade de erro de classificação, a função discriminante linear de Fisher e os escores da função discriminante, com e sem a observação considerada influente.

Em vários estudos sobre o tema, o tipo de perturbação mais utilizado para avaliar a influência de uma observação, é a omissão de observações (Muñoz et al, 2001), por isso é de interesse avaliar o efeito da *i*-ésima observação multivariada, $x_i = (x_{i1},...., x_{ip})'$, em cada uma das estatísticas envolvidas na questão da análise discriminante.

2.2 Medida de influência para a Distância de Mahalanobis

Para avaliar a possivel influência da observação multivariada x na Distância de Mahalanobis da amostra, $D^2 = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' S^{-1}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)$, Fung (1992) propôs a seguinte função de influência:

$$\hat{I}_M\left(x;\hat{\Delta}^2\right) = w_1\left[\hat{\psi} - w_1^{-1}\right]^2 \quad (7)$$

onde $\hat{\psi} = \hat{\alpha}'\left(x - \bar{x}^{(k)}\right)$ e w_k é o peso de cada grupo na formação da matriz de covariância; assim, $w_1 = \frac{(n_1 - 1)}{n_1 + n_2 - 2}$ e $w_2 = 1 - w_1$.

Esta medida depende em grande parte da estatística $\hat{\psi}$, que compara cada observação com o vetor de medições do grupo ao que pertence, ponderado pelos coeficientes da funcão discriminante linear de Fisher.

2.3 Medidas de Influência para a probabilidade do erro de classificação

A probabilidade de má classificação quantifica a probabilidade de alocar erroneamente o vetor com medidas $x = (x_{m_1},.....,x_{mp})'$, no grupo G_i quando na realidade pertence ao grupo G_j. Para uma regra de classificação R , a probabilidade de erro de classificação foi definido em (4) como $P(i/j;R) = \phi\left(-\frac{1}{2}D^2\right)$. Para avaliar a possível influência

da a *i*-ésima observação multivariada, sobre a probabilidade de erro de classificação, Hampel (1974) propôs a seguinte função de influência:

$$\hat{I}\ (x;MP) = (n_1 - 1)_1 \left[\phi\left(-\frac{1}{2}D\right) - \phi\left(-\frac{1}{2}D_{(i)}\right) \right]^2 \tag{8}$$

onde:

D : é a raiz quadrada da Distância de Mahalanobis com a amostra total, e

$D_{(i)}$: é a raiz quadrada da Distância de Mahalanobis omitindo a i -ésima observação.

Supondo-se que as estimativas dos vetores de médias omitindo a *i*-ésima observação do grupo k são, $\bar{x}_{(i)}^{(k)}$, a estimativa da função discriminante linear é:

$$Y = \left(\bar{x}_{(i)}^{(1)} - \bar{x}^{(2)} \right)' S^{-1} x$$

$$, \tag{9}$$

$$Y = \hat{\alpha}'_{(i)}\, x$$

onde: $\hat{\alpha}_{(i)} = \left(\bar{x}_{(i)}^{(1)} - \bar{x}^{(2)} \right)' S^{-1}$ são os coeficientes da função discriminante linear, quando se omitiu a *i*-ésima observação do Grupo 1. Nesse caso, a regra de classificação, omitindo a *i*-ésima observação é definido como:

$$\begin{aligned} &\text{Alocar } x \text{ ao grupo } G_1 \text{ quando } \hat{\alpha}'_{(i)} \left[x - \frac{1}{2}\left(\bar{x}_{(i)}^{(1)} + \bar{x}^{(2)} \right) \right] > 0 \\ &\qquad \text{caso contrário, alocar ao } G_2 . \end{aligned} \tag{10}$$

Fung (1992) propôs a seguinte medida de influência para avaliar o efeito da i-ésima observação sobre a probabilidade de erro de classificação:

$$DMP_i = \left[\frac{1}{2}\left(P_{(i)}^{(1)} + P_{(i)}^{(2)} \right) \right] - \left[\phi\left(-\frac{1}{2}D \right) \right] \tag{11}$$

onde:

$$P_{(i)}^{(1)} = \phi \left[\frac{-\hat{\alpha}'_{(i)} \left(\bar{x}^{(1)} - \bar{x}^{(2)} \right) - \hat{\alpha}'_{(i)} \left(\bar{x}^{(1)} - \bar{x}_{(i)}^{(1)} \right)}{2G} \right], \text{sendo } G^2 = \hat{\alpha}_{(i)} S \hat{\alpha}_{(i)} .$$

