Imputação múltipla para matriz de interação genótipo e ambiente

Welinton Y. Hirai & Carlos T. S. Dias

Departamento de Ciências Exatas
Estatística e Experimentação Agronômica

Piracicaba
2018

Estrutura da apresentação

1. Introdução
2. Problemática
3. Metodologia
4. Aplicação
5. Estudos Futuros

Estrutura da apresentação

1. Introdução
2. Problemática
3. Metodologia
4. Aplicação
5. Estudos Futuros

Abordagem estatística na genética

“As metodologias estatísticas são ferramentas cruciais para análises de dados genômicos, no qual atualmente está se tornando cada vez mais útil para avaliações para variedades de espécies, e ainda oferecer com precisão descrições sobre a variação genética que ocorre entre espécies, populações, famílias e indivíduos.”

Wu, R., Ma, C., & Casella, G. (2007). Statistical genetics of quantitative traits: linkage, maps and QTL. Springer Science & Business Media.

Métodos estatísticos	univariados paramétricos	Roemer (1917) Becker e Leon (1998)	Mensura os desvios para a média genética. Variância mínima em diferentes ambientes representa estabilidade genética.
		Eberhart e Russel (1966) Tai (1971)	Estuda a estabilidade de um experimental ótimo por meio de regressão.
		Shukla (1972)	Estima os componentes de variância da IGA atribuído por cada genótipo.
	univariados não-paramétricos	Hühn (1979)	Propôs métricas utilizando estatística não paramétrica, no qual valores baixo representa alta estabilidade.
		Ketata et al. (1989)	Representa graficamente a média dos "*rankeamentos"* em relação aos ambiente, em função dos desvios de *rank* para todos os genótipos .
		Flores (1993)	Gera polígonos para cada genótipos, no qual cada desenho representa os valores médios em cada ambiente.
	multivariados	Lin (1982)	Calcula valores de similaridades para cada par de genótipos, e mensura utilizando análise de agrupamentos.
		Westcott (1987)	Utiliza escalonamento multidimensional de cada genótipo em cada ambiente.

FLORES, F.; MORENO, M. .; CUBERO, J. . A comparison of univariate and multivariate methods to analyze G×E interaction. Field Crops Research, v. 56, n. 3, p. 271–286, 1 abr. 1998.

Interação genótipo x ambiente

“Para o pesquisador, este processo que diferencia a adaptabilidade da estabilidade é necessária, dado que, um caso implica a não perda de produção em diferentes ecossistemas, enquanto o outro caso pode providenciar uma vantagem significativa em ambientes específicos.”

Wade, L., C. McLaren, L. Quintana, D. Harnpichitvitaya, S. Rajatasereekul, A. Sarawgi, A. Kumar, H. Ahmed, A. Singh, R. Rodriguez, J. Siopongco, and S. Sarkarung, 1999 Genotype by environment interactions across diverse rainfed lowland rice environments. F. Crop. Res. **64**: 35–50.

“Há casos em que as respostas dos efeitos genotípicos e ambientais não são capazes de serem descritas separadamente, levando assim a necessidade de utilizar um termo que representa conjuntamente a contribuição destes efeitos.”

Gregorius, H.-R. and G. Namkoong, 1986 Joint analysis of genotypic and environmental effects. Theor Appl Genet **72**: 413–422.

Modelo AMMI

MANDEL, J. A New Analysis of Variance Model for Non-Additive Data. **Technometrics**, v. 13, n. 1, p. 1–18, 1971.

GAUCH, H. G.; ZOBEI, R. W. Predictive and postdictive success of statistical analyses of yield trials*. **Theor Appl Genet**, v. 76, p. 1–10, 1988.

Model

A ***dditive***

M ***ain effects and***

M ***ultiplicative***

I ***nteraction***

Análise de variância (ANOVA)

Análise de Componentes Principais (ACP)

Modelo estatístico para ANOVA conjunta

$$Y_{ijk} = \mu + g_i + e_j + (ge)_{ij} + \epsilon_{ijk}$$

$$\mathbb{GE} = \begin{bmatrix} (\widehat{ge})_{11} & (\widehat{ge})_{12} & \cdots & (\widehat{ge})_{1E} \\ (\widehat{ge})_{21} & (\widehat{ge})_{22} & & (\widehat{ge})_{2E} \\ \vdots & & \ddots & \vdots \\ (\widehat{ge})_{i1} & (\widehat{ge})_{i2} & & (\widehat{ge})_{ij} \\ \vdots & & \cdots & \vdots \\ (\widehat{ge})_{G1} & (\widehat{ge})_{G2} & & (\widehat{ge})_{GE} \end{bmatrix}$$

