HAIANY APARECIDA FERREIRA

COMPONENTES DE EFEITOS DE SAFRAS REPRESENTADOS EM BIPLOTS CORRIGIDOS POR PREDIÇÕES DE MODELOS GEE NA CLASSIFICAÇÃO GRANULOMÉTRICA DE CAFÉS

LAVRAS-MG
2019

HAIANY APARECIDA FERREIRA

COMPONENTES DE EFEITOS DE SAFRAS REPRESENTADOS EM BIPLOTS CORRIGIDOS POR PREDIÇÕES DE MODELOS GEE NA CLASSIFICAÇÃO GRANULOMÉTRICA DE CAFÉS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Análise Multivariada, para a obtenção do título de Mestre.

Prof. DSc. Marcelo Ângelo Cirillo
Orientador

Prof.ª DSc. Carla Regina Guimarães Brighenti
Coorientadora

LAVRAS-MG
2019

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Ferreira, Haiany Aparecida.
Componentes de efeitos de safras representados em biplots corrigidos por predições de modelos GEE na classificação granulométrica de cafés / Haiany Aparecida Ferreira. - 2019.
55 p. : il.

Orientador(a): Marcelo Ângelo Cirillo.
Coorientador(a): Carla Regina Guimarães Brighenti.
Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2019.
Bibliografia.

1. Granulometria. 2. Safra. 3. Binomial. I. Cirillo, Marcelo Ângelo. II. Brighenti, Carla Regina Guimarães. III. Título.

HAIANY APARECIDA FERREIRA

COMPONENTES DE EFEITOS DE SAFRAS REPRESENTADOS EM BIPLOTS CORRIGIDOS POR PREDIÇÕES DE MODELOS GEE NA CLASSIFICAÇÃO GRANULOMÉTRICA DE CAFÉS

COMPONENTS OF CROP EFFECTS REPRESENTED IN BIPLOTS CORRECTED BY PREDICTIONS OF GEE MODELS IN THE GRANULOMETRIC CLASSIFICATION OF COFFEE BEANS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Análise Multivariada, para a obtenção do título de Mestre.

APROVADA em 19 de Fevereiro de 2019.

Profª. DSc. Evelise Roman Corbalan Góis	UFLA
Profª. DSc. Izabela Regina Cardoso de Oliveira	UFLA
Profª. DSc. Luciane Teixeira Passos Giarola	UFSJ

Prof. DSc. Marcelo Ângelo Cirillo
Orientador

Profª. DSc. Carla Regina Guimarães Brighenti
Coorientadora

LAVRAS-MG
2019

Aos meus pais Arivaldo e Regiane por sempre me apoiarem nos bons e nos maus momentos.

Vocês são meu maior exemplo. Amo vocês incondicionalmente.

Dedico

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus, pois sem Ele nada disso seria possível. À minha família, meus pais Arivaldo e Regiane e meus irmãos Higor e Kaylon, por sempre me apoiarem ao longo desta caminhada. Aos meus avós Izabel e José, meu tio Reginaldo e meu padrinho Reinaldo que, apesar da distância, sempre torceram por mim.

Ao meu orientador professor Marcelo e a minha coorientadora professora Carla, por sempre confiarem em mim e dar todo suporte necessário ao longo do meu trabalho, são minha inspiração como profissionais e dedicação. Ao meu orientador da graduação professor Nogales e também à professora Karen, por estarem sempre abertos quando precisei e me incentivar a prosseguir meus estudos. As professoras Evelise, Luciane e Izabela, que fizeram parte da minha banca, por todas as contribuições feitas neste trabalho.

Aos meus amigos Carla, Lílian, Simone, Lui, Marcelo, Thaíla e Luciana, por sempre estarem comigo em todos os momentos, sejam eles bons ou ruins. Aos colegas e amigos que fiz no mestrado que estiveram juntos comigo nesta caminhada: Denize, Felipe, Lucas, Matheus, Kelly e Victor. A todos os colegas, em geral, que, de uma maneira ou outra, participaram deste processo da minha vida.

A todos do programa de pós-graduação em Estatística e Experimentação Agropecuária, por todo aprendizado e apoio durante o mestrado. À Universidade Federal de Lavras (UFLA) por proporcionar esta oportunidade e ao Conselho Nacional de Desenvolvimento Cientifico (Cnpq) pela bolsa concedida durante o mestrado e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro ao edital universal APQ -0242-16.

MUITO OBRIGADA!

“A ausência da evidência não significa evidência da ausência.”

(Carl Sagan)

RESUMO

Em uma análise granulométrica de cafés com diferentes categorias de defeitos, os dados podem ser organizados em tabelas de contingências e, ao considerar a discriminação por safra, as mesmas poderão ter uma estrutura que sugere um modelo mais complexo, no tocante, à interação das classificações de defeitos e porcentagens dos grãos de peneiras com efeitos de safra. Diante do exposto, surge a hipótese de que estruturas de correlação são viáveis de serem incorporadas em um modelo, a fim de aprimorar análises gráficas multidimensionais, como a técnica biplots. Com essa motivação, este trabalho tem por objetivo propor o uso de biplots corrigidos por predições de modelos GEE na classificação granulométrica de cafés, discriminada por componentes do efeito das safras. Para validação da proposta, realizações Monte Carlo foram feitas em diferentes estruturas de tabela de contingência em cenários com diferentes graus de correlação. Concluiu-se que o uso de modelos GEE com a técnica biplot corrigida pelas predições é viável de aplicação na análise granulométrica de grãos defeituosos de cafés, com uma eficiente discriminação dos efeitos de safras.

Palavras-chave: Granulometria. Safra. Logit. Complemento Log Log. Binomial.

ABSTRACT

In a granulometric analysis of coffee beans with different defect categories, the data can be organized in contingency tables and, considering discrimination by crops, they might present a structure that suggests a more complex model when it comes to the interaction of crop effects with the defect classifications and percentage of sieve beans. In view of the foregoing, the hypothesis that correlation structures may be incorporated to a model in order to improve multidimensional graphic analysis (such as the biplots technique) arises. Therefore, this work has as its objective to propose the use of biplots corrected by predictions of GEE models in the granulometric classification of coffee beans, discriminated by components of crop effects. To validate the proposal, Monte Carlo realizations were performed in different contingency table structures in scenarios with different degrees of correlation. It was concluded that the use of GEE models with the biplot technique corrected by the predictions is applicable in the granulometric analysis of defective coffee beans, with an efficient discrimination of crop effects.

Keywords: Granulometry. Crop. Logit. Complementary Log Log. Binomial.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

BC	Binomial Correlacionada
CCCMG	Centro de Comércio de Café do Estado de Minas Gerais
CLOGLOG	Complemento Log Log
COB	Classificação Oficial Brasileira
DVS	Decomposição em Valores Singulares
GEE	Equações de Estimação Generalizadas
GLM	Modelos Lineares Generalizados
LIMQL	Informação da Máxima Quase-Verossimilhança Limitada
PVA	Pretos, verdes e ardidos

SUMÁRIO

1. INTRODUÇÃO

A importância da comercialização e produtividade do café brasileiro vem ganhando espaço no setor econômico, graças ao volume de exportação e a dinâmica dos preços internacionais, associado ao forte crescimento do mercado consumidor. Nesse contexto, a cafeicultura brasileira tem contribuído na geração de renda e receitas cambiais, bem como, a formação de capital do setor agrícola.

Em particular, no ano de 2018, dados recentes divulgados pelo portal Valor Econômico (2018) apontam que a safra de café 2018/19 apresentam subsídios para afirmar que a qualidade de cafés colhidos, nesta temporada, é a melhor já vista em relação às safras anteriores e conduzem a uma estimativa de 58 milhões de sacas.

Naturalmente, as condições climáticas que ocorrem nas regiões cafeeiras supostamente influenciam tanto na qualidade como também em sua produção. O portal Café Point (2018), aponta como previsão uma perspectiva de perda de produtividade na safra futura de 2019, em razão da estiagem observada nas principais regiões cafeeiras do centro-sul influenciando um desfolhamento de plantas.

No tocante à avaliação da qualidade e aspecto físico dos grãos de café, mormente utiliza-se a granulometria. A análise granulométrica consiste na distribuição dos diversos tamanhos de grãos de café em porcentagem, a qual as dimensões das partículas do conjunto e das porcentagens de ocorrência. Por meio desse processo também é feita a classificação com base na contagem de grãos defeituosos e impurezas existentes na amostra, seguindo a Tabela de Classificação Oficial Brasileira.

Mediante a natureza dos dados, uma questão a ser ressaltada consiste na metodologia estatística a ser empregada com o propósito de detectar tendências e padrões de reconhecimento. Com esse enfoque, em consonância com os resultados de contagens dos grãos em diferentes categorias, obtidos na classificação dos defeitos, a análise estatística requer uma estrutura de dados organizada em tabelas de contingências.

Essas tabelas permitem a organização dos dados com base em duas ou mais variáveis categóricas, mostram quantos dados se encaixam em cada categoria e possibilitam ajustar diferentes modelos.

Evidentemente, a complexidade dessas tabelas decorre do número de variáveis categóricas a serem consideradas. Diante disso, uma possível estrutura é decorrente, por exemplo, pela contagem dos grãos defeituosos de café mediante diferentes safras interagindo

com a classificação de defeitos e porcentagens de grãos de peneiras, a qual caracteriza um delineamento em blocos com respostas multivariadas.

Contudo, em razão das técnicas embasadas na análise de variância, de forma a considerar que a estrutura de correlação uniforme para todas as parcelas, torna-se viável propor um modelo que permita contemplar diferentes estruturas entre as parcelas, associando os efeitos das safras aos defeitos no procedimento granulométrico aplicado ao café.

Frente ao exposto, conduziu-se este trabalho com o objetivo de propor o uso de modelos de equações de estimação generalizada (LIANG; ZEGER, 1986) em conjunto com a técnica biplot (GREENACRE, 2003) incorporando os efeitos das componentes de safras em uma aplicação referente à análise granulométrica de cafés. Como objetivo secundário foi feito um estudo de simulação quanto ao comportamento das estimativas das componentes.

Assim, espera-se que as predições obtidas pelo modelo GEE possam introduzir um procedimento inferencial à técnica biplot, agregando informações previsíveis em relação à estimação dos autovetores e autovalores, necessários ao entendimento e interpretação dos defeitos dos grãos de cafés em relação a diferentes efeitos de safras.

2. REFERENCIAL TEÓRICO

2.1 Granulometria de Cafés

O alto índice de consumo e exportação da bebida cafeeira tem incentivado a produção e a classificação do produto com o intuito de se obter uma bebida de qualidade mais elevada, considerando o café beneficiado. Segundo Silva et al. (2015), o café beneficiado é obtido por meio de um agrupamento de operações, que têm como foco obter lotes homogêneos que atendam às exigências de comercialização e ou industrialização, garantindo maior evidência de qualidade do produto. Para isso, o café passa por algumas etapas: limpeza, descascagem e classificação dos grãos.

A classificação, conforme ilustrado na Figura 1, tem como objetivo observar os parâmetros de qualidade do café. Segundo o Ministério da Agricultura, Pecuária e Abastecimento (MAPA) (2003), os critérios, segundo a espécie são: o formato do grão e a granulometria, o aroma e o sabor, a bebida, a cor e a qualidade dos grãos. Já, na prática, a classificação tem como base: a espécie; a qualidade, representada por um indicador numérico denominado tipo; e a bebida (SANTOS; NANTES, 2014).

Figura 1 - Classificação dos grãos de cafés.

Fonte: Rede Bahia (2014).

A espécie pode ser classificada em duas categorias: o *Coffea arabica* e o *Coffea canéfora*, mais conhecidos respectivamente como café arábico e café robusto (café conilon). O café arábico refere-se a cafés mais bem avaliados e com densidades mais homogêneas, se comparados ao robusto. Já o indicador numérico tipo refere-se ao precursor qualitativo do café e fornece um indicativo referente à quantidade de defeitos físicos no lote avaliado. Quanto ao indicador da bebida ele é obtido por meio de um teste conhecido como prova da

xícara que a classifica como: estritamente mole; mole; apenas mole; dura; riada; rio ou rio zona. (SANTOS; NANTES, 2014).