2.4 Medida de influência para a probabilidade do erro de classificação com a aproximação de Taylor

A proposta de Fung (1992) é uma medida alternativa á equação (11) considerando a aproximação de segunda ordem do polinômio de Taylor, em torno de $-\frac{1}{2}D$, dessa forma tem-se a medida DMP_i para a *i*-ésima observação:

$$DMP_i \cong \frac{\phi\left(-\frac{1}{2}D\right)}{4D(n_1-1)^2}\left[\left(1-w_k\hat{\psi}_i\right)^2\left(d_i^2-\frac{\hat{\psi}_i^2}{D^2}\right)+\frac{1}{4}\hat{\psi}_i^2\right] \quad (12)$$

onde: $d_i^2 = \left(x_i^{(k)} - \bar{x}^{(k)}\right)' S \left(x_i^{(k)} - \bar{x}^{(k)}\right)$

$\hat{\psi}_i = \hat{\alpha}' \left(x_i^{(k)} - \bar{x}^{(k)}\right)'$

$x_i^{(k)}$ é a i-ésima observação do grupo k, onde $k = 1, 2$.

2.5 Medida de influência para os escores da função discriminante

Fung (1995), propôs uma medida para os escores da função discriminante de Fisher, seguindo a metodologia proposta por Cook e Weisberg (1982), com base na quantificação do efeito da omissão de uma observação no vetor de parâmetros, considerando-se a relação de equivalência entre os coeficientes da função discriminante de Fisher e os coeficientes do modelo de regressão linear múltipla de Johnson3 (1987), onde:

$$\hat{\alpha}' x - \frac{1}{2}\hat{\alpha}'\left(\bar{x}^{(1)} + \bar{x}^{(2)}\right)$$

são os escores da função discriminante de Fisher, representado como $\beta' x$, onde :

$\beta' = \left[-\frac{1}{2}\hat{\alpha}'\left(\bar{x}^{(1)} + \bar{x}^{(2)}\right), \hat{\alpha}'\right]$, $x' = \left[1 \ , \ x'\right]$, $\beta_{(i)}$: é o vetor β, em que se omite a i-ésima observação do grupo 1.

O efeito da i - ésima observação é avaliado através da diferença dos escores da função discriminante, com e sem esta observação, ou seja, a diferença:

$$\beta' x - \beta'_{(i)} x .$$

Fung (1995) fez a proposta da seguinte medida:

$$E2 = t.\hat{\beta}_1^2 + (1-t).\hat{\beta}_2^2 + V \quad (13)$$

onde: $t = \frac{n_1}{n}$

[3] Referência em Fung (1992)

$$\beta_1 = \frac{\left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)' \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)}{2} - \frac{\hat{\alpha}'_{(i)} \left(\bar{x}^{(1)} - \bar{x}^{(1)}_{(i)}\right)}{2}$$

$$\hat{\beta}_2 = \frac{-\left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)' \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)}{2} - \frac{\hat{\alpha}_{(i)} \left(\bar{x}^{(1)} - x^{(1)}_{(i)}\right)}{2}$$

$$V = \left(\hat{\alpha} - \hat{\alpha}_{(i)}\right)' S \left(\hat{\alpha} - \hat{\alpha}_{(i)}\right).$$

3 Materiais e métodos

Para o presente trabalho foram utilizados os dados de 100 amostras de *Minthostachys tomentosa* que foram coletadas na Provincia de Cajatambo do Departamento de Lima, no ano 2005 (ver Figura 1) .

Figura 1 - Folhas de Minthostachys.

A amostragem foi feita entre janeiro e junho de 2005, a uma altitude de 2800 a 3600 m nas comunidades indígenas de Rancas e Cruzjirca localizadas em Cajatambo. As coletas dos dados ocorreram durante a estação chuvosa, janeiro-março e durante alguns meses de seca, abril-junho. As amostras foram inventariadas segundo o catálogo das Angiospermas e Gimnospermas do Peru (Brako; Zarucchi, 1993), que mostra a seguinte distribuição do gênero Minthostachys no Peru. (Figura 1) no Laboratorio de Etnobotânica e Botânica Econâmica do Museu de História Natural da UNMSM, em cujas instalações foram feitos os estudos de taxonomia e sistemática das amostras pelo sistema de classificação de Cronquist.

Tabela 1 - Distribuição das especies de *Minthostachys* no Perú (Brako;Zarucchi, 1993)

Espécies	Altitude (msnm)	Localização Geográfica
Minthostachys glabrescens (*Bentham*)	2500 - 4000	Apurimac, Cajamarca, Cuzco, Junín.
Minthostachys mollis (*Grisebach*)	500 - 3500	Amazonas, Arequipa, Cajamarca, Cuzco, Huanuco, Junín, Lima, La Libertad, Piura.
Minthostachys setosa (*Briquet*) *Epling*	1000 - 1500	Puno
Minthostachys tomentosa (*Bentham*)	2000 - 3500	Amazonas, Cajamarca, Cuzco, Huanuco, Junín, Lima, La Libertad, Ancash.
Minthostachys andina (*Britton*) *Epling*	2000 - 2500	Cuzco
Minthostachys mandoniana (*Briquet*) *Epling*	1000 - 1500	Ayacucho
Minthostachys salicifolia Epling	2500 - 3000	Ayacucho

As amostras foram divididas em dois grupos conforme a classificaçao obtida através da análise de componentes principais no estudo realizado por Gomez et al, (2008) sendo que o grupo 1 reúne 51 amostras de *Minthostachys* com pubescência abundante (pubescentes) que são representadas de 1 a 51 e o grupo 2, reúne 49 amostras de *Minthostachys* com pouca pubescência (não pubescentes) representadas de 52 a100.