DVS

$$(\widehat{ge})_{ij} = \sum_{k=1}^{K} \lambda_k \gamma_{ik} \delta_{jk} + \sum_{k=K+1}^{n} \lambda_k \gamma_{ik} \delta_{jk}$$

$$= \sum_{k=1}^{K} \lambda_k \gamma_{ik} \delta_{jk} + \rho_{ij}$$

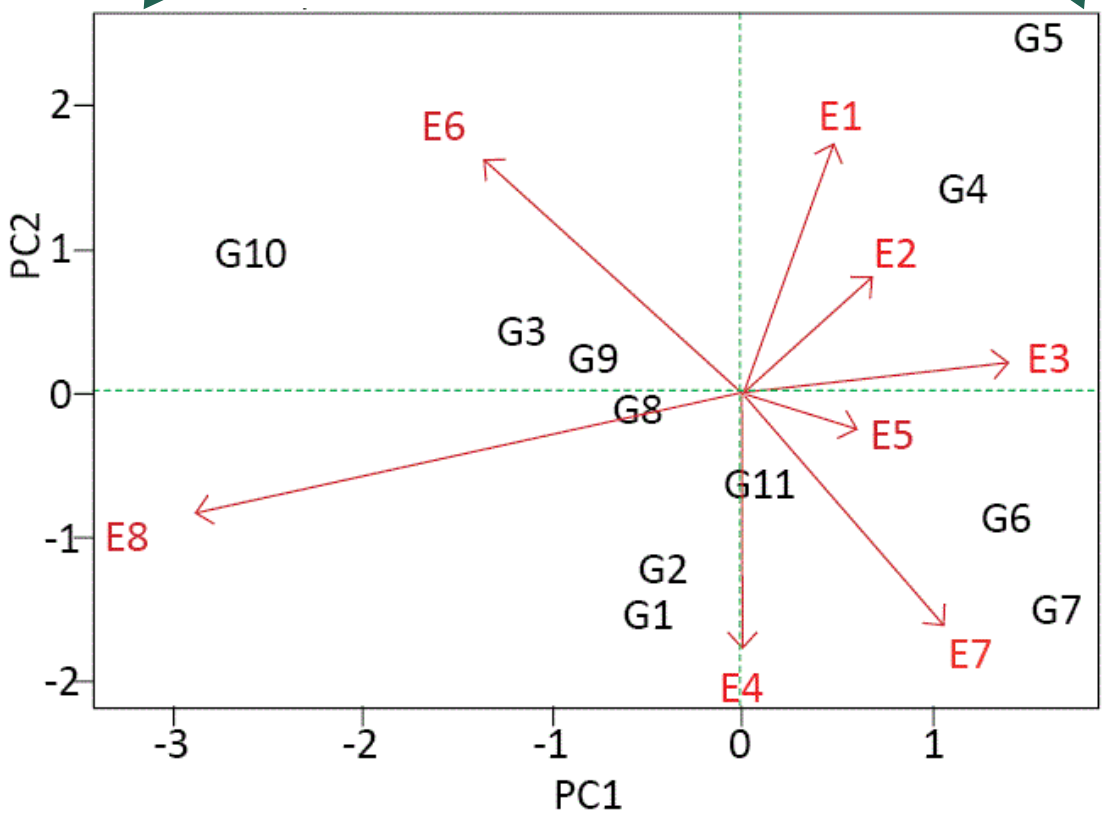
Parte determinística

Parte residual

Modelo AMMI

$$Y_{ijk} = \mu + g_i + e_j + \sum_{k=1}^{K} \lambda_k \gamma_{ik} \delta_{jk} + \rho_{ij} + \epsilon_{ijk}$$

Biplot



Estrutura da apresentação

1. Introdução
2. Problemática
3. Metodologia
4. Aplicação
5. Estudos Futuros

Ausência no conjunto de dados

$$\mathbb{GE} = \begin{bmatrix} (\widehat{ge})_{11} & (\widehat{ge})_{12} & \cdots & (\widehat{ge})_{1E} \\ (\widehat{ge})_{21} & (\widehat{ge})_{22} & & (\widehat{ge})_{2E} \\ NA & (\widehat{ge})_{32} & & (\widehat{ge})_{3E} \\ (\widehat{ge})_{41} & (\widehat{ge})_{42} & \cdots & (\widehat{ge})_{4E} \\ \vdots & & & \vdots \\ (\widehat{ge})_{i1} & (\widehat{ge})_{i2} & & NA \\ \vdots & & \cdots & \vdots \\ (\widehat{ge})_{G1} & (\widehat{ge})_{G2} & & (\widehat{ge})_{GE} \end{bmatrix}$$

Estrutura da apresentação

1. Introdução
2. Problemática
3. Metodologia
4. Aplicação
5. Estudos Futuros

Imputação Múltipla Livre de Distribuição

Missing value imputation in multivariate data using the singular value decomposition of a matrix	*Krzanowski, Wojtek Janusz*	1988
Imputação múltipla livre de distribuição utilizando a decomposição por valor singular em matriz de interação	*Bergamo, Genevile Carife*	2007
Imputação de dados em experimentos com interação genótipo por ambiente: uma aplicação a dados de algodão	*Arciniegas-Alarcón, Sergio*	2008
An alternative methodology for imputing missing data in trials with genotype-by-environment interaction: some new aspects	*Arciniegas-Alarcón, Sergio* *García-Peña, Marisol* *Krzanowski, Wojtek Janusz* *Dias, Carlos Tadeu dos Santos*	2014
Imputação de dados em experimentos multiambientais: novos algoritmos utilizando a decomposição por valores singulares	*Arciniegas-Alarcón, Sergio*	2015
Missing value imputation in multi-environment trials: Reconsidering the Krzanowski method Missing value imputation in multi-environment trials	*Arciniegas-Alarcón, Sergio* *García-Peña, Marisol* *Krzanowski, Wojtek Janusz*	2016

