Sumário Questão 2

1 Introdução

Os dados foram obtidos do arquivo moscas.txt, extraído do site http://www.ime.unicamp.br/~cnaber/Moscas.txt. Este se refere a sete variáveis medidas em duas espécies das moscas chamadas bitting fly (Leptoconops carteri e Leptoconops torrens), sendo elas espécie (0 - Leptoconops torrens e 1- Leptoconops carteri), comprimento da asa, largura da asa, comprimento do terceiro palpo, largura do terceiro palpo, comprimento do quarto palpo, comprimento do 12º segmento da antena e comprimento do 13º segmento da antena. Para ser mais eficiente, essas variáveis foram renomeadas como sendo: Especie (0 - torrens e 1- carteri), C.Asa, L.Asa, C3p, L3p, C4p, C12a e C13a, respectivamente.

Estas duas espécies são tão semelhantes (Johson e Wichern (2007)) que chegaram a ser consideradas, pelos pesquisadores, como uma uníca espécie. Sendo assim, objetivo do estudo é comparar as duas espécies de moscas em relação as variáveis citadas acima para saber se há diferença entre esses dois grupos. Neste relatório empregaremos a metodologia de componentes principais a fim de construir, através de transformações lineares (Azevedo 2017), variáveis não correlacionadas que retenham a maior parte da estrutura de variabilidade e correlação das variáveis originais e utiliza-las para identificar a existência ou não de diferença.

2 Análise Descritiva

A Figura 1 apresenta os autovalores das sete componentes e indica que a partir da componente 3 a variância de cada componente é parecida, ou seja, as componentes 4, 5, 6 e 7, contribuem muito pouco para o percentual de variância explicada acumulada. A tabela 1 mostra os valores dos desvios padrões, PVE (proporção da variância explicada) e PVEA (proporção da variância explicada acumulada) das componentes e é possível notar que 77% da variabilidade dos dados originais é explicada pelas 3 primeiras componentes, por este motivo apenas estas 3 setão consideradas a fim de análise.

Tabela 1: Desvios padrão, proporção da variância explicada (PVE) e proporção da variância explicada acumulada (PVEA) das componentes principais

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
Desvio Padrão	1,711	1,242	0,948	0,773	0,721	0,579	0,418
PVE	0,418	0,220	0,128	0,085	0,074	0,048	0,025
PVEA	0,418	0,639	0,767	0,853	0,927	0,975	1,000