Para realizar a análise discriminante considerou- se as seguintes variáveis: X_1=Comprimento do pecíolo (cm); X_2=Comprimento da folha (cm) e X_3=Largura da folha (cm), conforme descreve a Figura 2.

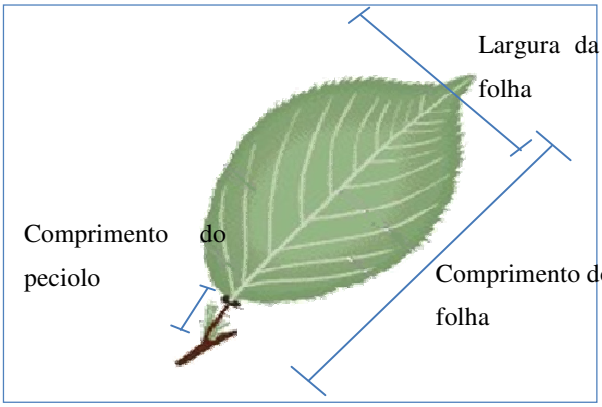

Figura 2 - Partes da folha de *Minthostachys.*

A análise dos dados foi processada com o uso software estatístico SPSS Statistical Package for the Social Sciences, versão 17 e o Matlab - Versão 7.1, foi adotado o nivel de 5% de significância.

4 Resultados e discussão

A Tabela 2 apresenta as estatísticas descritivas para cada uma das variáveis univariadas e a análise de variância simples para cada uma das variáveis univariadas. Os valores da estatística F e as probabilidades associadas (p valor) permitem rejeitar a hipótese de igualdade de médias de cada uma das variáveis univariadas ao nível de significância de 0,05.

Tabela 2 - Média aritmética e desvio padrão das variaveis e resultado do Teste de igualdade de médias para cada variável

Variáveis (cm)	*Minthostachys*		F(1,98)	P valor
	com abundante pubescência	com pouca pubescência		
Comprimento do pecíolo	0,475±0,125	1,202±0,507	105,35	< 0,001
Comprimento da folha	3,243±0,551	3,671±0,769	10,30	< 0,001
Largura da folha	1,726±0,349	2,167±0,524	23,56	< 0,001

A Tabela 3 mostra o valor de Lambda de Wilks, a relação entre a soma dos quadrados intra grupos e a soma dos quadrados total, o teste compara os vetores de médias multivariados ou as médias das funções discriminantes nos dois grupos e, é transformado em uma variável que assintoticamente tem distribuição qui-quadrado $\left(\chi^2 = -\left(n-k-\frac{1}{2}(p-k+2)\right)\ln(\Lambda)\right)$. Postulou-se, a hipótese de que as *Minthostachys* com abundante pubescência e pouca pubescência, vir de populações com vetores de médias significativamente diferentes, ou que as médias das funções discriminantes são significativamente diferentes. Observando-se o valor da estatística Lambda de Wilks (0,457) ou o valor de qui-quadrado, apresentados na Tabela 3, se rejeita a hipótese de igualdade de vetores média entre as *Minthostachys* de pubescência abundante e as *Minthostachys* de pouca pubescência. Ou seja, as diferenças dos vetores de médias são estatísticamente significativas ao nível de significância de 0,05.

Tabela 3 - Teste das funções discriminantes ou de igualdade de vetores de médias multivarida

Teste da Função	Wilks' Lambda	Qui Quadrado	df	P valor
1	0,457	75,468	3	<0,001

A seguir apresenta-se os resultados, os vetores de médias e as matrizes de covariâncias segundo a notação da análise discriminante :

$$\bar{x}^{(1)} = \begin{bmatrix} 0{,}475 \\ 3{,}243 \\ 1{,}736 \end{bmatrix}, \qquad \bar{x}^{(2)} = \begin{bmatrix} 1{,}202 \\ 3{,}671 \\ 2{,}167 \end{bmatrix},$$

$$S_1 = \begin{bmatrix} 0{,}016 & 0{,}017 & 0{,}015 \\ 0{,}017 & 0{,}304 & 0{,}061 \\ 0{,}015 & 0{,}061 & 0{,}122 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0{,}239 & 0{,}259 & 0{,}186 \\ 0{,}259 & 0{,}592 & 0{,}344 \\ 0{,}186 & 0{,}344 & 0{,}275 \end{bmatrix},$$