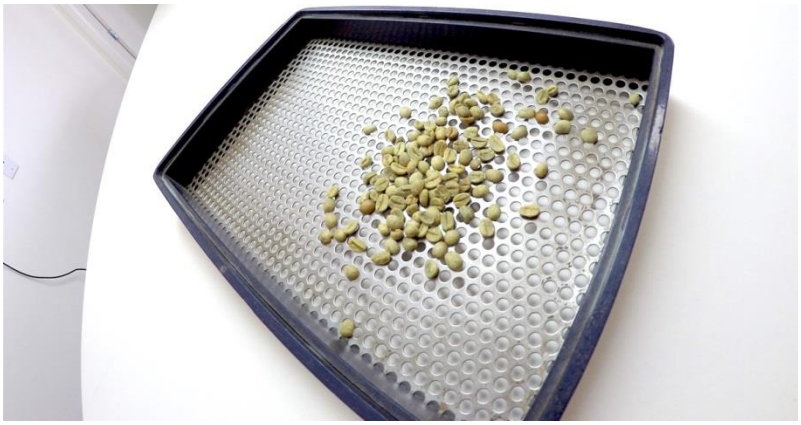
Os defeitos encontrados nos lotes de cafés podem ser intrínsecos, quando atribuídos às imperfeições do próprio grão; ou extrínsecos, quando ocasionados em razão da presença de impurezas. Defeitos extrínsecos são dados por frações estranhas presentes no café beneficiado, como por exemplo: coco, paus, marinheiro, casca e pedras. Defeitos intrínsecos advêm de grãos mal granados, quebrados, chochos, brocados, pretos, verdes, ardidos, de causa genética, fisiológica ou em decorrência de falhas nos procedimentos agrícolas ou industriais (REZENDE, 2015).

Segundo Esquivel e Jiménez (2012), os grãos defeituosos representam cerca de 15 a 20% da produção de café. Dentre esses defeitos, pode-se destacar como uma única categoria os grãos pretos, verdes e ardidos (PVA) e que são considerados os piores defeitos, pois afetam diretamente a qualidade e o tipo de café.

Segundo o Manual de Classificação do Centro de Comércio de Café do Estado de Minas Gerais (CCCMG, 2018), o número de grãos defeituosos avaliados, seguindo a tabela de Classificação Oficial Brasileira (COB) é contado a partir de cada amostra e irá determinar o tipo de café. Para isso, a apuração do número de grãos defeituosos é feita em amostras de 300 gramas cada e estes são convertidos em defeitos conforme a tabela de equivalência. Essa contagem de defeitos tem seu resultado comparado a uma escala de sete níveis, em que cada nível corresponde a um tipo de café. A indicação é numérica e varia desde o tipo 2, café menos defeituoso, até o tipo 8, café mais defeituoso.

A classificação dos grãos em peneiras remete-se a granulometria e ao formato dos grãos, que podem ser classificados como chatos ou mocas (CUSTÓDIO et al., 2007). Um exemplo ilustrativo é apresentado na Figura 2.

Figura 2 – Peneira com amostra de café.



Fonte: Coffee & Joy (2018).

A identificação dos grãos pode ser feita, por meio das dimensões dos crivos das peneiras que os retêm, a qual discrimina os grãos beneficiados (NASSER, 2000). Os crivos redondos são usados para medição e separação dos grãos chatos e os alongados para a separação dos grãos moca. Um grão é considerado chato quando tem uma face mais plana e a outra convexa e moca quando tem o formato arredondado.

Segundo Laviola et al. (2006) os grãos chatos podem se diferenciar dos grãos mocas também pelo processo da fecundação dos grãos, em que o moca se trata de um grão não fecundado. Visto que na formação do fruto têm-se dois grãos, nesse caso, apenas um se desenvolve, o que contribui pra o seu formato arredondado. Em razão disso, o moca não é tão bem aceito quanto o chato em mercados mais inflexíveis, que por si só, toleram apenas lotes de grãos chatos com no máximo 10% de grãos mocas.

No Brasil, a classificação dos grãos chatos se desdobra no intervalo entre peneiras de número 8 a 20. Contudo, para a comercialização internacional, as peneiras aceitas são apenas de números 13 a 20. Isso acontece porque os importadores tendem a preferir peneiras a partir do número 16, já que os cafés de peneiras de números maiores, incluindo outros fatores, têm um valor mais elevado no mercado (LAVIOLA et al., 2006).

Outro fator levado em consideração é o percentual dos grãos por peneiras, que contribui diretamente para que haja a maior uniformidade dos grãos quanto à coloração e à presença de defeitos. Nesse processo, também se evita que, no processo de torra, os grãos menores torrem primeiro que os maiores, podendo ocasionar a carbonização destes, influenciando assim o sabor e o aroma da bebida.

Assim, levando em consideração que a classificação de um grão de café tem como resposta natureza binária, o qual pode ser identificado como defeituoso ou não defeituoso, optou-se por utilizar o modelo binomial para a variável resposta.

2.2 Modelo Binomial e principais funções de ligação

A distribuição binomial está inclusa dentre as distribuições pertencentes à família exponencial para a variável resposta. Essa distribuição tem como característica uma natureza de resposta binária, a qual resulta da realização de *n* ensaios Bernoulli independentes, em que a variável aleatória *Y* é obtida da contagem do total de sucessos de um evento de interesse, sendo $0 < \pi_i < 1$ a probabilidade de sucesso do evento de interesse em cada ensaio e $1\text{-}\pi_i$, a probabilidade de fracasso. Assim, *Y*~binomial (n_i, π_i) tem a função de distribuição de probabilidade dada por:

$$P[Y_i = y_i] = \binom{n_i}{y_i} \pi_i^{y_i} (1\text{-}\pi_i)^{n_i - y_i} \tag{3}$$

em que, $\pi_i^{y_i} (1\text{-}\pi_i)^{n_i - y_i}$ é a probabilidade de obter y_i sucessos, tal que $y_i = 0, \ldots, n_i$. Logo, a média e a variância são dadas por:

$$\begin{aligned} &E[Y_i] = \mu_i = n_i \pi_i, \\ &\text{var}[Y_i] = \sigma_i^2 = n_i \pi_i (1\text{-}\pi_i). \end{aligned} \tag{4}$$

Esse modelo tem como característica uma modelagem dada em proporções e utiliza como funções de ligação o logit e o complemento log-log.

O modelo logit para proporções é construído sobre a representação do modelo binomial na família exponencial, na qual o parâmetro canônico representado por θ sugere que o modelo logit seja definido na sua forma funcional dada por:

$$\pi_i = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}}, \tag{8}$$

em que

$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}. \tag{7}$$

Outras funções de ligação podem ser adaptadas às respostas binomiais tais como a função de distribuição acumulada contínua de valores extremos ou gumbel, caracterizada pelo modelo complemento log log ou cloglog.

Em síntese, a diferenciação básica entre o modelo binomial logit e complemento log-log é perceptível na curva sigmodal. O modelo logit proporciona uma curva simétrica, ao

passo que o modelo complemento log-log resulta em uma curva assimétrica, com um aumento mais acentuado à medida que as probabilidades se aproximam de 1. Diante disso, o uso do complemento log-log é indicado para casos em que as probabilidades são muito pequenas ou muito grandes. Seguindo essas especificações, ele pode ser expresso por:

$$\hat{\pi}_i = 1 - \exp\left\{-\exp\left(\eta_i\right)\right\}.$$

Logo, a função de ligação é dada por

$$\eta_i = \log\left[-\log\left(1-\pi_i\right)\right]. \tag{9}$$

Maiores detalhes da implementação de outras funções de ligação podem ser vistos em Aranda-Ordaz (1981) e Agresti (2007).

2.3 Técnica Biplot

A técnica biplot, cuja fundamentação teórica é relatada por Gabriel (1971), consiste basicamente de um procedimento gráfico que permite visualizar relações entre múltiplas variáveis em um espaço dimensional reduzido.

Naturalmente, para que esse método seja implementado, considera-se uma amostra multivariada, disposta em um arranjo matricial, de modo que, usualmente, as variáveis são descritas nas colunas e as unidades amostrais nas linhas. Para uma melhor visualização, segue um layout da descrição dos dados na Tabela 1, em que u_i corresponde às unidades observadas, Y_j as variáveis, e y_{ij} as observações.

Tabela 1 - Layout de uma tabela de frequência.

u \ Y	Y_1	Y_2	$\cdots$	Y_J
u_1	y_{11}	y_{12}	$\cdots$	y_{1J}
u_2	y_{21}	y_{22}	$\cdots$	y_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
u_I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}

Fonte: Do autor (2019).

Denotando por $Y = \left(Y_{1i}, Y_{2i}, \ldots, Y_{ji}\right)$ a amostra multivariada em sua forma matricial e sendo $Y \in \Re^{I \times J}$ uma matriz de posto *r*, aplica-se uma fatoração, determinada pela decomposição em valores singulares, a qual produz:

$$Y_{(r)I\times J} = U_{(r)I\times r}\Lambda_{(r)r\times r}V^{t}_{(r)r\times J}, \tag{10}$$

em que as colunas das matrizes U e V são ortonormais, ou seja $U^{t}_{(r)I\times r}U_{(r)I\times r} = V^{t}_{(r)r\times J}V_{(r)r\times J} = I_{(r)}$, sendo U uma matriz de autovetores normalizados de $Y_{(r)}Y^{t}_{(r)}$ e V formada por autovetores normalizados obtidos de $Y^{t}_{(r)}Y_{(r)}$; $\Lambda = diag\left(\lambda_1,\ldots,\lambda_r\right)$ uma matriz diagonal formada pelos r valores singulares de Y_r, tais que $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \ldots \geq \sqrt{\lambda_r} \geq 0$ obtidos da matriz $Y^{t}_{(r)}Y_{(r)}$ ou $Y_{(r)}Y^{t}_{(r)}$.

Seguindo essas especificações, $Y_{(r)}$ poderá ser reproduzida como uma soma de r matrizes de posto 1, de tal forma que:

$$Y_{(r)} = \lambda_1 u_1 v^{t}_1 + \lambda_2 u_2 v^{t}_2 + \ldots + \lambda_r u_r v^{t}_r = \sum_{i=1}^{r}\lambda_i u_i v^{t}_i. \tag{11}$$

A redução do espaço dimensional é entendida ao aproximar $Y_{(r)I\times J} \approx Y_{(q)I\times J}$ sendo, $q < r$. Para isso, aplica-se a decomposição de valores singulares, de modo que

$$Y_{(r)I\times J} \approx Y_{(q)I\times J} = U_{(q)I\times q}\Lambda^{(\delta)}_{(q)q\times q}V^{t}_{(q)q\times J} = A_{(q)I\times q}B^{t}_{(q)q\times J} \tag{12}$$

em que $A_{(q)I\times q} = U_{(q)I\times q}\Lambda^{\delta}_{q\times q}$, $B^{t}_{(q)q\times J} = \Lambda^{1-\delta}_{(q)q\times q}V^{t}_{(q)q\times J}$ e δ uma potência $\in[0,1]$. Para certos valores de δ, Gabriel (1971) propõe três representações para os biplots, resumidas na Tabela 2 a seguir:

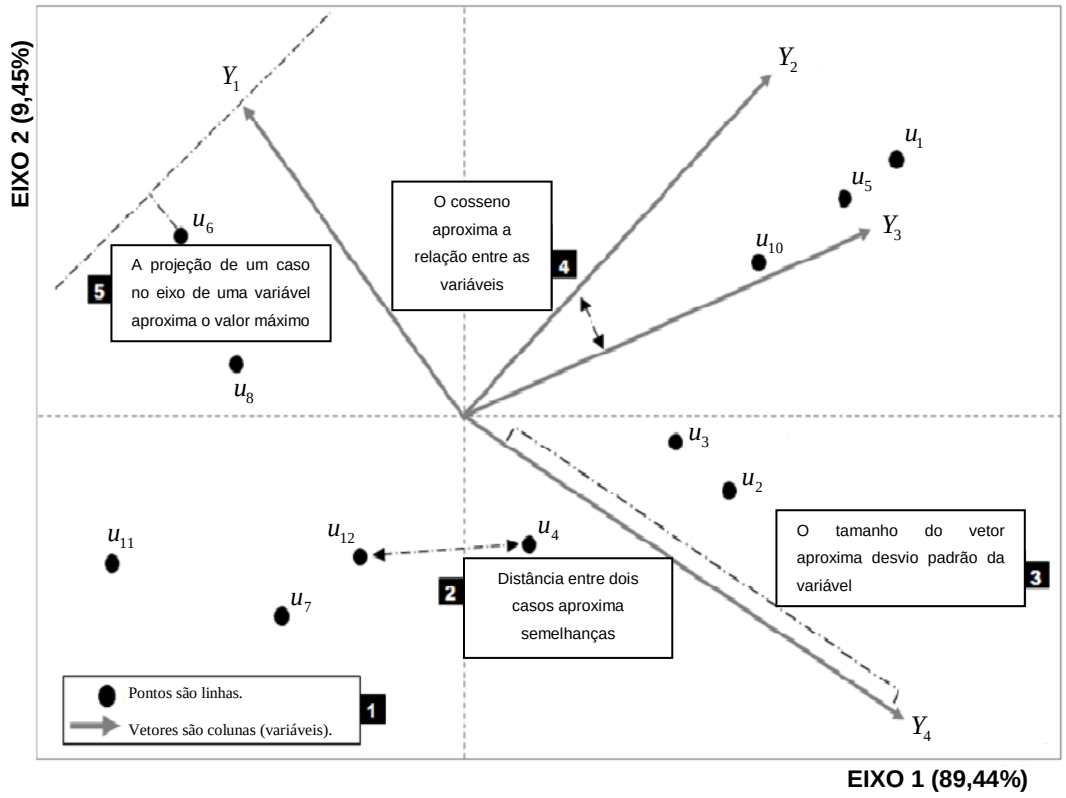
Tabela 2 - Resumo das três representações para os biplots propostas por Gabriel (1971).

