UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

Aprimoramento do poder discriminatório de funções elipsoidais modificadas por cargas fatoriais rotacionadas na formação otimizada de agrupamentos

Fabricio Alves de Almeida

Itajubá, Abril de 2021

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

Fabricio Alves de Almeida

Aprimoramento do poder discriminatório de funções elipsoidais modificadas por cargas fatoriais rotacionadas na formação otimizada de agrupamentos

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para obtenção do título de **Doutor em Ciências em Engenharia de Produção**.

Área de Concentração: Engenharia de Produção
Orientador: Prof. José Henrique de Freitas Gomes, Dr.
Co-orientador: Prof. Anderson Paulo de Paiva, Dr.

Itajubá, Abril de 2021

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

Fabricio Alves de Almeida

Aprimoramento do poder discriminatório de funções elipsoidais modificadas por cargas fatoriais rotacionadas na formação otimizada de agrupamentos

Tese aprovada por banca examinadora em 30 de abril de 2021, conferindo ao autor o título de ***Doutor em Ciências em Engenharia de Produção***.

Banca Examinadora:

Profa. Dra. Marcela Aparecida Guerreiro Machado (UNESP)

Prof. Dr. Roberto da Costa Quinino (UFMG)

Prof. Dr. Pedro Paulo Balestrassi (UNIFEI)

Prof. Dr. Antônio Fernando Branco Costa (UNIFEI)

Prof. Dr. Anderson Paulo de Paiva (Coorientador)

Prof. Dr. José Henrique de Freitas Gomes (Orientador)

Itajubá, Abril de 2021

DEDICATÓRIA

Aos meus pais, Helena e Aloísio, por acreditarem e me apoiarem nessa jornada, sendo exemplos de vida e dedicação.

AGRADECIMENTOS

Agradeço, inicialmente, a Deus pelo dom da vida e por me dar forças todos os dias.

Aos meus pais, Helena Maria Mendonça de Almeida e Aloísio Donizete de Almeida, por acreditaram e me incentivaram nessa jornada. À Fabianne e à Grécia, pelo apoio e incentivo durante essa fase, e também à Mel e à Amora.

Agradeço ao meu orientador e co-orientador, José Henrique de Freitas Gomes e Anderson Paulo de Paiva, pela amizade, ensinamentos, apoio, parceria no desenvolvimento de trabalhos científicos e confiança durante o meu doutorado.

Ao meu amigo e colega de pesquisa, Guilherme Ferreira Gomes, pelo incentivo, apoio e parceria no desenvolvimento de diversos estudos que ajudaram no meu desenvolvimento pessoal e acadêmico.

Aos professores Pedro Paulo Balestrassi, Carlos Henrique Pereira Mello, Carlos Eduardo Sanches da Silva e Rafael Coradi Leme pelos ensinamentos, orientações e contribuições para meu crescimento acadêmico.

Ao professor Jacques Miranda Filho, do Instituto Federal do Espírito Santo, pelo apoio e disponibilização dos dados e informações que auxiliaram no desenvolvimento deste trabalho.

Aos colegas e amigos da pós-graduação, em especial Vinícius Renó, Taynara Incerti, Rachel Sabioni, Gabriela Belinato, Estevão Romão, Alexandre Torres, Simone Streitenberger, Rodrigo Leite, Leandro Amorim, Mariangela Abans, Julio Cesar Mosquera, Diego Jean, Laila Alves, Mariana Bernardes, Franco Rocha, João Paulo Barbieri, João Ederson, Juliana Gaudêncio, Victor Valério, Aline Alvim, Ana Carolina, Lucas Chielli, Wesley Gabriel, Isac Areias, Gustavo Leal e Max Moreira, pela amizade, convivência e apoio.

Aos demais professores do Instituto de Engenharia de Produção e Gestão e demais professores da UNIFEI, pela formação e conhecimento adquirido.

A todos meus amigos que, direta ou indiretamente, contribuem para minha formação pessoal e profissional.

À CAPES, FAPEMIG, FAPEPE e CNPq pelo apoio financeiro.

“Verba volant, scripta manent”

Autor desconhecido

RESUMO

O advento tecnológico proporcionou a ascensão da coleta de dados em empresas, governos e diversos segmentos industriais. Nesse aspecto, técnicas que buscam realizar agrupamentos e discriminação de conglomerados são amplamente empregadas em dados que apresentam múltiplas variáveis, trazendo a necessidade de se utilizar ferramentas específicas, que contemplem a estrutura de variância-covariância existente. Com base nisso, esse trabalho apresenta uma proposta para aprimorar o poder discriminatório de regiões de confiança na formação e estimação de agrupamentos ótimos, utilizando técnicas multivariadas e experimentais para extrair informações de maneira otimizada em conjuntos de dados correlacionados. Como método multivariado exploratório, utilizou-se a análise fatorial, calibrando a rotação de cargas fatoriais através do arranjo de misturas e, em seguida, aglutinando as funções de variância total explicada pelo erro quadrático médio. A otimização dessa etapa é realizada através do algoritmo de programação quadrática sequencial. Conhecendo os escores ótimos, um arranjo fatorial multinível é formado para contemplar todas as combinações dos métodos de ligação e os tipos de análise, buscando encontrar a combinação de parâmetros que apresente a menor variabilidade e que, consequentemente, gere elipses de confiança com melhor discriminação entre os grupos. Uma estratégia para analisar os níveis de concordância e a existência de inversões na formação de clusters é proposta utilizando os indicadores de Kappa e Kendall. Motivado pela necessidade de estratégias para classificar subestações diante de fenômenos de afundamento de tensão, que causam quedas na distribuição de energia elétrica, o método foi aplicado em um conjunto de dados reais, representando os índices de qualidade de energia elétrica de subestações localizadas no sudeste do Brasil. Foram encontrados valores ótimos na rotação das cargas fatoriais e definiu-se a parametrização “Ward e análise de covariância” como as estratégias ideais para criar os clusters nesse conjunto de dados. Assim, gerou-se conglomerados de baixa variabilidade e elipses de confiança precisas para estimar os padrões de afundamentos de tensão, promovendo um melhor poder discriminatório na classificação dos clusters através das regiões de confiança. A análise confirmatória inferiu que o método de ligação “Ward” se mostrou o mais robusto para esse conjunto, mesmo sob influência de perturbações no conjunto original.

Palavras-Chaves: Elipses de confiança; Análise Fatorial; Análise de Cluster, Projeto de Experimentos; Variância; Afundamento de tensão.

ABSTRACT

The technological advent provided the rise of data collection in companies, governments and various industrial segments. In this respect, techniques that seek to perform groupings and discrimination of clusters are widely used in datasets with multiple variables, bringing the need to use specific tools, which contemplate the existing variance-covariance structure. Based on this, this work presents a proposal to improve the discriminatory power of confidence regions in the formation and estimation of optimal clusters, using multivariate and experimental techniques to extract information in an optimized way in correlated datasets. Factor analysis was used as the exploratory multivariate method, tuning the rotation for factor loads through the mixture design, and agglutinating the total variance explained functions by the mean square error afterwards. The optimization of this step is performed through the sequential quadratic programming algorithm. Knowing the optimal scores, a multilevel factorial design is formed to contemplate all combinations of the linkage methods and the types of analysis, seeking to find the parameter that presents the least variability, generating confidence ellipses with better discrimination between groups. A strategy to analyze the levels of agreement and the inversions existence in the formation of clusters is proposed using the Kappa and Kendall indicators. Motivated by the need for strategies to classify substations in the face of voltage sag phenomena, which cause faults in the distribution of electricity, the method was applied to a set of real data, representing the power quality indexes of substations located in southeastern Brazil. Optimum values were found in the factor loads rotation and the parameterization "Ward-analysis of covariance" was defined as the ideal strategies to create the clusters in this dataset. Thus, low variability clusters and precise confidence ellipses were generated to estimate the voltage sag patterns, promoting a better discriminatory power in the clusters' classification through the regions of confidence. The confirmatory analysis inferred that the "Ward" linkage proved to be the most robust method for this dataset, even under the influence of disturbances in the original data.

Keywords: Confidence ellipses; Factor analysis; Cluster Analysis, Design of Experiments; Variance; Voltage sag.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1 Contexto da pesquisa

A busca pelo aprimoramento de técnicas de análise e interpretação de dados se demonstra cada vez mais crescente, em que entidades de vários segmentos coletam e armazenam uma quantidade de informações que ultrapassa sua capacidade de processá-los [1]. Tais conjuntos se caracterizam de grande importância para setores estratégicos, visto que auxiliam a tomada de decisões baseada em dados [1]. Assim, tem-se a análise e interpretação de dados como uma estratégia fundamental na tomada de decisão, especialmente quando o desenvolvimento tecnológico promove o crescimento do volume e variedade de informações [2,3], trazendo a necessidade de técnicas estatísticas e algoritmos para essa finalidade.

Um conjunto de dados oriundo de um determinado processo ou segmento, usualmente apresenta uma estrutura com múltiplas variáveis e, segundo Ferreira [4], tais variáveis apresentam relações entre si, proporcionando uma característica multivariada ao conjunto de informações. A necessidade de compreender as relações existentes entre diversas variáveis de natureza correlacionada faz da análise multivariada um assunto intrinsecamente complexo [5], no qual permite avaliações mais informativas e robustas sobre o conjunto de dados.

Conhecendo a estrutura de variância-covariância significativa de múltiplas variáveis tem-se que, ao analisá-las de modo univariado, ou seja, separadamente, é possível encontrar resultados pouco satisfatórios [6,7] ou mesmo imprecisos, uma vez que a multicolinearidade existente no conjunto seria negligenciada [8]. Assim, a utilização de técnicas multivariadas se faz necessária para conjuntos com essas características. Entre as estratégias comumente utilizadas, destaca-se a análise de componentes principais (*PCA – Principal Component Analysis*), que foi introduzida por Pearson [9] e, posteriormente, atribuída de maneira distinta por Hotelling [10]. PCA se caracteriza como uma técnica multivariada exploratória que modela dados correlacionados a partir da estrutura de variância-covariância [4]. Além disso, essa técnica permite a redução da dimensionalidade dos dados [11,12], encontrando uma combinação linear de variáveis não correlacionadas que explica adequadamente as variáveis originais, com a menor perda de informação possível [13]. Deste modo, tem-se que os componentes principais podem ser obtidos por meio de uma diagonalização, especificamente, de matrizes simétricas semipositivas definidas [4]. A utilização dessa técnica pode ser encontrada em muitos estudos com diferentes aplicações, tais como: [14–23].

Outra estratégia multivariada exploratória que avalia a estrutura de variância-covariância dos dados é a análise fatorial (*FA – factor analysis*). Essa estratégia se difere da PCA, pois permite explicar todas covariâncias com poucas variáveis latentes (ou fatores comuns). FA pode ser explorada fazendo uso de dois tipos de extração: *i*) método da máxima verossimilhança e; *ii*) métodos dos componentes principais. Segundo Rencher [24], a FA busca reduzir a repetição de informações entre as variáveis por meio do uso de um número menor de variáveis latentes. No entanto, as cargas fatoriais estimadas pelos métodos de extração nem sempre permitem determinar os fatores, ou seja, nem sempre é possível identificar, com clareza, a qual fator uma determinada variável observável está associada. Diante dessa situação, a estratégia multivariada FA se sobressai, pois apresenta a opção de rotacionar os eixos para melhorar a explicação das informações latentes. Para Costello e Osborne [25], a finalidade da rotação de cargas do fator original se caracteriza por obter uma estrutura de dados mais simples e clara, com fácil interpretação, evitando, assim, o confundimento. Existem na literatura alguns trabalhos que fazem uso dessa estratégia, dentre os quais pode-se destacar: [26–32].

Dada a importância e amplo uso da rotação para a FA, pode-se verificar a existência de métodos de rotação que fornecem diferentes perspectivas na interpretação das cargas fatoriais. Segundo Browne [33], a rotação de uma matriz fatorial é um problema visto desde o início da FA, em que a rotação é realizada com o objetivo de minimizar um critério de simplicidade ou parcimônia. Ao rotacionar as cargas fatoriais, as mesmas tendem a reter as matrizes de correlação e a residual, além de variações e comunalidades específicas. Desta forma, esta rotação promove uma aproximação dos eixos ao maior número de pontos, associando os agrupamentos das variáveis originais a um fator. Entre os métodos de rotação ortogonal, pode-se destacar o método *quartimax*, sendo proposto, de forma independente, por quatro diferentes autores [34–37]. Essa abordagem promove a maximização da variação das cargas quadradas nas variáveis. Alguns anos depois, Kaiser [38] propôs uma nova alternativa de rotação denominada *varimax*. Este método maximiza a variação das cargas quadradas dentro dos fatores, simplificando as colunas da matriz de carregamento. Outros métodos baseados nessas abordagens têm sido propostos, conhecidos como família *orthomax* [33,39]. A rotação ortogonal pode ser representada pelo critério de rotação γ *orthomax*, variando este parâmetro entre *0* e *1*, representando um certo nível de rotação dos eixos [5]. Entre os métodos de rotação comumente utilizados, tem-se: *quartimax*, com valor de γ igual a *0*, e *varimax*, com valor de γ igual a *1*.

Apesar da técnica *varimax* ser a mais popularizada [40,41], as métricas de rotação são exploradas com diferentes técnicas em vários estudos da literatura [42–47]. Contudo Hair *et al.* [40] afirmam que não há um consenso formado sobre o melhor método de rotação a ser empregado na FA, em que a estratégia ideal a ser utilizada pode estar condicionada à característica do conjunto de dados analisados. Assim, existe uma necessidade de calibração dos parâmetros de rotação, com finalidade de aprimorar a escolha da rotação ortogonal para simplificar a intepretação das variáveis latentes e, consequentemente, promover uma melhor explicação das cargas fatoriais.

Além de estratégias para reduzir a dimensionalidade e melhorar a explicação dos dados, outra técnica multivariada comumente explorada é a análise de agrupamento, também denominada como análise de cluster. Essa estratégia faz uso de informações prévias dos dados originais para determinar e identificar uma categoria, ou classificação, significativa [1]. Essa classificação de indivíduos pode ser formada através do nível de similaridade existentes entre as variáveis em análise, em que os clusters buscam apresentar o maior nível de homogeneidade possível em relação as suas variáveis [48]. Assim, a finalidade dessa técnica, segundo Hair *et al.* [40], se dá por particionar *n* objetos em dois ou mais clusters, baseado no nível de similaridade, comentado anteriormente. Essa similaridade entre objetos está atribuída às características determinadas para esse agrupamento, chamada de "variável estatística de cluster" [40], fazendo dela uma medida empírica de correspondência. Os autores destacam também que um problema de cluster baseado em proximidade deve contemplar os tipos de medidas ou métricas de dissimilaridade, ou distância, transformando-as em similaridade. Entre elas, a distância Euclidiana se destaca por ser a mais utilizada ao avaliar vetores de informações independentes. Complementarmente, a distância de Mahalanobis é a métrica de distância utilizada quando há correlações significativas entre as variáveis. No caso de uso de escores de fator extraídos pelo método de componentes principais, a distância de Mahalanobis não se faz necessária, já que uma das características dessa técnica exploratória se dá por criar vetores de variáveis adimensionais e independentes. Em relação aos escores de fator na análise de cluster, Hair *et al.* [40] inferem que, apesar da FA criar grupos baseados na multicolinearidade das respostas em análise, as variáveis com impacto discriminantes nem sempre representam bem as soluções de FA, sendo esta discussão amplamente difundida na área acadêmica [40].

A etapa seguinte da análise de cluster se caracteriza pelos procedimentos de partição para formação dos agrupamentos. Esta técnica é caracterizada como uma estratégia de mineração de dados (*data mining*) [49], em que a escolha desse procedimento não é trivial, visto que existem

diferentes algoritmos disponíveis, baseados em critérios distintos [40]. Contudo, a essência desse método se dá pela maximização da variabilidade entre os clusters e, consequentemente, a minimização da variabilidade dentro dos clusters. Entre os inúmeros métodos de agrupamento, destacam-se o procedimento hierárquico e o não hierárquico [40,50]. O hierárquico, ou de amalgamação, busca classificar os indivíduos em determinados clusters através de uma "árvore de classificação", ou dendrogramas [4,51]. Dos algoritmos que contemplam a estratégia hierárquica, pode-se destacar os métodos de ligação: "Único", "Centroide", "Completa", "Média", "Mediana", "McQuitty" e "Ward", sendo estes amplamente empregados na literatura [52–59]. Já o procedimento não-hierárquico se caracteriza por realizar agrupamentos a partir da especificação da quantidade de clusters. Os algoritmos comumente utilizados nesse tipo de procedimento são chamados de "*k-médias*" [40].

Muitas vezes, a escolha desses métodos é feita de forma arbitrária, em que os autores escolhem um determinado método para trabalhar. Entretanto, é possível encontrar fontes de variações e erros na formulação de clusters, uma vez que essas técnicas são sensíveis a *outliers* [5]. Além disso, Pinel [60] afirma que o melhor método de agrupamento depende do conjunto de dados e da aplicação que o mesmo está vinculado. Nesse sentido, a configuração das técnicas utilizadas para gerar clusters deve ser examinada de forma detalhada e cuidadosa, avaliando a sensibilidade e a consistência de seus agrupamentos. Johnson e Wichern [5] afirmam que é uma boa medida aplicar vários métodos de clusters aliados a pequenas perturbações (pequenos erros) à unidade de dados, a fim de verificar se há inversões na formação dos clusters e analisar a variabilidade e concordância dos métodos para um caso particular. Deste modo, tem-se a oportunidade de criação de metodologias para análise, diagnóstico e calibração dos procedimentos de ligação.

Tais resultados de agrupamento, conhecidos como associações, podem ser avaliados confrontando-os com a variável de resposta principal ou a mais significativa. Essa avaliação pode ser conduzida através da análise de variância (ANOVA), criando intervalos de confiança (IC) para os agrupamentos, baseado na principal resposta [61]. De modo similar, a avaliação pode ser também realizada a partir da análise de covariância (ANCOVA), ajustando a resposta de interesse diante do efeito de uma segunda variável incontrolável que influencia nesse processo, sendo denominada de variável concomitante [62]. Esse ajuste tende a aprimorar a precisão dos resultados diante do seu intervalo de confiança, podendo assim favorecer na discriminação dos grupos criados. É importante ressaltar que os IC, bem como os valores médios, são influenciados pelas associações criadas pelos métodos de ligação. Assim, se faz

importante encontrar o método de ligação e de análise que produza uma melhor discriminação dos grupos, com menor variabilidade, além de apresentar estabilidade, visto que os agrupamentos são sensíveis a valores discrepantes [5].

Resultados criados a partir do uso dessas técnicas trazem a necessidade de estimar o grau de confiabilidade dos valores encontrados, em que, de acordo com Montgomery e Runger [63], essa confiabilidade deve ser especificada por um determinado nível de confiança, ou simplesmente, por intervalos de confiança. Os autores afirmam que, diante de uma amostra de uma determinada população, o IC expressa a estimativa de se encontrar um parâmetro. Assim, o comprimento desse intervalo univariado indica a precisão na estimação desse parâmetro. Em outras palavras, tem-se que um intervalo de confiança curto (ou pequeno) infere em uma estimação mais precisa [63].

Contudo, conhecendo a relação de multicolinearidade da variável principal com outras variáveis do processo (como indicado anteriormente sobre a variável concomitante), usualmente, o conjunto de variáveis correlacionadas apresenta uma estrutura geométrica elipsoidal. Johnson e Wichern [5] afirmam que, ao considerar duas variáveis x_1 e x_2, com diferentes valores de desvio padrão, sua distância estatística apresenta uma elipse como o lócus de todos os pontos a uma distância constante da origem. Esse conceito é amplamente utilizado em métodos de estimação e inferência dentro de estudos de análise multivariada [5]. Ao considerar as variâncias e covariâncias para estimar as distâncias entre os vetores aleatórios, tem-se uma distância quadrática que pode ser definida como a "distância generalizada de Mahalanobis", em que, segundo Ferreira [4], as distâncias constantes, diante de um vetor aleatório e um vetor fixado, caracterizam um elipsoide, que em uma visão bidimensional, é representado por uma elipse.

Baseado nas características existentes para dados multivariados, regiões de confiança podem ser formadas a fim de gerar elipses que representam toda a projeção de uma região, considerando um nível de confiança aceitável (como o de 95%). As elipses de confiança são capazes de representar, de maneira adequada, a separabilidade dos resultados em uma visão bidimensional, considerando não apenas a variável de interesse, mas também uma segunda variável incontrolável, que impacta consideravelmente na estimação dos resultados [62]. Esse tipo de estratégia para estimação pode ser encontrado em diferentes abordagens [64–69].

Diante do que foi discutido anteriormente, é possível verificar que as técnicas e procedimentos de estimação apresentam-se bem estruturados na literatura, sendo estratégias consolidadas e aplicadas para diversas finalidades. Além disso, tem-se que as pesquisas sobre

esses métodos são cada vez mais investigadas, a fim de consolidar e aprimorar metodologias para o advento científico e tecnológico. Contudo, é possível verificar algumas lacunas existentes, apresentando assim oportunidades de melhorias e contribuições no que se refere a aspectos de tratamento de dados multivariados. Com base nisso, o presente trabalho buscou se desenvolver com finalidade de contribuir com as carências apresentadas, motivadas pelas seguintes questões:

1) Diante da aplicação da FA, como determinar o melhor grau de rotação ortogonal para aprimorar a explicação das variáveis latentes, perante as cargas fatoriais?

2) Em relação aos procedimentos de cluster e os tipos de análise, como se deve determinar a melhor combinação de parâmetros para realizar agrupamentos com melhor discriminação dos mesmos?

3) Como representar, adequadamente, a região de confiança dos agrupamentos formados, a fim de criar uma discriminação confiável dos resultados?

4) Em relação aos resultados, como determinar se a escolha dos métodos de ligação é realmente robusta?

5) Por fim, como esses conceitos se comportam em um conjunto de dados reais, com estrutura de variância-covariância significativa?

Com finalidade de responder esses questionamentos, este trabalho considera a necessidade de se realizar discriminações de informação para a tomada de decisão, a partir de uma quantidade significativa de dados. Criando, portanto, uma metodologia que contempla diferentes aspectos da FA, dos métodos de clusters e dos tipos de análise, a fim de melhorar a discriminação de agrupamentos, gerando elipses de confiança não sobrepostas para uma região de confiança de 95%. Para isso, considerou-se a estratégia de planejamento de experimentos (*DOE – Design of Experiments*), a fim criar arranjos experimentais que contemplam as amplitudes dos parâmetros nesse problema. Para responder à questão 1, foi selecionado o arranjo de misturas do tipo *simplex-lattice*, pois promove um arranjo de proporção, visto que será abordado uma variação do valor de rotação γ entre 0 e 1. Esse DOE é aplicado para as variâncias totais explicadas das cargas fatoriais, no qual são aglutinadas em uma função objetivo única, modeladas através do Erro Quadrático Médio (EQM), no qual será, posteriormente, minimizada.

Para atender a questão 2, que trata apenas de parâmetros categóricos, selecionou-se o arranjo fatorial multiníveis, a fim de avaliar o comportamento e influência dos métodos de

ligação e do tipo de análise (ANOVA ou ANCOVA) para a formação e estimação de clusters, baseado em sua variância. No que se refere à etapa de otimização, problemas com estruturas mais complexas serão resolvidos a partir do algoritmo de programação quadrática sequencial (*SQP – Sequential quadratic programming*). Conhecendo a melhor escolha de parâmetros, é possível estimar a confiança para os valores médios encontrados. Tratando-se de dados com características multivariadas, os intervalos podem ser melhor representados por regiões de confiança, criando, assim, elipses que representam a projeção bidimensional entre a relação da variável de interesse com a variável concomitante. Considerando a combinação "método de ligação" e "tipo de análise" com menor variabilidade, tende-se a criar elipses de confiança estreitas e não sobrepostas, apresentando mais precisão e confiança na estimação dos resultados (questão 3).

Para confrontar a estabilidade dos métodos de ligação, considerando as afirmações de Johnson e Wichern [5], realizou-se um planejamento baseado no estudo de repetitividade e reprodutibilidade por atributos, também conhecida como análise de concordância por atributos. Para isso, réplicas com diferentes características precisam ser criadas, utilizando um pequeno nível de perturbação nos dados e, em seguida, aplicar a FA (com rotação otimizada). Posteriormente, os métodos de ligação (hierárquico e não hierárquico) são aplicados, armazenando suas associações para confrontar os resultados a partir das estatísticas de Kappa e do coeficiente de concordância de Kendall, proporcionando encontrar o método com maior estabilidade e robustez, consequentemente, sanando a problemática apresentada na questão 4.

Por fim, com o objetivo de demonstrar a aplicabilidade e resultados do método proposto, os aspectos teóricos foram empregados em conjunto de dados referentes ao setor de energia elétrica, mais especificamente para dados dos índices de qualidade de energia elétrica (QEE) de um conjunto de subestações localizadas no estado do Espírito Santo, sudeste do Brasil. A motivação de explorar esse conjunto se refere a uma necessidade de regulamentar os fenômenos relacionados a qualidade na distribuição de energia elétrica, em que a Agência Nacional de Energia Elétrica (ANEEL) do Brasil, busca quantificar o número de eventos de afundamento de tensão, ocasionada por variação de tensão de curta duração (VTCD), a fim de estabelecer normas e critérios para qualidade de energia elétrica [70]. A ANEEL [71] infere que a avaliação do fornecimento de energia está vinculada a qualidade de distribuição, em que a VTCD é um indicador crucial para determinar a qualidade desse serviço, visto que indústrias e segmentos similares apresentam alta vulnerabilidade a cargas sensíveis, nesse tipo de evento. É possível encontrar na literatura muitos trabalhos que fazem uso das informações de afundamentos de

tensão para análise da qualidade de energia elétrica. Santis *et al.* [72] avaliaram a origem do afundamento de tensão decorrente de falhas em redes reais e interligadas. Liao e Anani [73] utilizaram redes neurais artificiais para identificar falhas na estimativa do estado de afundamento de tensão. Branco *et al.* [74] introduziram um algoritmo evolutivo para otimizar as alocações de monitores QEE em sistemas de distribuição. Dentre vários objetivos específicos, como o monitoramento do custo, os autores enfatizam a minimização do número de incidências de afundamentos de tensão, confirmando a importância desta variável para a análise da QEE. Santos e Barros [75] propõem a predição da amplitude e duração do afundamento de tensão no planejamento da rede. Majumder *et al.* [76] propuseram uma metodologia que pode ser aplicada como solução de mitigação de afundamento de tensão para distribuição de concessionárias em um grupo de clientes, instalando um restaurador dinâmico de tensão. Moradi e Mohammadi [77] realizaram simulações em diferentes métodos para identificar fontes de afundamentos de tensão, além de proporem um método próprio utilizando informações de relés de sobrecorrente direcional. A abordagem dos autores visa auxiliar as concessionárias na operação e também no planejamento da rede. Em busca de suporte ao sistema de gerenciamento de falhas, Mokhlis e Li [78] apresentam uma aplicação para encontrar falhas na rede de distribuição baseada em perfis de afundamento de tensão não lineares e na medição de afundamento de tensão em subestação primária. Gencer *et al.* [79] utilizaram a transformada Wavelet para criar uma abordagem de detecção de eventos de afundamento de tensão, tendo em vista o aumento dos padrões de qualidade de energia. Também focado na necessidade de reconhecimento de eventos para qualidade de energia elétrica, Erişti *et al.* [80] propõem um novo sistema baseado no método de ligação não hierárquico "*k-médias*", em que os autores provam seus métodos em dados reais de eventos que impactam na QEE. Comumente importante, Sun *et al.* [81] propõem uma nova estratégia baseada em redes neurais convolucionais e no classificador de "*k-vizinho*" mais próximo ponderado para identificação de eventos de afundamento de tensão. Muitos outros estudos exploram essa temática [82–90], portanto, tem-se que o evento de afundamento de tensão se destaca como um parâmetro importante em sistemas de distribuição de energia elétrica, pois possibilita conhecer o nível de qualidade das subestações do sistema de distribuição de energia e, consequentemente, classificar os diferentes fornecedores desse sistema.

É importante destacar que a ANEEL estabelece distintos procedimentos referentes a qualidade de energia elétrica (descritos no Módulo 8 do Procedimento de Distribuição de Energia Elétrica (PRODIST), em [71]). Contudo, há uma carência de regulamentação nacional para avaliar esses fenômenos [70]. Uma proposta inicial foi explorada no estudo de Miranda *et*

al. [61], no qual os autores fazem uso da estratégia PCA e o procedimento de ligação "Ward". Contudo, os autores não avaliam o desempenho e estabilidade de outros métodos de ligação, nem mesmo a influência de variáveis concomitantes (que podem impactar diretamente na resposta principal), estimando os intervalos de confiança de modo univariado, negligenciando, assim, a estrutura de variância-covariância significativa dos dados. Além disso, sabe-se que o uso da estratégia PCA mesclada à análise de cluster deve ser cuidadosamente ajustada, visto que o PCA tende a alocar o máximo de informações possíveis no primeiro componente, fazendo com que esse vetor apresente um maior grau de importância, necessitando que os componentes sejam ajustados/ponderados previamente para o uso da análise de cluster [91]. Muitos estudos voltados, especificamente, ao setor elétrico e de energia exploram o uso de técnicas multivariadas como PCA [61,92–101] e os procedimentos de clusters [60,102–111]. Contudo, existem poucos estudos que utilizam a técnica FA para esse setor e nenhum deles propõe o uso de FA mesclado a análise de clusters para classificar e aprimorar a discriminação de conjuntos de subestações[1], inferindo uma oportunidade para a investigação dessa temática.

1.2 Objetivos

Com base na discussão apresentada anteriormente, é possível traçar o objetivo principal, aliado aos objetivos secundários, que complementam o presente estudo.

1.2.1 Objetivo geral

O objetivo principal deste trabalho é desenvolver e propor um método para o aprimoramento do poder de discriminação em regiões de confiança, através de funções elipsoidais, para formação de agrupamentos baseado em dados correlacionados. A estratégia é auxiliada pelo uso de escores de fator sob rotação ortogonal e outras estratégias, como os métodos de ligação e as variáveis concomitantes.

1.2.2 Objetivos específicos

Diante do objetivo majoritário, tem-se os seguintes objetivos específicos dessa proposta:

- Avaliar a influência da FA para reduzir a dimensionalidade dos dados e na geração de eixos independentes, agrupando adequadamente as características do conjunto de dados;

[1] É importante ressaltar que o único trabalho que faz uso dessas técnicas para a temática é, justamente, o artigo publicado por Almeida *et al*. [62], o qual é resultado direto do desenvolvimento deste trabalho.

- Desenvolver um método baseado em arranjo de misturas e no EQM para otimizar a rotação dos escores fatoriais, buscando auxiliar a tomada de decisão do nível γ de rotação *orthomax*;
- Investigar as associações de diferentes métodos de ligação para os escores de fator com rotação otimizada;
- Analisar a influência dos valores de média e desvio padrão à luz da variável principal, a fim de compreender o impacto utilizando e não utilizando a variável concomitante;
- Criar um arranjo fatorial multiníveis para investigar a influência da variabilidade na formação de clusters, utilizando os métodos de ligação aliados aos diferentes tipos de análise;
- Encontrar a parametrização ótima (rotação γ, método de ligação e tipo de análise) para agrupamentos de dados com múltiplas variáveis correlacionadas;
- Desenvolver regiões de confiança através de funções elipsoidais para aprimorar a análise discriminante de clusters em projeções bidimensionais;
- Investigar a análise de concordância por atributos para avaliar a estabilidade e robustez dos métodos de ligação com réplicas sob pequenas perturbações;
- Aplicar o método proposto para um conjunto de dados reais, a fim de avaliar o comportamento, contribuição e a adequação dos procedimentos apresentados.

Para atingir os objetivos indicados acima, essa tese fez uso de um método de pesquisa híbrido, contemplando o uso de técnicas de modelagem, simulação e experimentos no decorrer da pesquisa. Maiores detalhes sobre essa abordagem estão descritos no capítulo 3.

1.3 Contribuições esperadas

Diante das questões e propostas levantadas anteriormente, este estudo busca contribuir à metodologia de formação de agrupamentos em dados correlacionados, bem como na tomada de decisão de parâmetros de estratégias multivariadas, podendo ser devidamente combinadas para apresentar um melhor resultado. Com isso, apresenta-se regiões de confiança através de elipses, facilitando a interpretação e discriminação de grupos para posteriores análise, sob uma projeção bidimensional. A contribuição se estende a um método de confirmação para a escolha do procedimento de ligação, contemplando a lacuna teórica inferida por Johnson e Wichern [5]. Em relação a parte prática, este trabalho contribui com as diretrizes de análise e classificação da QEE para o setor energético brasileiro e para a agência reguladora ANEEL (favorecendo a Norma IEEE 1564 [112] e ao Módulo 8 do PRODIST [71]), devido às lacunas apresentadas

anteriormente. Contudo, se faz importante destacar que os procedimentos propostos podem ser aplicados em distintos conjuntos de dados, sendo uma estratégia potencial para todos segmentos que fazem uso de dados com uma estrutura de variância-covariância significativa.

1.4 Delimitações da pesquisa

No que se refere aos aspectos dessa pesquisa, tem-se que a mesma está sujeita a algumas delimitações, definidas a seguir:

- **Delimitações do método exploratório multivariado:** espera-se que o método proposto possa ser utilizado com distintos conjuntos de dados, contudo os mesmos estão sujeitos à adequação da estratégia de FA, como ao fato de que a quantidade de variáveis precisa ser $\leq$ ao número de observações, em que alguns autores [61,62] utilizam os mínimos quadrados parciais (PLS – *Partial Least-Squares*) para definir as variáveis mais significativas. Ainda sobre a estrutura dos dados, o conjunto precisa ser avaliado e estar adequado a partir do teste de esfericidade de Batlett e do índice KMO. Por fim, tem-se que a parametrização do nível γ de rotação não contempla o uso de diferentes números de fatores a serem utilizados na extração.
- **Delimitações do procedimento de agrupamento:** no que se refere ao uso e comparações de diferentes métodos de cluster, tem-se que, para os métodos de amalgamação (hierárquicos), a medida de distância será constante, não apresentando variação nessas parametrizações dentro do DOE. Em relação aos métodos de ligação, selecionou-se as opções mais comumente utilizadas na literatura. Ademais, para a estratégia de ligação não hierárquica (*k-médias*), sabe-se que o mesmo pode ser considerado como uma estratégia heurística, sendo necessário diversas replicações para avaliar e confirmar seus valores [113,114]. Contudo, neste caso particular, o mesmo será considerado como uma medida determinística, visto que a quantidade de cluster será previamente definida pela Regra de Sturges [115].
- **Delimitações dos dados reais a serem utilizados:** Os dados de subestações a serem utilizados nesse trabalho estão originalmente disponíveis em Miranda *et al.* [61], em que o número de variáveis (características de qualidade) é maior que o número de observações (subestações). Assim, a aplicação será condicionada a uma quantidade menor de variáveis, definidas previamente no estudo de Almeida *et al.* [62], contemplando aspectos técnicos e estatísticos.

1.5 Estrutura do trabalho

A estrutura do presente trabalho pode ser particionada em 6 capítulos, os quais apresentam todas informações necessárias. Neste primeiro capítulo, foi apresentada uma breve contextualização, tratando dos assuntos a serem abordados, bem como os problemas, objetivos, contribuições e delimitações vinculadas à pesquisa.

O segundo capítulo apresenta toda fundamentação teórica das estratégias empregadas, detalhando os conceitos e características dos métodos. Assim, tem-se a apresentação de diferentes tipos de análise de variabilidade, como a análise de variância, covariância e a análise de atributos. Em seguida, são apresentadas as técnicas de planejamento de experimentos, como o arranjo fatorial e a superfície de resposta, destacando sua classe especial denominada de arranjo de misturas. No que contempla as técnicas multivariadas, tem-se as estratégias de PCA, FA e a análise de cluster. A estratégia de aglutinação de funções objetivo, EQM, é apresentada, bem como o método de busca para sua minimização, o SQP. Por fim, apresentam-se as abordagens teóricas referentes a estimação dos intervalos e das regiões de confiança.

O terceiro capítulo detalha o método de aprimoramento do poder discriminatório de funções elipsoidais auxiliadas por escores de fator sob rotação ortogonal, apresentando todas as etapas do método proposto.

Em seguida, o quarto capítulo destaca o conjunto de dados utilizados para demonstrar a aplicabilidade do método proposto no capítulo 3. Esse conjunto se refere aos índices e características de qualidade de energia das subestações localizadas no Estado do Espírito Santo, detalhando todas as informações necessárias para a aplicação da abordagem proposta.

O capítulo 5 apresenta, de maneira completa, a aplicação e discussão do método proposto para os dados descritos no capítulo 4, detalhando todas as etapas e características necessárias.

Por fim, o capítulo 6 apresenta as conclusões do estudo, bem como as contribuições teóricas e práticas, além de sugestões para trabalhos a serem realizados futuramente, vinculados ao presente estudo. Os apêndices, ao final do documento, apresentam informações auxiliares e complementares ao que foi discutido ao longo do estudo.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo busca apresentar os conceitos e conhecimentos base para o desenvolvimento desta tese. Deste modo, busca-se detalhar todas as técnicas utilizadas a partir de referências clássicas e modernas da literatura. Assim, nesta etapa serão abordados: os conceitos sobre a variabilidade e técnicas para avaliá-las que compreendem a análise de variância, análise de covariância e sobre análise de concordância por atributos; o embasamento sobre análise e planejamento de experimentos, como o arranjo fatorial completo e o generalizado, além da classe especial de metodologia de superfície de resposta conhecida como arranjos de misturas; entre as estratégias de análise de dados, em especial as de natureza multivariada, serão abordadas a análise de componentes principais, análise fatorial e a análise de cluster; um breve levantamento sobre o método de otimização SQP e o método de aglutinação EQM; por fim, serão apresentados os intervalos e regiões de confiança.

2.1 Análise de variância

Sendo uma das estratégias mais utilizadas na área de inferência estatística, a ANOVA se destaca como o procedimento adequado para avaliar a hipótese de igualdade para duas ou mais médias [116]. Ao considerar níveis de um único fator na ANOVA, tem-se que o tratamento *a* é uma variável aleatória, em que, diante da Eq. (2.1), a *j-ésima* observação é realizada para o nível de fator *i*. Tal equação descreve o modelo experimental para um fator, em que μ_i e ε_{ij} representam a média do nível do fator e o componente de erro aleatório, respectivamente. De acordo com Montgomery [116], esse componente inclui todas as fontes de variabilidade (medição, fatores incontroláveis etc). Assim, para erros com média zero, $E(y_{ij}) = \mu_i$. Considerando, alternativamente que $\mu_i = \mu + \tau_i$, a Eq. (2.1) pode ser definida conforme a Eq. (2.2) [116].

$$y_{ij} = \mu_i + \varepsilon_{ij} \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., n \end{cases} \tag{2.1}$$

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \begin{cases} i = 1, 2, ..., a \\ j = 1, 2, ..., n \end{cases} \tag{2.2}$$

A partir desse novo modelo, tem-se o modelo de efeitos, em que μ e τ_i se definem como a média geral e o efeito do tratamento *i*, respectivamente. Assim, tais modelos são denominados como análise de variância [116], a qual considera um único fator.

Com base nisso, o teste de hipótese que contempla a ANOVA para um fator único pode ser descrita, conforme a Eq. (2.3).

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_a \\ H_1 : Ao\ menos\ uma\ \mu\ se\ difere \end{cases} \tag{2.3}$$

Assim, os métodos e formulações necessárias para se realizar a ANOVA para um fator único estão detalhados na Tabela 2.1, em que *MS*, *SS* e *DF* referem-se pelos quadrados médios, soma dos quadrados e graus de liberdade, respectivamente. n_T e *a* descrevem a quantidade de observações totais e a quantidade de níveis do fator, respectivamente. $\bar{y}_{i\bullet}$ e $\bar{y}_{i\bullet\bullet}$ representam a média de observações do fator para o *i-ésimo* nível e a média das observações, respectivamente. As Eqs. (2.4) e (2.5) apresentam as formulações para os valores de R^2 e R^2 *ajustado*, enquanto a Tabela 2.2 descreve os equacionamentos para o desvio padrão, desvio padrão combinado e intervalo de confiança individual.

$$1 - {SS_E}/{SS_T} \tag{2.4}$$

$$1 - {MS_E}/{SS_T / DF_T} \tag{2.5}$$

Tabela 2.1. Análise de variância para um único fator

Fonte de variação	**Soma dos Quadrados**	**Graus de Liberdade**	**Quadrado médio**	**F_0**
Entre	$n\sum_{i=1}^{a}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$	$a - 1$	${SS_{fator}}/{DF_{fator}}$	$F_0 = {MS_{fator}}/{MS_E}$
Erro (dentro)	$SS_T - SS_{fator}$	$n_T - a$	${SS_E}/{DF_E}$	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2$	$n_T - 1$		

Tabela 2.2. Formulações para os desvios e IC

	Sigla	Modelo
Desvio Padrão	S_i	$\sqrt{\dfrac{\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2}{n-1}}$
Desvio Padrão combinado	S^2	$\dfrac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2}{\sum_{i=1}^{a}(n-1)}$
Intervalo de confiança individual	IC	$\bar{y}_{i\bullet} \pm \dfrac{t_{1-\alpha/2;n_T-a}\, s}{\sqrt{n}}$

2.2 Análise de covariância

Análise de covariância é uma técnica muito eficaz que é utilizada para melhorar a precisão de um experimento removendo os efeitos de certos fatores. Esta técnica é baseada no ajuste da variável de resposta (*y*) para o efeito de uma variável concomitante (*x*) que não pode ser controlada no processo. Deste modo, tem-se que *y* está linearmente relacionado à *x* [116]. Se um experimento apresenta tais características, mas não faz uso da ANCOVA, tem-se que o erro quadrático médio pode ser inflacionado devido à ausência da variável concomitante (também conhecida como covariável). Além disso, de acordo com Montgomery [116], tem-se que a estratégia ANCOVA também é caracterizada como uma combinação da ANOVA e a análise de regressão.

Considerando a presença de apenas uma covariável, que seja linearmente relacionada à variável resposta, em um experimento de fator único, o seguinte modelo estatístico pode ser obtido, conforme descrito na Eq. (2.6).

$$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) + \varepsilon_{ij} \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \end{cases} \tag{2.6}$$

onde y_{ij} é a *j-ésima* observação na variável de resposta obtida no *i-ésimo* tratamento, x_{ij} é o valor da variável concomitante correspondente a y_{ij}, $\bar{x}_{\bullet\bullet}$ representa a média de todos os valores x_{ij}, μ é a média de todos os valores y_{ij} (média geral), τ_i indica o efeito do *i-ésimo* tratamento, β é o coeficiente de regressão que mostra a dependência linear entre y_{ij} e x_{ij}. Por fim, tem-se ε_{ij} representando o componente de erro aleatório. Os equacionamentos necessários para a análise estão descritos na Tabela 2.3, onde **S**, **T**, e **E** representam a soma dos quadrados e o produto vetorial para os valores totais, tratamentos e erros, respectivamente. Além disso, usualmente tem-se que **S** = **T** + **E** [116].

Considerando a ANCOVA para um experimento de fator único com uma única covariável, é necessário calcular o coeficiente de regressão linear, como descrito na Eq. (2.7).

$$\beta = \frac{E_{xy}}{E_{xx}} \tag{2.7}$$

A soma dos quadrados do erro, descritas na Eq. (2.8), considera $a(n-1)-1$ graus de liberdade. Consequentemente, a variância do erro experimental pode ser calculada a partir da Eq. (2.9).

Tabela 2.3. Equações da soma dos quadrados e o produto vetorial para **S**, **T** e **E**

S	***T***	***E***
$S_{yy} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\bullet\bullet})^2$	$T_{yy} = n\sum_{i=1}^{a}(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$	$E_{yy} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\bullet})^2$
$= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{\bullet\bullet}^2}{an}$	$= \frac{1}{n}\sum_{i=1}^{n} y_{i\bullet}^2 - \frac{y_{\bullet\bullet}^2}{an}$	$= S_{yy} - T_{yy}$
$S_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2$	$T_{xx} = n\sum_{i=1}^{a}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$	$E_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})^2$
$= \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - \frac{x_{\bullet\bullet}^2}{an}$	$= \frac{1}{n}\sum_{i=1}^{n} x_{i\bullet}^2 - \frac{x_{\bullet\bullet}^2}{an}$	$= S_{xx} - T_{xx}$
$S_{xy} = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})(y_{ij} - \bar{y}_{\bullet\bullet})$	$T_{xy} = n\sum_{i=1}^{a}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})$	$E_{xy} = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})(y_{ij} - \bar{y}_{i\bullet})$
$= \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}y_{ij} - \frac{(x_{\bullet\bullet})(y_{\bullet\bullet})}{an}$	$= \frac{1}{n}\sum_{i=1}^{n}(x_{i\bullet})(y_{i\bullet}) - \frac{(x_{\bullet\bullet})(y_{\bullet\bullet})}{an}$	$= S_{xy} - T_{xy}$

$$SS_E = E_{yy} - \frac{E_{xy}^{\ 2}}{E_{xx}} \tag{2.8}$$

$$MS_E = \frac{SS_E}{a(n-1)-1} \tag{2.9}$$

Considerando um modelo reduzido, ou seja, sem o efeito de tratamento, pode-se chegar à expressão descrita na Eq. (2.10). Assim, tem-se que a soma dos quadrados do erro pode ser calculada a partir da Eq. (2.11), adotando *an–2* graus de liberdade.

$$y_{ij} = \mu + \beta\left(x_{ij} - \bar{x}..\right) + \varepsilon_{ij} \tag{2.10}$$

$$SS_E^{'} = S_{yy} - \frac{S_{xy}^{\ 2}}{S_{xx}} \tag{2.11}$$

O termo $(S_{xy})^2/S_{xx}$ descrito na Eq. (2.11) indica a redução na soma dos quadrados de *y*. Deste modo, para verificar se o efeito é, ou não, significativo, tem-se a necessidade de calcular a estatística F_0, conforme descrito na Eq. (2.12). Assim, tem-se evidências para rejeitar a hipótese nula (H_0: $\tau_i = 0$) se F_0 for maior que $F_{\alpha,a-1,a(n-1)-1}$.

$$F_0 = \frac{\left(SS_E^{'} - SS_E\right)/(a-1)}{SS_E/\left[a(n-1)-1\right]} \tag{2.12}$$

Ao se utilizar a estratégia ANCOVA, tem-se a necessidade de verificar se a relação linear entre a resposta de interesse e a variável concomitante realmente existe. Para isso, um teste de hipótese pode ser realizado, a fim de verificar se o coeficiente β, que é calculado a partir da Eq. (2.7), apresenta valor igual a zero. A estatística desse teste pode ser calculada a partir da Eq. (2.13), onde a hipótese nula (H_0: $\beta = 0$) é rejeitada, caso o valor de F_0 seja maior que $F_{\alpha,1,a(n-1)-1}$.

$$F_0 = \frac{\left(E_{xy}\right)^2/E_{xx}}{MS_E} \tag{2.13}$$

Por fim, tem-se que o modelo de covariância pode ser verificado a partir da análise de resíduos, sendo os mesmos calculados a partir da diferença entre o valor real (y_{ij}) e o valor ajustado ($\hat{y}_{ij}$). Tal valor pode ser encontrado a partir da Eq. (2.14). A Tabela 2.4 apresenta a ANCOVA como uma ANOVA "ajustada", enquanto a Tabela 2.5 exemplifica a ANCOVA para um experimento de fator único com uma covariável.

$$\hat{y}_{ij} = \bar{y}_{i.} - \hat{\beta}\left(x_{ij} - \bar{x}_{i.}\right) \tag{2.14}$$

Tabela 2.4. Análise de covariância como um ajuste para o ANOVA

Fonte de Variação	**Soma dos Quadrados**	**Graus de Liberdade**	**Quadrado Médio**	**F_0**
Regressão	$\left(S_{xy}\right)^2 \Big/ S_{xx}$	*1*		
Tratamentos	$SS_E^{'} - SS_E = SS_Y - \left(S_{xy}\right)^2 / S_{xx} - \left[E_{yy} - \left(E_{xy}\right)^2 / E_{xx}\right]$	*a – 1*	$\frac{SS_E^{'} - SS_E}{a-1}$	$\frac{\left(SS_E^{'} - SS_E\right)/(a-1)}{MS_E}$
Erro	$SS_E = E_{yy} - \left(E_{xy}\right)^2 / E_{xx}$	$a(n-1) - 1$	$MS_E = \frac{SS_E}{a(n-1)-1}$	
Total	S_{yy}	*an – 1*		

Tabela 2.5. ANCOVA para um experimento de fator único com uma covariável

Fonte de Variação	**Graus de Liberdade**	**Soma dos Quadrados e Produtos**			**Ajustes para regressão**		
		x	***xy***	***y***	***y***	**Graus de Liberdade**	**Quadrado Médio**
Tratamentos	*a –1*	T_{xx}	T_{xy}	T_{yy}			
Erro	*a(n–1)*	E_{xx}	E_{xy}	E_{yy}	$SS_E = E_{yy} - \left(E_{xy}\right)^2 / E_{xx}$	*a(n–1)–1*	$MS_E = \frac{SS_E}{a(n-1)-1}$
Total	*an–1*	S_{xx}	S_{xy}	S_{yy}	$SS^{'}_E = S_{yy} - \left(S_{xy}\right)^2 / S_{xx}$	*an–2*	
Tratamentos ajustados					$SS^{'}_E = SS_E$	*a–1*	$\frac{SS^{'}_E - SS_E}{a-1}$

2.3 Análise de concordância por atributos

Ao analisar a variabilidade que apresentam variáveis discretas, o mais recomendado é o uso da estratégia intitulada análise de concordância por atributos (*AAA – Attribute Agreement Analysis*). Este método se caracteriza por ser uma estratégia estatística utilizada para verificar se os avaliadores apresentam consistências entre si e com padrões previamente conhecidos. Essa técnica permite reduzir ou eliminar o efeito sobre a subjetividade do julgamento, como, por exemplo, na classificação de aglomerados de subestações. Para criar adequadamente esse estudo, é necessária uma metodologia específica, definindo o número de amostras, avaliadores e réplicas a serem analisadas. Após isso, utilizam-se testes específicos para definir a variabilidade/concordância, como a estatística Kappa e os coeficientes de Kendall. Tais estatísticas buscam descrever o nível de concordância existente entre diferentes classificações, sendo calculada a partir das esquematizações de Fleiss [117,118].

O índice Kappa mostra a razão de proporções que os avaliadores concordam com as proporções máximas que poderiam concordar. A Eq. (2.15) indica o grau de concordância das classificações nominais ou padrão realizadas por vários avaliadores, quando analisam as

mesmas respostas. De acordo com AIAG [119], os valores de Kappa ≥ 0,75 indicam boa concordância, mas valores ≥ 0,9 são preferidos.

$$K = \left(\frac{P_o - P_e}{1 - P_e} \right) = \frac{\left[\frac{1}{N_k n_k (n_k - 1)} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} x_{ij} - N_k n_k \right) \right] - \sum_{j=1}^{l} p_j^2}{\left(1 - \sum_{j=1}^{l} p_j^2 \right)} \tag{2.15}$$

onde destacamos que p_j^2 é a proporção esperada de concordância para cada categoria; P_o é a proporção média de concordância observada; P_e é a proporção média de concordância esperada; N_k e n_k representam o número de itens avaliados e o número de avaliadores, respectivamente; k é o número de categorias da escala adotada; e x_{ij} representam o número de avaliadores que classificaram o *i-ésimo* item na *j-ésima* categoria.

O cálculo da variância da estatística Kappa pode ser definida a partir da Eq. (2.16).

$$Var(K) = \frac{\left[\left(\sum_{j=1}^{k} p_j (1 - p_j) \right)^2 - \sum_{j=1}^{k} p_j (1 - p_j)(1 - 2p_j) \right]}{N_k n_k (n_k - 1) \left(\sum_{j=1}^{k} p_j (1 - p_j) \right)^2} \tag{2.16}$$

Ao analisar o coeficiente de Kendall (descrito na Eq. (2.17)), verifica-se que o mesmo é mais adequado para realizar análises de dados ordinais, sendo utilizados como classificações, no estudo de Gisev *et al.* [120], ou mesmo em escala Likert, como no estudo de Lewis e Johnson [121]. Deste modo, diante da Eq. (2.17), pode-se mensurar o nível de concordância entre os avaliadores a partir do coeficiente de concordância de Kendall (W), tanto para avaliações dentro e entre avaliadores [122,123].

$$W = \frac{12 \sum_{i=1}^{n} R_i^2 - 3p^2 n (n+1)^2}{p^2 (n^3 - n) - p \left(\sum_{k=1}^{m} (t_k^3 - t_k) \right)} \tag{2.17}$$

sendo n o número de sujeitos, R_i^2 a estatística da soma dos quadrados para as somas das classificações R_i, p refere-se ao número de avaliadores, enquanto t_k indica o número de classificações com empate em cada k dos m grupos de empate.

Se for necessário comparar os avaliadores a um valor padrão ou referência determinada, é possível usar o coeficiente de correlação de Kendall (τ). Este indicador indica o grau de

associação entre as avaliações e um determinado padrão, conforme mostrado em Eq. (2.18) [124].

$$\tau = \frac{C - D}{\sqrt{\left(\left[n_{++}\left(n_{++} - 1\right)0.5 - T_X\right]\left[n_{++}\left(n_{++} - 1\right)0.5 - T_Y\right]\right)}} \tag{2.18}$$

onde C e D são os números de pares concordantes e discordantes calculados como $\sum_{i<k}\sum_{j<l} n_{ij}n_{kl}$ e $\sum_{i<k}\sum_{j>l} n_{ij}n_{kl}$, respectivamente. T_X e T_Y são respectivamente equivalentes a $0{,}5\sum_i n_{i+}\left(n_{i+} - 1\right)$ e $0{,}5\sum_j n_{+j}\left(n_{+j} - 1\right)$. n_{i+} pode ser definido como o número de observações na variável X, enquanto n_{+j} é o número de observações na variável Y. n_{ij} representa as observações em uma célula, que corresponde à *i-ésima* linha e à *j-ésima* coluna. De modo análogo, n_{kl} representa as observações em uma célula, que corresponde a *k-ésima* linha e *l-ésima* coluna. Por fim, tem-se o número total de observações representados por n_{++}.

A Tabela 2.6 apresenta os índices de concordância explorados neste estudo, enquanto a Tabela 2.7 indica os níveis de aceitabilidade de concordância de acordo com a AIAG [119] e Hinkle *et al.* [125].

Tabela 2.6. Índices de concordância de Kappa e Kendall

Índice de concordância	**Âmbito**	**Interpretação**
Estatística Kappa (K)	Entre -1 e 1	K = 1 significa uma concordância perfeita K = 0 significa que a concordância é a mesma que o esperado por acaso K = -1 significa que a concordância é menor do que o esperado por acaso
Coeficiente de concordância de Kendall (W)	Entre 0 e 1	W = 1 indica uma associação perfeita W = 0 indica que não há associação
Coeficiente de correlação de Kendall (τ)	Entre -1 e 1	τ = 1 significa uma associação positiva τ = 0 significa que não há associação τ = -1 significa uma associação negativa

Tabela 2.7. Níveis de aceitação de concordância

Aceitabilidade	**Valor da estatística Kappa (K)**	**Valores dos coeficientes de Kendall (*W e τ*)**
Pobre	$K < 0.40$	W or $\tau < 0.30$
Boa	$0.75 < K < 0.90$	$0.70 < W$ or $\tau < 0.90$
Excelente	$K > 0.90$	W or $\tau > 0.90$

2.4 Planejamento e análise de experimentos

Um experimento pode ser definido por uma sequência de testes que contemplam modificações propositais inferidas nas variáveis de controle de um processo, com finalidade de observar como as respostas de interesse são afetadas por tais modificações [116]. Nesse sentido, uma aplicação experimental busca contribuir para o aprimoramento e análise em diversos segmentos, sendo parte substancial para o método científico [116].

Com base nisso, pode-se destacar a estratégia de Planejamento e Análise de Experimentos. O DOE define-se como uma técnica que combina aplicações estatísticas e matemáticas para criar arranjos experimentais, com objetivo de facilitar a coleta e análise de dados experimentais. Deste modo, de acordo com Gomes [126], os problemas experimentais devem ser amparados por dois elementos, sendo eles: o planejamento experimental e a análise estatística dos resultados.

Perante a ampla aplicabilidade e uso do DOE, é possível destacar algumas vantagens de se utilizar esse tipo de estratégia, tais como: minimizar a variabilidade e, consequentemente, maximizar a conformidade das especificações da fabricação de produtos; melhorar a produtividade de processos; minimizar os custos e; reduzir o tempo para desenvolver processos ou produtos.

Nesse sentido, Montgomery [116] define três princípios para o DOE:

- Replicação: consiste em repetir um mesmo teste diversas vezes, proporcionando uma variância da resposta de interesse que será utilizada para analisar o erro experimental;
- Aleatorização: refere a realizar experimentos a partir de uma ordem aleatória, fazendo com que os efeitos não conhecidos sejam disseminados entre os fatores, proporcionando a credibilidade da investigação;
- Blocagem: utilizada no momento em que não for viável manter as condições experimentais de maneira homogênea.

Ao realizar a abordagem estatística, Gomes [126] infere a relevância do conhecimento, a priori, do fenômeno e da coleta de dados a serem estudados nos experimentos. Assim, para utilizar adequadamente essa estratégia, se faz necessário seguir as seguintes etapas [116]:

1. Definir o problema;
2. Escolher os fatores e definir os níveis experimentais;
3. Selecionar as respostas de interesse;

4. Executar (randomicamente) os experimentos;
5. Realizar a análise estatística;
6. Inferir as conclusões.

Diante disso, autores como [116,127–129], apresentam as técnicas mais utilizadas, tais como o arranjo fatorial, a metodologia de superfície de resposta, arranjos de misturas e arranjos de Taguchi, em que, os três primeiros citados, serão utilizados no presente trabalho.

2.4.1 Arranjo fatorial completo

Diversos experimentos analisam o estudo de efeitos com dois ou mais fatores [116], em que os experimentos do tipo fatorial promovem analisar essa relação. Por arranjo fatorial, implica-se que, em cada tentativa completa de um experimento, as combinações existentes para os níveis foram contempladas, em sua totalidade. Deste modo, segundo Montgomery [116], se existir *a* níveis para o fator **A** e *b* níveis para o fator **B**, cada replicação apresenta todas combinações possíveis para *ab*, assim, afirma-se que os fatores estão cruzados.

Cirilo [130] afirma que, usualmente, os experimentos fatoriais utilizam até 3 fatores, mas que, experimentações no setor industrial vislumbram inúmeros fatores. Quanto maior a quantidade de fatores adotados, maior será a quantidade de experimentos no arranjo fatorial. Assim, a quantidade de experimentos para fatores de dois níveis, pode ser definida por 2^k, onde *k* refere a quantidade de fatores. Uma quantidade relevante de fatores, tornaria a condução de um experimento real, inviável [130]. A Tabela 2.8 apresenta a quantidade de experimentos para diferentes níveis e fatores.

Considerando o efeito da resposta de interesse em relação a mudança de nível de fator, Montgomery [116] denomina essa interação de efeito principal, pois trata-se dos fatores mais importantes no experimento. A técnica mais utilizada para se analisar, de maneira generalizada, o experimento é a análise de variância. Considerando um experimento fatorial 2^k, tem-se *k* efeitos principais e $2^k–k–1$ efeitos de interação [130]. Assim, estima-se as somas dos quadrados para cada fonte utilizando a codificação, ou notação de Yates, representando o maior nível como +1 e o menor nível por –1. Considerando um arranjo com dois fatores a dois níveis, a Figura 2.1 apresenta a interpretação geométrica desse modelo fatorial.

Tabela 2.8. Quantidade de experimentos para fatoriais p^k

Níveis	Fatores				
	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
2	$2^2 = 4$	$2^3 = 8$	$2^4 = 16$	$2^5 = 32$	$2^6 = 64$
3	$3^2 = 9$	$3^3 = 27$	$3^4 = 81$	$3^5 = 243$	$3^6 = 729$
4	$4^2 = 16$	$4^3 = 64$	$4^4 = 256$	$4^5 = 1024$	$4^6 = 4096$
...	...	...	...	...	...
p	p^2	p^3	p^4	p^5	p^6

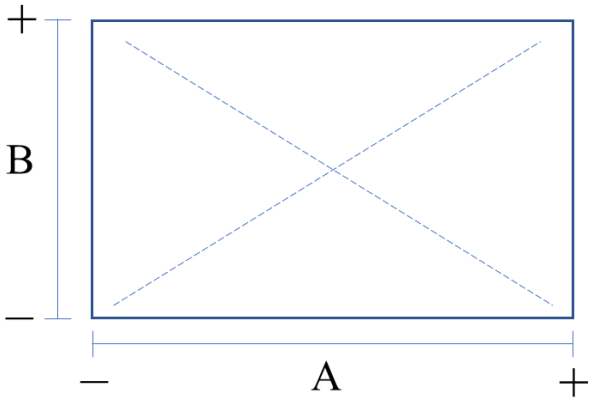


Figura 2.1. Representação geométrica de um fatorial $k = 2$

Em alguns modelos experimentais, verifica-se que um dos níveis se difere em relação a outros fatores (em todos níveis), essa ocorrência determina a existência de uma interação entre fatores. Essa interação pode ser representada graficamente pela Figura 2.2, onde Figura 2.2(a) apresenta linhas paralelas, indicando que não há evidências para determinar que a interação é significativa e a Figura 2.2(b) indica a ocorrência de uma interação entre os fatores. Contudo, apenas com os resultados da ANOVA é possível afirmar se as interações e efeito principal são, realmente, significativos ou não.

Segundo Montgomery [116], é possível verificar a interação através do modelo de regressão para o experimento fatorial, conforme descrito na Eq. (2.19), onde y representa a resposta de interesse, x_1 e x_2 são as variáveis que representam os fatores **A** e **B**, respectivamente, enquanto β's representam os parâmetros que devem ser determinados e o ε indica o termo de erro aleatório do modelo.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{2.19}$$

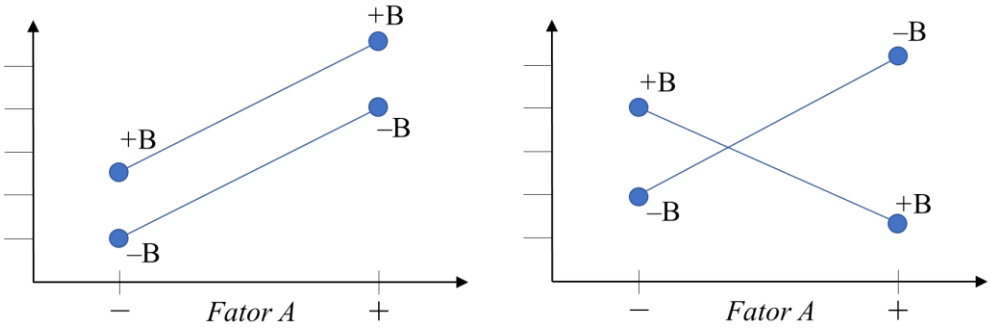


Figura 2.2. Experimento fatorial (a) sem interação; (b) com interação

Ao verificar as estimativas de parâmetro do modelo acima, tem-se que os mesmos apresentam relação direta com as estimativas de efeito, em que β_0 é estimado pela média de todas respostas, enquanto β_1 e β_2 apresentam as médias do valor do efeito principal correspondente. Deste modo, tem-se que as estimativas obtidas por meio do arranjo fatorial (com fatores em dois níveis – 2^k) são estimativas de mínimos quadrados [116]. Esse modelo pode ser representado através de gráficos de superfície e de contorno, conforme a Figura 2.3. Nesse sentido, Montgomery [116] afirma que conhecer a interação pode ser mais útil do que o efeito principal, uma vez que a interação significativa pode camuflar a relevância dos efeitos principais.

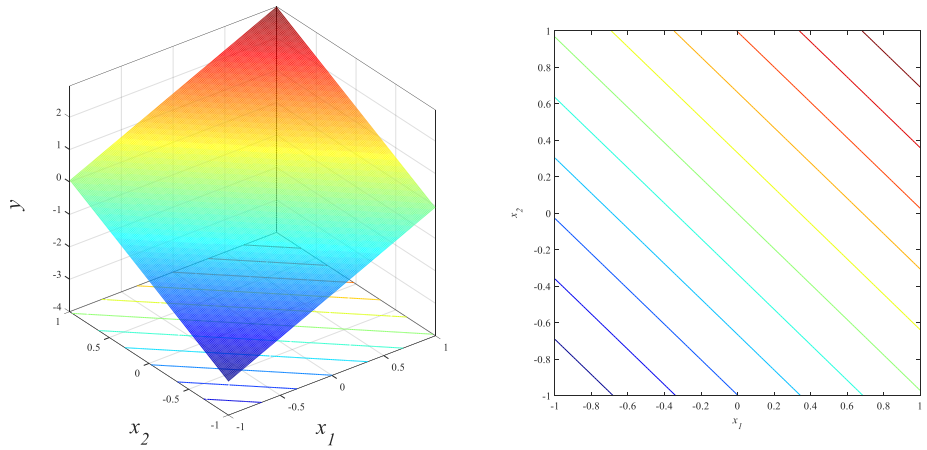
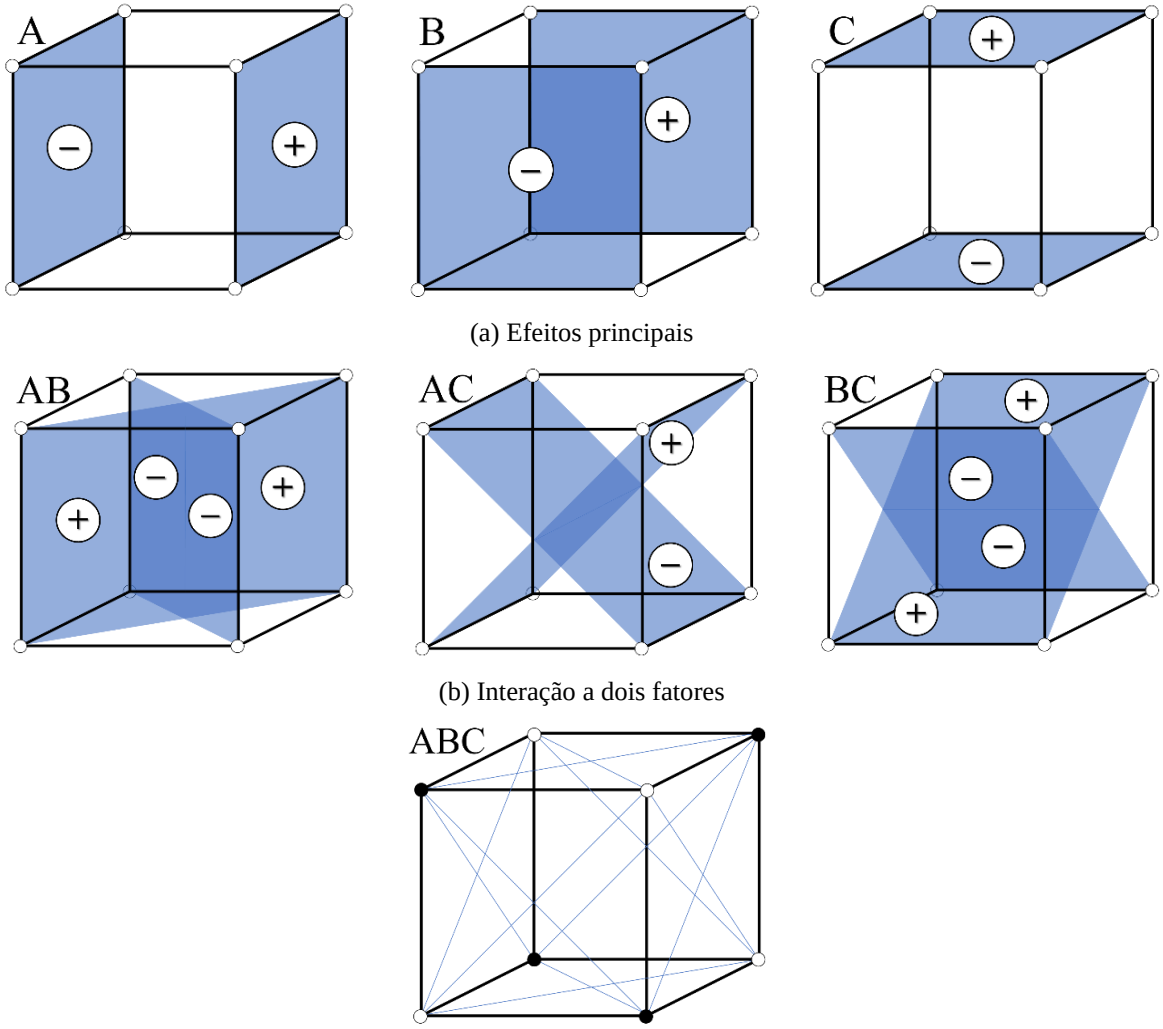


Figura 2.3. Gráficos de (a) Superfície de resposta e (b) Gráfico de contorno para um modelo linear

Expandindo a análise para um modelo de três fatores com dois níveis (2^3), a representação geométrica do arranjo fatorial apresenta formato cuboidal, indicando que cada uma das faces representa um fatorial duplo, favorecendo o entendimento para a estimação dos efeitos [130].

Considerando a posição dos efeitos (**A**, **B** e **C**), é possível estimar, através das diferenças entre as faces demarcadas, os efeitos principais [116], conforme detalhado na Figura 2.4.



(a) Efeitos principais

(b) Interação a dois fatores

(c) Interação a três fatores

Figura 2.4. Representação geométrica de contrastes correspondentes aos principais efeitos e interações (2^3)

2.4.2 Arranjo fatorial generalizado

Quando existe a necessidade de um modelo disposto em um experimento fatorial, que contemple fatores com diversos níveis, pode-se optar por uma extensão do modelo generalizado do arranjo fatorial. Segundo Montgomery [116], esse modelo pode ser utilizando onde exista *a* níveis para o fator **A**, *b* níveis para o fator **B**, *c* níveis para o fator **C** e quantos outros forem necessários.