$$S = \begin{bmatrix} 0{,}1252 & 0{,}1352 & 0{,}0989 \\ 0{,}1352 & 0{,}4449 & 0{,}1998 \\ 0{,}0989 & 0{,}1998 & 0{,}1971 \end{bmatrix},$$

onde; $S = \dfrac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$

O vetor de coeficientes da função discriminante linear de Fisher,

$\hat{\alpha} = S^{-1}\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) = \begin{bmatrix} -7{,}2491 \\ 1{,}0795 \\ 0{,}3566 \end{bmatrix}$, onde segundo a equação (1),

$Y = -7{,}25X_1 + 1{,}08X_2 + 0{,}36X_3$, é a função discriminante linear.

Da equação (3) temos o valor da Distância de Mahalanobis igual a 4,65 e a probabilidade do erro de classificação de acordo com a equação (4) teve o valor de 0,1405 conforme mostra a Tabela 4.

Tabela 4 - Valores das estatísticas com todas as observações

Vetor dos coeficientes	Distância de Mahalanobis	Prob. de erro de classificação	% de obs. classificadas erroneamente	observações classificadas erroneamente
$\begin{bmatrix} -7{,}2491 \\ 1{,}0795 \\ 0{,}3566 \end{bmatrix}$	4.65	0,1405	9%	53 62 74 75 90 91 98 99 100.

Cada observação ou cada uma das 100 amostras de Minthostachys foram avaliadas na equação (6) dando origem aos escores discriminantes. As maestros Minthostachys com pouca pubescência (2-Grupo 2), com os códigos 53, 62, 74, 75, 90, 91, 98, 99 e 100, foram classificadas erroneamente como Minthostachys com pubescência abundante (2-Grupo 1), representando 9% das amostras.

Na Tabela 5 apresentamos, parte dos resultados da análise discriminante. O número da amostra, o grupo verdadeiro ao qual pertence o indivíduo, o grupo ao qual os

indivíduos foram atribuídos de acordo com a equação de classificação (2) e as pontuações dos escores discriminantes para cada *Minthostachys*.

Tabela 5 - Estatísticas da classificação

No. da amostra	Grupo Verdadeiro	Classificado ao grupo	Escores discriminantes
1	1	1	-1,16
2	1	1	-1,43
3	1	1	-0,96
4	1	1	-0,81
5	1	1	-1,25
47	1	1	-1,19
48	1	1	-1,03
49	1	1	-1,82
50	1	1	-1,67
51	1	1	-1,35
52	2	2	1,93
53	2	1	-0,78
54	2	2	0,33
55	2	2	1,88
62	2	1	-0,31
74	2	1	-0,54
75	2	1	-0,16
90	2	1	-1,20
91	2	1	-0,45
98	2	1	-0,16
99	2	1	-0,10
100	2	1	-1,04

A Figura 3 mostra os escores da função discriminante linear de Fisher para cada uma das 100 observações.

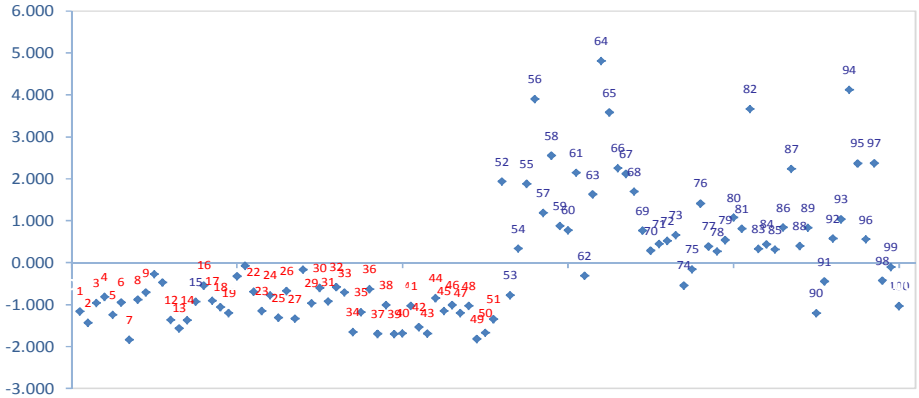

Figura 3 - Escores da função discriminante linear.

Removendo uma observação ou seja com 99 observações cada vez, encontraram-se o vetor de coeficientes da função discriminante linear de Fisher, a Distância de Mahalanobis e as observações classificadas erroneamente. Este procedimento é repetido, gerando para cada uma das repetiçoes os coeficientes de função discriminante, a Distância de Mahalanobis, a probabilidade do erro de classificação e as observações classificadas erroneamente. Entre todos os casos, as maiores mudanças nas estatísticas relacionadas com a análise discriminante foram para as observações 64, 90 e 100, conforme mostra a Tabela 6.