		Coordenadas		
δ	**Tipo de biplot**	**Linhas**	**Colunas**	**Recomendação**
0	GH-Biplot	U	ΛV	Prioriza coordenadas de coluna, obtém uma máxima qualidade de representação para as colunas e mínima para as linhas.
1	RMP-Biplot	$U\Lambda$	V	Prioriza coordenadas de linha, obtém uma máxima qualidade de representação para as linhas e mínima para as colunas.
$1/2$	SQRT-Biplot	$U\Lambda^{1/2}$	$V\Lambda^{1/2}$	Atribui a mesma importância das coordenadas de linhas e colunas. Recomenda-se seu uso em testes de diagnóstico em tabelas de contingência.

Fonte: Jolliffe (2002).

Para qualquer tipo, a interpretação de um biplot é ilustrada, por meio da Figura 3, em consonância com a identificação das unidades amostrais e variáveis descritas na Tabela 1.

Figura 3- Representação biplot.



Fonte: Adaptado de Salinas et al. (2013).

Pode-se certificar por meio da Figura 3, a correlação entre as variáveis que pode ser verificada, através dos ângulos formados entre os vetores e geralmente é representada através do cosseno dos ângulos. Desse modo, a correlação será positiva quando o ângulo formado entre as variáveis for agudo, negativa quando o ângulo formado entre as variáveis for obtuso e sem correlação quando o ângulo for reto.

Assim, nota-se que os pontos representam as unidades amostrais, os vetores as variáveis e a projeção de uma unidade no eixo de uma variável se aproxima do valor máximo. Outro fator que também é levado em consideração é o comprimento do vetor das colunas (variáveis) que exibe o valor aproximado do desvio padrão das variáveis. (SALINAS et al., 2013).

Importante enfatizar que não há restrições que impeçam o uso da técnica biplot em dados discretos. Para isso, é natural considerar uma estrutura de uma tabela de contingência similar a um delineamento inteiramente casualizado, conforme sugere o layout apresentado na Tabela 3 a seguir.

Tabela 3 - Layout de uma tabela de frequência com estrutura similar a um delineamento inteiramente casualizado.

Y / X	Y_1	Y_2	…	Y_j	…	Y_s	Total
X_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1.}$
X_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.s}$	$n_{..}$

Fonte: Bussab e Morettin (2010).

O entendimento da conexão de um delineamento experimental casualizado multivariado torna-se compreensível ao considerar as variáveis $Y_1, Y_2, \ldots, Y_s$ como variáveis aleatórias discretas e independentes, $X_1, X_2, \ldots, X_r$ como tratamentos. Assim, cada célula poderia representar uma frequência observada em cada parcela.

Em se tratando de um delineamento de blocos ao acaso, a representação em uma tabela de contingência é apresentada conforme o layout descrito na Tabela 4 a seguir.

Tabela 4 - Layout de uma tabela de frequência com estrutura similar a um delineamento em blocos ao acaso.

Bloco	Y / X	Y_1	Y_2	…	Y_s
A	X_1	$n_{A_{11}}$	$n_{A_{12}}$	…	$n_{A_{1s}}$
	X_2	$n_{A_{21}}$	$n_{A_{22}}$	…	$n_{A_{2s}}$
	⋮	⋮	⋮	⋮	⋮
	X_r	$n_{A_{r1}}$	$n_{A_{r2}}$	…	$n_{A_{rs}}$
B	X_1	$n_{B_{11}}$	$n_{B_{11}}$	…	$n_{B_{11}}$
	X_2	$n_{B_{21}}$	$n_{B_{21}}$	…	$n_{B_{21}}$
	⋮	⋮	⋮	⋮	⋮
	X_r	$n_{B_{r2}}$	$n_{B_{r2}}$	…	$n_{B_{r2}}$

Fonte: Do autor (2019).

Para esse caso, a decomposição em valores singulares (DVS) torna-se mais trabalhosa, cujo procedimento algébrico é descrito na seção seguinte.

2.4 Decomposição em Valores Singulares de uma tabela de contingência estruturada por um delineamento em blocos ao acaso

Seguindo o procedimento de Greenacre (2003), considere duas matrizes representando os blocos *A* e *B* com dimensões *m x n*. A diferença *A*-*B* e a soma *A*+*B* podem ser recuperadas na Decomposição em Valores Singulares (DVS) da matriz bloco

$$\begin{pmatrix} A & B \\ B & A \end{pmatrix}. \tag{13}$$

Com base nessa matriz, a DVS é aplicada de forma independente:

$$\begin{aligned} A+B &= U\Lambda_{\alpha}V^{\mathrm{T}} \\ A-B &= X\Lambda_{\beta}Y^{\mathrm{T}} \end{aligned} \tag{14}$$

sendo $U^{\mathrm{T}}U = V^{\mathrm{T}}V = I$ e $X^{\mathrm{T}}X = Y^{\mathrm{T}}Y = I$, em que *U* e *X* são matrizes de vetores singulares à esquerda, *V* e *Y* são matrizes de vetores singulares à direita, cada um com *k* colunas ortonormais, e Λ_{α} e Λ_{β} são as matrizes diagonais dos valores singulares positivos, em ordem decrescente de grandeza, para a soma e a diferença, respectivamente. Com essas especificações a matriz bloco é recuperada por:

$$\begin{pmatrix} A & B \\ B & A \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} U & X \\ U & -X \end{pmatrix}\begin{pmatrix} D_{\alpha} & 0 \\ 0 & D_{\beta} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} V & Y \\ V & -Y \end{pmatrix}^{\mathrm{T}}, \tag{15}$$

em que os vetores singulares esquerdo e direito na equação são ortonormais. Um ponto que deve ser levado em consideração é que os vetores singulares à esquerda e à direita são todos ortogonais entre si, em razão da ortogonalidade dos vetores nas DVSs e a matriz diferença é retratada pela alteração do sinal dos vetores singulares *X* e *Y*. O fator $\frac{1}{\sqrt{2}}$ garante a normalização correta da solução:

$$\frac{1}{\sqrt{2}}\begin{bmatrix} U \\ U \end{bmatrix}^{T}\frac{1}{\sqrt{2}}\begin{bmatrix} U \\ U \end{bmatrix} = \frac{1}{2}U^{T}U + \frac{1}{2}U^{T}U = I. \tag{16}$$

Logo, os vetores singulares à direita e à esquerda da equação são ortonormais.

Diferente do que acontece na equação (15) as DVSs da soma e da diferença não aparecem separadas, mas sim intercaladas, conforme a magnitude dos valores singulares correspondentes, os quais estão dispostos em ordem decrescente.

A distinção entre os vetores soma e diferença na DVS é facilmente detectada, visto que os vetores singulares à esquerda e à direita da soma têm duas cópias idênticas de um vetor

e são agrupados na mesma coluna. Já os vetores singulares da diferença o vetor agrupado ao vetor inicial têm sinais opostos a este (GREENACRE, 2003; BRIGHENTI; CIRILLO, 2018).

Em relação à decomposição da soma total de quadrados da matriz bloco, duas componentes poderão ser extraídas, uma decorrente da soma e outra da diferença, conforme a equação (17):

$$\underbrace{2\sum_i \sum_j a_{ij}^2 + 2\sum_i \sum_j b_{ij}^2}_{SQ_T} = \underbrace{\sum_i \sum_j \left(a_{ij} + b_{ij}\right)^2}_{SQ_{T_{A+B}}} + \underbrace{\sum_i \sum_j \left(a_{ij} - b_{ij}\right)^2}_{SQ_{T_{A-B}}}. \tag{17}$$

Para reproduzir as somas de quadrados dessas componentes, torna-se necessário trabalhar com a matriz bloco corrigida, envolvendo a centralização em relação à linha $\left[\bar{c}^{\mathrm{T}} \;\; \bar{c}^{\mathrm{T}}\right]$,

$$\begin{pmatrix} A - 1\bar{c}^{\mathrm{T}} & B - 1\bar{c}^{\mathrm{T}} \\ B - 1\bar{c}^{\mathrm{T}} & A - 1\bar{c}^{\mathrm{T}} \end{pmatrix}, \tag{18}$$

em que $\bar{c} = \left(\frac{1}{2}\right)\left(\bar{a} + \bar{b}\right)$ é a média de *A* e *B*. A diferença subtraindo $\bar{c}$ é feita para padronizar, retirando essa análise conjunta de *A* e *B*. Sendo assim, a decomposição da soma de quadrados é reescrita por:

$$\begin{aligned} &2\sum_i \sum_j \left(a_{ij} - \bar{c}_j\right)^2 + 2\sum_i \sum_j \left(b_{ij} - \bar{c}_j\right)^2 \\ &= \sum_i \sum_j \left(a_{ij} - \bar{c}_j + b_{ij} - \bar{c}_j\right)^2 + \sum_i \sum_j \left(a_{ij} - b_{ij}\right)^2. \end{aligned} \tag{19}$$

2.5 Fundamentos da metodologia de Estimação de equações generalizadas

O método de Equações de Estimação Generalizadas (*Generalized Estimating Equations* - GEE) foi proposto por Liang e Zeger (1986) e trata-se de uma extensão dos Modelos Lineares Generalizados (*Generalized linear model* - GLM). Esse método inclui uma estrutura de correlação que supostamente explica a dependência entre as medidas repetidas em um indivíduo e/ou unidade experimental e tem sua análise baseada na classe de modelos de quase-verossimilhança (WEDDERBURN, 1974), em que $V(\mu)$ não precisa ser necessariamente uma função de variância. Contudo, caso $V(\mu)$ pertença a família exponencial, então os modelos GLM e quasi-verossimilhança são equivalentes.

Hardin e Hilbe (2003) classificam didaticamente os modelos GEE em PA-GEE e SS-GEE, de modo que PA corresponde ao acrônimo de “Population Averaged” e SS refere-se ao acrônimo de “Subject Specific”. A diferenciação entre as duas divisões é dada nos seguintes aspectos:

- PA-GEE: o interesse é dado na resposta marginal média sobre a população de indivíduos e os coeficientes do modelo, dado em $\beta_{(PA)}$ são interpretados em termos da resposta média;
- SS-GEE: o interesse é dado em modelar a origem da heterocedasticidade, e os coeficientes $\beta_{(SS)}$ tem uma interpretação para estudar a “trajetória” do indivíduo.