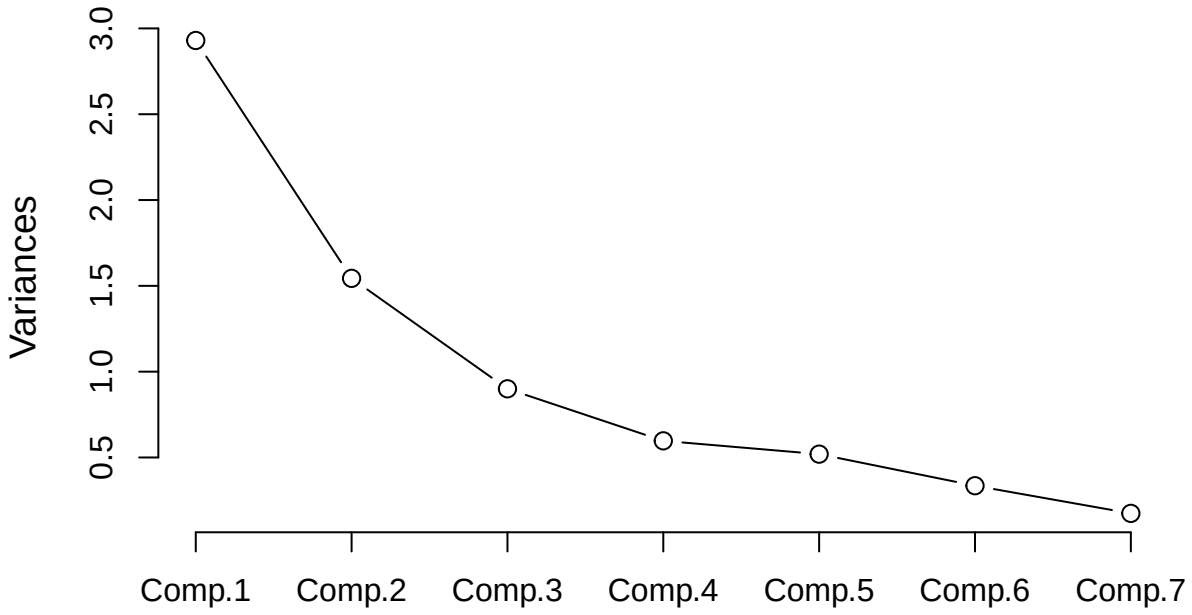

Figura 1: Gráfico de autovalores das componentes principais

A tabela 2, por sua vez, apresenta os coeficientes das três primeiras componentes principais e correlações com cada variável (entre parênteses) e através dela concluí-se que todas as variáveis estão bem representadas e correlacionadas com pelo menos uma das componentes consideradas. Pode-se, também, utilizar as informações da tabela para fazer uma interpretação das componentes. Neste casso diz-se que a componente 1 é um escore ponderado entre todas as variáveis, que a componente 2 é um contraste entre C12a e C13a e as demais variáveis e que a terceira componente é um contrate entre C3p e C4p e L.Asa e L3p.

Tabela 2: Coeficientes das três primeiras componentes principais e correlações com cada variável(entre parênteses)

Variável	Componente 1	Componente 2	Componente 3
C.Asa	-0,49 (-0,84)	-0,08 (-0,10)	0,09 (0,08)
L.Asa	-0,42 (-0,72)	-0,18 (-0,22)	0,30 (-0,28)
C3p	-0,32 (-0,54)	-0,30 (-0,37)	-0,65 (0,61)
L3p	-0,32 (-0,55)	-0,21 (-0,26)	0,67 (-0,64)
C4p	-0,37 (0,64)	-0,36 (-0,44)	-0,15 (0,15)
C12a	-0,35(-0,60)	0,58 (-0,72)	0,04 (0,04)
C13a	-0,34 (-0,58)	0,60 (-0,75)	0,07 (0,07)

Na figura 2 tem-se os gráficos de dispersão entre as componentes, por grupo. Nota-se que os valores referentes à espécie torrens tendem a se manter em torno de zero enquanto a espécie carteri não segue um padrão especifico. Além disso observa-se que, em geral, os pontos cinzas aparecem acima dos pontos pretos indicando que o grupo torrens apresenta valores mais altos do que Carteri.