Tabela 6 - Estimativas das medidas relacionadas com a análise discriminante com a omissão das observaçãoes: 90, 100, 64

Medidas	Omitindo a observação		
	90	100	64
Vetor de coeficientes da função discriminante linear	$\begin{bmatrix} -7,5909 \\ 0,9011 \\ 0,2986 \end{bmatrix}$	$\begin{bmatrix} -7,5371 \\ 0,9712 \\ 0,2710 \end{bmatrix}$	$\begin{bmatrix} -8,1096 \\ 1,2666 \\ 0,3888 \end{bmatrix}$
Distância de Mahalanobis	5,1191	5,0498	4,9960
Probabilidade de erro de classificação	0,1290	0,1306	0,1319
Porcentagem de observações classificadas erroneamente	7,1%	8,1%	9,1%
observações classificadas erroneamente	53 62 74 75 97 98 99	53. …62 74 75 90 91 98 99	53 62 74….75…. 89 …90…. 97….98…. 99.

A Tabela 7 apresenta os valores das observações identificadas como potencialmente influentes de acordo com a medida de influência avaliada. Os valores mais altos para a medida de influência da equação (7) correspondem as observações 53, 64, 90 e 100; para a medida de influência da equação (8) correspondem às observações 53, 90, 100 (positivo), 64 e 94 (negativo); para a medida de influência da equação (11) correspondem às observações 21, 62, 90 e 100; para a medida de influência da equação (12) correspondem às observações 21, 62, 90 e 100 e para a medida de influência da equação (13) os valores maiores correspondem as observações 64, 90 e 100.

Tabela 7 - Medidas e observações identificadas como potencialmente influentes segundo as diferentes medidas de influência (MI)

Medida de influência	Observações						
	21	53	62	64	90	94	100
Equação(7)	8,467	18,459	12,77	18,58	24,509	10,565	21,99
Equação(8)	0,112	0,208	0,158	-0,422	0,256	-0,342	0,237
Equação(11)	0,324	0,226	0,341	0,216	0,665	0,157	0,399
Equação(12)	0,178	0,137	0,192	0,127	0,405	0,091	0,253
Equação(13)	0,014	0,017	0,014	0,099	0,058	0,025	0,035

As Figuras 4, 5, 6 , 7 e 8, mostram as pontuações.

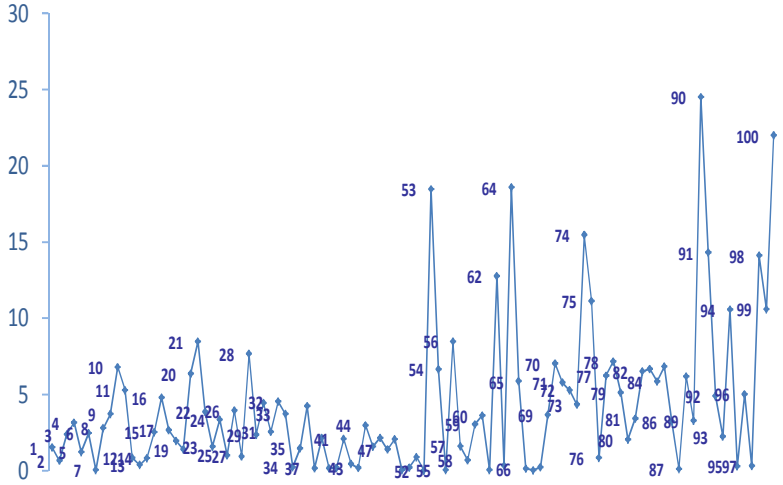

Figura 4 - Medida de influência para a Distância de Mahalanobis.

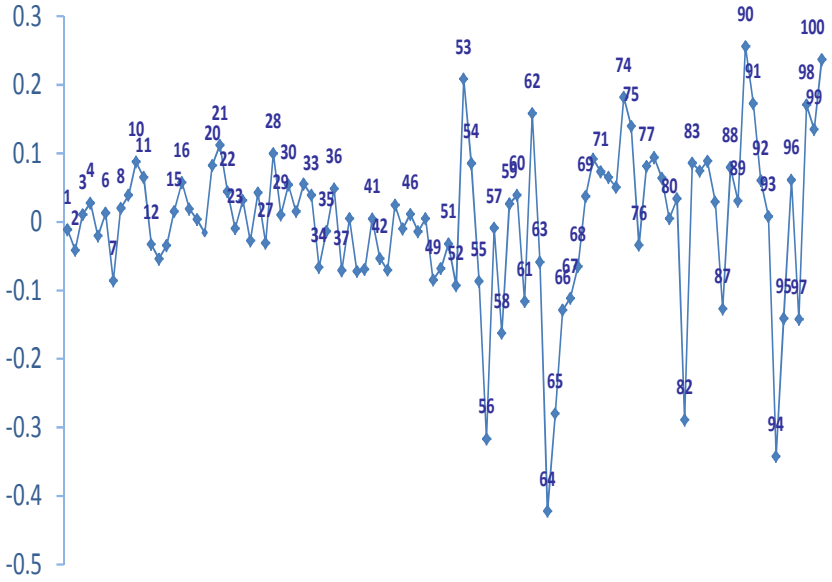

Figura 5 - Medida de influência para a probabilidade do erro de classificação.