O grupo de modelos derivado de GEE mais conhecido e mais utilizado pelos pesquisadores é o PA-GEE, o qual fornece uma introdução às equações de estimativas generalizadas, uma justificativa teórica e propriedades assintóticas para os estimadores resultantes (RAJAGOPALAN et al., 2015). Para escrever esse modelo, a princípio toma-se a equação de estimativa de Informação da Máxima Quase-Verossimilhança Limitada (LIMQL) para GLMs:

$$\Psi\left(\beta_{PA}\right)=\left[\left\{\sum_{i=1}^{n}\sum_{t=1}^{n_i}\frac{y_{it}-\mu_{it}}{a(\phi)V\left(\mu_{it}\right)}\left(\frac{\partial\mu}{\partial\eta}\right)_{it}x_{jit}\right\}_{j=1,\ldots,p}\right]_{p\times1}=\left[0\right]_{p\times1} \tag{20}$$

e sua forma reescrita em termos de matriz dos painéis

$$\Psi\left(\beta_{PA}\right)=\left[\left\{\sum_{i=1}^{n}x_{ji}^{\mathrm{T}}D\left(\frac{\partial\mu}{\partial\eta}\right)\left[V\left(\mu_i\right)\right]^{-1}\left(\frac{y_i-\mu_i}{a(\phi)}\right)\right\}_{j=1,\ldots,p}\right]_{p\times1}=\left[0\right]_{p\times1} \tag{21}$$

em que $\frac{\partial\mu}{\partial\eta}=\mu$; ϕ corresponde ao parâmetro de dispersão; $a(\phi)$ corresponde a uma função obtida na escrita da distribuição por meio da distribuição exponencial; y a resposta associada à i-ésima repetição na t-ésima parcela e/ou indivíduo; $D()$ denota a matriz diagonal e $V(\mu_i)$ é a matriz diagonal a qual pode ser decomposta:

$$V\left(\mu_i\right)=\left[D\left(V\left(\mu_{it}\right)\right)^{1/2}I_{(n_i\times n_i)}D\left(V\left(\mu_{it}\right)\right)^{1/2}\right]_{n_i\times n_i} \tag{22}$$

Por meio dessa apresentação, nota-se que a equação de estimação trata cada observação dentro de um painel, enumerado por *t=1,...,T*, com uma estrutura de correlação independente. Logo, o valor esperado e as funções da variância não mudam apesar da especificação atribuída para a equação de estimação do LIMQL para GLMs. Portanto, seja na abordagem LIMQL ou GLM as estimativas de $\beta_{(PA)}$ são equivalentes.

A incorporação de outras estruturas de correlação é dada na substituição da matriz identidade $I_{(n_i \times n_i)}$ por $R(\alpha)_{(n_i \times n_i)}$, denominada, então, como estrutura de correlação de trabalho. Dessa forma, $V(\mu)$ é reescrita como:

$$V(\mu_i) = \left[D\left(V(\mu_{it})\right)^{1/2} R(\alpha)_{(n_i \times n_i)} D\left(V(\mu_{it})\right)^{1/2} \right]_{n_i \times n_i}. \tag{23}$$

Note que a matriz de correlação $R(\alpha)_{(n_i \times n_i)}$ é escrita em função dos parâmetros de associação, definido por um escalar α ou vetor, dependendo da estrutura de correlação. Em particular, a matriz segue as especificações para estrutura de correlação uniforme, a qual considera que as observações dentro de um painel tenham alguma correlação comum. Tem-se α como um escalar e a matriz de correlação de trabalho uniforme dada por

$$R(\alpha)_{(n_i \times n_i)} = \begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix} \tag{24}$$

que pode ser reescrita como:

$$R_{uv} = \begin{cases} 1 \text{ se } u = v \\ \alpha \text{ caso contrário} \end{cases}. \tag{25}$$

Esse princípio é indicado para casos em que as medições repetidas dos conjuntos de dados não dependem do tempo e permitem qualquer permutação (HARDIN; HILBE, 2003). A estimação do parâmetro de correlação comum para um GEE com estrutura de correlação permutável, pode ser feita usando os resíduos estimados $\hat{r}_{it} = \left(y_{it} - \hat{\mu}_{it}\right) \Big/ \sqrt{V\left(\hat{\mu}_{it}\right)}$ a partir do ajuste atual do modelo. A estimativa de α usando esses resíduos é dada por:

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \sum_{i=1}^{n} \left\{ \frac{\sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \hat{r}_{iu} \hat{r}_{iv} - \sum_{v=1}^{n_i} \hat{r}_{iu}^2}{n_i \left(n_i - 1\right)} \right\} \tag{26}$$

Outro ponto a ser ressaltado, seguindo as argumentações de Sutradhar e Das (2000), é que na prática $R(\alpha)_{(n_i \times n_i)}$ é desconhecido e que a especificação incorreta de $R(\alpha)_{(n_i \times n_i)}$ afeta a eficiência de (β). Portanto, convém aplicar critérios que avaliam a adequacidade da imposição de $R(\alpha)$. Maiores detalhes podem ser vistos em Silva e Cirillo (2018).

Isso exposto, Silva (2017) descreve que a modelagem via GEE é dada basicamente em dois procedimentos de estimação. O primeiro método é conhecido como equações de estimação generalizada de primeira ordem, especificado por GEE1. Nessa abordagem, α é

tratado como parâmetro de perturbação e o interesse principal está na obtenção das estimativas de β, usualmente proposta por Liang e Zeger (1986).

A segunda abordagem, proposta por Prentice e Zhao (1991), conhecida na literatura por GEE2, utiliza as equações de estimação para obtenção das estimativas dos parâmetros de regressão (β) e de associação (α) conjuntamente. A vantagem dessa abordagem é que os parâmetros de associação (α) são estimados de forma mais precisa, porém, a consistência dos parâmetros de regressão (β) depende da especificação correta do modelo.

3. MATERIAIS E MÉTODOS

Em função dos objetivos propostos, a metodologia utilizada nesse trabalho será dividida em duas etapas: 3.1 Descrição dos dados experimentais utilizados na análise granulométrica e 3.2 Ajuste dos modelos GEE e estudos de simulação das componentes dos efeitos de safra: $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$.

3.1 Descrição dos dados experimentais obtidos da análise granulométrica

Em decorrência da proposta deste trabalho em realizar uma modelagem estatística que contemple o efeito de safras na análise granulométrica de cafés, utilizou-se um conjunto de dados referente a amostras de cafés da variedade catuaí, encontrado em Brighenti e Cirillo (2018).

Conforme descrevem os autores, a classificação de defeitos nos grãos de café de cada amostra seguiu a tabela de Classificação Oficial Brasileira (COB) considerando a proporção dos grãos defeituosos ou das impurezas contidas em uma amostra de 300g de café beneficiado provenientes de colheitas feitas por produtores rurais do Sul de Minas nos anos de 2014 e 2015.

A classificação dos grãos chatos de números 14 a 18 e mocas de números 9 a 13 foi feita utilizando peneiras intercaladas. O percentual de retenção inicialmente foi considerado de cada peneira individualmente. Em seguida do somatório das peneiras 17 e superiores foram atribuídos para grãos chatos graúdos advindos da retenção de peneiras circulares de crivo P17/18 (BRASIL, 2003).

Em razão dessas especificações, considerou-se a classificação das amostras em cinco defeitos (Tabela 5).

Tabela 5 – Descrição dos defeitos considerados na granulometria.

Codificação	Defeito	Descrição
d1	Brocados	Provenientes do ataque de pragas.
d2	Pretos, Verdes e Ardidos	Dos possíveis defeitos advindos da colheita têm-se os verdes, provenientes de colheita prematura, e os pretos ou ardidos de colheita atrasada e fermentações prorrogadas. Todos têm suas contagens agrupadas em uma categoria denominada PVA.
d3	Quebrados	Oriundos de descascador mal regulado ou seca inadequada.
d4	Conchas	Causados por fatores genéticos ou por possíveis causas fisiológicas.
d5	Impurezas	Ocasionados por frações estranhas presentes no café beneficiado, como por exemplo: casca, paus, pedras, dentre outros.

Fonte: Do autor (2019).

Mantendo essa classificação, obtiveram-se os dados organizados em uma tabela de contingência de tripla entrada com as contagens dadas em proporções, em relação aos tipos de defeitos e porcentagens dos grãos em peneiras, considerando as amostras coletadas nas safras de 2014 e 2015 (Tabela 6).

Tabela 6 - Proporções dos grãos das safras 2014 e 2015 quanto aos tipos de defeitos e porcentagens dos grãos em peneiras (17/18).

Safra	Defeitos	Porcentagens (p) dos grãos em peneiras (17/18)		
		$p < 20\%$	$20\% \leq p \leq 30\%$	$p > 30\%$
2014	d1	33	40	27
	d2	31	38	31
	d3	38	32	30
	d4	24	47	29
	d5	37	25	38
2015	d1	36	26	38
	d2	27	32	41
	d3	32	42	26
	d4	24	33	43
	d5	19	0	81

Fonte: Do autor (2019).

Seguindo a estrutura dos dados (Tabela 6) o efeito de safra associado aos defeitos (Tabela 5) foi incorporado por meio da técnica biplot, considerando os valores preditos obtidos pelo ajuste de diferentes modelos de equações de estimação generalizada. A decomposição em valores singulares foi feita nas matrizes blocos combinadas, em que cada bloco corresponde às avaliações granulométricas tendo em vista as amostras coletadas nas safras de 2014 e 2015, nomeadas em S_{2014} e S_{2015}, conforme procedimento sugerido por Greenacre (2003).

Nesse âmbito, as componentes referentes ao efeito de safras foram determinadas por: $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$, correspondendo ao efeito combinado e à diferença entre as referidas safras.

3.2 Ajuste dos modelos GEE e estudos de simulação das componentes dos efeitos de safra: $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$.

Em decorrência da estrutura da tabela de contingência (Tabela 6), o delineamento utilizado na obtenção dos parâmetros foi especificado nos fatores safra *(S)*, defeitos *(D)* e proporção de defeitos observados na peneira *(P)*.

Com essas especificações, a correlação entre as observações foi caracterizada pelas medidas repetidas dentro do bloco S_k, $(k=1,2)$ em cada nível de "linha" *(D)* e "coluna" *(P)*. Desta forma, cada unidade experimental y_{ijk} foi organizada em um vetor $Y_r=[Y_{r1},Y_{r2},..., Y_{rN}]$; $r=1,...,N$, sendo N o número total de parcelas para cada bloco, conforme o delineamento:

$$X = \begin{pmatrix} S & D & P \\ 1 & -2 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & -2 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & -2 & -1 \\ 1 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & -1 \\ 1 & 2 & -1 \\ -1 & -2 & 1 \\ -1 & -1 & 1 \\ -1 & 0 & 1 \\ -1 & 1 & 1 \\ -1 & 2 & 1 \\ -1 & -2 & 0 \\ -1 & -1 & 0 \\ -1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 2 & 0 \\ -1 & -2 & -1 \\ -1 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & 1 & -1 \\ -1 & 2 & -1 \end{pmatrix}, \quad (27)$$

em que 1 e -1 correspondem as safras, -2, -1, 0, 1, 2 aos 5 defeitos e 1, 0, -1 as porcentagens de grãos por peneira.

A estrutura dos dados sugere um delineamento em blocos ao acaso multivariado, de modo que as safras foram identificadas como efeito de bloco, os tipos de defeito tratamentos e as porcentagens dos grãos de peneira como respostas multivariadas.

Com o propósito de investigar o efeito do grau de correlação entre as classificações de defeitos em ambas as safras e seu impacto nas estimativas das somas de quadrado das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$, procedeu-se com um estudo de simulação. Para tanto, adotou-se uma tabela de contingência com estrutura similar aos dados (Tabela 6) e amostras binomiais correlacionadas, geradas pela distribuição binomial correlacionada. Dessa forma, a notação utilizada para especificação do modelo, seguiu as definições descritas na Tabela 7:

Tabela 7 - Layout das observações simuladas seguindo o modelo binomial correlacionado.

Bloco	P / D	P_1	P_0	P_{-1}	Total
S_1	D_{-2}	y_{111}	y_{121}	y_{131}	$n_{1.1}$
	D_{-1}	y_{211}	y_{221}	y_{231}	$n_{2.1}$
	D_0	y_{311}	y_{321}	y_{331}	$n_{3.1}$
	D_1	y_{411}	y_{421}	y_{431}	$n_{4.1}$
	D_2	y_{511}	y_{521}	y_{531}	$n_{5.1}$
S_{-1}	D_{-2}	y_{112}	y_{112}	y_{132}	$n_{1.2}$
	D_{-1}	y_{212}	y_{212}	y_{232}	$n_{2.2}$
	D_0	y_{312}	y_{312}	y_{332}	$n_{3.2}$
	D_1	y_{412}	y_{412}	y_{432}	$n_{4.2}$
	D_2	y_{512}	y_{512}	y_{532}	$n_{5.2}$

Fonte: Do autor (2019).