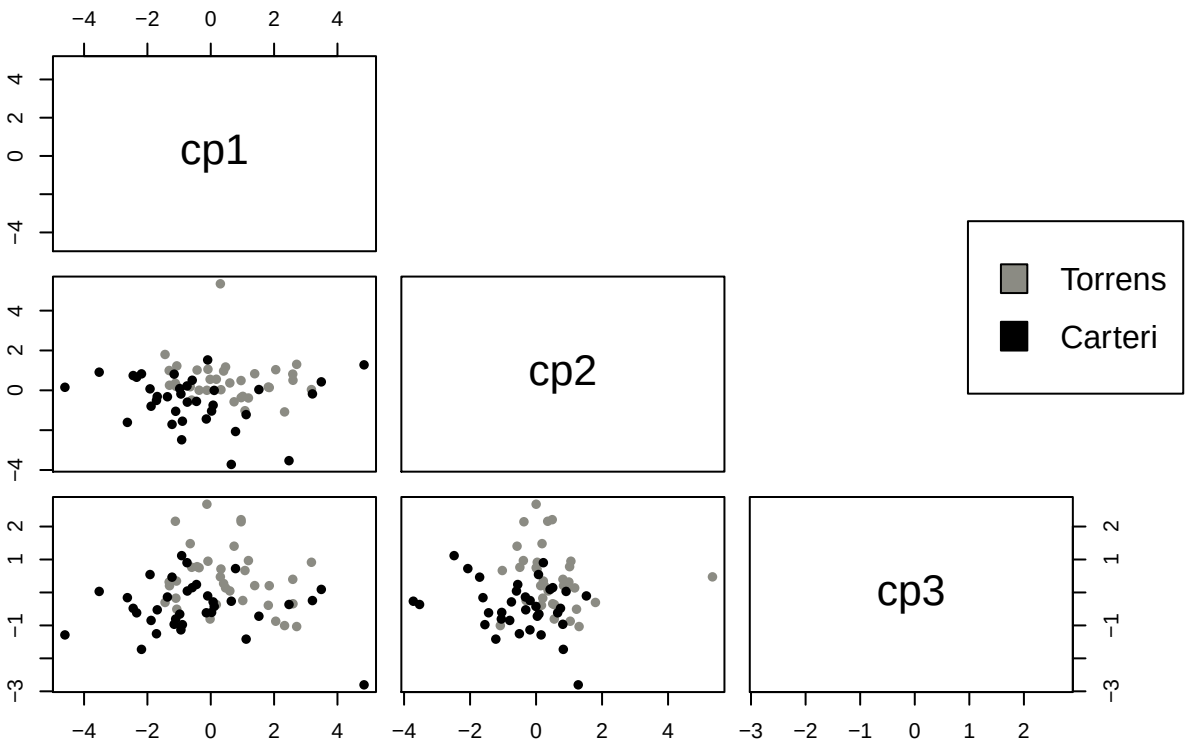

Figura 2: Gráfico de dispersão das Componentes principais 1, 2 e 3 para as espécies Torrens e Carteri

Os boxplots da figura 3, mostram que a espécie Torrens possui valores maiores que a espécie Carteri nas três componentes tendo, em todas elas, seu primeiro quartil maior que a mediana do grupo Carteri. Apesar disso, Carteri apresenta um variabilidade maior para as componetens 1 e 2 e conta com a presença de outliers em todas elas.