Considerando um modelo onde os fatores experimentais são fixos, é possível formular e testar as hipóteses para as interações e efeitos principais a partir da ANOVA, como apresentado na seção anterior. Para um modelo de efeitos fixos, divide-se o quadrado médio pelo erro

quadrado médio de cada efeito ou interação, a fim de construir as estatísticas de teste (para cada uma das interações e efeitos principais) [116]. Os graus de liberdade para qualquer um dos efeitos principais é indicado pelo número de níveis do fator menos um. Já os graus de liberdade da interação se dão pelo produto da quantidade de DF relacionados aos componentes individuais de interação. Por fim, para o teste *F*, utiliza-se testes de uma cauda e cauda superior [116]. Deste modo, um modelo de três fatores para análise de variância pode ser representado conforme a Eq. (2.20).

$$y_{ijkl} = \mu + \tau_i + \beta_j + \omega_k + (\tau\beta)_{ij} + (\tau\omega)_{ik} + (\beta\omega)_{jk} + (\tau\beta\omega)_{ijk} + \varepsilon_{ijkl} \begin{cases} i = 1,2,...,a \\ j = 1,2,...,b \\ k = 1,2,...,c \\ l = 1,2,...,n \end{cases} \quad (2.20)$$

Considerando que os fatores **A**, **B** e **C** sejam fixos, a análise de variância pode ser calculada a partir das equações detalhadas na Tabela 2.9, sendo os testes *F* calculados pelos quadrados médios (MS) esperados para os efeitos principais e interações.

Tabela 2.9. Análise de variância para o modelo de efeitos fixos de três fatores

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Médias Quadradas	Média Quadrada Esperada	F_0
A	SS_A	$a-1$	MS_A	$\sigma^2 + \frac{bcn\sum \tau_i^2}{a-1}$	$F_0 = \frac{MS_A}{MS_E}$
B	SS_B	$b-1$	MS_B	$\sigma^2 + \frac{acn\sum \beta_j^2}{b-1}$	$F_0 = \frac{MS_B}{MS_E}$
C	SS_C	$c-1$	MS_C	$\sigma^2 + \frac{abn\sum \omega_k^2}{c-1}$	$F_0 = \frac{MS_C}{MS_E}$
AB	SS_{AB}	$(a-1)(b-1)$	MS_{AB}	$\sigma^2 + \frac{cn\sum\sum (\tau\beta)_{ij}^2}{(a-1)(b-1)}$	$F_0 = \frac{MS_{AB}}{MS_E}$
AC	SS_{AC}	$(a-1)(c-1)$	MS_{AC}	$\sigma^2 + \frac{bn\sum\sum (\tau\omega)_{ik}^2}{(a-1)(c-1)}$	$F_0 = \frac{MS_{AC}}{MS_E}$
BC	SS_{BC}	$(b-1)(c-1)$	MS_{BC}	$\sigma^2 + \frac{an\sum\sum (\beta\omega)_{jk}^2}{(b-1)(c-1)}$	$F_0 = \frac{MS_{BC}}{MS_E}$
ABC	SS_{ABC}	$(a-1)(b-1)(c-1)$	MS_{ABC}	$\sigma^2 + \frac{n\sum\sum\sum (\tau\beta\omega)_{ijk}^2}{(a-1)(b-1)(c-1)}$	$F_0 = \frac{MS_{ABC}}{MS_E}$
Erro	SS_E	$abc(n-1)$	MS_E	σ^2	
Total	SS_T	$abcn-1$			

De modo complementar, a Tabela 2.10 apresenta os equacionamentos para cálculo da soma dos quadrados totais (SS_T) e soma dos quadrados para os efeitos principais $A(y_{i\ldots})$, $B(y_{.j\ldots})$

e $C(y_{..k.})$, sendo eles SS_A, SS_B e SS_C, respectivamente. Além disso, a soma dos quadrados para as combinações de segunda ordem (SS_{AB}, SS_{AC} e SS_{BC}) e de terceira ordem (SS_{ABC}) são contempladas, sendo que este último considera $\{y_{ijk.}\}$ para o cálculo. Por fim, a soma dos quadrados para o erro (SS_E) é calculada.

Tabela 2.10. Soma dos quadrados totais e dos efeitos principais

		Soma dos Quadrados
SS_T	=	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} y_{ijkl}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn}$
SS_A	=	$\frac{1}{bcn}\sum_{i=1}^{a} y_{i\bullet\bullet\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn}$
SS_B	=	$\frac{1}{acn}\sum_{j=1}^{b} y_{\bullet j\bullet\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn}$
SS_C	=	$\frac{1}{abn}\sum_{k=1}^{c} y_{\bullet\bullet k\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn}$
SS_{AB}	=	$\frac{1}{cn}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij\bullet\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn} - SS_A - SS_B = SS_{subtotais(AB)} - SS_A - SS_B$
SS_{AC}	=	$\frac{1}{bn}\sum_{i=1}^{a}\sum_{k=1}^{c} y_{i\bullet k\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn} - SS_A - SS_C = SS_{subtotais(AC)} - SS_A - SS_C$
SS_{BC}	=	$\frac{1}{an}\sum_{j=1}^{b}\sum_{k=1}^{c} y_{\bullet jk\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn} - SS_B - SS_C = SS_{subtotais(BC)} - SS_B - SS_C$
SS_{ABC}	=	$\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk\bullet}^2 - \frac{y_{\bullet\bullet\bullet\bullet}^2}{abcn} - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC}$ $= SS_{subtotais(ABC)} - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC}$
SS_E	=	$SS_T - SS_{subtotais(ABC)}$

É importante ressaltar que os modelos apresentados se referem a experimentos fatoriais fixos. Contudo, quando o modelo fatorial envolver pelo menos um dos fatores aleatórios, deve-se determinar os testes corretos analisando os quadrados médios esperados, conforme indicado por Montgomery [116].

2.4.3 Metodologia de superfície de resposta

A Metodologia de Superfície de Resposta (*RSM – Response Surface Methodology*) é uma estratégia amplamente difundida e utilizada nas mais variáveis aplicações, tais como [131–140].

Essa metodologia se classifica como uma coletânea de técnicas estatísticas e matemáticas utilizadas para modelar e analisar uma resposta de interesse, avaliando se a mesma é influenciada por várias variáveis [116,141].

Essa estratégia pode ser definida como uma extensão do arranjo fatorial, visto anteriormente, pois trata-se de um modelo completo com pontos adicionais que contemplam uma região experimental: os pontos axiais e pontos centrais. Assim, de maneira geométrica, a nova região experimental deixa de ser um cuboidal ($k = 3$) para ser uma esfera. A Figura 2.5 ilustra os modelos geométricos usuais da RSM para $k = 2$ e $k = 3$.

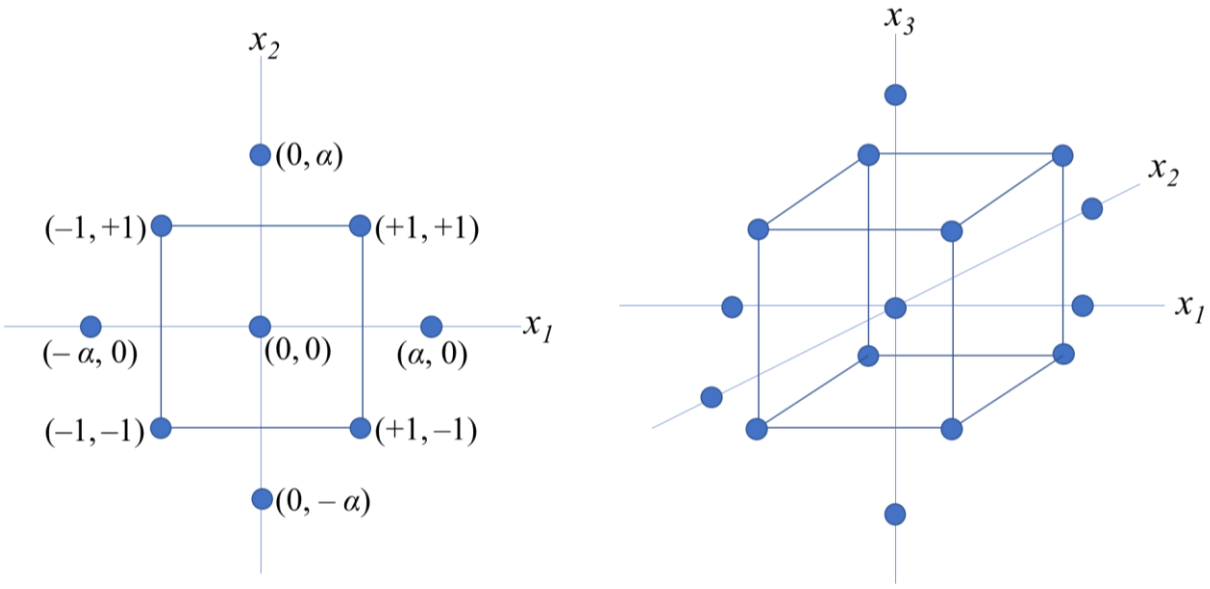


Figura 2.5. Representação geométrica da RSM para $k = 2$ e $k = 3$

Sendo desconhecida a relação "resposta × variáveis independentes", utiliza-se, inicialmente, um polinômio de ordem inferior em alguma região dessas variáveis [116]. Caso a resposta de interesse seja modelada adequadamente a partir de uma função linear, pode-se utilizar um modelo de primeira ordem, como o da Eq. (2,19), descrito na seção 2.2.1. Contudo, se existir curvatura significativa, um polinômio de maior ordem deve ser adotado, como o de segunda ordem, descrito na Eq. (2.21).

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \varepsilon \qquad (2.21)$$

onde y representa a resposta de interesse, x os parâmetros, β os coeficientes estimados, k o número de variáveis independentes e ε o termo de erro associado.

Um modelo RSM com curvatura significativa é usualmente representado conforme a Figura 2.6, tanto para o gráfico de superfície, quanto o gráfico de contorno. Assim como no arranjo fatorial, essa metodologia também permite analisar e gerar a interação e os efeitos principais.

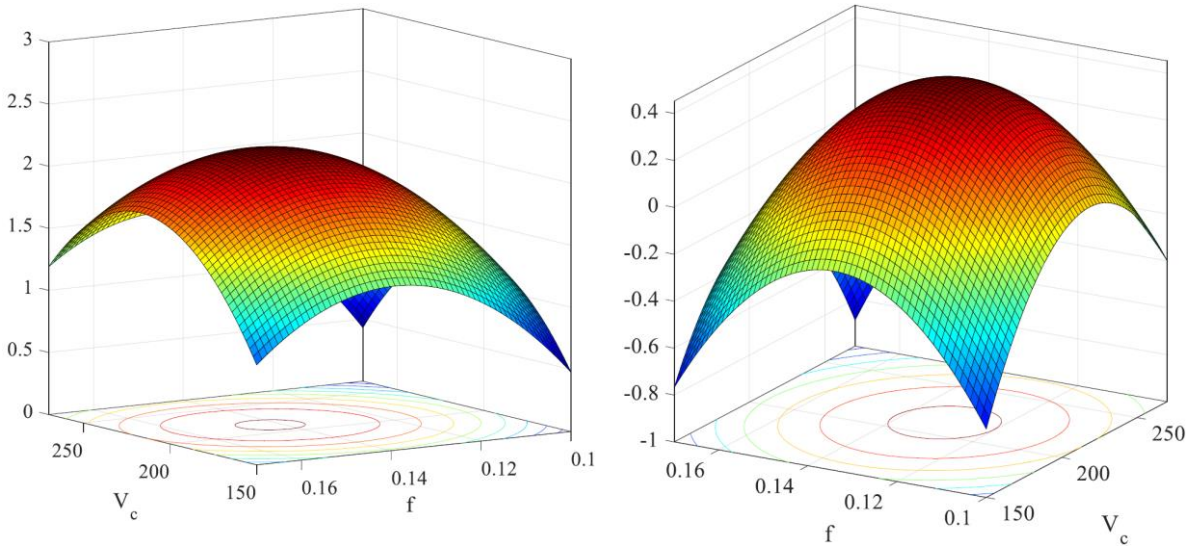


Figura 2.6. Exemplificação de gráficos de superfície e de contorno para RSM
Adaptada de [141]

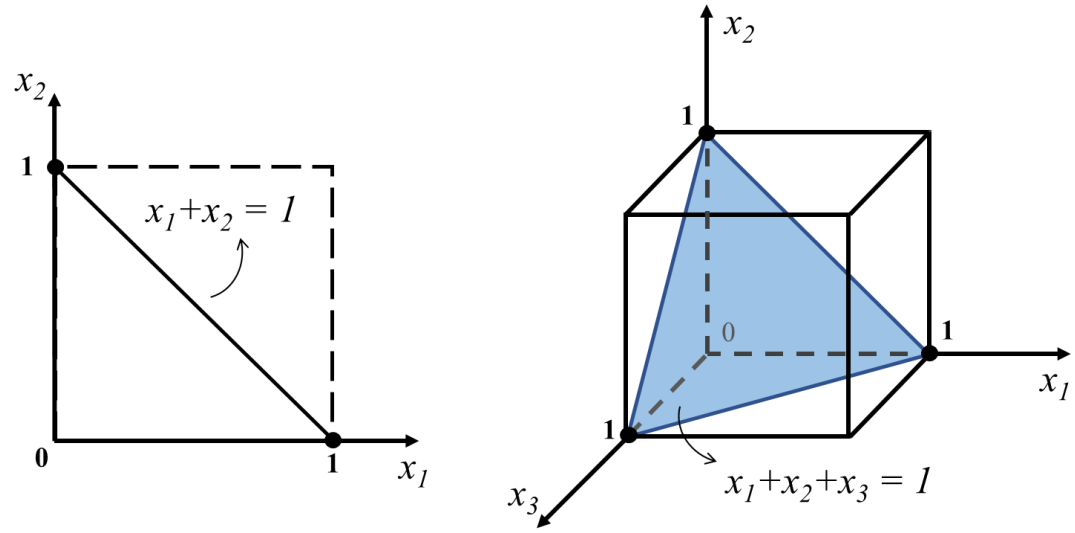
2.4.3.1 Arranjo de misturas

Após obter os conceitos básicos da RSM, é possível verificar os casos particulares desse método, como o arranjo de misturas. Inicialmente, os modelos apresentados indicavam arranjos com níveis distintos entre os fatores, contudo, os experimentos de misturas apresentam fatores não independentes. Isso se justifica pelo fato desses níveis serem complementares, como uma proporção.

O arranjo de misturas, também conhecido como método dos polinômios canônicos de mistura, referem-se a diferentes modelos usados neste tipo de planejamento experimental, sendo classificado como uma classe especial de RSM. Portanto, as variáveis de entrada são consideradas como componentes e as respostas de interesse são caracterizadas como funções das proporções de cada componente [116]. Consequentemente, os valores seguem uma restrição de totalidade, onde a soma é igual a 1 (Eq. (2.22)).

$$\sum_{i=1}^{k} x_i = x_1 + x_2 + \ldots + x_k = 1; \qquad com \;\; x_i \geq 0 \tag{2.22}$$

Para dois componentes, a região experimental considera valores ao longo de uma linha reta (Figura 2.7(a)), enquanto para três componentes, o espaço experimental é delimitado por um triângulo (Figura 2.7(b)). Tais características trazem a necessidade de que esses experimentos sejam planejados por projetos específicos, onde os tipos *simplex* são os mais comuns na literatura [113,114,127,128,141–143].



(a) (b)

Figura 2.7. Região experimental para projeto de mistura: (a) *k = 2* componentes; (b) *k = 3* componentes

Dentre os tipos de projetos *simplex*, pode-se destacar o *simplex-centroide*, *simplex-lattice* e o de *vértices extremos* [113,116,128,130]. Um dos principais e mais utilizados é o projeto conhecido como *simplex-lattice*. Nesse tipo de projeto, *k* variáveis de entrada são consideradas para um determinado grau *lattice* (*ld*). Desta forma, existem valores equidistantes entre 0 e 1, onde o número total de experimentos (*N*), pode ser definido pela Eq. (2.23).

$$N = \frac{(k + ld - 1)!}{ld! \cdot (k - 1)!} \tag{2.23}$$

O arranjo *simplex-lattice* para três variáveis de entrada e *lattice* igual a *2* (*ld = 2*) pode ser verificado graficamente através da Figura 2.8(a). No entanto, é importante destacar que também é necessário avaliar o ponto interno da matriz. Dessa forma, os experimentos podem ser incorporados adicionando pontos internos aos arranjos, como pontos centrais ou axiais. A Figura 2.8(b) ilustra um projeto de rede *simplex* embutido (*k = 3*; *ld = 2*).

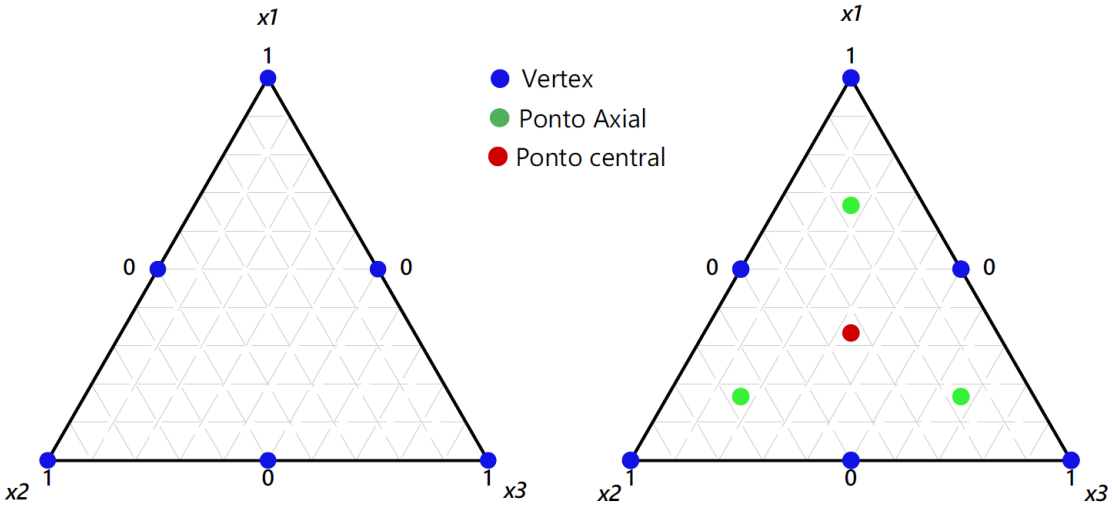


Figura 2.8. Projeto de rede simples para $k = 3$ componentes e $ld = 2$

Em relação ao modelo matemático usado para representar a resposta que acompanha a restrição ($\sum_{i=1}^{k} x_i = 1$), um dos mais aplicados com base nos polinômios usados no RSM é o modelo quadrático, onde β são os coeficientes e x são as variáveis de controle. Além do modelo quadrático, modelos com polinômios do modelo linear, cúbico e cúbico especial são representados na Tabela 2.11. Tais equacionamentos são também conhecidos como polinômios de Scheffé [116,126–128].

Tabela 2.11. Polinômios canônicos de misturas

Modelos matemáticos para o arranjo de misturas	
Linear	$y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i x_i$
Quadrático	$y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j}^{k} \sum \beta_{ij} x_i x_j$
Cúbico	$y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j}^{k} \sum \beta_{ij} x_i x_j + \sum_{i<j}^{k} \sum \delta_{ij} x_i x_j \left(x_i - x_j\right) + \sum_{i<j<l}^{k} \sum \sum \beta_{ijl} x_i x_j x_l$
Cúbico Especial	$y(\mathbf{x}) = \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j}^{k} \sum \beta_{ij} x_i x_j + \sum_{i<j<l}^{k} \sum \sum \beta_{ijl} x_i x_j x_l$

Conhecendo adequadamente as estratégias experimentais, a próxima subseção abordará elementos referentes a análises de dados multivariados, começando pela estratégia intitulada de análise de componentes principais.

2.5 Análise de componentes principais

A análise de componentes principais se caracteriza por ser uma técnica multivariada amplamente utilizada para interpretar e reduzir dados extensos e correlacionados. Assim, tem-se que o PCA minimiza a dimensionalidade das variáveis originais, de modo a absorver elementos significativos no eixo principal, enquanto mantém a variação do erro nos eixos secundários.

De maneira análoga, PCA também se destaca por ser difundida na literatura para reduzir o esforço computacional em análises que envolvem grandes conjuntos de dados. Essa estratégia faz uso de uma conversão ortogonal para transformar as observações em um conjunto de variáveis, não apresentando correlação entre si. A definição da quantidade ideal de componentes se dá por critérios definidos por Kaiser ([38]) e também detalhados por Johnson e Wichern [5], dos quais destaca-se que, quanto maior o grau de correlação das variáveis, menor a quantidade necessária de componentes a ser utilizado para representar as observações. Entre esses critérios, tem-se que os componentes devem explicar, pelo menos, 80% da variância acumulada. O uso dessa estratégia está presente em diversos estudos da literatura, como os destacados anteriormente, no capítulo 1.

Sabe-se que o PCA visa encontrar uma combinação de variáveis não correlacionadas que explica adequadamente as variáveis originais [144]. Para alcançar esse objetivo, considera-se o vetor aleatório $\boldsymbol{X}^T = [X_1, X_2, \ldots, X_p]$ que tem a matriz de covariância $\boldsymbol{\Sigma}$ com autovalores $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p \geq 0$. Então, as combinações lineares podem ser descritas como na Eq. (2.24).

$$\begin{aligned} Y_1 &= a_1^T X = a_{11}X_1 + a_{21}X_2 + \ldots + a_{p1}X_p \\ Y_2 &= a_2^T X = a_{12}X_1 + a_{22}X_2 + \ldots + a_{p2}X_p \\ &\vdots \\ Y_p &= a_p^T X = a_{1p}X_1 + a_{2p}X_2 + \ldots + a_{pp}X_p \end{aligned} \tag{2.24}$$

Considerando que Y_i for o *i-ésimo* componente principal, então, chega-se às Eqs. (2.25) e (2.26).

$$Var(Y_i) = a_i^T \sum a_i = e_i^T \sum e_i; \qquad \forall i = 1, 2, \ldots, p \tag{2.25}$$

$$CoVar(Y_i, Y_k) = a_i^T \sum a_k = e_i^T \sum e_k; \qquad \forall i, k = 1, 2, \ldots, p \tag{2.26}$$

Assim, tem-se que os componentes principais representam as combinações lineares não correlacionadas Y_1, Y_2, ..., Y_p, em que as variâncias descritas na Eq. (2.25) são as maiores possíveis. Ou seja, o *i-ésimo* componente pode ser definido a partir da Eq. (2.27), sendo esta, previamente obtida através da formulação escrita na Eq. (2.28) [145,146].

$$PC_i = e_i^T Y = e_{1i}Y_1 + e_{2i}Y_2 + \ldots + e_{qi}Y_q \qquad i = 1, 2, \ldots, q \tag{2.27}$$

$$\begin{aligned} &Max \quad Var\left[e_i^T Y\right] \\ &s.a: \quad e_i^T e_1 = 1 \\ &\qquad Cov\left[e_i^T Y, e_k^T Y\right] = 0 \\ &\qquad k < i \end{aligned} \tag{2.28}$$

Deste modo, é possível substituir as variáveis observáveis por um conjunto linear não-correlacionado que represente adequadamente essas observações, ou seja, por escores de componentes principais. A partir da matriz de autovetores **E** e também da matriz de dados padronizados **Z**, pode-se encontrar os escores dos componentes principais, a partir da Eq. (2.29).

$$PC_{escore} = \mathbf{Z}^T\mathbf{E} = \begin{bmatrix} \left(\frac{y_{11}-\overline{y_1}}{\sqrt{s_{11}}}\right) & \left(\frac{y_{12}-\overline{y_2}}{\sqrt{s_{22}}}\right) & \cdots & \left(\frac{y_{1q}-\overline{y_q}}{\sqrt{s_{qq}}}\right) \\ \left(\frac{y_{21}-\overline{y_1}}{\sqrt{s_{11}}}\right) & \left(\frac{y_{22}-\overline{y_2}}{\sqrt{s_{22}}}\right) & \cdots & \left(\frac{y_{2q}-\overline{y_q}}{\sqrt{s_{qq}}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{y_{n1}-\overline{y_1}}{\sqrt{s_{11}}}\right) & \left(\frac{y_{n2}-\overline{y_2}}{\sqrt{s_{22}}}\right) & \cdots & \left(\frac{y_{nq}-\overline{y_q}}{\sqrt{s_{qq}}}\right) \end{bmatrix}^T \times \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1q} \\ e_{21} & e_{22} & \cdots & e_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ e_{q1} & e_{q2} & \cdots & e_{qq} \end{bmatrix} \tag{2.29}$$

Segundo Johnson e Wichern [5], geometricamente, os componentes principais são encontrados na direção dos eixos de um elipsoide, que apresenta uma densidade constante.

2.6 Análise fatorial exploratória

A análise fatorial se caracteriza por uma técnica multivariada exploratória, ou mesmo de interdependência, utilizada para tratar de dados com múltiplas variáveis que apresentam coeficientes de correlação significativos, a fim de gerar novos vetores que representem, adequadamente, as variáveis originais. Ou seja, essa técnica busca representar variáveis mensuráveis, considerando suas características hipotéticas e latentes [147]. Segundo Fávero

[50], FA visa explicar o comportamento de um conjunto de variáveis independentes a partir de poucos fatores.

Essa técnica foi desenvolvida no início do século XX, mas com o advento da computação moderna, sua aplicação vem sendo amplamente investigada em mineração de dados, economia, psicologia, engenharia e outros setores. Entre as estratégias utilizadas para determinar e investigar os fatores, pode-se destacar dois métodos distintos de extração: através de componentes principais e através da máxima verossimilhança. Diante tais opções de extração, neste estudo, apenas a extração pelo método de componentes principais será abordada. Este método de extração permite representar o conjunto de dados por alguns fatores comuns, sem estrutura de variância-covariância significativas entre elas [50]. Deste modo, tem-se que a FA é amplamente utilizada tanto para fins exploratórios (reduzindo a dimensão dos dados), quanto para inferir uma determinada hipótese (em que os dados podem ser quantificados a um certo fator) [148]. Contudo, caso o objetivo principal seja investigar e confirmar relações, deve-se utilizar técnicas de análise fatorial confirmatória.

Fávero [50] determina quatro objetivos básicos para o uso da FA por componentes principais, sendo eles:

1. Determinar a existência de correlações significativas no conjunto de dados, com objetivo de criar novos fatores para representar a combinação linear das variáveis. Assim, pode-se realizar a redução da dimensionalidade dos dados a serem investigados;
2. Verificar a legitimidade de constructos criados, considerando o agrupamento das variáveis originais para cada um dos fatores;
3. Criar indicadores de desempenho, perante os fatores, para elaborar um escalonamento, ou classificação.
4. Por fim, extrair os fatores ortogonais para análises posteriores que podem carecer de variáveis independentes, ou seja, que não apresentem multicolinearidade.

Segundo Rencher [24], a técnica FA busca minimizar a repetição das informações existentes entre as variáveis observadas, utilizando uma menor quantidade de variáveis latentes. De maneira complementar, Johnson e Wichern [5] afirmam que que FA interpreta a estrutura de variância-covariância das variáveis originais em poucos fatores latentes. A representação básica desse objetivo pode ser visualizada na Figura 2.9.

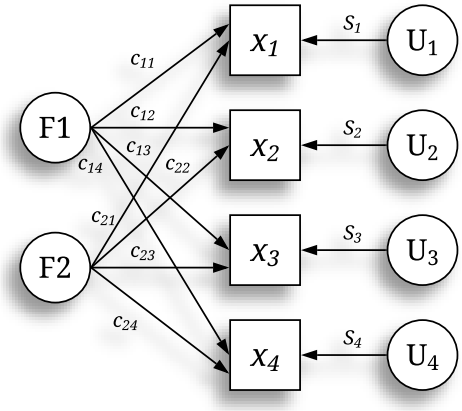


Figura 2.9. Diagrama de um modelo fatorial com dois fatores

Inicialmente, verifica-se o nível de correlação entre as variáveis, calculando o coeficiente de correlação de *Pearson*. Em seguida, se faz necessário avaliar se o conjunto de dados, de fato, está apto para aplicação dessa estratégia multivariada. Essa adequação global pode ser investigada a partir dos indicadores do teste de esfericidade de Bartlett e o índice Kaiser-Meyer-Olkin, que serão discutidos a seguir.

2.6.1 Análise de adequação dos dados

A aplicação da FA requer que as variáveis de resposta originais sejam adequadas [5]. Essa adequação pode ser avaliada por meio de testes específicos que serão discutidas nessa seção. Segundo Fávero [50], para uma extração adequada dos fatores, tem-se que a matriz de correlação populacional (**C**) apresente valores elevados e significativos. Para Hair *et al.* [40], não é possível afirmar, a partir de uma análise visual, se a matriz **C** é adequada, mas se a mesma apresentar uma quantidade considerável de valores menores que 0,3, isso seria um forte indício de que, usar FA nesse conjunto de dados, seria uma prática inadequada. Com finalidade de analisar a adequação dos dados para aplicação de FA, duas métricas são amplamente recomendadas e utilizadas na literatura, sendo o teste de esfericidade de Bartlett e o índice Kaiser-Meyer-Olkin (KMO).

O teste de esfericidade de Bartlett, proposto por Bartlett [149], analisa a matriz de correlações, comparando-a com a matriz identidade (**I**) de dimensão idêntica. Para isso, considera-se uma estatística de teste $\chi^2_{\alpha;v}$ para verificar se a matriz de correlação é uma matriz identidade, adotando um nível de significância α e um número $v = p\ (p - 1)/2$ de graus de liberdade [7]. Assim, caso as diferenças entre os valores fora da diagonal principal não sejam

estatisticamente distintas de zero, tem-se que os dados não são adequados para o uso da FA [50]. A hipótese nula e a hipótese alternativa deste teste estão descritas na Tabela 2.12.

Tabela 2.12. Teste de hipótese para adequação dos dados pela esfericidade de Bartlett

	Hipótese		
Nula	H_0:	$\mathbf{C} = \mathbf{I}$	$\begin{pmatrix} 1 & C_{12} & \cdots & C_{1k} \\ C_{21} & 1 & \cdots & C_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ C_{k1} & C_{k2} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$
Alternativa	H_1:	$\mathbf{C} \neq \mathbf{I}$	$\begin{pmatrix} 1 & C_{12} & \cdots & C_{1k} \\ C_{21} & 1 & \cdots & C_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ C_{k1} & C_{k2} & \cdots & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$

Este teste também assume que o conjunto de dados $\boldsymbol{Y} = [Y_1, Y_2, ..., Y_p]^T$ segue uma distribuição multivariada normal. Nesse sentido, a hipótese nula de que a matriz de correlação é igual à matriz identidade não é rejeitada. Em outras palavras, os dados são considerados não correlacionados quando $\chi^2 > \chi^2_{\alpha;[p(p-1)/2]}$, e o valor de χ^2 é obtido da Eq. (2.30) [24].

$$\chi^2 = -\left[n - 1 - \frac{(2p+5)}{6} \right] \ln |\mathbf{R}| \tag{2.30}$$

onde n representa o número de observações por amostra e $|\mathbf{R}|$ representa o determinante da matriz de correlação amostral.

Outra forma de verificar a adequação dos dados é por meio do índice KMO. Segundo Fávero [50], essa estatística apresenta o grau de proporção de variância que é comum para as variáveis originais analisadas. O índice contempla a amplitude entre 0 e 1, em que os valores próximos a zero concluem que o conjunto de dados é inadequado para aplicação de FA (sendo os coeficientes de Pearson são baixos), enquanto valores próximos a 1, indicam que os dados são adequados. É desejável que o nível KMO seja ≥ 0,5 [7,8].

Kaiser [150] apresentou, incialmente, o equacionamento para calcular esse índice, que pode ser verificado a partir da Eq. (2.31).

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} q_{ij}^2} \tag{2.31}$$

onde r_{ij} representa a matriz de correlação amostral **R** e q_{ij} representa a matriz de correlação anti-imagem **Q** [17], sendo $\mathbf{Q} = \mathbf{D}\mathbf{R}^{-1}\mathbf{D}$ para $\mathbf{D} = \left[\left(diag\mathbf{R}^{-1}\right)^{1/2}\right]^{-1}$.

Considerando os dois testes para verificar a adequação dos dados, tem-se que o teste de esfericidade de Bartlett é preferível em relação ao KMO, uma vez que o primeiro teste considera um nível de significância diante de uma distribuição de probabilidade, enquanto o segundo indicador não apresenta as mesmas particularidades, tratando-se apenas de um coeficiente mensurado [50].

Embora amplamente recomendado e, por alguns autores, mandatório, os testes de adequação são muitas vezes negligenciados ou omitidos em estudos e pesquisas científicas. Muitos autores fazem uso de estratégias multivariadas sem verificar se o conjunto de dados é realmente apto para essa análise, podendo gerar resultados poucos satisfatórios ou mesmo incorretos. Conhecendo os testes necessários para adequação global dos dados, pode-se realizar a aplicação da FA. A próxima seção descreve o procedimento de definição do modelo da FA.

2.6.2 Modelo fatorial

De acordo com Johnson e Wichern [50], o objetivo da FA é descrever as várias variáveis aleatórias (y_i, $i = 1, 2, \ldots, p$), que são observáveis e linearmente independentes em termos de características comuns entre elas (f_j, $j = 1, 2, \ldots, m$). Quando $m < p$, eles são conhecidos como fatores comuns ou variáveis latentes, ou seja, variáveis não observáveis.

A relação linear que representa o modelo fatorial, descrevendo a relação entre as variáveis de resposta e as variáveis latentes, pode ser expressa, matricialmente, de acordo com a Eq. (2.32).

$$\mathbf{Y} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon} \tag{2.32}$$

onde $\mathbf{Y}_{(p\times1)} = [y_1, y_2, ..., y_p]^T$ representa um vetor de variáveis aleatórias observáveis, $\boldsymbol{\mu}_{(p\times1)}$ é o vetor de médias populacionais, $\mathbf{L}_{(p\times m)}$ é a matriz de cargas fatoriais (Eq. (2.33)), $\mathbf{F}_{(m\times1)} = [F_1, F_2, ..., F_m]^T$ é um vetor aleatório de variáveis latentes e $\boldsymbol{\varepsilon}_{(p\times1)} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_p]^T$ representa um vetor aleatório de erros, também conhecido como *fatores específicos*.

$$\mathbf{L} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix} \tag{2.33}$$

A partir de um sistema de equações, é possível representar a Eq. (2.32) conforme descrito na Eq. (2.34).

$$\begin{aligned} y_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 \\ y_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ y_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p \end{aligned} \tag{2.34}$$

onde l_{ij} são coeficientes conhecidos como cargas fatoriais, sendo estas, medidas de influência de y_i para o fator comum F_j [24].

Deste modo, para Johnson e Wichern [5], deve-se assumir que:

Para $E(\mathbf{F}) = \underset{(m\times 1)}{0}$, tem-se $Cov(\mathbf{F}) = E\left[\mathbf{F}\mathbf{F}^T\right] = \underset{(m\times m)}{\mathbf{I}}$ (2.35)

E para $E(\boldsymbol{\varepsilon}) = \underset{(p\times 1)}{0}$, a $Cov(\boldsymbol{\varepsilon}) = E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \underset{(p\times p)}{\boldsymbol{\Psi}} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$ (2.36)

onde ψ_i representa a variância específica, sendo a proporção da variância de y_i explicada por ε_i. Considerando que **F** e **ε** são independentes, pode-se chegar à Eq. (2.37).

$$Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = E\left[\boldsymbol{\varepsilon}\mathbf{F}^T\right] = \underset{(p\times m)}{0} \tag{2.37}$$

Deste modo, tem-se que a relação das equações apresentadas anteriormente promovem o modelo fatorial ortogonal. De fato, essa estratégia busca inferir **Σ** para **L** e **Ψ** [24], sendo $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi}$, em que, diante do que foi considerado nas Eqs. (2.35), (2.36) e (2.37), tem-se a expressão descrita na Eq. (2.38) [151].

$$\begin{aligned} \boldsymbol{\Sigma} = Cov(\mathbf{y}) &= E\left[\left(\mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}\right) + \left(\mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}\right)^T\right] \\ &= E\left[\mathbf{L}\mathbf{F}\left(\mathbf{L}\mathbf{F}\right)^T + \mathbf{L}\mathbf{F}\boldsymbol{\varepsilon}^T + \boldsymbol{\varepsilon}\left(\mathbf{L}\mathbf{F}\right)^T + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] \\ &= \mathbf{L}E\left(\mathbf{F}\mathbf{F}^T\right)\mathbf{L}^T + \mathbf{L}E\left(\mathbf{F}\boldsymbol{\varepsilon}^T\right) + E\left(\boldsymbol{\varepsilon}\mathbf{F}^T\right)\mathbf{L}^T + E\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right) \\ &= \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi} \end{aligned} \tag{2.38}$$

Ainda considerando $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T + \boldsymbol{\Psi}$, tem-se que $Var(X_i) = \sigma_{ii}$, em que a mesma é particionada em duas, sendo conhecidas como *comunalidade* e *variância específica* [5]. Considerando a *i-ésima* comunalidade por h_i^2, tem-se, na Eq. (2.39).

$$\begin{aligned} \overbrace{\sigma_{ii}}^{Var(X_i)} &= \overbrace{l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2}^{Comunalidade} + \overbrace{\psi_i}^{Variância\ \ específica} \\ h_i^2 &= l_{i1}^2 + l_{i2}^2 + \ldots + l_{im}^2 \\ \sigma_{ii} &= h_i^2 + \psi_i; \qquad i = 1, 2, \ldots, p. \end{aligned} \tag{2.39}$$

onde a comunalidade (h_i^2) representa uma proporção da variância de y_i que é contribuída pelos m fatores comuns. A outra proporção é representada pela variância específica ψ_i, apresentada anteriormente. Johnson e Wichern [5] denotam que a soma dos quadrados da *i-ésima* variável dos m fatores comuns é representada pela *i-ésima* comunalidade.

A extração dos fatores para essa estratégia pode ser feita por diferentes métodos, contudo, alguns autores apresentam o método de extração por componentes principais como a estratégia mais utilizada, pois não necessita de uma distribuição de probabilidade específica para sua aplicação [7,8,151], como o método de máxima verossimilhança, por exemplo. Deste modo, o modelo de extração por componentes principais será detalhado nesse estudo.

2.6.3 Estimação de parâmetros por componentes principais

De acordo com Leite [151], a estimação através de componentes principais se dá pela decomposição espectral de $\mathbf{\Sigma}$, com finalidade de estimar $\mathbf{L}$ e $\mathbf{\Psi}$. Considerando a relação de $m < p$, onde m é o número de fatores e p o número de variáveis observáveis, tem-se a matriz ($p \times m$) para os autovetores de $\mathbf{\Sigma}$ representado por $\mathbf{P}_m = [e_1, e_2, ..., e_m]$, enquanto a matriz diagonal ($p \times m$) dos autovalores é representada por $\mathbf{\Lambda}_m = [\lambda_i]$ [151]. Sendo estas matrizes o resultado da decomposição espectral, apresentadas anteriormente, pode-se assumir, de maneira análoga, que $\mathbf{\Sigma} = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots + \lambda_2 e_2 e_2^T$, ou seja, $\mathbf{\Sigma} = \mathbf{\Lambda}_m \mathbf{P}_m \mathbf{P}^{\mathbf{T}}_m = \mathbf{L}\mathbf{L}^T$. Segundo Ferreira [4], a matriz $\mathbf{L}$ pode ser descrita, de acordo com a Eq. (2.40).

$$\mathbf{L} = \mathbf{P}_m \mathbf{\Lambda}_m^{1/2} = \left[\sqrt{\lambda_1}\mathbf{e}_1, \sqrt{\lambda_2}\mathbf{e}_2, \ldots, \sqrt{\lambda_m}\mathbf{e}_m \right] \tag{2.40}$$

Avaliando a matriz $\mathbf{\Sigma}$, a mesma pode ser expressa, de maneira incompleta, por $\mathbf{\Sigma} \cong \mathbf{L}\mathbf{L}^T$, em que o modelo não considera a contribuição dos últimos autovalores e autovetores. Além disso, os fatores específicos (erros) também não são representados nesse modelo. Assim, pode-se assumir o seguinte modelo de $\mathbf{\Sigma} \cong \mathbf{L}\mathbf{L}^T + \mathbf{\Psi}$, onde a representação da matriz de variâncias específicas pode ser definida conforme a Eq. (2.41) [151].

$$\mathbf{\Psi} = diag(\mathbf{\Sigma} - \mathbf{L}\mathbf{L}^T) \tag{2.41}$$

Como os parâmetros da população são desconhecidos, pode-se estimar a matriz **Σ** perante a matriz de covariância da amostra $\mathbf{S}_{(p\times p)}$. Considerando as mesmas diretrizes analisadas anteriormente, encontra-se o equacionamento para os modelos estimados para a matriz **S**.

No entanto, muitos dos problemas que envolvem análises multivariadas, as escalas entre as variáveis de resposta são distintas. Deste modo, se faz mais apropriado modelar a matriz de correlação de amostra $\mathbf{R}_{(p\times p)}$, uma vez que esta matriz é insensível à discrepância entre as escalas originais e produz resultados mais precisos quando comparada ao caso em que a matriz **S** é utilizada [8]. Assim, tem-se que a matriz **R** pode ser expressa de acordo com a Eq. (2.42)

$$\mathbf{R} \cong \hat{\mathbf{L}}\hat{\mathbf{L}}^T + \hat{\mathbf{\Psi}} \tag{2.42}$$

Nesse sentido, considerando $m < p$, para n amostras aleatórias com p observações e m fatores, Ferreira [4] estima as cargas fatoriais a partir da Eq. (2.43), enquanto as variâncias específicas podem ser estimadas de acordo com a Eq. (2.44), considerando a matriz diagonal de autovalores de **R** sendo $\hat{\mathbf{\Lambda}}_m = \left[\hat{\lambda}_i\right]$ e a matriz de autovalores normalizadas de **R** iguais a $\hat{\mathbf{P}}_m = \left[\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_m\right]$.

$$\hat{\mathbf{L}} = \hat{\mathbf{P}}_m \hat{\mathbf{\Lambda}}_m^{1/2} = \left[\sqrt{\hat{\lambda}_1}\hat{\mathbf{e}}_1, \sqrt{\hat{\lambda}_2}\hat{\mathbf{e}}_2, \ldots, \sqrt{\hat{\lambda}_m}\hat{\mathbf{e}}_m\right] \tag{2.43}$$

$$\hat{\mathbf{\Psi}} = diag\left(\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}^T\right) = \begin{bmatrix} 1-\hat{h}_1^2 & 0 & \cdots & 0 \\ 0 & 1-\hat{h}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-\hat{h}_p^2 \end{bmatrix} = \begin{bmatrix} 1-\sum_{j=1}^{m}\hat{l}_{1j}^2 & 0 & \cdots & 0 \\ 0 & 1-\sum_{j=1}^{m}\hat{l}_{2j}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-\sum_{j=1}^{m}\hat{l}_{pj}^2 \end{bmatrix} \tag{2.44}$$

De maneira análoga ao que foi observado para a matriz **S**, a variância total do *j-ésimo* fator e a sua respectiva proporção atribuída para a matriz **R**, podem ser expressas pelas Eqs. (2.45) e (2.46), respectivamente. Da Eq. (2.42), é derivado que $\sum_{j=1}^{m}\hat{l}_{ij}^2 = \hat{h}_i^2$, sendo esta a *i-ésima* comunalidade que compreende uma fração da variância total de y_i, explicada por todas as variáveis latentes [5]. O mesmo é observado para a estimativa da *i-ésima* variância específica $\hat{\psi}_i = 1-\sum_{j=1}^{m}\hat{l}_{ij}^2$, *i=1,2,...,p*.

$$Var\left(F_j\right) = \sum_{i=1}^{p}\hat{l}_{ij}^2 = \hat{\lambda}_j \tag{2.45}$$

$$Proporção\left(\sum_i s_{ii}\,\middle|\,F_j\right)=\frac{\sum_{i=1}^{p}\hat{l}_{ij}^{2}}{tr(\mathbf{R})}=\frac{\hat{\lambda}_j}{m} \tag{2.46}$$

2.6.4 Quantidade de fatores

Ao se utilizar técnicas exploratórias como FA, deve-se determinar a quantidade de fatores que serão utilizados no estudo. A partir de uma quantidade *p* de observações, tem-se uma quantidade *m* de fatores comuns, em que um alto grau de correlação entre as *p* observações determina que $m < p$. Para se estimar uma quantidade para o valor de *m*, Rencher [24] estabelece o critério relacionado ao percentual de variância explicada pelos *m* fatores devem ser de pelo menos 80% da variância total, neste caso, de *tr*(**R**).

A partir deste critério, tem-se a capacidade de estabelecer a quantidade *m* de fatores a serem utilizados no estudo, dependendo apenas do conjunto de dados utilizados. Contudo, alguns autores como Visinescu e Evangelopoulos [152] defendem que, ao utilizar fatores rotacionados, tem-se que a interpretação dos mesmos se mantem mais estável a medida que se aumenta a quantidade de fatores. Além disso, esses autores recomendam explorar uma maior quantidade de fatores ao se utilizar técnicas de rotação em FA.

2.6.5 Rotação dos fatores e estruturas simplificadas

A estratégia de rotação em FA é uma prática muito difundida para lidar com a dificuldade de interpretação das cargas fatoriais, pois facilita a associação dos fatores comuns às variáveis de resposta, utilizando uma estrutura de carga mais simples [25]. De forma antagônica ao método PCA, a estratégia FA permite a rotação do fator de cargas (quando $m \geq 2$). Darton [153] afirma que é possível submeter a solução fatorial perante uma rotação por uma matriz não singular. Tal aplicação, muitas vezes, é necessária pois, normalmente, as cargas fatoriais não permitem que o fator explique bem as variáveis observadas.

Diante dessa situação, Thurstone [154] estabeleceu o conceito de simplificação da estrutura de cargas fatoriais, criando algumas especificações para a matriz de cargas fatoriais **L**. Essa simplificação ficou conhecida como "*solução de estrutura simples*" [155], onde Thurstone considerava o princípio da parcimônia na explicação científica. Tais critérios serviram como base para criação de métodos de rotação ortogonal como o *varimax* [38], o qual será detalhado posteriormente. Contudo, esses critérios só podem ser totalmente respeitados em casos particulares, em que tais implicações raramente são encontradas para dados reais. Por este

motivo, os critérios não são totalmente apresentados aqui e maiores detalhes podem ser encontrados na obra do autor, Thurstone [154], além dos trabalhos de Harman [39], Darton [153] e Mulaik [155].

Darton [153] estabelece, diagramaticamente, as cargas fatoriais em uma situação inicial, comparando-a a uma segunda situação, com rotações das cargas fatoriais. Na Tabela 2.13, verifica-se que os asteriscos "*****" representam as maiores cargas fatoriais, enquanto os espaços em branco se referem a cargas fatoriais nulas, ou "*quase zero*".

Tabela 2.13. Cargas fatoriais para uma solução inicial e após a rotação

Observações	*Solução inicial*			*Solução rotacionada*		
	F_1	F_2	F_3	F_1	F_2	F_3
y_1	*	*		*		
y_2	*	*		*		
y_3	*	*	*	*		
y_4	*		*	*	*	
y_5	*	*			*	
y_6	*		*		*	*
y_7	*	*	*			*
y_8	*		*			*

Adaptado de [156]

Analogamente, uma estrutura simples e de fácil interpretação deve promover cargas fatoriais elevadas em um único fator, enquanto os demais fatores apresentam valores moderados [5]. Contudo, tal comportamento nem sempre é verificado, em que a rotação das cargas fatoriais originais se faz necessária, visando uma estrutura de fácil interpretação [25]. É importante ressaltar que as rotações podem ser ortogonais e oblíquas, sendo a primeira capaz de promover vetores de fatores não correlacionados, enquanto os oblíquos apresentam fatores com níveis de correlação significativa. Nesse estudo, como inferido anteriormente, apenas os métodos ortogonais serão detalhados. As obras de Johnson e Wichern [5], Hair *et al.* [40] e Tabachnick e Fidell [41] apresentam maiores detalhes sobre os métodos oblíquos.

Ao analisar o comportamento dos carregamentos de maneira geométrica, tem-se que as cargas da *i-ésima* linha de **L**, correspondem a um ponto no espaço do fator y_i. Assim, diante de uma matriz ortogonal **T**, onde $\mathbf{L}^\circ = \mathbf{LT}$ e $\mathbf{F}^\circ = \mathbf{T}^T\mathbf{F}$, a rotação exibe as coordenadas, mantendo as propriedades geométricas básicas (sendo $\mathbf{TT}^T = \mathbf{I}$). O objetivo principal desta rotação é encontrar uma matriz de carregamento **L** com interpretação simples, uma vez que a matriz

contém as cargas fatoriais representadas pela covariância entre as variáveis originais e os fatores [5,24]. Deste modo, tem-se, na Eq. (2.47) [151], que as comunalidades e variâncias específicas se mantêm idênticas.

$$\mathbf{S} \cong \hat{\mathbf{L}}^{\circ}\hat{\mathbf{L}}^{\circ T} + \hat{\boldsymbol{\Psi}} = \hat{\mathbf{L}}\mathbf{T}\mathbf{T}^{T}\hat{\mathbf{L}}^{T} + \hat{\boldsymbol{\Psi}} = \hat{\mathbf{L}}\hat{\mathbf{L}}^{T} + \hat{\boldsymbol{\Psi}} \tag{2.47}$$

Diante a rotação ortogonal, considerando métodos analíticos e gráficos para encontrar uma estrutura simplificada, Johnson e Wichern [5] concluem que, para pares de cargas de fator (ou seja, $\hat{l}_{i1}, \hat{l}_{i2}$), tem-se *p* pontos, cada um correspondendo a uma variável. Com isso, pode-se rotacionar os eixos das coordenadas através de um ângulo θ, em que as cargas rotacionadas podem ser encontradas a partir da relação descrita na Eq. (2.48) [5].

$$\underset{(p\times 2)}{\hat{\mathbf{L}}} = \underset{(p\times 2)}{\hat{\mathbf{L}}}\,\underset{(2\times 2)}{\mathbf{T}}$$

$$\textit{sendo} \begin{cases} \mathbf{T} = \begin{bmatrix} \cos\theta & \textit{sen}\,\theta \\ -\textit{sen}\,\theta & \cos\theta \end{bmatrix} & \textit{Rotação no sentido horário} \\ \mathbf{T} = \begin{bmatrix} \cos\theta & -\textit{sen}\,\theta \\ \textit{sen}\,\theta & \cos\theta \end{bmatrix} & \textit{Rotação no sentido anti-horário} \end{cases} \tag{2.48}$$

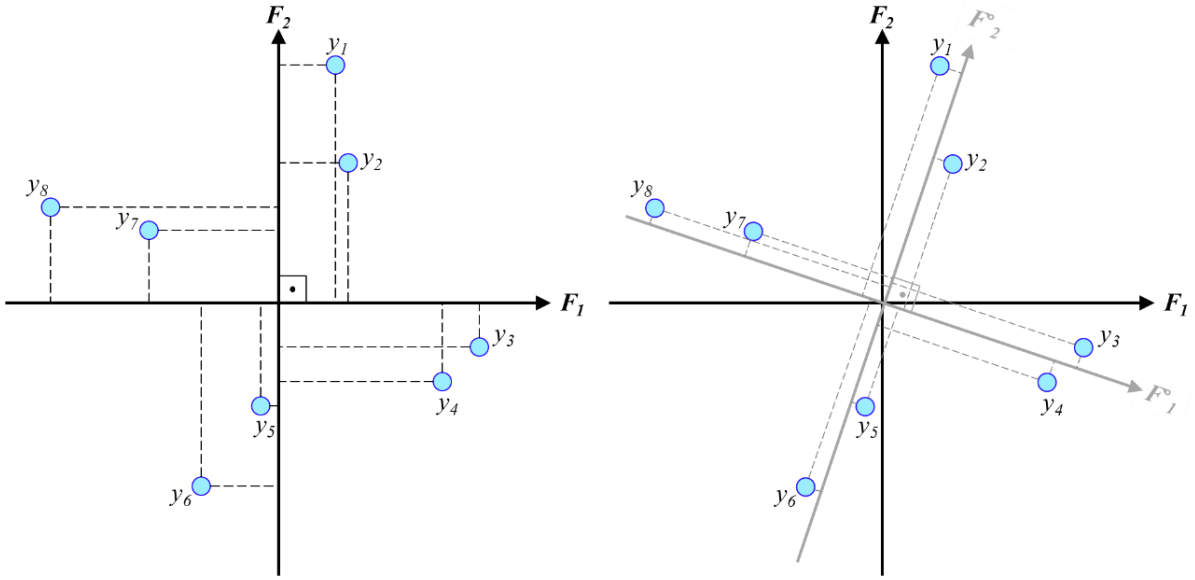
Deste modo, é possível expressar a rotação ortogonal **L°** por meio da Eq. (2.49). Consequentemente, as cargas fatoriais rotacionadas podem ser descritas de acordo com a Eq. (2.50).

$$\mathbf{L}^{\circ} = \mathbf{L}\mathbf{T} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \\ \vdots & \vdots \\ l_{i1} & l_{i2} \\ \vdots & \vdots \\ l_{p1} & l_{p2} \end{bmatrix} \times \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} l_{11}^{\circ} & l_{12}^{\circ} \\ l_{21}^{\circ} & l_{22}^{\circ} \\ \vdots & \vdots \\ l_{i1}^{\circ} & l_{i2}^{\circ} \\ \vdots & \vdots \\ l_{p1}^{\circ} & l_{p2}^{\circ} \end{bmatrix} \tag{2.49}$$

$$\begin{cases} l_{i1}^{\circ} = l_{i1} \times \cos\theta + l_{i2} \times \cos\theta \\ l_{i2}^{\circ} = -l_{i1} \times \sin\theta + l_{i2} \times \cos\theta \end{cases} \tag{2.50}$$

Usualmente, as relações para *m = 2* (indicadas na Eq. (2.48)), são de fácil interpretação. Contudo, quando *m > 2*, as dimensões das cargas rotacionadas ficam mais complexas e sua investigação se faz necessária para encontrar a melhor interpretação dos dados originais. Assim, deve-se definir uma matriz ortogonal que apresente uma estrutura simples e de fácil interpretação. A Figura 2.10 ilustra as variáveis diante do comportamento dos eixos dos fatores

sem rotação (Figura 2.10(a)) e após a rotação (Figura 2.10(b)), onde é possível verificar que os fatores se aproximaram das variáveis originais.



(a) (b)

Figura 2.10. Posição relativa das variáveis (a) sem rotação e (b) com rotação

Com base nesses conceitos, diversos autores propuseram seus métodos de rotação, buscando reduzir a complexidade na explicação das variáveis originais. Uma abordagem comumente utilizada para rotacionar os eixos é o método *quartimax*. A abordagem *quartimax* é caracterizada como um tipo de rotação ortogonal que visa simplificar as colunas de uma matriz fatorial [40], minimizando o termo de produto cruzado, conforme a Eq. (2.51) [153].

$$Quartimax = \sum_{i=1}^{p}\sum_{j=1}^{q}\tilde{l}_{ij}^{\circ 4} + \sum_{i=1}^{p}\sum_{j\neq k}^{q}\tilde{l}_{ij}^{\circ 2}\tilde{l}_{ik}^{\circ 2} \tag{2.51}$$

No entanto, alguns métodos de rotação podem funcionar melhor do que outros, dependendo da estrutura de dados. Deste modo, outras alternativas de rotação são demonstradas na literatura, como o método *varimax*, no qual seleciona uma matriz ortogonal **T** para criar cargas de fator de rotação que promovem a maximização da função objetivo indicada na Eq. (2.52), onde $\tilde{l}_{ij}^{\circ} = l_{ij}^{\circ} / \sqrt{h_i^2}$. Em outras palavras, o método *varimax* representa a relação entre a carga fatorial rotacionada e a *i-ésima* comunalidade.

$$Varimax = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{l}_{ij}^{\circ 4} - \left(\sum_{i=1}^{p}\tilde{l}_{ij}^{\circ 2}\right)^2 \Big/ p\right] \tag{2.52}$$

Segundo Tabachnick e Fidell [41], junto às duas técnicas citadas anteriormente, existem diversas outras estratégias de rotação, como *parimax*, *equamax*, *orthoblique*, entre outras. Além disso, como já inferido anteriormente no capítulo 1, tem-se a técnica denominada *orthomax*, a qual faz uso de um nível de rotação γ para definir a amplitude da rotação, a fim de simplificar as cargas fatoriais. Os valores de γ variam de 0 (variáveis simplificadas, referente a rotação *quartimax*) até 1 (fatores simplificados, referente a rotação *varimax*) [41].

De acordo com Hair *et al.* [40], não há consenso sobre qual o método de rotação que melhor interpreta as variáveis latentes, em que essa escolha pode ser condicionada ao conjunto de dados a ser explorado. Em seu estudo, Visinescu e Evangelopoulos [152], sugerem que, ao se utilizar FA, deve-se realizar uma investigação de diferentes tipos de rotação ortogonal para melhorar a interpretação dos dados. Assim, há necessidade de uma metodologia que permita encontrar o valor ótimo da rotação que melhor interprete as variáveis latentes de um conjunto de dados com uma estrutura de variância-covariância significativa, criando uma estrutura de cargas fatoriais simplificada.

2.6.6 Extração dos escores de fator

Como indicado anteriormente, um dos objetivos da FA por componentes principais se dá pela geração de vetores de respostas não-correlacionados, a partir de um conjunto de dados com estrutura de variância-covariância significativa, sendo essas variáveis conhecidas como "fatores comuns". Segundo Ferreira [4], é possível estimar os fatores comuns para análises subsequentes, sendo estes valores aleatórios não-observáveis conhecidos como escores fatoriais ou escores de fatores [5].

A estimação dos escores fatoriais se dá pelo uso de métodos de regressão. Ao se utilizar a extração por componentes principais, como explanado neste estudo, Johnson e Wichern [5] afirmam que é comumente utilizado o método dos mínimos quadrados ordinários (*OLS – ordinary least squares*). Contudo, outros métodos de estimação podem ser encontrados para essa aplicação, como o dos mínimos quadrados ponderados, por exemplo.

Segundo Johnson e Wichern [5], no OLS, minimiza-se a soma dos quadrados dos resíduos do modelo fatorial (Eq. (2.32)) para se estimar os escores de fator. Considerando o vetor de resíduos de $\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{\mu} - \mathbf{LF}$, tem-se a minimização da Eq. (2.53), em que sua derivada de primeira ordem em relação a $\mathbf{F}$ é igualada a zero, permitindo encontrar a matriz de estimativas para o escores de fator [151]. Assim, diante da utilização da matriz de correlação amostral $\mathbf{R}$, que

considera a matriz de variáveis padronizadas **Z**, pode-se encontrar a expressão para se estimar os escores fatoriais, como descrito na Eq (2.54).

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{y} - \boldsymbol{\mu} - \mathbf{LF})^T(\mathbf{y} - \boldsymbol{\mu} - \mathbf{LF}) \tag{2.53}$$

$$\hat{\mathbf{F}} = \mathbf{Z}\left[\hat{\mathbf{L}}^{\circ}\left(\hat{\mathbf{L}}^{\circ T}\hat{\mathbf{L}}^{\circ}\right)^{-1}\right] \tag{2.54}$$

sendo $\hat{\mathbf{F}}$ a matriz de escores de fator estimado e $\hat{\mathbf{L}}^{\circ}$ representando a matriz de cargas rotacionadas de fator estimados, ambos relacionados à matriz **R**.

Diante de todos os detalhes apresentados sobre as análises exploratórias, tanto a de fatores quanto a de PCA, a próxima subseção irá detalhar o método de análise de cluster, apresentando as métricas de distância, métodos hierárquicos e não-hierárquico, bem como seus respectivos métodos de ligação.

2.7 Análise de cluster

A análise de cluster, também conhecida como análise de conglomerados ou mesmo de agrupamentos, se caracteriza por uma coletânea de técnicas que visam explorar um conjunto de dados, identificando similaridades entre as observações e as variáveis determinadas [50]. Assim, essa estratégia busca promover grupos, alocando as observações que apresentem homogeneidade interna e heterogeneidade entre eles [50]. De modo análogo, a análise de cluster busca fragmentar um grupo com grandes observações em pequenos grupos com características internas semelhantes e dissimilares entre si. Assim, pode-se concluir que essa estratégia busca maximizar a similaridade dentro dos grupos e, consequentemente, minimizar a similaridade entre os grupos.

De acordo com Lattin *et al.* [91], a análise de cluster se relaciona com outra estratégia multivariada, conhecida como escalonamento multidimensional (*MDS – multidimensional scaling*), a qual representa objetos/observações baseados no seu nível de similaridade. Contudo, os autores concluem que a MDS proporciona valores contínuos, enquanto a análise de cluster apresenta valores discretos, também chamados de associações, ou "*memberships*".

Fávero [50] destaca que a análise de cluster é interdependente, ou seja, exploratória. Deste modo, seu uso não apresenta uma característica preditiva para observações externas. Assim, a inclusão de observações adicionais traz a necessidade de realizar uma nova análise. Para estender a análise para confirmar os clusters formados e avaliar predições, pode-se utilizar

estratégias como regressão logística, redes neurais artificiais ou análise discriminante, por exemplo [50].

Para a aplicação adequada da análise de cluster, tem-se a necessidade de estabelecer alguns critérios para realizar a análise, interpretação e comparação dos resultados obtidos, sendo eles [50]:

- Definir os objetivos do estudo;
- Escolher a métrica de distância/semelhança;
- Determinar o esquema de aglomeração.

Este último item se refere ao tipo de abordagem em que a estrutura se subdividirá para criar os grupos. Entre os distintos métodos, pode-se destacar os métodos hierárquicos e não hierárquicos. Os métodos hierárquicos apresentam uma solução *k* grupos, formados pela combinação de dois grupos de solução *k+1*, representando, assim, uma estrutura de árvore [91]. O uso de métodos hierárquicos é amplamente explorado para diversas aplicações, como em estudos apresentados, anteriormente, no capítulo 1. Já os métodos não hierárquicos buscam dividir as observações a partir de um número prévio de clusters, em que as soluções dos grupos (*k* e *k+1*) não são, essencialmente, aninhadas [91]. Pode-se verificar uma gama de aplicações em diversas áreas utilizando métodos não hierárquicos de cluster [157–162].

Muitos dos problemas de aglomerações, independentemente do tamanho, carecem de uma solução heurística. Tal solução se faz necessária, pois a quantidade de soluções de grupos possíveis aumenta à medida que a quantidade de objetos do conjunto de observações cresce. Assim, as possibilidades de divisão de *n* objetos em *m* aglomerados de tamanho n_1, n_2, ... n_m, se dá perante a Eq. (2.55) [91].

$$\frac{n!}{\left[n_1!n_2!n_3!\cdots n_m!m!\right]} \tag{2.55}$$

Fávero [50] ressalta a importância de poder comparar as soluções encontradas por diferentes esquemas de conglomerados (métodos hierárquicos e não hierárquicos), em que o autor afirma que a análise também pode ser cíclica (Figura 2.11 ilustra a estrutura lógica dessa aplicação). Além disso, o mesmo complementa que para escolher a métrica de distância, bem como a de esquema de conglomerado, deve considerar critérios previamente estabelecidos (discutidos, complementarmente, em Bussab *et al.* [51]), concluindo que a escolha de distintas métricas de distância e de esquema de conglomerados podem interferir na formação dos clusters. Contudo, Johnson e Wichern [5] afirmam que a melhor escolha dos métodos de ligação (inferidos nos esquemas de conglomerados), pode variar de acordo com o conjunto de dados

utilizados. Assim, os autores afirmam que é uma boa medida aplicar vários métodos de clusters aliados a pequenas perturbações na unidade de dados para investigar possíveis inversões, além de analisar a variabilidade e concordância dos métodos para um caso particular.

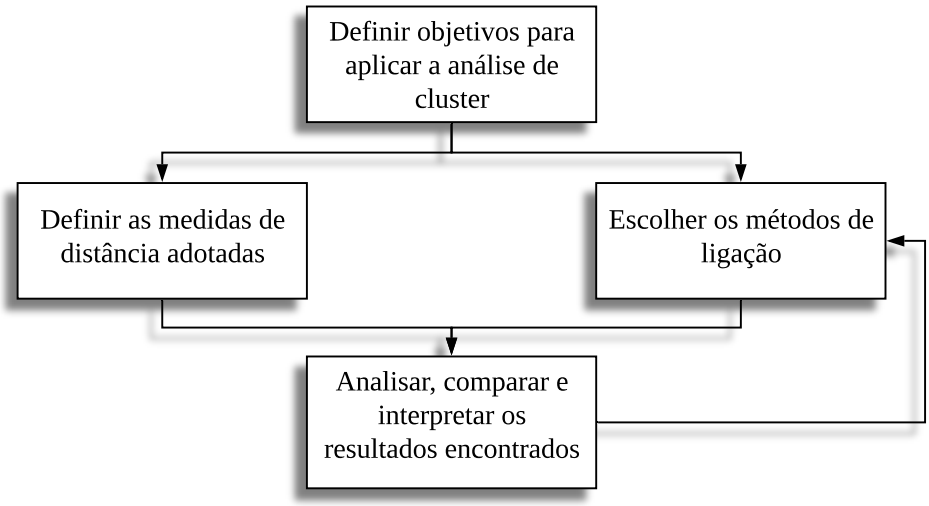


Figura 2.11. Lógica para elaboração da análise de agrupamentos

Adaptado de [50]

As próximas subseções irão abordar, de maneira detalhada, as estratégias que contemplam a análise de cluster, sendo elas: as métricas de distância (como a distância euclidiana e a distância de *Mahalanobis*); os métodos hierárquicos (como os métodos de ligação Único, Centroide, Completa, Média, Mediana, McQuitty e Ward); além do método não hierárquico (*k-médias*).

2.7.1 Métricas de distância

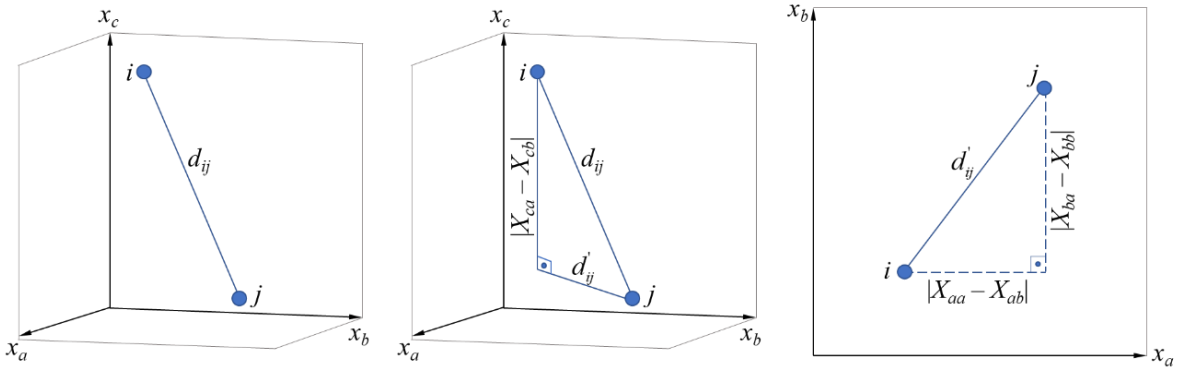
Ao aprofundar na estratégia de análise de cluster, verifica-se que o assunto começa a se caracterizar como uma técnica complexa. Inicialmente, deve-se definir uma métrica, ou medida, de distância para a alocação das observações em determinados grupos. Deste modo, Lattin *et al.* [91] afirmam que ao se utilizar métodos aglomerativos, ou de ligação, de caráter hierárquico, tem-se a necessidade de criar grupos baseado na similaridade (ou em sua mútua proximidade).

Tratando-se de cálculos voltados a métricas de distância, alguns métodos podem ser destacados, os quais serão tratados a seguir.

2.7.1.1 Distância euclidiana

A distância Euclidiana busca apresentar a distância entre dois pontos. Tal formulação pode ser alcançada a partir da consolidada distância de Pitágoras para triângulos retângulos.

Conforme indicado por Fávero [50], ao considerar um caso hipotético para duas observações com três variáveis, conforme ilustrado na Figura 2.12(a), a métrica de distância pode ser encontrada pela projeção dos eixos X_a e X_b diante um plano horizontal, ou seja, a distância d_{ij} (Figura 2.12(b)).



(a) (b) (c)

Figura 2.12. Gráfico de projeção dos pontos d_{ij}

Considerando a distância de Pitágoras $d_{ij} = \sqrt{\left(d'_{ij}\right)^2 + \left(X_{ca} - X_{cb}\right)^2}$, em que $|X_{ca} - X_{cb}|$ se dá pela distância das projeções verticais (X_c) para *i* e *j*, e visto que essa distância não é apresentada, deve-se utilizar as distâncias *i* e *j* nos demais eixos (X_a e X_b) [50]. Tal aplicação pode ser verificada na Figura 2.12(c).

Deste modo, tem-se $d'_{ij} = \sqrt{\left(X_{aa} - X_{ab}\right)^2 + \left(X_{ba} - X_{bb}\right)^2}$ e, substituindo-a pela distância de Pitágoras inicial, é possível encontrar a métrica de distância para este caso particular (entre *i* e *j*), denominada distância Euclidiana (Eq. (2.56)) [50].

$$d'_{ij} = \sqrt{\left(X_{aa} - X_{ab}\right)^2 + \left(X_{ba} - X_{bb}\right)^2 + \left(X_{ca} - X_{cb}\right)^2} \tag{2.56}$$

Assim, tem-se a distância Euclidiana como a métrica mais comum para se projetar a distância de dois objetos, que de maneira sintética, pode ser definida conforme a Eq. (2.57), sendo d_{ij} a distância euclidiana entre objetos *i* e *j*.

$$d_{ij} = \left[\sum_k \left(x_{ik} - x_{jk}\right)\right]^{1/2} \tag{2.57}$$

Lattin *et al.* [91] afirmam que as variáveis, usualmente, apresentam unidades distintas e, consequentemente, o uso dessa métrica se dá por dados padronizados (com finalidade de manter

o mesmo grau de importância para cada variável). Contudo, os autores afirmam que, ao se utilizar estratégias como PCA, deve-se conceder um peso maior para o primeiro componente principal, visto que o mesmo apresenta um maior grau de explicação dos dados.

2.7.1.2 Distância de Mahalanobis

Ao se verificar as distâncias determinadas pela estratégia euclidiana, tem-se que as mesmas não consideram a estrutura de variância-covariância existente nos dados. Assim, *Mahalanobis* propôs um cálculo para ajustar a covariância ao definir a distância [91]. Tal métrica é descrita conforme a Eq. (2.58), determinada por euclidiana generalizada, em que **Σ** representa a matriz de variância-covariância da população de dados da matriz **x**.

$$D_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right) \tag{2.58}$$

Considerando a Figura 2.13, onde Φ e Ω representam pontos de uma distribuição normal multivariada de covariância positiva e centrada na origem, tem-se o locus dos pontos equiespaçados da origem, representados pela elipse [91]. Assim, pode-se dizer que ambos pontos apresentam, para a origem, uma mesma distância de *Mahalanobis*, mesmo que Φ tenha uma distância euclidiana comum menor. Deste modo, a distância de *Mahalanobis* indica que os pontos Φ e Ω apresentam uma probabilidade igual de terem sido escolhidos de uma distribuição normal multivariada com centro na origem, ou mesmo que ambos os pontos apresentam contorno de isodensidade [91].