Nessa Tabela 7, tem–se que $Y_{ijk} \sim BC(\pi_{i..}, n_{i..}, \rho)$ para i=1,...,5; j=1,...,3 e k=1 ou 2 com distribuição de probabilidade dada por (CIRILLO; RAMOS, 2014):

$$P\left(Y_{ijk} \middle| n_{i..}, \pi_{i..}\right) = \binom{n_{i..}}{y_{ijk}} \pi_{i..}^{y_{ijk}} \left(1-\pi_{i..}\right)^{n_{i..}-y_{ijk}} \left(1-\rho\right) I_{\theta_1}\left(y_{ijk}\right) + \pi^{y_{ijk}/n_{i..}} \rho I_{\theta_2}\left(y_{ijk}\right), \qquad (28)$$

sendo, ρ a taxa de mistura entre as distribuições binomial $(n_{i..}, \pi_{i..})$, com probabilidade $(1-\rho)$, e uma distribuição bernoulli modificada, representada pela variável $\mathrm{BeM}(\pi)$, assumindo 0 ou $n_{i..}$ valores com probabilidade ρ, fixado o vetor $\pi = (0{,}5; 0{,}5; 0{,}5; 0{,}5; 0{,}5)$. Assim, os valores paramétricos utilizados na geração das observações binominais encontram-se na Tabela a seguir.

Tabela 8 - Valores paramétricos utilizados como cenários para gerar as tabelas de contingência.

Cenário	***n***	**Grau de correlação (ρ)**
1	50	0,2
2		0,5
3		0,7
4	100	0,2
5		0,5
6		0,7
7	150	0,2
8		0,5
9		0,7

Fonte: Do autor (2019).

Com a realização desse procedimento, a correlação (ρ) entre as variáveis "linha" (Tabela 8), justifica supor o efeito de superdispersão (STONER; LEROUX, 2002), uma vez que observações correlacionadas é uma das causas que gera uma variabilidade amostral superior a uma variabilidade esperada pelo modelo ajustado. Assim, torna-se inapropriado o uso de modelos generalizados, como alternativa, utilizou-se o modelo GEE, dado pela solução do sistema:

$$\sum_{i=1}^{N}\left(\frac{\partial\mu_i(\beta)}{\partial\beta}\right)R_i^{-1}\left(Y_i - \hat{\mu}_i(\beta)\right)=0 \tag{29}$$

em que $\hat{\mu}_i(\beta)$ correspondeu ao vetor de médias ajustados considerando os modelos binomiais com função de ligação logit (30) e cloglog (31)

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)=\eta_{ijk}\ ;\ \ \mu_{ijk}=\frac{\exp\left(\eta_{ijk}\right)}{1+\exp\left(\eta_{ijk}\right)} \tag{30}$$

$$\log\left\{-\log\left(1-\mu_{ijk}\right)\right\}=\eta_{ijk};\ \ \mu_{ijk}=1-\exp\left\{-\exp\left(\eta_{ijk}\right)\right\}. \tag{31}$$

R_i^{-1}, corresponde à inversa da matriz de correlação uniforme, encontrada em Barnett et al. (2010) e dada por:

$$R_i=\begin{bmatrix}1 & \hat{\alpha} & \cdots & \hat{\alpha}\\ \hat{\alpha} & 1 & \cdots & \hat{\alpha}\\ \vdots & \cdots & \ddots & \vdots\\ \hat{\alpha} & \hat{\alpha} & \cdots & 1\end{bmatrix},$$

$$R_{rs}^{-1}=\begin{cases}\left[1+(m-2)\rho\right]/\gamma\ ;\ \ r=s=1,...,m\\ -\rho/\gamma\ \ ;\ \ r,s=1,...,m,r\neq s\end{cases}. \tag{32}$$

Seja m o número de covariáveis consideradas no modelo e $\gamma = \sigma^2[1 + (m-2)\rho + (m-1)\rho^2]$. A estimativa do parâmetro de associação α foi dada por:

$$\hat{\alpha} = \frac{\phi \sum_{r=1}^{N} \sum_{i \geq r} \hat{e}_{r_i} \hat{e}_{r_i}}{\left\{ \sum_{r=1}^{N} \frac{1}{2} n_r \left(n_r - 1 \right) - d \right\}}, \tag{33}$$

sendo d o número de parâmetros e $\hat{\phi}$ a estimativa do parâmetro de dispersão em função do resíduo de Pearson

$$\hat{e}_{r_i} = \frac{y_{r_i} - \hat{\mu}_{r_i}}{V\left(\hat{\mu}_{r_i}\right)}. \tag{34}$$

Após a obtenção de 1000 realizações Monte Carlo (R CORE TEAM, 2016), em ambos os modelos, para o estudo da aplicação em safras obtiveram-se valores preditos g_{ijk}, sendo esses organizados em uma tabela de tripla entrada, conforme layout ilustrado na Tabela 7 tornando-se possível computar a média das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$. Os códigos utilizados para a obtenção dos resultados se encontram nos apêndices A e B.

4. RESULTADOS E DISCUSSÃO

4.1 Avaliação do comportamento das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$ em dados correlacionados ajustados pelo modelo GEE considerando a estrutura de correlação uniforme

Os resultados descritos, nas Tabelas 9 e 10, correspondem à média das estimativas das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$ em 1000 realizações Monte Carlo. As estimativas são obtidas, por meio da decomposição de valores singulares, aplicada nas tabelas geradas com frequência, em diferentes graus de correlação em relação aos níveis organizados em "linhas". Para isso, seguiu-se o layout da Tabela 8 para cada bloco.

Tabela 9 - Cenários de simulação e média das estimativas das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$ obtidas pela decomposição de valores singulares para o modelo GEE com função de ligação logit.

Modelo GEE-Logit			
n	**Grau de correlação (ρ)**	**Componente $S_{2014}+S_{2015}$**	**Componente $S_{2014}-S_{2015}$**
50	0,2	0,0271	0,0055
	0,5	0,0178	0,0037
	0,8	0,0073	0,0023
100	0,2	0,0151	0,0024
	0,5	0,0086	0,0018
	0,8	0,0036	0,0013
150	0,2	0,0094	0,0017
	0,5	0,0059	0,0012
	0,8	0,0022	8e-04

Fonte: Do autor (2019).

Tabela 10 - Cenários de simulação e média das estimativas das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$ obtidas pela decomposição de valores singulares para o modelo GEE com função de ligação cloglog.

Modelo GEE-Cloglog			
n	**Grau de correlação (ρ)**	**Componente $S_{2014}+S_{2015}$**	**Componente $S_{2014}-S_{2015}$**
50	0,2	0,0287	0,0053
	0,5	0,0174	0,0034
	0,8	0,0072	0,0024
100	0,2	0,0141	0,0027
	0,5	0,0089	0,0021
	0,8	0,0033	0,0012
150	0,2	0,0091	0,0018
	0,5	0,0060	0,0013
	0,8	0,0024	7e-04

Fonte: Do autor (2019).

Em se tratando do modelo GEE-logit, notou-se que o aumento do grau de correlação ρ resultou em uma redução na estimativa da componente $S_{2014}+S_{2015}$, para todos os tamanhos amostrais avaliados. No tocante a componente $S_{2014}-S_{2015}$, esse comportamento foi menos expressivo, o que conduz a afirmar que a decomposição da variância amostral pelos dois primeiros componentes principais extraídos do efeito de igualdade entre os blocos ($S_{2014}-S_{2015}$) é mais robusto em relação ao aumento do grau de correlação, independente do tamanho amostral considerado.

Especificamente, para o modelo GEE-cloglog, evidenciou-se que o comportamento das estimativas é análogo ao do modelo logit.

4.2 Aplicação da técnica de biplots na granulometria de cafés

Após esse estudo preliminar, procedeu-se com a aplicação aos dados referentes a granulometria de cafés. Os biplots foram construídos considerando o efeito aditivo das contagens dos defeitos realizados nas safras 2014 e 2015, representada por $S_{2014}+S_{2015}$ e da homogeneidade dada por $S_{2014}-S_{2015}$. Dessa forma, os valores preditos, nos quais os biplots foram centrados e dados pelos modelos GEE-logit e GEE-cloglog, respectivamente, seguem as expressões a seguir.

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)=\eta_{ijk}$$
$$\begin{aligned}\eta_{ijk} = &-0,0196S_2 - 0,2199D_2 + 0,0187D_3 - 0,4495D_4 - 0,2633D_5\\ &- 0,1157P_2 - 0,3779P_3 + 0,2950D_2\times P_2 + 0,1283D_3\times P_2\\ &+ 0,6998D_4\times P_2 - 0,8597D_5\times P_2 + 0,3445D_2\times P_3 - 0,1956D_3\times P_3\\ &+ 0,5741D_4\times P_3 + 1,0537D_5\times P_3 + 0,0003S_2\times P_2 + 0,0020S_2\times P_3\end{aligned} \quad (35)$$

$$\log\left\{-\log\left(1-\mu_{ijk}\right)\right\}=\eta_{ijk}$$
$$\begin{aligned}\eta_{ijk} = &-0,0199S_2 - 0,1957D_2 + 0,0166D_3 - 0,4045D_4 - 0,2345D_5\\ &- 0,1076P_2 - 0,3456P_3 + 0,2623D_2\times P_2 + 0,1132D_3\times P_2\\ &+ 0,6245D_4\times P_2 - 0,8100D_5\times P_2 + 0,3088D_2\times P_3\\ &- 0,1792D_3\times P_3 + 0,5176D_4\times P_3 + \ 0,9279D5\times P_3\\ &+ 0,0129S_2\times P_2 + 0,0131S_2\times P_3\end{aligned} \quad (36)$$

Para cada modelo, utilizou-se a estrutura de correlação de trabalho uniforme, $R(\alpha)$, cujas estimativas obtidas pelo ajuste encontram-se descritas a seguir e referem-se, respectivamente, aos modelos logit e cloglog.

$$R(\alpha)=\begin{bmatrix} 1 & -0,0301 & -0,0301 & -0,0301\\ -0,0301 & 1 & -0,0301 & -0,0301\\ -0,0301 & -0,0301 & 1 & -0,0301\\ -0,0301 & -0,0301 & -0,0301 & 1\end{bmatrix} \quad (37)$$

$$R(\alpha)=\begin{bmatrix} 1 & -0,0294 & -0,0293 & -0,0293\\ -0,0293 & 1 & -0,0293 & -0,0293\\ -0,0293 & -0,0293 & 1 & -0,0293\\ -0,0293 & -0,0293 & -0,0293 & 1\end{bmatrix} \quad (38)$$

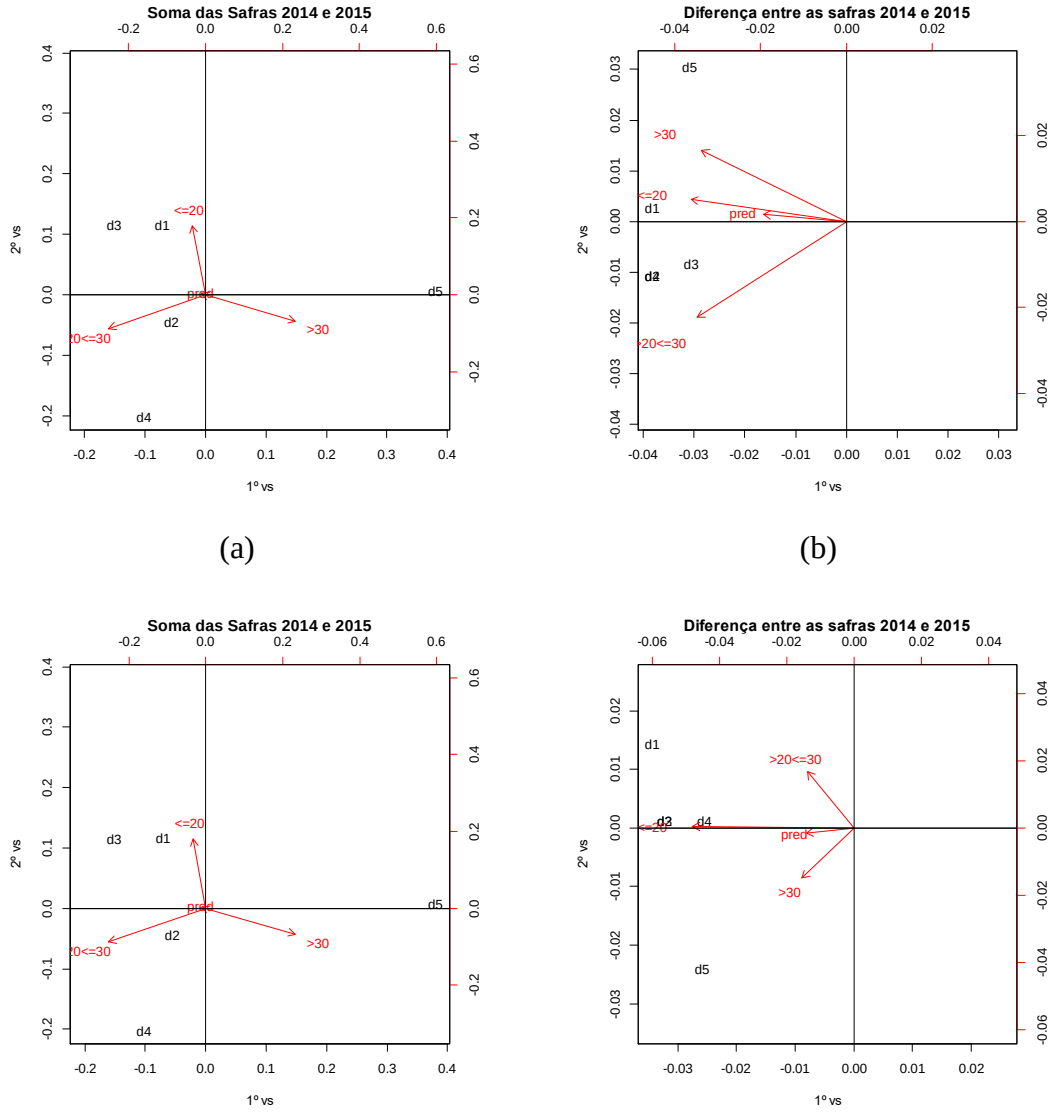
Em síntese, considerando ambos os modelos, nota-se a existência de uma fraca correlação entre as avaliações dadas na classificação dos defeitos. Tal fato sugere que não houve tendências na realização das classificações.