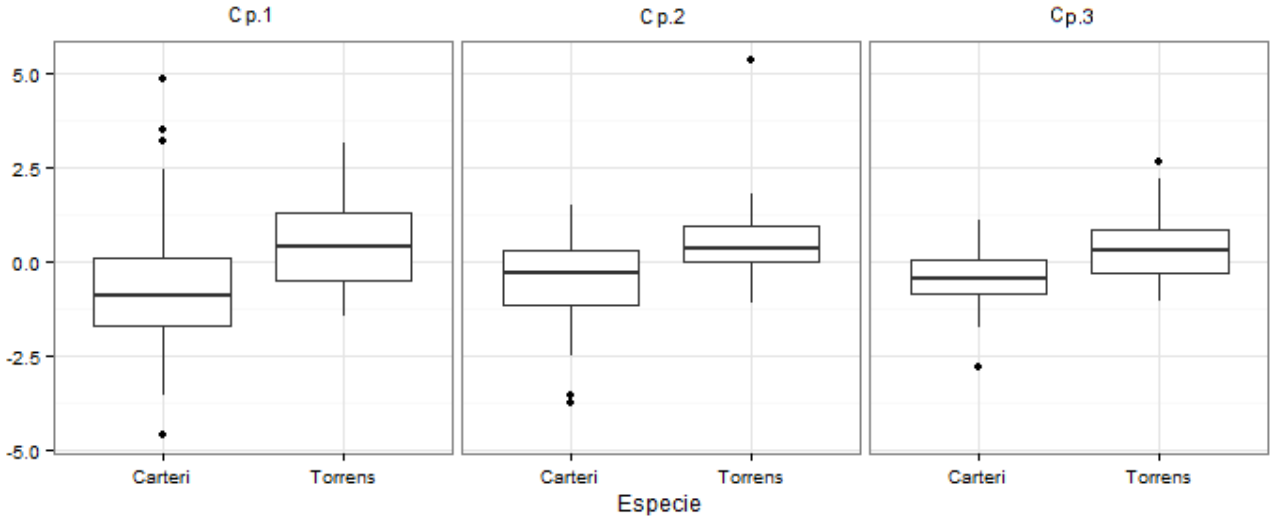

Figura 3: Boxplots para as distribuições das componentes 1, 2 e 3 por espécie

Os gráfico de quantis-quantis com envelopes das componentes por espécie encontram-se na figura 4 e através deles é possivel observar que, os gráficos para a espécie Carteri apresentam claras tendências, como por exemplo uma convexidade na componente 2 e cinco pontos fora dos envelope para a componente 1. Para o grupo Torrens, existe um ponto aberrante bastante evidente na componente dois e apesar de não haver pontos fora dos envelopes para a componente 1, a tendêndcia em formato de "**S**" é explícita e também indica uma possível fuga da normalidade. Sendo assim, a conclusão é de que, para nenhuma das compontes, em nenhuma das espécies, a suposição de normalidade parece razoavel.

Portanto, ao final da análise descritiva, pode-se conjecturar que exista diferença entre as espécies e que, em geral, Torrens apresenta valores maiores do que Carteri.Além disso também é possivel conjecturar que os dados não assumem normalidade.