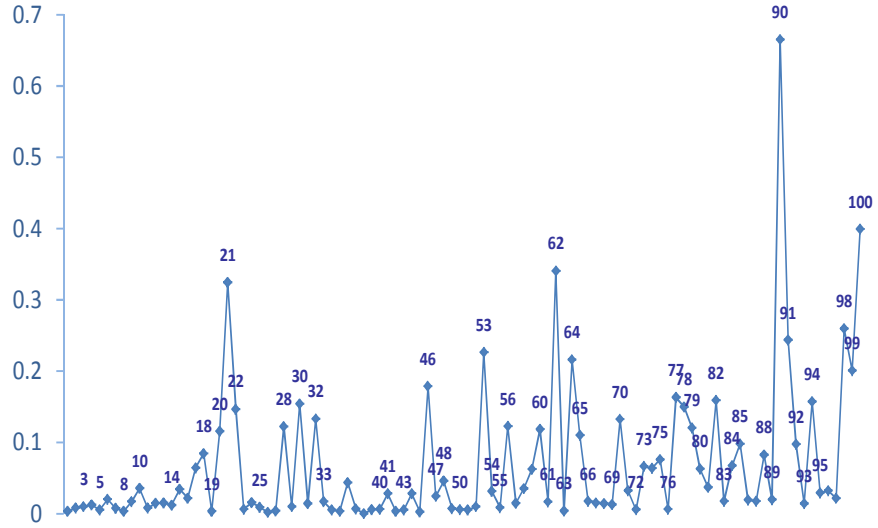

Figura 6 - Medida de influência alternativa para a probabilidade de classificação errônea

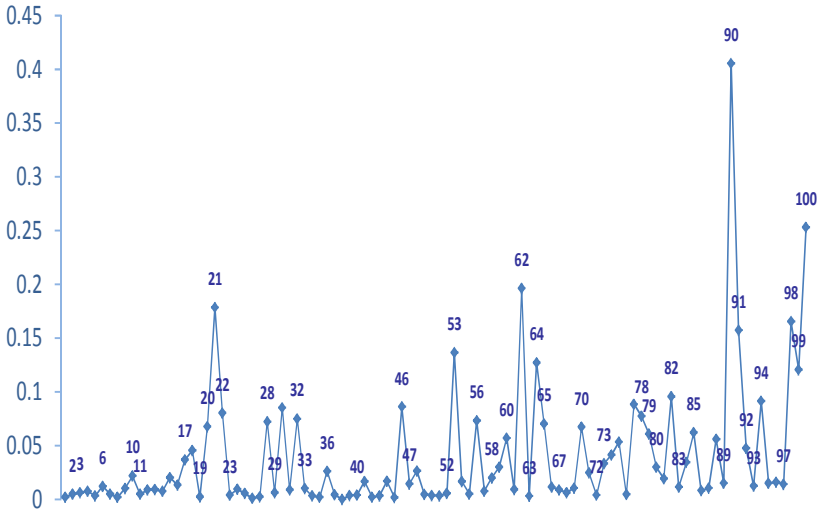

Figura 7 - Medida de influência segundo a aproximação de Taylor.

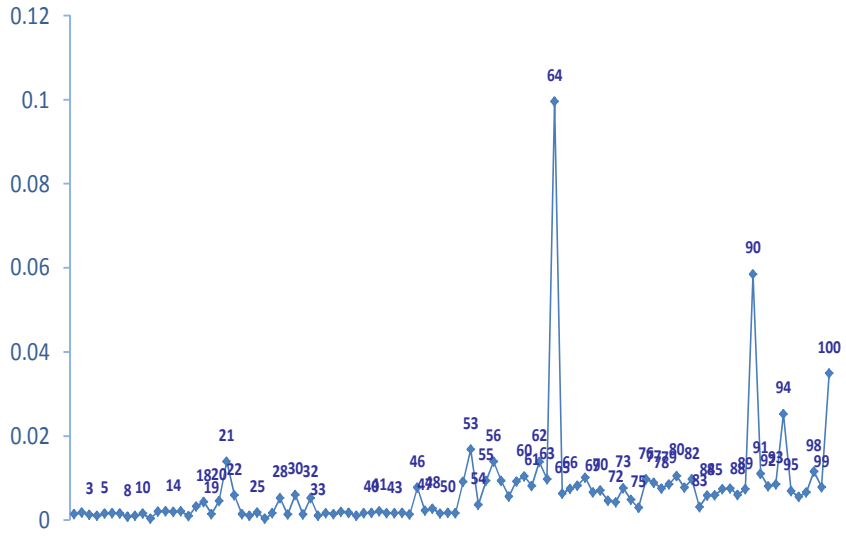

Figura 8 - Medida de influência para os escores da função discriminante linear.