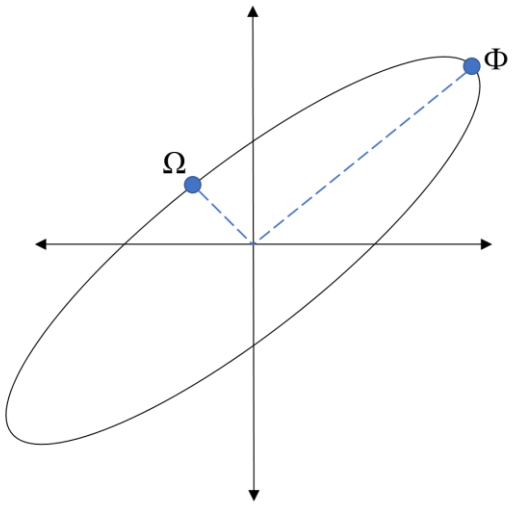


Figura 2.13. Contorno de isodistância entre duas dimensões Φ e Ω

2.7.1.3 Outras métricas de distância e procedimento de padronização

Além das métricas de distância euclidiana e distância D^2 de Mahalanobis, é possível verificar diversas outras estratégias para o cálculo de distância em análise de cluster, com suas particularidades. Algumas das alternativas mais conhecidas estão descritas na Tabela 2.14, com suas respectivas características. Tais métricas não serão exploradas nesse trabalho, contudo é possível encontrar maiores informações nos trabalhos de [163–167].

Tabela 2.14. Medidas de distância alternativas a Euclidiana

Métrica de distância	Formulação	Característica
Manhattan	$d_{ij} = \sum_{p=1}^{k} \left\| X_{pi} - X_{pj} \right\|$	*Conhecida como distância absoluta. Não considera a geometria, apenas a diferença dos valores.*
Minkowski	$d_{ij} = \left[\sum_{p=1}^{k} \left(\left\| X_{pi} - X_{pj} \right\| \right)^{m} \right]^{1/m}$	*Métrica de distância mais geral, conhecida como "métrica do quarteirão".*
Chebychev	$d_{ij} = máx \left\| X_{pi} - X_{pj} \right\|$	*Conhecida como "distância infinita", é um caso particular para distância de Manhattan (e também para a de Minkowski, quando m tende ao infinito).*
Canberra	$d_{ij} = \sum_{p=1}^{k} \frac{\left\| X_{pi} - X_{pj} \right\|}{\left(X_{pi} + X_{pj} \right)}$	*Utilizada quando existem apenas valores positivos (quantidade de variáveis de 0 a p).*
Pearson	$d_{ij} = \sqrt{\sum_{p=1}^{k} \frac{\left(X_{ip} - X_{jp} \right)^2}{\sigma_p^2}}$	*Considerada uma métrica de similaridade onde* σ_p^2 *é a variância da variável p.*

Antes de realizar o procedimento de aplicação da análise de cluster e das métricas de distância, deve-se verificar a unidade dos dados utilizados no estudo, conforme destacado na seção 2.3.1.1. Caso os dados apresentem unidades distintas, tem-se a necessidade de padronizar as variáveis, a fim de evitar que as variáveis que apresentem uma maior escala, influenciem de maneira arbitrária nas distâncias entre as observações [50]. O método mais utilizado para padronizar as variáveis se dá através do procedimento chamado $Z_{escores}$ [50]. O procedimento $Z_{escores}$ é definido conforme a Eq. (2.59), em que $\bar{X}$ é a média e s o desvio padrão da variável X_j. É importante ressaltar que esse procedimento não altera a distribuição da variável original.

$$ZX_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j} \tag{2.59}$$

Após definir as medidas de distância, deve-se definir o método aglomerativo a ser utilizado. Como inferido anteriormente no começo da seção, dois métodos podem ser

destacados: o método hierárquico e o método não hierárquico. Deste modo, nas próximas seções serão detalhadas cada um dos métodos com suas respectivas particularidades.

2.7.2 Métodos hierárquicos

O método hierárquico de cluster (*HCA – Hierarchical Clustering Analysis*) consiste em técnicas de aglomeração cujo objetivo é agrupar objetos com um determinado nível de similaridade. Esses métodos de agrupamento começam com objetos únicos, o que significa que, inicialmente, o número de clusters é igual ao número total de objetos. Em seguida, objetos semelhantes formam grupos, que são mesclados de acordo com suas semelhanças. À medida em que a similaridade é reduzida, os subgrupos tendem a formar um único cluster [5]. Nesse contexto, diferentes métodos de ligação são apresentados.

2.7.2.1 Método de ligação Único

O método de ligação Único (*Single linkage method*), consiste na formação de grupos pela fusão de aglomerados que apresentam maior similaridade, ou seja, menor separação [5]. Assim, para dois clusters, *A* e *B*, é necessário minimizar a distância entre os vetores de observação $\mathbf{y}_i$ (d_{ij}) e $\mathbf{y}_k$ (d_{kl}), para *A* e *B*, respectivamente. Assim, tem-se *n* observações ($i,j,k,l = 1, 2, \ldots, n$), que pode ser descrito pela Eq. (2.60).

$$D_{único}(A,B) = \min\left\{d\left(\mathbf{y}_i, \mathbf{y}_k\right)\right\} \tag{2.60}$$

onde $D_{único}(A, B)$ se refere a distância entre os clusters *A* e *B*; $\mathbf{y}_i$ a distância entre *i* e *j* (d_{ij}) e $\mathbf{y}_k$ a distância entre *k* e *l* (d_{kl}).

2.7.2.2 Método de ligação Completa

O método de ligação Completa (*Complete linkage method*), também é conhecido como método do vizinho mais distante (*farthest neighbor method*). Diferentemente da abordagem "Único", este método busca maximizar a distância de dois objetos entre os clusters [168]. Novamente, sejam *A* e *B* dois clusters diferentes, o método de ligação Completa separa os objetos mais distantes [24], conforme mostrado na Eq. (2.61).

$$D_{completa}(A,B) = \max\left\{d\left(\mathbf{y}_i, \mathbf{y}_k\right)\right\} \tag{2.61}$$

2.7.2.3 Método de ligação Média

O método de ligação Média (*Average linkage method*) define a distância entre dois clusters como a distância média entre todos os pares de objetos, em que cada item pertence a um cluster específico [5,24,62]. Segundo Rencher [24], a distância entre dois clusters pode ser expressa como mostrado na Eq. (2.62), onde n_A e n_B, representam o número de objetos no cluster *A* e *B*, respectivamente.

$$D_{média}(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d\left(\mathbf{y}_i, \mathbf{y}_k\right) \tag{2.62}$$

2.7.2.4 Método de ligação Centroide

O Centroide de um cluster é definido como seu centro de massa e a distância entre os centroides dos clusters define a similaridade entre eles [169]. Portanto, sejam *A* e *B* dois clusters diferentes, a distância entre eles depende da distância euclidiana entre seus centroides, ou seja, os vetores médios $\bar{\mathbf{y}}_A$ e $\bar{\mathbf{y}}_B$,, respectivamente. Esta formulação é indicada nas Eqs. (2.63) e (2.64), apresentando a média ponderada, que calcula o centroide do novo cluster *AB*.

$$D_{centroide}(A,B) = d(\bar{\mathbf{y}}_A, \bar{\mathbf{y}}_B) = \sqrt{(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)^T (\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)} \tag{2.63}$$

$$\bar{\mathbf{y}}_{AB} = \frac{n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B}{n_A + n_B} \tag{2.64}$$

onde $\bar{\mathbf{y}}_A = \sum_{i=1}^{n_A} \mathbf{y}_i / n_A$ e $\bar{\mathbf{y}}_B = \sum_{i=1}^{n_B} \mathbf{y}_i / n_B$.

2.7.2.5 Método de ligação Mediana

O método de ligação Mediana (*median linkage method*) calcula a distância mediana entre os elementos de diferentes grupos e a Eq. (2.65) mostra como a matriz de distância é obtida [62]. As variáveis D_{mj}, D_{kj}, D_{lj} e D_{kl} são definidas como as distâncias entre os aglomerados *mi* e *j*; *k* e *j*; *l* e *j*; e *k* e *l*, respectivamente. Destacamos que *m* representa o grupo mesclado que consiste nos clusters *k* e *l*, com *mi* = (*k*, *i*).

$$D_{(mediana)mij} = \frac{D_{kj} + D_{lj}}{2} - \frac{D_{kl}}{4} \tag{2.65}$$

2.7.2.6 Método de ligação McQuitty

O método de ligação de McQuitty e o método de ligação Média são estratégias de agrupamento muito semelhantes. A principal diferença entre eles é que o primeiro aplica a média aritmética para definir os clusters, enquanto o último usa a média ponderada para o mesmo propósito. Assim, a distância entre um cluster *AB* e um dado cluster *C* é calculada conforme a Eq. (2.66) [170].

$$D_{McQuitty(AB-C)} = \frac{D_{AC} + D_{BC}}{2} \tag{2.66}$$

2.7.2.7 Método de ligação Ward

O método de ligação Ward (*Ward linkage method*) mescla dois clusters distintos para minimizar a perda de informação, que é descrita como um aumento no critério da soma dos quadrados dos erros (*ESS – error sum-of-squares*). Agrupando os clusters em um grupo específico de *n* variáveis, *ESS* pode ser descrito conforme a Eq. (2.67) [171].

$$ESS_i = \sum_{j=1}^{n_i} (X_{ij} - \overline{X_i})' (X_{ij} - \overline{X_i}) \tag{2.67}$$

onde $\overline{X_i}$ é a média dos objetos e X_{ij} é a medição multivariada associada ao *j-ésimo* objeto. Uma explicação mais profunda sobre este método pode ser encontrada no estudo de Ward [171].

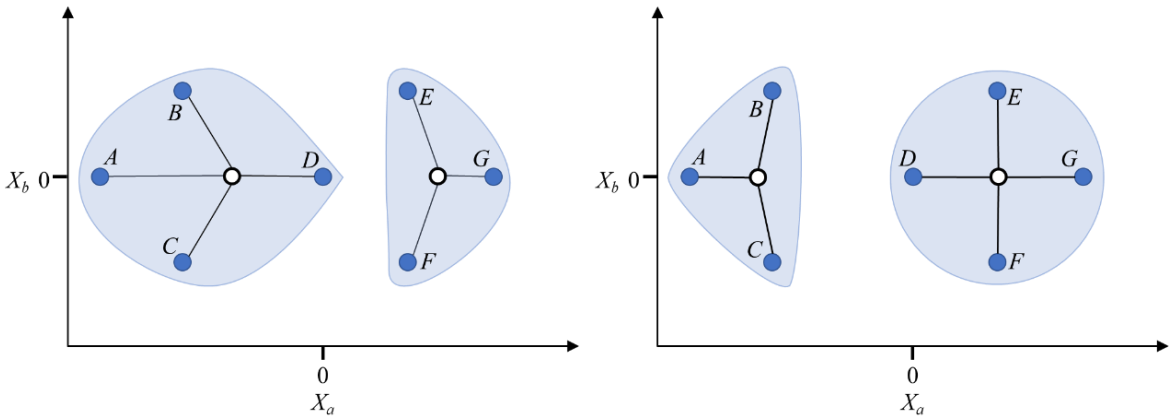
2.7.3 Métodos não hierárquicos

As técnicas de cluster não hierárquico (*NHCA – Non-hierarchical Clustering Analysis*) visam agrupar os itens em um determinado número (*k*) de agrupamentos não sobrepostos. Deste modo, as variáveis entre os grupos serão mais dissimilares possíveis, enquanto as variáveis dentro apresentem um maior nível de similaridade [91]. Uma consideração importante para esses métodos, é que o número de grupos finais deve ser determinado antes de iniciar o procedimento de agrupamento [172]. Além disso, essas técnicas podem ser aplicadas em situações onde a coleta de dados é consideravelmente grande, uma vez que não há necessidade de calcular matrizes de distâncias nem de armazenar dados básicos durante a execução do computador. Uma das técnicas mais populares neste contexto é o método *k-médias* (*k-means*) [5].

O algoritmo descrito pelo método *k-médias* atribui a cada item de um cluster uma média mais próxima [61]. De acordo com Johnson e Wichern [5], a versão mais simples deste procedimento consiste nas três etapas principais listadas abaixo:

1. Inicialmente, particiona-se os itens em *k* clusters distintos e calcula-se as coordenadas dos centroides dos clusters;
2. Em seguida, atribui-se um item ao cluster cujo centroide é o mais próximo. Para isso, calcula-se a distância euclidiana.
3. É necessário recalcular os centroides dos clusters que recebem e perdem um item;
4. Por fim, executa-se a segunda etapa até que não haja mais reatribuições a serem feitas.

Diante do detalhamento desse algoritmo, é possível verificar o comportamento das iterações do agrupamento perante uma situação hipotética, ilustrada na Figura 2.14. A partir de sete observações de duas dimensões (X_a e X_b), verifica-se, na Figura 2.14(a), que as observações *A* à *D* formam o primeiro cluster, enquanto as observações *E* à *G* formam o segundo. Contudo, após uma iteração do método *k-médias* (detalhado anteriormente), tem-se que o objeto *D* é realocado para o segundo cluster (Figura 2.14(b)). Isso acontece devido a sua maior proximidade com o centroide do cluster 2, ou seja, menor valor de soma dos quadrados dos erros em relação ao centroide do cluster 2 do que em relação ao centroide do cluster 1.



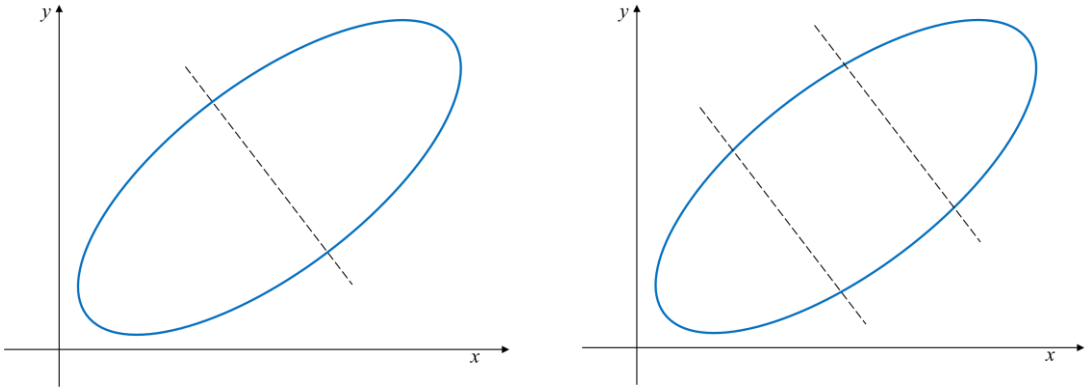
(a) (b)

Figura 2.14. Comportamento do cluster: (a) situação inicial; (b) situação final

Adaptado de [91]

Deste modo, diferentemente do que fora observado nos métodos hierárquicos, em que as soluções "*no topo da árvore*" podem ser realocadas para um nível de similaridade maior, a abordagem não hierárquica, como o *k-médias*, não permite ir de uma solução de dois clusters

para uma de três cluster sem realocar as observações. A Figura 2.15 ilustra esse comportamento para um exemplo hipotético. Portanto, a estratégia *k-médias* busca produzir soluções convexas e compactas, assim como o método "Ward", reduzindo o *ESS* e a heterogeneidade dos agrupamentos [91].



(a) (b)

Figura 2.15. Soluções de partições não hierárquicas para (a) $k = 2$ e (b) $k = 3$

Por fim, após o processo de iteração finalizar e não haver mais observações a serem realocadas, o procedimento *k-médias* apresenta o agrupamento final, em que não há observações com maiores proximidades diante a outros centroides. A Figura 2.16 ilustra um exemplo de agrupamento final pelo método *k-médias*.

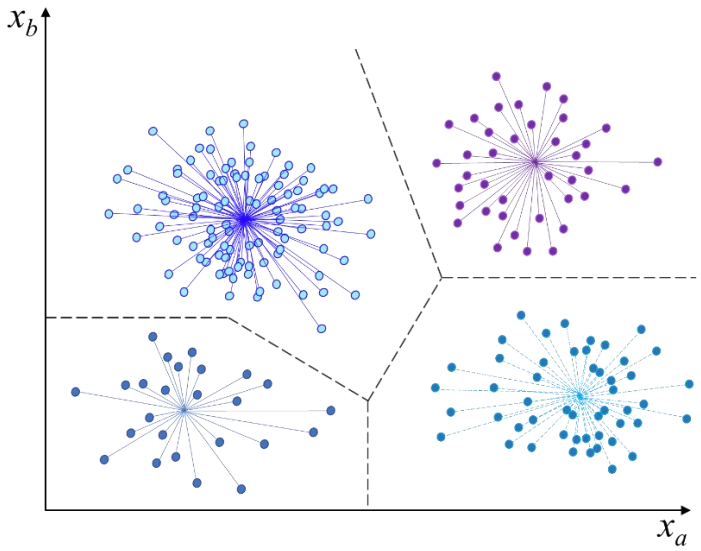


Figura 2.16. Situação hipotética que representa o término do procedimento *k-médias*

2.8 Erro Quadrático Médio

Ao se tratar de múltiplas funções objetivo, a estratégia denominada *erro quadrático médio* se destaca em aplicações de otimização robusta, sendo amplamente investigada em distintos segmentos [12,173–175]. De acordo com Köksoy [176], essa técnica aglutina funções objetivos considerando a soma da variância e o quadrado das diferenças entre o valor alvo e a média da resposta. A minimização dessa função aglutinação, ou seja, o EQM, permite que o valor médio das respostas se aproxime do valor alvo estipulado e, consequentemente, reduza o valor da variabilidade [141]. A formulação do EQM pode ser descrita, conforme a Eq. (2.68) [177,178]. É importante ressaltar que essa abordagem de otimização apresenta como restrição, apenas, a condição de espaço de solução viável [126].

$$Min\left\{EQM = \left[\hat{y}\left(\mathbf{x}\right) - T\right]^2 + \hat{\sigma}^2\left(\mathbf{x}\right)\right. \tag{2.68}$$

onde $\hat{y}$ é o modelo estabelecido para a média, T o valor alvo da resposta (podendo ser definido por diferentes critérios) e σ^2 a variância do modelo.

A expressão apresentada na Eq. (2.68) se trata de uma modelagem de EQM para apenas uma resposta em análise. Tratando de problemas com múltiplas respostas (sendo uma característica de conjuntos de dados multivariados), Köksoy [176] inferiu uma abordagem para o EQM que visa aglutinar diversas funções. Além disso, nessa abordagem, o autor possibilita a implementação de pesos, caso necessário, para tratar de diferentes graus de importância. A Eq. (2.69) descreve esse modelo proposto, sendo w_i os pesos que podem ser atribuídos na função.

$$EQM = \sum_{i=1}^{m} w_i \times EQM_i = \sum_{i=1}^{m} w_i \times \left\{\left[\hat{y}_i\left(\mathbf{x}\right) - T_i\right]^2 + \hat{\sigma}_i^{\,2}\left(\mathbf{x}\right)\right\} \tag{2.69}$$

Diante dessa modelagem, é possível encontrar o valor ótimo a partir de um algoritmo de otimização adequado, visando a minimização da função objetivo.

2.9 Programação quadrática sequencial

Com base nas estratégias estatísticas e matemáticas apresentadas, para aplicar, de fato, a estratégia de otimização, tem-se a necessidade de um algoritmo eficiente e adequado para a convergência. Entre as opções numéricas existentes, o método denominado de *Programação Quadrática Sequencial* se destaca sendo um dos melhores métodos para resolução de problemas não lineares de otimização restrita [179–182]. O SQP se caracteriza como um dos métodos mais recentes, sendo uma abordagem eficiente e precisa comparada a outros métodos de

programação não linear (PNL) para vários problemas [179,180,183]. De acordo com Rao [179], o SQP apresenta uma base teórica vinculada a dois pilares:

- Solucionar um determinado conjunto de equações não lineares a partir do método de Newton;
- Derivar, simultaneamente, equações não lineares a partir das condições Kuhn-Tucker para o Lagrangiano do problema de otimização restrita.

Assim, tem-se que essa abordagem apresenta características diretas em relação ao método de Newton. A partir de cada uma das iterações, realiza-se uma aproximação pela Hessiana da função Lagrangiana, utilizada para gerar um subproblema de programação quadrática. Com a solução desse subproblema, é possível definir uma direção de busca considerando um problema de busca linear [180]. Deste modo, o SQP transforma um problema não linear em subproblemas quadráticos, resolvendo-os conforme a Eq. (2.70) [181,184–187].

$$\begin{cases} Min \quad S(d) = \nabla f(x)^T d + \frac{1}{2} d^T \mathbf{B}_k d \\ sujeito \;\; a: g(x_k) + \nabla g(x_k)^T d \le 0 \end{cases} \tag{2.70}$$

onde ∇_x se caracteriza como a derivada parcial de primeira ordem, *g* representa o vetor da função de restrição, *d* se caracteriza como a direção de busca do processo iterativo, enquanto **B** representa a matriz Hessiana. A partir das Eqs. (2.71), (2.72) e (2.73), a matriz Hessiana (**B**) pode ser encontrada utilizando a fórmula de Broyden – Fletcher – Goldfarb – Shanno (BFGS).

$$\mathbf{B}^{(k+1)} = \mathbf{B}^{(k)} - \frac{\mathbf{B}^{(k)} \delta \delta^k \mathbf{B}^{(k)}}{\delta^T \mathbf{B}^{(k)} \delta} + \frac{\gamma_L {\gamma_L}^T}{\delta^T} \tag{2.71}$$

$$\delta = x^{(k+1)} - x^{(k)} \tag{2.72}$$

$$\gamma_L = \nabla_x L\left(x^{(k+1)}, \lambda^{(k)}\right) - \nabla_x L\left(x^{(k)}, \lambda^{(k)}\right) \tag{2.73}$$

onde γ_L representa a diferença de gradiente das funções de Lagrange.

A partir dessa abordagem, é possível encontrar o ponto ótimo de funções não lineares, em que o SQP se destaca, sendo utilizado em diversas abordagens. Esse método carece de uma quantidade menor de avaliações para definir soluções ótimas, comparadas a algoritmos que fazem uso de meta-heurísticas. Contudo, não apresenta a capacidade de se deslocar de ótimos locais, além de apresentar ruídos nas funções objetivos ou restrições, por utilizar informações de gradiente no método de busca [180]. O algoritmo para o método SQP pode ser verificado, passo a passo, a partir do fluxograma ilustrado na Figura 2.17.

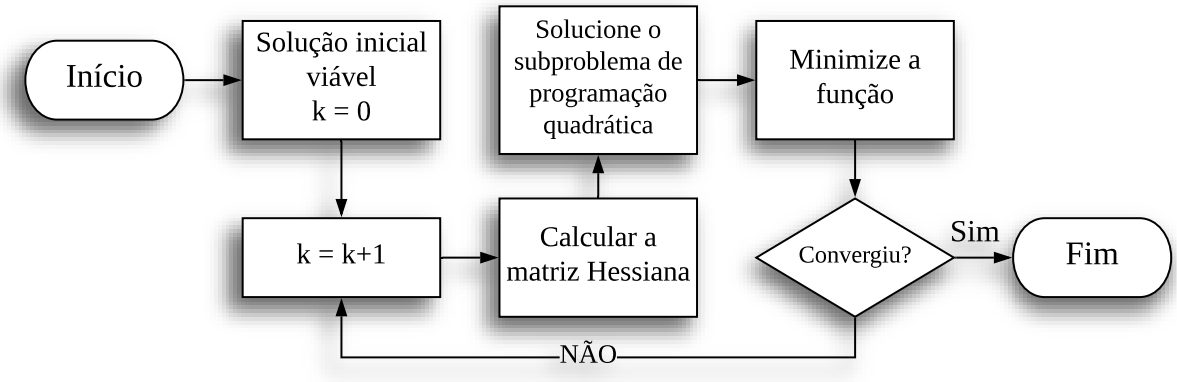


Figura 2.17. Fluxograma para o algoritmo SQP

2.10 Intervalos e regiões de confiança

Um modo básico para apresentar os valores estimados, resultantes de diversas resoluções, se dá por caracterizar uma amplitude de possíveis valores, que são comumente chamados de *intervalos de confiança*. Ao se tratar de uma análise univariada, tem-se intervalos com uma única dimensão, contudo, ao se tratar de duas ou mais variáveis, diante da sua natureza correlacionada, tem-se o que é conhecido como *regiões de confiança*. Segundo Montgomery e Runger [63], a precisão de um intervalo está diretamente relacionada ao comprimento do mesmo, em que um intervalo curto infere em uma estimação com maior precisão. De modo antagônico, um intervalo maior, caracteriza em um menor nível de precisão. Assim, para fins discriminatórios, busca-se intervalos, ou regiões, não sobrepostas, com maior nível de distinção e precisão, promovendo uma melhor separabilidade e qualidade na tomada de decisão. Deste modo, os próximos tópicos visam abordar sobre os assuntos de intervalos e regiões de confiança, apresentando os tratamentos para dados univariados e multivariados, respectivamente.

2.10.1 Intervalo de confiança

O desenvolvimento básico de um IC fica evidente ao se considerar dados populacionais normalmente distribuídos, assumindo uma média desconhecida com uma variância conhecida [63]. Considerando uma amostra aleatória X_1, X_2, ..., X_n, em que $\bar{X}$ é a média amostral, sendo padronizada conforme a Eq. (2.74). A estimativa do IC para μ, se dá pela variação de $l \leq \mu \leq u$, em que estes valores dependem da amostra, considerados valores de variáveis aleatórias de L e U [63]. Supondo a determinação desses valores, em que $P\{L \leq \mu \leq U\} = 1 - \alpha$ (onde α varia de

0 a *1*), Montgomery e Runger [63] concluem que existe uma probabilidade $1 - \alpha$ de se escolher uma amostra onde μ é verdadeira, dentro de um IC. Nesta hipótese, l e u são os limites inferiores e superiores desse intervalo, sendo $1 - \alpha$ o coeficiente de confiança.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \tag{2.74}$$

De modo subsequente, a partir do que foi apresentado e considerado na Eq. (2.74), tem-se a Eq. (2.75) e, por consequência, chega-se a Eq. (2.76).

$$P\left\{-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right\} = 1 - \alpha \tag{2.75}$$

$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha \tag{2.76}$$

Montgomery e Runger [63] afirmam que tal equacionamento se refere a um intervalo aleatório, pois os limites $\bar{X} \pm z_{\alpha/2}\,\sigma/\sqrt{n}$ que contemplam a variável $\bar{X}$, referem-se aos limites de confiança de *L* e *U*. Assim, os autores afirmam que, ao considerar $\bar{X}$ como uma média amostral aleatória de um conjunto *n*, sendo esta de uma população com variância conhecida, o IC de $100(1 - \alpha)\%$ de μ é definido conforme a Eq. (2.77).

$$\bar{x} - z_{\alpha/2}\,\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2}\,\sigma/\sqrt{n} \tag{2.77}$$

onde $z_{\alpha/2}$ refere-se ao valor superior, considerando $100\alpha/2\%$ de uma distribuição normal. A Figura 2.18 ilustra o comportamento de vários intervalos de confiança para a média μ.

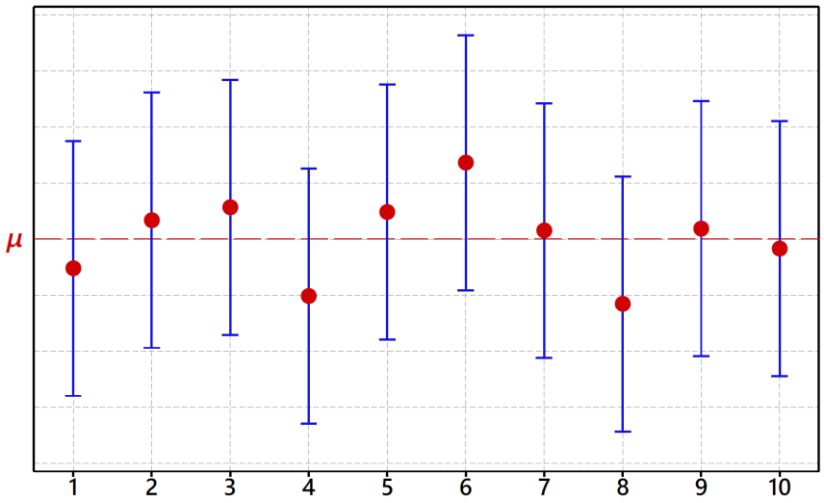


Figura 2.18. Construção de intervalos de confiança univariados para μ

Por fim, Montgomery e Runger [63] finalizam que o tamanho do IC se caracteriza como uma métrica de precisão para a estimação, em que visa-se encontrar um IC estreito o suficiente para favorecer a tomada de decisão, como em classificações de subestação para qualidade na distribuição de energia [61,62].

2.10.2 Regiões de confiança

A natureza multivariada está presente nos mais diversos campos de estudo, em que, tratar um conjunto de dados (com uma correlação significativa) de maneira univariada, pode negligenciar informações essenciais em processos decisórios. De acordo com Johnson e Wichern [5], para a realização de inferências diante uma amostra desse tipo, inicialmente é preciso estender o conceito de intervalos de confiança para o conceito de regiões de confiança multivariadas. Regiões de confiança visam apresentar o espaço viável para a precisão de um resultado com duas ou mais variáveis. Tais regiões são usualmente representadas por elipses de confiança (em um caso bidimensional), baseadas na distância de Mahalanobis. Este conceito é amplamente utilizado em diversos estudos, contemplando diferentes segmentos e aplicações [7,8,62,188].

Segundo Ferreira [4], a densidade de uma distribuição normal multivariada está diretamente relacionada com a expressão $(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, sendo a mesma referente à distância de Mahalanobis entre a média $\boldsymbol{\mu}$ e a observação $\boldsymbol{x}$, como descrita anteriormente na Eq. (2.58). Tal métrica infere em hiperelipsoides para um espaço dimensional *p*. Diante do comportamento intrínseco de dados multivariados, tem-se que conjuntos de dados apresentam comportamento elipsoidal quando existe uma correlação significativa e, de modo antagônico, quando as variáveis são independentes (correlação igual a zero), os dados apresentam um comportamento esférico. A Figura 2.19 ilustra este comportamento para uma situação bidimensional.

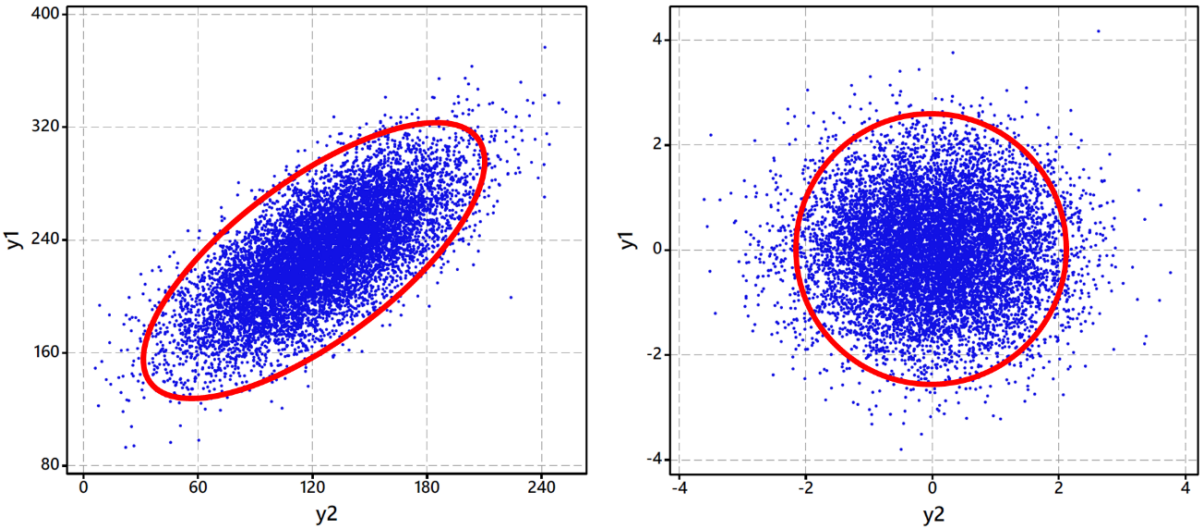
Perante o equacionamento referente a distância de Mahalanobis, Johnson e Wichern [5] afirmam que a densidade da variável aleatória $\boldsymbol{x}$ é constante no elipsoide $\boldsymbol{\mu}$ centrado, conforme a Eq. (2.78).

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)=c^{2} \tag{2.78}$$

O hiperelipsoide apresentado tem o comprimento de suas direções definidos pela $\sqrt{\lambda_i}$, dos autovalores de $\boldsymbol{\Sigma}$ [4,5]. Considerando um modelo com duas dimensões, o hiperelipsoide é

representado por uma elipse, em que, a partir da Eq. (2.78), os contornos podem ser definidos conforme descrito na Eq. (2.79).

$$c^2 = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}}\right)\right] \tag{2.79}$$



(a) (b)

Figura 2.19. Comportamento de dados com correlação: (a) *r = +0.705*; (b) *r = 0*

Deste modo, diante da matriz **Σ**, pode-se definir os autovalores (Eq. (2.80)), em que os valores positivos e negativos se referem ao maior e menor autovalor, respectivamente [4]. De modo similar e, considerando que os semieixos são iguais a $\pm c\sqrt{\lambda_i}e_i \quad \forall i = 1, 2, ..., p$, é possível determinar os autovetores de **Σ**, como descrito na Eq. (2.81) [8].

$$\lambda_i = \frac{\sigma_{11}+\sigma_{22} \pm \sqrt{(\sigma_{11}-\sigma_{22})^2 + 4\sigma_{12}^2}}{2} \quad ; \quad i = 1, 2 \tag{2.80}$$

$$\cos(\theta) = \frac{\sigma_{12}}{\sqrt{(\lambda_1-\sigma_{11})^2 + \sigma_{12}^2}} \tag{2.81}$$

Assim, diante do que foi apresentado na Eq. (2.78), Johnson e Wichern [5] definem que a elipse pode ser descrita conforme a Eq. (2.82).

$$\mathbf{x}^{\mathbf{T}}\left[\Sigma^{-1}\right]\mathbf{x} = \left(\frac{x_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2}{\sqrt{\sigma_{22}}}\right)^2 = \frac{x_1^2}{\sigma_{11}} + \frac{x_2^2}{\sigma_{22}} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} (1/\sigma_{11}) & 0 \\ 0 & (1/\sigma_{22}) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c^2 \tag{2.82}$$

A partir disso, tem-se, pela Eq. (2.83), que:

$$\begin{aligned}\mathbf{x}^{\mathbf{T}}\left[\Sigma^{-1}\right]\mathbf{x} &= \mathbf{x}^{\mathbf{T}}\left[\frac{\left(e_1e_1^T\right)}{\lambda_1}+\frac{\left(e_2e_2^T\right)}{\lambda_2}\right]\mathbf{x} = \frac{\left(\mathbf{x}^T e_1\right)^2}{\lambda_1}+\frac{\left(\mathbf{x}^T e_2\right)^2}{\lambda_2} = c^2 \\ &= \mathbf{x}^{\mathbf{T}}\left[\frac{1}{\lambda_1}e_1e_1^T+\frac{1}{\lambda_2}e_2e_2^T\right]\mathbf{x} = \frac{1}{\lambda_1}\left(\mathbf{x}^{\mathbf{T}}e_1\right)^2+\frac{1}{\lambda_2}\left(\mathbf{x}^{\mathbf{T}}e_2\right)^2 = c^2 \\ &= \frac{1}{c^2\lambda_1}\left(\mathbf{x}^{\mathbf{T}}e_1\right)^2+\frac{1}{c^2\lambda_2}\left(\mathbf{x}^{\mathbf{T}}e_2\right)^2 = 1\end{aligned} \tag{2.83}$$

Deste modo, se considerar o autovetor e_2 igual a 0, tem-se que $1/\lambda_1(\mathbf{x}^{\mathbf{T}}e_1)^2 = c^2$, logo $\mathbf{x}^{\mathbf{T}}e_1 = c\sqrt{\lambda_1}$. De modo similar, considerando o autovetor $e_1 = 0$, $1/\lambda_2(\mathbf{x}^{\mathbf{T}}e_2)^2 = c^2$, conclui-se que $\mathbf{x}^{\mathbf{T}}e_2 = c\sqrt{\lambda_2}$. Por consequência, chega-se ao comprimento dos semieixos da elipse de densidade constante, conforme ilustrado na Figura 2.20.

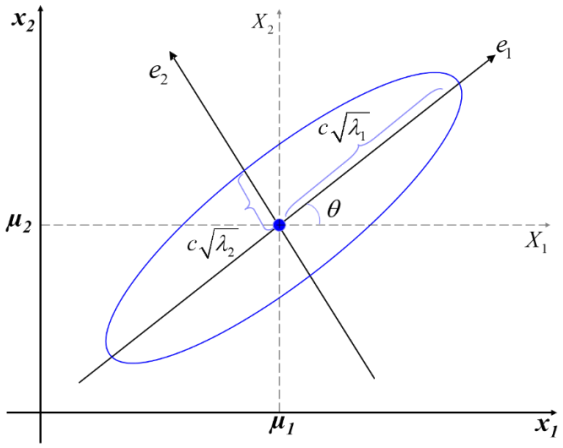


Figura 2.20. Elipse de densidade constante

De modo similar, Segundo Almeida *et al.* [188], o elipsoide de **Σ** pode ser definido a partir da decomposição espectral da Eq. (2.78), como descrito na Eq. (2.84).

$$\begin{aligned}&\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) = c^2 \\ &\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathbf{T}}\mathbf{P}\Lambda^{-1}\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right) = c^2 \\ &\left[\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{\mathbf{T}}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}^{-1/2}\left[\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right] = c^2 \\ &\sqrt{\left[\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]^{\mathbf{T}}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}^{-1/2}\left[\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]} = \sqrt{c^2} \\ &\boldsymbol{\Lambda}^{-1/2}\left[\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right] = \sqrt{\chi^2}\end{aligned} \tag{2.84}$$

Sendo ortonormal a matriz para autovetores de **Σ** [188], ou seja, $\mathbf{P}^{T} = \mathbf{P}^{-1}$, pode-se chegar a Eq. (2.85):

$$\begin{aligned}
\left[\mathbf{P}^{\mathbf{T}}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right] &= \sqrt{\chi^2}\,\boldsymbol{\Lambda}^{1/2} \\
\left[\mathbf{P}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right] &= \sqrt{\chi^2}\,\boldsymbol{\Lambda}^{1/2} \\
\left(\mathbf{x}-\boldsymbol{\mu}\right) &= \mathbf{P}\left[\sqrt{\chi^2}\,\boldsymbol{\Lambda}^{1/2}\right] \\
\mathbf{x} &= \boldsymbol{\mu} + \mathbf{P}\left[\sqrt{\chi^2}\,\boldsymbol{\Lambda}^{1/2}\right]
\end{aligned} \tag{2.85}$$

Sabendo que: $\mathbf{x} = \boldsymbol{\mu} + \mathbf{P}\left[\sqrt{\chi^2}\,\boldsymbol{\Lambda}^{1/2}\right]$; $\mathbf{P} = \begin{bmatrix} \cos\theta & -\text{sen}\theta \\ \text{sen}\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix}$ e $\boldsymbol{\Lambda}^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix}$, é possível obter a equação para as elipses paramétricas (α) e as elipses rotacionadas (θ), a partir da Eq. (2.86). Os detalhes matemáticos das relações trigonométricas, para a rotação das elipses de confiança, estão descritos no Apêndice A.

$$\begin{aligned}
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + c \times \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \times \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \times \begin{bmatrix} \cos\alpha \\ \text{sen}\alpha \end{bmatrix} \\
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \cos\theta & -\text{sen}\theta \\ \text{sen}\theta & \cos\theta \end{bmatrix} \times \begin{bmatrix} \sqrt{\chi^2\lambda_1} & 0 \\ 0 & \sqrt{\chi^2\lambda_2} \end{bmatrix} \\
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \cos\theta & -\text{sen}\theta \\ \text{sen}\theta & \cos\theta \end{bmatrix} \times \begin{bmatrix} \sqrt{\chi^2\lambda_1} & 0 \\ 0 & \sqrt{\chi^2\lambda_2} \end{bmatrix} \times \begin{bmatrix} \cos\alpha \\ \text{sen}\alpha \end{bmatrix} \\
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \times \begin{bmatrix} \sqrt{\chi^2\lambda_1} & 0 \\ 0 & \sqrt{\chi^2\lambda_2} \end{bmatrix} \times \begin{bmatrix} \cos\alpha \\ \text{sen}\alpha \end{bmatrix}
\end{aligned} \tag{2.86}$$

A partir da Eq. (2.86), pode-se obter o modelo generalizado, como descrito na Eq. (2.87).

$$\begin{aligned}
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} c\sqrt{\lambda_1}e_{11}\cos\alpha - c\sqrt{\lambda_2}e_{12}\text{sen}\alpha \\ c\sqrt{\lambda_1}e_{21}\cos\alpha + c\sqrt{\lambda_2}e_{22}\text{sen}\alpha \end{bmatrix} \\
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + c \times \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \times \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \times \begin{bmatrix} \cos\alpha \\ \text{sen}\alpha \end{bmatrix} \\
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \sqrt{\chi^2_{(p,\alpha)}} \times \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \times \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \times \begin{bmatrix} \cos\theta \\ \text{sen}\theta \end{bmatrix} \\
\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{(p,n-p)}(\alpha)} \times \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \times \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} \times \begin{bmatrix} \cos\theta \\ \text{sen}\theta \end{bmatrix}
\end{aligned} \tag{2.87}$$

onde $c = \sqrt{\chi^2_{(p,\alpha/2)}}$; $0 \le \alpha \le 2\pi$. $\begin{bmatrix}\mu_1 & \mu_2\end{bmatrix}^T$ representa o vetor de médias, $\begin{bmatrix}\lambda_1 & \lambda_2\end{bmatrix}^T$ e $\begin{bmatrix}e_{11} & e_{12}\\ e_{21} & e_{22}\end{bmatrix}$ representam o vetor de autovalores e a matriz de autovetores de **Σ**, respectivamente. *p* indica o número de variáveis analisadas e *n* é o número de observações no conjunto de dados.

Por fim, a região de confiança, no caso multivariado, pode ser definida considerando o vetor de médias e a matriz de variância-covariância conhecida, em que, segundo Ferreira [4], uma região de $100(1-\alpha)\%$ de confiança para **μ**, é definida pela Eq. (2.88).

$$\text{Região de confiança} = \left\{ n(\boldsymbol{\mu}-\bar{\mathbf{y}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}-\bar{\mathbf{y}}) \le \frac{\nu p}{\nu+1-p} F_{\alpha,p,\nu+1-p} \right\} \tag{2.88}$$

sendo $v = n - 1$ e $F_{\alpha, p, v+1-p}$ os graus de liberdade e o quantil $100\alpha\%$ superior da distribuição *F*. Tal região representa o hiperelipsoide de distância $\chi^2_{\alpha,p}/n$ constante entre **μ** e $\bar{\mathbf{y}}$ [4]. Graficamente é possível gerar essas regiões de confiança através de elipsoides ($p = 3$) e elipses ($p = 2$), sendo a última, um reflexo bidimensional de um elipsoide (considerando as variáveis dos seus respectivos eixos). A Figura 2.21 apresenta ambos comportamentos. Conhecendo as projeções da região de confiança, uma boa medida é verificar os limites bilaterais de Bonferroni, conforme discutido em Almeida *et al*. [8].

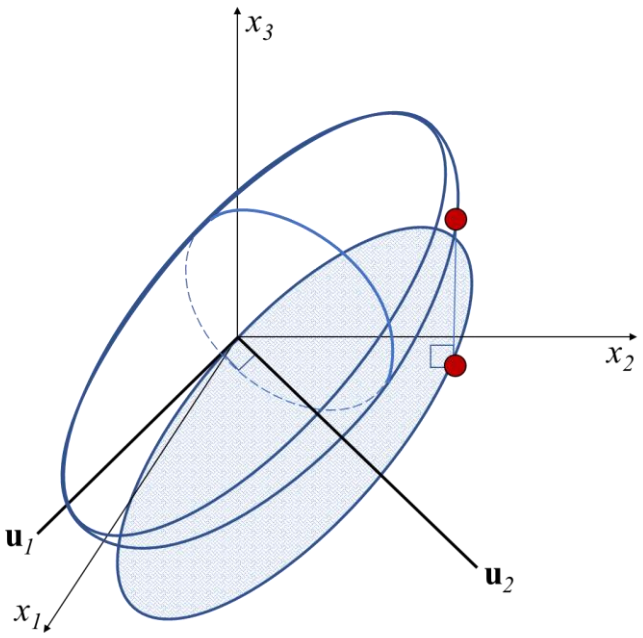


Figura 2.21. Projeção de um elipsoide tridimensional para um plano bidimensional $\mathbf{u}_1$ e $\mathbf{u}_2$

2.11 Considerações finais

As seções descritas neste capítulo apresentam os fundamentos teóricos necessários para o desenvolvimento do presente estudo. Deste modo, diante das definições apresentadas, é possível compreender a proposta principal desse trabalho, a qual será apresentada no capítulo seguinte. É importante destacar que as técnicas de FA e análise de cluster foram apresentadas com maiores detalhes, dado a grande relevância das mesmas na proposta deste trabalho. No que se refere à essas estratégias, verifica-se a oportunidade de desenvolver métodos que auxiliam a calibração dos parâmetros das técnicas multivariadas, em que as abordagens de planejamento de experimentos se fazem adequadas. Além disso, através das discussões apresentadas em diversos estudos, verifica-se a oportunidade de criar métodos para agrupamentos ótimos, bem como para verificar a robustez dos agrupamentos com base nos métodos de ligação, baseado na instabilidade desses métodos com dados discrepantes. Assim, novas estratégias podem ser desenvolvidas a fim de contribuir com as questões teóricas e práticas das técnicas multivariadas, utilizando de outras estratégias disponíveis na literatura, as quais este tudo também irá investigar.

3. MÉTODO PARA O APRIMORAMENTO DO PODER DISCRIMINATÓRIO DE FUNÇÕES ELIPSOIDAIS

Diante dos aspectos discutidos nos capítulos anteriores, verifica-se a ascensão do uso de técnicas que buscam analisar uma grande extensão de dados, visando auxiliar a tomada de decisão baseado na maior quantidade de informações disponíveis. Muitas dessas estratégias são utilizadas para cumprir uma certa necessidade, contudo, algumas combinações de técnicas são pouco exploradas ou mesmo negligenciadas, diante do comportamento do conjunto de dados e do objetivo principal de cada pesquisa. Com base nessas constatações, o presente capítulo visa apresentar uma abordagem nova para aprimorar a exploração e análise de dados correlacionados que podem, ou não, apresentar grandes quantidades de informações, melhorando a identificação e precisão de padrões similares, estimando regiões de confiança mais estreitas e concisas. Em outras palavras, busca-se aprimorar o poder discriminatório de funções elipsoidais, utilizadas para representar a região de confiança de cluster formados, com base em escores de fatores sob rotações ortogonais. Para tal, considera-se estratégias experimentais, como o arranjo fatorial multiníveis e o arranjo de misturas, métodos hierárquicos e não hierárquicos de clusters (Único, Centroide, Completa, Média, Mediana, McQuitty, Ward e *k-médias*), métodos de análise, como ANOVA e ANCOVA, aglutinação e otimização de funções objetivo e a técnica exploratória multivariada de análise fatorial.

3.1 Modelagem do método proposto

O método proposto combina o uso de diferentes técnicas estatísticas e matemáticas para estimar agrupamentos otimizados, melhorando a discriminação dos mesmos ao gerar elipses de confiança estreitas. Para isso, combina-se a estratégia de planejamento experimental e otimização para analisar e encontrar a melhor combinação de métodos de cluster, ao estimar agrupamentos otimizados, representando-os através de regiões de confiança. Com base nas discussões sobre os métodos de agrupamento, apresentada no capítulo anterior, essa estratégia busca apresentar um método conciso para realizar agrupamentos das informações de um conjunto de dados com estrutura de variância-covariância significativa, estimando as regiões de confiança através de funções elipsoidais, com melhor nível de precisão.

A extensão total dessa abordagem é complementada por outros dois métodos auxiliares, desenvolvidos para atender as carências da proposta inicial de aprimoramento do poder discriminatório de elipses de confiança para agrupamentos. Entre eles, o primeiro refere-se a

um método para encontrar o melhor nível de rotação ortogonal das cargas fatoriais, baseado em arranjos de misturas, que está inserido na proposta inicial. Essa abordagem irá auxiliar a calibração da rotação, visto que existe, na literatura, uma discussão sobre o melhor nível de rotação, que depende da estrutura dos dados empregados, conforme apresentado no capítulo 2.

O segundo desenvolvimento se dá pela necessidade de uma confirmação da robustez dos resultados de cluster para diferentes cenários/conjunto de dados, visto que alguns métodos de ligação apresentam a sensibilidade à valores discrepantes [5,8]. Assim, esta segunda análise será detalhada separadamente na subseção posterior (Subseção 3.2). A abordagem completa do método pode ser verificada através do pseudocódigo disponível no Apêndice B. Deste modo, tem-se a primeira etapa do método proposto descrito em 9 passos, destacadas detalhadamente a seguir e delineada, de modo ilustrativo, na Figura 3.1.

Passo 1: A partir de um conjunto de dados a serem analisados, inicialmente é necessário investigar o comportamento dos dados através de critérios estatísticos básicos, avaliando sua estrutura de variância-covariância e, consequentemente, se há redundância e dependência entre as variáveis analisadas. Caso os dados sejam totalmente independentes, não há necessidade de uma exploração dos mesmos utilizando técnicas multivariadas. Contudo, se os dados apresentarem estrutura de variância-covariância significativa, deve-se utilizar uma estratégia adequada para a análise, como a FA.

Passo 2: Conforme estipulado por diversos autores, destacado no capítulo 2, antes de se aplicar a técnica de FA, existe a necessidade de verificar se os dados são adequados para esse tipo de estratégia. Para isso, testes específicos devem ser utilizados, como o teste de esfericidade de Bartlett e o índice Kaiser-Meyer-Olkin. Caso o conjunto de dados esteja apto para a aplicação dessa estratégia multivariada, deve-se seguir para o próximo passo. Se o conjunto de dados não for adequado, deve-se utilizar uma estratégia exploratória alternativa, que contemple a estrutura de variância-covariância significativa.

Passo 3: Sendo o conjunto de dados adequados para aplicação de FA, tem-se, nesta etapa, a abordagem para encontrar a rotação ótima. Esse passo irá detalhar um método para encontrar, por meio de aplicações de caráter experimental, o melhor valor gama (γ) de rotação ortogonal que irá proporcionar uma estrutura simplificada de cargas fatoriais, promovendo uma interpretação e explicação ótima dos fatores, diante das variáveis analisadas. Como indicado anteriormente, o valor γ para rotação ortogonal das cargas fatoriais, pode variar de 0 a 1, em que as extremidades representam as rotações *quartimax* e *varimax*, respectivamente. Essa

variação é conhecida como família *orthomax*. As etapas, que consistem nestas análises, são destacadas a seguir:

- A partir de um conjunto de dados adequado para aplicação de FA, inicialmente, deve-se estimar a quantidade de fatores a serem utilizados. Como quantidade mínima, tem-se que o percentual de explicação acumulado deve ser de, pelo menos, 80% [5]. Contudo, Visinescu e Evangelopoulos [152] afirmam que utilizar uma quantidade maior de fatores promove uma melhor explicação dos dados e, consequentemente, resultados melhores e mais consistentes. Assim, além do critério de percentual de explicação, apresentado anteriormente, estabelece-se que, quando há rotação, a quantidade *m* de fatores pode ser definida por até 50% da quantidade de *p* variáveis, se o percentual de incremento for de, pelo menos, 10%. Deste modo, evita-se negligenciar fatores que possam apresentar um incremento significativo de explicação, além de evitar uma acumulação de vetores de resposta que possam apresentar baixa porcentagem de contribuição na interpretação das variáveis latentes. Uma definição similar fora utilizada no estudo de Almeida *et al.* [62].
- O próximo passo do procedimento é delinear um arranjo experimental para contemplar a amplitude dos valores de rotação γ. Tratando-se de uma variabilidade de proporções, a classe especial de superfície de resposta, chamada de arranjos de misturas, será utilizada. Entre as estratégias de misturas, o arranjo do tipo *Simplex-Lattice* se faz o mais adequado para esta finalidade. Deste modo, cria-se um DOE para 2 componentes (valor γ + complemento de proporção), com *Lattice* igual a 10 ($ld = 10$) e 2 pontos axiais. Com isso, o arranjo proporciona uma matriz experimental de 13 combinações, capaz de contemplar toda a amplitude de rotação dos eixos de modo significativo.
- Após definir o arranjo experimental e a quantidade de fatores a serem utilizados, aplica-se a estratégia FA extraída por componentes principais, considerando o valor γ para cada uma das linhas experimentais. Para cada parâmetro, deve-se registrar os indicadores que explicam as variáveis, assim, calcula-se a variância total explicada (VTE) de cada um dos fatores, para cada experimento. Quanto mais alto for o valor de VTE, maior será o nível de explicação desse respectivo fator. Entretanto, espera-se que a diferença entre as VTE's deva ser mínima, ou seja, os valores da VTE, entre os fatores, devem ser os mais homogêneos possíveis. Portanto, as respostas de interesse podem ser definidas como: a proporção da variância amostral explicada pelo *j-ésimo* fator (Eq. (3.1)) e a variabilidade existente entre esses valores (σ^2_{VTEj}). Tem-se que $tr(\mathbf{R})$ é o traço da matriz de correlação, conforme indicado no capítulo 2.

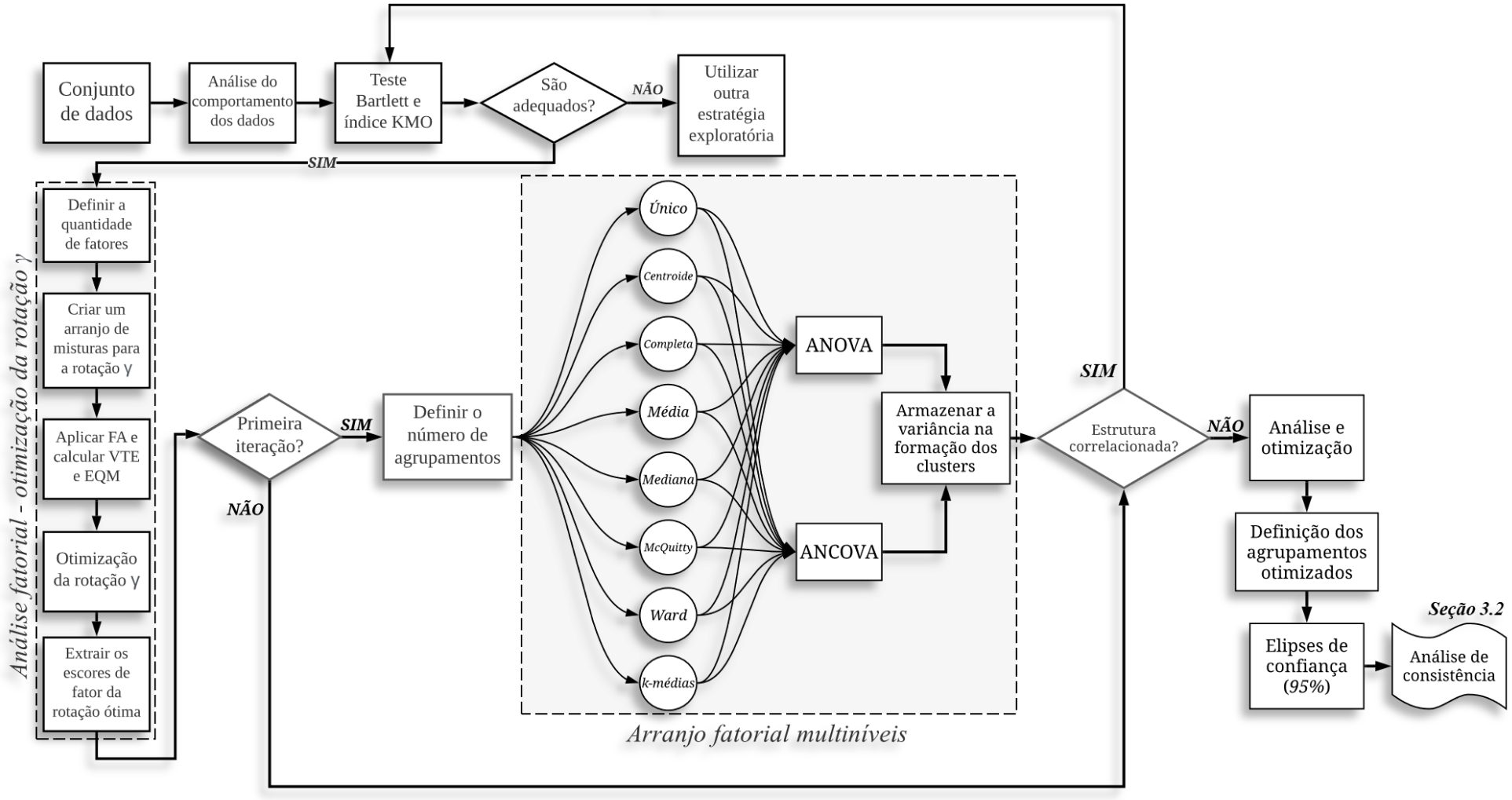


Figura 3.1. Fluxograma do método proposto

$$\frac{\mathbf{L}_{1j}^{\circ 2} + \mathbf{L}_{2j}^{\circ 2} + ... + \mathbf{L}_{pj}^{\circ 2}}{tr(\mathbf{R})} = \frac{\lambda_j}{tr(\mathbf{R})} \tag{3.1}$$

Diante dos valores VTE armazenados, pode-se aglutinar todas essas informações criando uma parametrização única baseado na técnica EQM, calculando a média amostral para cada um dos valores de VTE (sendo este, o alvo definido) e a variância entre os valores. Assim, esse modelo matemático pode ser descrito conforme a Eq. (3.2), que, posteriormente, deve ser minimizado, considerando as restrições do espaço experimental. Tal métrica representa o erro quadrático médio dos valores VTE, considerando a variância do modelo.

$$EQM_{VTE} = \sum_{i=1}^{m}\left[\left(VTE_i - \bar{X}_{VTE}\right)^2 + \sigma_{VTE}^2\right] \tag{3.2}$$

Sabendo que os valores γ, definidos pelo DOE, irão contemplar toda amplitude de rotação γ, é possível realizar a análise dos experimentos, verificando se os resultados apresentam ajuste adequado, além de analisar o comportamento dos efeitos. Após essa avaliação e ajuste dos fatores, quando necessário, é possível calcular e armazenar os coeficientes a serem utilizados na próxima etapa.

- Com os coeficientes das equações de regressão do arranjo experimental, deve-se buscar o valor de rotação γ ótimo. Para tal, utiliza-se da abordagem de otimização para encontrar o ponto ótimo que represente o melhor valor γ, ou seja, a rotação que apresente a melhor estrutura de cargas fatoriais perante as variáveis latentes, considerando o princípio de parcimônia, explicando-as de maneira simples e de fácil interpretação.

  A etapa de otimização se dá pela minimização do valor de EQM_{VTE}, gerado pela Eq. (3.2), respeitando a restrição de espaço experimental definido pelo arranjo de misturas. Assim, tem-se o modelo matemático simplificado, conforme descrito na Eq. (3.3). Como método de busca para o ponto ótimo, será utilizado o algoritmo SQP.

$$\begin{cases} Min \quad \sum_{i=1}^{m} EQM_{VTEi} = \begin{array}{r} \beta_i \times \gamma_1 + \beta_j \times \gamma_2 + \beta_{ij} \times \gamma_1 \times \gamma_2 + \beta_{ij} \times \gamma_1 \times \gamma_2 \times \left(\gamma_1 - \gamma_2\right) \\ + \beta_{ij} \times \gamma_1 \times \gamma_2 \times \left(\left(\gamma_1 - \gamma_2\right)^2\right) \end{array} \\ sujeito\ a: \\ \qquad \gamma_1 + \gamma_2 = 1 \end{cases} \tag{3.3}$$

- Por fim, após a convergência do valor ótimo de γ, realiza-se a aplicação da FA, definindo o valor de rotação γ com o valor encontrado pela otimização da etapa anterior. Deste modo, deve-se extrair e armazenar os escores de fator com rotação γ *orthomax* otimizada e seguir para etapa posterior.

Passo 4: Após definir os valores otimizados para encontrar a melhor explicação das variáveis latentes, deve-se utilizar os escores rotacionados para a análise de cluster. Como inferido no capítulo 2, a definição da quantidade clusters pode variar a cada objetivo. Contudo, alguns autores como Miranda Filho *et al.* [61] e Almeida *et al.* [62], indicam uma sugestão de métrica para a quantidade de cluster, em que o valor ideal é definido pela regra de categorização do grupo, conhecida como Regra de Sturges [115], sendo a quantidade (k_c) definida por $k_c = 1 + 3{,}322log(\zeta)$, onde ζ é a quantidade de objetos no estudo.

Conhecendo a quantidade de clusters a serem avaliados, aplica-se, aos escores com rotação γ otimizada, cada um dos principais métodos de ligação existentes na literatura e fundamentados no Capítulo 2, compondo os métodos hierárquicos (Único, Centroide, Completa, Média, Mediana, McQuitty e Ward) e o método não hierárquico (*k-médias*). Tratando-se de escores de fator, de rotação ortogonal e com extração por componentes principais, sabe-se que os escores são independentes entre si, ou seja, não há correlação entre os vetores de escores. Portanto, descarta-se o uso da distância de Mahalanobis, como métrica de distância, e define-se a estratégia de distância Euclidiana. Conhecendo o número de clusters e a parametrização, armazena-se os valores de agrupamento, ou associações (*memberships*), encontrados para cada método de ligação.

Passo 5: Com as associações definidas, realiza-se a análise dos agrupamentos com as principais variáveis. Assim, considera-se as associações de cada um dos agrupamentos realizando uma análise de duas maneiras distintas: *i*) de maneira usual, através da ANOVA, considerando apenas uma variável por vez e; *ii*) considerando, além da variável principal, uma variável concomitante através da ANCOVA. Esta última, por sua vez, implica em melhorar a precisão da explicação de uma resposta considerando outra variável que também explica, significativamente, a resposta de interesse, pois realiza um ajuste na ANOVA convencional. Deste modo, armazena-se os valores de média e desvio padrão relatados por cada uma das análises (ANOVA e ANCOVA) para cada uma das associações, geradas pelos métodos de cluster do passo anterior.

Passo 6: Considerando as análises realizadas, calcula-se a variância para cada combinação entre método de cluster e tipo de análise. Deste modo, é possível gerar um novo DOE para modelar, analisar e encontrar a melhor configuração das estratégias, baseado em análises estatísticas.

Para isso, considera-se um DOE do tipo fatorial completo generalizado (também conhecido como arranjo fatorial multiníveis), com dois fatores, sendo um deles com oito níveis, contemplando todos os métodos de ligação utilizados (hierárquicos e não hierárquico), enquanto o segundo fator apresenta dois níveis, referentes aos métodos de análise (ANOVA e ANCOVA). Define-se, como respostas de interesse, as variâncias calculadas na formação dos clusters. A partir disso, é possível realizar a análise dos efeitos principais e dos ajustes de regressão do modelo, removendo os efeitos não significativos quando necessários.

Passo 7: Antes de realizar a otimização para definir os parâmetros ótimos de agrupamento e análise, se faz necessário verificar a estrutura dos dados que contemplam as variâncias dos clusters. Caso as respostas de interesse deste arranjo fatorial (sendo as variâncias na formação dos clusters detalhados no passo anterior) apresentem uma estrutura de variância-covariância significativa, deve-se tratar as mesmas com uma técnica multivariada adequada [6–8,142,188–190]. Assim, volta-se ao **Passo 2** para verificar se os dados são adequados para aplicação da estratégia multivariada FA. Em caso positivo, segue-se para o **Passo 3** a fim de encontrar o valor γ otimizado para esse novo conjunto de respostas, de modo que as cargas fatoriais apresentem a estrutura mais simples possível. A partir disso, extrai-se os escores rotacionados e segue-se para a segunda iteração (Passo 3.1), que encaminha novamente ao DOE multiníveis para analisar os escores rotacionados, que representam a variância na formação dos clusters.

Passo 8: A partir das respostas de variância, ou mesmo dos escores que representem essas respostas, realiza-se uma nova otimização para encontrar a parametrização ótima, ou seja, o método de análise e de ligação que apresente menor variabilidade. Assim, tem-se o sentido de otimização definido para encontrar o ponto ótimo deste experimento fatorial, conforme indicado na Eq. (3.4).

$$\begin{cases} Min\ y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \\ Sujeito\ a: \mathbf{X}^T\mathbf{X} \end{cases} \tag{3.4}$$

Passo 9: Conhecendo a parametrização ótima, calcula-se o volume das elipses para cada cluster, considerando a resposta principal do conjunto de dados e sua respectiva variável concomitante. Essas elipses representam a região de confiança dos dados, conforme apresentado no capítulo 2. Para calcular essas regiões (Eq. 3.5), se faz necessário os vetores de média e de autovalores,

além das matrizes de variância-covariância e de autovetores, calculadas a partir das principais variáveis, com base na Eq. (2.87).

$$n\left(\boldsymbol{\mu}-\bar{\mathbf{y}}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\bar{\mathbf{y}}\right)\leq\frac{\nu p}{\nu+1-p}F_{\alpha,p,\nu+1-p} \tag{3.5}$$

Com isso, calcula-se as elipses de confiança para um nível α de 5%, ou seja, para uma região de confiança de 95%, com finalidade de estimar e definir a discriminação dos agrupamentos. Elipses sobrepostas inferem que não é possível rejeitar a hipótese nula de que as médias são iguais, enquanto a não sobreposição rejeita tal hipótese. A partir dos resultados, é possível categorizar os níveis dos agrupamentos, diante das características individuais da resposta de interesse.

3.2 Análise confirmatória da estabilidade dos métodos de ligação

A partir dos resultados encontrados pelo método proposto na seção anterior, pode-se definir a combinação de estratégias que resultam em agrupamentos com menor variabilidade, representados por regiões de confianças estreitas, precisas e não sobrepostas. Contudo, ao se referir a métodos de ligação, Johnson e Wichern [5] afirmam que existem fontes de erros e variações na formação dos clusters, que podem promover inversões nos agrupamentos, uma vez que essa estratégia é sensível a *outliers*. Baseado nisso, os mesmos autores afirmam a necessidade de gerar pequenas perturbações no conjunto de dados, com finalidade de verificar o comportamento dos métodos. Deste modo, como forma de inferir e encontrar um método de ligação robusto em diferentes cenários do mesmo objeto de estudo, a segunda parte do método propõe uma abordagem para confirmar a parametrização ótima dos métodos de ligação, utilizados na seção anterior. Para isso, será aplicada a técnica de análise de concordância por atributos, conforme descrito na Figura 3.2.

A partir das variáveis originais, deve-se gerar diferentes cenários que apresentem pequenas perturbações distintas nos conjuntos de dados (na faixa de 1%). Em seguida, pelo menos quatro réplicas devem ser geradas de maneira aleatória, contemplando o grau de perturbação dos dados. É importante ressaltar que as replicações devem manter uma estrutura de variância-covariância similar ao conjunto de dados original.

Em seguida, deve-se verificar se os dados são adequados para aplicação de FA, para cada uma das réplicas (o que é esperado, caso o conjunto de dados inicial também esteja apto a essa

aplicação). Sendo adequadas, aplica-se toda etapa definida no **Passo 3** da sessão 3.1, encontrando o valor de rotação γ otimizado para o conjunto de dados, extraindo os escores de fator para cada uma das replicações. Com base nesses valores, realiza-se as associações para cada método de ligação, armazenando os valores de agrupamento e realizando a análise pela estratégia definida no resultado da aplicação da seção 3.1 (ANOVA ou ANCOVA). Deste modo, tem-se todos os critérios necessários para conduzir uma análise de concordância por atributos com 4 cenários distintos (réplicas com perturbações) e 8 avaliadores, representadas pelos métodos de ligação.

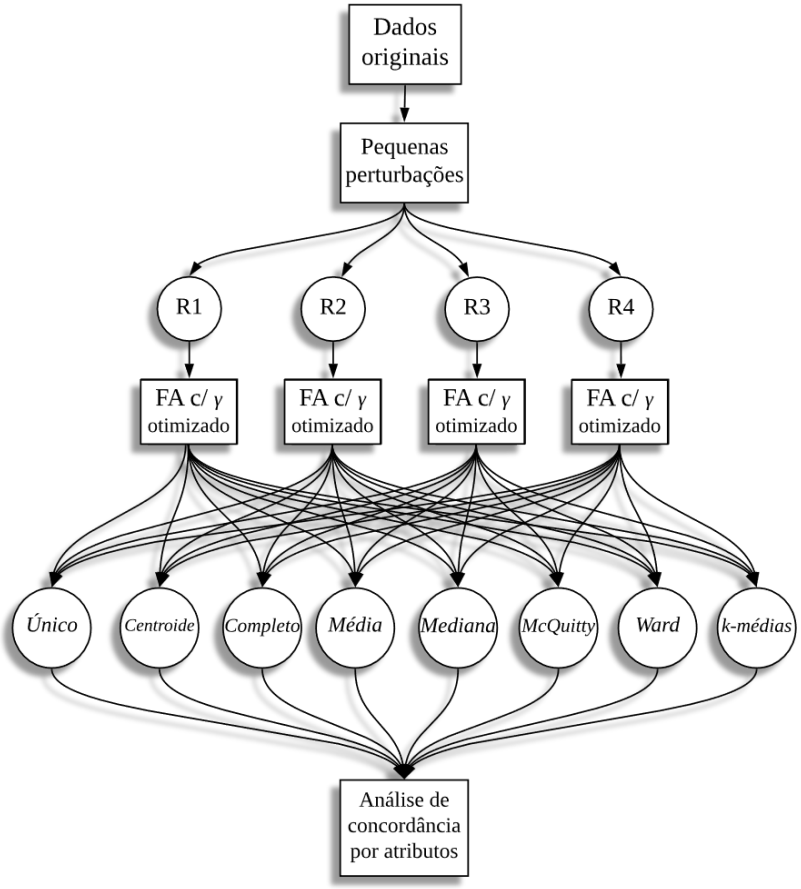


Figura 3.2. Fluxograma para verificar a robustez dos métodos de ligação

Assim, pode-se conduzir os estudos, calculando as estatísticas de Kappa e os coeficientes de concordância de Kendall, definindo os métodos com melhor comportamento e robustez perante os critérios de classificação mundiais, estabelecidos pela AIAG [119].