Teorema da Decomposição em Valores Singulares (DVS)

$$\mathbb{X}_{(n,p)} = \mathbb{U}_{(n,p)}\Phi_{(p,p)}\mathbb{V}^{T}_{(p,p)}$$

Suponha-se 3 valores ausentes, numa matriz $\mathbb{X}_{(n,p)}$

$$\left[x_{miss}^{[1]}, x_{miss}^{[2]}, x_{miss}^{[3]}\right]$$

$$\mathbb{X}_{miss} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{miss}^{[1]} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{miss}^{[2]} & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & x_{miss}^{[3]} \\ \vdots & & & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

1º processo iterativo da imputação múltipla $(impu_1)$

$$\mathbb{X}_{miss} = \begin{bmatrix} \bar{x}_1\ s_1 & \bar{x}_2\ s_2 & \bar{x}_3\ s_3 & \bar{x}_4\ s_4 & & \bar{x}_p\ s_p \\ x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{miss}^{[1]} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{miss}^{[2]} & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & x_{miss}^{[3]} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

Substitui-se os valores ausente pelas média referente à coluna

$$\mathbb{X}_{miss} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ \bar{x}_1 & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & \bar{x}_3 & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & \bar{x}_p \\ \vdots & & & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

Para **a primeira observação** que será imputada, tem-se os passos:

$$\mathbb{X}_{miss} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{miss}^{[1]} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{miss}^{[2]} & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & x_{miss}^{[3]} \\ \vdots & & & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

1ª Passo: Gerar duas matriz a partir da matriz com os valores médios

1.1 Matriz sem a linha que está o primeiro valor ausente

1.2 Matriz sem a coluna que está o primeiro valor ausente

$$\mathbb{X}_{-i_{miss}} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ \\ x_{31} & x_{32} & \bar{x}_3 & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & \bar{x}_p \\ \vdots & & & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

$$\mathbb{X}_{-j_{miss}} = \begin{bmatrix} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ & x_{32} & \bar{x}_3 & x_{34} & \cdots & x_{3p} \\ & x_{42} & x_{43} & x_{44} & \cdots & \bar{x}_p \\ \vdots & & & \vdots & & \vdots \\ & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

2ª Passo: Aplicar a padronização nas duas matrizes

$$\mathbb{X}^*_{-i_{miss}} = \begin{bmatrix} x^*_{11} & x^*_{12} & x^*_{13} & x^*_{14} & \cdots & x^*_{1p} \\ x^*_{31} & x^*_{32} & \bar{x}_3 & x^*_{34} & \cdots & x^*_{3p} \\ x^*_{41} & x^*_{42} & x^*_{43} & x^*_{44} & \cdots & \bar{x}_p \\ & \vdots & & \vdots & & \vdots \\ x^*_{n1} & x^*_{n2} & x^*_{n3} & x^*_{n4} & \cdots & x^*_{np} \end{bmatrix}$$

$$\mathbb{X}^*_{-i_{miss}} = \begin{bmatrix} & x^*_{12} & x^*_{13} & x^*_{14} & \cdots & x^*_{1p} \\ & x^*_{22} & x^*_{23} & x^*_{24} & \cdots & x^*_{2p} \\ & x^*_{32} & \bar{x}_3 & x^*_{34} & \cdots & x^*_{3p} \\ & x^*_{42} & x^*_{43} & x^*_{44} & \cdots & \bar{x}_p \\ & \vdots & & \vdots & & \vdots \\ & x^*_{n2} & x^*_{n3} & x^*_{n4} & \cdots & x^*_{np} \end{bmatrix}$$

3ª Passo: Calcular a DVS para as duas matrizes padronizadas

$$\mathbb{X}^*_{-i_{miss}} = \begin{bmatrix} x^*_{11} & x^*_{12} & x^*_{13} & x^*_{14} & \cdots & x^*_{1p} \\ & & & & & \\ x^*_{31} & x^*_{32} & \bar{x}_3 & x^*_{34} & \cdots & x^*_{3p} \\ x^*_{41} & x^*_{42} & x^*_{43} & x^*_{44} & \cdots & \bar{x}_p \\ \vdots & & & \vdots & & \vdots \\ x^*_{n1} & x^*_{n2} & x^*_{n3} & x^*_{n4} & \cdots & x^*_{np} \end{bmatrix}$$

$$\mathbb{X}^*_{(n-1,p)} = \mathbb{U}_{(n-1,p)}\Phi_{(p,p)}\mathbb{V}^T_{(p,p)}$$

$$\mathbb{X}^*_{-i_{miss}} = \mathbb{U}_{-i_{miss}}\Phi_{-i_{miss}}\mathbb{V}^T_{-i_{miss}}$$

$$\mathbb{X}^*_{-i_{miss}} = \begin{bmatrix} & x^*_{12} & x^*_{13} & x^*_{14} & \cdots & x^*_{1p} \\ & x^*_{22} & x^*_{23} & x^*_{24} & \cdots & x^*_{2p} \\ & x^*_{32} & \bar{x}_3 & x^*_{34} & \cdots & x^*_{3p} \\ & x^*_{42} & x^*_{43} & x^*_{44} & \cdots & \bar{x}_p \\ \vdots & & & \vdots & & \vdots \\ & x^*_{n2} & x^*_{n3} & x^*_{n4} & \cdots & x^*_{np} \end{bmatrix}$$

$$\mathbb{X}^*_{(n,p-1)} = \mathbb{U}_{(n,p-1)}\Phi_{(p-1,p-1)}\mathbb{V}^T_{(p-1,p-1)}$$

$$\mathbb{X}^*_{-j_{miss}} = \mathbb{U}_{-j_{miss}}\Phi_{-j_{miss}}\mathbb{V}^T_{-j_{miss}}$$