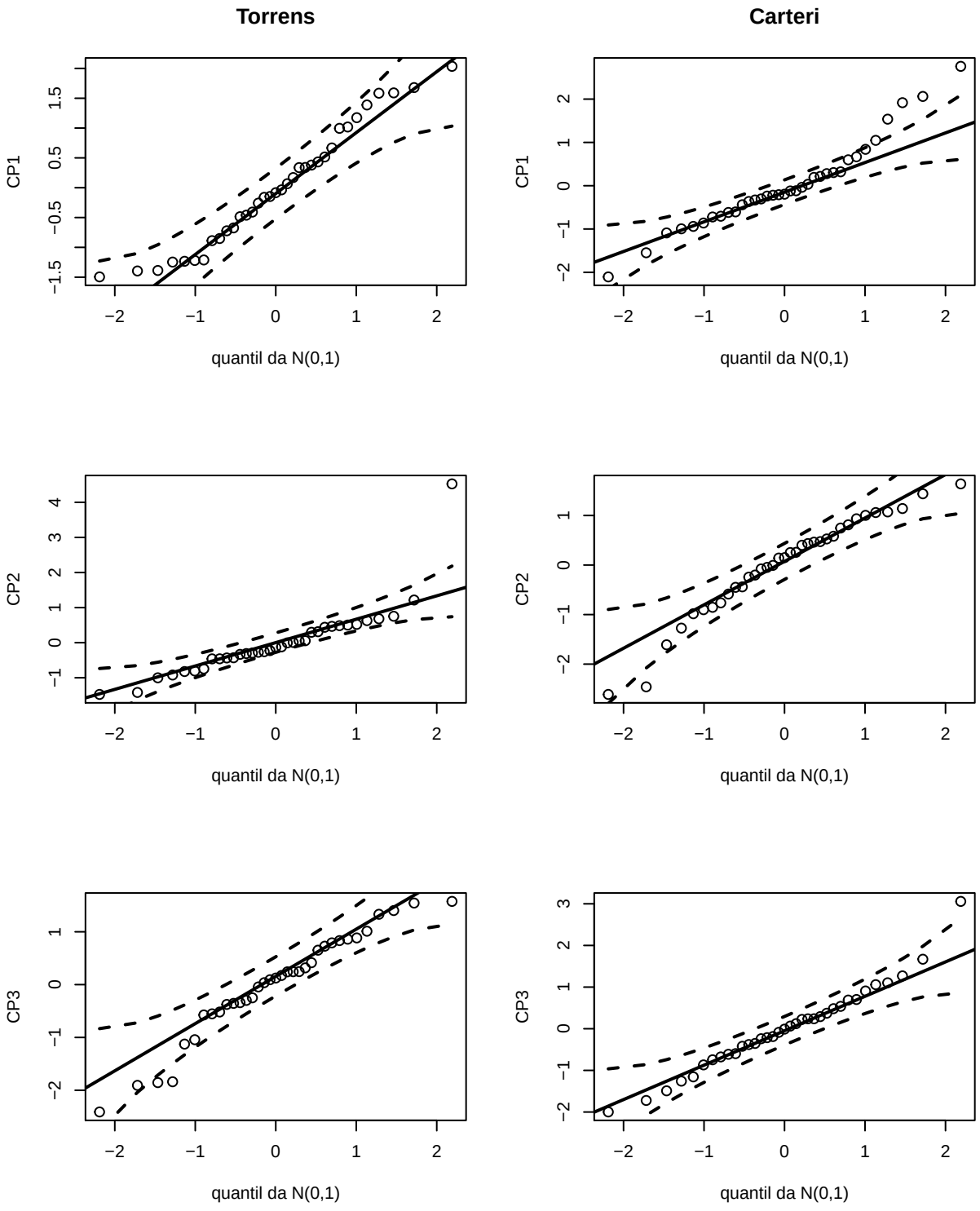

Figura 4: Gráfico de quantis-quantis com envelopes das componentes por espécie

3 Análise Inferencial

Considerando o fato já anteriormente mensionado de que a componente 1 pode ser interpretada como um escore ponderado entre todas as variáveis, foi ajustado um modelo de regressão linear normal para essa componente. O modelo considerado foi:

$$Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij} \; com \; \alpha_1 = 0 \; e \; \boldsymbol{\varepsilon_{ij}} \sim N(0, \sigma)$$

onde:

- i = 1,2 (grupo, 1-Torrens, 2-Carteri)
- j = 1, 2,..., 35 (indivíduo)

Como o parâmetro α_2 representa a diferença entre as médias do grupo Torrens e Carteri na componente 1, há interesse em testar a significancia do mesmo. Para isso, utilizou-se um teste t usual ($H_0 : \alpha_2 = 0 \; vs \; H_1 : \alpha_2 \neq 0$) e os resultados são mostrados na tabela 3.

Tabela 3: Estimativas dos parâmetros do modelo de regressão para a componente 1

Componente	Parâmetro	Estimativa	Erro Padrão	Valor F	p-valor
CP1	μ_1	0,51	0,28	0,84	0,07
	α_1	-1,03	0,39	-2,60	0,01

Como observado na tabela acima, considerando-se um teste de nível de significância igual a 5%, tem-se que o parâmetro α_2 é diferente de zero e, portanto, estatísticamente significativo. Sendo assim, conclui-se que as espécies são diferentes. Apesar disso, nota-se que os intervalos de confiança para as médias preditas pelo modelo se inteseptam (tabela 4), o que não era esperado uma vez que a igualdade de médias foi rejeitada. A tabela 4 ainda mostra que o grupo carteri tem valores menores que Torrens.

Tabela 4: Médias preditas e intervalos de confiança por espécie

Espécie	Estimativa	Erro Padrão	IC
Torrens	0,51	0,28	[-0,04;1,07]
Carteri	-0,51	0,28	[-1,07;0.04]

Fazendo análise dos resíduos, nota-se que as suposições do modelo de normalidade e homocedasticade não são satisfeitas. A figura 5 apresentam os gráficos de diagnóstico para as componentes 1. Observando o gráfico “índice x resíduo studentizado”, tem-se 4 pontos fora dos limites de confiança e é possivel detectar que índices maiores estão associados a maior variabilidade o que indica dependência. Além disso, pelo gráfico “valores ajustados x resíduo studentizado”, as variâncias residuais não parecem constantes, indicando fuga da suposição de heterocedasticidade e a suposição de normalidade também não é adequada uma vez que o histograma apresenta uma assimetria positiva e há um padrão nos gráficos de quantil da Normal(0,1) além de algumas observações estarem fora do envelope.