Os coeficientes de correlação de Pearson obtidos entre os valores com as diferentes medidas de influência indicam uma relação muito boa, ou seja, todas as medidas coincidem em identificá-las as mesmas observações como observações potencialmente influentes. Os resultados são apresentados na Tabela 8.

Tabela 8 - Medidas de associação entre diferentes medidas

Pontuações das medidas	Coeficiente de Correlação
Pontuações das equações (7) e (11)	0,899
Pontuações das equações (7) e (12)	0,8286
Pontuações das equações (7) e (13)	0,6187
Pontuações das equações (7) e (12)	0,9963
Pontuações das equações (7) e (13)	0,6031

Conclusões

Aplicando a metodologia da análise discriminante, as *Minthostachys* com pouca pubescência, 53, 62, 74, 75, 90, 91, 98, 99 e 100 ou seja 9% das amostras foram classificadas erroneamente como *Minthostachys* com pubescência abundante. Considerando-se todas as medidas de influência representadas nas equações (7), (8), (11), (12) e (13), as amostras 21, 53, 62, 64, 90, 94 e 100 foram identificadas como potencialmente influentes.

As maiores mudanças nos valores das várias medidas envolvidas na análise discriminante, cada vez que ocorre a eliminação de uma observação, ocorreram quando foram retiradas as observações, 64, 90 e 100, cujos valores são apresentados na Tabela 6. Assim, pode-se concluir que essas observações foram influentes.

Os valores dos coeficientes de correlação entre os escores obtidos com as diferentes medidas de influência são maiores de 0,6 e há um caso com valor de 0,99, Tabela 8, indicando que há concordância muito boa entre os escores das medidas de influência.

Agradecimentos

Os autores agradecem ao Consejo Superior de Investigaciones de la Universidad Nacional Mayor de San Marcos- Perú, pelo apoio financeiro.

CAÑARI, D; GÓMEZ, D.; SOLANO, O.L.; QUINTEROS, Y.; ALBAN, J. Identification of influential observations on Minthostachys gender samples. *Rev. Bras. Biom.*, São Paulo, v29, n.3, p.493-511, 2011.

- *ABSTRACT: This paper explores the possibility of identifying influential observations in discriminant analysis framework, 100 botanical specimens of the genus* Minthostachys*, pubescent and pubescent not collected in the province of Cajatambo department of Lima. The evaluation of morphological variables in the main branch of each* Minthostachys *being studied were: length of petiole, leaf length and width of the blade. Taxonomic and systematic studies of the samples were performed at the Laboratory of Ethnobotany and Economic Botany of the Natural History Museum and the determination of the species are held in the herbarium of the San Marcos University, using the Cronquist classification system, which marked 51 plants such as non-pubescent and pubescent 49. For the full sample and removing each time one of the samples or observations, we calculated the value of the Mahalanobis Distance, the probability of misclassification, the weightings and scores of discriminant function of Fisher (Campell, 1978; Fung, 1992 , 1995). Comparison of the values of the estimates, with and without the observation under evaluation, it was concluded that observations 64, 90 and 100 were identified as influential.*

- *KEYWORDS*: *Influence measures; influential observation; linear discriminant analysis; Gender* Minthostachys*.*

Referências

ANDERSON T. W. *An introduction to Multivariate Statistical Analysis*. 2. ed. New York: Wiley e Sons, 1984. 373p.

BECKMAN, R. J.; COOK, R. D. Outliers. *Technometrics*, v.25, n.2, p.119-149, 1983.

BRAKO L.; ZARUCCHI J. *Catálogo de Angiospermas y Gimnospermas del Perú*. Missouri Botanical Garden. USA. 1993.

CAMPBELL, N. A. The Influence function as an aid in outlier detection in discriminant analysis. *Applied. Statistics*, v.27, n.3, p.251-258, 1978.

FUNG, W.K. Diagnostics in linear discriminant analysis. *Statistics and Probability Letters,* v.13, p.279–285, 1992.

FUNG, W. K. Some diagnostic measures in discriminant analysis. *J. Am. Stat. Assoc,.* v. 90, p.952-956, 1995.