4ª Passo: Calcular a matriz com o valor imputado, utilizando as decomposições das duas matrizes

$$\mathbb{Y}^{*}_{impu1} = \mathbb{U}_{-j_{miss}} \Phi^{\left(\frac{a}{b}\right)}_{-j_{miss}} \Phi^{\left(\frac{b-a}{b}\right)}_{-i_{miss}} \mathbb{V}^{T}_{-i_{miss}}$$

Segundo Bergamo (2007) são utilizados cinco processos iterativos na imputação, com valores para $a = 8, 9, 10, 11, 12$ e $b = 20$

$$\mathbb{Y}^{*}_{(n,p)} = \mathbb{U}_{(n,p-1)} \Phi^{\left(\frac{a}{b}\right)}_{(p-1,p-1)} \Phi^{\left(\frac{b-a}{b}\right)}_{(p,p)} \mathbb{V}^{T}_{(p,p)}$$

Desta forma, é descartado o último autovalor

Em por consequência, é retirada a última coluna

5ª Passo: “Despadronizar” a matriz calculada utilizando os valores de média e desvio-padrão encontradas no início

$$\mathbb{Y}_{impu_1} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & \cdots & y_{1p} \\ y_{impu_1}^{[1]} & y_{22} & y_{23} & y_{24} & \cdots & y_{2p} \\ y_{31} & y_{32} & y_{33} & y_{34} & \cdots & y_{3p} \\ y_{41} & y_{42} & y_{43} & y_{44} & \cdots & y_{4p} \\ \vdots & & & \vdots & & \vdots \\ y_{n1} & y_{n2} & y_{n3} & y_{n4} & \cdots & y_{np} \end{bmatrix}$$

Para a todas as observações que serão imputadas, tem-se os passos:

1ª Passo: Gerar duas matriz a partir da matriz com os valores médios
- 1.1 Matriz sem a linha que está o primeiro valor ausente
- 1.2 Matriz sem a coluna que está o primeiro valor ausente

2ª Passo: Aplicar a padronização nas duas matrizes

3ª Passo: Calcular a DVS para as duas matrizes padronizadas

4ª Passo: Calcular a matriz com o valor imputado, utilizando as decomposições das duas matrizes

5ª Passo: "Despadronizar" a matriz calculada utilizando os valores de média e desvio-padrão encontradas no início

$$\left[x_{miss}^{[1]}, x_{miss}^{[2]}, x_{miss}^{[3]}\right] \Rightarrow \left[y_{impu_1}^{[1]}, y_{impu_1}^{[2]}, y_{impu_1}^{[3]}\right]$$

No 1º processo iterativo da imputação múltipla $(impu_1)$

$$\mathbb{Y}_{impu_1} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ y_{impu_1}^{[1]} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & y_{impu_1}^{[2]} & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & y_{impu_1}^{[3]} \\ \vdots & & \vdots & & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

No 2º processo iterativo da imputação múltipla $(impu_2)$

$$\mathbb{Y}_{impu_1} = \begin{array}{cccccc} \bar{x}_1\ s_1 & \bar{x}_2\ s_2 & \bar{x}_3\ s_3 & \bar{x}_4\ s_4 & & \bar{x}_p\ s_p \\ \left[\begin{array}{c} x_{11} \\ y_{impu_1}^{[1]} \\ x_{31} \\ x_{41} \\ \vdots \\ x_{n1} \end{array}\right. & \begin{array}{c} x_{12} \\ x_{22} \\ x_{32} \\ x_{42} \\ \\ x_{n2} \end{array} & \begin{array}{c} x_{13} \\ x_{23} \\ y_{impu_1}^{[2]} \\ x_{43} \\ \\ x_{n3} \end{array} & \begin{array}{c} x_{14} \\ x_{24} \\ x_{34} \\ x_{44} \\ \vdots \\ x_{n4} \end{array} & \begin{array}{c} \cdots \\ \cdots \\ \cdots \\ \cdots \\ \\ \cdots \end{array} & \left.\begin{array}{c} x_{1p} \\ x_{2p} \\ x_{3p} \\ y_{impu_1}^{[3]} \\ \vdots \\ x_{np} \end{array}\right] \end{array}$$

Para o 2º processo iterativo, segue-se os passos:

1ª Passo: Gerar duas matriz a partir da matriz com os valores médios

1.1 Matriz sem a linha que está o primeiro valor ausente

1.2 Matriz sem a coluna que está o primeiro valor ausente

2ª Passo: Aplicar a padronização nas duas matrizes

3ª Passo: Calcular a DVS para as duas matrizes padronizadas

4ª Passo: Calcular a matriz com o valor imputado, utilizando as decomposições das duas matrizes

5ª Passo: "Despadronizar" a matriz calculada utilizando os valores de média e desvio-padrão encontradas no início

$$\left[x_{miss}^{[1]}, x_{miss}^{[2]}, x_{miss}^{[3]}\right] \Rightarrow \left[y_{impu_2}^{[1]}, y_{impu_2}^{[2]}, y_{impu_2}^{[3]}\right]$$