3.3 Método de pesquisa

Com base no que foi apresentado, tem-se a caracterização do método de pesquisa da tese baseado nas diretrizes de classificação de pesquisas científicas, destacando a área de conhecimento da Engenharia de Produção. Assim, o método de pesquisa pode ser avaliado quanto a natureza, o tipo de abordagem e os objetivos [191].

Deste modo, tem-se que o presente estudo se classifica como de natureza aplicada, visto que apresenta objetivos práticos, podendo ser apresentados em problemas reais [192]. A sua abordagem é caracterizada como quantitativa, uma vez que analisa relações entre as variáveis de desempenho e as controláveis [193], fazendo uso de um procedimento híbrido, o qual contempla os aspectos de modelagem e simulação, além de experimentos. Essa abordagem proporciona um estudo com característica axiomática normativa, visto que busca encontrar soluções em um delineamento definido.

3.4 Considerações finais

Diante do que foi apresentado nesse capítulo, tem-se que o método proposto apresenta uma nova estratégia que busca aprimorar o poder discriminatório de funções elipsoidais auxiliadas por escores de fatores sob rotação ortogonal. Essa abordagem se difere por apresentar artifícios que buscam sanar algumas dificuldades entre as técnicas apresentadas no capítulo 2, auxiliando na definição de parâmetros ideais para a rotação de cargas fatoriais. Em seguida, uma abordagem para realizar agrupamentos ótimos, com base na variabilidade na formação dos clusters, gerando elipses de confianças precisas, auxiliando na discriminação das informações. Os passos apresentados na seção 3.1 remetem a um formato generalizado, ou seja, que pode ser aplicado a diversos problemas de agrupamento e de discriminação de dados extensos. É importante ressaltar que o método de otimização da rotação γ pode ser estendido, separadamente, para outras aplicações que necessitam de técnicas multivariadas de exploração, não se restringindo apenas para discriminação de informações e formações de agrupamentos de dados.

A abordagem da seção 3.2 proporciona avaliar a robustez do método de ligação, verificando qual método apresenta melhor grau de concordância sob pequenas perturbações, contemplando a discussão sobre a estabilidade dos agrupamentos. O capítulo seguinte irá tratar do conjunto de dados a ser investigado, perante a proposta apresentada neste capítulo.

4. ÍNDICES DE QUALIDADE DE ENERGIA DE SUBESTAÇÕES NO SUDESTE DO BRASIL

A partir dos conceitos e etapas propostas para aprimorar o poder discriminatório em conjuntos de dados com múltiplas respostas correlacionadas, detalhados na seção anterior, este capítulo busca descrever o objeto de estudo que será utilizado para demonstrar o comportamento do método proposto. Para isso, será investigado um conjunto de dados de subestações de uma distribuidora de energia elétrica localizada no sudeste do Brasil. Tais dados são parte de um projeto de Pesquisa e Desenvolvimento (P&D) realizado na Universidade Federal de Itajubá (UNIFEI), e explorado, incialmente, nos trabalhos de [61,70], os quais visam analisar e classificar as subestações de distribuição de energia elétrica diante do seu grau de qualidade, avaliando o fenômeno de variações de tensão de curta duração em afundamentos momentâneos de tensão, perante as necessidade regulatórias da ANEEL [61]. Segundo Miranda Filho [70], VTCD se caracteriza como um distúrbio de qualidade que afeta sistemas industriais, podendo gerar curtos-circuitos em grandes áreas, além de poder danificar equipamentos e energizar transformadores. Tais prejuízos podem interferir diretamente no nível de produtividade das indústrias, paralisando a linha de produção, danificando equipamentos e outros custos inerentes desta causa. Devido a essas ocorrências, tem-se a necessidade de um sistema de distribuição mais confiável, com um nível de qualidade conhecido.

Baseado nisso, um exemplo real de modelagem de rede e simulação de falhas em níveis de transmissão e distribuição localizados no estado do Espírito Santo (ES), Brasil, é investigado com o objetivo de validar a proposta do presente trabalho. Considerou-se um sistema de distribuição de energia elétrica composto por 17 subestações cuja área total é de aproximadamente 41.241 km^2, cerca de 90% do Estado do ES. A Figura 4.1 ilustra a distribuição dos conjuntos de subestações para a região em análise. Esse sistema contempla também o uso de 96 alimentadores com 25.769 km de linhas de distribuição. O conjunto de dados, que será detalhado posteriormente (originalmente disponível em [61]), foram obtidos em um projeto de 30 meses gerenciado pela *EDP ES Concessionária de Distribuição*, empresa de energia elétrica distribuidora.

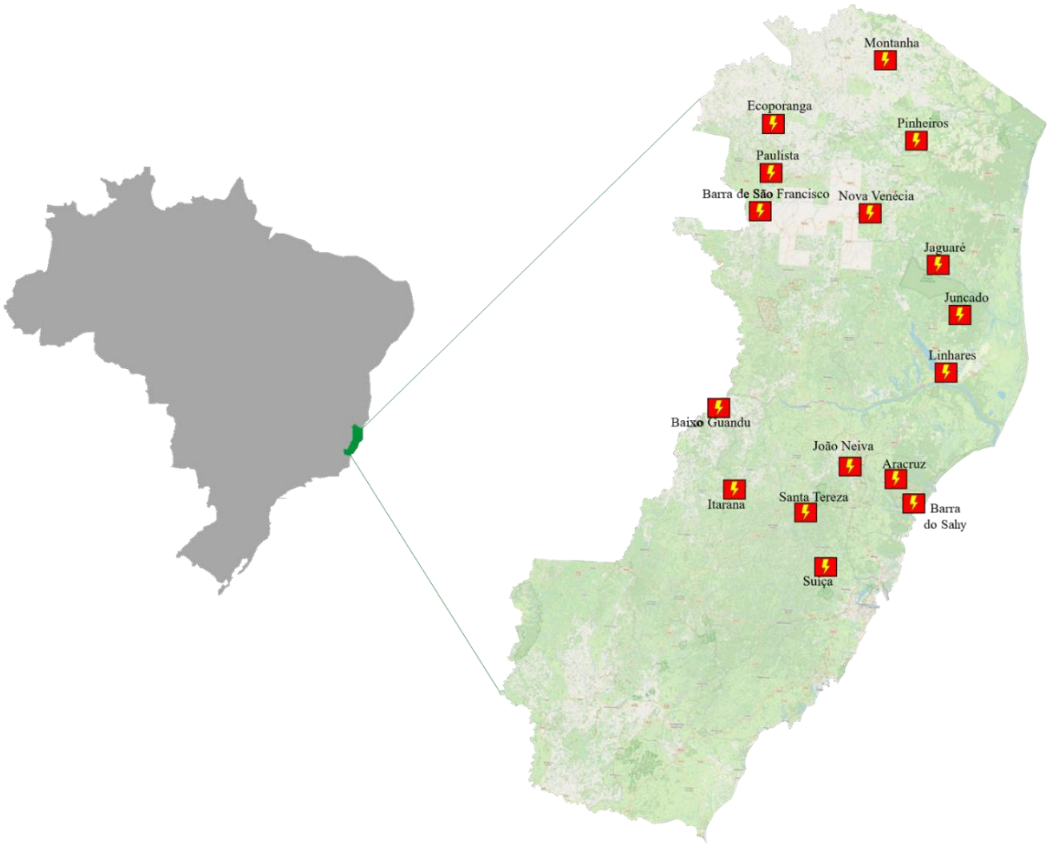


Figura 4.1. Localização das subestações investigadas no Estado do Espírito Santo
Gráfico gerado parcialmente pelo *software QGIS-3*®

Nesse contexto, os afundamentos de tensão ocorrem por ocorrências de raios e curtos-circuitos, uma vez que a maior parte das linhas aéreas e alimentadores não são cobertas. As informações dos alimentadores de tensão nominal, comprimento das linhas de distribuição e as estatísticas de falha (falhas por 100 km/ano), utilizadas nas simulações de curto-circuito causando os afundamentos investigados, podem ser visualizados na Tabela 4.1. Os monitores de qualidade de energia registraram os eventos que foram coletados nas barras dos secundários dos transformadores. Obteve-se uma linha de distribuição equivalente a 13,8kV a partir das taxas totais de longo prazo (maior ou igual a 3 minutos) e de curta duração das estatísticas, na qual foi utilizada a Taxa de Falha de Média Tensão (*MVFR – Medium Voltage Failure Rate*). Pode-se verificar também a existência do tipo trifásico com menor incidência, enquanto uma única fase apresenta maior incidência entre as ocorrências. Por fim, como destacado em Miranda Filho *et al*. [61], considerou-se uma distribuição normal para estimar os sistemas de transmissão e sub transmissão, em que a média (μ) fosse igual a 5Ω e o desvio padrão (σ) fosse igual a 1Ω. Também foi utilizado uma distribuição uniforme variando de zero a um para a rede de distribuição, em que um valor máximo de 30Ω foi atribuído para falhas 1LG, 30Ω para falhas

2LG, 10Ω para falhas 3LG e 20 Ω para falhas de 2L. Maiores detalhes sobre as barras da subestação em análise estão disponíveis em [61,62,70].

Tabela 4.1. Comprimento das linhas de distribuição, nível de tensão dos alimentadores e estatísticas de falhas utilizadas nas simulações de curtos-circuitos

Nível de tensão	Comprimento	Taxa de falha	1LG	2LG	LL	LLL
[*kV*]	[*km*]		[%]	[%]	[%]	[%]
138	2125,5	2,33	75	13	10	2
69	1033	6,34	58	25	11	6
34.5	619	43,13	70	15	10	5
13.8	22.750	MVFR	78	10	9	3

Entre as subestações analisadas, destacadas na Figura 4.1, considerou-se 32 características de projeto e qualidades de energia. Essas características estão detalhadas na Tabela 4.2. Para a finalidade de classificação, Miranda Filho *et al.* [61] e Almeida *et al.* [62] destacam que a principal característica de qualidade é a que se refere ao número total de eventos de afundamentos de tensão por ano (*TNE – Total Number of Sag events*), que é obtido por meio de simulação, enquanto o número de eventos de monitorados (*MNE – Monitored Number of Events*) é obtido por monitoramento durante um ano. As Tabelas 4.3 a 4.6 apresentam os valores das 32 variáveis utilizadas neste estudo, coletadas originalmente em [61]. As medições de qualidade de energia foram coletadas ao longo de um ano, com finalidade de cobrir diferentes sazonalidades que influenciam no desempenho da rede de distribuição de energia elétrica (como chuva, ventos, entre outros fenômenos). Além disso, 30 monitores de qualidade de energia da *Schweitzer Engineering Laboratories*, modelo SEL 734, foram utilizados para adquirir esses dados. O comportamento do conjunto de dados perante as subestações está ilustrado na Figura 4.2.

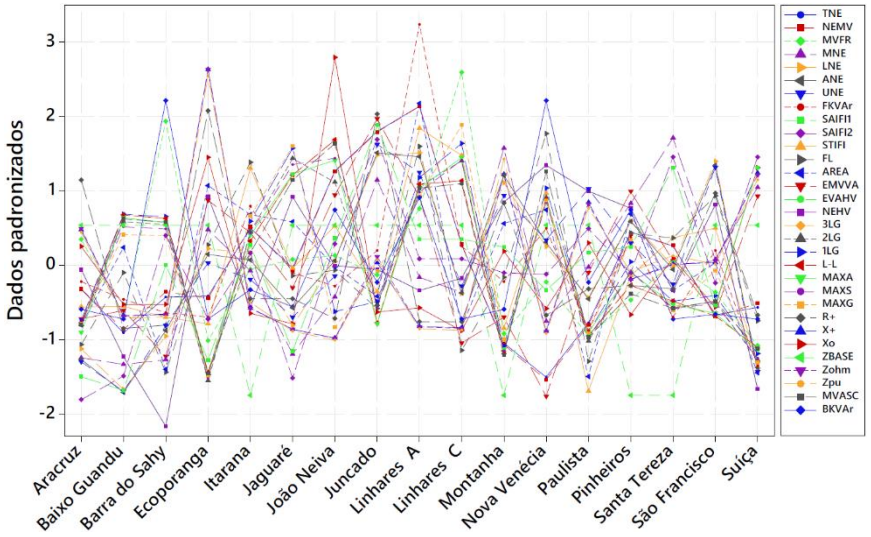


Figura 4.2. Relação entre os índices de qualidade de energia e as subestações

Tabela 4.2. Características de qualidade analisadas na distribuição de energia elétrica

Sigla	Variável	Significado
HVFR	*High Voltage Failure Rate*	Taxa de falha de alta tensão (falhas por linha de 100 km por ano)
TNE	*Total number of Sag events*	Número total de eventos de afundamento por ano (falhas voltagem média e na alta voltagem)
NEMV	*Number of Events in Medium Voltage*	Quantidade de Eventos em Média Tensão (falhas em voltagem média)
MVFR	*Medium Voltage Failure Rate*	Taxa de falha de média tensão (falhas por 100 km alimentador por ano)
MNE	*Monitored Number of Events*	Número de eventos monitorados (um ano de barramento de 13,8 kV)
LNE	*Lower Number of Events*	Número inferior de eventos (100 cenários simulados)
ANE	*Average Number of Events*	Número médio de eventos (100 cenários simulados)
UNE	*Upper Number of Events*	Número superior de eventos (100 cenários simulados)
FKVAr	*Shunt Capacitor KVAr*	Capacitor shunt KVAr instalado nos alimentadores
SAIFI1	*System Average Interruption Frequency Index*	Índice de frequência de interrupção média do sistema (sem dia crítico)
SAIFI2	*System Average Interruption Frequency Index*	Índice de frequência de interrupção média do sistema (com dia crítico)
STIFI	*System Total Interruption Frequency Index*	Índice de frequência de interrupção total do sistema (número de eventos)
FL	*Feeders Length*	Comprimento dos alimentadores (km)
AREA	*Cluster Area*	Área de agrupamento (km^2), segundo ANEEL [194]
EMVVA	*Equivalent Medium Voltage Vulnerability Area*	Área de vulnerabilidade de média tensão equivalente (km), em que os curtos-circuitos causam quedas no barramento da subestação
BMVAr	*Shunt Capacitor MVAr*	Capacitor shunt MVAr instalado no barramento
NEHV	*Number of Events High Voltage*	Número de eventos de alta tensão (falhas em linhas de alta tensão)
3LG	*Three phase to ground short-circuit current*	Corrente de curto-circuito trifásico para terra (no barramento)
2LG	*Double phase to ground short-circuit current*	Corrente de curto-circuito de fase dupla para terra (no barramento)
1LG	*Single phase to ground short-circuit current*	Corrente de curto-circuito monofásico para terra (no barramento)
L-L	*Phase to phase short-circuit current*	Corrente de curto-circuito fase a fase
MAXA	*Maximum asymmetric short circuit Current*	Corrente máxima de curto-circuito assimétrico (no barramento)
MAXS	*Maximum symmetric short circuit Current*	Corrente máxima simétrica de curto-circuito (no barramento)
MAXG	*Maximum symmetric short circuit Current*	Corrente máxima de curto-circuito simétrico para a terra (no barramento)
R+	*Positive Sequence Resistance*	Resistência de sequência positiva (no barramento)
X+	*Positive Sequence Reactance*	Reatância de sequência positiva (no barramento)
Xo	*Zero Sequence Reactance*	Reatância de sequência zero (no barramento)
ZBASE	*Base Impedance ohm*	Ohm de impedância de base (no barramento)
Zohm	*Equivalent Impedance ohm*	Ohm de impedância equivalente (no barramento)
Zpu	*Per Unit Impedance*	Impedância por unidade (no barramento)
EVAHV	*Equivalent High Voltage Vulnerability Area*	Área de Vulnerabilidade de Alta Tensão Equivalente (km), em que os curtos-circuitos causam quedas no barramento da subestação
MVASC	*Short Circuit Power MVA*	MVA de energia de curto-circuito (1000/Zpu no barramento)

Tabela 4.3. Índices de qualidade de energia da subestação (Parte I)

Subestação	HVFR*	TNE	NEMV	MVFR	MNE	LNE	ANE	UNE
Aracruz	2.33	210,8953	179,60	235	79	47	61,22	82
Baixo Guandu	2.33	131,93174	105,70	126	72	28	43,69	64
Barra do Sahy	2.33	196,8775	174,70	404	77	53	76,7	104
Ecoporanga	2.33	198,81387	163,27	90	216	94	117,38	142
Itarana	2.33	322,34765	293,25	192	172	92	114,3	132
Jaguaré	2.33	249,6614	214,14	206	83	58	88,81	109
João Neiva	2.33	426,1293	394,61	212	144	88	114,02	134
Juncado	2.33	498,07767	466,78	184	269	138	171,59	214
Linhares A	2.33	543,97146	513,89	279	165	139	169,45	197
Linhares C	2.33	292,0779	261,28	474	149	73	96,79	128
Montanha	2.33	113,1813	77,98	100	303	125	159,5	183
Nova Venécia	2.33	51,20909	13,84	174	118	95	123,79	156
Paulista	2.33	149,87619	114,02	97	176	71	93,52	187
Pinheiros	2.33	314,99604	280,13	148	244	107	134,94	154
Santa Tereza	2.33	289,72731	259,65	203	314	86	109,26	129
São Francisco	2.33	164,39265	129,33	159	184	135	164,92	200
Suíça	2.33	177,98019	153,63	83	261	38	56,83	76

*Variável constante (não simulada)

Adaptado de [61]

Tabela 4.4. Índices de qualidade de energia da subestação (Parte II)

Subestação	FKVAr	SAIFI1	SAIFI2	STIFI	FL	AREA	EMVVA	EVAHV
Aracruz	6900	5,567	6,531	1130	365,22	555,414	76,42	1343,24
Baixo Guandu	5700	5,070	7,392	1136	885,65	1080,046	83,89	1125,83
Barra do Sahy	4800	9,592	12,531	1014	164,74	255,178	43,24	951,75
Ecoporanga	4500	6,163	9,481	941	1086,75	1499,906	181,41	1525,50
Itarana	12000	10,329	11,885	2672	1685,29	1303,753	152,74	1248,75
Jaguaré	7500	6,490	7,317	1561	863,08	1254,651	103,95	1524,50
João Neiva	6600	10,575	12,225	2029	903,17	950,368	186,14	1352,75
Juncado	9000	14,668	16,040	1298	625,00	749,424	253,68	1343,25
Linhares A	24300	10,538	11,677	3117	1800,09	2057,083	184,19	1290,89
Linhares C	9300	10,538	11,677	2814	323,23	537,861	55,12	1321,75
Montanha	6900	10,253	11,166	891	853,20	1242,832	77,98	1511,00
Nova Venécia	10500	8,695	11,118	2392	1892,82	1335,288	7,95	1603,90
Paulista	3600	10,062	12,781	193	244,11	205,801	117,54	1539,00
Pinheiros	9900	10,253	11,166	1477	1173,21	1307,187	189,28	1496,50
Santa Tereza	4500	13,109	15,398	1899	1136,70	789,833	127,91	1290,75
São Francisco	9000	8,135	10,796	2006	1441,63	977,427	81,34	1504,70
Suíça	1200	13,109	15,398	521	536,59	231,392	185,09	1045,25

Adaptado de [61]

Diante da importância dos afundamentos de tensão para a qualidade de energia em indústrias e distribuidoras, se faz necessário estimar, com um alto nível de confiança, os padrões de queda de tensão, além de verificar os agrupamentos de subestações de distribuição com diferentes níveis de qualidade. Assim, a aplicação do método proposto, para esse conjunto de dados, será detalhada no capítulo seguinte.

Tabela 4.5. Índices de qualidade de energia da subestação (Parte III)

Subestação	NEHV	3LG	2LG	1LG	L-L	MAXA	MAXS	MAXG
Aracruz	31	3155	3513	3478	2732	5295	3513	3862
Baixo Guandu	26	6485	6578	6632	5616	11676	6632	6786
Barra do Sahy	22	6377	6504	6564	5522	11430	6564	6761
Ecoporanga	36	1631	1839	1854	1412	2941	1854	2145
Itarana	29	5674	6281	6425	4914	10718	6425	7401
Jaguaré	36	7714	8469	8595	6681	14120	8595	9694
João Neiva	32	8798	7761	3737	7620	14895	8798	3737
Juncado	31	3151	4111	4034	2729	5769	4111	5584
Linhares A	30	7433	7643	7724	6437	13326	7724	8038
Linhares C	31	7536	8397	8736	6526	15037	8736	10390
Montanha	35	2293	2669	2718	1986	4364	2718	3334
Nova Venécia	37	6964	7299	7408	6031	12647	7408	7910
Paulista	36	2755	3020	3113	2386	5300	3113	3579
Pinheiros	35	4311	5076	5216	3734	8657	5216	6597
Santa Tereza	30	3827	4022	4046	3314	6850	4046	4290
São Francisco	35	3368	4204	4211	2917	6363	4211	5604
Suíça	24	2400	2456	2484	2078	4324	2484	2574

Adaptado de [61]

Tabela 4.6. Índices de qualidade de energia da subestação (Parte IV)

Subestação	R+	X+	Xo	ZBASE	Zohm	Zpu	MVASC	BMVAr
Aracruz	0,460	2,480	1,840	1,904	2,522	1,324	75,503	2,4
Baixo Guandu	0,010	1,230	1,000	1,904	1,230	0,646	154,824	0,0
Barra do Sahy	0,020	1,250	1,000	1,904	1,250	0,656	152,333	53,4
Ecoporanga	0,670	4,820	3,130	1,904	4,866	2,555	39,134	0,0
Itarana	0,100	1,340	0,870	1,742	1,344	0,771	129,669	7,2
Jaguaré	0,100	1,030	0,720	1,904	1,035	0,543	184,028	3,0
João Neiva	0,040	0,900	4,580	1,904	0,901	0,473	211,391	26,7
Juncado	0,660	2,000	0,890	1,904	2,106	1,106	90,424	9,0
Linhares A	0,030	1,070	0,950	1,904	1,070	0,562	177,911	29,7
Linhares C	0,030	1,050	0,620	1,904	1,050	0,552	181,297	0,0
Montanha	0,390	3,290	1,770	1,742	3,313	1,901	52,592	2,4
Nova Venécia	0,050	1,000	0,940	1,904	1,001	0,526	190,202	53,4
Paulista	0,130	2,880	1,890	1,904	2,883	1,514	66,058	9,0
Pinheiros	0,140	1,760	0,850	1,742	1,766	1,013	98,688	26,7
Santa Tereza	0,130	1,980	1,670	1,742	1,984	1,139	87,811	0,0
São Francisco	0,420	2,000	0,960	1,904	2,044	1,073	93,187	1,2
Suíça	0,050	3,310	2,980	1,904	3,310	1,738	57,528	0,0

Adaptado de [61]

5. APLICAÇÃO DO MÉTODO PROPOSTO EM ÍNDICES DE QUALIDADE DE ENERGIA

Este capítulo busca desenvolver e apresentar a aplicação do método proposto no capítulo 3, baseando-se nas estratégias detalhadas no capítulo 2, a partir das informações de subestações de distribuição de energia elétrica. Neste sentido, o presente capítulo irá realizar uma investigação sobre o comportamento dos dados, com finalidade de justificar o uso do método proposto e, em seguida, os demais passos que detalham a aplicabilidade deste método serão apresentados, subdividindo-os para um melhor entendimento. Após isso, uma análise confirmatória será realizada utilizando dados com pequenas perturbações (como sugerido por Johnson e Wichern [5]) e estratégias de análise do sistema de medição. O desenvolvimento e detalhes inerentes a essa aplicação, bem como as discussões, também serão apresentadas neste capítulo. Com finalidade de contemplar a aplicação do método proposto, os *softwares* *Minitab*®, *Matlab*®, *R studio*® e o *Visual Basic for Applications*® (VBA) foram utilizados.

5.1 Análise e adequação do conjunto de dados

Perante as informações das subestações apresentadas anteriormente, inicialmente foi preciso verificar o comportamento dos dados, buscando compreender suas características e, consequentemente, verificar se o conjunto estava apto para a aplicação do método proposto (que é baseado na análise fatorial). Verificando o conjunto, observou-se que a variável *Taxa de Falha da Alta Tensão* (HVFR) era uma constante, portanto foi removida do conjunto. Com finalidade de verificar a natureza multivariada, aplicou-se a análise de correlação entre as características disponíveis. Devido a quantidade de variáveis e, consequentemente, de informações, os resultados das análises (*Pearson* e *p-value*) estão descritas nas Tabelas C.1 a C.4 (Apêndice C). Analisando os resultados, foi possível verificar a existência de variáveis com correlações significativas e mesmo com dependência linear (no caso das variáveis TNE e NEMV). Assim, pôde-se inferir a existência de uma estrutura multivariada entre os dados, o que já era esperado, tratando-se de um conjunto real com múltiplas respostas.

Ao analisar o conjunto, notou-se que a estrutura dos dados brutos não permitia a aplicação da FA, visto que o número de colunas era maior que o número de linhas (sendo esta, uma delimitação da técnica). Assim, se fez necessário encontrar as variáveis principais a serem consideradas, visto que esse tipo de conjunto pode apresentar informações redundantes [70]. Frente a esta situação, os autores Almeida *et al.* [62], realizaram a filtragem desses dados

considerando os aspectos técnicos auxiliados pela análise através da estratégia de regressão dos mínimos quadrados parciais. Considerando essa filtragem, definiu-se as variáveis a serem investigadas, diante de suas caracterizações práticas, de modo a contornar a delimitação da técnica FA diante da estrutura do conjunto de dados original. As variáveis consideradas no estudo foram: TNE; NEMV; MVFR; UNE; FKVAr; SAIFI1; SAIFI2; AREA; EMVVA; EVAHV; NEHV; 3LG; 1LG; L-L e; R+.

De modo prático, teve-se que o número de eventos e a área de vulnerabilidade na média tensão (NEMV e EMVVA, respectivamente) foram as variáveis mais significativas para o TNE. Com essa filtragem prévia, foi possível verificar, novamente, a estrutura de variância-covariância dos dados, conforme ilustrado na Figura 5.1. Assim, mesmo após a redução de variáveis redundantes, o conjunto manteve uma estrutura multivariada significativa. O comportamento da nova relação dos indicadores de qualidade de energia está ilustrado na Figura 5.2.

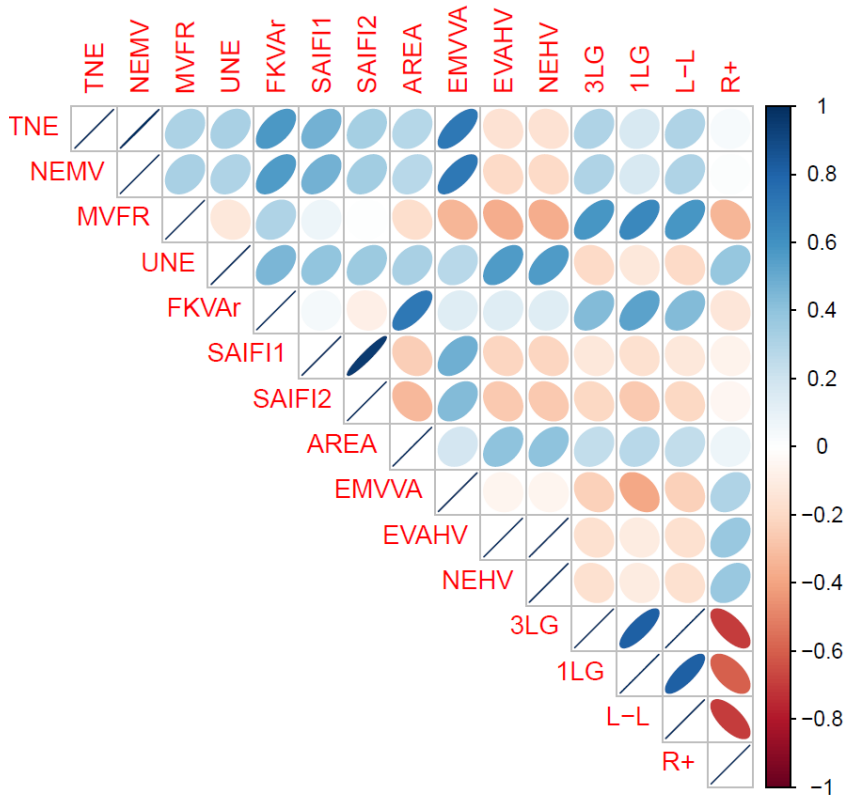


Figura 5.1. Análise gráfica da correlação de *Pearson*

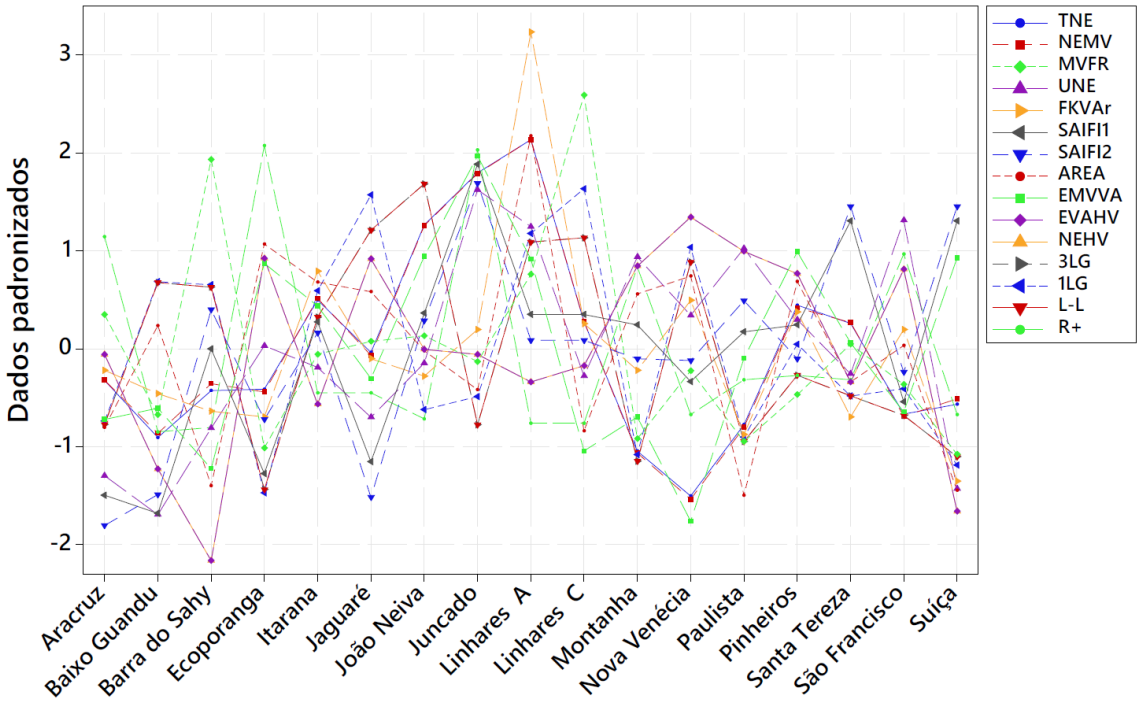


Figura 5.2. Comportamento das variáveis utilizadas na aplicação

Além dos dados apresentarem uma correlação adequada, foi necessário confirmar se o conjunto estava apto a ser receber a aplicação da FA. Assim, realizou-se os testes de adequação, detalhados no capítulo 2. Antes de iniciar esses testes, a normalidade multivariada do conjunto de dados original precisou ser investigada. O teste de normalidade multivariada de Mardia indicou que os dados não seguem uma distribuição de probabilidade multivariada normal, pois a medida de curtose não foi significativa (*p-value* = 0,000), apesar da medida de assimetria (*skewness*) ter sido significativa (*p-value* = 0.9999). Assim, descartou a aplicação do teste de esfericidade de Bartlett, pois este teste parte da suposição de que o conjunto de dados segue uma distribuição multivariada normal. Consequentemente, o resultado do teste de Mardia também sinalizou que o método de extração pela Máxima Verossimilhança não deveria ser utilizado, uma vez que o uso deste método supõe à normalidade multivariada dos dados.

Diante do que foi apresentado, a métrica do índice KMO precisou ser considerada, a fim de testar a adequação das variáveis à aplicação de FA. A aplicação deste indicador (KMO) proporcionou um valor generalizado de 0,5 para o conjunto, que indica um grau apropriado de adequação das variáveis para aplicação da FA, baseado nos critérios de análise estabelecidos na literatura. Destaca-se, também, que os valores individuais de KMO para cada variável foi de 0,5.

Conhecendo o comportamento das variáveis mais significativas e que a estrutura dos dados é adequada para aplicação da FA, seguiu-se para a etapa seguinte, que se refere ao método para encontrar o valor γ de rotação *orthomax* otimizado, baseado no critério de simplicidade e na melhor explicação de variáveis latentes.

5.2 Otimização da rotação γ *orthomax* e extração dos escores

Sabendo que o conjunto apresentou estrutura adequada para aplicação da FA, foi possível seguir com estratégia proposta, a fim de criar vetores de escores que representassem, de modo apropriado, as variáveis originais. Contudo, diante das possíveis diversificações de parâmetros da FA, se fez necessário o uso de um método para encontrar a melhor calibração dessa técnica. Inicialmente, foi definida a quantidade de fatores a serem utilizados. Como o estudo fez uso da técnica de componentes principais para extração do modelo fatorial (visto que o conjunto não apresentou comportamento significativo de uma distribuição multivariada normal), foi necessário que, para um conjunto com correlação significativa, o modelo apresentasse um número de fatores menor que o número de variáveis observáveis, sendo $m < p$. Para isso, foi considerado, pelo menos 80% de explicação e a métrica de complemento estabelecida no método proposto, no capítulo 3.

5.2.1 Quantidade de fatores

Ao analisar o conjunto de dados, foi possível verificar que a quantidade de quatro fatores foi capaz de explicar, pelo menos, 80% dos dados, sendo este um requisito mínimo, baseado no critério de Kaiser. Contudo, Visinescu e Evangelopoulos [152] afirmam que utilizar poucos fatores pode comprometer a interpretação das variáveis latentes, sugerindo o uso mais fatores para melhorar a explicação dos dados. Além disso, uma quantidade maior de fatores pode favorecer a alocação das cargas fatoriais. Deste modo, analisando a contribuição dos demais fatores, foi possível verificar que os mesmos adicionam uma quantidade expressiva de informações, aprimorando a explicação e representação por esses fatores. A partir desse critério, teve-se que o uso de sete fatores apresentou uma contribuição considerável, dentro dos critérios estabelecidos no método proposto, além de respeitar o critério de Kaiser para explicação (de, pelo menos, 80%). Assim, o uso de sete fatores apresentou um percentual de 97,1% de explicação dos dados, além de proporcionarem uma redução de dimensionalidade de 53,34% do conjunto, já reduzido na análise preliminar. A Figura 5.3 apresenta a evolução do percentual de explicação de cada um dos fatores, e de seus respectivos autovalores, através de uma carta

de Pareto, sendo, a quantidade de sete fatores, ideal para representar as variáveis originais com validação estatística adequada.

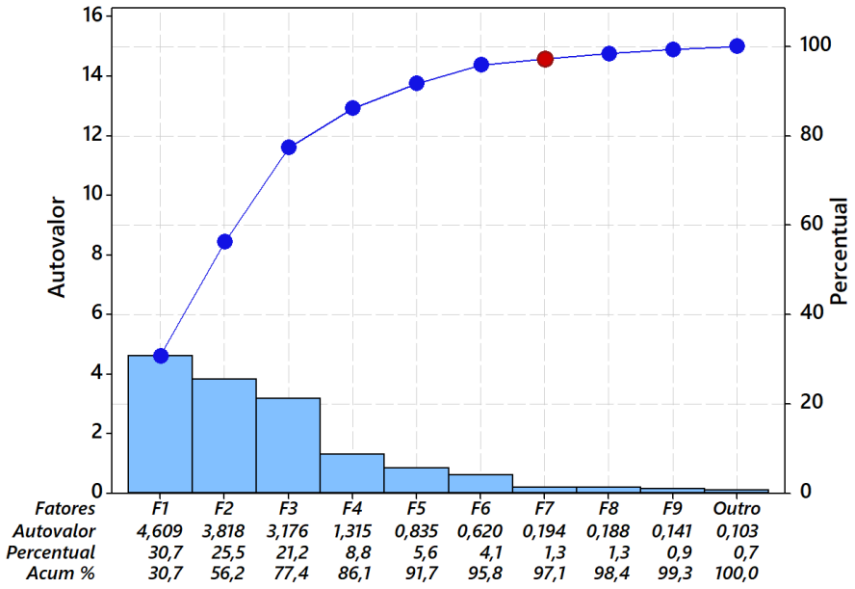
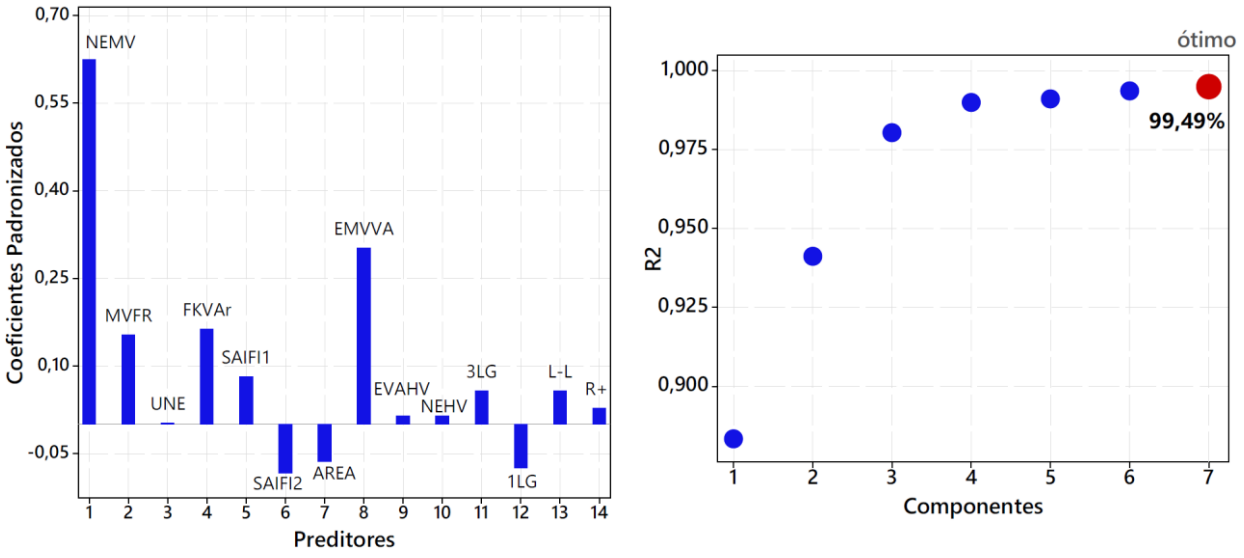


Figura 5.3. Carta de Pareto para os fatores dos índices

De maneira complementar, ao analisar os dados utilizando os mínimos quadrados parciais, foi possível verificar que o conjunto apresentou um modelo com valor de R^2 igual a 99,49%, quando representado por 7 componentes, conforme ilustra o gráfico de seleção do PLS (Figura 5.4). É importante ressaltar que a variável TNE foi considerada como resposta principal, nessa análise.



(a) (b)

Figura 5.4. Gráfico por PLS para (a) coeficientes e (b) seleção do modelo (resposta: TNE)

5.2.2 Arranjo experimental – *simplex-lattice*

Conhecendo a quantidade de fatores, a etapa seguinte se deu por criar um arranjo de misturas do tipo *simplex-lattice* de grau 10, que considerou valores axiais, gerando um total de 13 experimentos. A Tabela 5.1 apresenta os parâmetros experimentais para a rotação γ (representado por γ_1) e, por se tratar de um DOE de proporções, um valor complementar (representado por γ_2). É importante ressaltar que, independentemente da extensão do conjunto em análise, o arranjo experimental para essa etapa será sempre o mesmo.

Tabela 5.1. Variáveis de controle e níveis para o arranjo de misturas

Parâmetros	Níveis	
γ_1	0	1
γ_2	0	1

Em seguida, realizou-se a aplicação da FA para cada um dos parâmetros experimentais definidos, considerando a extração de 7 fatores. Os valores calculados de VTE, conforme a Eq. (3.1), para cada fator em seu respectivo experimento, estão descritos na Tabela 5.2. Com base nos valores de VTE, calculou-se a função EQM_{VTE}, conforme descrito na Eq. (3.2), também disponíveis na Tabela 5.2. Para realização dessa formulação, encontrou-se o valor alvo (representado pela média dos valores VTE igual a 2,0811) e a variância calculada para cada uma das linhas.

Tabela 5.2. Matriz experimental para o *simplex-lattice* – 1ª iteração

Teste	Controle		Respostas								
	γ_1	γ_2	VTE_1	VTE_2	VTE_3	VTE_4	VTE_5	VTE_6	VTE_7	σ^2	**EQM_{VTE}**
1	1	0	3,448	2,793	2,522	2,320	1,850	1,387	0,250	1,086	14,116
2	0,9	0,1	3,504	2,806	2,527	2,315	1,828	1,343	0,245	1,131	14,703
3	0,8	0,2	3,560	2,819	2,533	2,310	1,806	1,300	0,240	1,178	15,308
4	0,7	0,3	3,616	2,834	2,539	2,304	1,781	1,259	0,236	1,225	15,930
5	0,6	0,4	3,670	2,850	2,545	2,298	1,754	1,220	0,232	1,274	16,562
6	0,5	0,5	3,721	2,866	2,551	2,292	1,726	1,183	0,229	1,322	17,191
7	0,4	0,6	3,770	2,883	2,558	2,286	1,696	1,149	0,225	1,371	17,821
8	0,3	0,7	3,817	2,902	2,565	2,279	1,665	1,119	0,223	1,419	18,441
9	0,2	0,8	3,860	2,921	2,573	2,272	1,631	1,091	0,220	1,466	19,052
10	0,1	0,9	3,900	2,941	2,581	2,265	1,597	1,066	0,218	1,511	19,648
11	0	1	3,937	2,962	2,590	2,257	1,561	1,044	0,216	1,556	20,232
12	0,75	0,25	3,588	2,827	2,536	2,307	1,794	1,279	0,238	1,201	15,619
13	0,25	0,75	3,839	2,911	2,569	2,276	1,648	1,105	0,221	1,442	18,747

A partir desses resultados, realizou-se a análise dos experimentos considerando os valores calculados de EQM_{VTE} para um modelo de quarta ordem completo. Os coeficientes de regressão apresentaram valores de ajustes de R^2 e R^2_{adj} iguais a 100%. Além disso, o ajuste preditivo também apresentou um valor elevado, indicando que o ajuste experimental foi capaz de realizar previsões adequadas. A análise de variância das proporções indicou que todos os termos (linear, quadrático, cúbico e de quarta ordem), apresentaram valores significativos, com *p-values* iguais a 0.000 para os modelos. A Tabela 5.3 apresenta análise do EQM_{VTE} para o arranjo experimental.

Tabela 5.3. Análise de variância para EQM_{VTE} (proporções de componente)

Fonte	Graus de Liberdade	Soma dos Quadrados (seq.)	Soma dos Quadrados (ajust.)	Quadrado Médio (ajust.)	F	P
Regressão	4	46,7775	46,7775	11,6944	7436376,1	0,000
Linear	1	46,7725	21,6021	21,6021	13736615	0,000
Quadrático $\gamma_1*\gamma_2$	1	0,0004	0,0005	0,0005	295,71	0,000
Cúbico completo $\gamma_1*\gamma_2*(-)$	1	0,0045	0,0045	0,0045	2860,3	0,000
Quártico completo $\gamma_1*\gamma_2*(-)^2$	1	0,0001	0,0001	0,0001	45,69	0,000
Erro de Resíduos	8	0,0000	0,0000	0,0000		
Total	12	46,7775				

Os resultados estatísticos, apresentados anteriormente, podem ser também verificados de maneira gráfica, em que o comportamento das mudanças do valor γ nos valores das VTE's e no EQM_{VTE} podem ser ilustrados pelo traço de resposta Cox (Figura 5.5). Ao verificar os gráficos das VTE's para os fatores 1 a 7, notou-se que as mesmas apresentaram um comportamento sutilmente quadrático. Já para a Figura 5.5(h), que representa os valores para o EQM_{VTE}, foi possível verificar um comportamento mais linear, mesmo apresentando valores significativos para análises quadráticas, cúbicas e de quarta ordem. Tais afirmações podem ser, estatisticamente, confirmadas através da Tabela 5.3. A Figura 5.6 apresenta o comportamento dos gráficos de superfície de resposta e contornos das variáveis VTE's. É importante ressaltar que os gráficos de superfície representam o comportamento do modelo de regressão das VTE's, contemplando toda amplitude dos parâmetros de γ_1 e γ_2. Contudo, na etapa de otimização, foi preciso considerar a restrição de igualdade ($\gamma_1 + \gamma_2 = 1$).

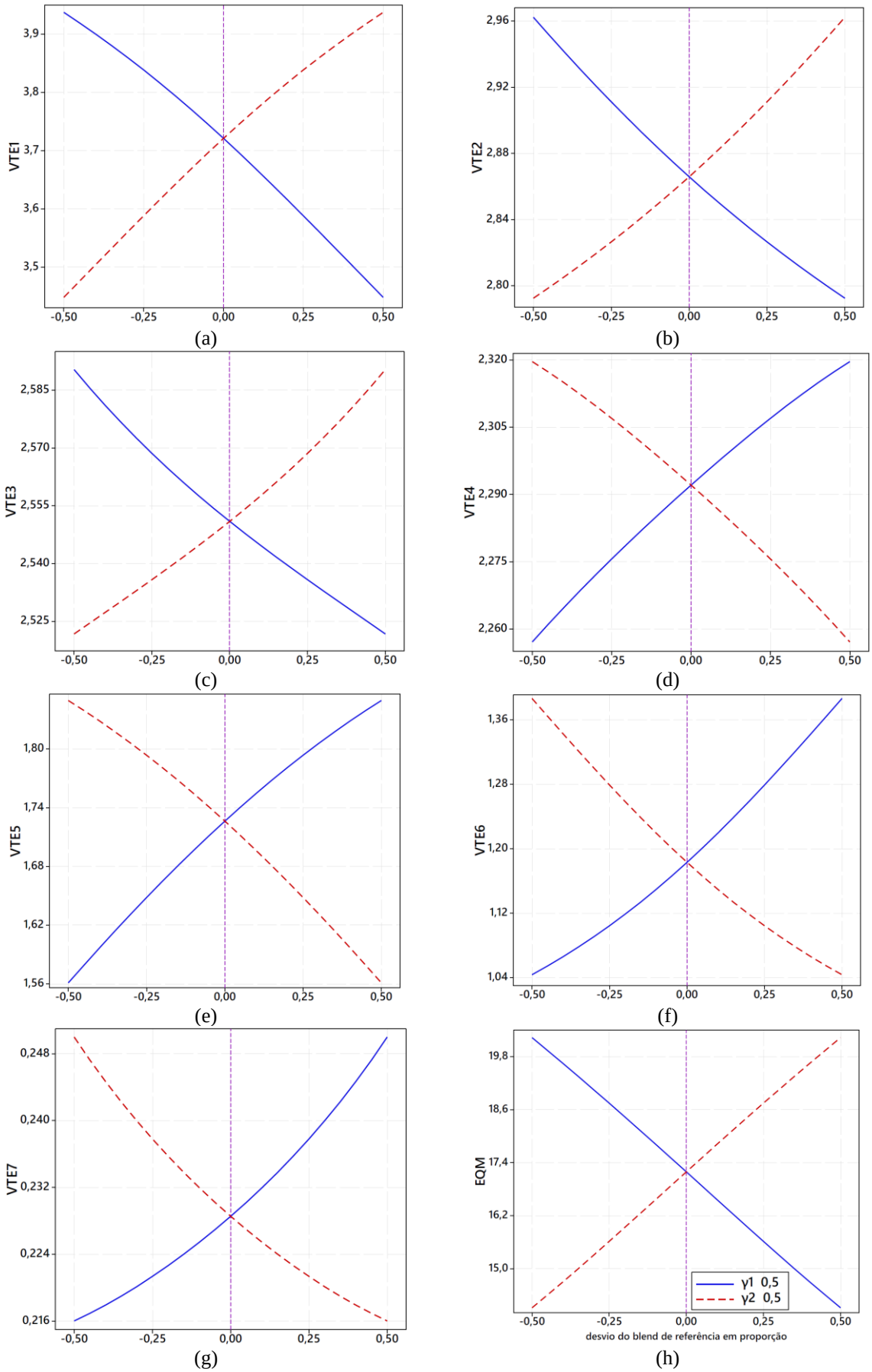


Figura 5.5. Gráficos de traço de resposta Cox para (a) VTE_1, (b) VTE_2, (c) VTE_3, (d) VTE_4, (e) VTE_5, (f) VTE_6, (g) VTE_7 e (g) EQM_{VTE}– 1ª iteração

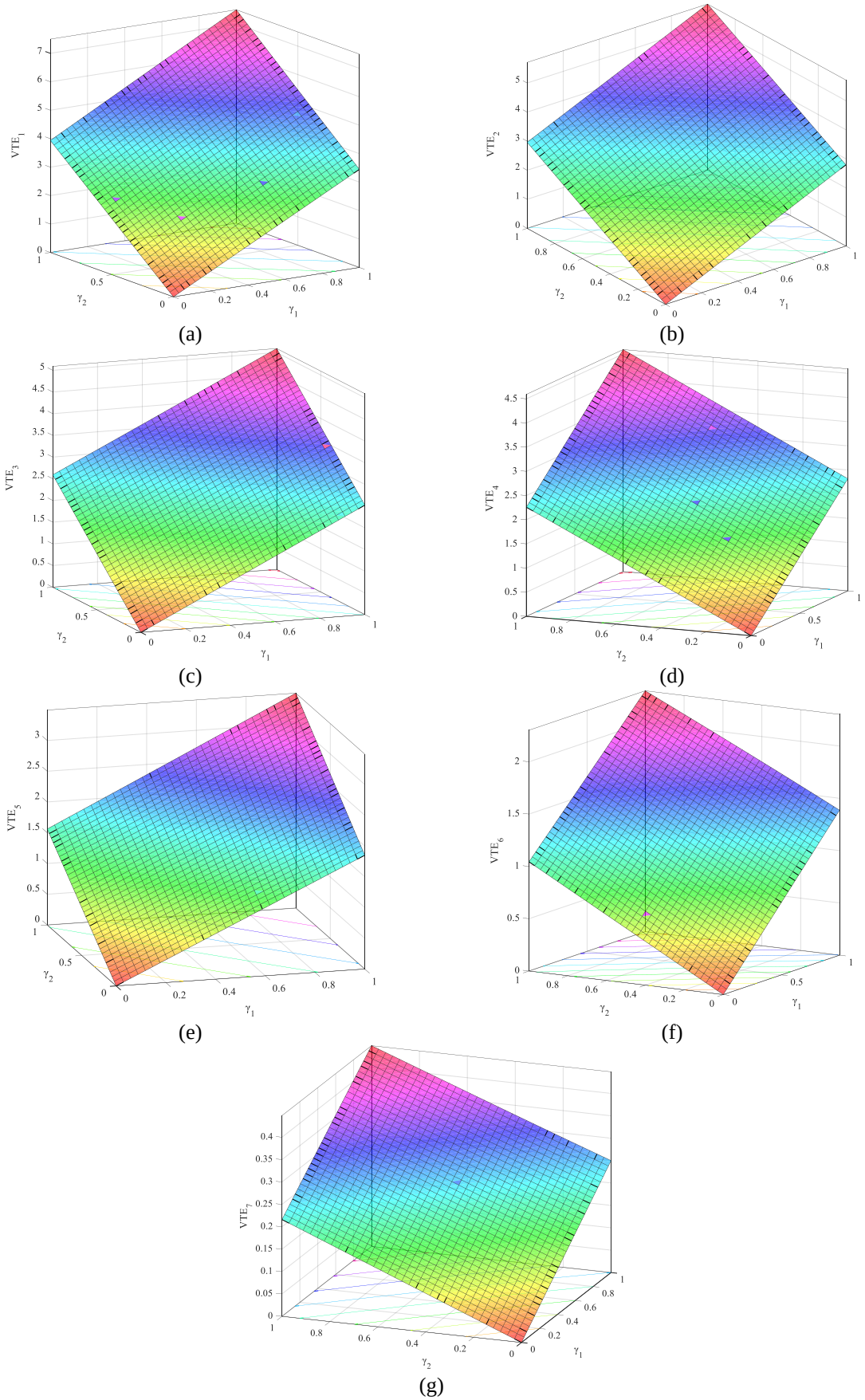


Figura 5.6. Gráficos de superfície e contorno para (a) *VTE₁*, (b) *VTE₂*, (c) *VTE₃*, (d) *VTE₄*, (e) *VTE₅*, (f) *VTE₆* e (g) *VTE₇* – 1ª iteração

5.2.3 Otimização do EQM para o valor *orthomax γ*

A partir dos resultados e análises experimentais apresentados nessa etapa, foi possível ajustar o modelo para realizar a otimização, buscando encontrar o valor ótimo da rotação γ. Assim, a partir dos coeficientes calculados para o EQM_{VTE} no arranjo de misturas, obteve-se a equação de regressão que representa seu comportamento, considerando o modelo descrito na Eq. (4.1).

$$\begin{aligned} EQM_{VTE} = {} & 14{,}1167\times\gamma_1 + 20{,}2317\times\gamma_2 + 0{,}07270\times\gamma_1\times\gamma_2 - 0{,}38278\times\gamma_1\times\gamma_2\times(\gamma_1-\gamma_2) \\ & - 0{,}1004\times\gamma_1\times\gamma_2\times\left((\gamma_1-\gamma_2)^2\right) \end{aligned} \tag{4.1}$$

Com base nessa equação, foi possível realizar a otimização de objetivo único, com sentido de minimização, visto que, buscou-se o menor valor para o erro quadrático médio desse modelo. Para alcançar esse resultado, utilizou-se o algoritmo SQP, detalhado no capítulo 2. É importante ressaltar que a função objetivo estava sujeita a uma restrição de igualdade, que representa o espaço experimental, delimitada pelo arranjo de misturas, conforme detalhado na Eq. (4.2).

$$\gamma_1 + \gamma_2 = 1 \tag{4.2}$$

A aplicação do algoritmo proporcionou encontrar um ponto ótimo para EQM_{VTE} de valor igual 14,1167, sendo os parâmetros experimentais de $\boldsymbol{\gamma}$ = [1; 0]. A Figura 5.7 apresenta a superfície de resposta dessa modelagem destacando o ponto ótimo encontrado pelo SQP. Assim, o valor de rotação otimizado, que melhor explicou as variáveis latentes e apresentou o melhor nível de simplicidade para as cargas fatoriais, se deu pelo valor de γ igual a *1*, também representado pela rotação intitulada de *varimax*.

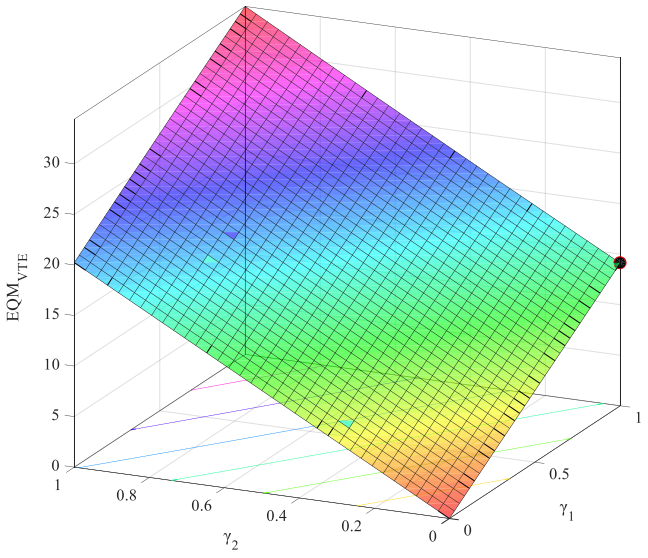


Figura 5.7. Gráfico de superfície de resposta e contorno do EQM_{VTE} – 1ª iteração

5.2.4 Extração dos escores de fator rotacionados

Conhecendo o valor ótimo para o parâmetro γ, foi possível extrair os escores de fator rotacionado para representar, de maneira adequada, os dados originais estabelecidos. Assim, tem-se na Tabela 5.4 os escores de fatores otimizados que foram utilizados nas etapas subsequentes.

Tabela 5.4. Escores dos fatores com rotação otimizada ($\gamma = 1$)

F_1	F_2	F_3	F_4	F_5	F_6	F_7
-1,2356	0,0952	0,3908	-1,9595	-0,6494	-1,0273	0,0405
0,6704	-0,6477	1,4893	-1,5731	0,1818	1,0797	-0,0390
0,3369	-0,8376	1,9395	0,2237	-0,4696	-1,5422	1,0523
-1,7542	0,4494	-0,5446	-1,3847	0,0363	0,6221	-0,1150
0,2734	0,2875	0,6511	0,1780	0,9267	0,3212	-0,6648
1,2270	0,3358	-1,0811	-1,3065	-0,4756	-0,0422	-1,5607
1,7073	1,6521	-0,3521	-0,1993	-1,3726	0,9861	2,1674
-1,2981	1,9039	-0,3871	1,2611	-0,1828	-0,7744	-0,0086
0,6556	1,2124	0,5218	0,2503	2,8976	-0,2101	0,5408
0,9135	-0,1081	-0,2584	0,3280	-0,6076	-2,4517	-0,9129
-1,0326	-1,1936	-0,5769	0,3490	0,6776	0,4366	0,7186
1,1513	-1,8761	-1,2935	0,3730	0,6886	0,1984	-0,1497
-0,2096	-0,7568	-1,0992	0,8816	-1,0809	0,5240	0,7216
0,0383	0,5890	-0,7021	0,1491	0,2927	0,6327	-1,3078
-0,1203	0,0076	0,2481	1,2679	-0,4701	0,1378	-0,7404
-0,8376	-0,8330	-0,7058	-0,0663	0,5353	-0,3788	1,3171
-0,4855	-0,2800	1,7602	1,2276	-0,9279	1,4881	-1,0594

A importância de melhorar a rotação γ, para a interpretação dos dados, foi visível ao se analisar o comportamento dos agrupamentos das variáveis a partir da alocação das cargas nos fatores. As informações de cargas fatoriais e comunalidades da extração dos fatores com rotação otimizada ($\gamma = 1$) estão descritas na Tabela C.5 (Apêndice C), comparando-as com o uso da rotação *quartimax* ($\gamma = 0$) e também da aplicação de FA sem rotação. Os resultados com rotação γ das cargas fatoriais proporcionaram uma melhor separabilidade das características com maior similaridade, o que significou uma melhor interpretação da informações latentes [5]. Ao analisar os dados com rotação *quartimax*, observou-se que a interpretação das informações latentes foi mais confusa, uma vez que as variáveis de resposta não estavam separadas de modo adequado, apresentando uma grande quantidade de cargas e informações em um único fator.

Por fim, pôde-se verificar que a variabilidade das VTE nas cargas fatoriais atribuídas ao valor γ otimizado foi igual a 1,085, apresentando um valor de variância inferior em relação à rotação *quartimax* (com variabilidade igual a 1,556), ou mesmo sem o uso de rotação ortogonal (variabilidade de 3,073). Assim, foi possível concluir que a método proposto nesta etapa auxiliou na decisão de escolha de um valor de rotação γ adequado, a partir das opções da família

orthomax. A rotação ideal forneceu uma estrutura simples que promoveu uma análise mais concisa na interpretação dos dados, com menor confundimento ao investigar dados correlacionados para a tomada de decisão.

A partir dos valores de escores de fator extraídos (Tabela 5.4), partiu-se para a etapa seguinte, que considerou os métodos de ligação hierárquicos e não hierárquico, além de diferentes tipos de análise dentro de um único planejamento experimental (arranjo fatorial multiníveis). Com finalidade de demonstrar o comportamento da abordagem para otimização da rotação γ *orthomax*, o Apêndice D descreve uma aplicação alternativa, com dados de grande extensão referentes a análise da degradação de motores *turbofan*.

5.3 Aplicação do arranjo fatorial multiníveis

5.3.1 Definição do arranjo experimental multiníveis

Considerando os vetores de escores rotacionados que foram extraídos, os quais representaram de maneira adequada a estrutura de dados originais, foi possível realizar as aplicações referentes aos métodos de amalgamação, além do método não hierárquico *k-médias*. Ademais, os resultados coletados, para cada um dos métodos de ligação, foram investigados por diferentes estratégias de análise (ANOVA e ANCOVA). Tal cenário se caracterizou como um outro tipo de planejamento de experimentos, sendo este o arranjo fatorial. Contudo, esse modelo teve uma particularidade, em que um dos fatores (métodos de ligação) apresentou uma amplitude maior de níveis (neste caso, oito), que precisaram ser consideradas. Assim, deparou-se com um arranjo fatorial completo do tipo generalizado, ou simplesmente, arranjo fatorial multiníveis. Inicialmente, foi feita a delineação desse DOE, em que foram consideradas duas variáveis de controle: Métodos de Ligação e Tipo de Análise, apresentando 8 e 2 níveis, respectivamente. Esses parâmetros estão descritos na Tabela 5.5.

Tabela 5.5. Variáveis de controle e níveis do arranjo fatorial multiníveis

Parâmetros	Níveis							
Método de Ligação	Único	Centroide	Completa	Média	Mediana	McQuitty	Ward	k-médias
Tipo de Análise	ANOVA				ANCOVA			

A partir da definição dos parâmetros, criou-se o arranjo experimental necessário para esta análise. Considerando os níveis dos parâmetros apresentados na Tabela 5.5, o arranjo fatorial obtido contemplou um total de 16 experimentos. É importante ressaltar que, tratando-se de um

experimento com respostas calculadas, que faz uso de técnicas de análises não-heurísticas para um mesmo conjunto de dados, a criação de réplicas não proporcionaria a variabilidade necessária para uma análise, como, por exemplo, em um processo industrial, em que há variabilidade nos experimentos devido a diversos fatores (peça, operador, temperatura, ambiente etc). Deste modo, este arranjo não contemplou o uso de réplicas.

5.3.2 Aplicação dos métodos de ligação e tipo de análise

Após conhecer o arranjo experimental, as associações geradas pela análise de cluster foram confrontadas com a resposta de interesse, TNE. Assim, inicialmente, foram aplicadas e detalhadas a amalgamação dos sete métodos hierárquicos e também do método não hierárquico. Como apresentado anteriormente, os escores rotacionados, extraídos pelo método de componentes principais, não apresentaram uma estrutura multivariada. Em tese, esses escores apresentaram correlação de *Pearson* igual a zero, com *p-values* iguais a 1. Tal resultado pode ser conferido na Tabela 5.6, em que se realizou o teste de correlação, provando esta premissa. Assim, diante das características dos escores, verificou-se que a medida de distância *Mahalanobis* não deve ser aplicada, sendo a medida euclidiana definida como padrão para todas análises de amalgamação.

Tabela 5.6. Correlação dos escores de fator rotacionados da 1ª iteração

	F_1	F_2	F_3	F_4	F_5	F_6
F_2	0,000[(1)] 1,000[(2)]					
F_3	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]				
F_4	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]			
F_5	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]		
F_6	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	
F_7	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]	0,000[(1)] 1,000[(2)]

[(1)]Correlação de Pearson
[(2)]*p-value*

A seguir, foi definida, para ambos métodos de ligação (hierárquicos e não hierárquicos), a quantidade de clusters, ou agrupamentos, formados. Baseado na Regra de Sturges, em que k_c

$= 1 + 3{,}322log(17)$, sendo 17 a quantidade de subestações do conjunto de dados, o número de clusters adequados para a análise foi $k_c = 5{,}0875 \cong 5$ clusters. A partir dessas definições foi possível aplicar cada um dos métodos de amalgamação e armazenar os valores de suas associações. A Tabela 5.7 apresenta estas associações geradas por cada estratégia de amalgamação. Devido sua caracterização hierárquica, esses métodos de amalgamação podem ser representados, graficamente, através de dendrogramas, conforme apresentados na Figura 5.8, em que é possível verificar o comportamento nas ramificações e separações realizadas a partir do nível de similaridade entre as subestações.

Tabela 5.7. Associações dos agrupamentos gerados pelos métodos de ligação

Subestações	Único	Centroide	Completa	Média	Mediana	McQuitty	Ward	k-*médias*
Aracruz	1	1	1	1	1	1	1	1
Baixo Guandu	2	1	1	1	1	1	1	2
Barra do Sahy	3	2	2	2	1	1	1	3
Ecoporanga	1	1	1	1	1	1	1	4
Itarana	2	1	3	1	1	2	2	5
Jaguaré	2	1	4	3	1	3	3	5
João Neiva	4	3	5	4	2	4	4	5
Juncado	2	1	3	1	3	2	2	5
Linhares A	2	4	3	5	4	5	2	5
Linhares C	5	1	4	3	5	3	3	3
Montanha	2	1	4	1	1	3	5	4
Nova Venécia	2	1	4	1	1	3	5	5
Paulista	2	1	4	1	1	3	5	5
Pinheiros	2	1	3	1	1	2	2	5
Santa Tereza	2	1	3	1	1	2	2	5
São Francisco	2	1	4	1	1	3	5	4
Suíça	2	5	2	1	1	2	2	5

Após a realização dos métodos hierárquicos, o mesmo foi feito ao método não hierárquico *k-médias*. A parametrização dessa abordagem se diferenciou das demais, visto que o método apresenta características distintas em sua aplicação. Sabe-se que o método *k-médias* pode ser considerado uma meta-heurística, visto que seus resultados podem apresentar diferentes valores, dependendo da partição inicial adotada. Contudo, como a quantidade de clusters para esta aplicação foi definida (conforme indicado através da *Regra de Sturges*), observou-se que a estratégia *k-médias* não apresentou esse comportamento, logo, não houve a necessidade de várias réplicas, pois trata-se de um resultado determinístico. A Tabela 5.7 também apresenta as associações criadas a partir dessa aplicação. Como não se trata de uma amalgamação, os métodos não hierárquicos não apresentaram o processo de distinção através dos dendrogramas. As tabelas e demais estatísticas geradas pela aplicação dos métodos hierárquicos e do método não hierárquico, tais como valores e distâncias de centroides, estão disponíveis no Apêndice C.

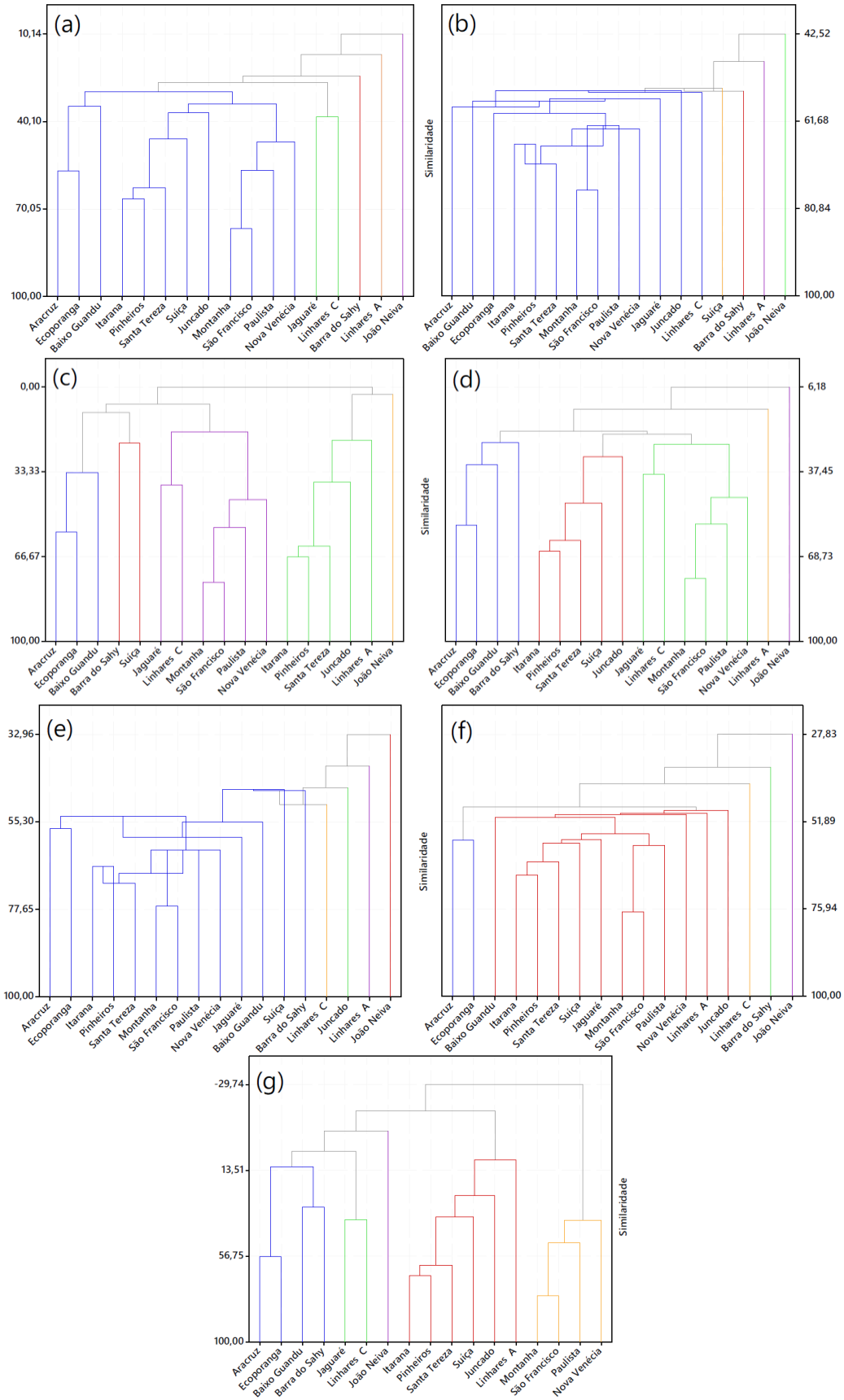


Figura 5.8. Dendrogramas das amalgamações (a) Único, (b) Centroide, (c) Completa, (d) Média, (e) Mediana, (f) McQuitty e (g) Ward

Com as associações geradas pelos diferentes métodos de ligação, foi possível seguir para as análises ANOVA e ANCOVA, que contemplam o segundo fator de controle do arranjo fatorial multiníveis. Esses procedimentos foram realizados considerando as associações dos agrupamentos, confrontando-os com os valores originais de afundamento de tensão, TNE.