A relação entre os vetores, representadas pelas porcentagens de grãos de peneiras nos biplots construídos em função dos parâmetros de escala ($\delta = 0; 0,5$ e 1), em ambos os modelos, foi avaliada em função das similaridades detectadas pela proximidade das coordenadas dos tipos de defeitos descritos na Tabela 5.

Tendo por base os valores preditos pelos modelos GEE ajustados, a associação das avaliações dadas nas porcentagens de peneira em relação aos defeitos foi determinada por meio dos biplots. As coordenadas destes foram obtidas por meio da decomposição dos valores singulares (GREENACRE, 2003) aplicadas à combinação de tabelas de frequência, nesse

caso, caracterizada pelas estruturas individuais a cada safra. Assim, diferentes biplots foram obtidos nas componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$, discutidos a seguir.

Figura 4 - Biplots centrados nos valores preditos do modelo GEE, $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$, com função de ligação logit (a) e (b) e cloglog (c) e (d) considerando δ=0,5.



Fonte: Do autor (2019).

Em termos práticos, convém ressaltar que o maior interesse na interpretação dos vetores é focado no 1º e 3º quadrante, uma vez que, em todos os eixos, temos os escores positivos e negativos, respectivamente. Assim favorece indicar interpretações que sugerem uma discriminação entre um aspecto positivo ou negativo das associações.

As coordenadas necessárias para a composição dos biplots encontram-se no Apêndice C. Mediante ao exposto, observou-se que, ao considerar o efeito aditivo das safras, $S_{2014}+S_{2015}$

(Figura 4 (a)), percebe-se que o defeito d2 (PVA) está caracterizado na classificação obtida a $20\% \leq p \leq 30\%$, indicando do ponto de vista prático uma maior incidência de grãos pretos, verdes e ardidos. Tal resultado também é confirmado ao ajustar o modelo GEE considerando a função de ligação cloglog (Figura 4 (c)).

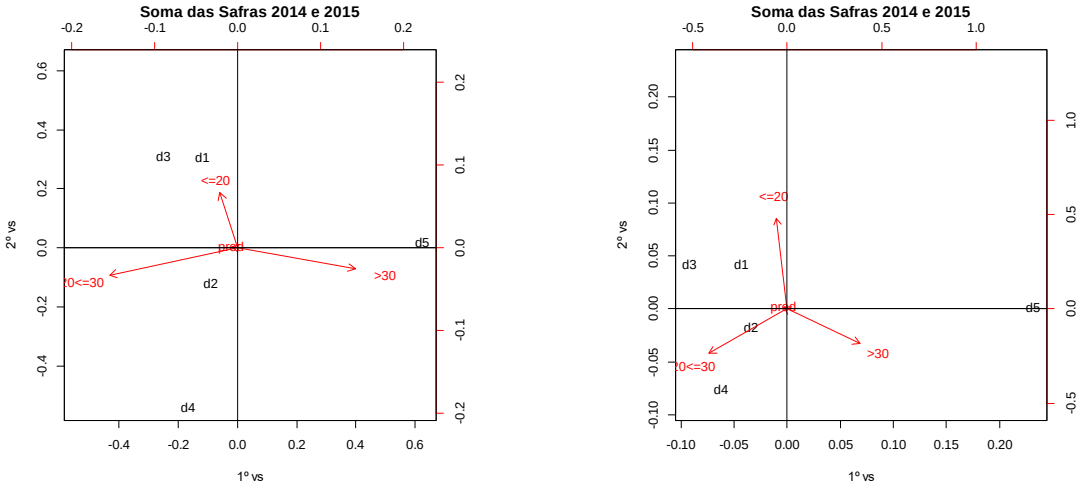
Diante da mesma situação, Brighentti e Cirillo (2018) utilizaram o modelo log-linear hierárquico com estrutura de correlação independente e verificaram que o defeito d2 (PVA) estava associado às classificações de peneira $p < 20\%$ e $20\% \leq p \leq 30\%$. Confrontando esses resultados, há evidências de que uma modelagem estatística que considere uma estrutura de correlação entre as unidades experimentais é preferível, decorrente do fato de que a classificação sugerida na escala de defeitos possa apresentar alguma correlação nas avaliações, no sentido de que algum grão possa ocasionar dúvida em relação a sua classificação.

Assim, em uma escala mais discriminada, Costa et al. (2018), ao utilizarem a análise de correspondência simples com a incorporação dos resíduos de Pearson, concluíram a mesma associação, considerando os grãos verdes em uma proporção de defeitos na amostra próximo a 20%. Segundo Carvalho et al. (1970) os grãos PVA são considerados os mais indesejáveis e tidos como o pior defeito, em que o preto é o pior, seguido dos verdes e ardidos. O valor destes, apesar de ser considerado o maior dos defeitos, tem sua porcentagem de defeitos equivalente ao período das colheitas, apresentando maior incidência de verdes no início e de pretos no final.

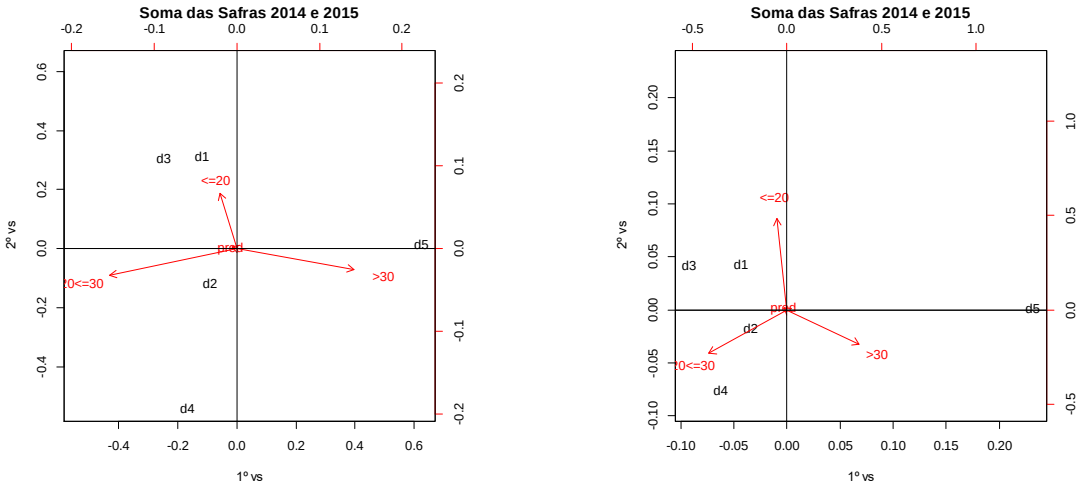
Em relação ao efeito da diferença das safras, S_{2014}-S_{2015}, os resultados das associações dos defeitos em relação a contagens em peneiras foram distintos quando utilizados modelos com diferentes funções de ligação (Figuras 4(b) e 4(d)). De acordo com essas características, ressalta-se que os biplots com efeito da soma das safras (S_{2014}+S_{2015}) são recomendados, por manterem o mesmo comportamento, frente às duas especificações do modelo.

Em se tratando dos tipos de biplots a serem utilizados na construção dos gráficos obteve-se também o GH-biplot e o RMP-biplot referindo-se respectivamente a δ=0 e δ=1, conforme a definição na página 16 na Tabela 2. Logo, os gráficos são ilustrados a seguir.

Figura 5 - Biplots centrados nos valores preditos considerando o modelo GEE com função de ligação logit (a) e (b) e cloglog (c) e (d) para a componente S_{2014}+S_{2015} considerando δ=0 e δ=1, respectivamente em ambos os modelos.



(a) (b)



(c) (d)

Fonte: Do autor (2019).

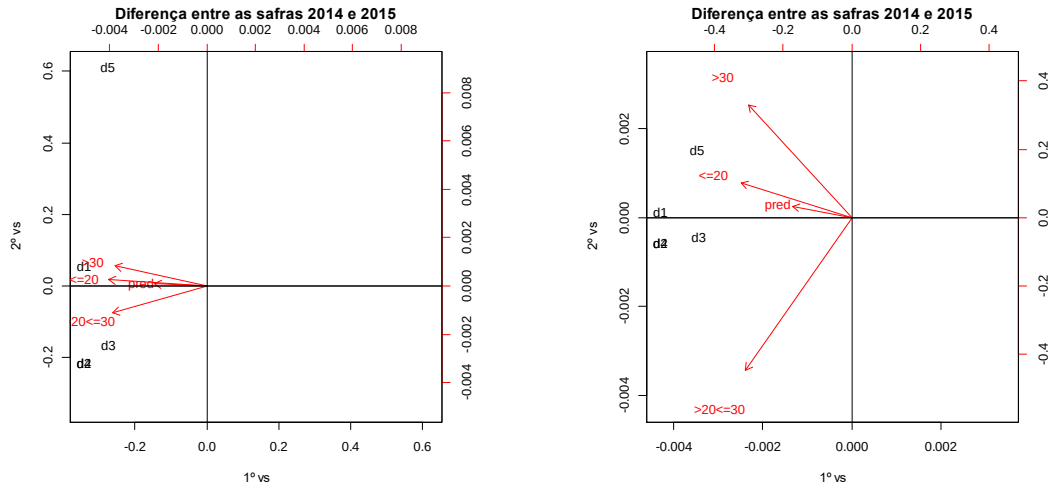
Pelos resultados ilustrados nas Figuras 5(b) e 5(d), em ambas as funções de ligação especificadas no modelo, nota-se que no RMP-biplot, caracterizado por priorizar a máxima qualidade de representação para as linhas, evidencia-se que o defeito d4 (Conchas) está associado à proporção de 20% ≤ p ≤ 30%.

Nesse caso, os defeitos d2(PVA) e d4(Conchas) estão mais associados à porcentagem de peneiras 20% ≤ p ≤ 30%. De um modo geral, os resultados do RMP-biplot podem ser mais informativos por sugerir uma classificação mais básica, tendo em vista que os defeitos d2(PVA) e d4(Conchas) podem combinar em uma classe de defeitos intrínsecos, ou seja, os

defeitos para essa peneira são atribuídos às imperfeições dos próprios grãos, enquanto, outros defeitos são denominados por extrínsecos, ou seja, caracterizado pela presença de impurezas (SILVA et al., 2015).

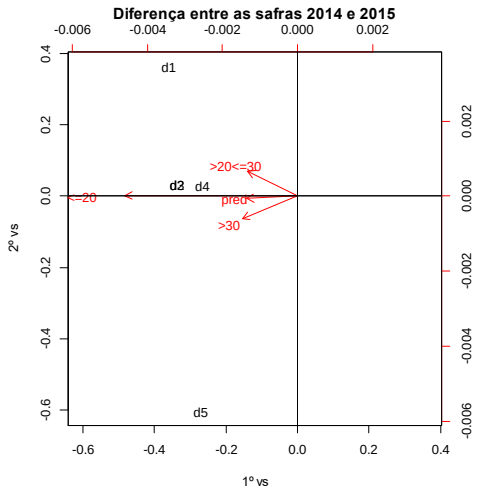
Em se tratando dos GH-biplots (δ=0) e RMP-biplots (δ =1) obtidos para a componente S_{2014}-S_{2015}, os resultados são dados a seguir.