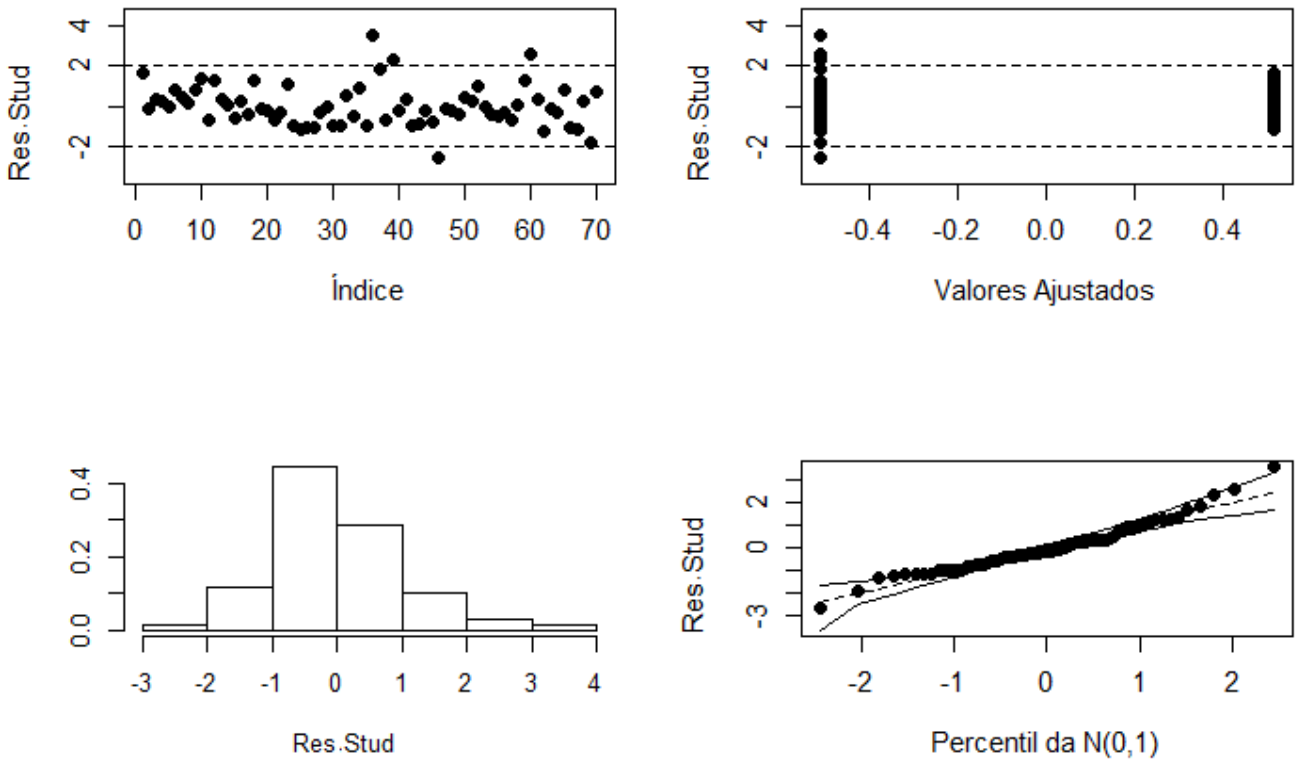

Figura 5: Gráficos de Resíduos para a componente 1

Na figura 6 apresenta-se o gráfico de biplot para a combinação das três componentes, a análise do primeiro gráfico, biplot entre as componentes 1 e 2, indica que as variáveis CA, LA, CP3, LP3 e CP4 possuem uma correlação positiva entre si, já as demais variáveis, SA12 e SA13, são intensamente correlacionadas entre si, porém não são correlacionadas com as outras variáveis.