A partir da aplicação da estratégia ANOVA, foi possível criar intervalos de confiança univariados para cada um dos clusters formados. Esses intervalos proporcionaram entender o comportamento da TNE dentro de cada um dos agrupamentos para um determinado nível de confiança. A Figura 5.9 e a Figura 5.10 apresentam, à esquerda, os valores médios e os respectivos IC gerados pela ANOVA. Destaca-se a mudança de comportamento para cada um dos métodos de ligação, visto que certas estratégias apresentaram uma determinada similaridade entre algumas subestações, enquanto manteve outras isoladas. Esse resultado pode indicar que há uma possível dificuldade (para esse determinado método) em realizar a separabilidade do conjunto de dados em análise. Clusters com apenas uma subestação utilizam o desvio padrão combinado para criar os IC. É importante destacar que, para todas as análises, foi considerado um α de 0,05, ou seja, um IC de 95%.

A aplicação do método ANCOVA apresentou algumas similaridades com o método ANOVA, contudo, para realizar a análise, foi preciso considerar uma variável concomitante. Essa variável deve estar relacionada com a variável de interesse TNE. Para essa definição, foi possível utilizar as características da estrutura de variância-covariância, em que pôde-se verificar algumas das variáveis que mais impactam nos afundamentos de tensão, conforme já discutido no trabalho de Almeida *et al.* [62]. Deste modo, notou-se que a NEMV foi a variável que teve mais influência no TNE, contudo, ao investigá-la de maneira detalhada, observou-se que a NEMV apresentou uma dependência linear com o TNE, uma vez que a variável foi utilizada em sua formulação (apresentando um coeficiente de *Pearson = 1,000*). Assim, tal variável não foi a mais adequada para essa análise, devido a sua característica. Após a NEMV, a variável que mais impactou nos afundamentos de tensão foi a "área de vulnerabilidade na média tensão", ou simplesmente EMVVA. Essa variável foi adquirida pela média das áreas pela sua extensão (em quilômetros) e se caracteriza pelo uso na definição de regiões em que os curtos-circuitos podem influenciar cargas de uma barra específica. De acordo com Miranda Filho [70], a região em análise (do ES) apresenta nós e segmentos de linha que podem levar a eventos de afundamento de tensão na barra, devido a ocorrência de curtos-circuitos. Tais caracterizações estão descritas na norma IEEE 1564 [112]. A variável EMVVA apresentou uma multicolinearidade significativa com TNE, proporcionando um valor de *Pearson* de 0,712, com *p-value* igual a 0,001.

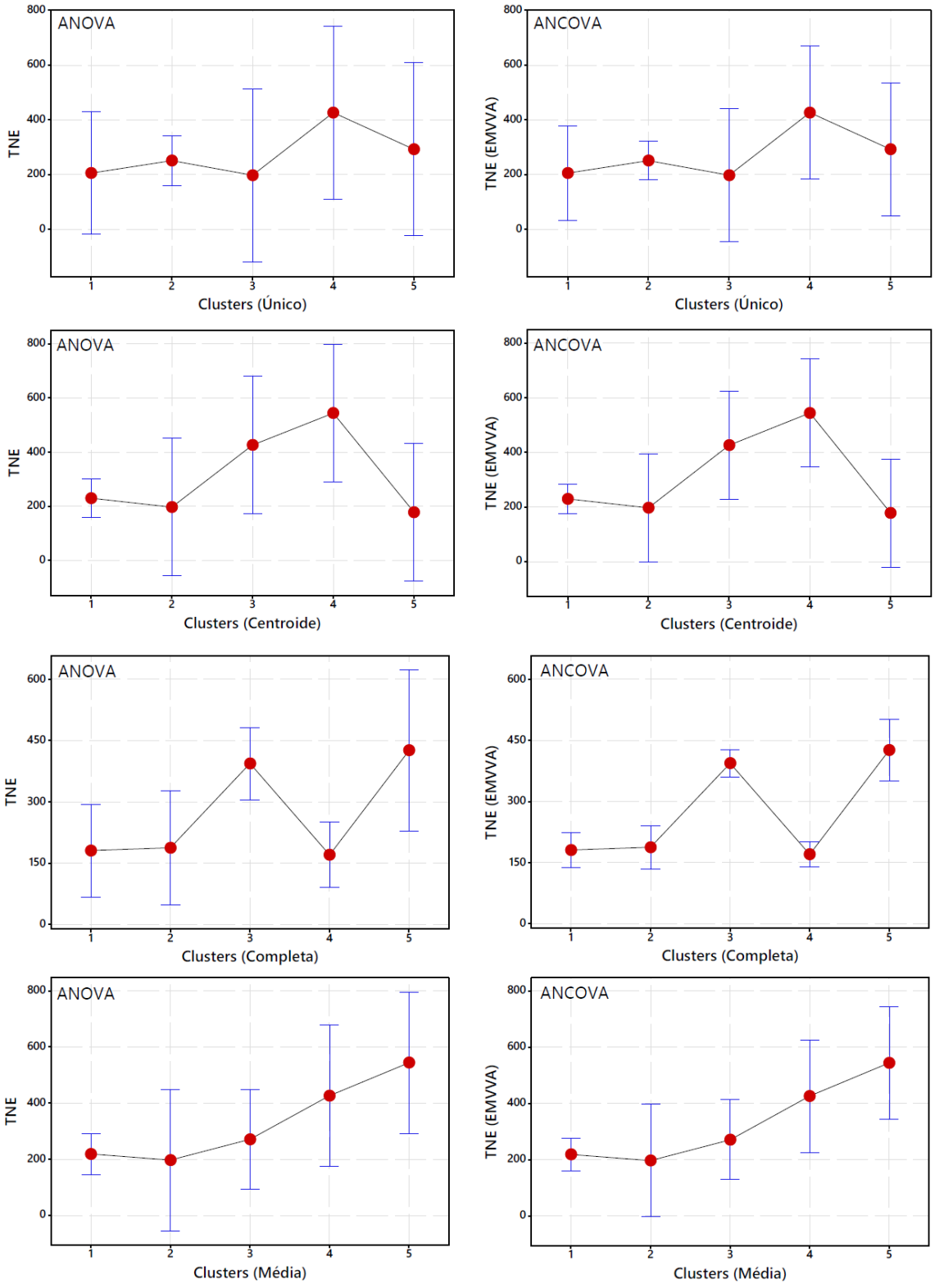


Figura 5.9. Intervalos de ANOVA e ANCOVA para os métodos de ligação (parte I)

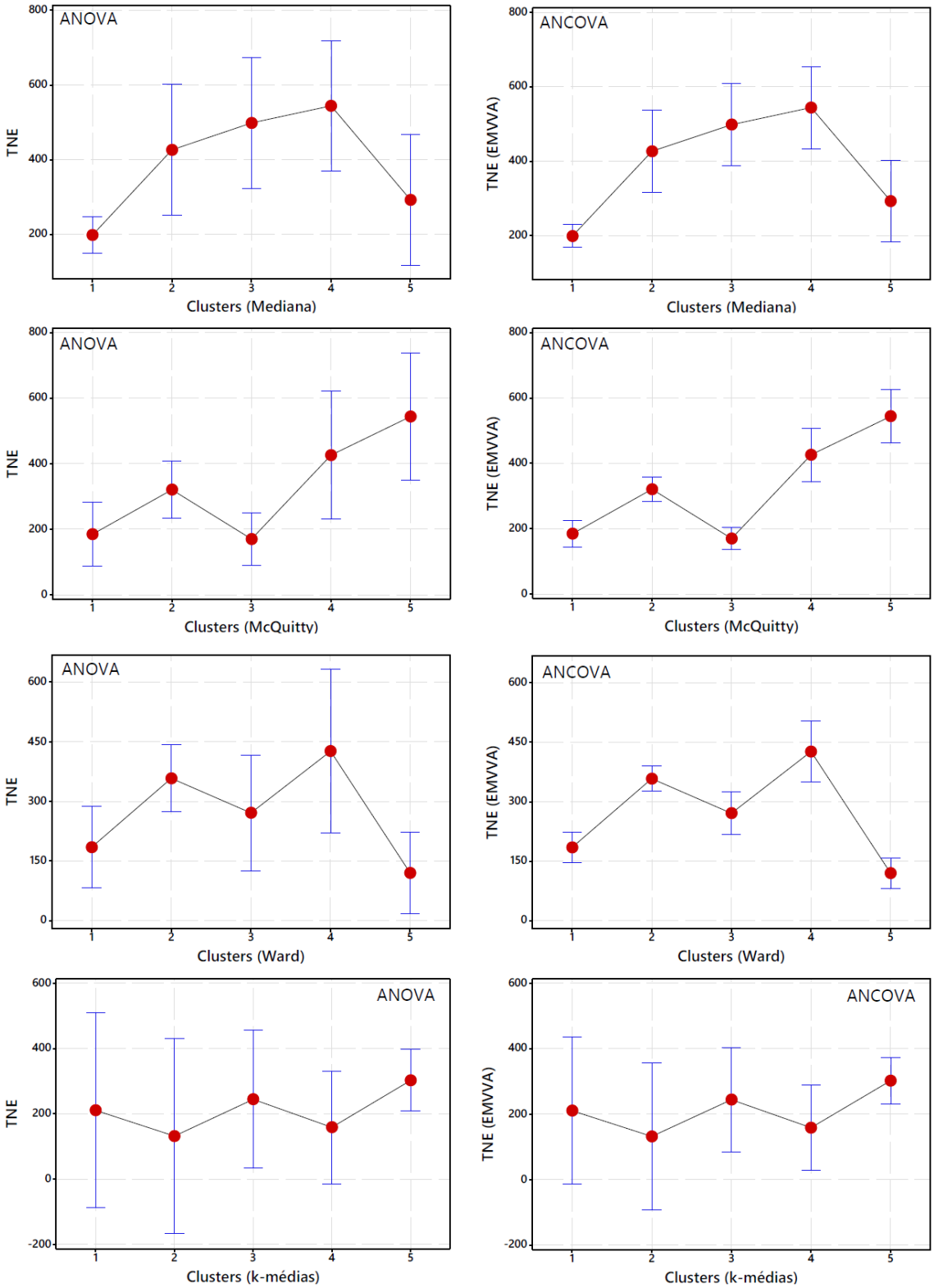


Figura 5.10. Intervalos de ANOVA e ANCOVA para os métodos de ligação (parte II)

Com a definição da variável concomitante para o estudo, foi possível realizar a ANCOVA e gerar seus respectivos intervalos de confiança, considerando as mesmas características anteriores. A partir das Figura 5.9 e Figura 5.10 foi possível verificar, à direita, os gráficos de

intervalo para cada um dos métodos de ligação, utilizando a ANCOVA. Para um melhor comparativo de ambas estratégias (ANOVA e ANCOVA), os gráficos de intervalo foram posicionados horizontalmente, para cada método de ligação, além disso, as escalas do eixo das ordenadas foram padronizadas. Deste modo, foi possível verificar, incialmente, que os valores médios de cada cluster se mantiveram inalterados e que, no uso da ANCOVA, os clusters apresentaram intervalos de confiança mais estreitos, indicando que os resultados apresentaram menor variabilidade, logo, foram mais precisos. Os valores de TNE, ajustados para EMVVA, na a análise ANCOVA, estão disponíveis na Tabela C.10 do Apêndice C.

5.3.3 Análise do arranjo experimental

Considerando as aplicações anteriores, foi possível armazenar os valores de média e desvio padrão na formação de clusters, a partir dos diagnósticos da ANOVA e ANCOVA. Tais informações foram utilizadas para o complemento do método proposto. Assim, a partir dos desvios padrão coletados, calculou-se as variâncias na formação de cada cluster, representando, assim, as respostas de interesse para o arranjo fatorial multiníveis, conforme detalhado na Tabela 5.8.

Tabela 5.8. Matriz experimental do arranjo multiníveis

Teste	**Parâmetros**		**Respostas**							
			σ^2 dos Clusters					Escores dos Clusters		
	Ligação	Análise	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	FC_1	FC_2	FC_3
1	Único	ANOVA	72,98114	23064,71	21148,72	21148,72	21148,72	2,2410	1,9046	-0,9619
2	Único	ANCOVA	15341,92	12206,29	12467,51	12467,51	12467,51	0,9421	-0,1979	1,1403
3	Centroide	ANOVA	13559,74	13559,74	13559,74	13559,74	13559,74	1,1050	0,0588	0,8947
4	Centroide	ANCOVA	8258,593	8258,593	8258,593	8258,593	8258,593	0,3643	-0,3972	0,1873
5	Completa	ANOVA	1809,073	178,5537	13892,98	7804,998	8199,465	1,6139	-1,6201	-0,3913
6	Completa	ANCOVA	1474,706	4323,681	964,5994	651,6533	1199,147	-0,8588	-0,6093	-0,7951
7	Média	ANOVA	14511,74	13377,47	899,58	13377,47	13377,47	-0,6094	0,8519	0,9618
8	Média	ANCOVA	8982,073	8429,554	2351,92	8429,554	8429,554	-0,4344	0,0114	0,2546
9	Mediana	ANOVA	6446,163	6446,163	6446,163	6446,163	6446,163	0,1111	-0,5531	-0,0546
10	Mediana	ANCOVA	2569,78	2569,78	2569,78	2569,78	2569,78	-0,4306	-0,8865	-0,5719
11	McQuitty	ANOVA	1272,72	13206,58	7804,998	7972,454	7972,454	0,0638	0,5952	-0,9031
12	McQuitty	ANCOVA	2179,471	1384,249	934,5922	1395,702	1395,702	-0,6629	-0,9921	-0,6259
13	Ward	ANOVA	1272,72	18879,14	899,58	8896,244	2547,181	-1,2712	1,8616	-1,4545
14	Ward	ANCOVA	1949,152	990,9274	657,1481	1244,185	1156,979	-0,6887	-1,0262	-0,6550
15	k-médias	ANOVA	18795,04	18795,04	4531,565	1856,731	24143,85	-0,7527	0,5368	2,3977
16	k-médias	ANCOVA	10626,52	10626,52	192,1911	9407,758	12056,65	-0,7324	0,4621	0,5770

Tratando-se de um arranjo experimental sem réplicas, observou-se que a quantidade de experimentos foi igual ao número de termos (16). Sabendo que o grau de liberdade para o erro experimental foi definido pelo número de experimentos subtraídos pela quantidade de termos,

para este caso, o modelo completo apresentaria zero graus de liberdade, o que impossibilitaria a estimação da variância experimental e a significância dos efeitos. Diante dos critérios de hierarquia e espacialidade dos efeitos, removeu-se a interação de ordem superior, podendo, assim, calcular o erro e testar a significância dos efeitos de primeira ordem.

Considerando os dados que complementam o arranjo fatorial multiníveis, foi possível realizar, previamente, a análise individual de cada um dos clusters, mediante os diferentes parâmetros. A não inserção de réplicas, por motivos justificados anteriormente, promoveu a remoção do modelo de segunda ordem para realizar a análise. Deste modo, estabeleceu-se os modelos de primeira ordem utilizando os mínimos quadrados ponderados (*WLS – weighted least squares*) para cada resposta de clusters. Com essa análise, foi possível verificar que ambos fatores de análise do DOE foram significativos para os Clusters 2, 3, 4 e 5, apresentando *p-values* menores que 0,05 e ajustes adequados. Contudo, a análise do Cluster 1 (R^2_{adj} de 74,68%) indicou que, para essa resposta, a escolha do método de análise não se mostrou significativa, apresentando *p-value* igual a 0,435. Assim, removeu-se este termo e uma nova análise foi realizada, apresentando um sutil aumento no ajuste do modelo (R^2_{adj} = 75,68%). As informações dessas análises, além de outras interpretações estatísticas, estão descritas na Tabela 5.9. Para demonstrar os coeficientes completos dos termos, a Tabela 5.9 apresenta os valores para o "Tipo de Análise", com seus respectivos valores de ajuste.

Tabela 5.9. Coeficientes e ajustes de regressão dos Clusters

Termos		Coeficientes				
		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Constante		6820,15	9768,56	6098,7	7842,95	9058,06
Ligação	*Único*	887,3	7866,94	10709,4	8965,16	7750,06
	Centroide	4089,02	1140,6	4810,4	3066,21	1851,11
	Completa	-5178,26	-7517,45	1330,1	-3614,63	-4358,75
	Média	4926,76	1134,95	-4473	3060,56	1845,45
	Mediana	-2312,18	-5260,59	-1590,8	-3334,98	-4550,09
	McQuitty	-5094,05	-2473,15	-1728,9	-3158,88	-4373,98
	Ward	-5209,21	166,47	-5320,4	-2772,74	-7205,98
	k-médias	7890,63	4942,22	-3736,9	-2210,71	9042,19
Análise	ANOVA	170,419	3794,57	2611,8	2460,4	3209,88
	ANCOVA	-170,419	-3794,57	-2611,8	-2460,4	-3209,88
R^2		88,18%	95,92%	99,46%	98,95%	99,11%
R^2_{adj}		74,68%	91,26%	98,83%	97,75%	98,09%
R^2_{pred}		50,92%	72,32%	94,04%	95,03%	95,21%

Conhecendo as adequações do modelo do DOE, foi possível verificar, através da Figura 5.11, os gráficos de efeitos principais para cada um dos clusters. Nestes gráficos, verificou-se

a influência dos métodos de ligação, em que o método “Completa” inferiu uma diminuição na variabilidade nos Clusters 1, 2 e 4, enquanto o método “Ward” se destacou por apresentar bons resultados para os Clusters 1, 3 e 5. Tal comportamento se diferiu do método de ligação não hierárquico, *k-médias*, o qual apresentou uma alta variabilidade nos modelos para os Clusters 1, 2, 4 e 5. O mesmo comportamento foi verificado para o método “Único”, que apresentou alta variabilidade na formação dos Clusters 2, 3, 4 e 5. Em relação ao tipo de análise, com exceção do Cluster 1 que não foi significativo, a investigação através da ANCOVA promoveu uma redução significativa na variabilidade para os demais Clusters. Esse resultado mostrou que o uso da variável concomitante, neste caso a EMVVA, promoveu resultados mais estáveis e precisos, com menor variabilidade, conforme o comportamento dos IC demonstrados anteriormente na Figura 5.9 e Figura 5.10.

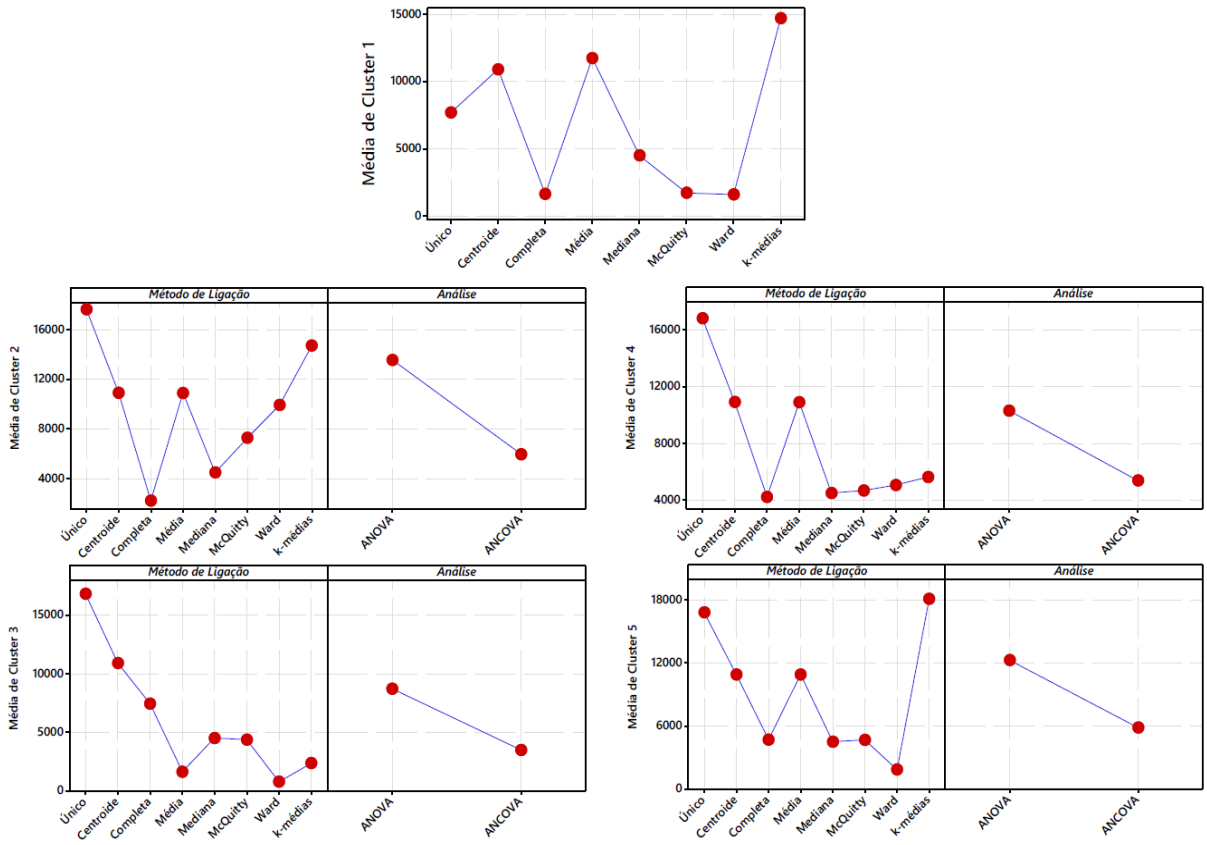
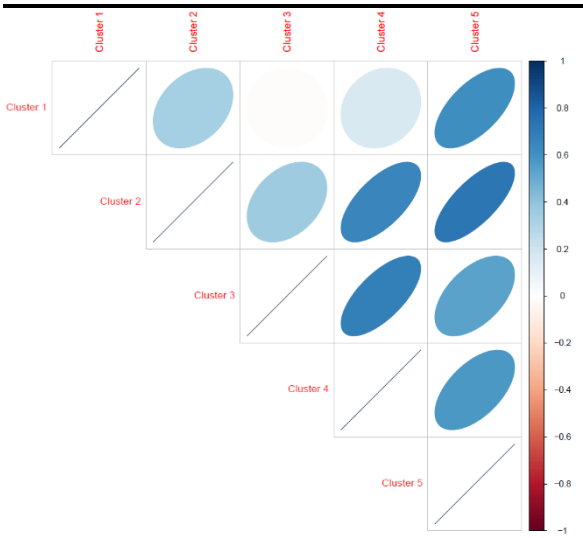


Figura 5.11. Gráficos de efeitos principais para os Clusters

Tratando-se de múltiplas respostas (referentes aos valores de variância na formação dos cinco clusters), se fez necessário verificar a estrutura e comportamento dos dados. Ao realizar a análise de correlação, observou-se que os dados apresentaram uma estrutura de variância-covariância significativa, inferindo que as informações dos clusters não são totalmente independentes, logo, apresentam uma natureza multivariada. Os valores de *Pearson* e *p-value*

desse teste estão descritos na Tabela 5.10. Tal comportamento é também verificado conforme a matriz de dispersão, ilustrada na Figura 5.12.

Tabela 5.10. Análise de correlação de Pearson para os Clusters



	Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
Cluster 2	0,331[(1)] 0,211[(2)]			
Cluster 3	-0,017[(1)] 0,949[(2)]	0,359[(1)] 0,172[(2)]		
Cluster 4	0,163[(1)] 0,546[(2)]	0,653[(1)] 0,006[(2)]	0,686[(1)] 0,003[(2)]	
Cluster 5	0,618[(1)] 0,011[(2)]	0,725[(1)] 0,001[(2)]	0,531[(1)] 0,034[(2)]	0,570[(1)] 0,021[(2)]

[(1)]Correlação de *Pearson*
[(2)]*p-value*

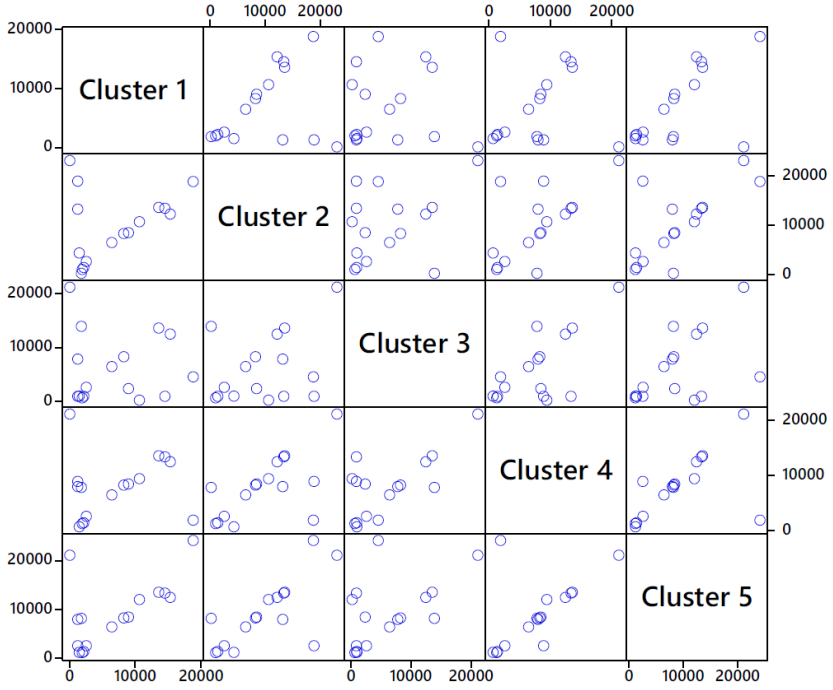


Figura 5.12. Matriz de dispersão para os dados da variabilidade na formação dos Clusters

5.3.4 Otimização da rotação *orthomax* para os clusters

Diante do comportamento multivariado desses dados, tem-se o *ciclo* (ou *loop*) do método proposto, conforme descrito no capítulo 3, em que foi necessário verificar se esses novos dados

em análise estavam aptos a aplicação da FA e, em caso positivo, encontrar o melhor valor γ de rotação *orthomax* para representar essas informações. Para cumprir esse *ciclo*, voltou-se a etapa para analisar a adequação dos dados. Pelo teste de normalidade multivariada de Mardia, observou-se que, apesar da medida de curtose ter sido significativa (*p-value* = 0,0843), o conjunto não seguiu uma distribuição multivariada normal, pois a medida de assimetria apresentou *p-value* igual a 0,000. Portanto, descartou-se o teste de esfericidade de Bartlett para essa análise. Seguiu-se, então, para a verificação através do índice KMO. A partir desta análise, obteve-se que o indicador apresentou valor igual a 0,5, concluindo que esse conjunto apresentou um comportamento adequado para o uso da estratégia multivariada de FA.

Para a próxima etapa do *ciclo*, realizou-se a otimização do valor γ de rotação *orthomax* para a FA. Seguindo o mesmo princípio apresentado anteriormente, determinou-se a quantidade de fatores necessários para análise a partir dos critérios de incremento apresentados no capítulo 3. Então, a quantidade adequada pôde ser definida pelo uso de 3 fatores, os quais satisfizeram o critério de incremento. A Figura 5.13 apresenta a carta de Pareto, trazendo as informações dos autovalores e porcentagem de explicação de cada fator.

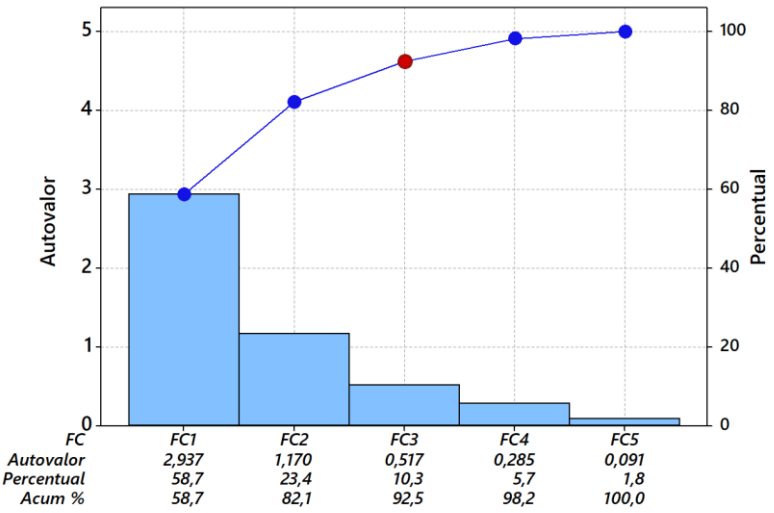


Figura 5.13. Carta de Pareto para os fatores dos Clusters

Conhecendo a quantidade de fatores, gerou-se um novo arranjo de misturas (que seguiu a mesma estrutura, como já foi apresentado), armazenando as informações de VTE para o cálculo do EQM desta etapa. A partir dos valores de VTE, determinou-se o alvo (igual a 1,5412) e também a variância de cada linha do arranjo de misturas. Os valores calculados de EQM_{VTE} e das VTE's estão descritos na Tabela 5.11.

Tabela 5.11. Matriz experimental do *simplex-lattice* para os fatores dos Clusters – 2ª iteração

Teste	Controle		Respostas dos fatores				
	γ_1	γ_2	VTE_1	VTE_2	VTE_3	σ^2	**EQM_{VTE}**
1	1	0	1,61990	1,55480	1,44900	0,00744	0,03720
2	0,9	0,1	1,62290	1,55440	1,44640	0,00792	0,03959
3	0,8	0,2	1,62620	1,55360	1,44400	0,00841	0,04207
4	0,7	0,3	1,63000	1,55190	1,44180	0,00894	0,04470
5	0,6	0,4	1,63460	1,55020	1,43890	0,00963	0,04817
6	0,5	0,5	1,63980	1,54770	1,43610	0,01041	0,05203
7	0,4	0,6	1,64660	1,54340	1,43370	0,01134	0,05668
8	0,3	0,7	1,65580	1,53620	1,43170	0,01257	0,06287
9	0,2	0,8	1,67030	1,52280	1,43060	0,01462	0,07309
10	0,1	0,9	1,69890	1,49150	1,43330	0,01949	0,09745
11	0	1	1,79500	1,45140	1,37730	0,04967	0,24835
12	0,75	0,25	1,62800	1,55290	1,44280	0,00868	0,04338
13	0,25	0,75	1,66220	1,53020	1,43120	0,01343	0,06715

A análise do arranjo de misturas para o EQM indicou que todos os termos foram significativos, considerando um modelo de quarta ordem completo, com *p-values* < 0,05. Com isso, o modelo proporcionou valores de R^2 e R^2_{adj} iguais a 96,76%, 95,14%, respectivamente. A Tabela 5.12 apresenta a análise de variância para o EQM, considerando as proporções de componente. O comportamento dos modelos para as VTE's e EQM podem ser visualizadas graficamente a partir dos traços de resposta Cox (Figura 5.14), em que o comportamento não linear do modelo foi visível para as três VTE's. O modelo de quarta ordem significativo foi visível na análise do EQM, conforme ilustrado na Figura 5.14(d). A Eq. (4.3) apresenta o modelo matemático de regressão dos coeficientes para o EQM, que foi utilizada na otimização, a seguir.

$$\begin{aligned} EQM_{VTE} = {} & 0,0425\times\gamma_1 + 0,2377\times\gamma_2 - 0,3319\times\gamma_1\times\gamma_2 + 0,4420\times\gamma_1\times\gamma_2\times(\gamma_1-\gamma_2) \\ & - 0,576\times\gamma_1\times\gamma_2\times\left((\gamma_1-\gamma_2)^2\right) \end{aligned} \tag{4.3}$$

Seguindo a mesma etapa do *ciclo*, realizou-se a otimização utilizando o algoritmo SQP diante das restrições experimentais do arranjo de misturas e a restrição de igualdade do modelo ($\gamma_1 + \gamma_2 = 1$). A partir dessa aplicação, encontrou-se o valor ótimo para EQM, sendo igual a 0,0308, além dos parâmetros otimizados, proporcionando um resultado de $\boldsymbol{\gamma}$ = [0,9022; 0,0978]. Esse resultado implicou em uma rotação *orthomax* com um valor γ igual a 0,9022, mesclando as rotações *varimax* e *quartimax*, estando mais próxima da primeira. Ressalta-se que essa função, em específico, apresentou pontos de mínimo local e global, em que a escolha dos pontos iniciais pode influenciar no resultado. Neste caso, o método foi capaz de encontrar o ótimo

global, mas em caso de funções mais complexas, o uso de meta-heurísticas pode ser recomendado.

Tabela 5.12. Análise de variância para EQM dos FC's (proporções de componente) 2ª iteração

Fonte	Graus de Liberdade	Somaa dos Quadrados (seq.)	Somaa dos Quadrados (ajust.)	Quadrado Médio (ajust.)	F	P
Regressão	4	0,036515	0,036515	0,009129	59,71	0,000
Linear	1	0,017992	0,022026	0,022026	144,07	0,000
*Quadrático γ1*γ2*	1	0,01016	0,009696	0,009696	63,42	0,000
*Cúbico completo γ1*γ2*(-)*	1	0,005998	0,005998	0,005998	39,23	0,000
*Quártico1 completo γ1*γ2*(-)2*	1	0,002366	0,002366	0,002366	15,47	0,004
Erro de Resíduos	8	0,001223	0,001223	0,000153		
Total	12	0,037738				

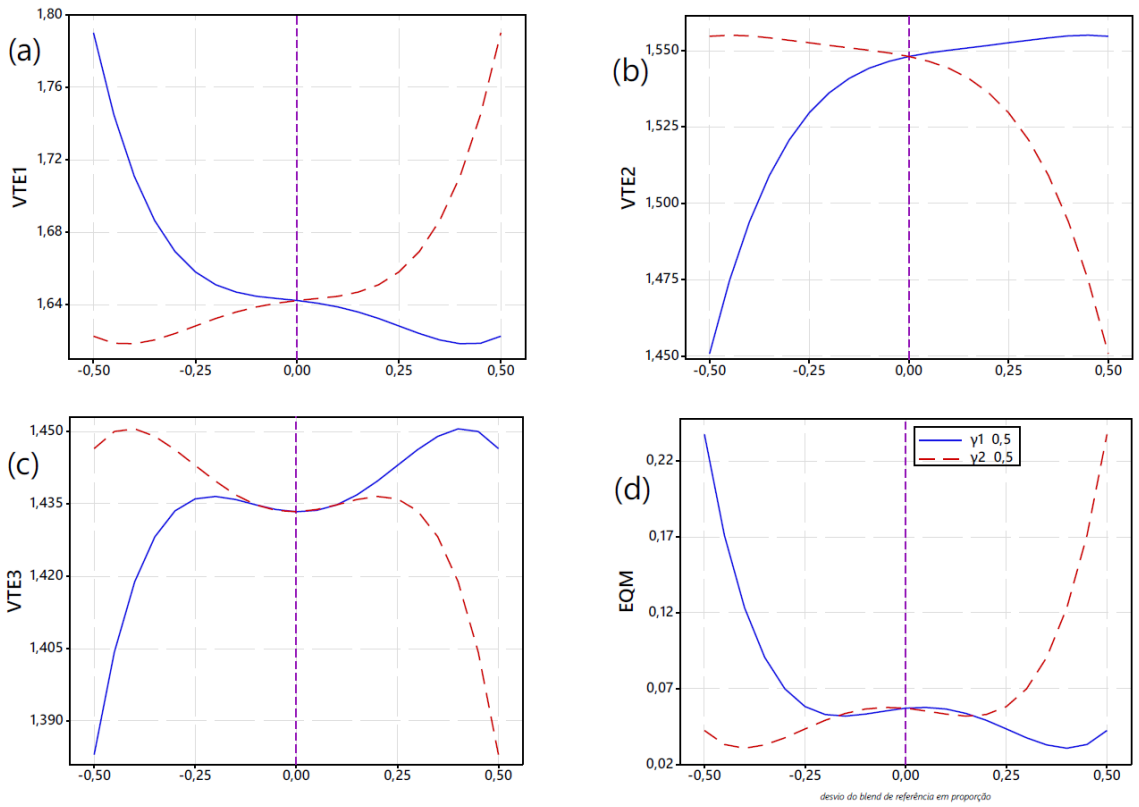


Figura 5.14. Gráficos de traço de resposta Cox para (a) VTE_1, (b) VTE_2, (c) VTE_3 e (d) EQM_{VTE} dos Clusters – 2ª iteração

O resultado da otimização, bem como o comportamento da função objetivo, pode ser visualizado a partir dos gráficos de superfície de resposta e contorno, conforme a Figura 5.15.

Com base nisso, foi possível extrair os escores dos fatores rotacionados com γ otimizado, descritos na Tabela 5.8 (*seção 5.3.3*). As informações de cargas fatoriais, diante do valor γ otimizado, estão descritas na Tabela 5.13, em que foi possível verificar que nenhum dos fatores dos clusters (FC) apresentaram altas cargas de sentido oposto. Tal resultado permitiu inferir que os FC's apresentaram um sentido único de otimização, contemplando o comportamento dos valores de variância dos clusters. Essas informações permitiram também verificar que FC_1 apresentou um maior grau de explicação dos dados referentes ao Clusters 3 e 4, enquanto FC_2 explicou, majoritariamente, o Clusters 2 e FC_3 apresentou explicação dos Clusters 1 e 5.

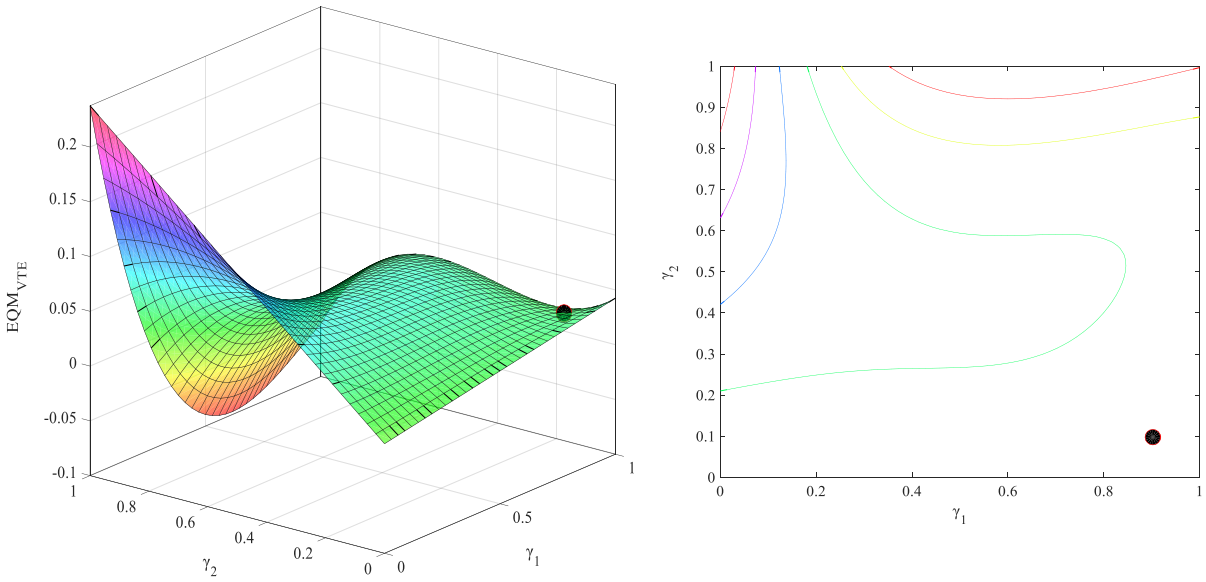


Figura 5.15. Gráfico de (a) superfície de resposta e (b) contorno com ponto ótimo para EQM_{VTE} – 2ª iteração

Tabela 5.13. Cargas fatoriais e comunalidades da rotação *orthomax* otimizada dos clusters

Variável	Rotação γ = 0,9022			Comum
	Fator1	Fator2	Fator3	
Cluster 1	-0,056	0,123	0,968	0,955
Cluster 2	0,181	0,935	0,244	0,966
Cluster 3	0,971	0,153	0,028	0,967
Cluster 4	0,669	0,62	0,062	0,836
Cluster 5	0,442	0,508	0,667	0,899
VTE	1,6228	1,5544	1,4465	4,6237
% Var	0,325	0,311	0,289	0,925

5.3.5 Análise e parametrização ótima pelos escores fatoriais dos clusters

Conforme detalhado na Figura 3.1, seguiu-se para a etapa condicional, em que foi preciso retornar a análise do arranjo multiníveis com os escores de rotação otimizada, após a primeira

iteração. Deste modo, realizou-se novamente a análise do modelo experimental, mas considerando os três vetores de escores dos clusters, **FC** = [FC_1; FC_2; FC_3]. Fazendo uso do WLS, foi possível verificar que o Método de Ligação se mostrou significativo para os três fatores dos clusters, com valores de *p-value* < *0,05*. O Tipo de Análise não se mostrou significativo apenas para o FC_3, mas apresentou valores significativos aos demais, FC_1 e FC_2. Esse comportamento era esperado, visto que, na análise anterior, esse parâmetro experimental não foi significativo para o Cluster 1. Além disso, FC_3 apresentou um alto valor de carga fatorial para o Cluster 1 (Tabela 5.13). Contudo, tal fator não foi removido do modelo para a análise, visto que FC_3 também explicou o Cluster 5. Faz-se importante destacar que os valores FC_1, FC_2 e FC_3 confirmaram a pré-suposição de não rejeitar a hipótese de normalidade dos resíduos, através da estatística de Ryan-Joiner (similar ao teste de normalidade de Shapiro-Wilk), conforme ilustrado nos gráficos da Figura 5.16. Os coeficientes e ajustes de regressão estão descritos na Tabela 5.14. Demais estatísticas disponíveis nas Tabelas C.11 e C.12 do Apêndice C.

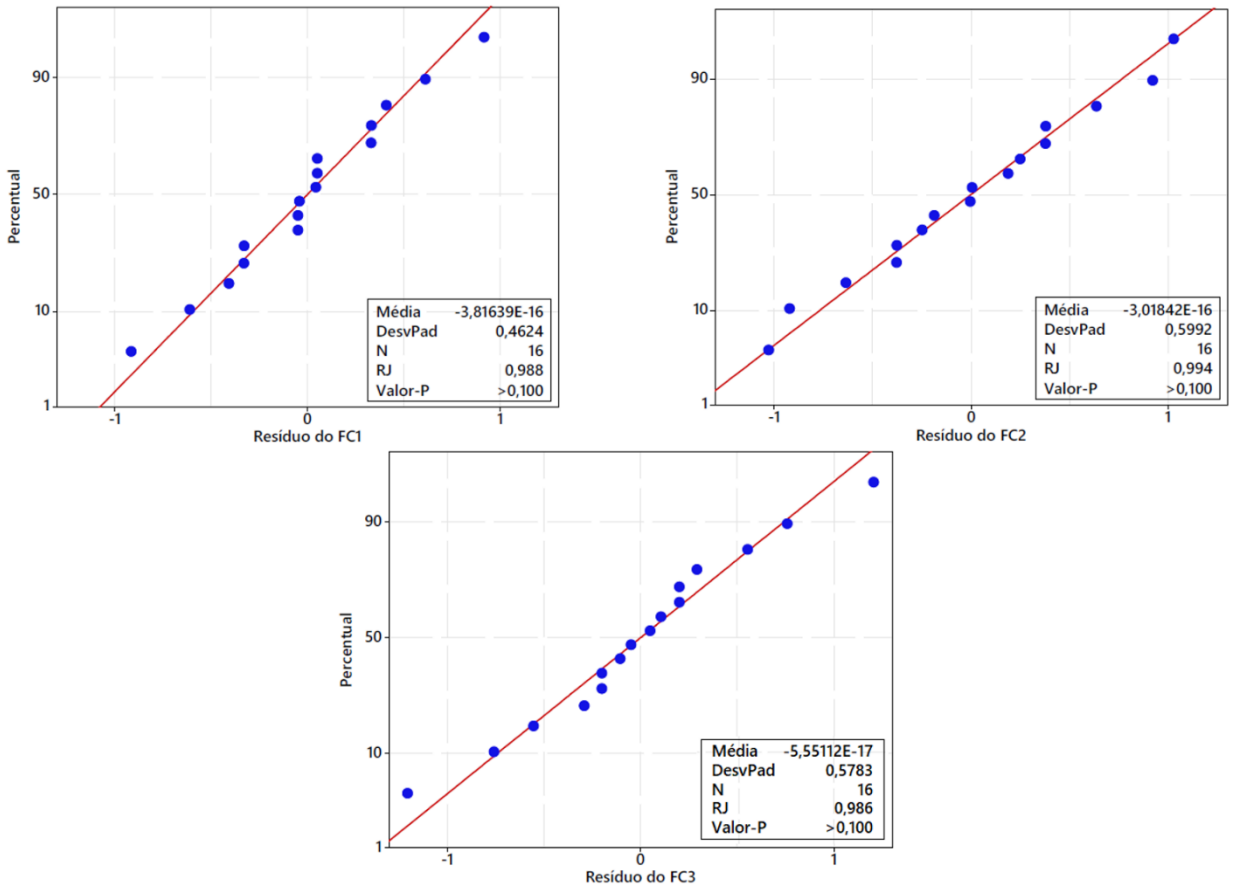


Figura 5.16. Teste Ryan-Joiner de normalidade dos resíduos

Tabela 5.14. Coeficientes de regressão do DOE multiníveis para os escores fatoriais dos Clusters

Termos		Coeficientes		
		FC_1	FC_2	FC_3
Constante		0,00000	0,00000	0,00000
Ligação	*Único*	1,59154	0,85334	0,08917
	Centroide	0,73466	-0,16922	0,54097
	Completa	0,37753	-1,11468	-0,59321
	Média	-0,5219	0,43166	0,60818
	Mediana	-0,15976	-0,71982	-0,31324
	McQuitty	-0,29954	-0,19845	-0,76447
	Ward	-0,97993	0,41774	-1,05476
	k-médias	-0,74259	0,49944	1,48737
Análise	ANOVA	0,32042	0,41586	0,15318
	ANCOVA	-0,32042	-0,41586	-0,15318
R^2		97,94%	96,53%	84,52%
R^2_{adj}		95,58%	92,56%	66,82%
R^2_{pred}		87,12%	75,22%	21,80%

Os gráficos de efeitos principais, ilustrados na Figura 5.17, demonstraram o comportamento dos escores, em que FC_1 apresentou um comportamento significativo entre os métodos de ligação, sendo a abordagem *"*Ward*"* a que apresentou um menor valor para esses escores. Para o tipo de análise, a ANCOVA é o método que apresentou o melhor comportamento, com menor valor de escore fatorial. Um desempenho similar foi visto para FC_3, em que *"*Ward*"* se destacou e, apesar do tipo de análise não ter sido significativo para FC_3, foi possível verificar seu comportamento no gráfico. Para FC_2, o método de ligação que apresentou melhor desempenho foi o "Completa", mantendo a análise ANCOVA como a que apresentou menor variabilidade na formação de clusters em dados de subestação. De modo geral, os métodos de ligação "Único" e "*k-médias*" demonstraram, para essa análise, serem incrementos de alta variabilidade.

A partir dessas análises, foi possível criar as equações de regressão dos coeficientes para **FC**, a fim de encontrar o ponto ótimo que apresentasse uma parametrização adequada para criação e classificação dos clusters de subestação. Os valores dos coeficientes estão descritos na Tabela 5.14. Considerando o sentido de otimização para os três fatores, observou-se que o valor ótimo para os escores de fatores rotacionados foi de **FC*** = [-1,3003; 0,0019; -1,2079] para FC_1, FC_2 e FC_3 respectivamente. Tais valores referem-se à parametrização ótima do *Método de Ligação*, como *"*Ward*"* e o tipo de análise, como "ANCOVA", apresentando valores de variância na formação de clusters iguais a **Y*** = [1949,15; 990,927; 657,148; 1244,18; 1156,98] para o Cluster 1, 2, 3, 4 e 5, respectivamente.

Conhecendo a parametrização ótima para a classificação dos clusters de subestação, foi possível estimar as regiões de confiança a partir de funções elipsoidais.

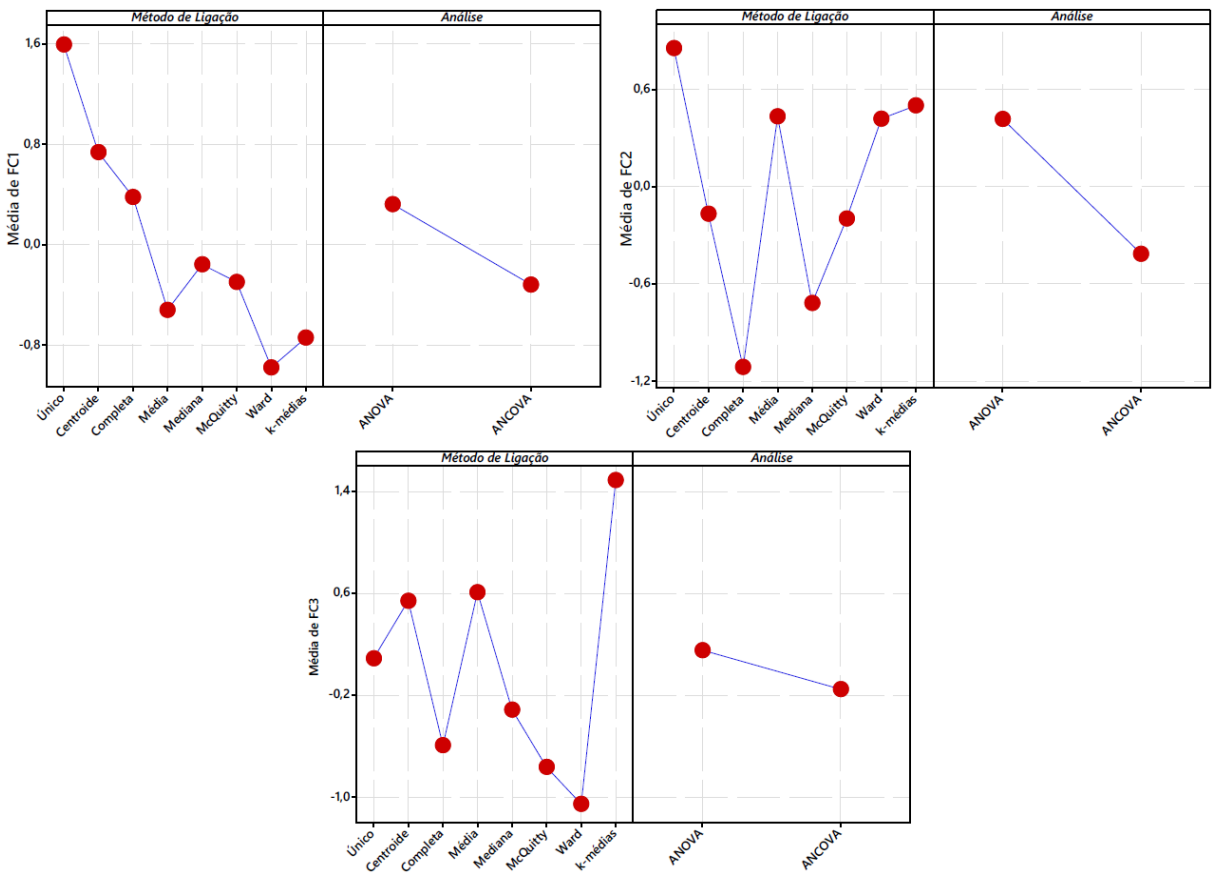


Figura 5.17. Gráfico de efeitos principais para os escores de fator rotacionados dos Clusters

5.4 Elipses de confiança (95%)

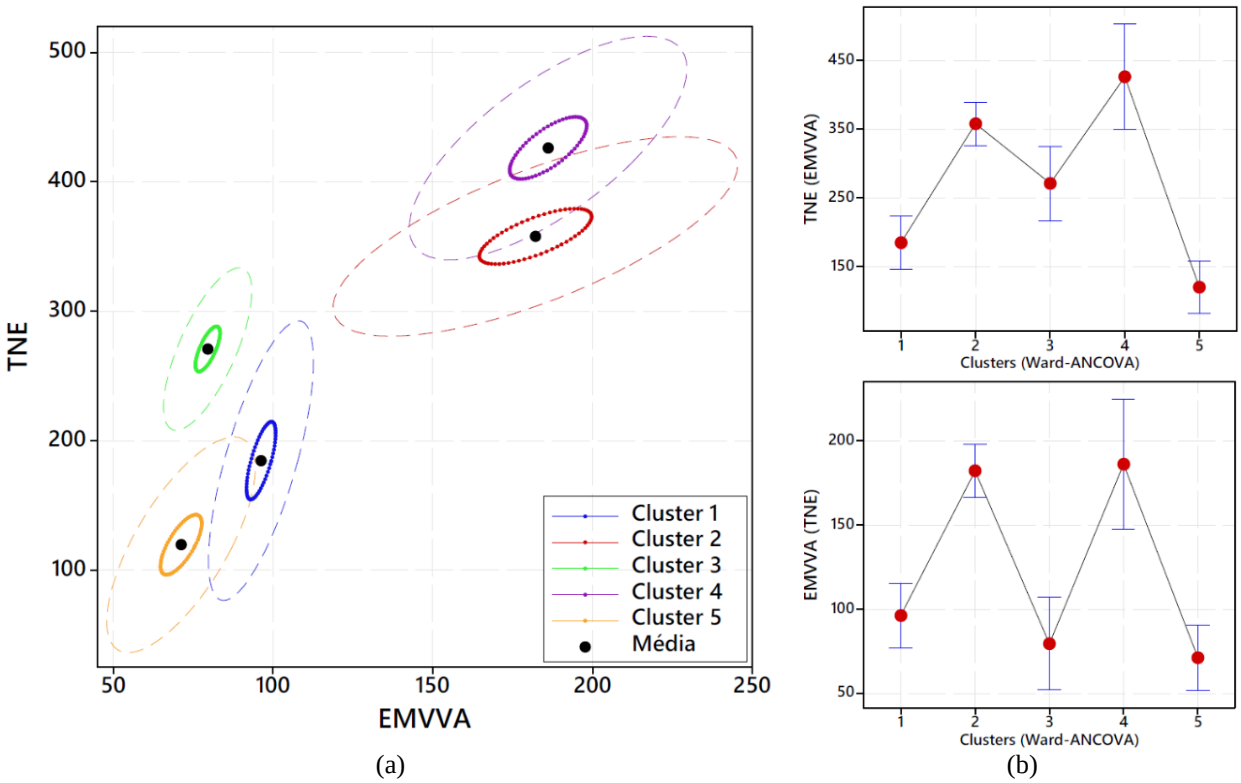
A partir da combinação ótima de parâmetros, teve-se a necessidade de estimar os intervalos de maneira adequada. Tratando-se de um conjunto dados em que as características apresentavam uma estrutura de variância-covariância significativa, as estimativas precisaram ser representadas através de regiões de confiança, assim como detalhado anteriormente, no capítulo 2. Para representar as regiões de maneira adequada, elipses de confiança, para um nível α de 5%, foram modeladas considerando a parametrização ótima e a variável concomitante EMVVA.

Com as análises realizadas na etapa anterior, foi possível definir as informações necessárias para estimar as elipses de confiança dos clusters formados pelos escores de fator rotacionados. Baseado nas formulações matemáticas, as informações de vetor de médias ($\boldsymbol{\mu}$), matriz sigma ($\boldsymbol{\Sigma}$), autovalor ($\boldsymbol{\Lambda}$) e autovetor ($\mathbf{P}$) se fizeram necessárias para estimar cada um dos clusters, diante das técnicas definidas (Ward-ANCOVA). A Tabela 5.15 detalha os valores necessários para essa aplicação.

Tabela 5.15. Vetores e matrizes para estimar as elipses TNE×EMVVA (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 96,24 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 182,14 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 79,537 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 186,13 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 71,203 \end{bmatrix}$
Σ	$\begin{bmatrix} 1949,15 & 211,24 \\ 211,24 & 45,16 \end{bmatrix}$	$\begin{bmatrix} 990,92 & 580,09 \\ 580,09 & 669,883 \end{bmatrix}$	$\begin{bmatrix} 657,14 & 103,11 \\ 103,11 & 31,91 \end{bmatrix}$	$\begin{bmatrix} 1244,18 & 446,2 \\ 446,2 & 315,66 \end{bmatrix}$	$\begin{bmatrix} 1156,97 & 230,24 \\ 230,24 & 90,38 \end{bmatrix}$
Λ	$\begin{bmatrix} 1972,31 \\ 22,0 \end{bmatrix}$	$\begin{bmatrix} 1432,3 \\ 228,51 \end{bmatrix}$	$\begin{bmatrix} 673,72 \\ 15,35 \end{bmatrix}$	$\begin{bmatrix} 1423,85 \\ 136 \end{bmatrix}$	$\begin{bmatrix} 1204,56 \\ 42,8 \end{bmatrix}$
P	$\begin{bmatrix} 0,994 & -0,109 \\ 0,109 & 0,994 \end{bmatrix}$	$\begin{bmatrix} 0,795 & -0,605 \\ 0,605 & 0,795 \end{bmatrix}$	$\begin{bmatrix} 0,987 & -0,158 \\ 0,158 & 0,987 \end{bmatrix}$	$\begin{bmatrix} 0,927 & -0,373 \\ 0,373 & 0,927 \end{bmatrix}$	$\begin{bmatrix} 0,979 & -0,2024 \\ 0,202 & 0,979 \end{bmatrix}$

Diante dos elementos da Tabela 5.15 e dos equacionamentos para regiões de confiança (em que n =17 e p = 2), foi possível gerar todas as informações necessárias para estimar as elipses de confiança e outras análises complementares. Considerando essas análises, foi possível verificar as informações χ^2 = 5,991; t = 2,813; F = 3,682 e $T^2_{crítico}$ = 2,803. Os valores de c para média e dados foram 0,680 e 2,448, respectivamente. A partir desses resultados, foi possível calcular os valores que projetaram as regiões de confiança. A Tabela 5.16 apresenta os pontos equiespaçados de contorno das elipses de confiança (95%). De modo complementar, a Figura 5.18 ilustra o comportamento das elipses, considerando um nível α de 5%, para as médias e para os dados, sendo esta última, representada pelas linhas pontilhadas da figura.



(a) (b)

Figura 5.18. Gráficos: (a) Elipses de confiança e (b) intervalos univariados de confiança (95%)

Tabela 5.16. Pontos equiespaçados de contorno das elipses de confiança (95%)

Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
TNE	EMVVA	TNE	EMVVA	TNE	EMVVA	TNE	EMVVA	TNE	EMVVA
184,282	99,411	351,628	190,326	270,447	82,167	423,168	193,491	118,765	75,558
187,280	99,724	353,703	191,84	272,188	82,433	425,559	194,410	121,076	76,013
190,251	100,002	355,819	193,257	273,917	82,670	427,955	195,247	123,373	76,420
193,166	100,242	357,956	194,564	275,614	82,876	430,332	195,993	125,633	76,775
195,996	100,442	360,092	195,746	277,264	83,049	432,668	196,641	127,833	77,074
198,712	100,600	362,205	196,793	278,851	83,187	434,938	197,184	129,952	77,314
201,288	100,715	364,275	197,693	280,357	83,288	437,121	197,616	131,968	77,494
203,697	100,785	366,281	198,438	281,769	83,352	439,193	197,933	133,861	77,610
205,915	100,810	368,202	199,020	283,072	83,377	441,135	198,133	135,612	77,663
207,921	100,789	370,020	199,434	284,253	83,364	442,927	198,213	137,204	77,651
209,694	100,722	371,717	199,675	285,300	83,313	444,552	198,172	138,620	77,574
211,217	100,611	373,274	199,74	286,204	83,224	445,992	198,011	139,848	77,434
212,474	100,456	374,678	199,630	286,953	83,099	447,234	197,731	140,873	77,231
213,453	100,259	375,914	199,346	287,543	82,938	448,264	197,336	141,687	76,969
214,143	100,022	376,969	198,889	287,965	82,742	449,074	196,828	142,280	76,649
214,539	99,747	377,833	198,265	288,217	82,515	449,654	196,214	142,648	76,274
214,637	99,437	378,497	197,480	288,295	82,258	450,000	195,499	142,786	75,849
214,434	99,095	378,955	196,542	288,20	81,974	450,107	194,690	142,693	75,377
213,933	98,725	379,202	195,460	287,931	81,665	449,974	193,796	142,370	74,863
213,140	98,330	379,236	194,245	287,492	81,336	449,603	192,826	141,820	74,313
212,062	97,914	379,056	192,910	286,886	80,988	448,998	191,788	141,048	73,732
210,710	97,481	378,665	191,466	286,121	80,626	448,164	190,694	140,063	73,126
209,097	97,036	378,065	189,930	285,203	80,253	447,109	189,555	138,874	72,500
207,239	96,583	377,263	188,316	284,142	79,872	445,846	188,382	137,494	71,862
205,156	96,127	376,268	186,640	282,948	79,489	444,385	187,186	135,935	71,217
202,868	95,671	375,088	184,919	281,634	79,106	442,742	185,979	134,213	70,572
200,397	95,222	373,736	183,171	280,212	78,727	440,932	184,774	132,346	69,933
197,769	94,782	372,226	181,412	278,697	78,356	438,975	183,583	130,353	69,306
195,010	94,357	370,571	179,661	277,103	77,997	436,890	182,417	128,253	68,699
192,147	93,951	368,79	177,934	275,448	77,653	434,697	181,289	126,066	68,117
189,208	93,568	366,899	176,250	273,746	77,329	432,418	180,209	123,816	67,565
186,225	93,212	364,918	174,625	272,016	77,026	430,077	179,188	121,525	67,050
183,225	92,886	362,867	173,074	270,274	76,749	427,696	178,236	119,215	66,577
180,239	92,593	360,765	171,615	268,539	76,499	425,300	177,364	116,909	66,149
177,297	92,337	358,634	170,261	266,826	76,279	422,912	176,579	114,631	65,773
174,428	92,120	356,495	169,025	265,154	76,093	420,555	175,890	112,403	65,450
171,661	91,944	354,370	167,921	263,539	75,940	418,255	175,303	110,248	65,185
169,024	91,811	352,279	166,958	261,998	75,824	416,033	174,824	108,186	64,980
166,543	91,722	350,244	166,148	260,545	75,745	413,913	174,458	106,240	64,837
164,242	91,679	348,286	165,497	259,195	75,703	411,914	174,209	104,428	64,758
162,145	91,681	346,422	165,013	257,962	75,700	410,057	174,080	102,767	64,743
160,273	91,728	344,673	164,700	256,857	75,735	408,361	174,070	101,276	64,793
158,644	91,821	343,056	164,561	255,893	75,808	406,842	174,181	99,969	64,907
157,275	91,958	341,586	164,598	255,079	75,919	405,517	174,412	98,858	65,084
156,179	92,137	340,279	164,810	254,422	76,065	404,397	174,760	97,955	65,321
155,368	92,358	339,148	165,196	253,929	76,247	403,494	175,221	97,269	65,618
154,848	92,618	338,203	165,751	253,606	76,461	402,817	175,792	96,807	65,970
154,627	92,913	337,455	166,470	253,455	76,706	402,374	176,466	96,573	66,375
154,705	93,242	336,910	167,345	253,479	76,979	402,167	177,236	96,570	66,828
155,082	93,601	336,575	168,368	253,676	77,278	402,200	178,096	96,798	67,325
155,754	93,987	336,452	169,529	254,044	77,599	402,473	179,036	97,254	67,860
156,715	94,394	336,543	170,816	254,581	77,940	402,981	180,047	97,934	68,429
157,955	94,821	336,847	172,217	255,281	78,296	403,721	181,118	98,831	69,026
159,461	95,261	337,361	173,716	256,137	78,665	404,685	182,240	99,937	69,644
161,219	95,711	338,079	175,300	257,139	79,043	405,863	183,401	101,239	70,278
163,210	96,167	338,995	176,952	258,279	79,426	407,243	184,589	102,726	70,921
165,416	96,623	340,100	178,656	259,545	79,810	408,812	185,792	104,382	71,567
167,814	97,076	341,382	180,395	260,924	80,191	410,555	186,999	106,190	72,209
170,379	97,520	342,828	182,152	262,402	80,565	412,452	188,198	108,133	72,842
173,088	97,952	344,424	183,908	263,964	80,930	414,487	189,375	110,192	73,458
175,911	98,366	346,155	185,647	265,596	81,280	416,638	190,521	112,345	74,051
178,821	98,759	348,002	187,351	267,281	81,613	418,883	191,622	114,571	74,616
181,790	99,127	349,948	189,003	269,001	81,925	421,202	192,669	116,848	75,147
184,282	99,411	351,628	190,326	270,447	82,167	423,168	193,491	118,765	75,558

As regiões de confiança foram estabelecidas pelas elipses das médias, em que foi possível verificar se os clusters apresentaram diferenças significativas, considerando a variável principal (TNE) e sua respectiva variável concomitante, determinada para a análise (EMVVA). Além disso, foi possível perceber que o uso do parâmetro ANCOVA promoveu a redução da variabilidade, criando clusters com regiões de confiança mais estreitas e precisas, possibilitando uma melhor separação dos grupos de qualidade. Com as informações das elipses das médias, se fez necessário analisar os limites bilaterais de Bonferroni, calculados com as informações disponíveis na Tabela 5.15. A Tabela 5.17 apresenta os limites superiores (LS) e os limites inferiores (LI) calculados a partir das elipses das médias de cada cluster e, complementarmente, tem-se os valores calculados para os limites bilaterais de Bonferroni. Para ilustrar graficamente esse comportamento, a Figura 5.19 apresenta as elipses de confiança junto aos limites calculados. Com base nas informações descritas e ilustradas, foi possível verificar que as elipses respeitaram os limites dos intervalos bilaterais de Bonferroni, indicando uma análise adequada, em conjunto às demais realizadas.

Tabela 5.17. Limites dos intervalos bilaterais de Bonferroni

Cluster	**TNE**				
	Média	*LI*	*LS*	$LI_{Bonferroni}$	$LS_{Bonferroni}$
1	184,6296	154,6266	214,6365	154,5081	214,7511
2	357,8501	336,4520	379,2363	336,3730	379,3272
3	270,8696	253,4553	288,2954	253,3798	288,3594
4	426,1293	402,1672	450,1067	402,0637	450,1949
5	119,6648	96,5700	142,7859	96,4579	142,8717
Cluster	**EMVVA**				
1	96,2418	91,6786	100,8097	91,6569	100,8267
2	182,1477	164,5611	199,7403	164,4892	199,8062
3	79,5371	75,6999	83,3771	75,6825	83,3917
4	186,1369	174,0702	198,2128	174,0151	198,2587
5	71,2033	64,7432	77,6626	64,7171	77,6895

LI: Limite Inferior
LS: Limite Superior

Em função dos agrupamentos formados por meio da parametrização ótima, foi possível identificar diferentes categorias para o evento afundamento de tensão, causado pela VTCD, uma vez que as elipses de confiança não apresentaram intervalos sobrepostos (diferente do que fora apresentado pelos intervalos univariados, na Figura 5.9 e Figura 5.10). Na Figura 5.20 é possível verificar que os Clusters 1 e 5 apresentaram menor número de eventos de afundamento de tensão, enquanto os Clusters 2 e 4 foram classificados como os que apresentaram maior número de eventos. As subestações pertencentes aos clusters com maior incidência de eventos (Itarana, Juncado, Linhares A, Pinheiros, Santa Tereza, Suíça e João Neiva) apresentaram níveis

elevados de TNE e EMVVA, justificados pelos seus afundamentos de tensão e zonas de vulnerabilidade existentes em suas localizações.

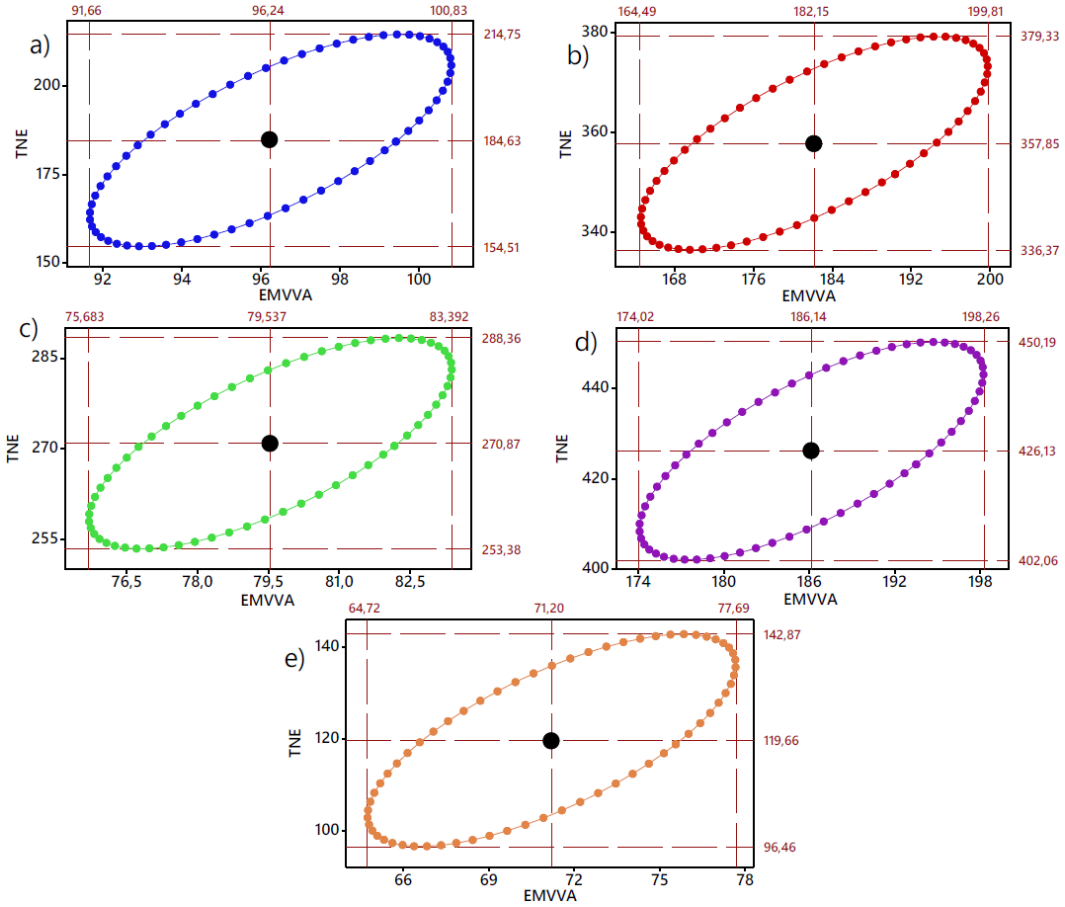


Figura 5.19. Intervalos bilaterais de Bonferroni sobre as elipses de confiança dos Clusters (a) 1, (b) 2, (c) 3, (d) 4 e (e) 5

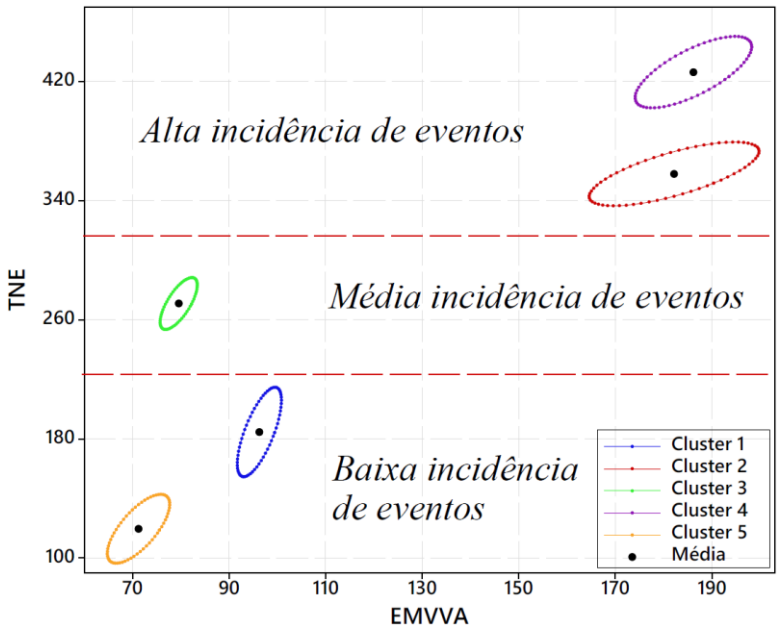


Figura 5.20. Elipses de confiança (95%) com discriminação de incidência de eventos de afundamento de tensão

O Cluster 3 possui uma quantidade intermediária de eventos, diferindo dos demais, indicando uma incidência média de afundamento de tensão (mas com baixos valores de EMVVA), representando as subestações Jaguaré e Linhares C, com valores de TNE de 249,661 e 292,078, respectivamente. As classificações de baixa ocorrência em eventos de afundamento de tensão estão concentradas no Clusters 1 (representados pelas subestações Aracruz, Baixo Guandu, Barra Sahy e Ecoporanga) e no Cluster 5 (representados pelas subestações Montanha, Nova Venécia, Paulista e São Francisco), sendo estas, as subestações que apresentaram uma melhor qualidade na distribuição de energia elétrica.