No final do processo iterativo da imputação múltipla, tem-se que os valores imputados serão:

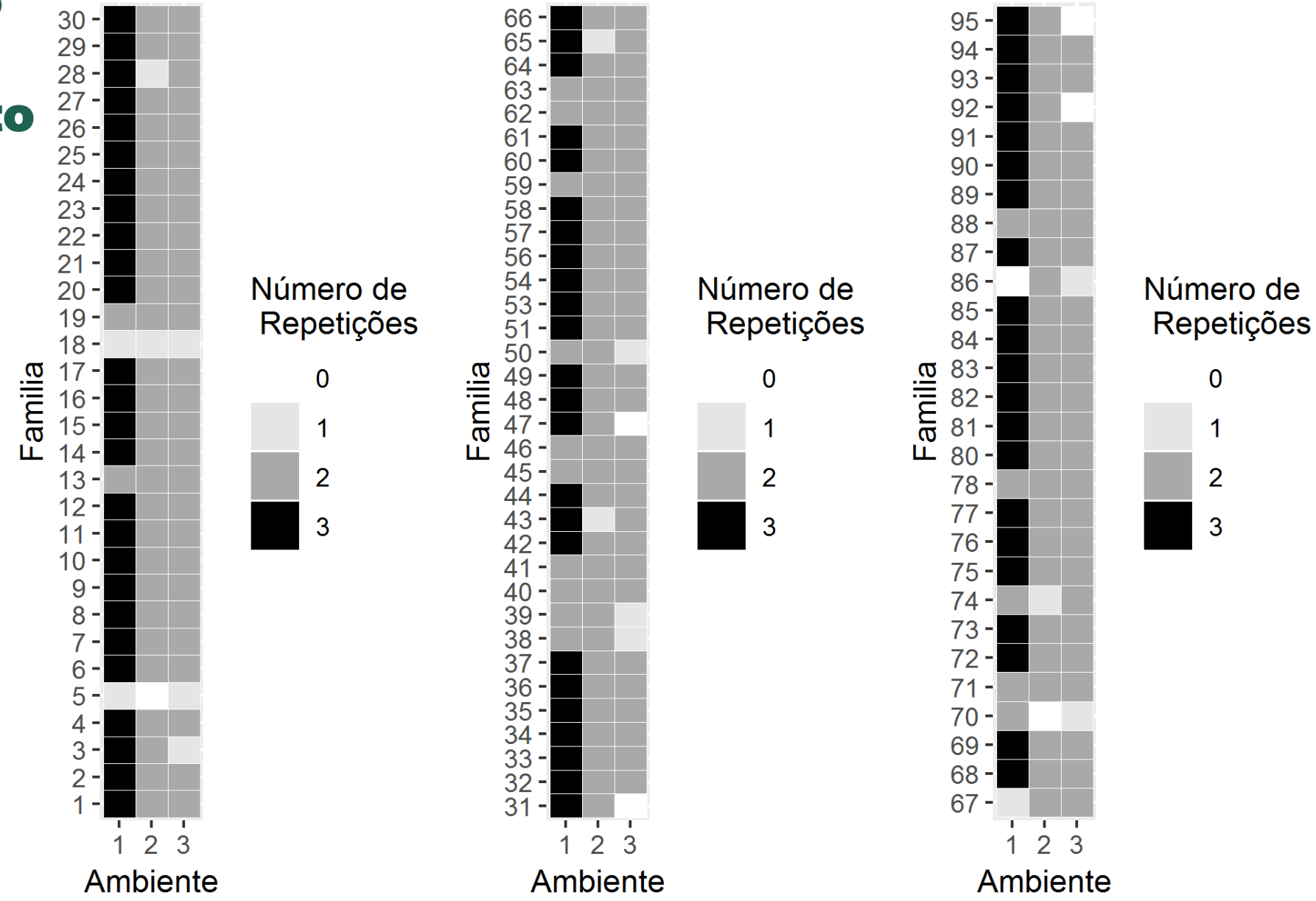
$$\hat{y}_{impu}^{[1]} = \sum_{l=1}^{M} \frac{y_{impu_k}^{[1]}}{M} \qquad \hat{y}_{impu}^{[2]} = \sum_{l=1}^{M} \frac{y_{impu_k}^{[2]}}{M} \qquad \hat{y}_{impu}^{[3]} = \sum_{l=1}^{M} \frac{y_{impu_k}^{[3}}{M}$$

$$\mathbb{X}_{miss} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ \hat{y}_{impu}^{[1]} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ x_{31} & x_{32} & \hat{y}_{impu}^{[2]} & x_{34} & \cdots & x_{3p} \\ x_{41} & x_{42} & x_{43} & x_{44} & \cdots & \hat{y}_{impu}^{[3]} \\ \vdots & & & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix}$$