Figura 6 - Biplots centrados nos valores preditos considerando o modelo GEE com função de ligação logit (a) e (b) e clogog (c) e (d) para a componente S_{2014}-S_{2015} considerando δ=0 e δ=1, respectivamente em ambos os modelos.

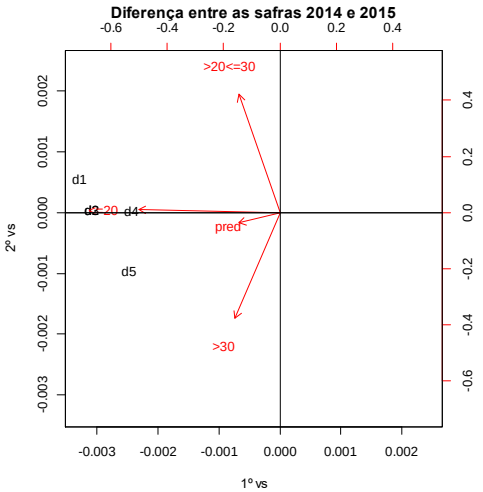


(a)

(b)



(c)



(d)

Fonte: Do autor (2019).

Para essas componentes, ao comparar-se com a Figura 4(b), nota-se que as variações representadas por δ=0 e δ=1 resultaram em biplots que deturparam a interpretação, de modo que, ao considerar δ=0 os vetores representativos indicaram correlação entre as porcentagens

de peneira, confundindo as associações com os defeitos. No caso de δ=1 os resultados não foram informativos, de modo a sugerir novas classificações.

Em relação à composição das somas de quadrados, os resultados descritos na Tabela 11 referem-se aos valores singulares obtidos para as componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$, com ressalva de que λ_4 foi extraído considerando o vetor dos valores preditos no qual o centroide dos biplots foi centrado.

Tabela 11 - Valores singulares utilizados para estimação das somas dos quadrados das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$ em ambos os modelos.

	Valores Singulares									
	$S_{2014}+S_{2015}$					$S_{2014}-S_{2015}$				
	λ_1	λ_2	λ_3	λ_4	SQ	λ_5	λ_6	λ_7	λ_8	SQ
Modelo Logit	1,1689	0,7143	0,2827	0,1054	1,9676	0,1118	0,0500	0,0300	0,0030	0,0159
Modelo Cloglog	1,1688	0,7162	0,3214	0,1072	1,9938	0,0959	0,0400	0,0245	0,0045	0,0114

Fonte: Do autor (2019).

Em razão das similaridades entre as estimativas das somas de quadrados nota-se uma confirmação em relação aos resultados anteriores de que ambos os modelos podem ser utilizados junto aos biplots corrigidos pelas suas predições.

5. CONCLUSÃO

Portanto, por meio deste trabalho, ao considerar os efeitos de associação e diferença entre safras, verificou-se que no estudo de simulação o aumento do grau de correlação resultou na redução das estimativas das somas de quadrados das componentes $S_{2014}+S_{2015}$ e $S_{2014}-S_{2015}$, em todos os tamanhos amostrais para os modelos logit e cloglog.

Em relação à aplicação decorrente da componente soma $S_{2014}+S_{2015}$, observou-se que todos os biplots corrigidos pelas predições dos modelos GEE conduziram a identificar o defeito de grãos PVA associados a porcentagem de peneiras entre 20% e 30% condizendo com a literatura existente. No caso da componente $S_{2014}-S_{2015}$ os biplots apresentaram interpretações diferenciadas, logo não se obtiveram resultados conclusivos.

REFERÊNCIAS BIBLIOGRÁFICAS

BARNETT, A. G. et al. Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. **Methods in Ecology and Evolution**, [S.l.], v.1, p.15–24, 2010.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Instrução Normativa n. 8, de 11 de junho de 2003. **Diário Oficial da República Federativa do Brasil**, Brasília, DF, 20 ago. 2003. Seção 1, p. 22-29.

BRIGHENTI, C. R. G.; CIRILLO, M. A. Analysis of defects in coffee beans compared to biplots for simultaneous tables. **Revista Ciência Agronômica**, Fortaleza, v. 49, n. 1, p. 62-69, jan/mar. 2018.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 6. ed. São Paulo: Saraiva, 2010.

CARVALHO, A. et al. Ocorrência dos principais defeitos do café em várias fases de maturação dos frutos. **Bragantia**, Campinas, v.29, n.20, p. 207-220, jun. 1970.

CENTRO DE COMERCIO DE CAFÉ DO ESTADO DE MINAS GERAIS. **Manual de Classificação**: **Métodos de Classificação de Café utilizados pelo CCCMG.** Varginha, 2018.

CIRILLO, M.A.; RAMOS, P. S. Goodness-of-fit tests for modified multinomial logit models. **Chilean Journal of Statistics**, [S.l.], v.5, n.1, p.73-85, abr. 2014.

CLASSIFICAÇÃO quanto a peneira. **Coffee & Joy**, [S.l.], 2018. Disponível em: < http://sobrecafe.coffeeandjoy.com.br/article/classificacao-quanto-a-peneira/>. Acesso em: 15 nov. 2018.

COSTA, A. L.; BRIGHENTI, C. R. G.; CIRILLO, M. A. A new approach to simple correspondence analysis with emphasis on the violation of the independence assumption of the levels of categorical variables, **Acta Scientiarum. Technology**, Maringá, v.40, 2018.

CUSTODIO, A. A. P.; GOMES, N. M.; LIMA, L. A. Efeito da irrigação sobre a classificação do café. **Engenharia Agrícola**, Jaboticabal, v.27, n.3, p.691-701, set/dez. 2007

ESQUIVEL, P.; JIMÉNEZ, V. M. Functional properties of coffee and coffee by-products. **Food Research International**, [S.l.], v. 46, n. 2, p. 488–495, maio. 2012.

GABRIEL, K.R. The biplot graphic display of matrices with application to principal componente analysis. **Biometrika**, [S.l.], v.58, n.3, p.453-467, 1971.

GREENACRE, M. Singular value decomposition of matched matrices. **Journal of Applied Statistics**, [S.l.], v. 30, n.10, 1101-1113, dec. 2003.

HARDIN, J. W.; HILBE, J. M. **Generalized Estimating Equations**. 6. ed. Boca Raton: Chapman & Hall/CRC, 2003.

JOLLIFFE, I.T. **Principal Component Analysis**. 2. ed. New York: Springer, 2002.

LAVIOLA, B. G. et al. Influência da adubação na formação de grãos mocas e no tamanho dos grãos de café (Coffea arábica L.). **Coffee Science**, Lavras, v. 1, n. 1, p. 36-42, abr/jun. 2006.

LIANG, K. Y.; ZEGER, S. L. Longitudinal Data Analysis Using Generalized Linear Models. **Biometrika**, [S.l.], v. 73, n. 1, p. 13-22, abr. 1986.

MATIELLO, J. B. Boa qualidade do café nessa safra sinaliza perda de produtividade na próxima. **Café Point**, 01 ago. 2018. Disponível em: <https://www.cafepoint.com.br/noticias/tecnicas-de-producao/boa-qualidade-do-cafe-nessa-safra-sinaliza-perda-de-produtividade-na-proxima-209518/>. Acesso em: 14 nov. 2018.

NASSER, P. P. **Influência da separação de grãos de café (Coffea arábica L.) por tamanho na qualidade e ocorrência de ocratoxina A**. 2001. 128 p. Dissertação (Mestrado em Ciência dos Alimentos)-Universidade Federal de Lavras, Lavras, 2001.

PRENTICE, R. L.; ZHAO, L. P. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. **Biometrics**, [S.l.], v. 47, n.3, p. 825-839, set. 1991.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2016. ISBN 3-900051-07-0. Disponível em: <http://www.R-project. org>.

RAJAGOPALAN, V.; VIJAYASANKAR, M.; LAKSHMI, S. Generalized Estimating Equations for Alternating Logistic Regression Model for Analysis of the CKD Patients In Type-2 Diabetes**. Indo American Journal of Pharmaceutical Sciences**, [S.l.], v. 2, n. 12, p. 1665-1672, 2015.

REDE BAHIA . Conheça o processo de classificação e beneficiamento dos grãos de café. **G1**, Vitória da Conquista, 20 fev. 2019. Disponível em: <http://g1.globo.com/bahia/bahia-agora/videos/t/edicoes/v/conheca-o-processo-de-classificacao-e-beneficiamento-dos-graos-de-cafe/3643058/>. Acesso em: 15 mar. 2019.

REZENDE, J. E. Série Tecnológica Cafeicultura: Defeitos do café. **Emater-MG**, Reduto, 2015.

ROCHA, A. A. Uma safra de café que ficará na memória. **Valor Econômico**, São Paulo, 21 ago. 2018. Disponível em: < https://www.valor.com.br/agro/5753819/uma-safra-de-cafe-que-ficara-na-memoria >. Acesso em: 15 nov. 2018.

TORRES-SALINAS, D. et al. On the use of biplot analysis for multivariate bibliometric and scientific indicators. **Journal of the American Society for Information Science and Technology**, [S.l.], v. 64, n. 7, p. 1468-1479, 2013.

SANTOS, F. L.; NANTES, J. F. D. Coordenação no mercado do café brasileiro: o desserviço da classificação por defeitos. **Gestão e Produção**, São Carlos, v. 21, n. 3, p. 586-599, jul./set. 2014.

SILVA, J. A. D.; CIRILLO, M. A**.** Selection criterion of work matrix as a function of limiting estimates of the covariance matrix of correlated data in GEE. **Biometrical Journal**, [S.l.], v.60, n. 5, p. 979-990, jul. 2018.

SILVA, J. A. **Equações de estimações generalizadas para dados ordinais em análise sensorial de cafés especiais e critérios de seleção para matrizes de correlação de trabalho**. 2017. 94 p. Tese (Doutorado em Estatística e Experimentação Agropecuária)-Universidade Federal de Lavras, Lavras, 2017.

SILVA, L. C.; MORELI, A. P.; JOAQUIM, T. N. M. Café: beneficiamento e industrialização. **Embrapa, Rondônia-Capítulo em livro científico (ALICE)**, 2015.

STONER, J. A.; LEROUX, B. G. Analysis of clustered data: A combined estimating equations approach. **Biometrika**, [S.l.], v.89, n.3, p.567-578, ago. 2002.

SUTRADHAR, B. C.; DAS, K. On the accuracy of efficiency of estimating equation approach. **Biometrics**, [S.l.], v.56, n.2, p.622-625, maio. 2000.

WEDDERBURN, R. W. M. Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. **Biometrika**, [S.l.], v.61, n.3, p.439-447, dec. 1974.

ZEGER, S. L.; LIANG, K. Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. **Biometrics**, [S.l.], v. 42, p.121-130, mar. 1986.