O uso de escores de fator com rotação otimizada para a avaliação da variável principal para o problema (TNE), em conjunto com o auxílio da variável concomitante (EMVVA), permitiram a criação de regiões de confiança com menor variabilidade, aprimorando o poder discriminatório das elipses de confiança. Tal resultado promoveu uma interpretação mais precisa e confiável para a classificação dos clusters subestações. Além disso, foi possível distinguir adequadamente os clusters, criando três categorias distintas para a avaliação de eventos para os clusters, mediante a qualidade na distribuição de energia elétrica.

5.4.1 Análise de variáveis relacionadas

Com a definição dos parâmetros, pôde-se realizar o mesmo procedimento para analisar outras variáveis que apresentaram impacto nos eventos de afundamento de tensão. Diante de um critério técnico estabelecido em estudos de TNE [61,70], selecionou-se as características que apresentaram maior influência: o número de eventos na média tensão (NEMV), a taxa de falha na média tensão (MVFR) e a área de vulnerabilidade equivalente na alta tensão (EVAHV). Além das características citadas, destacou-se o número de eventos monitorados (MNE) que, mesmo não sendo estatisticamente significativo, foi o indicador utilizado por agências regulatórias que não estabeleceram TNE como métrica principal.

Após definir as variáveis a serem consideradas, repetiu-se o procedimento para projeção das elipses de confiança, mantendo o mesmo nível de α, além de criar os intervalos univariados para as distintas combinações de variáveis concomitantes. As análises prévias de vetores e matrizes, necessárias para estimar as regiões de confiança dessa etapa, estão detalhadas no Apêndice C. É importante ressaltar que todas as relações foram baseadas na parametrização ótima: “Ward-ANCOVA”.

Com base nessas informações, calculou-se os intervalos e elipses de confiança para as combinações definidas. A Figura 5.21(a) apresenta a relação TNE × NEMV, em que é possível

verificar o comportamento de dependência linear entre as variáveis (discutida anteriormente), com correlação de *Pearson* = 1,000. Esse comportamento pode ser verificado nos resultados de autovalor descritos na Tabela C.14 (Apêndice C). Deste modo, devido a dependência linear entre as variáveis, tem-se um comportamento extremamente estreito para as elipses. Nessa relação, a discriminação dos Clusters 1 e 5 para os demais, ainda foi visível em ambos eixos. Contudo, não foi possível realizar a discriminação entre o Cluster 2, Cluster 3 e Cluster 5. Tal resultado, implica na necessidade de investimentos para aprimorar a QEE nos níveis desses agrupamentos, visto a relação da média tensão e a quantidade de afundamentos de tensão [61].

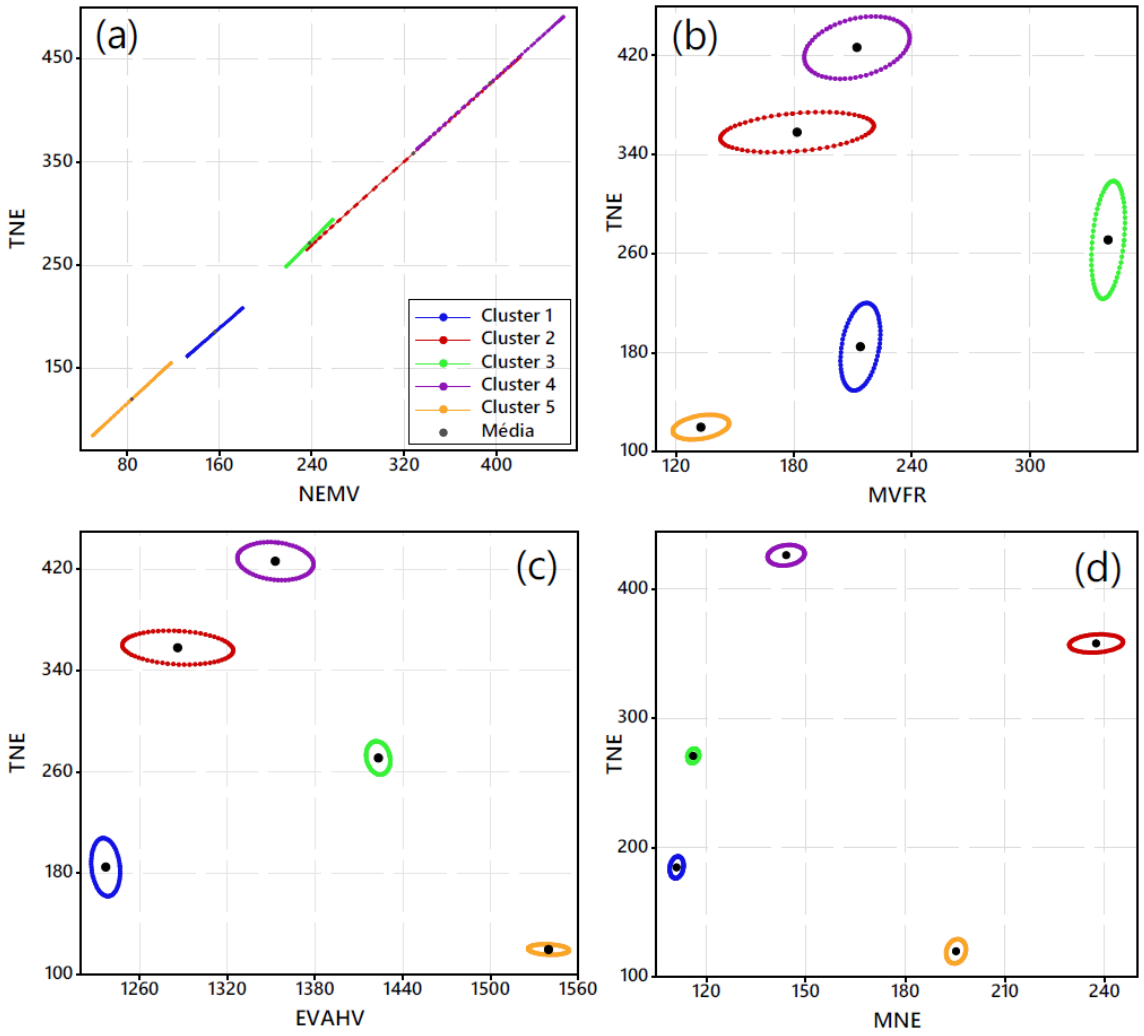


Figura 5.21. Elipses de confiança (95%) com as concomitantes (a)NEMV, (b) MVFR, (c) EVAHV e (d) MNE

Para a relação TNE × MVFR, os resultados apresentados na Figura 5.21(b) confirmam a classificação e descriminação inferida na análise principal do estudo, em que os Clusters 1 e 5 apresentaram baixa incidência. Além disso, tal resultado se destacou da relação com NEMV,

uma vez que foi possível discriminar os clusters de alta incidência, no Cluster 3 (com quantidade de eventos intermediários). Contudo, ao se analisar o comportamento da variável MVFR, referente a taxa de falha, notou-se que investimentos relacionados às subestações do Cluster 3 deveriam ser priorizados, apresentando uma taxa de falha no valor de 340 na média tensão.

A terceira relação desta análise se deu pelo número total de eventos pela área de vulnerabilidade equivalente na alta tensão (TNE × EVAHV). Ao avaliar o comportamento das elipses de confiança na Figura 5.21(c), foi visível a alta discriminação entre os clusters, principalmente em relação ao TNE. A precisão das elipses nessa dimensão, permitiu separar cada cluster em uma determinada categoria. Ao analisar a dimensão da EVAHV, a discriminação foi um pouco menor, mas ainda significativa. Diferente da EMVVA (variável concomitante principal), a EVAHV referiu-se a área de vulnerabilidade na alta tensão, em que os clusters analisados apresentaram médias de quilometragem próximas (entre 1.236,58km e 1.539,65km). Deste modo, investimentos para melhoria dessa variável podem não trazer mudanças muito significativas, em um primeiro momento.

Por fim, ao analisar a relação TNE × MNE (Figura 5.21(d)), foi possível verificar que todos os clusters apresentaram um alto nível de discriminação das elipses de confiança, em que nenhuma região ficou sobreposta nas dimensões avaliadas. Nesse aspecto, destacou-se o Cluster 1, que além de apresentar uma baixa incidência para TNE, também apresentou um baixo valor para MNE, com média de $\boldsymbol{\mu}_{\mathbf{TNE\times MNE}}$ = [184,63; 111]. Em situação oposta, pôde-se destacar o Cluster 2, o qual se apresentou em um quadrante com alta incidência de TNE e MNE, sendo um conjunto de subestações que necessitam de investimentos para aprimorar a qualidade de energia elétrica.

Baseado nos resultados, tem-se que a parametrização ótima permitiu identificar os clusters com maior precisão, diferindo, estatisticamente, os conjuntos de subestações com melhor qualidade, das subestações que precisam de um aprimoramento técnico. Assim, oportunidades promissoras de melhorias nas subestações devem ser analisadas pelos gestores de agências reguladoras, além de permitir a criação de políticas para gestão das concessionárias, com base técnica.

5.4.2 Influência da variável concomitante nas regiões de confiança

Com finalidade de verificar a influência das variáveis concomitantes dentro das funções elipsoidais de confiança para as médias e para os dados, gerou-se essas regiões de confiança,

estabelecendo a relação de TNE com EMVVA, MVFR, EVAHV e MNE. A Figura 5.22 apresenta tal relação, em pares, das características citadas, mas desconsiderando a influência da variável concomitante. Em outras palavras, estimou-se as regiões de confiança a partir dos resultados pela estratégia ANOVA.

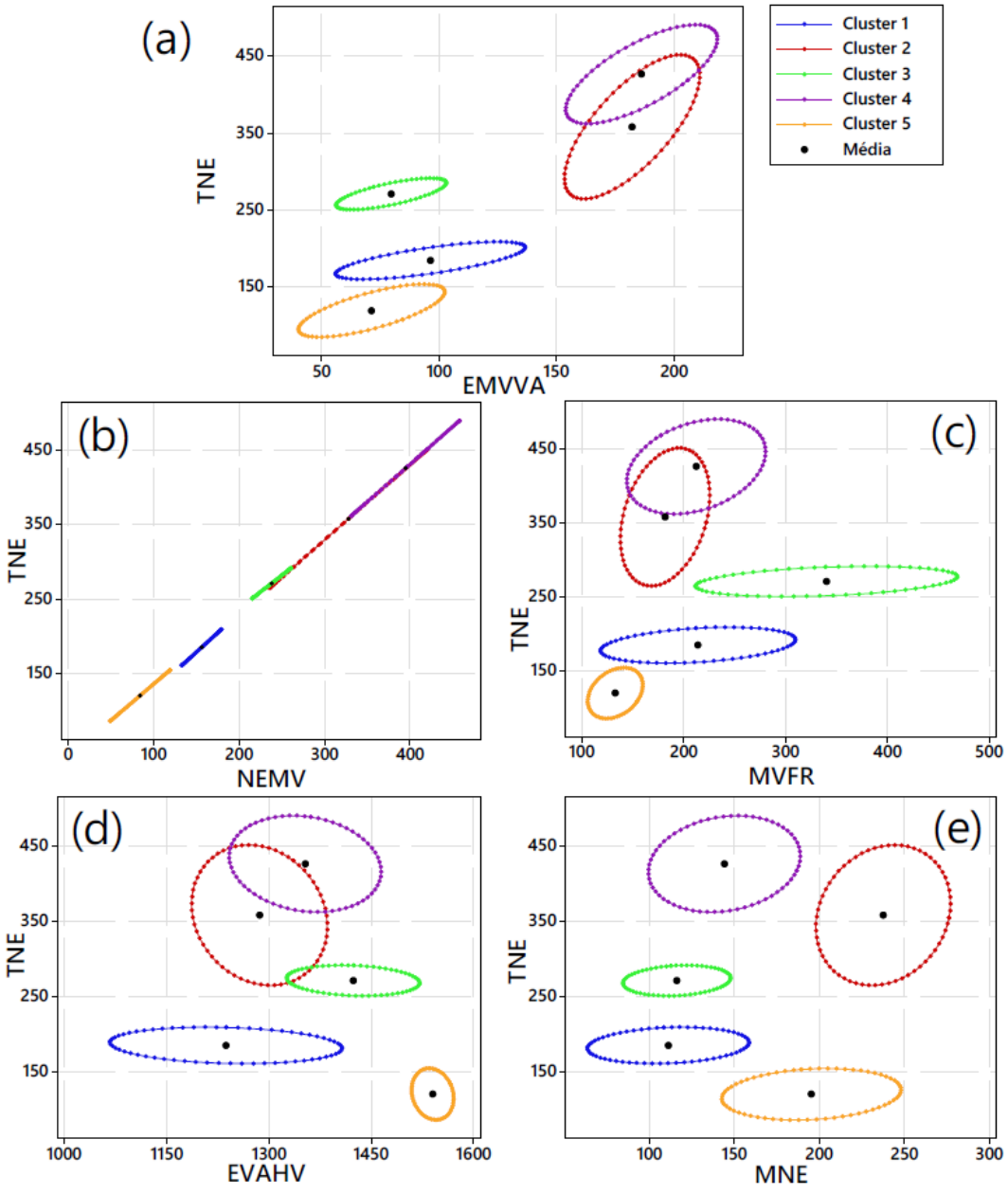


Figura 5.22. Elipses de confiança (95%) estimadas sem a influência da covariável para (a)NEMV, (b) MVFR, (c) EVAHV e (d) MNE

A projeções das elipses de confiança apresentaram o impacto da ausência do ajuste criado pela variável concomitante para todos os casos, em que as regiões encontradas na Figura 5.22 apresentaram maior variabilidade, comparadas às da Figura 5.21. Considerar a covariável no momento de realizar as análises, proporcionou um melhor ajuste, gerando um modelo mais

preciso e com melhor discriminação ao terminar as regiões, além de proporcionar um melhor entendimento no que se refere ao comportamento e previsão dos resultados. Tal resultando também explanou a importância de se considerar a estrutura de variância-covariância em conjuntos de natureza multivariada.

Em destaque, tem-se na Figura 5.22(a) a relação principal do estudo, em que foi possível verificar o impacto da EMVVA ao estimar os intervalos para TNE dos agrupamentos que apresentaram menor incidência de eventos. Tal resultado, proporcionou a estimação de elipses não sobrepostas para os Clusters 1, 3 e 5. No que se refere aos agrupamentos que apresentaram alta incidência (Clusters 2 e 4), a alta variabilidade atribuída foi visível, gerando regiões que não puderam ser discriminadas individualmente, mostrando a importância de se identificar e considerar as variáveis secundárias para estimar a influência da variável principal. Contudo, a discriminação entre os agrupamentos de baixa e alta incidência foram visíveis, mostrando que, mesmo sem a influência da variável concomitante, a FA promoveu resultados satisfatórios para a análise de cluster, auxiliando a discriminação e agrupamentos de dados multivariados.

A relação e impacto das variáveis concomitantes pode ser verificada também nos intervalos de confiança para cada uma das respostas de interesse. A Figura 5.23 apresenta a diferença da influência das variáveis secundárias, considerando TNE como a variável concomitante (a variável NEMV, como já indicado anteriormente, apresentou dependência linear com TNE, logo, não foi utilizada nessa análise, pois o resultado similar já foi apresentado). Nessa análise, a minimização da variabilidade foi explicita, em que, o uso da covariável permitiu criar intervalos de confiança mais estreitos e, consequentemente, mais precisos.

5.4.3 Influência dos escores de fator rotacionados nas regiões de confiança

Para verificar a influência do uso de escores de fator sob rotação ortogonal na classificação e variabilidade dos clusters, gerou-se duas análises comparativas: a primeira referente a formulação de clusters sem o uso de técnicas exploratórias (aplicando o método de ligação com os dados originais) e a segunda referente a análise utilizando uma técnica multivariada exploratória alternativa, que é a análise de componentes principais (PCA), sem a presença de rotações (ortogonais ou oblíquas). É importante ressaltar que, em ambas análises, considerou-se o método de ligação “Ward” e a variável concomitante EMVVA, a fim de manter os demais parâmetros constantes e favorecer a comparação justa aos resultados encontrados pelo método proposto. O nível de confiança foi mantido em 95%.

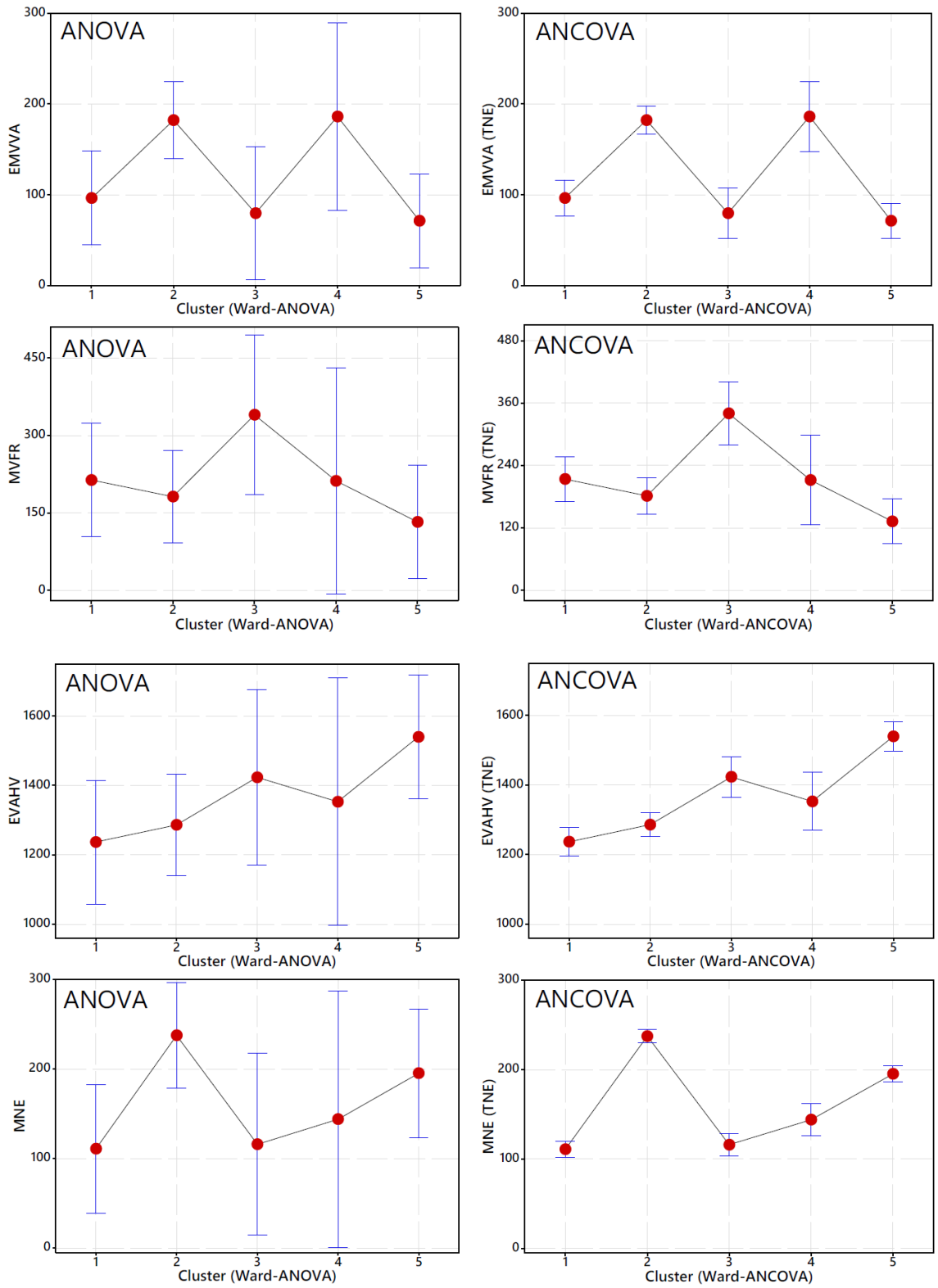


Figura 5.23. Intervalos de confiança (95%) das variáveis relacionadas pela ANOVA e ANCOVA

Para análise sem técnicas exploratórias, gerou-se os agrupamentos com método de ligação "Ward", em que foi possível verificar os clusters formados na Tabela 5.18. De modo análogo, a Figura 5.24(a) apresenta o dendrograma formado para esta análise. No segundo comparativo, inicialmente, houve a necessidade de extrair os componentes principais do estudo. Sete

componentes foram extraídos, descritos na Tabela 5.18, os quais explicam 92,9% dos dados. Em seguida, realizou-se os agrupamentos com o método "Ward". As associações do mesmo estão disponíveis na Tabela 5.18 e representados graficamente na Figura 5.24(b).

Tabela 5.18. Escores dos componentes principais e associações para os dados brutos e PCA

Subestação	Escores dos componentes principais							Associações - Ward	
	PC_1	PC_2	PC_3	PC_4	PC5	PC_6	PC_7	Originais	PCA
Aracruz	-2,1258	-3,0295	-1,1436	1,2341	1,4553	0,9374	0,2898	1	1
Baixo Guandu	2,1344	-4,0708	-0,9106	0,8273	-0,7671	1,5175	0,1239	2	2
Barra do Sahy	2,6416	-3,6254	2,1978	-1,4331	0,9585	-0,5965	2,0523	3	2
Ecoporanga	-6,3751	0,4462	-2,0497	2,4059	0,2684	0,2254	0,6330	1	3
Itarana	1,9658	1,3641	0,4091	-0,4139	-1,3886	1,9166	-0,0754	2	4
Jaguaré	3,9443	-1,3587	-2,1183	0,8330	0,4732	0,3218	-1,6898	2	2
João Neiva	3,0613	0,3670	1,9197	2,8620	-1,3553	-2,6129	-0,7266	4	5
Juncado	-1,9910	3,8291	2,7458	-0,0227	2,2567	-0,2821	-0,2311	1	4
Linhares A	4,7762	4,5737	0,5620	1,3817	0,2367	0,9477	1,6115	5	5
Linhares C	4,6491	-1,0068	0,7524	-1,1178	1,8212	0,2892	-1,4779	2	2
Montanha	-4,6366	1,5484	-1,5868	-1,4168	-0,6783	0,0156	-0,2074	1	3
Nova Venécia	3,4862	0,2495	-3,2729	-1,5456	-1,1786	-1,9733	1,0542	3	2
Paulista	-3,7507	-1,1968	-0,3081	-0,7236	0,6726	-1,8018	-0,6928	1	1
Pinheiros	-0,1929	2,1042	-0,3349	-0,6258	-0,9239	0,5353	-0,4698	4	4
Santa Tereza	-1,6582	0,8496	1,9484	-1,4832	-1,4875	0,6885	-0,9029	1	4
São Francisco	-1,3482	1,7131	-2,1303	-0,8242	0,9264	-0,2642	0,3534	1	3
Suíça	-4,5808	-2,7568	3,3202	0,0625	-1,2899	0,1358	0,3556	1	1

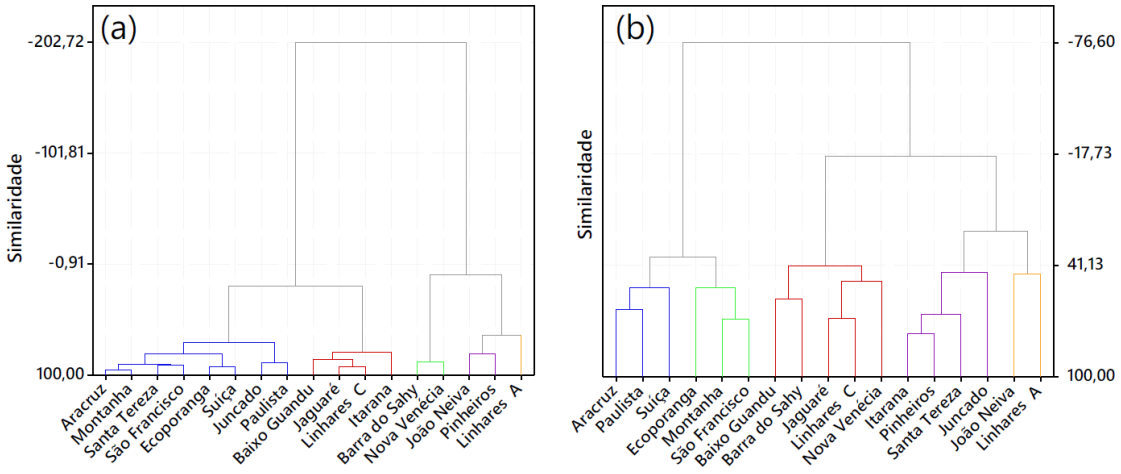
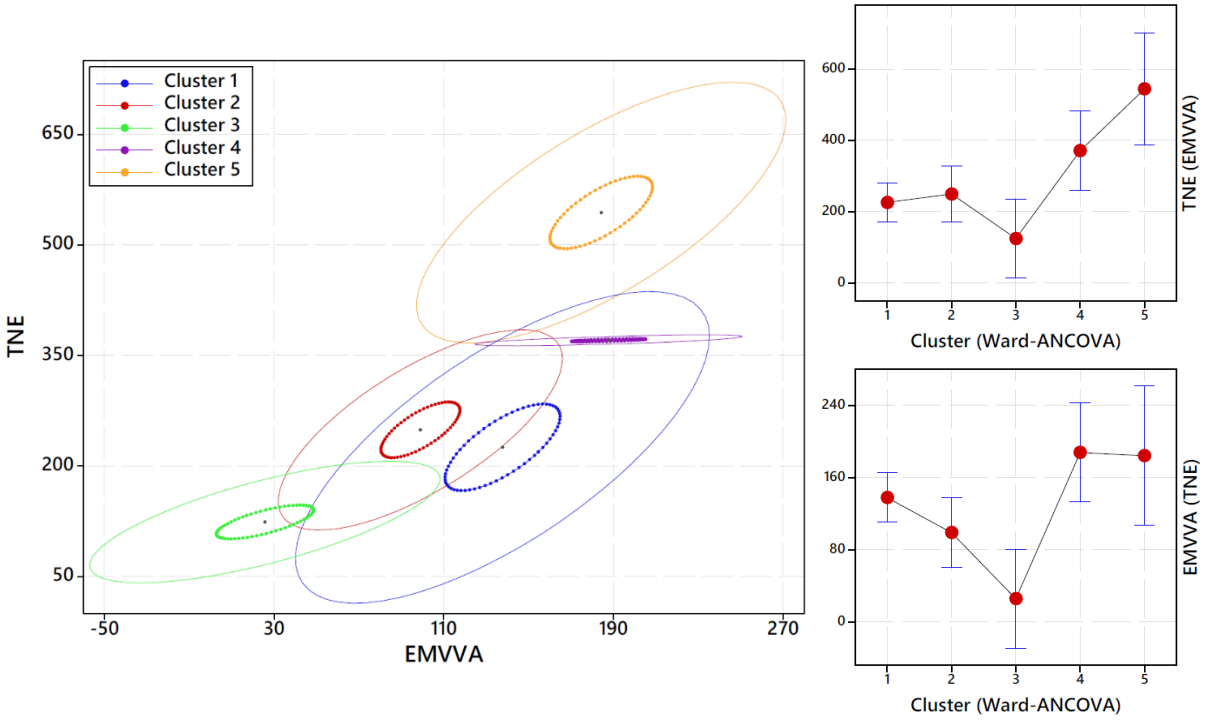


Figura 5.24. Dendrogramas pelo método Ward para (a) os dados brutos e (b) para PCA

Para estimar as regiões de confiança de ambos os casos, necessitou-se dos dados de vetor de $\boldsymbol{\mu}$ e $\boldsymbol{\Lambda}$, além das matrizes $\boldsymbol{\Sigma}$ e $\mathbf{P}$, conforme as etapas anteriores. Tais informações estão disponíveis nas Tabelas C.23 e C.24 (Apêndice C). Ao estimar as regiões de confiança para a primeira situação (Figura 5.25), foi possível verificar que os agrupamentos das subestações se diferiram e, mesmo que as elipses de confiança (representadas pelas elipses pontilhadas) não estejam sobrepostas, o intervalo de confiança dos Clusters 1 e 2 não permitu uma discriminação adequada para a variável principal, TNE. Esse comportamento foi visível para os intervalos de

confiança univariados, ilustrados à direita do gráfico. As linhas continuas de maior extensão, ilustram as elipses dos dados.



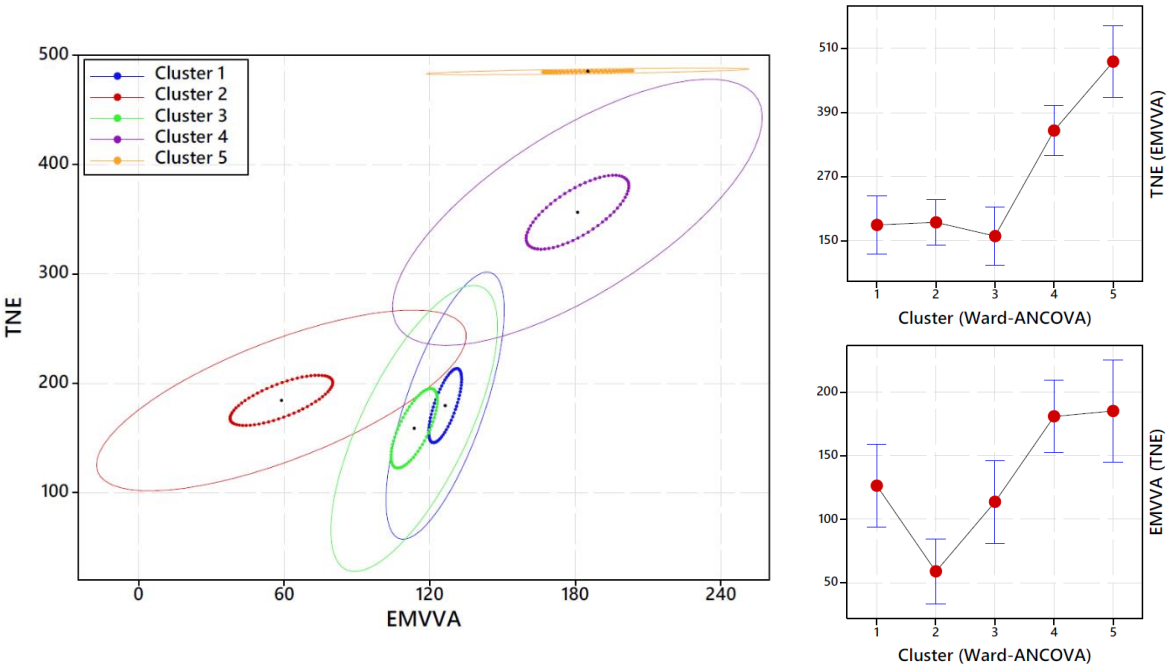
(a) (b)

Figura 5.25. Gráficos da análise dos dados brutos: (a) Elipses de confiança e (b) intervalos de confiança (95%)

A aplicação com a técnica exploratória PCA demonstrou um novo comportamento para os grupos, conforme visualizado anteriormente. As regiões de confiança foram calculadas de modo similar a etapa anterior, em que foi possível verificar, na Figura 5.26, que as elipses de confiança (para as médias) apresentaram sobreposições em ambas dimensões, nos agrupamentos dos Cluster 1 e 3. Tal resultado foi prejudicial em uma etapa classificatória, visto que não foi possível concluir que a média desses grupos são estatisticamente diferentes, promovendo um confundimento estatístico ao analisar esses dados. Esse comportamento foi também visível nos intervalos de confiança univariados, na mesma figura.

A partir dessas variações, foi possível verificar o comportamento de diferentes meios de classificação, atribuindo a variável concomitante. Com finalidade de mostrar o comportamento da extração pelos dados originais e pelo PCA, sem considerar a covariável, a Figura 5.27 apresenta as elipses de confiança para os dados originais e PCA com a análise comumente utilizada para esse fim, a ANOVA (similar ao estudo de [61]). Com base nos gráficos, foi possível verificar que, em ambas situações, existem agrupamentos que apresentaram confundimento para classificação entre os Clusters 1 e 2 (Figura 5.27(a)) e entre os Clusters 1 e 3 (Figura 5.27(b)). Essas sobreposições impedem a discriminação dos agrupamentos,

prejudicando a tomada de decisão que implicam a estas variáveis. Um comportamento diferente ocorreu com o uso da FA (com ou sem a ANCOVA), como visto em análises anteriores (Figura 5.20 e Figura 5.22(a), respectivamente). Assim, o uso da estratégia FA auxiliou na discriminação dos dados, uma vez que a rotação ortogonal ajudou a separar os dados de maneira prévia na análise exploratória (como também foi detalhado anteriormente), favorecendo a interpretação e explicação de variáveis latentes. As informações necessárias para estimar as regiões de confiança nessa análise, estão detalhadas nas Tabela C.25 e C.26 do Apêndice C.



(a) (b)

Figura 5.26. Gráficos da análise por PCA: (a) Elipses de confiança e (b) intervalos de confiança (95%)

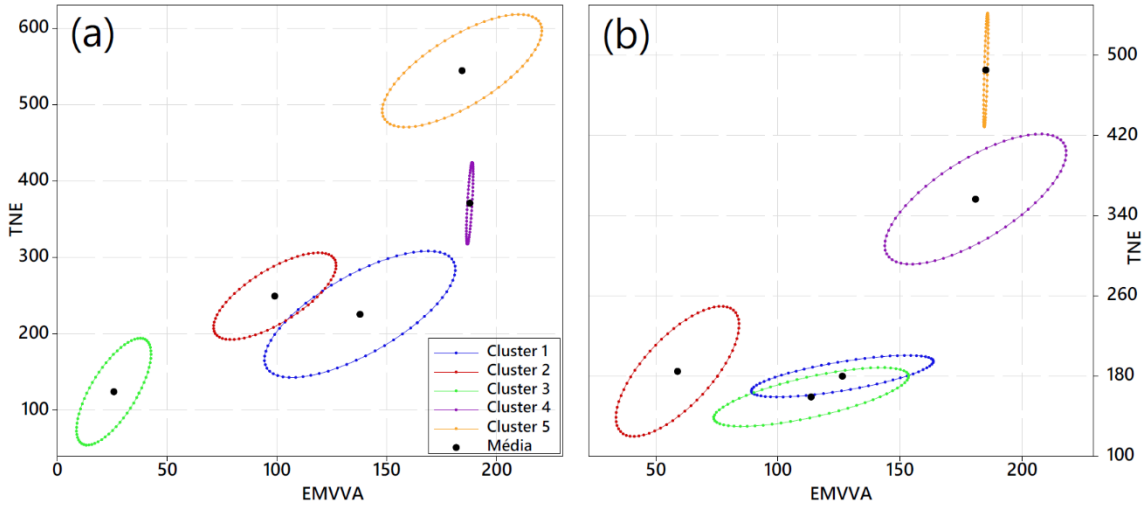


Figura 5.27. Elipses de confiança (95%) sem a influência da covariável para (a) dados brutos e (b) PCA

5.5 Confirmação do parâmetro ótimo em cenários com perturbações

Através das análises anteriores, verificou-se e confirmou-se a influência das variáveis concomitantes e das técnicas exploratórias. A proposta deste trabalho concluiu que o método de ligação "Ward" apresentou o melhor comportamento para auxiliar na discriminação nesse determinado conjunto de dados. Contudo, isso não infere que esse método de ligação seja sempre a melhor opção, pois isso depende também da estrutura dos dados a serem investigados. Além disso, tem-se que os métodos de ligação são sensíveis a pequenos erros, podendo apresentar inversões de agrupamentos na presença de valores discrepantes [5].

Partindo dessa premissa, a última etapa dessa proposta se referiu à confirmação da robustez do método, aplicando pequenas perturbações ao conjunto de dados sobre os índices de qualidade de energia, analisando o grau de consistência dos métodos de ligações a partir da análise de concordância por atributos, conforme descrito no capítulo 3.

5.5.1 Réplicas com pequenas perturbações

Considerando o conjunto original de dados, criou-se quatro réplicas com pequenas perturbações (faixa de 1%), como sugerido por Johnson e Wichern [5], buscando gerar diferentes cenários que pudessem representar dados coletados em diferentes períodos. É importante ressaltar que as réplicas com perturbações foram geradas randomicamente, além de ser mantido um grau significativo da estrutura de variância-covariância dos dados. Devido à grande extensão dessas informações, as réplicas formadas estão disponíveis no Apêndice C.

5.5.2 Otimização γ e formação de agrupamentos das réplicas

Todas as etapas referentes à otimização da rotação *orthomax* foram realizadas para as quatro réplicas, gerando um grau de rotação igual a $\boldsymbol{\gamma^*}$ = [1; 1; 1; 1] (**EQM*** = [14,2755; 14,1058; 13;8918; 14,1794]) para a réplica R1, R2, R3 e R4, respectivamente. A Figura 5.28 ilustra o comportamento das funções com seu respectivo ponto ótimo. Os coeficientes de regressão utilizados estão descritos na Tabela 5.19. Demais informações como arranjo experimental, estatísticas e escores rotacionados estão descritas no Apêndice C.

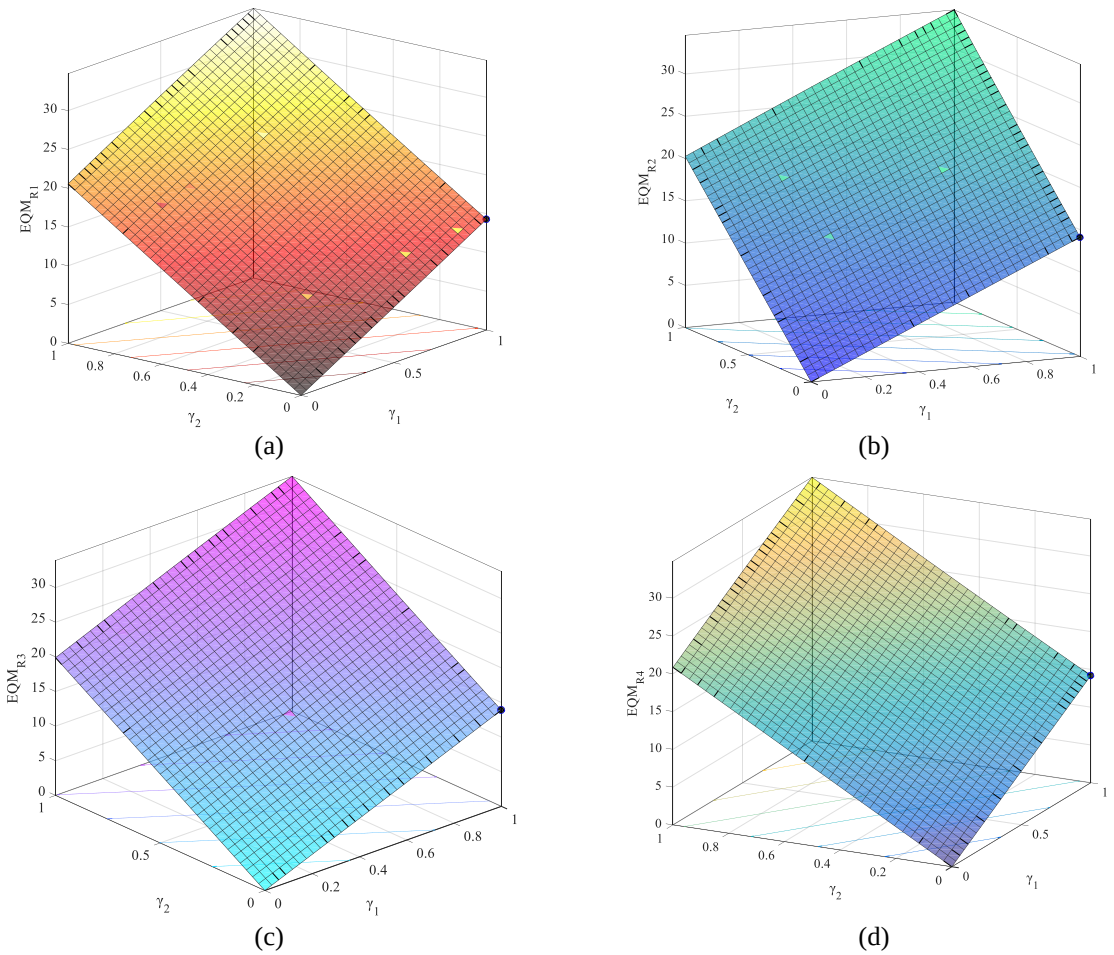


Figura 5.28. Gráfico de superfície e contorno para (a) R1, (b) R2, (c) R3 e (d) R4

Tabela 5.19. Coeficientes de regressão do EQM para as réplicas

Termo	**EQM$_{R1}$**		**EQM$_{R2}$**		**EQM$_{R3}$**		**EQM$_{R4}$**	
	Coef	P	*Coef*	P	*Coef*	P	*Coef*	P
γ_1	14,2755	*	14,1058	*	13,8918	*	14,1794	*
γ_2	20,5565	*	20,2517	*	19,8814	*	20,8334	*
$\gamma_1 \times \gamma_2$	-0,0727	0,000	0,2216	0,000	0,14251	0,000	-0,14284	0,000
$\gamma_1 \times \gamma_2 \times (-)$	-0,3882	0,000	-0,3751	0,000	-0,4041	0,000	-0,48	0,000
$\gamma_1 \times \gamma_2 \times (-)^2$	-0,0682	0,002	-0,0980	0,013	-0,1243	0,002	-0,116	0,001
R^2	100,00%		100,00%		100,00%		100,00%	
R^2_{adj}	100,00%		100,00%		100,00%		100,00%	
R^2_{pred}	100,00%		100,00%		100,00%		100,00%	

Conhecendo a quantidade de clusters a serem gerados (a partir da regra de Sturges), aplicou-se todos os métodos de ligação apresentados anteriormente, agrupando-os em cinco diferentes clusters, a fim de realizar um comparativo do comportamento sob pequenas perturbações. As associações de cada método, em cada réplica, estão descritas na Tabela 5.20 e na Tabela 5.21.

Tabela 5.20. Associações dos clusters formados pelos métodos de ligação nas réplicas (Parte I)

Amostras	Único				Centroide				Completa				Média			
	Replica				Replica				Replica				Replica			
	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4
Aracruz	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Baixo Guandu	1	1	2	2	1	1	1	1	1	2	1	1	1	1	1	1
Barra do Sahy	2	2	3	3	2	2	2	2	2	2	2	2	2	2	2	2
Ecoporanga	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Itarana	1	1	2	2	1	1	1	1	3	1	3	3	1	1	1	1
Jaguaré	1	1	2	2	1	1	1	1	4	3	4	4	3	3	3	1
João Neiva	3	3	4	4	3	3	3	3	5	4	5	5	4	4	4	3
Juncado	4	4	2	2	1	1	1	1	3	1	3	3	1	1	1	4
Linhares A	1	1	2	2	4	4	4	4	3	5	3	3	5	5	5	5
Linhares C	5	5	5	5	1	5	1	5	4	3	4	4	3	3	3	1
Montanha	1	1	2	2	1	1	1	1	4	3	4	4	1	1	1	1
Nova Venécia	1	1	2	2	1	1	1	1	4	3	4	4	1	1	1	1
Paulista	1	1	2	2	1	1	1	1	4	3	4	4	1	1	1	1
Pinheiros	1	1	2	2	1	1	1	1	3	1	3	3	1	1	1	1
Santa Tereza	1	1	2	2	1	1	1	1	3	1	3	3	1	1	1	1
São Francisco	1	1	2	2	1	1	1	1	4	3	4	4	1	1	1	1
Suíça	1	1	2	2	5	1	5	1	2	2	2	2	1	1	1	1

Tabela 5.21. Associações dos clusters formados pelos métodos de ligação nas réplicas (Parte II)

Amostras	Mediana				McQuitty				Ward				k-médias			
	Replica				Replica				Replica				Replica			
	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4
Aracruz	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Baixo Guandu	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2
Barra do Sahy	1	1	1	1	2	1	1	2	1	1	1	1	3	3	3	3
Ecoporanga	1	1	1	1	1	1	1	1	1	1	1	1	4	4	4	4
Itarana	1	1	1	1	1	2	2	1	2	2	2	2	5	5	5	5
Jaguaré	1	1	1	1	1	3	3	1	3	3	3	3	5	5	5	5
João Neiva	2	2	2	2	3	4	4	3	4	4	4	4	2	5	2	2
Juncado	3	3	3	3	4	2	2	4	2	2	2	2	5	5	5	5
Linhares A	4	4	4	4	4	5	5	5	2	2	2	2	5	5	5	5
Linhares C	5	5	1	5	2	3	3	1	3	3	3	3	3	3	3	3
Montanha	1	1	1	1	1	3	3	1	5	5	5	5	4	4	4	4
Nova Venécia	1	1	1	1	1	3	3	1	5	5	5	5	5	5	5	5
Paulista	1	1	1	1	1	3	3	1	5	5	5	5	5	5	5	5
Pinheiros	1	1	1	1	1	2	2	1	2	2	2	2	5	5	5	5
Santa Tereza	1	1	1	1	1	2	2	1	2	2	2	2	5	5	5	5
São Francisco	1	1	1	1	1	3	3	1	5	5	5	5	4	4	4	4
Suíça	1	1	5	1	5	2	2	1	2	2	2	2	5	5	5	5

5.5.3 Análise de concordância dos métodos de ligação

A partir das associações, gerou-se um planejamento para o estudo de concordância, definindo as 17 subestações em análise (“número de amostras”), 8 métodos de ligação (como os “avaliadores”) e 4 cenários com diferentes perturbações (como as “réplicas”). Tal planejamento gerou um total de 544 combinações de agrupamento para variáveis contendo as características de qualidade na distribuição de energia elétrica, disponíveis na Tabela 5.20 e na Tabela 5.21.

Com base na aplicação dessa estratégia, considerando um IC de 95%, foi possível verificar o grau de consistência e precisão com que os métodos mantiveram o agrupamento das subestações, por meio da estatística Kappa de Fleiss e do Coeficiente de Concordância de Kendall. Avaliando inicialmente a concordância dentro dos avaliadores (ou repetibilidade), foi possível verificar, por meio da Tabela 5.22, que o único método que apresentou 100% de consistência para todos os clusters foi o método de "Ward" (Figura 5.29). Este método de ligação não mostrou inversão na formação dos clusters em nenhum dos quatro cenários com perturbação. O nível de concordância para o "Ward" pode ser validado por meio da Tabela 5.23 e Tabela 5.24, em que todos os clusters (e a também a avaliação geral) apresentaram índice Kappa e Kendall igual a 1, indicando um nível de concordância excelente segundo os critérios da AIAG [119] (Tabela 2.7).

Tabela 5.22. Concordância de avaliação dentro dos métodos de ligação

Avaliador	Nº de Inspecionados	Nº de Correspondências	%	IC de 95%
Centroide	17	15	88,24	(63,56; 98,54)
Completa	17	4	23,53	(6,81; 49,90)
k-médias	17	16	94,12	(71,31; 99,85)
McQuitty	17	3	17,65	(3,80; 43,43)
Média	17	13	76,47	(50,10; 93,19)
Mediana	17	15	88,24	(63,56; 98,54)
Único	17	3	17,65	(3,80; 43,43)
Ward	17	17	100	(83,84; 100,00)

Nº de Concordâncias: O avaliador concorda com os ensaios

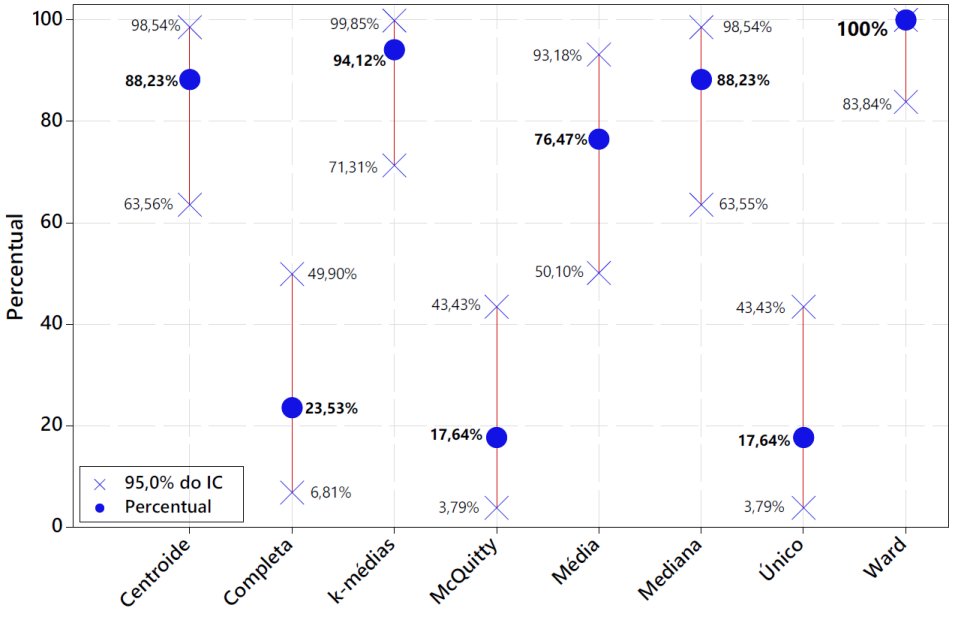


Figura 5.29. Grau de concordância para os métodos de ligação nos cenários com perturbações

Tabela 5.23. Resultados para estatísticas Kappa de Fleiss dentro dos avaliadores

Método	Cluster	Kappa	EP_{Kappa}	Z	P	Método	Cluster	Kappa	EP_{Kappa}	Z	P
Centroide	1	0,7821	0,099	7,898	0,000	*Média*	1	0,7809	0,099	7,887	0,000
	2	1,0000	0,099	10,100	0,000		2	1,0000	0,099	10,100	0,000
	3	1,0000	0,099	10,100	0,000		3	0,5223	0,099	5,274	0,000
	4	1,0000	0,099	10,100	0,000		4	0,4688	0,099	4,734	0,000
	5	0,2917	0,099	2,946	0,002		5	1,0000	0,099	10,100	0,000
	Global	0,8046	0,061	13,210	0,000		*Global*	0,7441	0,061	12,191	0,000
Completa	1	0,5723	0,099	5,780	0,000	*Mediana*	1	0,8365	0,099	8,449	0,000
	2	0,8719	0,099	8,806	0,000		2	1,0000	0,099	10,100	0,000
	3	0,2422	0,099	2,446	0,007		3	1,0000	0,099	10,100	0,000
	4	0,4887	0,099	4,936	0,000		4	1,0000	0,099	10,100	0,000
	5	0,4688	0,099	4,734	0,000		5	0,4688	0,099	4,734	0,000
	Global	0,4949	0,054	9,092	0,000		*Global*	0,8535	0,061	14,012	0,000
k-médias	1	1,0000	0,099	10,100	0,000	*Único*	1	0,1252	0,099	1,264	0,103
	2	0,8408	0,099	8,491	0,000		2	-0,0794	0,099	-0,802	0,789
	3	1,0000	0,099	10,100	0,000		3	0,2917	0,099	2,946	0,002
	4	1,0000	0,099	10,100	0,000		4	0,2917	0,099	2,946	0,002
	5	0,9407	0,099	9,501	0,000		5	1,0000	0,099	10,100	0,000
	Global	0,9544	0,058	16,485	0,000		*Global*	0,1538	0,066	2,313	0,010
McQuitty	1	0,1736	0,099	1,753	0,040	*Ward*	1	1,0000	0,099	10,100	0,000
	2	0,1441	0,099	1,455	0,073		2	1,0000	0,099	10,100	0,000
	3	0,1605	0,099	1,621	0,053		3	1,0000	0,099	10,100	0,000
	4	0,2085	0,099	2,105	0,018		4	1,0000	0,099	10,100	0,000
	5	0,4688	0,099	4,734	0,000		5	1,0000	0,099	10,100	0,000
	Global	0,1910	0,058	3,292	0,001		*Global*	1,0000	0,055	18,164	0,000

Tabela 5.24. Coeficiente de concordância de Kendall dentro dos avaliadores

Método	Coef.	Qui-quadrado	DF	P
Centroide	0,7788	49,8407	16	0,0000
Completa	0,8834	56,5369	16	0,0000
k-médias	0,9406	60,2003	16	0,0000
McQuitty	0,5842	37,3878	16	0,0018
Média	0,8474	54,2305	16	0,0000
Mediana	0,8341	53,3805	16	0,0000
Único	0,8537	54,6381	16	0,0000
Ward	1,0000	64,0000	16	0,0000

O método não hierárquico (*k-médias*) apresentou o segundo melhor comportamento, com 94,12% de concordância, com intervalo de confiança entre 71,31% e 99,85%. Avaliando a estatística Kappa, verificou-se que o método "*k-médias*" apresentou concordância geral de 95,43%, obtendo-se uma inversão entre os clusters 2 e 5 para a subestação João Neiva (com número de eventos original de afundamento de tensão igual a 426). A partir do Coeficiente de Concordância de Kendall, verificou-se que "*k-médias*" apresentou um valor igual a 0,9406, representando um ótimo nível de concordância para os critérios estabelecidos.

Os resultados para os métodos de ligação "Centroide" e "Mediana" apresentaram níveis de concordância aceitáveis, com valores de avalição geral para a estatística Kappa iguais a 0,8046 e 0,8535, respectivamente. Os coeficientes de concordância de Kendall para esses métodos apresentaram valores de 0,7788 e 0,8341, respectivamente. Para Kendall, verificou-se

também que a abordagem “Completa” apresentou coeficiente igual a 0,8834, mas com baixos valores gerais de Kappa (igual a 0,4949). Os demais métodos de ligação apresentaram concordâncias baixas, sendo não recomendadas [119].

5.5.4 Elipses de confiança (95%) para as réplicas

Para demonstrar o comportamento do método nos diferentes cenários, estimou-se as regiões de confiança para cada réplica, a fim de verificar as elipses de confiança para os diferentes cenários de afundamentos de tensão. A Figura 5.30 mostra as regiões de confiança, com nível α de 5%, para quatro cenários diferentes. Nesses gráficos foi possível verificar que, apesar de sutis deslocamentos no vetor de médias, o método se manteve estável na presença de perturbações, resultando em elipses de confiança precisas e consistentes, como na aplicação com os dados originais. Tal resultado confirmou a parametrização ótima encontrada, utilizando o método “Ward” para esse conjunto de dados, que possibilitou a estimativa robusta de agrupamentos de subestações, fornecendo uma divisão adequada e precisa para avaliar a distribuição da qualidade de energia em subestações baseadas em afundamentos de tensão.

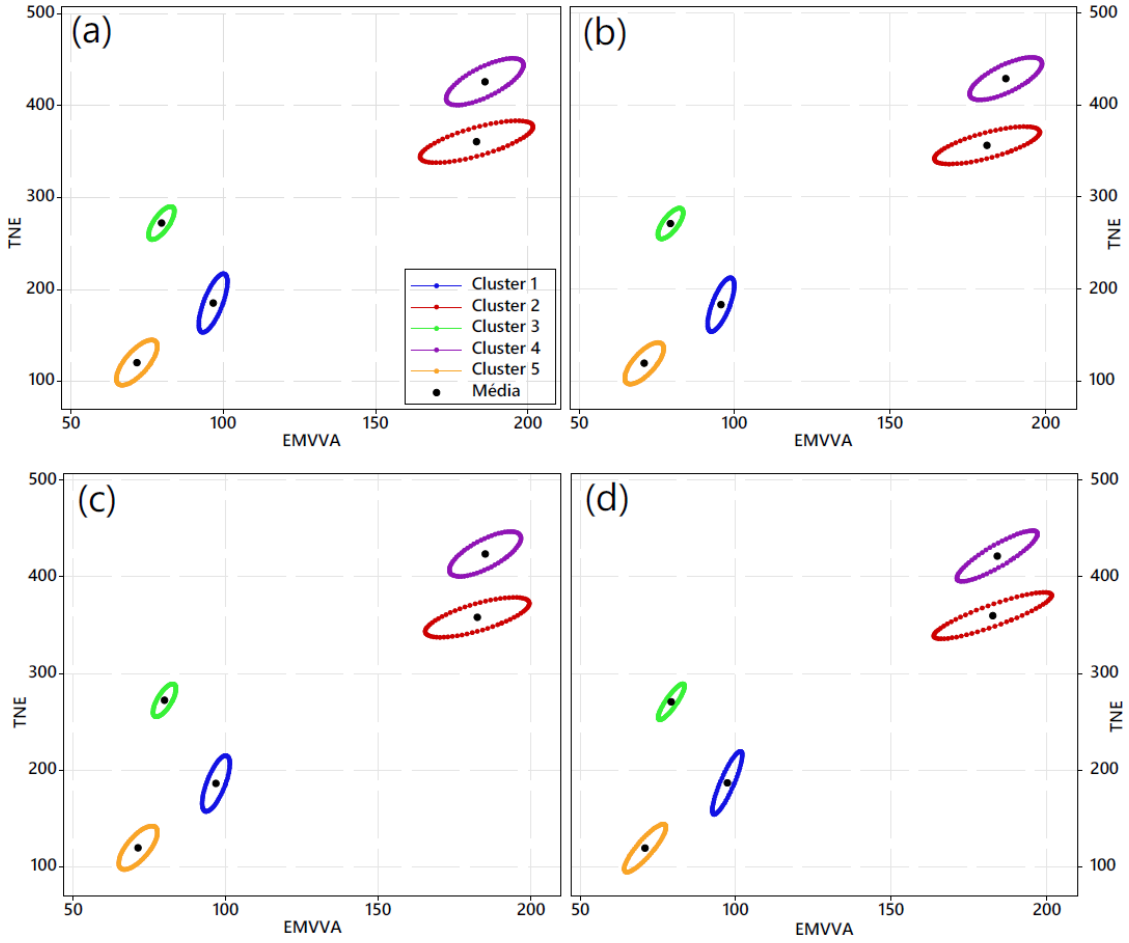


Figura 5.30. Elipses de confiança (95%) para (a) R1, (b) R2, (c) R3 e (d) R4

É importante destacar que as elipses de confiança para R1 e R2 apresentaram um comportamento muito similar às elipses de confiança para os dados originais, enquanto as elipses para R3 e R4 apresentaram uma distinção mais acentuada. R3 proporcionou um nível de correlação sutilmente menor, com valor de *Pearson* igual a 0,708, comparada a R1 e R2 (com *Pearson* de 0,711 e 0,715, respectivamente). Pelo comportamento de R4, verificou-se uma estrutura elipsoidal com maior inclinação e estreitamento positivo, justificado pelo nível de correlação, gerado aleatoriamente para estudo, com *Pearson* igual a 0,871.

5.6 Considerações finais

Os resultados apresentados e discutidos nesse capítulo descrevem a aplicação do método de aprimoramento do poder discriminatório de funções elipsoidais, auxiliadas por escores de fatores sob rotação ortogonal, para os dados de indicadores de qualidade na distribuição de energia elétrica de subestações situadas no estado do Espírito Santo. Considerando que os dados são adequados para aplicações multivariadas, partiu-se para a estratégia exploratória de FA. Com essa aplicação, uma calibração da rotação se fez necessária para aprimorar a explicação das cargas fatoriais e atender o princípio de parcimônia. Nessa etapa, foi possível verificar o comportamento das cargas de fator em todas combinações de rotação γ, geradas pelo arranjo de misturas. A estratégia EQM proporcionou aglutinar as VTE's em uma única função, proporcionando um valor ótimo de rotação. Essa etapa foi repetida na segunda iteração, com um delineamento distinto, proporcionando um novo valor γ. Essa abordagem contribui com uma discussão da literatura quanto a escolha do método de rotação, detalhada ao longo do trabalho. Na parametrização dos métodos de ligação e tipo de análise, um modelo de experimentos fatoriais multiníveis foi utilizado para encontrar o método de menor variância na construção de agrupamentos. O método "Ward" e análise ANCOVA apresentaram menor variabilidade nos clusters, assim, criou-se regiões de confiança bidimensionais a partir de funções de elipses, podendo classificar as subestações em diferentes categoriais de qualidade de energia. Por fim, confirmou-se a adequação das técnicas, inferindo a robustez e precisão da parametrização ótima, utilizando o índice Kappa e o coeficiente de Kendall, mostrando-se uma estratégia adequada para avaliar os métodos de ligação, sendo essa uma discussão existente na literatura.

6. CONCLUSÃO

A crescente procura pelo aperfeiçoamento de técnicas que favorecem a interpretação de dados correlacionados, faz das estratégias multivariadas uma opção valiosa com o advento da computação moderna. Um conjunto de dados, formado por diversos vetores de informação, pode apresentar covariâncias que influenciam de maneira direta no comportamento da variável de interesse, dificultando a formação adequada de grupos e a discriminação dos mesmos. Assim, o presente trabalho foi desenvolvido com finalidade de propor uma nova abordagem para a análise e interpretação de dados com estrutura de variância-covariância significativa, visando aprimorar a formação de agrupamentos e a discriminação de funções elipsoidais ao estimar regiões de confiança. Para atingir esse objetivo, diferentes técnicas estatísticas e matemáticas foram empregadas, como DOE, FA, análise de cluster entre outras. Considerando os trabalhos disponíveis na literatura, destacados no desenvolvimento desse trabalho, constatou-se oportunidades de contribuições em diferentes segmentos, no que se refere à calibração e validação de métodos, discriminação e análise voltados a conjunto de dados correlacionados, destacando a relevância dessa pesquisa. Deste modo, foi proposto o método de aprimoramento do poder discriminatório de funções elipsoidais auxiliadas por escores de fator sob rotação ortogonal, o qual contempla contribuições para definir o nível de rotação das cargas fatoriais, favorece a tomada de decisão dos métodos de ligação mesclados ao tipo de análise e também na estimação de elipses de confiança mais precisas e não sobrepostas. Além disso, foi apresentado um método para analisar a consistência e robustez de métodos de ligação, utilizando a análise de concordância por atributos baseado na discussão inserida por Johnson e Wichern [5]. Motivado pela necessidade desse tipo de análise no setor elétrico, o método proposto foi aplicado em índices de QEE de subestações de uma concessionária de distribuição localizada no sudeste brasileiro. Todas etapas do método foram aplicadas com êxito, permitindo acompanhar todo comportamento em dados reais, como verificado nas discussões apresentadas nos capítulos anteriores. Pode-se concluir que o método proposto promoveu resultados adequados e satisfatórios nos quesitos teóricos e práticos, sendo devidamente confirmado nos comparativos realizado nas últimas sessões do capítulo 5.

Deste modo, considerando os objetivos específicos do método proposto, detalhados no capítulo 1 deste estudo, pode-se concluir as seguintes considerações:

- Com base na necessidade de encontrar uma rotação ótima para as cargas fatoriais, o método de otimização do nível γ se mostrou eficiente para definir a rotação que apresente valores balanceados de VTE, a partir da minimização do EQM. A abordagem

foi utilizada em dois momentos da aplicação, apresentando valor γ igual a 1 na primeira iteração (equivalente à rotação *varimax*) e um valor γ de 0,9022 para a segunda iteração. Confirmando, assim, as premissas encontradas na literatura, de que o nível de rotação ideal pode variar de acordo com o conjunto analisado. O algoritmo SQP se mostrou um método de busca eficiente, mas que pode ser substituído por alguma meta-heurística, dependendo da complexidade da função modelada. Os valores para EQM_{VTE} foram de 14,1167 e 0,0308 para a primeira e segunda iteração, respectivamente. Com esses resultados, foi possível extrair escores de fator com rotação otimizada, gerando eixos independentes, além de reduzir em 53,34% a dimensionalidade dos dados;

- A partir dos escores com rotação ótima inseridos para a formulação do DOE multiníveis (para os métodos de ligação e tipo de análise), foi possível verificar o comportamento da variabilidade na formação dos clusters pelos efeitos principais. Contudo, devido sua estrutura de variância-covariância significativa, foi necessário utilizar técnicas multivariadas para sua análise (referindo-se à segunda iteração do método de otimização γ, comentado no item anterior). Com os novos escores rotacionados, foi possível analisar adequadamente a influência dos parâmetros, em que o método de ligação "Ward" promoveu menor variabilidade na formação de clusters para FC_1 (que detêm a explicação dos Clusters 3 e 4) e FC_3 (que representa os Clusters 1 e 5). Ao avaliar FC_2 (que explica apenas o Cluster 2), verifica-se que o mesmo apresenta uma variabilidade mais elevada para o método "Ward", em que, nesse fator, o método "Completa" apresentou melhores valores de variabilidade. Em relação ao tipo de análise, todos os clusters (representados pelos fatores) apresentaram baixa variabilidade utilizando o método ANCOVA. A parametrização ótima desse DOE multiníveis indicou a combinação "Ward – ANCOVA", a qual considerou a variável TNE como principal e a EMVVA como concomitante.
- Conhecendo a parametrização ótima, foi possível calcular e aplicar as informações para estimar as regiões de confiança, através das elipses, considerando 95% de confiança. A partir dos resultados, pode-se verificar que a parametrização ótima proporcionou uma separabilidade adequada entre os grupos de subestações, criando elipses não sobrepostas, estreitas e, consequentemente, mais precisas. Essa condição proporcionou uma discriminação mais eficiente, visto que é possível separar os clusters em três diferentes categorias, alta, média e baixa incidência de eventos de afundamento de tensão. Essa discriminação visa favorecer órgãos regulatórios como a ANEEL e também podendo ser atribuídas a diferentes problemas de classificação e discriminação em

dados multivariados. As elipses proporcionaram uma visão bidimensional dos limites de confiança, respeitando os intervalos bilaterais de Bonferroni;

- A influência do uso da ANCOVA e da FA, com rotação, foi verificada e confrontada com diferentes alternativas, o que indicou que a parametrização fornecida pelo método proposto se faz a melhor opção para aprimorar a classificação e separabilidade das informações, favorecendo, diretamente, possíveis tomadas de decisão. ANCOVA se sobressaiu sobre o método ANOVA, além de apresentar um comportamento desejável ao utilizar diferentes variáveis concomitantes. Já a técnica FA mostrou-se como a melhor alternativa ao se tratar de exploração de dados, a qual foi comparada com a aplicação direta aos dados originais e com a técnica PCA (essas alternativas não proporcionaram uma discriminação desejável dos clusters, apresentando elipses sobrepostas e com menor precisão);
- A comparação com diferentes cenários permitiu determinar a robustez dos métodos de ligação, em que o método "Ward" apresentou estabilidade em todas as réplicas com perturbações. O método "*k-médias*" apresentou uma estabilidade adequada, com base nas estatísticas Kappa e o coeficiente de Kendall (com valores de 0,9543 e 0,944, respectivamente), mas apresentou dificuldades ao classificar a subestação João Neiva, apresentando inversões na formação de clusters. Segundo a AIAG [119], os métodos de ligação "Mediana" e "Centroide" também apresentaram concordância aceitável. Contudo, apenas o método "Ward" apresentou concordância absoluta em todos os cenários. Esse método de ligação resultou em elipses de confiança estreitas com comportamento homogêneo para os diferentes cenários de afundamento de tensão, similares aos resultados encontrados inicialmente, utilizando os dados originais;

Por fim, tem-se que a metodologia proposta se mostrou consistente e robusta, proporcionando encontrar uma combinação de métodos para aprimorar o poder discriminatório na formação de regiões de confiança e em sua classificação, aplicando-as em dados reais.

6.1 Contribuições do trabalho

A contribuição majoritária do presente estudo se dá pela proposição de um método estruturado para aprimorar o poder discriminatório de regiões de confiança, através de funções elipsoidais, auxiliadas por técnicas multivariadas e de DOE. O método se destaca por proporcionar a criação de agrupamentos robustos (confrontados em cenários com perturbações), com alto poder discriminatório das regiões de confiança, comparadas com

diferentes abordagens da literatura, como PCA e a aplicação direta no conjunto de dados. Além disso, favorece a tomada de decisão para classificações, como no exemplo numérico, motivado pela necessidade de regulamentação do setor de distribuição de energia elétrica, baseado em quantidade de afundamentos de tensão.

Assim, tem-se uma contribuição direta para aplicações da FA, em que é possível encontrar o grau de rotação γ *orthomax* ideal com a minimização do EQM, calculado a partir das informações de VTE de cada fator. A contribuição se estende ao uso de um arranjo fatorial multiníveis para analisar e determinar diferentes métodos de ligação e tipos de análise, que proporcionam menor variabilidade na formação dos clusters, contemplando métodos de ligação hierárquicos e não hierárquico. O uso da ANCOVA permite o ajuste da explicação da variável principal a partir de uma variável secundária, melhorando a precisão dos resultados e, consequentemente, auxiliando a estimação de elipses mais estreitas.