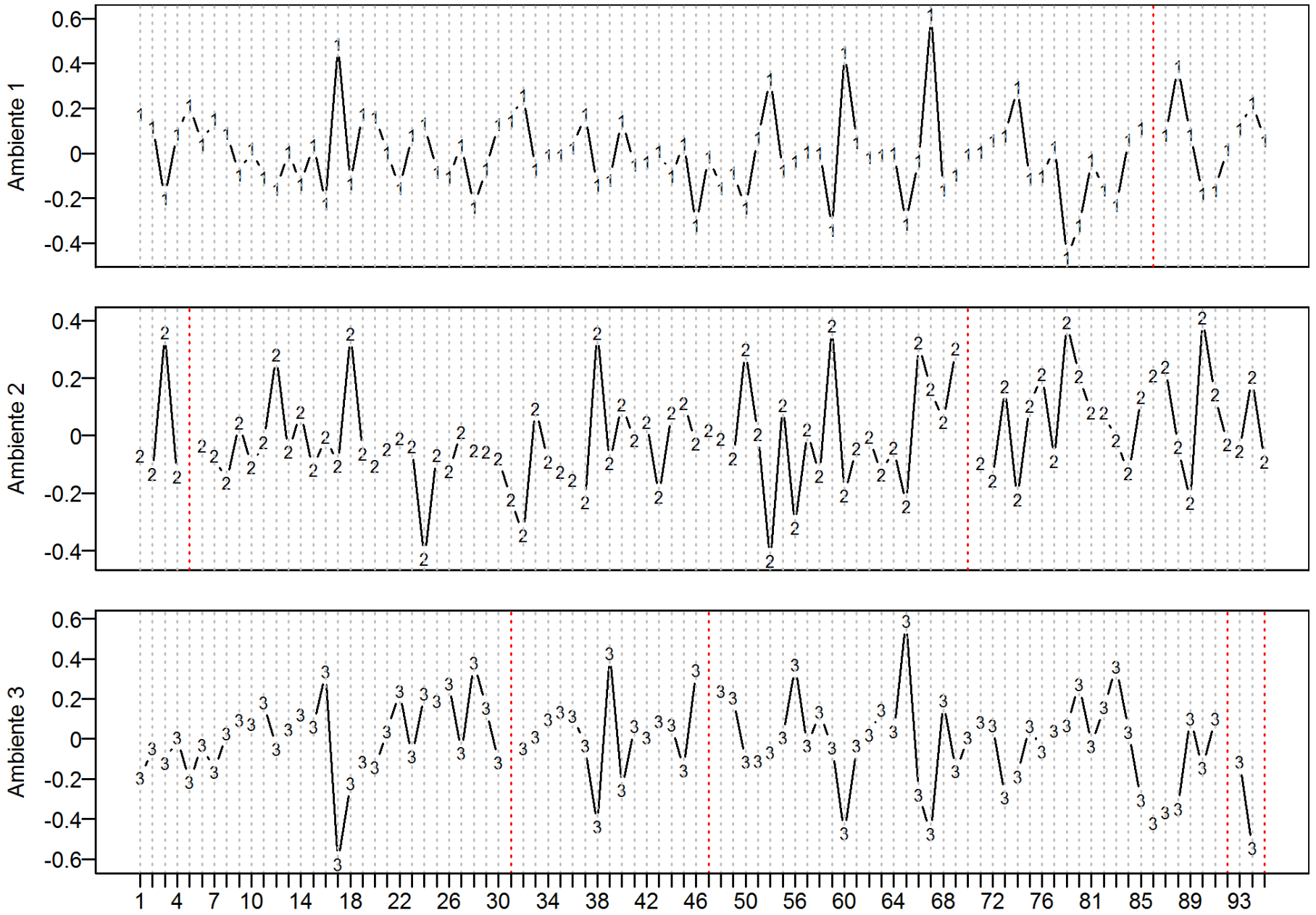
Estrutura da apresentação

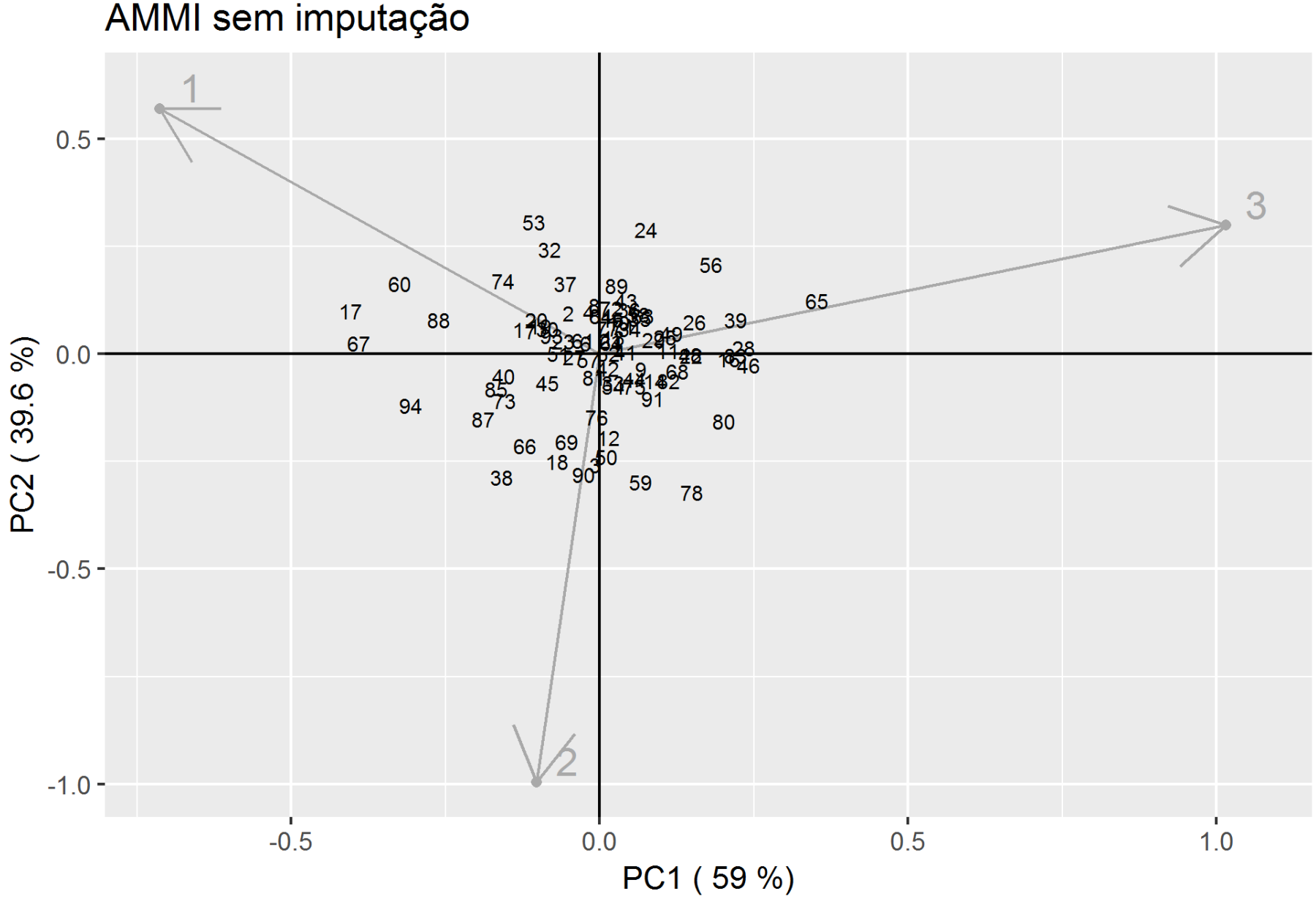
1. Introdução
2. Problemática
3. Metodologia
4. Aplicação
5. Estudos Futuros