APÊNDICE A – Código da Simulação Monte Carlo para ajuste dos Modelos GEE com respostas Binomiais

```
options(show.error.messages = FALSE)
source("exec_biplot.txt")
source("Part_matriz.txt")
require (bindata)
require (gee)
dados <- read.table("c3.txt",h=T)
attach(dados)
n=150
rho=0.8
prop_par <- c(0.5,0.5,0.5,0.5,0.5)
### prop_par indica os valores parametricos, da binomial para cada linha
gerabin <- function(n,rho,prop)
{ trials <- rmvbin(n, prop, bincorr=(1-rho)*diag(5)+rho)
freq=colSums(trials)
return(freq) }
sim_mod=function(dados,rho,n)
{namos=n
r=rho
# ######### par: Vetor paramétrico ######### #
# ######### dados: matriz de delineamento ## #
ch_at <- matrix(0,5,1)
grupo <-as.factor(G)
 linha <-as.factor(L)
coluna <-as.factor(C)
delin=cbind(grupo,linha,coluna)
# ##### gerando amostras dependente entre as linhas # ##### #
for (i in 1:6)
{ # ######### Atualizar n ######## #
ch <- gerabin (namos,r,prop_par)
aux_ch <- as.matrix(ch)
ch_at <- cbind(ch_at,aux_ch) }
ch_at <- ch_at[,2:7]
yaux=matrix(0,1,6)
for (i in 1:5)
{ a=t(ch_at[i,])
yaux=cbind(a,yaux)}
yaux=t(yaux) ; arq <- as.matrix(dados)
dad=cbind(arq,yaux[1:30,1])
dad=cbind(dad,n)
resp <- cbind(dad[,4],dad[,5])
```

```
ajust<- gee(cbind(dad[,4],dad[,5]-dad[,4])~ 1 + grupo + linha + coluna,id=G,family=binomial(link=logit),corstr
= "exchangeable")
alfag <- ajust$working.correlation[2,1]
beta <- coef(ajust)
pred <- fitted.values(ajust)
return(list(dados=dad,est_beta=beta,ajust=pred,alfagee=alfag)) }

# ############### Inicio do Programa ################## #

nsim=500  ; alfa=1
estcoef=matrix(0,nsim,8)
tamanho=matrix(0,nsim,2)
resavs=matrix(0,1,3)
resavd=matrix(0,1,3)

# ##### Definição das matrizes para armazenar resultados ##### #

distcols=distcold=matrix(0,3,4)  # ### dist. das coordenadas de coluna ### #
distlins=distlind=matrix(0,5,4)  # ### dist. das cooordenadas de linha ### #
resd=matrix(0,nsim,2)            # res. da diferença #
estcoef=matrix(0,nsim,8)         # ### estimativa dos coeficientes ##### #
#tamanho=matrix(0,nsim,2)
resavs=matrix(0,1,3)             # ### Resultados autovalores - soma ### #
resavd=matrix(0,1,3)             # ### Resultados autovalores - dif #### #
for (h in 1:nsim)
{ expcol=matrix(c(h,h,h),3,1)
explin=matrix(c(h,h,h,h,h),5,1)
varcol=matrix(c(1,2,3),3,1)
varlin=matrix(c(1,2,3,4,5),5,1)
ch_sim_mod=sim_mod(dados,rho,n)
estcoef[h,1:8]=ch_sim_mod$est_beta   ####
resd[h,1]=ch_sim_mod$alfagee
dados_pred <- cbind(dados,ch_sim_mod$ajust)
ch_mat_part <- mat_part(dados_pred)
res = ch_mat_part    #####
ch_info
infobiplot(alfa,ch_mat_part$somau,ch_mat_part$avs,ch_mat_part$somav,ch_mat_part$difu,ch_mat_part$difv,
ch_mat_part$avd)
resavs=rbind(ch_mat_part$avs,resavs)
resavd=rbind(ch_mat_part$avd,resavd)

##### coordenadas da soma A+B #####

mccs=cbind(ch_info$ccs,expcol,varcol)
```

```
distcols <- rbind(distcols,mccs)
coordsc<-distcols[-c(1,2,3),]
mcls=cbind(ch_info$cls,explin,varlin)
distlins <- rbind(distlins,mcls)
coordsl<-distlins[-c(1,2,3,4,5),]

# coordenadas da diferenca A-B

mccd=cbind(ch_info$ccd,expcol,varcol)
distcold <- rbind(distcold,mccd)
coorddc<-distcold[-c(1,2,3),]
mcld=cbind(ch_info$cld,explin,varlin)
distlind <- rbind(distlind,mcld)
coorddl<-distlind[-c(1,2,3,4,5),] }

# coordenadas medias da soma A+B

z1=as.matrix(coordsc)
var1sc <- coordsc[coordsc[,4]==1,]
var2sc <- coordsc[coordsc[,4]==2,]
var3sc <- coordsc[coordsc[,4]==3,]
scv1=apply(var1sc,2,mean)
scv2=apply(var2sc,2,mean)
scv3=apply(var3sc,2,mean)
sc=rbind(scv1,scv2,scv3)
c1c2sc<-sc[,-c(3,4)]
z2=as.matrix(coordsl)
var1sl <- coordsl[coordsl[,4]==1,]
var2sl <- coordsl[coordsl[,4]==2,]
var3sl <- coordsl[coordsl[,4]==3,]
var4sl <- coordsl[coordsl[,4]==4,]
var5sl <- coordsl[coordsl[,4]==5,]
slv1=apply(var1sl,2,mean)
slv2=apply(var2sl,2,mean)
slv3=apply(var3sl,2,mean)
slv4=apply(var4sl,2,mean)
slv5=apply(var5sl,2,mean)
sl=rbind(slv1,slv2,slv3,slv4,slv5)
c1c2sl<-sl[,-c(3,4)]

# coordenadas medias da diferenca A-B

z3=as.matrix(coorddc)
```

```
var1dc <- coorddc[coorddc[,4]==1,]
var2dc <- coorddc[coorddc[,4]==2,]
var3dc <- coorddc[coorddc[,4]==3,]
dcv1=apply(var1dc,2,mean)
dcv2=apply(var2dc,2,mean)
dcv3=apply(var3dc,2,mean)
dc=rbind(dcv1,dcv2,dcv3)
c1c2dc<-dc[,-c(3,4)]
z4=as.matrix(coorddl)
var1dl <- coorddl[coorddl[,4]==1,]
var2dl <- coorddl[coorddl[,4]==2,]
var3dl <- coorddl[coorddl[,4]==3,]
var4dl <- coorddl[coorddl[,4]==4,]
var5dl <- coorddl[coorddl[,4]==5,]
dlv1=apply(var1dl,2,mean)
dlv2=apply(var2dl,2,mean)
dlv3=apply(var3dl,2,mean)
dlv4=apply(var4dl,2,mean)
dlv5=apply(var5dl,2,mean)
dl=rbind(dlv1,dlv2,dlv3,dlv4,dlv5)
c1c2dl<-dl[,-c(3,4)]

# ######### Construção dos Biplots ###################### #

par(mfrow = c(1, 2))
rownames(c1c2sl) <- c("V1","V2","V3","V4","V5")
rownames(c1c2sc) <-c("T1","T2","T3")
biplot(c1c2sl,c1c2sc,var.axes = TRUE, xlab = "1º axis",ylab ="2º axis",main="A1+A2",col="black")
abline(0,0,0,0)
rownames(c1c2dl) <- c("V1","V2","V3","V4","V5")
rownames(c1c2dc) <-c("T1","T2","T3")
biplot(c1c2dl,c1c2dc,var.axes = TRUE, xlab = "1º axis",ylab ="2º axis",main="A1-A2",col="black")
abline(0,0,0,0)
matrix(ch_sim_mod$dados[,4],5,6)  #exemplo de tabela gerada sem correlação
ch_sim_mod ; ch_info ; ch_mat_part
alfa_med=round(mean(resd[,1]),digits = 3)
mresavs=round(apply(resavs,2,mean),digits=4)
mresavd=round(apply(resavd,2,mean),digits=4)
sqsoma=round((sum((mresavs)^2)),digits = 4)
sqmedia=round(sqsoma/3,digits = 4)
sqdif=round((sum((mresavd)^2)),digits = 4)
sqtotal=round((sum((mresavs)^2))+sum((mresavd)^2),digits = 4)
alfa_med
```

APÊNDICE B – Código da Aplicação da técnica de biplots na granulometria de cafés

```
dados <-read.table("safrajunto.txt",h=T)
attach(dados)
source ("dvs.txt")
source ("biplot.txt")

# ##### programa Biplot ###### #

# prepara os dados para o ajuste da binomial #
dad=dados
tot_lin=as.matrix(apply(dad[,2:4],1,sum))
dif_lin= as.matrix(tot_lin-dad[2,4])
dadcomb=cbind(dad,tot_lin,dif_lin)
# ajusta o modelo para a proporção das variáveis no sentido "linha" #
mod=gee(cbind(dadcomb[,6],dadcomb[,5])~-1 + safra + menor + medio+ maior + safra*menor + safra*medio
+ safra*maior,family=binomial(link="logit"))
# ### incorpora os valores ajustados pelo modelo no biplot # ###
dad_biplot <- cbind(dad[,1],dadcomb[2:4],mod$fitted.values)
n1=5 # ### Tamanho de G1 # ###
p=5  # ### número de variáveis #### #

# ######## Executa Funções ######## #

ch <- mat_part(n1,4,dad_biplot)
ch_biplot <- infobiplot(0.5,ch$somau,ch$avs,ch$somav,ch$difu,ch$difv,ch$avd)

# ###### Gráficos Biplots ######### #

par(mfrow = c(1, 2))
rownames(ch_biplot$cls) <- c("d1","d2","d3","d4","d5")
rownames(ch_biplot$ccs) <-c("<=20",">20<=30",">30","pred")
biplot(ch_biplot$cls,ch_biplot$ccs,var.axes = TRUE, xlab = "1º vs",ylab ="2º vs",main="Soma das Safras 2014
e 2015")
abline(0,0,0,0)
rownames(ch_biplot$cld) <- c("d1","d2","d3","d4","d5")
rownames(ch_biplot$ccd) <- c("<=20",">20<=30",">30","pred")
vs",main="Diferença entre as safras 2014 e 2015")
abline(0,0,0,0)
```

APÊNDICE C – Vetores Singulares

Tabela 11 - Vetores singulares à esquerda e à direita das somas e diferenças entre as safras logit.

Vetores singulares à esquerda					
$S_{2014}+S_{2015}$	d1	-0,1166	0,3135	-0,4147	0,3407
	d2	-0,0911	-0,1153	0,4877	0,3749
	d3	-0,2479	0,3156	0,1743	-0,4566
	d4	-0,1676	-0,5369	-0,2446	-0,1544
	d5	0,6233	0,0230	-0,0027	-0,1046
$S_{2014}-S_{2015}$	d1	-0,3418	0,0592	0,4583	0,2507
	d2	-0,3411	-0,2131	-0,2885	0,4386
	d3	-0,2729	-0,1627	0,3114	-0,3350
	d4	-0,3413	-0,2129	-0,2892	-0,3544
	d5	-0,2753	0,6158	-0,1617	-0,0833
Vetores singulares à direita					
$S_{2014}-S_{2015}$	p<20%	-0,0700	0,6002	-0,3466	
	20%≤p≤30%	-0,5160	-0,2947	-0,3633	
	p>30%	0,4779	-0,2296	-0,4522	
$S_{2014}-S_{2015}$	p<20%	-0,4033	0,1268	0,5610	
	20%≤p≤30%	-0,3865	-0,5578	-0,1672	
	p>30%	-0,3762	0,4136	-0,3907	

Tabela 12 - Vetores singulares à esquerda e à direita das somas e diferenças entre as safras cloglog.

Vetores singulares à esquerda					
$S_{2014}+S_{2015}$	d1	-0,1164	0,3171	-0,1222	0,5205
	d2	-0,0912	-0,1140	-0,5517	-0,2726
	d3	-0,2477	0,3122	0,3351	-0,3590
	d4	-0,1680	-0,5371	0,2436	0,1547
	d5	0,6233	0,0218	0,0953	-0,0437
$S_{2014}-S_{2015}$	d1	-0,3596	0,3644	0,4138	-0,2375
	d2	-0,3371	0,0348	-0,3635	-0,2550
	d3	-0,3373	0,0345	-0,3611	0,1416
	d4	-0,2656	0,0314	0,1406	0,5867
	d5	-0,2698	-0,6032	0,2156	-0,1195

Vetores singulares à direita				
$S_{2014}+S_{2015}$	p<20%	-0,0681	0,6035	0,3523
	20%≤p≤30%	-0,5169	-0,2898	0,3769
	p>30%	0,4773	-0,2271	0,4643
$S_{2014}-S_{2015}$	p<20%	-0,6269	0,0122	-0,3057
	20%≤p≤30%	-0,1817	0,5259	0,4269
	p>30%	-0,2008	-0,4701	0,4577

Tabela 13 - Vetores singulares à esquerda e à direita correspondentes às safras individuais de 2014 e 2015.

Vetores singulares à esquerda				
S_{2014}	d1	0,1932	0,5154	0,7757
	d2	0,1017	-0,1223	0,2024
	d3	-0,2887	0,5339	0,02167
	d4	0,6561	-0,4149	0,2816
	d5	-0,6623	-0,5121	0,5268
S_{2015}	d1	-0,1393	0,8120	-0,1817
	d2	-0,1293	0,2224	0,7815
	d3	-0,4813	0,0569	-0,0848
	d4	-0,0999	0,5365	-0,5902
	d5	0,8498	-0,0039	-0,0283
Vetores singulares à direita				
S_{2014}	p<20%	-0,4965	0,6482	-0,5774
	20%≤p≤30%	0,8096	0,1059	-0,5774
	p>30%	-0,3131	-0,7541	-0,5774
S_{2015}	p<20%	-0,1921	-0,7936	-0,5774
	20%≤p≤30%	-0,5912	0,5631	-0,5774
	p>30%	0,7833	0,2305	-0,5774