No segundo gráfico biplot, tem-se que as variáveis SA12, SA13, CA e CP4 são correlacionadas entre si, e não são correlacionadas com as demais variáveis (CP3, LA e LP3, que não são correlacionadas entre si).

No último gráfico, observa-se que as variáveis se dividem em 3 grupos não correlacionados. O primeiro grupo inclui as variáveis SA12 e SA13, que se correlacionam. Já o segundo contém as variáveis LP3 e LA (que são correlacionadas entre si). E por fim, o terceiro grupo abrange as variáveis CP4 e CP3 que não se correlacionam.

Esta figura, também permite concluir que a espécie torrens tende a se manter abaixo da média para as variáveis C.Asa, C3p e C4p e a espécie Carteri apresenta valores mais altos para estas variáveis. Tal conclusão vai ao encontro com aquela feita na questão 1, ou seja, corrobora os resultados nela obtidos.

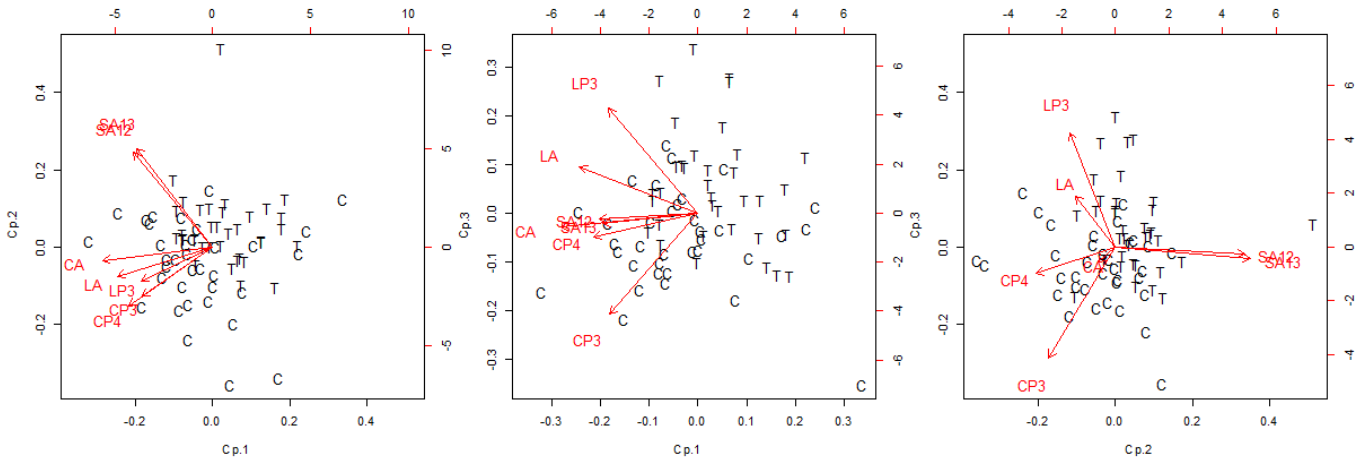

Figura 6: Biplots para a combinação das três componentes

4 Conclusão

Feita a devida ressalva de que o modelo escolhido como base para as análises nao se ajustou bem aos dados, e considerando válidos os resultados através dele obtidos, conclui-se que as espécies Torrens e Carteri apresentma diferenças em relação à componente 1. Além disso, a análise inferencial valida as conjecturas feitas após a análise descritiva e, por fim, a análise dos biplots levou a conclusões semelhantes àquelas tidas na questão 1 o que corrobora o resultado obtido.

5 Referências bibliográficas

1. AZEVEDO, C. L. N. (2017). **Notas de Aula - Métodos em Análise Multivariada**. Disponível em < http://www.ime.unicamp.br/~cnaber/Material_AM_2S_2017.htm >.
2. JOHNSON, R. A., WICHERN, D. W. (2002). **Applied Multivariate Statistical Analysis**, 5^a edição, Upper Saddle River, NJ: Prentice-Hall.
3. R CORE TEAM (2017). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Viena, Áustria. Disponível em < https://www.R-project.org/ >.