Descrição do experimento

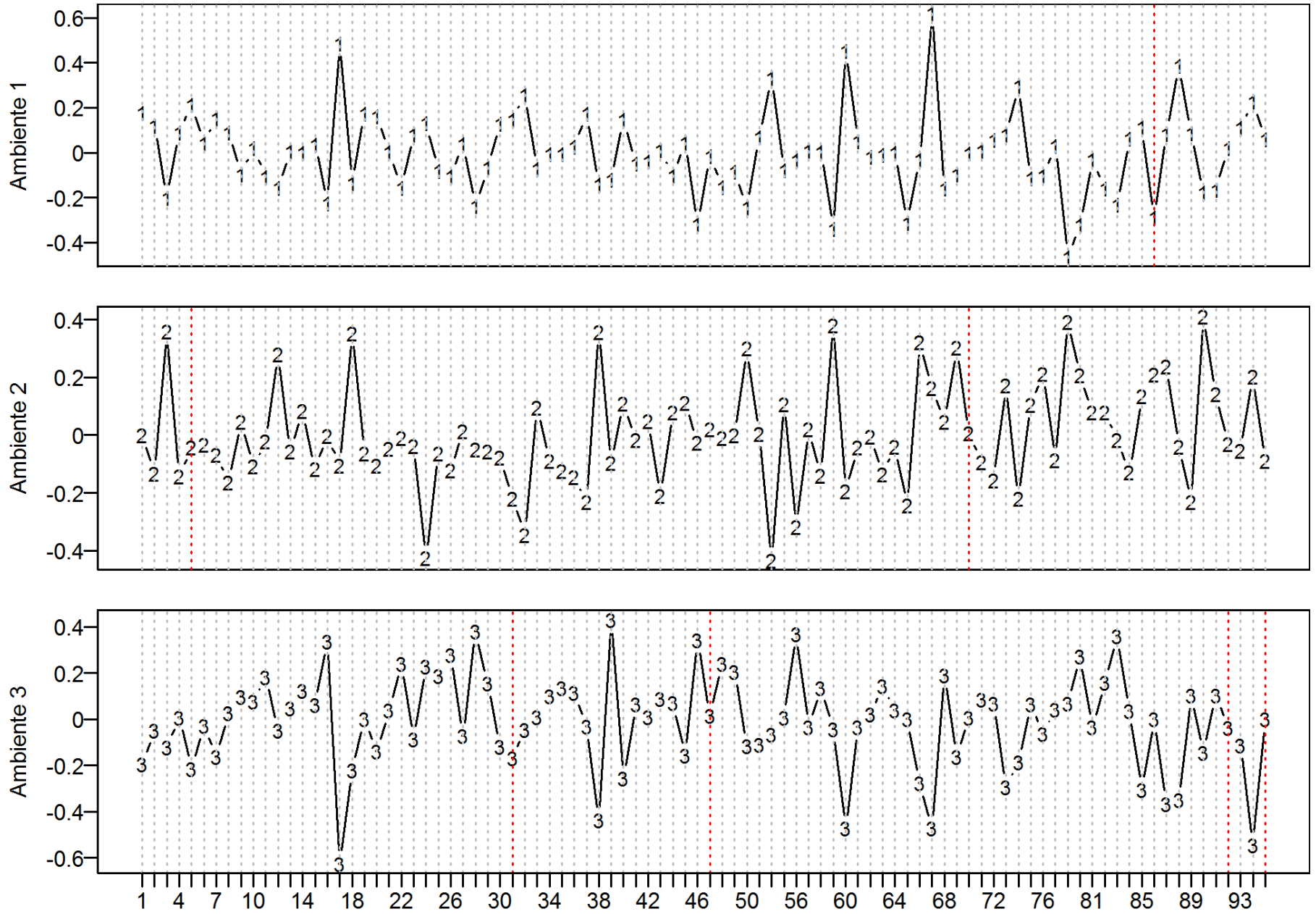


Matriz da interação genótipo e ambiente com valores ausentes

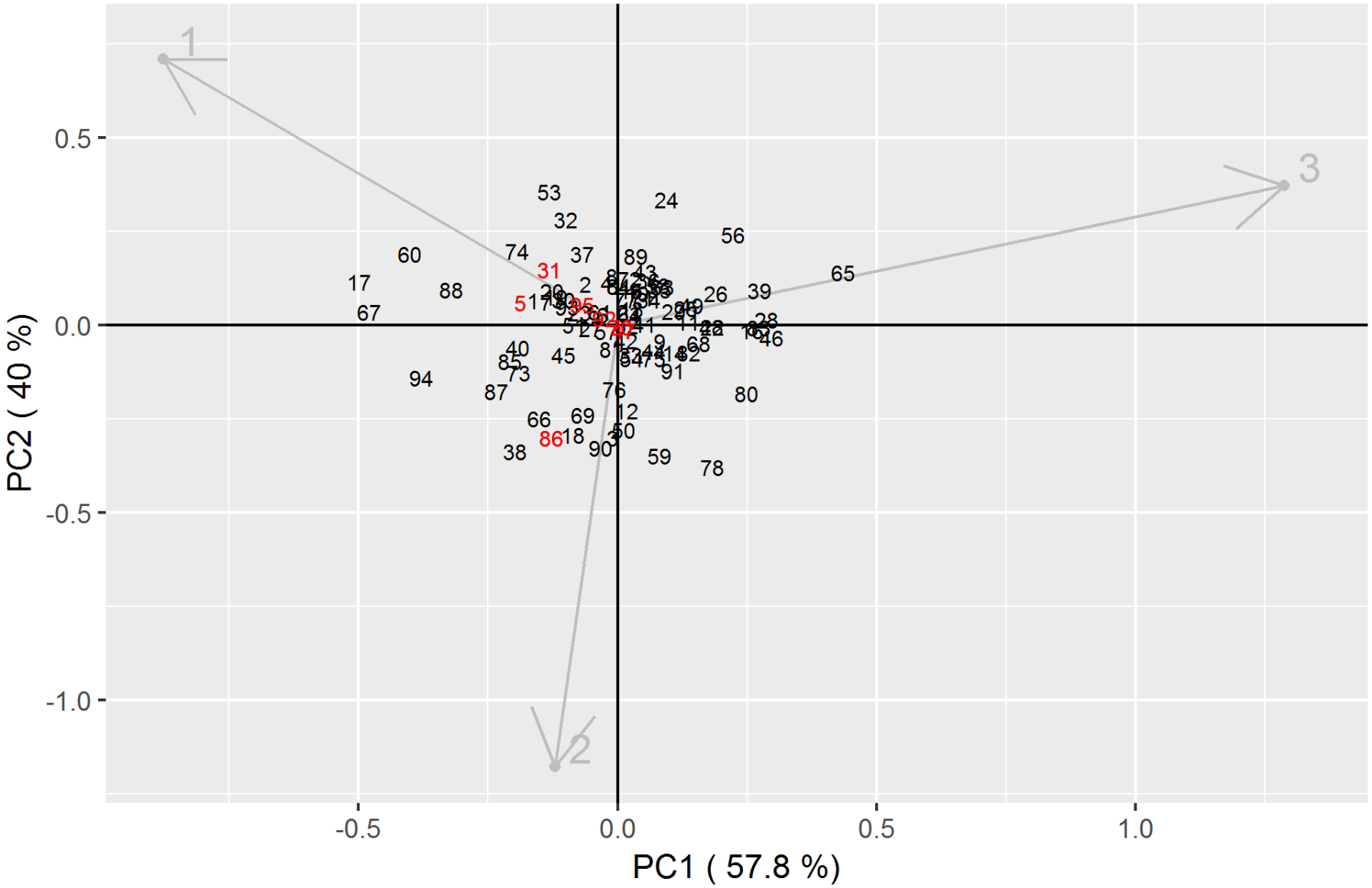



AMMI sem imputação
1
2
3
0.5
0.0
-0.5
-1.0
PC2 (39.6 %)
-0.5
0.0
0.5
1.0
PC1 (59 %)

Matriz da interação genótipo e ambiente com valores imputados



AMMI com imputação



Estrutura da apresentação

1. Introdução
2. Problemática
3. Metodologia
4. Aplicação
5. Estudos Futuros

1. Fazer testes em relação ao número de iterações, e sobre os coeficientes;
2. Estudar sobre a teoria de imputação e encontrar um critério de teste para os valores imputados;
3. Outros métodos de imputação;

VIII ENCONTRO DOS ALUNOS

Estatística e Experimentação Agronômica

Obrigado pela atenção ^^