A contribuição secundária, se dá pelo uso de técnicas voltadas à análise do sistema de medição para avaliar a estabilidade de métodos de ligação, mediante a questionamentos existentes na literatura. Assim, foi possível estabelecer uma estratégia capaz de avaliar e encontrar o melhor método de ligação a partir de critérios de classificação estabelecidos por critérios internacionais [119].

Em relação ao uso de um conjunto de dados, referente a índices de QEE de subestações, foi possível verificar o comportamento do método em dados reais, favorecendo de modo direto aos critérios estabelecidos pela ANEEL para classificar e regulamentar as subestações diante da quantidade de afundamento de tensão. A proposta contribui para diretrizes necessárias como ao PRODISC [71] e IEEE 1564 [112], destacados anteriormente. Além disso, conforme mencionado anteriormente, espera-se que o método proposto possa ser aplicável em diferentes conjuntos de dados, favorecendo em distintas áreas de pesquisa que apresentam dados com estrutura de variância-covariância significativa.

No que se refere às contribuições científicas/acadêmicas, o presente estudo proporcionou uma contribuição direta, referente ao artigo "*Enhancement of discriminatory power by ellipsoidal functions for substation clustering in voltage sag studies*" [62], além de contribuições indiretas, com diferentes aplicações realizadas na linha de pesquisa de "Modelagem, otimização e controle", as quais foram convertidas em publicações de artigos em periódicos internacionais e nacionais. Além disso, realizou-se publicações em anais de congressos nacionais e internacionais de grande relevância na área. Ainda como contribuições indiretas ao método proposto, pode-se destacar a criação de três diferentes abordagens para a

análise do sistema de medição mesclado a técnica multivariada FA com o uso de rotações [8] e com ponderação por autovalores [7]. Por fim, tem-se uma contribuição voltada a estudos de medição multivariados, criando número de categorias distintas baseadas em funções elipsoidais [188]. Deste modo, tem-se um pequeno acervo de publicações (27 artigos em periódicos internacionais, 3 artigos em periódicos nacionais e 19 artigos em anais de congressos), que estão disponíveis no final deste documento, destacando as importantes contribuições deste trabalho.

6.2 Sugestões para trabalhos futuros

Considerando o desenvolvimento do método proposto, diferentes estudos podem ser realizados com finalidade de vislumbrar a estratégia proposta. Deste modo, as seguintes sugestões podem ser destacadas:

- Aplicação do método desenvolvido em diferentes conjuntos de dados, como no setor econômico, de processos industriais e diferentes instituições que fazem uso de *big data*;
- Contemplar o uso de diferentes parâmetros e arranjos experimentais para o método de otimização dos parâmetros de rotação em FA;
- Considerar diferentes meta-heurísticas, como algoritmo de busca, para a otimização do EQM_{VTE}, em caso de funções mais complexas, com diferentes ótimos locais;
- Combinação de diferentes parâmetros para os métodos de ligação, podendo contemplar outras estratégias para verificar o seu comportamento;
- Criar modelos de aprendizado de máquina para automatizar a classificação e estimação de elipses de confiança baseado no modelo proposto.

APÊNDICE A – Relações trigonométricas para a rotação das elipses de confiança

Assumindo que o $P_1(x_1, x_2)$ tenha um ângulo β com o eixo das abscissas, conforme ilustrado na Figura A.1. Com base nisso, considere uma rotação, em sua posição inicial, de amplitude α, que forme um novo ângulo que resulte em $\alpha + \beta$. Assim, é possível concluir que as coordenadas de P_1 podem ser representadas conforme a Eq. (A.1).

$$\begin{aligned} x_1 &= r\cos\beta \\ x_2 &= r sen\beta \end{aligned} \tag{A.1}$$

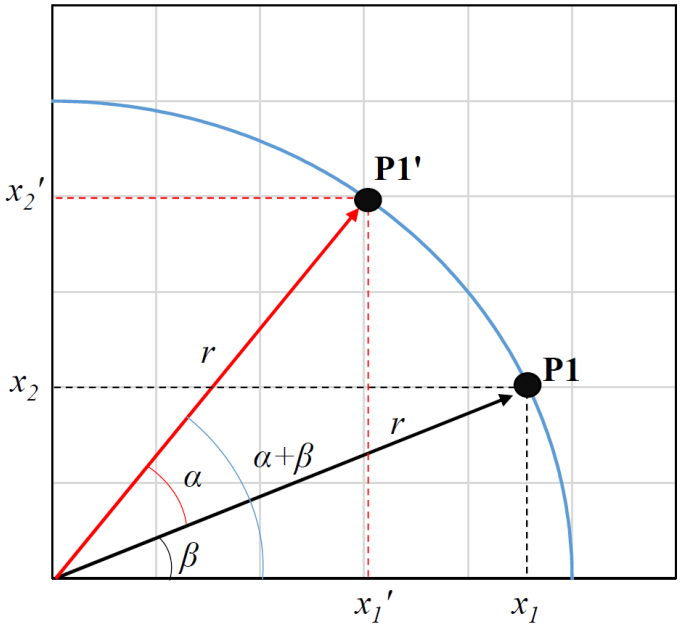


Figura A.1. Rotação do ponto P1 para o P1'

Deste modo, tem-se que as novas coordenadas, referentes ao ponto P_1', podem ser definidas pelas Eqs. (A.2) e (A.3):

$$\begin{aligned} x_1' &= r\left[\cos(\alpha+\beta)\right] = r\cos\alpha(\cos\beta) - rsen\alpha(sen\beta) \\ x_1' &= r\left[\cos(\alpha+\beta)\right] = (r\cos\beta)\cos\alpha - (rsen\beta)sen\alpha \\ x_1' &= r\left[\cos(\alpha+\beta)\right] = x_1\cos\alpha - x_2 sen\alpha \end{aligned} \tag{A.2}$$

De maneira similar, tem-se, para x_2' que:

$$
\begin{aligned}
x_2^{'} &= r\left[sen\left(\alpha+\beta\right)\right] = rsen\alpha\left(\cos\beta\right) - rsen\beta\left(\cos\alpha\right) \\
x_2^{'} &= r\left[sen\left(\alpha+\beta\right)\right] = \left(r\cos\beta\right)sen\alpha - \left(rsen\beta\right)\cos\alpha \\
x_2^{'} &= r\left[\cos\left(\alpha+\beta\right)\right] = x_1 sen\alpha + x_2\cos\alpha
\end{aligned}
\tag{A.3}
$$

Assim, inferindo matricialmente, tem-se, na Eq. (A.4):

$$
\begin{bmatrix} x_1^{'} \\ x_2^{'} \end{bmatrix} = \begin{bmatrix} \cos\alpha & -sen\alpha \\ sen\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{A.4}
$$

A partir da Fig. A.3, é possível verificar que o ponto **P** pode ser definido, conforme a Eq. (A.5).

$$
\mathbf{P} = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & -sen\theta \\ sen\theta & \cos\theta \end{bmatrix} \tag{A.5}
$$

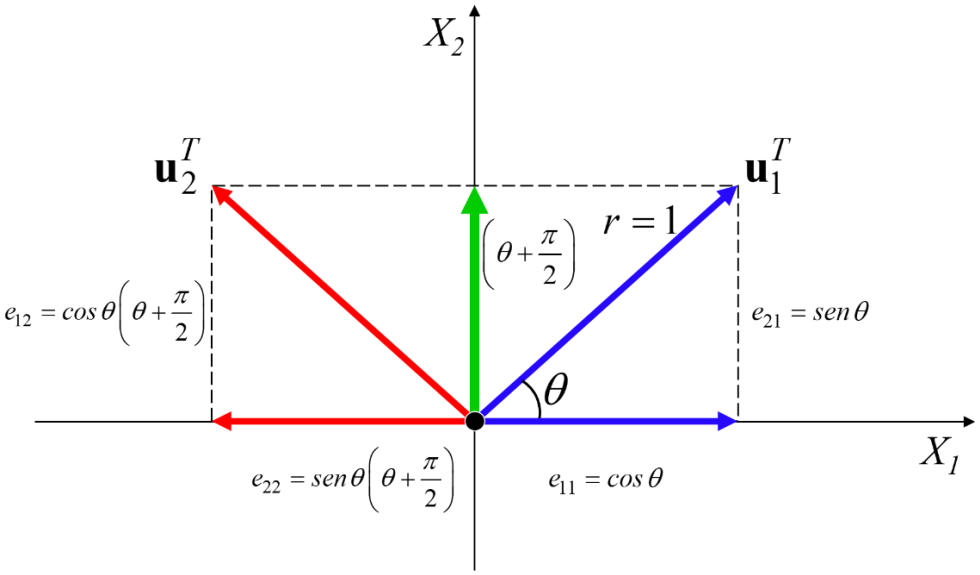


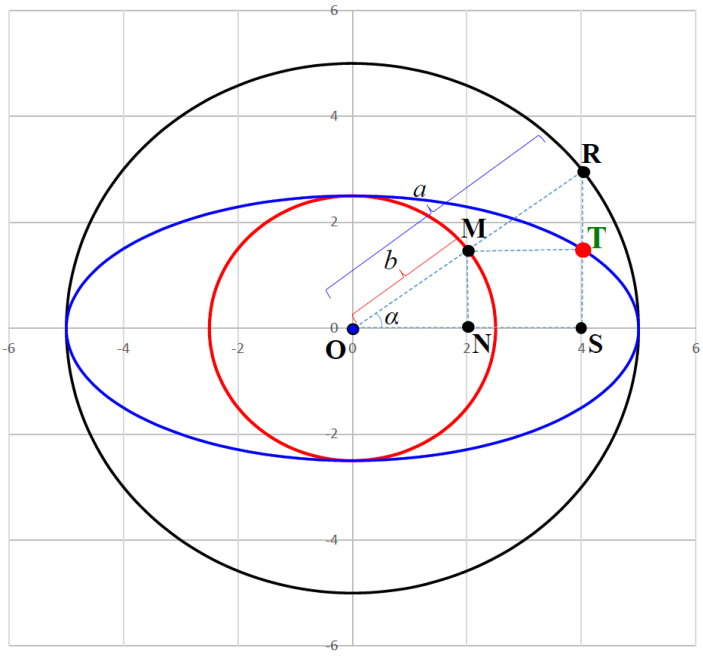
Fig. A.3. Comportamento do ângulo de rotação da elipse

Deste modo, é possível verificar o comportamento do ângulo de rotação com os autovetores da matriz **Σ**, descritos na Eq. (A.6).

$$
\begin{aligned}
sen\left(\theta+\frac{\pi}{2}\right) &= sen\theta\cos\frac{\pi}{2} + sen\frac{\pi}{2}\cos\theta = \cos\theta \\
\cos\left(\theta+\frac{\pi}{2}\right) &= \cos\theta\cos\frac{\pi}{2} - sen\theta.sen\frac{\pi}{2} = -sen\theta
\end{aligned}
\tag{A.6}
$$

Em virtude da sua relação trigonométrica, Almeida *et al.* [188] e [62] afirmam que, para transformar uma elipse em forma não paramétrica para uma em forma paramétrica, se faz necessário assumir que a elipse é o lócus geométrico de duas circunferências de raios, respectivamente iguais ao meio eixo maior (a) e ao meio eixo menor (b). Deste modo, pode-se delinear a elipse, diante de suas relações trigonométricas entre as regiões, para um ponto T aleatório. A Tabela A.1 ilustra e descreve a figura e as equações, respectivamente.

Tabela A.1. relações trigonométricas para a elipse

	$\overline{OS} = x \therefore x = (a)\cos\alpha$ $\overline{MN} = y \therefore y = (b)\sin\alpha$ $T = \begin{bmatrix} (a)\cos\alpha \\ (b)\sin\alpha \end{bmatrix} = \begin{bmatrix} c\sqrt{\lambda_1}\cos\alpha \\ c\sqrt{\lambda_2}\sin\alpha \end{bmatrix}$ $0 \le \alpha \le 2\pi$

APÊNDICE B – Pseudocódigo do método proposto

PSEUDOCÓDIGO

FUNÇÃO PRINCIPAL

Entrada: Variáveis de resposta originais do sistema

Saída: Agrupamentos otimizados

$Y \leftarrow y_1, y_2, \ldots, y_p \leftarrow$ variáveis de resposta $(i = 1,2, \ldots, p)$

$N \leftarrow$ número de linhas dos dados originais

$R \leftarrow \{r_{11}, r_{12}, \ldots, r_{1N}, \ldots, r_{p1}, r_{p2}, \ldots, r_{pN}\}$ //sample correlation matrices R

$Q \leftarrow \{q_{11}, q_{12}, \ldots, q_{1N}, \ldots, q_{p1}, q_{p2}, \ldots, q_{pN}\}$ //anti-image correlation matrices Q

$indiceAdequacao \leftarrow 0$

$iteracao \leftarrow 1$

$correlDados \leftarrow correlacao(Y)$

//**Fluxograma 1**

se $(correlDados = 0)$

    //Utilizar outra estratégia exploratória

senão

    enquanto $(correlDados = 1)$

        se $(p < N)$

            //Teste de Barlett

            $X_2 \leftarrow -\left[n - 1\frac{(2p+5)}{6}\right]\ln|R|$

            $Chi_2 \leftarrow testeChi(alfa; \left[p * \frac{p-1}{2}\right])$

            se $(X_2 \leq Chi_2)$

                $indiceAdequacao \leftarrow 1$

            fim se

            //Índice KMO

            $somaQuad\_r \leftarrow 0$

            $somaQuad\_q \leftarrow 0$

            para $i = 1$ **até** $i = p$

                para $j = 1$ **até** $j = N$

                    se $i \neq j$

                        $somaQuad\_r \leftarrow somaQuad\_r + r_{ij}^2$

                        $somaQuad\_q \leftarrow somaQuad\_q + q_{ij}^2$

                    fim se

                fim para

            fim para

            $indiceKMO \leftarrow somaQuad\_r/(somaQuad\_r + somaQuad\_q)$

            se $(indiceAdequacao > 0$ **ou** $indiceKMO \geq 5)$

                $escores \leftarrow rotacaoOtima(dados)$

            fim se

            se $(iteracao = 1)$

                $numAgrup \leftarrow 1 + 3{,}322log(\zeta)$ // ζ é o número de objetos

                //DOE multinível para 2 fatores (método de ligação + tipo de análise)

                $af_1, af_2, \ldots, af_{16} \leftarrow$ arranjo experimental

```
                                        //Armazena a variância na formação dos agrupamentos
                                        para w = 1 até w = 16
                                            para s = 1 até s = numAgrup
                                                Var_ws ← variância do agrupamento
                                            fim para
                                        fim para
                                fim se

                                //w = 1, 2, … , 16; s = 1, 2, … , numAgrup
                                VarCluster ← {Var_1_1, Var_1_2, …, Var_1_s, …, Var_ws}
                                correlDados ← correlacao(VarClusters)

                                se (correlDados = 0)
                                        //Otimização parâmetros
                                        metodoLigacao_método, tipoAnalise_método
                                                ← metodoLigacao_ótimo, tipoAnalise_ótimo

                                        //Calcular as elipses de confiança
                                fim se
                        fim se
                fim enquanto
                iteracao ← iteracao + 1
                indiceAdequacao ← 0
                indiceKMO ← 0
        fim se

        //Fluxograma 2
        //Define percentual de perturbação as ser aplicado nos dados originais
        perturb ← percentual de perturbação

        //Aplica-se o percentual a 4 réplicas distintas
        YR_1 ← perturb * Y
        YR_2 ← perturb * Y
        YR_3 ← perturb * Y
        YR_4 ← perturb * Y

        //Otimização nível gama
        escoresR_1 ← rotacaoOtima(YP_1)
        escoresR_2 ← rotacaoOtima(YP_2)
        escoresR_3 ← rotacaoOtima(YP_3)
        escoresR_4 ← rotacaoOtima(YP_4)

        //Aplicação dos métodos de ligação
        Me_1 ← metodosLigacao(escoresR_1)
        Me_2 ← metodosLigacao(escoresR_2)
        Me_3 ← metodosLigacao(escoresR_3)
        Me_4 ← metodosLigacao(escoresR_4)

        //Índices de Kappa e Kendall
        para z = 1 até z = 8
                i_Kappa_z, cc_Kendall_z ← KappaKendall(Me_1, Me_2, Me_3, Me_4)
        fim para

        // Define os métodos com melhor comportamento e robustez perante os critérios de classificação mundiais, estabelecidos pela AIAG
        para z = 1 até z = 8
                se (i_Kappa_z > 0,9 e cc_Kendall_z > 0,9)
                        metodoLigacao_conf ← melhor método de confirmação
```

fim se
fim para

se $metodoLigacao_{método} = metodoLigacao_{conf}$
//Método proposto confirmado
fim se

FUNÇÕES AUXILIARES

Função Correlação
Entrada: l variáveis
Saída: Existência ou não de correlação

$d_1, d_2, \ldots, d_l \leftarrow$ dados para análise de correlação

//Avaliação da correlação
se (DadosCorrelacionados)
retorna 1
Senão
retorna 0

Função Rotação Ótima
Entrada: l variáveis
Saída: Escores dos fatores obtidos com g ótimo

//Percentual de explicação acumulado deve ser, pelo menos, 80%
//Quantidade de fatores pode ser definida por até 50% da quantidade de l variáveis, se o
//percentual de incremento for de, pelo menos, 10%
$nF \leftarrow$ quantidade de fatores a serem extraídos

//DOE para 2 componentes (valor γ + complemento de proporção)
// *Lattice* igual a 10 e 2 pontos axiais
$am_1, am_2, \ldots, am_{13} \leftarrow$ arranjo experimental

//Análise fatorial extraída por componentes principais
para $z = 1$ **até** $z = 13$
$f_{1_{am_z}}, f_{2_{am_z}}, \ldots, f_{nF_{am_z}} \leftarrow \text{FA}(\gamma_{am_z})$
$VTE_{f_{1_{am_z}}}, VTE_{f_{2_{am_z}}}, \ldots, VTE_{f_{nF_{am_z}}} \leftarrow$ variância total explicada de cada fator para cada experimento
$$EQM_{VTE_{am_z}} \leftarrow \sum_{i=1}^{nF} \left[\left(VTE_z - \bar{X}_{VTE_z} \right)^2 + \sigma^2_{VTE_z} \right]$$

fim para

//Otimização (minimização) da rotação γ
$\gamma \leftarrow \gamma_{ótimo}$

//Análise fatorial
$Escores \leftarrow e_1, e_2, \ldots, e_{13} \leftarrow \text{FA}(\gamma_{ótimo})$

retorna $Escores$

Função Métodos de Ligação
Entrada: l variáveis
Saída: memberships

$meUn_1, meUn_2, \ldots, meUn_{numAgrup} \leftarrow$ memberships método de ligação Único
$meCe_1, meCe_2, \ldots, meCe_{numAgrup} \leftarrow$ memberships método de ligação Centróide
$meCo_1, meCo_2, \ldots, meCo_{numAgrup} \leftarrow$ memberships método de ligação Completo
$meMe_1, meMe_2, \ldots, meMe_{numAgrup} \leftarrow$ memberships método de ligação Média

163: $meMa_1, meMa_2, \dots, meMa_{numAgrup}$ ← memberships método de ligação Mediana
164: $meMc_1, meMc_2, \dots, meMc_{numAgrup}$ ← memberships método de ligação McQuilty
165: $meWa_1, meWa_2, \dots, meWa_{numAgrup}$ ← memberships método de ligação Ward
166: $meKM_1, meKM_2, \dots, meKM_{numAgrup}$ ← memberships método de ligação K-Médias
167:
168: //Matriz com todos os memberships calculados
169: ME = $\{meUn_1, \dots, meUn_{numAgrup}, \dots, meCe_1, \dots, meCo_1, \dots, meKM_{numAgrup}\}$
170:
171: **retorna** ME

172:
173: **Função** Kappa e Kendall
174: **Entrada:** Memberships
175: **Saída:** índice de Kappa e coeficiente de Kendall

176: i_{Kappa} ← estatística de Kappa
177: $cc_{Kendall}$ ← coeficiente de concordância de Kendall
178:
179: **retorna** indiceKappa, ccKendall

APÊNDICE C – Análises e informações complementares

Tabela C.1. Análise de correlação dos índices de qualidade de energia (Parte I)

	TNE	NEMV	MVFR	MNE	LNE	ANE	UNE	FKVAr	SAIFI1	SAIFI2
NEMV	1,000									
	0,000									
MVFR	0,314	0,324								
	0,219	0,204								
MNE	0,169	0,163	-0,409							
	0,516	0,531	0,103							
LNE	0,442	0,424	-0,086	0,510						
	0,076	0,090	0,741	0,036						
ANE	0,431	0,413	-0,078	0,503	0,996					
	0,084	0,100	0,765	0,040	0,000					
UNE	0,329	0,309	-0,122	0,449	0,916	0,919				
	0,198	0,228	0,641	0,070	0,000	0,000				
FKVAr	0,576	0,569	0,301	-0,114	0,586	0,575	0,459			
	0,015	0,017	0,240	0,662	0,013	0,016	0,064			
SAIFI1	0,475	0,479	0,075	0,691	0,380	0,385	0,394	0,047		
	0,054	0,052	0,775	0,002	0,133	0,126	0,117	0,857		
SAIFI2	0,336	0,343	0,010	0,673	0,317	0,317	0,369	-0,086	0,955	
	0,187	0,178	0,971	0,003	0,215	0,215	0,145	0,744	0,000	
STIFI	0,469	0,465	0,522	-0,150	0,403	0,393	0,211	0,763	0,095	-0,005
	0,057	0,060	0,032	0,566	0,109	0,119	0,416	0,000	0,717	0,984
FL	0,137	0,125	-0,242	0,108	0,533	0,520	0,337	0,628	-0,048	-0,055
	0,600	0,631	0,349	0,681	0,027	0,032	0,186	0,007	0,856	0,835
AREA	0,289	0,275	-0,171	0,034	0,538	0,538	0,324	0,716	-0,245	-0,322
	0,260	0,285	0,511	0,897	0,026	0,026	0,205	0,001	0,342	0,207
EMVVA	0,712	0,710	-0,337	0,511	0,342	0,325	0,272	0,132	0,485	0,433
	0,001	0,001	0,187	0,036	0,179	0,202	0,291	0,612	0,049	0,082
EVAHV	-0,159	-0,190	-0,364	0,161	0,484	0,497	0,560	0,135	-0,217	-0,260
	0,541	0,464	0,151	0,538	0,049	0,042	0,019	0,606	0,404	0,313
NEHV	-0,159	-0,190	-0,364	0,161	0,484	0,497	0,560	0,135	-0,217	-0,260
	0,541	0,464	0,151	0,538	0,049	0,042	0,019	0,606	0,404	0,313

Correlação de Pearson
Valor-p

Tabela C.2. Análise de correlação dos índices de qualidade de energia (Parte II)

	TNE	NEMV	MVFR	MNE	LNE	ANE	UNE	FKVAr	SAIFI1	SAIFI2
3LG	0,306	0,310	0,589	-0,612	-0,136	-0,108	-0,193	0,431	-0,127	-0,207
	0,232	0,227	0,013	0,009	0,602	0,681	0,457	0,084	0,627	0,425
2LG	0,309	0,311	0,632	-0,604	-0,081	-0,049	-0,135	0,489	-0,125	-0,225
	0,228	0,225	0,006	0,010	0,759	0,851	0,605	0,046	0,633	0,386
1LG	0,163	0,165	0,641	-0,566	-0,092	-0,061	-0,123	0,531	-0,169	-0,263
	0,533	0,526	0,006	0,018	0,725	0,815	0,638	0,028	0,518	0,308
L-L	0,306	0,310	0,588	-0,612	-0,136	-0,108	-0,193	0,431	-0,127	-0,207
	0,232	0,227	0,013	0,009	0,603	0,681	0,457	0,084	0,627	0,425
MAXA	0,293	0,295	0,621	-0,592	-0,122	-0,093	-0,172	0,446	-0,114	-0,201
	0,255	0,250	0,008	0,012	0,641	0,724	0,509	0,072	0,664	0,438
MAXS	0,327	0,329	0,622	-0,590	-0,078	-0,048	-0,134	0,463	-0,108	-0,208
	0,200	0,198	0,008	0,013	0,765	0,856	0,609	0,061	0,681	0,424
MAXG	0,194	0,193	0,633	-0,479	0,014	0,046	-0,013	0,535	-0,115	-0,229
	0,456	0,458	0,006	0,052	0,956	0,862	0,960	0,027	0,661	0,376
R+	0,031	0,019	-0,340	0,360	0,423	0,404	0,387	-0,137	-0,063	-0,045
	0,905	0,942	0,182	0,156	0,090	0,108	0,124	0,601	0,810	0,864
X+	-0,323	-0,328	-0,579	0,506	0,029	0,007	0,074	-0,461	-0,041	0,066
	0,206	0,199	0,015	0,038	0,912	0,979	0,779	0,063	0,876	0,800
Xo	0,030	0,031	-0,347	0,181	-0,152	-0,167	-0,171	-0,418	0,062	0,139
	0,909	0,907	0,173	0,487	0,559	0,522	0,511	0,095	0,815	0,596
ZBASE	-0,022	-0,019	0,200	-0,577	-0,265	-0,258	-0,112	-0,036	-0,295	-0,201
	0,933	0,943	0,442	0,015	0,303	0,317	0,669	0,892	0,250	0,438
Zohm	-0,314	-0,319	-0,579	0,509	0,041	0,018	0,084	-0,457	-0,037	0,069
	0,220	0,213	0,015	0,037	0,877	0,944	0,748	0,065	0,888	0,792
Zpu	-0,316	-0,321	-0,588	0,551	0,065	0,044	0,098	-0,451	-0,014	0,081
	0,217	0,209	0,013	0,022	0,803	0,867	0,710	0,069	0,956	0,757
MVASC	0,280	0,282	0,573	-0,616	-0,089	-0,059	-0,136	0,440	-0,117	-0,186
	0,276	0,272	0,016	0,008	0,735	0,822	0,602	0,077	0,655	0,474
BKVAr	0,050	0,052	0,320	-0,328	0,117	0,137	0,127	0,307	0,075	0,116
	0,850	0,843	0,211	0,199	0,654	0,600	0,628	0,230	0,774	0,659

Correlação de Pearson
Valor-p

Tabela C.3. Análise de correlação dos índices de qualidade de energia (Parte III)

	STIFI	FL	AREA	EMVVA	EVAHV	NEHV	3LG	2LG	1LG	L-L
FL	0,638 0,006									
AREA	0,542 0,025	0,820 0,000								
EMVVA	-0,072 0,783	0,075 0,775	0,185 0,476							
EVAHV	0,068 0,797	0,343 0,177	0,400 0,111	-0,057 0,828						
NEHV	0,068 0,797	0,343 0,177	0,400 0,111	-0,057 0,828	1,000 *					
3LG	0,647 0,005	0,212 0,415	0,249 0,335	-0,236 0,362	-0,161 0,538	-0,161 0,538				
2LG	0,694 0,002	0,230 0,374	0,270 0,294	-0,262 0,309	-0,110 0,675	-0,110 0,675	0,979 0,000			
1LG	0,655 0,004	0,240 0,354	0,276 0,284	-0,386 0,126	-0,106 0,685	-0,106 0,685	0,810 0,000	0,899 0,000		
L-L	0,647 0,005	0,212 0,414	0,249 0,335	-0,236 0,362	-0,160 0,538	-0,160 0,538	1,000 0,000	0,979 0,000	0,810 0,000	
MAXA	0,672 0,003	0,207 0,425	0,246 0,342	-0,258 0,318	-0,138 0,597	-0,138 0,597	0,992 0,000	0,991 0,000	0,857 0,000	0,992 0,000
MAXS	0,687 0,002	0,215 0,408	0,255 0,322	-0,233 0,368	-0,101 0,699	-0,101 0,699	0,988 0,000	0,996 0,000	0,857 0,000	0,988 0,000
MAXG	0,660 0,004	0,226 0,383	0,262 0,309	-0,335 0,189	-0,014 0,958	-0,014 0,958	0,730 0,001	0,850 0,000	0,977 0,000	0,730 0,001
R+	-0,311 0,224	-0,075 0,776	0,073 0,780	0,309 0,228	0,377 0,135	0,377 0,135	-0,691 0,002	-0,646 0,005	-0,590 0,013	-0,691 0,002
X+	-0,649 0,005	-0,224 0,388	-0,122 0,642	0,251 0,331	0,195 0,454	0,195 0,454	-0,882 0,000	-0,910 0,000	-0,817 0,000	-0,882 0,000
Xo	-0,323 0,206	-0,174 0,504	-0,151 0,562	0,349 0,169	-0,024 0,927	-0,024 0,927	-0,156 0,550	-0,331 0,194	-0,667 0,003	-0,156 0,550
ZBASE	-0,098 0,710	-0,289 0,261	-0,227 0,381	-0,113 0,667	-0,101 0,701	-0,101 0,701	0,226 0,383	0,199 0,444	0,133 0,611	0,226 0,383
Zohm	-0,648 0,005	-0,224 0,387	-0,121 0,644	0,259 0,316	0,198 0,445	0,198 0,445	-0,886 0,000	-0,912 0,000	-0,819 0,000	-0,886 0,000
Zpu	-0,639 0,006	-0,206 0,428	-0,101 0,698	0,255 0,324	0,209 0,421	0,209 0,421	-0,895 0,000	-0,919 0,000	-0,823 0,000	-0,895 0,000
MVASC	0,658 0,004	0,245 0,344	0,249 0,335	-0,261 0,313	-0,112 0,668	-0,112 0,668	0,990 0,000	0,976 0,000	0,811 0,000	0,991 0,000
BKVAr	0,231 0,373	0,225 0,385	0,124 0,635	-0,226 0,382	-0,081 0,756	-0,081 0,756	0,467 0,059	0,429 0,086	0,348 0,171	0,467 0,059

Correlação de Pearson
Valor-p

Tabela C.4. Análise de correlação dos índices de qualidade de energia (Parte IV)

	MAXA	MAXS	MAXG	R+	X+	Xo	ZBASE	Zohm	Zpu	MVASC
MAXS	0,995 0,000									
MAXG	0,793 0,000	0,807 0,000								
R+	-0,702 0,002	-0,648 0,005	-0,474 0,055							
X+	-0,887 0,000	-0,901 0,000	-0,784 0,000	0,649 0,005						
Xo	-0,234 0,366	-0,253 0,326	-0,740 0,001	0,114 0,664	0,429 0,086					
ZBASE	0,192 0,460	0,199 0,443	0,088 0,738	0,030 0,910	-0,068 0,797	0,148 0,572				
Zohm	-0,892 0,000	-0,904 0,000	-0,784 0,000	0,666 0,004	1,000 0,000	0,424 0,090	-0,063 0,810			
Zpu	-0,899 0,000	-0,911 0,000	-0,785 0,000	0,659 0,004	0,997 0,000	0,409 0,103	-0,138 0,597	0,997 0,000		
MVASC	0,982 0,000	0,983 0,000	0,738 0,001	-0,648 0,005	-0,889 0,000	-0,178 0,495	0,287 0,265	-0,892 0,000	-0,905 0,000	
BKVAr	0,443 0,075	0,427 0,087	0,268 0,298	-0,382 0,130	-0,456 0,066	-0,100 0,702	0,129 0,621	-0,459 0,064	-0,464 0,061	0,508 0,038

Correlação de Pearson
Valor-p

Tabela C.5. Cargas fatoriais e comunalidades para rotação *orthomax* otimizada ($\gamma = 1$)

Variável	Rotação Ortomax otimizada ($\gamma = 1$)							Comum
	F1	F2	F3	F4	F5	F6	F7	
TNE	0,142	0,911	0,051	0,216	0,246	-0,174	0,053	0,992
NEMV	0,145	0,908	0,081	0,218	0,241	-0,177	0,053	0,993
MVFR	0,411	0,099	0,226	0,025	-0,001	-0,856	0,023	0,965
UNE	-0,233	0,170	-0,604	0,462	0,422	-0,106	0,337	0,964
FKVAr	0,266	0,308	-0,116	0,018	0,851	-0,225	0,051	0,958
SAIFI1	-0,041	0,318	0,075	0,928	-0,057	-0,022	-0,039	0,975
SAIFI2	-0,098	0,198	0,131	0,943	-0,114	0,039	0,052	0,972
AREA	0,138	0,223	-0,294	-0,277	0,797	0,240	-0,060	0,929
EMVVA	-0,261	0,836	0,026	0,248	0,039	0,375	-0,076	0,977
EVAHV	-0,111	-0,087	-0,966	-0,123	0,114	0,111	-0,014	0,995
NEHV	-0,111	-0,087	-0,966	-0,123	0,114	0,111	-0,014	0,995
3LG	0,932	0,154	0,048	-0,155	0,123	-0,217	0,076	0,987
1LG	0,724	-0,063	0,026	-0,117	0,343	-0,449	-0,301	0,953
L-L	0,932	0,154	0,048	-0,155	0,123	-0,217	0,076	0,987
R+	-0,853	0,222	-0,317	-0,173	-0,008	-0,091	0,110	0,928
VTE	3,448	2,793	2,522	2,320	1,850	1,387	0,250	14,568
% Var	0,230	0,186	0,168	0,155	0,123	0,092	0,017	0,971

Tabela C.5. Cargas fatoriais e comunalidades para rotação *quartimax* ($\gamma = 0$)

Variável	Rotação Quartimax ($\gamma = 0$)							Comum
	F1	F2	F3	F4	F5	F6	F7	
TNE	0,185	0,931	0,039	0,195	0,183	-0,136	0,032	0,992
NEMV	0,188	0,928	0,070	0,198	0,179	-0,138	0,033	0,993
MVFR	0,551	0,105	0,232	0,039	-0,027	-0,771	-0,005	0,965
UNE	-0,209	0,213	-0,625	0,450	0,413	-0,165	0,289	0,964
FKVAr	0,357	0,361	-0,154	0,004	0,804	-0,168	0,028	0,958
SAIFI1	-0,060	0,337	0,078	0,919	-0,056	-0,018	-0,058	0,975
SAIFI2	-0,134	0,216	0,134	0,937	-0,096	0,025	0,037	0,972
AREA	0,157	0,260	-0,331	-0,295	0,751	0,269	-0,057	0,929
EMVVA	-0,316	0,840	0,023	0,217	0,007	0,342	-0,079	0,977
EVAHV	-0,126	-0,086	-0,970	-0,126	0,083	0,078	-0,032	0,995
NEHV	-0,126	-0,086	-0,970	-0,126	0,083	0,078	-0,032	0,995
3LG	0,963	0,151	0,035	-0,134	0,048	-0,059	0,106	0,987
1LG	0,825	-0,051	0,016	-0,105	0,287	-0,297	-0,296	0,953
L-L	0,963	0,151	0,035	-0,134	0,048	-0,059	0,106	0,987
R+	-0,824	0,227	-0,309	-0,197	0,023	-0,242	0,059	0,928
VTE	3,937	2,962	2,590	2,257	1,561	1,044	0,216	14,568
% Var	0,262	0,197	0,173	0,150	0,104	0,070	0,014	0,971

Tabela C.5. Cargas fatoriais e comunalidades não rotacionadas

Variável	Não Rotacionadas							Comum
	F1	F2	F3	F4	F5	F6	F7	
TNE	0,506	0,802	-0,019	-0,230	-0,107	-0,171	-0,001	0,992
NEMV	0,515	0,796	0,007	-0,238	-0,108	-0,161	0,002	0,993
MVFR	0,730	-0,092	0,176	0,215	-0,583	-0,060	-0,044	0,965
UNE	-0,163	0,645	-0,485	0,450	-0,137	0,145	0,210	0,964
FKVAr	0,586	0,387	-0,543	-0,046	-0,077	0,401	0,018	0,958
SAIFI1	0,049	0,729	0,453	0,454	0,130	0,042	-0,108	0,975
SAIFI2	-0,056	0,654	0,530	0,478	0,154	0,088	-0,008	0,972
AREA	0,236	0,218	-0,765	-0,296	0,238	0,308	-0,031	0,929
EMVVA	-0,100	0,838	0,100	-0,399	0,227	-0,191	-0,088	0,977
EVAHV	-0,379	0,052	-0,828	0,296	0,046	-0,259	-0,079	0,995
NEHV	-0,379	0,052	-0,828	0,296	0,046	-0,259	-0,079	0,995
3LG	0,913	-0,193	-0,152	0,074	0,150	-0,216	0,131	0,987
1LG	0,840	-0,265	-0,235	0,168	-0,055	0,093	-0,284	0,953
L-L	0,913	-0,193	-0,152	0,074	0,150	-0,216	0,131	0,987
R+	-0,672	0,330	-0,243	-0,245	-0,496	-0,050	0,002	0,928
VTE	4,609	3,819	3,176	1,316	0,835	0,620	0,194	14,568
% Var	0,307	0,255	0,212	0,088	0,056	0,041	0,013	0,971

Tabela C.6. Partição final criada pelos métodos de ligação

	Cluster	Observações	Dentro da soma de quadrados do agrupado	Distância média do centroide	Distância máxima do centroide
Único	1	2	2,407	1,097	1,097
	2	12	66,010	2,261	3,209
	3	1	0,000	0,000	0,000
	4	1	0,000	0,000	0,000
	5	1	0,000	0,000	0,000
Centroide	1	13	67,681	2,239	2,823
	2	1	0,000	0,000	0,000
	3	1	0,000	0,000	0,000
	4	1	0,000	0,000	0,000
	5	1	0,000	0,000	0,000
Completa	1	3	8,994	1,709	2,096
	2	2	7,939	1,992	1,992
	3	5	17,652	1,794	2,634
	4	6	27,060	2,084	2,679
	5	1	0,000	0,000	0,000
Média	1	12	60,456	2,189	2,822
	2	1	0,000	0,000	0,000
	3	2	4,943	1,572	1,572
	4	1	0,000	0,000	0,000
	5	1	0,000	0,000	0,000
Mediana	1	13	68,809	2,255	2,965
	2	1	0,000	0,000	0,000
	3	1	0,000	0,000	0,000
	4	1	0,000	0,000	0,000
	5	1	0,000	0,000	0,000
McQuitty	1	4	17,992	2,086	2,598
	2	5	15,218	1,651	2,340
	3	6	27,060	2,084	2,679
	4	1	0,000	0,000	0,000
	5	1	0,000	0,000	0,000
Ward	1	4	17,992	2,086	2,598
	2	6	25,757	1,936	2,963
	3	2	4,943	1,572	1,572
	4	1	0,000	0,000	0,000
	5	4	8,302	1,408	1,854
k-médias	1	1	0,000	0,000	0,000
	2	1	0,000	0,000	0,000
	3	2	5,207	1,614	1,614
	4	3	5,439	1,304	1,777
	5	10	62,320	2,377	3,499

Tabela C.7. Valores calculados dos centroides do grupo (Parte I)

Variável	Único					
	Cluster					Centroide
	1	2	3	4	5	global
F1	-1,495	0,003	0,337	1,707	0,914	0,000
F2	0,272	-0,104	-0,838	1,652	-0,108	0,000
F3	-0,077	-0,098	1,940	-0,352	-0,258	0,000
F4	-1,672	0,249	0,224	-0,199	0,328	0,000
F5	-0,307	0,255	-0,470	-1,373	-0,608	0,000
F6	-0,203	0,284	-1,542	0,986	-2,452	0,000
F7	-0,037	-0,186	1,052	2,167	-0,913	0,000
	Centroide					
F1	-0,170	0,337	1,707	0,656	-0,485	0,000
F2	-0,134	-0,838	1,652	1,212	-0,280	0,000
F3	-0,298	1,940	-0,352	0,522	1,760	0,000
F4	-0,116	0,224	-0,199	0,250	1,228	0,000
F5	-0,010	-0,470	-1,373	2,898	-0,928	0,000
F6	-0,056	-1,542	0,986	-0,210	1,488	0,000
F7	-0,208	1,052	2,167	0,541	-1,059	0,000
	Completa					
F1	-0,773	-0,074	-0,090	0,202	1,707	0,000
F2	-0,034	-0,559	0,800	-0,739	1,652	0,000
F3	0,445	1,850	0,066	-0,836	-0,352	0,000
F4	-1,639	0,726	0,621	0,093	-0,199	0,000
F5	-0,144	-0,699	0,693	-0,044	-1,373	0,000
F6	0,225	-0,027	0,021	-0,286	0,986	0,000
F7	-0,038	-0,004	-0,436	0,022	2,167	0,000
	Média					
F1	-0,403	0,337	1,070	1,707	0,656	0,000
F2	-0,188	-0,838	0,114	1,652	1,212	0,000
F3	-0,064	1,940	-0,670	-0,352	0,522	0,000
F4	0,059	0,224	-0,489	-0,199	0,250	0,000
F5	0,002	-0,470	-0,542	-1,373	2,898	0,000
F6	0,272	-1,542	-1,247	0,986	-0,210	0,000
F7	-0,107	1,052	-1,237	2,167	0,541	0,000

Tabela C.8. Valores calculados dos centroides do grupo (Parte II)

Variável	Cluster					Centroide
	1	2	3	4	5	global
Mediana						
F1	-0,152	1,707	-1,298	0,656	0,914	0,000
F2	-0,358	1,652	1,904	1,212	-0,108	0,000
F3	0,037	-0,352	-0,387	0,522	-0,258	0,000
F4	-0,126	-0,199	1,261	0,250	0,328	0,000
F5	-0,057	-1,373	-0,183	2,898	-0,608	0,000
F6	0,188	0,986	-0,774	-0,210	-2,452	0,000
F7	-0,137	2,167	-0,009	0,541	-0,913	0,000
McQuitty						
F1	-0,496	-0,318	0,202	1,707	0,656	0,000
F2	-0,235	0,502	-0,739	1,652	1,212	0,000
F3	0,819	0,314	-0,836	-0,352	0,522	0,000
F4	-1,173	0,817	0,093	-0,199	0,250	0,000
F5	-0,225	-0,072	-0,044	-1,373	2,898	0,000
F6	-0,217	0,361	-0,286	0,986	-0,210	0,000
F7	0,235	-0,756	0,022	2,167	0,541	0,000
Ward						
F1	-0,496	-0,156	1,070	1,707	-0,232	0,000
F2	-0,235	0,620	0,114	1,652	-1,165	0,000
F3	0,819	0,349	-0,670	-0,352	-0,919	0,000
F4	-1,173	0,722	-0,489	-0,199	0,384	0,000
F5	-0,225	0,423	-0,542	-1,373	0,205	0,000
F6	-0,217	0,266	-1,247	0,986	0,195	0,000
F7	0,235	-0,540	-1,237	2,167	0,652	0,000
k-médias						
F1	-1,236	0,670	0,625	-1,208	0,294	0,000
F2	0,095	-0,648	-0,473	-0,526	0,308	0,000
F3	0,391	1,489	0,841	-0,609	-0,173	0,000
F4	-1,960	-1,573	0,276	-0,367	0,408	0,000
F5	-0,649	0,182	-0,539	0,416	0,030	0,000
F6	-1,027	1,080	-1,997	0,227	0,326	0,000
F7	0,041	-0,039	0,070	0,640	-0,206	0,000

Tabela C.9. Distância entre os centroides do cluster

	Único					Mediana				
Cluster	1	2	3	4	5	1	2	3	4	5
1	0,000	2,579	3,906	4,670	3,987	0,000	3,916	3,081	3,584	3,061
2	2,579	0,000	3,193	3,865	3,102	3,916	0,000	4,525	4,957	5,090
3	3,906	3,193	0,000	4,686	3,227	3,081	4,525	0,000	4,031	3,693
4	4,670	3,865	4,686	0,000	5,090	3,584	4,957	4,031	0,000	4,674
5	3,987	3,102	3,227	5,090	0,000	3,061	5,090	3,693	4,674	0,000
	Centroide					McQuitty				
1	0,000	3,143	3,913	3,513	3,180	0,000	2,476	2,273	4,151	3,922
2	3,143	0,000	4,686	4,437	3,985	2,476	0,000	2,163	4,187	3,556
3	3,913	4,686	0,000	4,957	5,089	2,273	2,163	0,000	4,036	3,848
4	3,513	4,437	4,957	0,000	5,108	4,151	4,187	4,036	0,000	4,957
5	3,180	3,985	5,089	5,108	0,000	3,922	3,556	3,848	4,957	0,000
	Completa					Ward				
1	0,000	2,950	2,705	2,522	4,319	0,000	2,432	2,930	4,151	2,628
2	2,950	0,000	2,677	2,867	4,468	2,432	0,000	2,823	4,117	2,527
3	2,705	2,677	0,000	2,096	4,096	2,930	2,823	0,000	4,497	3,219
4	2,522	2,867	2,096	0,000	4,036	4,151	4,117	4,497	0,000	4,216
5	4,319	4,468	4,096	4,036	0,000	2,628	2,527	3,219	4,216	0,000
	Média					k-médias				
1	0,000	3,142	2,609	3,945	3,535	0,000	3,268	3,152	2,644	3,265
2	3,142	0,000	3,754	4,686	4,437	3,268	0,000	3,724	3,263	2,892
3	2,609	3,754	0,000	4,497	4,405	3,152	3,724	0,000	3,473	2,750
4	3,945	4,686	4,497	0,000	4,957	2,644	3,263	3,473	0,000	2,149
5	3,535	4,437	4,405	4,957	0,000	3,265	2,892	2,750	2,149	0,000

Tabela C.10. Valores de TNE ajustados para ANCOVA

TNE com EMVVA como concomitante							
Único	Centroide	Completa	Média	Mediana	McQuitty	Ward	k-médias
117,271	174,414	155,974	146,267	167,595	169,071	169,916	210,895
172,907	184,833	160,867	156,751	174,338	174,931	175,458	131,932
196,877	196,877	140,933	196,877	137,62	143,021	145,281	234,675
292,439	320,953	224,8	293,729	262,434	251,495	247,863	270,749
287,776	280,928	374,928	253,452	236,53	297,855	336,013	305,514
206,381	212,836	189,714	305,162	192,461	193,596	288,996	225,003
426,129	426,129	426,129	426,129	426,129	426,129	426,129	360,639
456,209	421,832	441,107	395,244	498,078	377,11	410,963	472,117
340,26	543,971	395,55	543,971	543,971	543,971	359,367	357,428
292,078	144,681	157,703	236,577	292,078	155,261	252,743	254,281
163,039	176,578	172,685	148,445	168,995	173,202	124,692	100,042
46,209	78,842	126,78	50,093	105,741	118,228	72,705	66,57
229,059	231,808	198,625	204,022	204,74	204,267	154,07	247,435
348,743	331,93	398,883	304,775	269,538	326,543	363,142	365,818
246,352	246,274	358,652	218,579	214,102	278,363	317,58	264,539
168,657	181,277	174,892	153,173	172,037	175,845	127,192	105,598
341,762	177,98	233,924	298,899	265,759	323,258	360,036	358,913

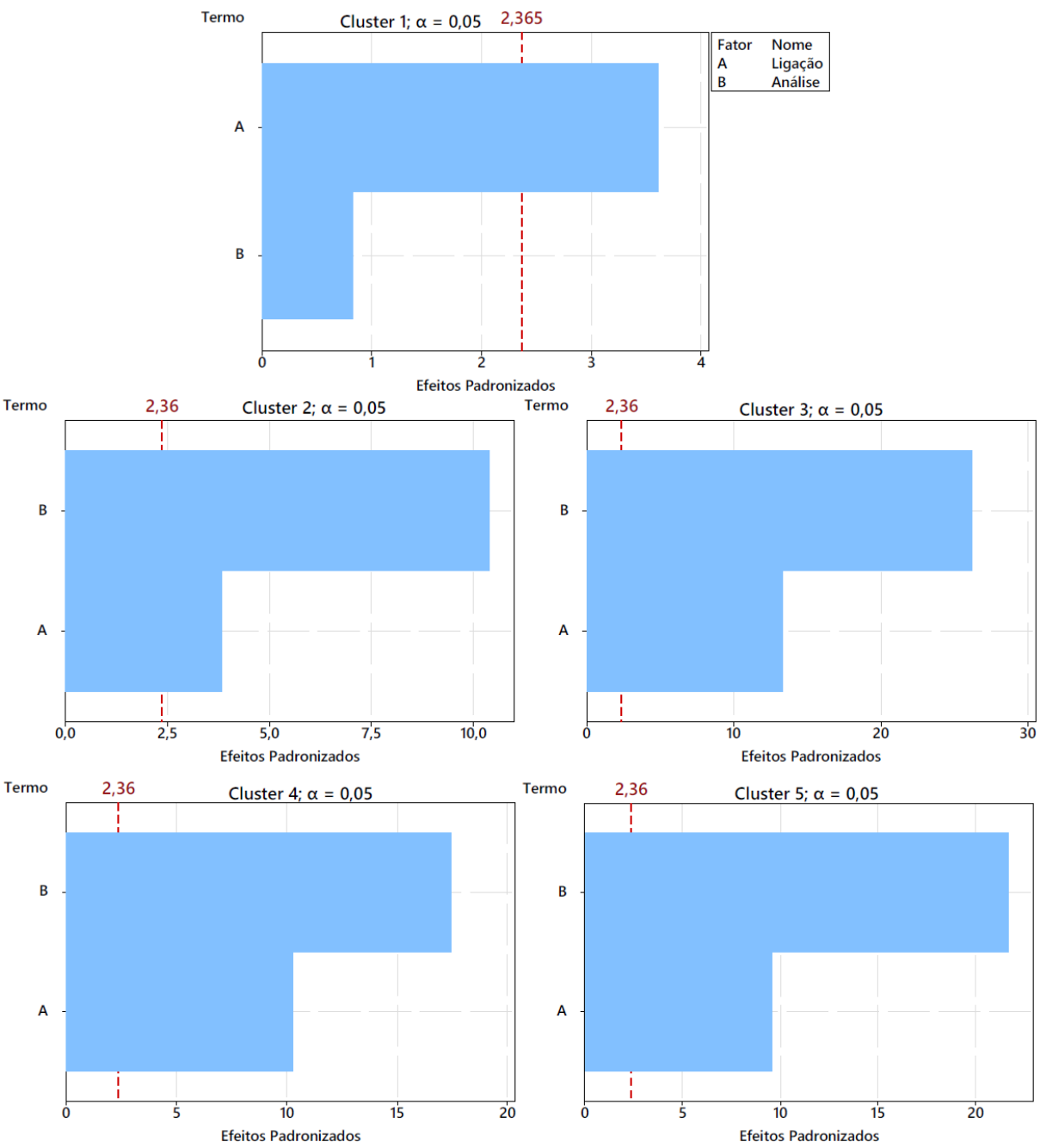


Figura C.1. Gráfico de Pareto dos Clusters (DOE multiníveis)

Tabela C.11. Resíduos dos clusters e dos escores fatoriais

Teste	Parâmetros		Resíduos							
	Ligação	Análise	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	FC1	FC2	FC3
1	Único	ANOVA	-8031,84	1759,34	1791,42	2050,74	1224,29	0,3368	0,5968	-1,1122
2	Único	ANCOVA	8031,84	-1759,34	-1791,42	-2050,74	-1224,29	-0,3368	-0,5968	1,1122
3	Centroide	ANOVA	2253,2	-1019,29	101,39	360,71	-465,75	0,0577	-0,2265	0,2926
4	Centroide	ANCOVA	-2253,2	1019,29	-101,39	-360,71	465,75	-0,0577	0,2265	-0,2926
5	Completa	ANOVA	-230,19	-5742,43	3915	1286,81	383,84	0,9237	-0,9599	0,1408
6	Completa	ANCOVA	230,19	5742,43	-3915	-1286,81	-383,84	-0,9237	0,9599	-0,1408
7	Média	ANOVA	2367,46	-1195,91	-3275,36	184,1	-642,36	-0,4001	-0,0342	0,2925
8	Média	ANCOVA	-2367,46	1195,91	3275,36	-184,1	642,36	0,4001	0,0342	-0,2925
9	Mediana	ANOVA	1540,82	-1731,67	-611	-351,67	-1178,13	-0,0419	-0,2877	0,1976
10	Mediana	ANCOVA	-1540,82	1731,67	611	351,67	1178,13	0,0419	0,2877	-0,1976
11	McQuitty	ANOVA	-850,75	2241,3	886,02	998,51	172,06	0,0507	0,3391	-0,1997
12	McQuitty	ANCOVA	850,75	-2241,3	-886,02	-998,51	-172,06	-0,0507	-0,3391	0,1997
13	Ward	ANOVA	-735,59	5274,25	-2427,97	1536,17	-2421,22	-0,6039	0,9894	-0,4608
14	Ward	ANCOVA	735,59	-5274,25	2427,97	-1536,17	2421,22	0,6039	-0,9894	0,4608
15	k-médias	ANOVA	3686,89	414,4	-379,5	-6065,38	2927,28	-0,3228	-0,4171	0,8493
16	k-médias	ANCOVA	-3686,89	-414,4	379,5	6065,38	-2927,28	0,3228	0,4171	-0,8493

Tabela C.12. Pesos para WLS dos clusters e dos escores fatoriais

Teste	Parâmetros		Pesos (w)							
	Ligação	Análise	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	FC1	FC2	FC3
1	Único	ANOVA	0,0000000	0,0000003	0,0000003	0,0000002	0,0000007	8,8170	2,8080	0,8084
2	Único	ANCOVA	0,0000000	0,0000003	0,0000003	0,0000002	0,0000007	8,8170	2,8080	0,8084
3	Centroide	ANOVA	0,0000002	0,0000010	0,0000973	0,0000077	0,0000046	300,583	19,499	11,679
4	Centroide	ANCOVA	0,0000002	0,0000010	0,0000973	0,0000077	0,0000046	300,583	19,499	11,679
5	Completa	ANOVA	0,0000189	0,0000000	0,0000001	0,0000006	0,0000068	1,1720	1,0850	50,448
6	Completa	ANCOVA	0,0000189	0,0000000	0,0000001	0,0000006	0,0000068	1,1720	1,0850	50,448
7	Média	ANOVA	0,0000002	0,0000007	0,0000001	0,0000295	0,0000024	6,2460	853,826	11,6893
8	Média	ANCOVA	0,0000002	0,0000007	0,0000001	0,0000295	0,0000024	6,2460	853,826	11,6893
9	Mediana	ANOVA	0,0000004	0,0000003	0,0000027	0,0000081	0,0000007	570,699	12,078	25,6237
10	Mediana	ANCOVA	0,0000004	0,0000003	0,0000027	0,0000081	0,0000007	570,699	12,078	25,6237
11	McQuitty	ANOVA	0,0000014	0,0000002	0,0000013	0,0000010	0,0000338	389,720	8,6940	25,0777
12	McQuitty	ANCOVA	0,0000014	0,0000002	0,0000013	0,0000010	0,0000338	389,724	8,6940	25,0777
13	Ward	ANOVA	0,0000018	0,0000000	0,0000002	0,0000004	0,0000002	2,7420	1,0210	4,7089
14	Ward	ANCOVA	0,0000018	0,0000000	0,0000002	0,0000004	0,0000002	2,7420	1,0210	4,7089
15	k-médias	ANOVA	0,0000001	0,0000058	0,0000069	0,0000000	0,0000001	9,5950	5,7490	1,3865
16	k-médias	ANCOVA	0,0000001	0,0000058	0,0000069	0,0000000	0,0000001	9,5950	5,7490	1,3865

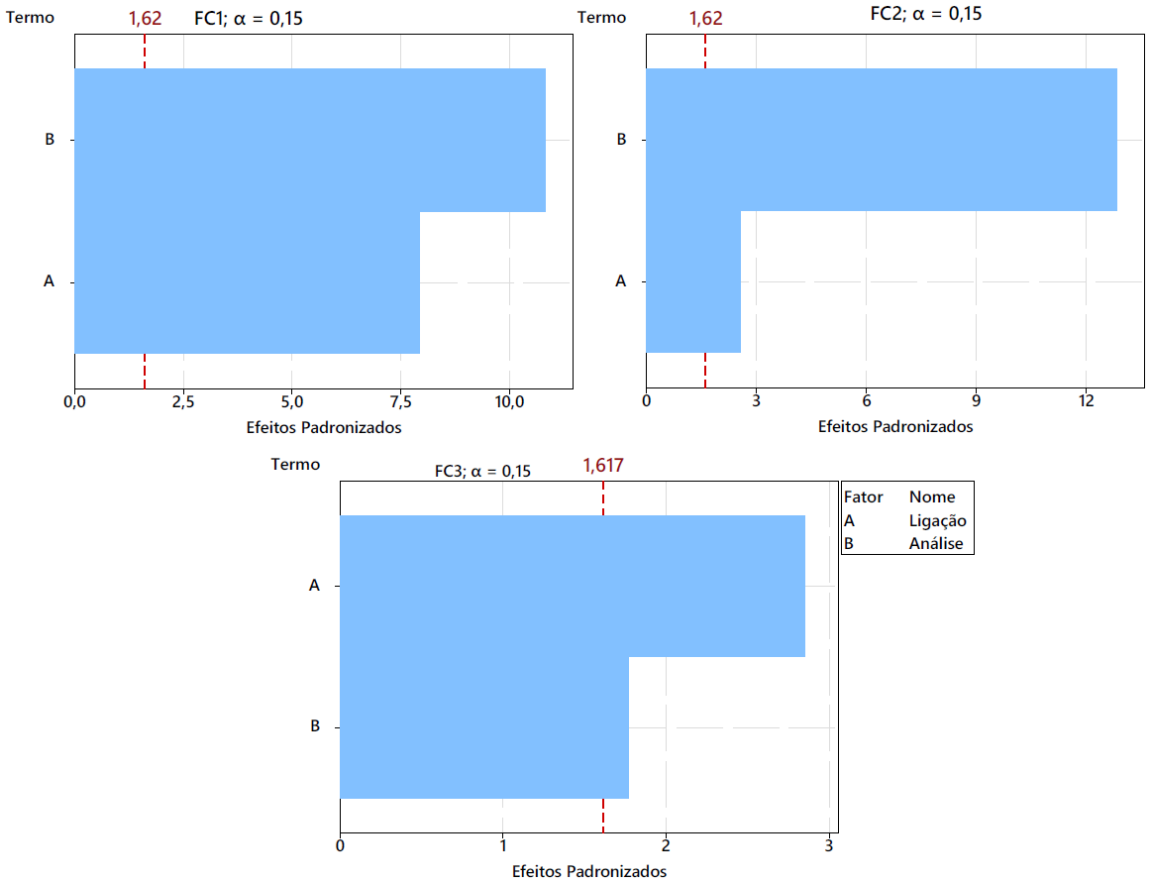


Figura C.2. Gráfico de Pareto dos escores fatoriais (DOE multiníveis)

Tabela C.13. Vetores e matrizes para estimar as elipses TNE×NEMV (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 155,81 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 327,88 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 237,71 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 394,61 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 83,79 \end{bmatrix}$
Σ	$\begin{bmatrix} 1181,74 & 1214,77 \\ 1214,77 & 1248,73 \end{bmatrix}$	$\begin{bmatrix} 18774,34 & 18648,39 \\ 18648,39 & 18523,4 \end{bmatrix}$	$\begin{bmatrix} 1128,89 & 998,19 \\ 998,19 & 882,63 \end{bmatrix}$	$\begin{bmatrix} 8882,23 & 8805,09 \\ 8805,09 & 8728,62 \end{bmatrix}$	$\begin{bmatrix} 2680,51 & 2588,26 \\ 2588,26 & 2499,18 \end{bmatrix}$
Λ	$\begin{bmatrix} 2430,48 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 37297,6 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 2011,52 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 17610,9 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 5179,7 \\ 0,00 \end{bmatrix}$
P	$\begin{bmatrix} 0,697 & 0,716 \\ 0,716 & -0,697 \end{bmatrix}$	$\begin{bmatrix} 0,7095 & -0,7047 \\ 0,7047 & 0,7095 \end{bmatrix}$	$\begin{bmatrix} 0,7491 & -0,6624 \\ 0,6624 & 0,7491 \end{bmatrix}$	$\begin{bmatrix} 0,7102 & -0,704 \\ 0,704 & 0,7102 \end{bmatrix}$	$\begin{bmatrix} 0,719 & -0,694 \\ 0,694 & 0,719 \end{bmatrix}$

Tabela C.14. Vetores e matrizes para estimar as elipses TNE×MVFR (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 213,75 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 181,5 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 340 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 212 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 132,5 \end{bmatrix}$
Σ	$\begin{bmatrix} 2708,43 & 244,3 \\ 244,3 & 223,5 \end{bmatrix}$	$\begin{bmatrix} 569,53 & 431,47 \\ 431,47 & 3315,34 \end{bmatrix}$	$\begin{bmatrix} 4890,91 & 276 \\ 276 & 157,97 \end{bmatrix}$	$\begin{bmatrix} 1375,8 & 460,34 \\ 460,34 & 1562,25 \end{bmatrix}$	$\begin{bmatrix} 215,22 & 97,42 \\ 97,42 & 447,3 \end{bmatrix}$
Λ	$\begin{bmatrix} 2732,22 \\ 199,71 \end{bmatrix}$	$\begin{bmatrix} 3381,55 \\ 503,33 \end{bmatrix}$	$\begin{bmatrix} 4906,96 \\ 141,93 \end{bmatrix}$	$\begin{bmatrix} 1938,72 \\ 999,34 \end{bmatrix}$	$\begin{bmatrix} 482,78 \\ 179,75 \end{bmatrix}$
P	$\begin{bmatrix} 0,995 & -0,096 \\ 0,096 & 0,995 \end{bmatrix}$	$\begin{bmatrix} 0,151 & 0,988 \\ 0,988 & -0,151 \end{bmatrix}$	$\begin{bmatrix} 0,998 & -0,058 \\ 0,058 & 0,998 \end{bmatrix}$	$\begin{bmatrix} 0,633 & 0,774 \\ 0,774 & -0,633 \end{bmatrix}$	$\begin{bmatrix} 0,342 & 0,939 \\ 0,939 & -0,342 \end{bmatrix}$

Tabela C.15. Vetores e matrizes para estimar as elipses TNE×EVAHV (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 1236,57 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 1285,89 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 1423,12 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 1352,75 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 1539,65 \end{bmatrix}$
Σ	$\begin{bmatrix} 1138,77 & -77,768 \\ -77,768 & 210,07 \end{bmatrix}$	$\begin{bmatrix} 388,31 & -174,9 \\ -174,9 & 3116,21 \end{bmatrix}$	$\begin{bmatrix} 373,02 & -37,42 \\ -37,42 & 148,48 \end{bmatrix}$	$\begin{bmatrix} 486,9 & -134,44 \\ -134,44 & 1468,42 \end{bmatrix}$	$\begin{bmatrix} 37,33 & -19,919 \\ -19,919 & 420,439 \end{bmatrix}$
Λ	$\begin{bmatrix} 1145,25 \\ 203,61 \end{bmatrix}$	$\begin{bmatrix} 3127,39 \\ 377,14 \end{bmatrix}$	$\begin{bmatrix} 379,1 \\ 142,41 \end{bmatrix}$	$\begin{bmatrix} 1486,51 \\ 468,83 \end{bmatrix}$	$\begin{bmatrix} 421,47 \\ 36,3 \end{bmatrix}$
P	$\begin{bmatrix} 0,996 & 0,082 \\ -0,082 & 0,996 \end{bmatrix}$	$\begin{bmatrix} -0,063 & 0,998 \\ 0,998 & 0,063 \end{bmatrix}$	$\begin{bmatrix} 0,987 & 0,16 \\ -0,16 & 0,987 \end{bmatrix}$	$\begin{bmatrix} -0,133 & 0,991 \\ 0,991 & 0,133 \end{bmatrix}$	$\begin{bmatrix} -0,051 & 0,998 \\ 0,998 & 0,051 \end{bmatrix}$

Tabela C.16. Vetores e matrizes para estimar as elipses TNE×MNE (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 111 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 237,5 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 116 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 144 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 195,25 \end{bmatrix}$
Σ	$\begin{bmatrix} 156,41 & 6,551 \\ 6,551 & 9,609 \end{bmatrix}$	$\begin{bmatrix} 108,245 & 20,991 \\ 20,991 & 142,533 \end{bmatrix}$	$\begin{bmatrix} 69,401 & 3,669 \\ 3,669 & 6,792 \end{bmatrix}$	$\begin{bmatrix} 137,983 & 16,269 \\ 16,269 & 67,165 \end{bmatrix}$	$\begin{bmatrix} 191,978 & 10,268 \\ 10,268 & 19,231 \end{bmatrix}$
Λ	$\begin{bmatrix} 156,7 \\ 9,32 \end{bmatrix}$	$\begin{bmatrix} 152,49 \\ 98,29 \end{bmatrix}$	$\begin{bmatrix} 69,61 \\ 6,58 \end{bmatrix}$	$\begin{bmatrix} 141,54 \\ 63,61 \end{bmatrix}$	$\begin{bmatrix} 192,59 \\ 18,62 \end{bmatrix}$
P	$\begin{bmatrix} 0,999 & -0,044 \\ 0,044 & 0,999 \end{bmatrix}$	$\begin{bmatrix} 0,428 & 0,903 \\ 0,903 & -0,428 \end{bmatrix}$	$\begin{bmatrix} 0,998 & -0,058 \\ 0,058 & 0,998 \end{bmatrix}$	$\begin{bmatrix} 0,976 & -0,213 \\ 0,213 & 0,976 \end{bmatrix}$	$\begin{bmatrix} 0,998 & -0,059 \\ 0,059 & 0,998 \end{bmatrix}$

Tabela C.17. Vetores e matrizes para estimar as elipses TNE×EMVVA (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 96,24 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 182,14 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 79,53 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 186,13 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 71,2 \end{bmatrix}$
Σ	$\begin{bmatrix} 1272,72 & 1510,43 \\ 1510,43 & 3535,97 \end{bmatrix}$	$\begin{bmatrix} 18879,14 & 4147,85 \\ 4147,85 & 1797,65 \end{bmatrix}$	$\begin{bmatrix} 899,58 & 737,33 \\ 737,33 & 1192,14 \end{bmatrix}$	$\begin{bmatrix} 8896,24 & 3190,48 \\ 3190,48 & 2257,08 \end{bmatrix}$	$\begin{bmatrix} 2547,18 & 1646,28 \\ 1646,28 & 2098,89 \end{bmatrix}$
Λ	$\begin{bmatrix} 4291,67 \\ 517,03 \end{bmatrix}$	$\begin{bmatrix} 19833,1 \\ 843,7 \end{bmatrix}$	$\begin{bmatrix} 1797,56 \\ 294,16 \end{bmatrix}$	$\begin{bmatrix} 10180,9 \\ 972,4 \end{bmatrix}$	$\begin{bmatrix} 3984,51 \\ 661,56 \end{bmatrix}$
P	$\begin{bmatrix} 0,447 & 0,894 \\ 0,894 & -0,447 \end{bmatrix}$	$\begin{bmatrix} 0,974 & -0,224 \\ 0,224 & 0,974 \end{bmatrix}$	$\begin{bmatrix} 0,634 & 0,772 \\ 0,772 & -0,634 \end{bmatrix}$	$\begin{bmatrix} 0,927 & -0,373 \\ 0,373 & 0,927 \end{bmatrix}$	$\begin{bmatrix} 0,753 & -0,657 \\ 0,657 & 0,753 \end{bmatrix}$

Tabela C.18. Vetores e matrizes para estimar as elipses TNE×NEMV (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 155,81 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 327,88 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 237,71 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 394,61 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 83,79 \end{bmatrix}$
Σ	$\begin{bmatrix} 1272,72 & 1216,69 \\ 1216,69 & 1163,13 \end{bmatrix}$	$\begin{bmatrix} 18879,14 & 18677,81 \\ 18677,81 & 18478,62 \end{bmatrix}$	$\begin{bmatrix} 899,58 & 999,77 \\ 999,77 & 1111,12 \end{bmatrix}$	$\begin{bmatrix} 8896,24 & 8818,97 \\ 8818,97 & 8742,38 \end{bmatrix}$	$\begin{bmatrix} 2547,18 & 2592,34 \\ 2592,34 & 2638,312 \end{bmatrix}$
Λ	$\begin{bmatrix} 2435,86 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 37357,8 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 2010,7 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 17638,6 \\ 0,00 \end{bmatrix}$	$\begin{bmatrix} 5185,49 \\ 0,00 \end{bmatrix}$
P	$\begin{bmatrix} 0,722 & -0,691 \\ 0,691 & 0,722 \end{bmatrix}$	$\begin{bmatrix} 0,7109 & -0,7033 \\ 0,7033 & 0,7109 \end{bmatrix}$	$\begin{bmatrix} 0,668 & 0,743 \\ 0,743 & -0,668 \end{bmatrix}$	$\begin{bmatrix} 0,7102 & -0,704 \\ 0,704 & 0,7102 \end{bmatrix}$	$\begin{bmatrix} 0,7009 & 0,7133 \\ 0,7133 & -0,7009 \end{bmatrix}$

Tabela C.19. Vetores e matrizes para estimar as elipses TNE×MVFR (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 213,75 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 181,5 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 340 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 212 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 132,5 \end{bmatrix}$
Σ	$\begin{bmatrix} 1272,72 & 1579,71 \\ 1579,71 & 19886,92 \end{bmatrix}$	$\begin{bmatrix} 18879,14 & 2790,02 \\ 2790,02 & 4181,89 \end{bmatrix}$	$\begin{bmatrix} 899,58 & 1784,71 \\ 1784,71 & 35911,99 \end{bmatrix}$	$\begin{bmatrix} 8896,24 & 2976,69 \\ 2976,69 & 10101,85 \end{bmatrix}$	$\begin{bmatrix} 2547,18 & 629,99 \\ 629,99 & 1580,33 \end{bmatrix}$
Λ	$\begin{bmatrix} 20020 \\ 1139,60 \end{bmatrix}$	$\begin{bmatrix} 19391 \\ 3670,1 \end{bmatrix}$	$\begin{bmatrix} 36002,7 \\ 808,8 \end{bmatrix}$	$\begin{bmatrix} 12536,2 \\ 6461,9 \end{bmatrix}$	$\begin{bmatrix} 2857,85 \\ 1269,66 \end{bmatrix}$
P	$\begin{bmatrix} 0,084 & 0,996 \\ 0,996 & -0,084 \end{bmatrix}$	$\begin{bmatrix} 0,9836 & -0,1804 \\ 0,1804 & 0,9836 \end{bmatrix}$	$\begin{bmatrix} 0,0508 & 0,9987 \\ 0,9987 & -0,0508 \end{bmatrix}$	$\begin{bmatrix} 0,633 & 0,774 \\ 0,774 & -0,633 \end{bmatrix}$	$\begin{bmatrix} 0,8969 & -0,4423 \\ 0,4423 & 0,8969 \end{bmatrix}$