GOMEZ, D. et al. Determinación de patrones de variación morfológica del género *Minthostachys* en Unchos y Cajatambo mediante métodos estadísticos multivariantes de reducción de datos. *Pesquimat - Revista de investigación de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos*, Lima, Perú, v.11, n.1, p.53-66, 2008.

HAMPEL, F. R. Influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* v.69, p.383-393, 1974.

INEI - *Censo de Población y Vivienda*. Instituto Nacional de Estadística.2005.

MANLY, B. *Multivariate statistical methods.* 3.ed. New York: *Chapman* & Hall/CRC, 2005. 214p.

MUÑOZ, J.M; MORENO, J.L; GÓMEZ, T; ENGUIX, A. El sesgo condicionado en el análisis de influencia: una Revisión. Facultad de Matemática, Universidad de Sevilla. *Questiió,* v. 25, n. 2, p. 263-284, 2001.

SCHMID, T.; LEBUHN, A. N. Ethnobotany, biochemistry and pharmacology of *Minthostachys*(Lamiaceae). *J. Ethnopharmacol*. v.118, n.3, p.343-353, 2008.

Recebido em 01.04.2011

Aprovado após revisão em 20.01.2012

ANEXO- Banco de dados utilizado no estudo.

Comprimento do pecíolo	Largura da folha	Comprimento da folha	Grupo	Comprimento do pecíolo	Largura da folha	Comprimento da folha	Grupo
0.50	3.60	1.80	I	0.40	3.40	1.50	I
0.30	3.00	1.20	I	1.50	3.90	2.50	II
0.50	3.30	1.50	I	0.50	3.00	1.30	II
0.50	2.90	1.80	I	1.00	3.80	2.30	II
0.50	3.70	2.00	I	1.40	3.50	2.00	II
0.50	3.30	1.40	I	2.10	4.00	2.50	II
0.30	3.70	1.50	I	1.20	3.60	1.80	II
0.50	3.10	1.60	I	1.60	3.50	2.00	II
0.50	2.80	1.50	I	1.30	4.60	2.70	II
0.70	3.10	2.00	I	1.20	4.40	1.90	II
0.70	3.50	2.00	I	1.40	3.00	1.90	II
0.40	3.20	2.20	I	0.60	2.80	1.10	II
0.40	3.90	1.30	I	1.30	3.30	2.10	II
0.40	3.50	1.35	I	2.40	4.10	2.80	II
0.50	3.30	1.30	I	2.10	4.60	2.60	II
0.60	3.00	1.90	I	1.70	4.60	2.50	II
0.70	4.40	1.90	I	1.70	4.60	3.30	II
0.70	4.70	1.90	I	1.50	4.50	2.10	II
0.40	3.00	1.80	I	1.00	3.10	1.80	II
0.60	2.80	1.20	I	1.10	4.50	2.50	II
0.60	2.10	1.80	I	0.90	3.00	2.00	II
0.70	3.70	2.70	I	1.00	3.50	2.10	II
0.40	2.90	1.80	I	1.20	4.40	2.60	II
0.60	3.40	2.10	I	0.60	3.10	1.60	II
0.50	3.90	1.80	I	0.70	2.90	1.90	II
0.60	3.30	1.80	I	1.30	3.60	2.50	II
0.30	2.70	1.50	I	1.10	4.20	2.80	II
0.60	2.40	1.45	I	1.10	4.50	2.60	II
0.60	3.80	2.05	I	1.20	4.60	2.70	II
0.40	1.90	1.50	I	1.40	4.80	2.90	II
0.60	3.80	1.80	I	1.20	4.10	2.60	II
0.40	2.00	1.10	I	2.10	4.50	2.40	II
0.50	2.80	1.50	I	0.90	3.30	1.80	II
0.35	3.60	1.70	I	1.10	4.20	2.50	II
0.50	3.60	1.90	I	0.80	2.80	1.40	II
0.50	2.70	1.30	I	1.20	4.10	2.40	II
0.40	3.90	2.10	I	1.60	4.00	2.40	II
0.50	3.30	1.80	I	0.80	2.60	1.50	II
0.30	3.30	1.90	I	1.00	3.00	1.70	II
0.40	3.90	2.00	I	0.20	1.90	1.10	II
0.50	3.20	2.20	I	0.50	2.30	1.40	II
0.30	3.10	1.50	I	1.20	4.60	2.50	II
0.20	2.80	1.30	I	1.10	3.10	2.20	II
0.60	3.70	1.60	I	2.20	4.10	2.90	II
0.40	3.00	1.50	I	1.50	3.10	2.30	II
0.50	3.00	2.70	I	1.10	4.00	2.40	II
0.40	2.90	2.10	I	1.70	4.40	2.40	II
0.40	2.60	2.00	I	0.50	2.30	1.30	II
0.30	3.50	2.00	I	0.80	3.30	2.40	II
0.30	3.40	1.40	I	0.30	2.20	1.20	II