Tabela C.20. Vetores e matrizes para estimar as elipses TNE×EVAHV (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 1236,57 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 1285,89 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 1423,12 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 1352,75 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 1539,65 \end{bmatrix}$
Σ	$\begin{bmatrix} 1272,72 & -1420,907 \\ -1420,907 & 62748,647 \end{bmatrix}$	$\begin{bmatrix} 18879,14 & -3195,65 \\ -3195,65 & 21396,46 \end{bmatrix}$	$\begin{bmatrix} 899,58 & -683,7 \\ -683,7 & 20554,18 \end{bmatrix}$	$\begin{bmatrix} 8896,24 & -2456,44 \\ -2456,44 & 26829,457 \end{bmatrix}$	$\begin{bmatrix} 2547,181 & -363,94 \\ -363,94 & 2056,94 \end{bmatrix}$
Λ	$\begin{bmatrix} 62781,5 \\ 1239,9 \end{bmatrix}$	$\begin{bmatrix} 23572,4 \\ 16703,2 \end{bmatrix}$	$\begin{bmatrix} 20577,9 \\ 875,8 \end{bmatrix}$	$\begin{bmatrix} 27159,8 \\ 8565,9 \end{bmatrix}$	$\begin{bmatrix} 2740,86 \\ 1863,27 \end{bmatrix}$
P	$\begin{bmatrix} -0,023 & 0,999 \\ 0,999 & 0,0231 \end{bmatrix}$	$\begin{bmatrix} -0,5628 & 0,8266 \\ 0,8266 & 0,5628 \end{bmatrix}$	$\begin{bmatrix} -0,0347 & 0,999 \\ 0,999 & 0,0347 \end{bmatrix}$	$\begin{bmatrix} -0,133 & 0,991 \\ 0,991 & 0,133 \end{bmatrix}$	$\begin{bmatrix} 0,882 & 0,469 \\ -0,4698 & 0,882 \end{bmatrix}$

Tabela C.21. Vetores e matrizes para estimar as elipses TNE×MNE (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,63 \\ 111 \end{bmatrix}$	$\begin{bmatrix} 357,85 \\ 237,5 \end{bmatrix}$	$\begin{bmatrix} 270,87 \\ 116 \end{bmatrix}$	$\begin{bmatrix} 426,12 \\ 144 \end{bmatrix}$	$\begin{bmatrix} 119,66 \\ 195,25 \end{bmatrix}$
Σ	$\begin{bmatrix} 1272,7 & 422,4 \\ 422,4 & 49,08,6 \end{bmatrix}$	$\begin{bmatrix} 18879,1 & 1353,4 \\ 1353,4 & 3397,09 \end{bmatrix}$	$\begin{bmatrix} 899,5 & 236,5 \\ 236,5 & 2177,9 \end{bmatrix}$	$\begin{bmatrix} 8896,2 & 1048,9 \\ 1048,9 & 4330,3 \end{bmatrix}$	$\begin{bmatrix} 2547,1 & 662,05 \\ 662,05 & 6024,9 \end{bmatrix}$
Λ	$\begin{bmatrix} 4957,1 \\ 1224,29 \end{bmatrix}$	$\begin{bmatrix} 18996,6 \\ 3279,7 \end{bmatrix}$	$\begin{bmatrix} 2220,3 \\ 857,2 \end{bmatrix}$	$\begin{bmatrix} 9125,6 \\ 4100,9 \end{bmatrix}$	$\begin{bmatrix} 6146,6 \\ 2425,4 \end{bmatrix}$
P	$\begin{bmatrix} 0,113 & 0,993 \\ 0,993 & -0,113 \end{bmatrix}$	$\begin{bmatrix} 0,996 & -0,086 \\ 0,086 & 0,996 \end{bmatrix}$	$\begin{bmatrix} 0,176 & 0,984 \\ 0,984 & -0,176 \end{bmatrix}$	$\begin{bmatrix} 0,976 & -0,213 \\ 0,213 & 0,976 \end{bmatrix}$	$\begin{bmatrix} 0,1809 & 0,9835 \\ 0,9835 & -0,1809 \end{bmatrix}$

Tabela C.22. Valores ajustados para ANCOVA

Principal:	**TNE**				**NEMV**	**MVFR**	**EVAHV**	**MNE**
	Concomitantes:							
Subestação	*NEMV*	*MVFR*	*EVAHV*	*MNE*	*TNE*	*TNE*	*TNE*	*TNE*
Aracruz	208,6	192,472	198,999	190,342	181,834	224,757	1247,25	108,718
Baixo Guandu	134,113	152,246	169,709	191,591	103,618	191,667	1215,17	115,579
Barra do Sahy	203,664	254,84	146,259	190,699	167,949	218,883	1241,55	109,936
Ecoporanga	192,141	138,961	223,552	165,887	169,867	219,694	1242,34	109,768
Itarana	322,937	361,725	352,846	369,542	292,722	166,622	1271,47	240,585
Jaguaré	247,112	221,418	284,527	276,76	216,703	331,113	1414,51	117,843
João Neiva	426,129	426,129	426,129	426,129	394,61	212	1352,75	144
Juncado	497,848	358,773	365,576	352,227	466,789	240,263	1342,87	225,316
Linhares A	545,337	393,832	358,523	370,792	512,248	259,495	1361,52	221,328
Linhares C	294,628	320,321	257,213	264,979	258,718	348,887	1431,74	114,157
Montanha	113,803	107,671	115,805	100,431	77,369	129,783	1537,02	195,813
Nova Venécia	49,155	134,98	128,32	133,454	15,983	103,813	1511,84	201,198
Paulista	150,132	106,564	119,577	123,101	113,716	145,16	1551,92	192,625
Pinheiros	309,709	345,487	386,221	356,69	285,44	163,542	1268,49	241,224
Santa Tereza	289,071	365,784	358,504	344,194	260,411	152,953	1258,22	243,419
São Francisco	165,57	129,444	114,956	121,673	128,095	151,244	1557,82	191,364
Suíça	182,199	321,499	325,431	353,655	149,721	106,124	1212,82	253,129

Tabela C.23. Vetores e matrizes para estimar as elipses da análise com dados originais (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 225,36 \\ 137,67 \end{bmatrix}$	$\begin{bmatrix} 249,005 \\ 98,925 \end{bmatrix}$	$\begin{bmatrix} 124,04 \\ 25,59 \end{bmatrix}$	$\begin{bmatrix} 370,56 \\ 187,7 \end{bmatrix}$	$\begin{bmatrix} 543,97 \\ 184,19 \end{bmatrix}$
Σ	$\begin{bmatrix} 7464,7 & 2454,41 \\ 2454,41 & 1591,92 \end{bmatrix}$	$\begin{bmatrix} 3083,97 & 1083,04 \\ 1083,04 & 750,27 \end{bmatrix}$	$\begin{bmatrix} 1137,3 & 810,85 \\ 810,85 & 1140,36 \end{bmatrix}$	$\begin{bmatrix} 8,99 & 55,01 \\ 55,01 & 663,74 \end{bmatrix}$	$\begin{bmatrix} 5220,91 & 1830,88 \\ 1830,88 & 1266,53 \end{bmatrix}$
Λ	$\begin{bmatrix} 8355,39 \\ 701,24 \end{bmatrix}$	$\begin{bmatrix} 3509,14 \\ 325,11 \end{bmatrix}$	$\begin{bmatrix} 1949,69 \\ 327,98 \end{bmatrix}$	$\begin{bmatrix} 668,34 \\ 4,41 \end{bmatrix}$	$\begin{bmatrix} 5938,43 \\ 549,02 \end{bmatrix}$
P	$\begin{bmatrix} 0,94 & -0,341 \\ 0,341 & 0,94 \end{bmatrix}$	$\begin{bmatrix} 0,9308 & -0,3654 \\ 0,3654 & 0,9308 \end{bmatrix}$	$\begin{bmatrix} 0,706 & 0,707 \\ 0,707 & -0,706 \end{bmatrix}$	$\begin{bmatrix} 0,083 & 0,996 \\ 0,996 & -0,083 \end{bmatrix}$	$\begin{bmatrix} 0,931 & -0,364 \\ 0,364 & 0,931 \end{bmatrix}$

Tabela C.24. Vetores e matrizes para estimar as elipses da análise com PCA (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 179,58 \\ 126,35 \end{bmatrix}$	$\begin{bmatrix} 184,35 \\ 58,83 \end{bmatrix}$	$\begin{bmatrix} 158,79 \\ 113,57 \end{bmatrix}$	$\begin{bmatrix} 356,28 \\ 180,9 \end{bmatrix}$	$\begin{bmatrix} 485,05 \\ 185,16 \end{bmatrix}$
Σ	$\begin{bmatrix} 2487,96 & 353,67 \\ 353,67 & 99,17 \end{bmatrix}$	$\begin{bmatrix} 1138,28 & 748,26 \\ 748,26 & 970,29 \end{bmatrix}$	$\begin{bmatrix} 2854,67 & 534,51 \\ 534,51 & 197,42 \end{bmatrix}$	$\begin{bmatrix} 2470,94 & 1102,76 \\ 1102,76 & 970,82 \end{bmatrix}$	$\begin{bmatrix} 1,564 & 24,195 \\ 24,195 & 738,27 \end{bmatrix}$
Λ	$\begin{bmatrix} 2539,23 \\ 47,91 \end{bmatrix}$	$\begin{bmatrix} 1807,26 \\ 301,32 \end{bmatrix}$	$\begin{bmatrix} 2958,17 \\ 93,93 \end{bmatrix}$	$\begin{bmatrix} 3054,56 \\ 387,22 \end{bmatrix}$	$\begin{bmatrix} 739,07 \\ 0,77 \end{bmatrix}$
P	$\begin{bmatrix} 0,989 & -0,143 \\ 0,143 & 0,989 \end{bmatrix}$	$\begin{bmatrix} 0,7455 & -0,6665 \\ 0,6665 & 0,7455 \end{bmatrix}$	$\begin{bmatrix} 0,981 & -0,19 \\ 0,19 & 0,981 \end{bmatrix}$	$\begin{bmatrix} 0,883 & -0,467 \\ 0,467 & 0,883 \end{bmatrix}$	$\begin{bmatrix} 0,032 & 0,999 \\ 0,999 & -0,032 \end{bmatrix}$

Tabela C.25. Vetores e matrizes para estimar as elipses da análise com dados originais (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 225,36 \\ 137,67 \end{bmatrix}$	$\begin{bmatrix} 249,005 \\ 98,925 \end{bmatrix}$	$\begin{bmatrix} 124,04 \\ 25,59 \end{bmatrix}$	$\begin{bmatrix} 370,56 \\ 187,7 \end{bmatrix}$	$\begin{bmatrix} 543,97 \\ 184,19 \end{bmatrix}$
Σ	$\begin{bmatrix} 14810,81 & 5539,53 \\ 5539,53 & 4087,03 \end{bmatrix}$	$\begin{bmatrix} 6980,3 & 2444,3 \\ 2444,3 & 1688,5 \end{bmatrix}$	$\begin{bmatrix} 10609,6 & 1830,06 \\ 1830,06 & 622,69 \end{bmatrix}$	$\begin{bmatrix} 6175,3 & 124,1 \\ 124,1 & 4,92 \end{bmatrix}$	$\begin{bmatrix} 11783,5 & 4132,2 \\ 4132,2 & 2858,5 \end{bmatrix}$
Λ	$\begin{bmatrix} 17158,4 \\ 1739,4 \end{bmatrix}$	$\begin{bmatrix} 7936,63 \\ 732,22 \end{bmatrix}$	$\begin{bmatrix} 10934,4 \\ 297,9 \end{bmatrix}$	$\begin{bmatrix} 6177,8 \\ 2,43 \end{bmatrix}$	$\begin{bmatrix} 13402,9 \\ 1239,1 \end{bmatrix}$
P	$\begin{bmatrix} 0,92 & -0,39 \\ 0,39 & 0,92 \end{bmatrix}$	$\begin{bmatrix} 0,931 & -0,364 \\ 0,364 & 0,931 \end{bmatrix}$	$\begin{bmatrix} 0,984 & -0,174 \\ 0,174 & 0,984 \end{bmatrix}$	$\begin{bmatrix} 0,999 & -0,02 \\ 0,02 & 0,999 \end{bmatrix}$	$\begin{bmatrix} 0,931 & -0,364 \\ 0,364 & 0,931 \end{bmatrix}$

Tabela C.26. Vetores e matrizes para estimar as elipses da análise com PCA (Ward-ANOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 179,58 \\ 126,35 \end{bmatrix}$	$\begin{bmatrix} 184,35 \\ 58,83 \end{bmatrix}$	$\begin{bmatrix} 158,79 \\ 113,57 \end{bmatrix}$	$\begin{bmatrix} 356,28 \\ 180,9 \end{bmatrix}$	$\begin{bmatrix} 485,05 \\ 185,16 \end{bmatrix}$
Σ	$\begin{bmatrix} 932,7 & 1193,08 \\ 1193,08 & 3010,3 \end{bmatrix}$	$\begin{bmatrix} 9125,5 & 2524,1 \\ 2524,1 & 1377,2 \end{bmatrix}$	$\begin{bmatrix} 1856,7 & 1803,08 \\ 1803,08 & 3454,01 \end{bmatrix}$	$\begin{bmatrix} 9130,5 & 3720 \\ 3720 & 2989,7 \end{bmatrix}$	$\begin{bmatrix} 6943,3 & 81,6 \\ 81,6 & 1,89 \end{bmatrix}$
Λ	$\begin{bmatrix} 3553,48 \\ 389,61 \end{bmatrix}$	$\begin{bmatrix} 9875,2 \\ 627,5 \end{bmatrix}$	$\begin{bmatrix} 4627,4 \\ 683,3 \end{bmatrix}$	$\begin{bmatrix} 10883.6 \\ 1236,7 \end{bmatrix}$	$\begin{bmatrix} 6944,3 \\ 0,93 \end{bmatrix}$
P	$\begin{bmatrix} 0,414 & 0,91 \\ 0,91 & -0,414 \end{bmatrix}$	$\begin{bmatrix} 0,958 & -0,284 \\ 0,284 & 0,958 \end{bmatrix}$	$\begin{bmatrix} 0,545 & 0838 \\ 0,838 & -0,545 \end{bmatrix}$	$\begin{bmatrix} 0,904 & -0,426 \\ 0,426 & 0,904 \end{bmatrix}$	$\begin{bmatrix} 0,999 & -0,011 \\ 0,011 & 0,999 \end{bmatrix}$

Tabela C.27. Dados da réplica R1 (Parte I)

Subestação	TNE	NEMV	MVFR	UNE	FKVAr
Aracruz	211,3504	179,9854	235,5072	82,1770	6914,8910
Baixo Guandu	132,3513	106,0362	126,4007	64,2036	5718,1288
Barra do Sahy	195,3590	173,3543	400,8840	103,1979	4762,9778
Ecoporanga	200,4716	164,6311	90,7504	143,1840	4537,5213
Itarana	319,4567	290,6217	190,2780	130,8161	11892,3771
Jaguaré	249,2079	213,7515	205,6258	108,8020	7486,3752
João Neiva	425,5538	394,0773	211,7137	133,8190	6591,0868
Juncado	500,8071	469,3379	185,0083	215,1727	9049,3200
Linhares A	554,2647	523,6178	284,2794	200,7277	24759,8149
Linhares C	294,5477	263,4905	478,0081	129,0824	9378,6397
Montanha	112,3466	77,4000	99,2625	181,6505	6849,1153
Nova Venécia	51,1755	13,8291	173,8859	155,8977	10493,1127
Paulista	152,7435	116,1988	98,8557	190,5775	3668,8722
Pinheiros	319,5669	284,1924	150,1476	156,2347	10043,6563
Santa Tereza	290,2330	260,1060	203,3543	129,2252	4507,8543
São Francisco	164,2684	129,2354	158,8798	199,8488	8993,1963
Suíça	177,5178	153,2267	82,7844	75,8026	1196,8825

Tabela C.28. Dados da réplica R1 (Parte II)

Subestação	SAIFI1	SAIFI2	AREA	EMVVA	EVAHV
Aracruz	5,5787	6,5450	556,6126	76,5895	1346,1389
Baixo Guandu	5,0863	7,4154	1083,4811	84,1557	1129,4057
Barra do Sahy	9,5178	12,4339	253,2098	42,9095	944,4092
Ecoporanga	6,2147	9,5600	1512,4123	182,9234	1538,2197
Itarana	10,2365	11,7784	1292,0602	151,3655	1237,5505
Jaguaré	6,4786	7,3037	1252,3718	103,7629	1521,7325
João Neiva	10,5611	12,2089	949,0845	185,8855	1350,9231
Juncado	14,7489	16,1283	753,5308	255,0749	1350,6110
Linhares A	10,7376	11,8982	2096,0080	187,6766	1315,3168
Linhares C	10,6273	11,7760	542,4091	55,5887	1332,9266
Montanha	10,1772	11,0840	1233,6666	77,4000	1499,8570
Nova Venécia	8,6889	11,1105	1334,4121	7,9478	1602,8480
Paulista	10,2546	13,0251	209,7382	119,7925	1568,4429
Pinheiros	10,4016	11,3284	1326,1552	192,0219	1518,2153
Santa Tereza	13,1318	15,4252	791,2116	128,1311	1293,0029
São Francisco	8,1290	10,7879	976,6881	81,2801	1503,5625
Suíça	13,0749	15,3583	230,7909	184,6105	1042,5345

Tabela C.29. Dados da réplica R1 (Parte III)

Subestação	NEHV	3LG	1LG	L-L	R+
Aracruz	31,3650	3161,8088	3485,5059	2737,8960	0,4610
Baixo Guandu	26,3152	6505,6254	6653,0930	5633,8616	0,0100
Barra do Sahy	22,0047	6327,8144	6513,3721	5479,4090	0,0198
Ecoporanga	35,8405	1644,5994	1869,4588	1423,7733	0,6756
Itarana	28,8349	5623,1123	6367,3769	4869,9284	0,0991
Jaguaré	35,4564	7699,9864	8579,3860	6668,8630	0,0998
João Neiva	31,4765	8786,1184	3731,9532	7609,7093	0,0399
Juncado	31,4692	3168,2675	4056,1063	2743,9549	0,6636
Linhares A	30,6469	7573,6504	7870,1568	6558,8036	0,0306
Linhares C	31,0572	7599,7235	8809,8706	6581,1831	0,0303
Montanha	34,9467	2276,0901	2697,9559	1971,3541	0,3871
Nova Venécia	37,3464	6959,4321	7403,1409	6027,0441	0,0500
Paulista	36,5447	2807,7063	3172,5553	2431,6469	0,1325
Pinheiros	35,3744	4373,5558	5291,6880	3788,1831	0,1420
Santa Tereza	30,1270	3833,6796	4053,0619	3319,7843	0,1302
São Francisco	35,0330	3365,4539	4207,8166	2914,7948	0,4197
Suíça	24,2911	2393,7649	2477,5467	2072,6014	0,0499

Tabela C.30. Dados da réplica R2 (Parte I)

Subestação	TNE	NEMV	MVFR	UNE	FKVAr
Aracruz	207,7332	176,9050	231,4765	80,7705	6796,5437
Baixo Guandu	132,5288	106,1783	126,5702	64,2896	5725,7936
Barra do Sahy	193,9826	172,1329	398,0597	102,4708	4729,4218
Ecoporanga	197,5827	162,2586	89,4426	141,1206	4472,1323
Itarana	320,7345	291,7843	191,0392	131,3394	11939,9485
Jaguaré	248,0745	212,7794	204,6906	108,3072	7452,3275
João Neiva	428,5148	396,8193	213,1868	134,7501	6636,9469
Juncado	491,1084	460,2486	181,4254	211,0056	8874,0683
Linhares A	542,0890	512,1154	278,0345	196,3183	24215,9085
Linhares C	294,1104	263,0994	477,2985	128,8907	9364,7179
Montanha	112,2150	77,3093	99,1463	181,4376	6841,0916
Nova Venécia	51,3727	13,8824	174,5560	156,4985	10533,5496
Paulista	149,7064	113,8883	96,8901	186,7881	3595,9207
Pinheiros	314,8973	280,0398	147,9536	153,9517	9896,8978
Santa Tereza	290,1800	260,0585	203,3172	129,2015	4507,0305
São Francisco	163,8362	128,8954	158,4618	199,3230	8969,5359
Suíça	176,8877	152,6828	82,4905	75,5335	1192,6338

Tabela C.31. Dados da réplica R2 (Parte II)

Subestação	SAIFI1	SAIFI2	AREA	EMVVA	EVAHV
Aracruz	5,4832	6,4330	547,0863	75,2787	1323,0999
Baixo Guandu	5,0931	7,4253	1084,9334	84,2685	1130,9196
Barra do Sahy	9,4507	12,3463	251,4259	42,6072	937,7557
Ecoporanga	6,1251	9,4222	1490,6174	180,2874	1516,0529
Itarana	10,2775	11,8255	1297,2286	151,9710	1242,5009
Jaguaré	6,4491	7,2705	1246,6760	103,2909	1514,8118
João Neiva	10,6346	12,2939	955,6882	187,1789	1360,3227
Juncado	14,4632	15,8160	738,9378	250,1351	1324,4547
Linhares A	10,5018	11,6369	2049,9644	183,5539	1286,4228
Linhares C	10,6116	11,7585	541,6039	55,5062	1330,9479
Montanha	10,1653	11,0710	1232,2214	77,3093	1498,0999
Nova Venécia	8,7223	11,1533	1339,5545	7,9784	1609,0248
Paulista	10,0507	12,7662	205,5678	117,4106	1537,2561
Pinheiros	10,2496	11,1628	1306,7774	189,2161	1496,0311
Santa Tereza	13,1294	15,4224	791,0670	128,1076	1292,7666
São Francisco	8,1076	10,7595	974,1185	81,0663	1499,6067
Suíça	13,0285	15,3038	229,9716	183,9552	1038,8337

Tabela C.32. Dados da réplica R2 (Parte III)

Subestação	NEHV	3LG	1LG	L-L	R+
Aracruz	30,8282	3107,6950	3425,8520	2691,0373	0,4531
Baixo Guandu	26,3504	6514,3459	6662,0111	5641,4135	0,0100
Barra do Sahy	21,8497	6283,2339	6467,4843	5440,8056	0,0197
Ecoporanga	35,3240	1620,8995	1842,5185	1403,2557	0,6659
Itarana	28,9503	5645,6056	6392,8474	4889,4089	0,0995
Jaguaré	35,2951	7664,9672	8540,3673	6638,5333	0,0994
João Neiva	31,6955	8847,2513	3757,9198	7662,6569	0,0402
Juncado	30,8598	3106,9099	3977,5546	2690,8147	0,6508
Linhares A	29,9737	7407,2777	7697,2707	6414,7244	0,0299
Linhares C	31,0111	7588,4424	8796,7931	6571,4139	0,0302
Montanha	34,9057	2273,4236	2694,7952	1969,0446	0,3867
Nova Venécia	37,4903	6986,2514	7431,6701	6050,2703	0,0502
Paulista	35,8181	2751,8782	3109,4725	2383,2963	0,1299
Pinheiros	34,8575	4309,6491	5214,3655	3732,8299	0,1400
Santa Tereza	30,1215	3832,9790	4052,3212	3319,1775	0,1302
São Francisco	34,9408	3356,5997	4196,7462	2907,1263	0,4186
Suíça	24,2048	2385,2676	2468,7520	2065,2442	0,0497

Tabela C.33. Dados da réplica R3 (Parte I)

Subestação	TNE	NEMV	MVFR	UNE	FKVAr
Aracruz	213,8886	182,1469	238,3354	83,1639	6997,9341
Baixo Guandu	132,9096	106,4834	126,9339	64,4743	5742,2462
Barra do Sahy	198,4829	176,1263	407,2943	104,8480	4839,1404
Ecoporanga	198,8666	163,3131	90,0239	142,0377	4501,1946
Itarana	321,5005	292,4811	191,4954	131,6531	11968,4627
Jaguaré	250,9114	215,2127	207,0314	109,5457	7537,5509
João Neiva	423,4414	392,1212	210,6628	133,1548	6558,3696
Juncado	497,8580	466,5741	183,9189	213,9056	8996,0315
Linhares A	542,5800	512,5792	278,2863	196,4961	24237,8427
Linhares C	293,2057	262,2900	475,8302	128,4942	9335,9086
Montanha	112,8374	77,7381	99,6962	182,4440	6879,0362
Nova Venécia	50,9758	13,7752	173,2073	155,2893	10452,1676
Paulista	150,8889	114,7879	97,6554	188,2635	3624,3242
Pinheiros	314,1173	279,3462	147,5871	153,5704	9872,3831
Santa Tereza	291,0947	260,8783	203,9581	129,6088	4521,2383
São Francisco	163,7299	128,8117	158,3590	199,1937	8963,7165
Suíça	180,3241	155,6490	84,0931	77,0009	1215,8034

Tabela C.34. Dados da réplica R3 (Parte II)

Subestação	SAIFI1	SAIFI2	AREA	EMVVA	EVAHV
Aracruz	5,6457	6,6236	563,2972	77,5093	1362,3051
Baixo Guandu	5,1077	7,4466	1088,0509	84,5107	1134,1692
Barra do Sahy	9,6700	12,6327	257,2588	43,5956	959,5108
Ecoporanga	6,1649	9,4834	1500,3042	181,4590	1525,9050
Itarana	10,3020	11,8538	1300,3266	152,3339	1245,4681
Jaguaré	6,5229	7,3536	1260,9328	104,4722	1532,1349
João Neiva	10,5087	12,1483	944,3734	184,9628	1344,2173
Juncado	14,6620	16,0333	749,0935	253,5729	1342,6577
Linhares A	10,5113	11,6474	2051,8212	183,7202	1287,5880
Linhares C	10,5789	11,7224	539,9378	55,3354	1326,8535
Montanha	10,2217	11,1324	1239,0560	77,7381	1506,4092
Nova Venécia	8,6549	11,0672	1329,2051	7,9168	1596,5935
Paulista	10,1300	12,8670	207,1915	118,3380	1549,3986
Pinheiros	10,2243	11,1352	1303,5405	188,7474	1492,3254
Santa Tereza	13,1708	15,4710	793,5607	128,5115	1296,8419
São Francisco	8,1024	10,7526	973,4865	81,0137	1498,6338
Suíça	13,2816	15,6011	234,4393	187,5290	1059,0154

Tabela C.35. Dados da réplica R3 (Parte III)

Subestação	NEHV	3LG	1LG	L-L	R+
Aracruz	31,7417	3199,7800	3527,3645	2770,7762	0,4665
Baixo Guandu	26,4261	6533,0644	6681,1539	5657,6237	0,0101
Barra do Sahy	22,3566	6428,9997	6617,5245	5567,0278	0,0202
Ecoporanga	35,5536	1631,4330	1854,4922	1412,3748	0,6702
Itarana	29,0194	5659,0881	6408,1144	4901,0855	0,0997
Jaguaré	35,6987	7752,6223	8638,0333	6714,4503	0,1005
João Neiva	31,3203	8742,5054	3713,4284	7571,9358	0,0397
Juncado	31,2839	3149,6106	4032,2212	2727,7967	0,6597
Linhares A	30,0008	7413,9870	7704,2427	6420,5347	0,0299
Linhares C	30,9157	7565,0975	8769,7309	6551,1978	0,0301
Montanha	35,0993	2286,0333	2709,7421	1979,9661	0,3888
Nova Venécia	37,2006	6932,2757	7374,2531	6003,5260	0,0498
Paulista	36,1010	2773,6148	3134,0337	2402,1216	0,1309
Pinheiros	34,7712	4298,9741	5201,4495	3723,5837	0,1396
Santa Tereza	30,2164	3845,0620	4065,0956	3329,6408	0,1306
São Francisco	34,9182	3354,4219	4194,0233	2905,2401	0,4183
Suíça	24,6751	2431,6068	2516,7130	2105,3662	0,0507

Tabela C.36. Dados da réplica R4 (Parte I)

Subestação	TNE	NEMV	MVFR	UNE	FKVAr
Aracruz	215,0290	183,1180	239,6060	83,6070	7035,2
Baixo Guandu	134,0270	107,3790	128,0010	65,0160	5790,5
Barra do Sahy	197,8200	175,5380	405,9340	104,4980	4823,0
Ecoporanga	200,1050	164,3300	90,5850	142,9220	4529,2
Itarana	321,1530	292,1650	191,2880	131,5110	11955,5
Jaguaré	248,1980	212,8860	204,7930	108,3610	7456,0
João Neiva	421,4080	390,2380	209,6510	132,5150	6526,9
Juncado	505,5440	473,7770	186,7580	217,2080	9134,9
Linhares A	551,3850	520,8970	282,8020	199,6850	24631,2
Linhares C	292,8610	261,9820	475,2710	128,3430	9324,9
Montanha	113,3160	78,0680	100,1190	183,2180	6908,2
Nova Venécia	50,8360	13,7370	172,7310	154,8620	10423,4
Paulista	148,2600	112,7880	95,9540	184,9840	3561,2
Pinheiros	317,0710	281,9730	148,9750	155,0150	9965,2
Santa Tereza	288,5590	258,6060	202,1810	128,4800	4481,9
São Francisco	164,2100	129,1890	158,8230	199,7780	8990,0
Suíça	174,6190	150,7250	81,4330	74,5650	1177,3

Tabela C.37. Dados da réplica R4 (Parte II)

Subestação	SAIFI1	SAIFI2	AREA	EMVVA	EVAHV
Aracruz	5,6758	6,6590	566,3000	77,9230	1369,57
Baixo Guandu	5,1507	7,5093	1097,2000	85,2210	1143,71
Barra do Sahy	9,6377	12,5905	256,4000	43,4500	956,31
Ecoporanga	6,2033	9,5425	1509,6500	182,5890	1535,41
Itarana	10,2909	11,8409	1298,9200	152,1690	1244,12
Jaguaré	6,4523	7,2741	1247,3000	103,3430	1515,57
João Neiva	10,4582	12,0900	939,8400	184,0750	1337,76
Juncado	14,8884	16,2809	760,6600	257,4880	1363,39
Linhares A	10,6819	11,8364	2085,1200	186,7020	1308,48
Linhares C	10,5665	11,7086	539,3000	55,2700	1325,29
Montanha	10,2651	11,1796	1244,3100	78,0680	1512,80
Nova Venécia	8,6311	11,0367	1325,5500	7,8950	1592,20
Paulista	9,9536	12,6428	203,5800	116,2760	1522,41
Pinheiros	10,3204	11,2399	1315,8000	190,5220	1506,36
Santa Tereza	13,0561	15,3362	786,6500	127,3920	1285,55
São Francisco	8,1261	10,7841	976,3400	81,2510	1503,03
Suíça	12,8614	15,1075	227,0200	181,5960	1025,51

Tabela C.38. Dados da réplica R4 (Parte III)

Subestação	NEHV	3LG	1LG	L-L	R+
Aracruz	31,9110	3216,8400	3546,1700	2785,5500	0,4690
Baixo Guandu	26,6483	6587,9900	6737,3300	5705,1900	0,0102
Barra do Sahy	22,2819	6407,5300	6595,4200	5548,4300	0,0201
Ecoporanga	35,7750	1641,6000	1866,0400	1421,1700	0,6744
Itarana	28,9880	5652,9700	6401,1900	4895,7900	0,0996
Jaguaré	35,3127	7668,8000	8544,6300	6641,8500	0,0994
João Neiva	31,1699	8700,5300	3695,6000	7535,5800	0,0396
Juncado	31,7669	3198,2400	4094,4700	2769,9100	0,6699
Linhares A	30,4877	7534,3000	7829,2700	6524,7300	0,0304
Linhares C	30,8793	7556,2000	8759,4200	6543,4900	0,0301
Montanha	35,2482	2295,7300	2721,2400	1988,3700	0,3905
Nova Venécia	37,0983	6913,2100	7353,9800	5987,0200	0,0496
Paulista	35,4721	2725,3000	3079,4400	2360,2700	0,1286
Pinheiros	35,0982	4339,4000	5250,3600	3758,6000	0,1409
Santa Tereza	29,9532	3811,5700	4029,6800	3300,6400	0,1295
São Francisco	35,0205	3364,2600	4206,3200	2913,7600	0,4195
Suíça	23,8944	2354,6800	2437,0900	2038,7600	0,0491

Tabela C.39. Matriz experimental *simplex-lattice* para as réplicas

Teste	Controle		Respostas			
	γ_1	γ_2	EQM_{R1}	EQM_{R2}	EQM_{R3}	EQM_{R4}
1	1	0	14,276	14,104	13,890	14,179
2	0,9	0,1	14,865	14,713	14,472	14,793
3	0,8	0,2	15,477	15,326	15,065	15,434
4	0,7	0,3	16,111	15,961	15,680	16,101
5	0,6	0,4	16,752	16,598	16,301	16,782
6	0,5	0,5	17,396	17,235	16,923	17,471
7	0,4	0,6	18,045	17,865	17,541	18,160
8	0,3	0,7	18,688	18,484	18,143	18,844
9	0,2	0,8	19,322	19,088	18,737	19,516
10	0,1	0,9	19,947	19,677	19,318	20,184
11	0	1	20,556	20,253	19,881	20,833
12	0,75	0,25	15,793	15,642	15,371	15,765
13	0,25	0,75	19,005	18,787	18,441	19,185

Tabela C.40. Escores de fatores com rotação otimizada para R1

F1	F2	F3	F4	F5	F6	F7
-1,2527	0,1250	0,3429	-1,9440	-0,6439	-1,0771	-0,2068
0,6792	-0,6247	1,4437	-1,5450	0,1920	1,0762	-0,0083
0,3209	-0,8776	1,9700	0,2571	-0,4356	-1,5110	1,0850
-1,7321	0,4506	-0,5498	-1,3873	0,0694	0,6454	0,2617
0,2408	0,2732	0,7387	0,1105	0,9130	0,3357	-0,6007
1,2238	0,3537	-1,0098	-1,4100	-0,4959	0,0056	-1,2810
1,6830	1,5532	-0,2611	-0,1254	-1,3065	1,0132	2,3450
-1,2766	1,8724	-0,3306	1,2403	-0,1803	-0,7377	0,3358
0,6864	1,2402	0,3875	0,3452	2,8839	-0,2917	0,1414
0,9482	-0,0836	-0,3023	0,2892	-0,6712	-2,4594	-0,8996
-1,0486	-1,2402	-0,4653	0,3461	0,7240	0,4561	0,8111
1,1566	-1,9049	-1,2224	0,3429	0,6911	0,2526	0,0920
-0,2130	-0,7166	-1,2958	0,9848	-1,1324	0,4426	0,0308
0,0547	0,6677	-0,8061	0,1171	0,2352	0,6029	-1,5320
-0,1148	0,0159	0,2808	1,2211	-0,4785	0,1627	-0,5970
-0,8509	-0,8850	-0,6704	-0,0149	0,5704	-0,3853	1,2151
-0,5049	-0,2193	1,7501	1,1721	-0,9347	1,4693	-1,1924

Tabela C.41. Escores de fatores com rotação otimizada para R2

F1	F2	F3	F4	F5	F6	F7
-1,2095	0,0737	0,4559	-1,9732	0,6556	-0,9801	-0,0942
0,7396	-0,6436	1,4156	-1,5517	-0,1590	1,0998	0,1024
0,3449	-0,8452	2,0055	0,1686	0,4788	-1,4617	-1,0872
-1,7509	0,4589	-0,5796	-1,3638	-0,0148	0,5946	0,3199
0,2740	0,2757	0,6567	0,1858	-0,9390	0,3172	0,6400
1,1769	0,3051	-1,0274	-1,3047	0,4251	-0,0774	1,5094
1,7121	1,7070	-0,4098	-0,1992	1,4228	0,9989	-1,9904
-1,3429	1,8699	-0,3019	1,2161	0,1615	-0,7730	-0,0440
0,6568	1,2233	0,5373	0,2284	-2,8859	-0,1790	-0,6490
0,9168	-0,0992	-0,3194	0,3781	0,5973	-2,5287	0,9239
-1,0439	-1,1801	-0,5440	0,3174	-0,6620	0,4466	-0,6898
1,1307	-1,8770	-1,3223	0,3823	-0,7095	0,1724	0,1731
-0,2154	-0,7505	-1,0788	0,8551	1,0860	0,5515	-0,9302
0,0244	0,5845	-0,7254	0,1794	-0,3204	0,5960	1,2733
-0,1222	0,0118	0,1923	1,3247	0,4604	0,0905	0,8587
-0,8398	-0,8100	-0,6971	-0,1021	-0,5057	-0,3487	-1,3808
-0,4513	-0,3044	1,7425	1,2588	0,9089	1,4810	1,0648

Tabela C.42. Escores de fatores com rotação otimizada para R3

F1	F2	F3	F4	F5	F6	F7
-1,2460	0,1134	0,3093	-1,9372	-0,6853	-1,0782	0,1507
0,6769	-0,6698	1,5049	-1,5782	0,1537	1,0774	0,0367
0,3412	-0,8502	1,9730	0,2233	-0,4966	-1,5341	-1,0873
-1,7586	0,4242	-0,5007	-1,3965	0,0475	0,6399	0,0222
0,2590	0,2783	0,6839	0,1560	0,9270	0,3151	0,6451
1,2541	0,3543	-1,1447	-1,2906	-0,4608	-0,0458	1,4697
1,6845	1,6262	-0,2709	-0,2286	-1,3532	1,0374	-2,1835
-1,2830	1,9067	-0,3635	1,2388	-0,1597	-0,7361	-0,1881
0,6487	1,2160	0,5206	0,2597	2,8781	-0,2188	-0,4104
0,9229	-0,0831	-0,2983	0,3385	-0,5916	-2,4557	0,8777
-1,0376	-1,2103	-0,5257	0,3369	0,6883	0,4469	-0,7380
1,1459	-1,8740	-1,2590	0,3620	0,7108	0,2091	0,0614
-0,1936	-0,7364	-1,1925	0,9278	-1,0855	0,5057	-0,5804
0,0292	0,5941	-0,7121	0,1440	0,3130	0,6117	1,3348
-0,1173	0,0257	0,2143	1,2673	-0,4650	0,1222	0,7500
-0,8402	-0,8515	-0,6426	-0,0721	0,5436	-0,3481	-1,3517
-0,4863	-0,2637	1,7039	1,2491	-0,9642	1,4515	1,1911

Tabela C.43. Escores de fatores com rotação otimizada para R4

F1	F2	F3	F4	F5	F6	F7
-1,2546	0,1347	0,2790	-1,9554	0,6174	-1,1152	-0,0081
0,6669	-0,5846	1,4057	-1,5798	-0,2212	1,0015	0,1012
0,3020	-0,8314	1,9300	0,2288	0,4329	-1,5908	-1,0521
-1,7718	0,4551	-0,5407	-1,3919	-0,0953	0,5975	0,0030
0,2450	0,2816	0,7070	0,1395	-0,9327	0,3144	0,6314
1,2313	0,3206	-1,0243	-1,3252	0,4915	0,0071	1,4874
1,6825	1,5986	-0,2610	-0,2183	1,3694	1,0026	-2,2415
-1,2330	1,9233	-0,4916	1,3522	0,2496	-0,7467	0,0433
0,6862	1,2243	0,4537	0,2941	-2,9061	-0,2570	-0,5056
0,9433	-0,1302	-0,2859	0,3489	0,6524	-2,4080	0,9203
-1,0257	-1,2061	-0,5885	0,3777	-0,6837	0,4390	-0,7124
1,1586	-1,8854	-1,2627	0,3841	-0,6435	0,2416	0,1449
-0,1919	-0,7747	-1,0483	0,8781	1,1080	0,5863	-0,6199
0,0801	0,6049	-0,7597	0,1701	-0,2454	0,6597	1,3791
-0,1229	-0,0019	0,2828	1,2460	0,4869	0,1698	0,7342
-0,8312	-0,8620	-0,6940	-0,0381	-0,5457	-0,3666	-1,3237
-0,5647	-0,2668	1,8985	1,0894	0,8655	1,4650	1,0185

Tabela C.44. Vetores e matrizes para estimar as elipses de R1 (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 184,88 \\ 96,64 \end{bmatrix}$	$\begin{bmatrix} 360,3 \\ 183,14 \end{bmatrix}$	$\begin{bmatrix} 271,87 \\ 79,67 \end{bmatrix}$	$\begin{bmatrix} 425,55 \\ 185,88 \end{bmatrix}$	$\begin{bmatrix} 120,13 \\ 71,6 \end{bmatrix}$
Σ	$\begin{bmatrix} 2194,94 & 231,69 \\ 231,69 & 47,83 \end{bmatrix}$	$\begin{bmatrix} 1126,14 & 655,07 \\ 655,07 & 745,36 \end{bmatrix}$	$\begin{bmatrix} 701,66 & 117,8 \\ 117,8 & 38,68 \end{bmatrix}$	$\begin{bmatrix} 1404,04 & 501,44 \\ 501,44 & 350,3 \end{bmatrix}$	$\begin{bmatrix} 1310,44 & 256,49 \\ 256,49 & 98,2 \end{bmatrix}$
Λ	$\begin{bmatrix} 2219,66 \\ 23,12 \end{bmatrix}$	$\begin{bmatrix} 1617,93 \\ 253,58 \end{bmatrix}$	$\begin{bmatrix} 721,97 \\ 18,38 \end{bmatrix}$	$\begin{bmatrix} 1604,52 \\ 149,83 \end{bmatrix}$	$\begin{bmatrix} 1362,49 \\ 46,17 \end{bmatrix}$
P	$\begin{bmatrix} ,994 & -0,106 \\ 0,106 & 0,994 \end{bmatrix}$	$\begin{bmatrix} 0,799 & -0,6 \\ 0,6 & 0,799 \end{bmatrix}$	$\begin{bmatrix} 0,985 & -0,169 \\ 0,169 & 0,985 \end{bmatrix}$	$\begin{bmatrix} 0,928 & -0,371 \\ 0,371 & 0,928 \end{bmatrix}$	$\begin{bmatrix} 0,98 & -0,198 \\ 0,198 & 0,98 \end{bmatrix}$

Tabela C.45. Vetores e matrizes para estimar as elipses de R2 (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 182,95 \\ 95,61 \end{bmatrix}$	$\begin{bmatrix} 355,98 \\ 181,15 \end{bmatrix}$	$\begin{bmatrix} 271,09 \\ 79,39 \end{bmatrix}$	$\begin{bmatrix} 428,51 \\ 187,17 \end{bmatrix}$	$\begin{bmatrix} 119,28 \\ 70,94 \end{bmatrix}$
Σ	$\begin{bmatrix} 1845,09 & 191,44 \\ 191,44 & 39,29 \end{bmatrix}$	$\begin{bmatrix} 894,62 & 530,64 \\ 530,64 & 622,61 \end{bmatrix}$	$\begin{bmatrix} 600,63 & 104,21 \\ 104,21 & 35,76 \end{bmatrix}$	$\begin{bmatrix} 1158,83 & 414,67 \\ 414,67 & 293,52 \end{bmatrix}$	$\begin{bmatrix} 1098,96 & 217,55 \\ 217,55 & 85,19 \end{bmatrix}$
Λ	$\begin{bmatrix} 1865,17 \\ 19,22 \end{bmatrix}$	$\begin{bmatrix} 1306,41 \\ 210,83 \end{bmatrix}$	$\begin{bmatrix} 619,25 \\ 17,16 \end{bmatrix}$	$\begin{bmatrix} 1325,46 \\ 126,9 \end{bmatrix}$	$\begin{bmatrix} 1143,69 \\ 40,48 \end{bmatrix}$
P	$\begin{bmatrix} 0,994 & -0,104 \\ 0,104 & 0,994 \end{bmatrix}$	$\begin{bmatrix} 0,79 & -0,613 \\ 0,613 & 0,79 \end{bmatrix}$	$\begin{bmatrix} 0,984 & -0,175 \\ 0,175 & 0,984 \end{bmatrix}$	$\begin{bmatrix} 0,927 & -0,372 \\ 0,379 & 0,927 \end{bmatrix}$	$\begin{bmatrix} 0,979 & -0,201 \\ 0,201 & 0,979 \end{bmatrix}$

Tabela C.46. Vetores e matrizes para estimar as elipses de R3 (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 186,03 \\ 96,76 \end{bmatrix}$	$\begin{bmatrix} 357,91 \\ 182,4 \end{bmatrix}$	$\begin{bmatrix} 272,05 \\ 79,9 \end{bmatrix}$	$\begin{bmatrix} 423,44 \\ 184,96 \end{bmatrix}$	$\begin{bmatrix} 119,6 \\ 71,25 \end{bmatrix}$
Σ	$\begin{bmatrix} 1809,06 & 200,43 \\ 200,43 & 44,3 \end{bmatrix}$	$\begin{bmatrix} 922,51 & 539,97 \\ 539,7 & 630,52 \end{bmatrix}$	$\begin{bmatrix} 622,73 & 97,31 \\ 97,31 & 30,34 \end{bmatrix}$	$\begin{bmatrix} 1162,18 & 416,68 \\ 416,68 & 298,03 \end{bmatrix}$	$\begin{bmatrix} 1094,58 & 218,3 \\ 218,3 & 86,86 \end{bmatrix}$
Λ	$\begin{bmatrix} 1831,54 \\ 21,82 \end{bmatrix}$	$\begin{bmatrix} 1335,88 \\ 217,16 \end{bmatrix}$	$\begin{bmatrix} 638,31 \\ 14,76 \end{bmatrix}$	$\begin{bmatrix} 1130,37 \\ 129,85 \end{bmatrix}$	$\begin{bmatrix} 1139,84 \\ 41,6 \end{bmatrix}$
P	$\begin{bmatrix} 0,993 & -0,111 \\ 0,111 & 0,993 \end{bmatrix}$	$\begin{bmatrix} 0,794 & -0,607 \\ 0,607 & 0,794 \end{bmatrix}$	$\begin{bmatrix} 0,987 & -0,158 \\ 0,158 & 0,987 \end{bmatrix}$	$\begin{bmatrix} 0,927 & -0,374 \\ 0,374 & 0,927 \end{bmatrix}$	$\begin{bmatrix} 0,979 & -0,203 \\ 0,203 & 0,979 \end{bmatrix}$

Tabela C.47. Vetores e matrizes para estimar as elipses de R4 (Ward-ANCOVA)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
μ	$\begin{bmatrix} 186,74 \\ 97,29 \end{bmatrix}$	$\begin{bmatrix} 359,72 \\ 182,64 \end{bmatrix}$	$\begin{bmatrix} 270,53 \\ 79,3 \end{bmatrix}$	$\begin{bmatrix} 421,4 \\ 184,07 \end{bmatrix}$	$\begin{bmatrix} 119,15 \\ 70,87 \end{bmatrix}$
Σ	$\begin{bmatrix} 2303,18 & 295,97 \\ 295,97 & 50,13 \end{bmatrix}$	$\begin{bmatrix} 1248,49 & 859,59 \\ 859,59 & 780,13 \end{bmatrix}$	$\begin{bmatrix} 746,49 & 147,96 \\ 147,96 & 38,66 \end{bmatrix}$	$\begin{bmatrix} 1491,29 & 642,86 \\ 642,86 & 365,28 \end{bmatrix}$	$\begin{bmatrix} 1332,36 & 314,6 \\ 314,6 & 97,91 \end{bmatrix}$
Λ	$\begin{bmatrix} 2341,42 \\ 11,9 \end{bmatrix}$	$\begin{bmatrix} 1905,23 \\ 123,39 \end{bmatrix}$	$\begin{bmatrix} 776,18 \\ 8,97 \end{bmatrix}$	$\begin{bmatrix} 1782,84 \\ 73,75 \end{bmatrix}$	$\begin{bmatrix} 1407,91 \\ 22,37 \end{bmatrix}$
P	$\begin{bmatrix} 0,991 & -0,128 \\ 0,128 & 0,991 \end{bmatrix}$	$\begin{bmatrix} 0,794 & -0,607 \\ 0,607 & 0,794 \end{bmatrix}$	$\begin{bmatrix} 0,98 & -0,196 \\ 0,196 & 0,98 \end{bmatrix}$	$\begin{bmatrix} 0,91 & -0,413 \\ 0,413 & 0,91 \end{bmatrix}$	$\begin{bmatrix} 0,972 & -0,233 \\ 0,233 & 0,972 \end{bmatrix}$

APÊNDICE D – Otimização γ em dados sobre degradação de motores *turbofan*

Compreender e predizer danos em motores aeronáuticos é amplamente investigado por indústrias e pesquisadores, como nos artigos de [195–201], os quais contemplam a coleta de inúmeras informações, geradas por diversos sensores distintos em posições específicas. Usualmente, o diagnóstico dessa magnitude pode gerar dados extensos que apresentam características multivariadas. Assim, um conjunto dessa natureza pode ser ideal para investigar o comportamento do método de otimização da rotação γ *orthomax*, apresentado no capítulo 3. Com base nisso, dados referentes a prognósticos de degradação de motores a reação, *turbofan*, serão analisados, sendo este conjunto disponibilizado pela *National Aeronautics and Space Administration* (NASA), referente ao estudo de Saxena *et al.* [195] e disponíveis em [202]. Tais conjuntos foram gerados através de *softwares* de simulação (*C-MAPSS – Commercial Modular AeroPropulsion System Simulation*), baseado no comportamento de motores do tipo *turbofan*, apresentado no diagrama da Figura D.1 e descrito na Tabela D.1. Maiores detalhes podem ser verificados em [203].

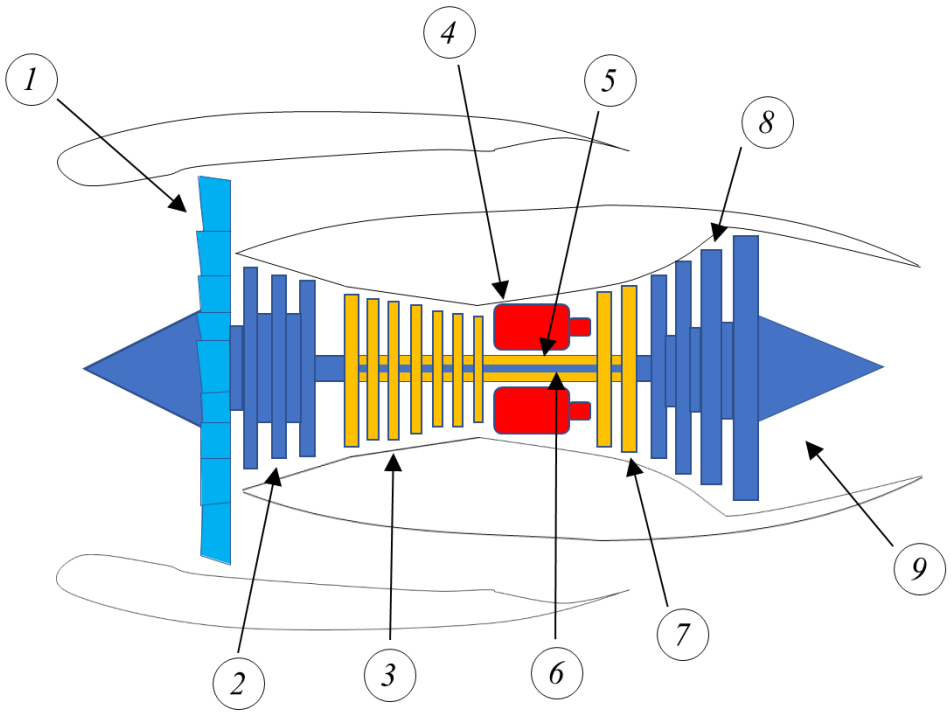


Figura D.1. Diagrama simplificado do motor *turbofan* para o C-MAPSS®

Tabela D.1. Descrição dos componentes do motor tipo *turbofan*

Item	Descrição
1:	Entrada de ar (*Fan*)
2:	Compressor de baixa pressão (*LPC*)
3:	Compressor de alta pressão (*HPC*)
4:	Câmera de combustão
5:	Eixo das turbinas
6:	Eixo dos compressores
7:	Turbina de alta pressão (*HPT*)
8:	Turbina de baixa pressão (*LPT*)
9:	Aumento do fluxo de ar para o empuxo (*Nozzle*)

O conjunto de dados dos autores constitui da simulação de 218 motores distintos, os quais foram mantidos em execução até o momento de falha (variando de 127 a 356 ciclos). A coleta de dados conta com 21 diferentes sensores distribuídos de maneira adequada para analisar o comportamento dos motores. As informações de saída dos sensores estão descritas na Tabela D.2.

Tabela D.2. Respostas de saída para o C-MAPPS

Sensor	Unidade	Descrição	Sensor	Unidade	Descrição
T2	[°R]	Temperatura total na entrada do *Fan*	phi	[pps/psi]	Razão de fluxo de combustível para Ps30
T24	[°R]	Temperatura total na saída na LPC			
T30	[°R]	Temperatura total na saída na HPC	NRf	[rpm]	Velocidade corrigida do *Fan*
T50	[°R]	Temperatura total na saída na LPT	NRc	[rpm]	Velocidade de núcleo corrigida
P2	[psia]	Pressão na entrada do *Fan*	BPR	*	Razão de diluição
P15	[psia]	Pressão total no duto de diluição	farB	*	Razão de mistura combustível/ar
P30	[psia]	Pressão total na saída HPC	htBleed	*	Entalpia de sangria
Nf	[rpm]	Velocidade física do *Fan*	Nf_dmd	[rpm]	Velocidade exigida do *Fan*
Nc	[rpm]	Velocidade do núcleo físico	PCNfR_dmd	[rpm]	Velocidade corrigida exigida do *Fan*
epr	*	Razão de pressão do motor (P50/P2)	W31	[lbm/s]	Sangria do refrigerante HPT
Ps30	[psia]	Pressão estática na saída HPC	W32	[lbm/s]	Sangria de refrigerante LPT

Adaptado de [195].

O método de rotação γ proposto será investigado com dados de 259 trajetórias de teste, que apresentam apenas um modo de falha (degradação do compressor de alta pressão). Deste modo, tem-se um total de 713.811 dados (33.991 linhas experimentais com 21 sensores). Pode-se notar a grande extensão dos dados, que exigem um grande esforço computacional para avaliar e interpretar os resultados para possíveis predições e padronizações.

Através dos testes de adequação, tem-se que os dados apresentaram uma estrutura apta para a técnica multivariada FA, apresentando, para o teste de esfericidade de Bartlett, *p-value* igual a 0,000 e KMO igual a 0,93. Sabendo que o conjunto se caracteriza como uma série temporal multivariada adequada, é preciso definir a quantidade de fatores necessários para o estudo. A partir dos critérios de Kaiser, estabelecidos no capítulo 2, e das diretrizes apresentadas

no capítulo 3, estabelece-se o uso de 2 fatores, sendo estes capazes de representar adequadamente os 21 sensores, explicando 97.6% dos dados originais.

Em seguida, se faz necessário extrair os escores, assim, o melhor nível de rotação γ deve ser definido para se obter a interpretação mais fidedigna e que considere a estrutura de variância-covariância dos dados. Com base nisso, o DOE, do tipo *simplex-lattice* é criado a fim de gerar todas as variações significativas para os níveis de rotação γ. Utilizando um *ld* igual a 10 e considerando pontos axiais, gera-se 13 testes experimentais, como descrito na Tabela D.3. Obtendo-se as configurações dos valores experimentais, é possível aplicar a estratégia FA para cada nível de rotação γ, em que os carregamentos fatoriais são analisados. Esta etapa permite encontrar os valores VTE para os dois fatores extraídos.

Tabela D.3. Matriz experimental para os dados de análise de degradação do motor *turbofan*

Experimento	Controle		Respostas			
	γ1	γ2	VTE1	VTE2	Variância	EQM
1	1	0	12,428	8,058	9,545	28,635
2	0,9	0,1	12,661	7,825	11,690	35,069
3	0,8	0,2	12,933	7,553	14,470	43,409
4	0,7	0,3	13,249	7,238	18,066	54,198
5	0,6	0,4	13,609	6,877	22,660	67,981
6	0,5	0,5	14,009	6,477	28,365	85,094
7	0,4	0,6	14,433	6,053	35,114	105,341
8	0,3	0,7	14,857	5,630	42,568	127,704
9	0,2	0,8	15,250	5,236	50,145	150,435
10	0,1	0,9	15,592	4,894	57,216	171,649
11	0	1	15,871	4,615	63,343	190,029
12	0,75	0,25	13,085	7,401	16,154	48,461
13	0,25	0,75	15,059	5,427	46,383	139,148

A análise dos valores de VTE coletados, referentes aos 21 sensores do motor *turbofan*, foi baseada no modelo quadrático completo para verificar o comportamento dos experimentos. A partir da Tabela D.4 é possível verificar que todos valores VTE são significativos para os parâmetros experimentais, bem como para os demais modelos. Ainda nesta tabela, tem-se que os valores de VTE apresentam um ótimo ajuste experimental, com valores R^2_{adj} iguais a 100% para as VTE's de ambos fatores. Além disso, o ajuste preditivo também apresentou valores altos, indicando que o ajuste experimental é capaz de realizar previsões adequadas. O comportamento das mudanças do valor γ nas VTE's pode ser verificado também nos gráficos de resposta de traço Cox (Figura D.2) e nos gráficos de superfície de resposta (Figura D.3).

Tabela D.4. Análise de variância para EQMVTE (*turbofan*)

Fonte	Graus de Liberdade	Soma dos Quadrados (seq.)	Soma dos Quadrados (ajust.)	Quadrado Médio (ajust.)	F	P
Regressão	4	36409,1	36409,1	9102,3	315961,05	0,000
Linear	1	35503,1	15037,1	15037,1	521971,91	0,000
Quadrático $\gamma1*\gamma2$	1	800,5	810,1	810,1	28118,79	0,000
Cúbico completo $\gamma1*\gamma2*(-)$	1	86,7	86,7	86,7	3010,21	0,000
Quártico completo $\gamma1*\gamma2*(-)^2$	1	18,8	18,8	18,8	650,86	0,000
Erro de Resíduos	8	0,2	0,2	0,0		
Total	12	36409,3				

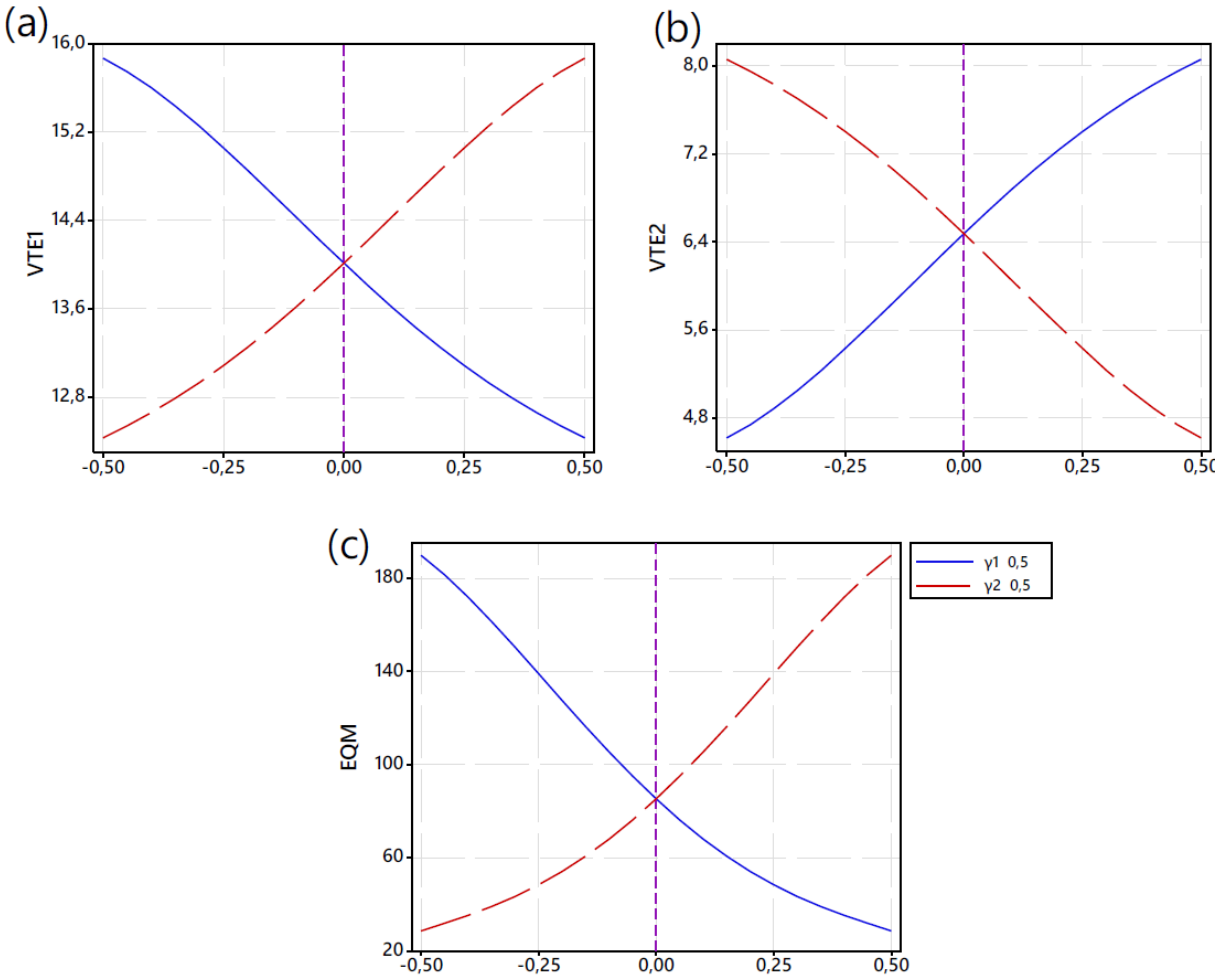
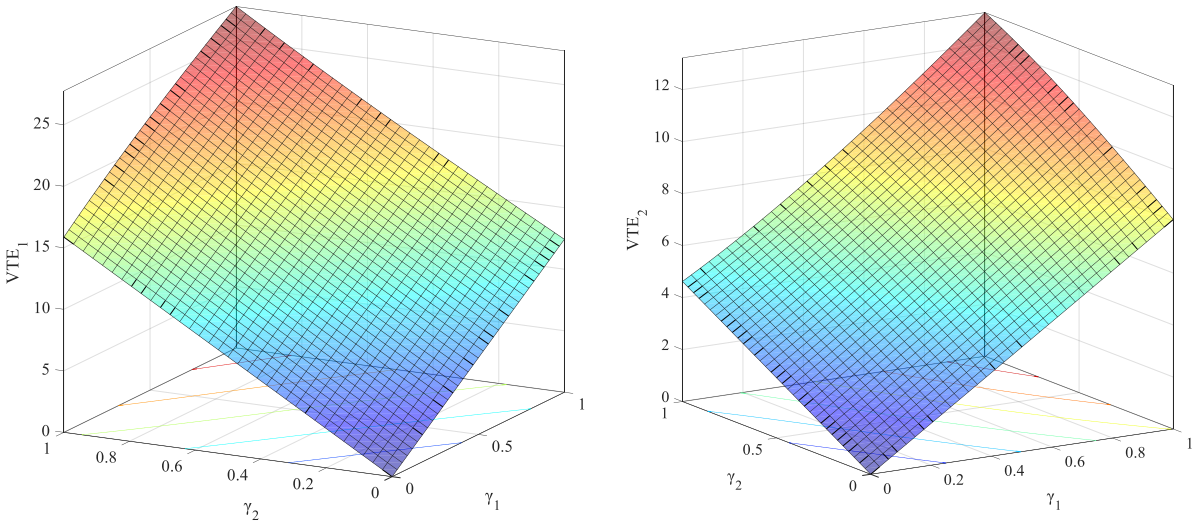


Figura D.2. Traço de resposta Cox para (a) VTE_1, (b) VTE_2 e (c) *EQM*. (dados *turbofan*)



(a) (b)

Figura D.3. Gráfico de contorno e superfície de resposta para (a) VTE_1 e (b) VTE_2. (dados *turbofan*)

Como apresentado anteriormente, sabe-se que os valores de VTE devem ser os mais próximos possíveis. Assim, a partir dessas informações, tem-se a necessidade de aglutinar todos os resultados para criar uma função objetivo única, que represente o erro quadrático médio entre os vetores, que deve ser minimizado. Em seguida, deve-se utilizar a Eq. (3.2), a qual se refere a uma adaptação da modelagem matemática do EQM. Inicialmente, calcula-se a média (μ) entre as VTE's para cada uma das linhas, assim como a sua respectiva variância (σ^2). A partir disso, calculou-se o EQM para cada uma das linhas experimentais. Tais valores estão descritos na Tabela D.3.

Baseado nos valores EQM, é possível realizar a análise do arranjo de misturas considerando um modelo quadrático completo. A Tabela D.4 apresenta que todos efeitos principais e de iteração são significativos (para um IC de 95%). Além disso, a tabela indica que o ajuste do modelo é adequado, apresentando valores de R^2_{adj} e R^2_{pred} iguais a 100% e 99,9% respectivamente. O comportamento de γ para o EQM pode ser verificado na Figura D.2. A partir destes resultados, pode-se modelar a equação com os coeficientes de regressão do EQM, para realizar a otimização (Eq. D.1).

$$EQM_{VTE} = 28,593\times\gamma_1 + 189,928\times\gamma_2 - 95,946\times\gamma_1\times\gamma_2 - 53,148\times\gamma_1\times\gamma_2\times(\gamma_1-\gamma_2) + 51,29\times\gamma_1\times\gamma_2\times\left((\gamma_1-\gamma_2)^2\right) \quad \text{(D.1)}$$

A partir dessa formulação, pode-se realizar a otimização, considerando que os valores experimentais gerados pelo EQM_{VTE} devem ser minimizados. Essa função EQM_{VTE} está sujeita

ao espaço experimental predominante ao DOE (sendo as restrições da otimização), em que as proporções dos parâmetros representam 100%, ou seja, $\gamma_1 + \gamma_2 = 1$.

Para encontrar o ponto ótimo, utilizou-se o algoritmo SQP. Deste modo, através dessa aplicação, chegou-se ao ponto ótimo de rotação γ igual a 1, apresentando EQM_{VTE} de 28,593. Essa condição ótima se refere ao ponto em que o erro quadrático médio apresenta seu menor valor, indicando assim, a melhor possibilidade para a extração dos escores fatoriais. Em outras palavras, considerando esse valor ótimo, pode-se gerar a estrutura com maior nível de simplificação para esse conjunto de dados, em que 2 vetores de escores fatoriais podem explicar um conjunto de 21 sensores com 33.991 linhas experimentais. Tal simplificação condiz em uma redução de dimensionalidade de 90.47%, em que 713.811 dados podem ser representados adequadamente por 67.982 dados. A Figura D.4 ilustra o comportamento da função objetivo através da superfície de resposta e gráficos de contorno, além de indicar o ponto ótimo encontrado pelo SQP.

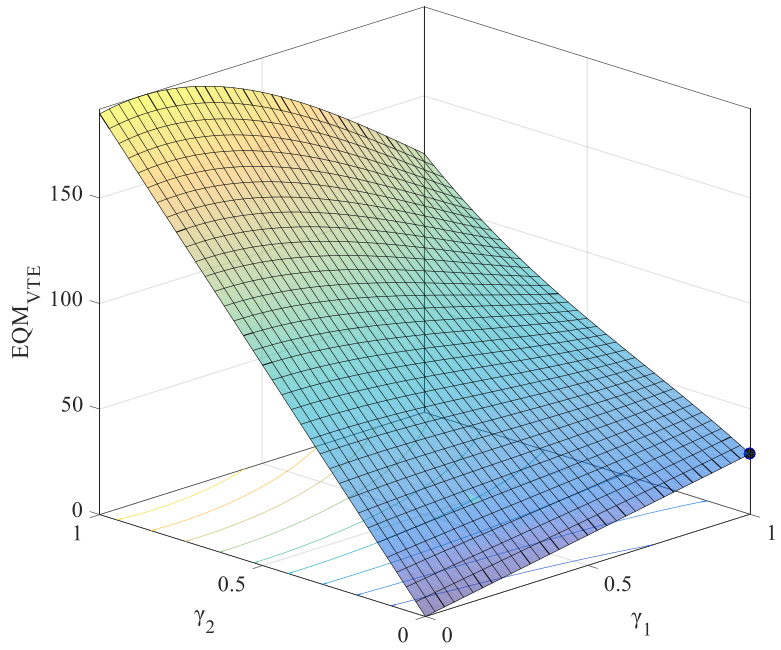


Figura D.4. Gráfico de contorno e superfície de resposta para EQM_{VTE} - *turbofan*

A importância de melhorar a rotação γ *orthomax* para interpretação dos dados é visível ao se analisar o comportamento dos agrupamentos realizados dentro dos fatores com esta rotação. Considerando uma análise comparativa, é possível verificar, na Tabela D.5, o comportamento da FA com diferentes padronizações: sem rotação, com rotação *quartimax*, com γ igual a 0,5 e com rotação γ otimizada. Inicialmente, pode-se verificar que a aplicação não

rotacionada tende a trazer uma maior quantidade de explicação apenas no F_1, criando um valor de VTE expressivo, comparada com o F_2 da mesma aplicação. Assim, tem-se 19 sensores sendo explicados pelo primeiro fator, enquanto apenas 2 sensores são explicados pelo segundo fator. Tal implicação é confirmada a partir do valor de EQM, sendo este extremamente alto para essa hipótese. O mesmo comportamento é verificado ao analisar as cargas fatoriais para rotação *quartimax*, na qual 18 sensores são explicados no primeiro fator, enquanto o segundo explica 3. Este aumento no poder discriminatório (em relação as cargas não rotacionadas) é verificado nos valores de VTE e também no seu erro quadrático médio (190,02).

Tabela D.5. Comparativo das alocações das cargas fatoriais em diferentes rotações

Sensores	***Sem rotação***		***Quartimax***		***Orthomax γ = 0,5***		***γ otimizado***		*Comunalidade*
	F1	*F2*	*F1*	*F2*	*F1*	*F2*	*F1*	*F2*	
T2	**0,874**	-0,458	**0,967**	-0,196	**0,986**	0,002	**0,977**	0,133	0,9728
T24	**0,978**	-0,143	**0,979**	0,135	**0,932**	0,329	**0,880**	0,450	0,9761
T30	**0,995**	0,024	**0,949**	0,301	**0,869**	0,486	**0,797**	0,597	0,991
T50	**0,996**	-0,065	**0,974**	0,216	**0,911**	0,407	**0,849**	0,524	0,9953
P2	**0,859**	-0,508	**0,967**	-0,248	**0,997**	-0,049	**0,995**	0,084	0,9964
P15	**0,899**	-0,433	**0,985**	-0,165	**0,998**	0,037	**0,984**	0,169	0,9963
P30	**0,943**	-0,328	**0,997**	-0,052	**0,987**	0,149	**0,959**	0,279	0,9969
Nf	**0,881**	0,462	**0,717**	0,689	0,564	**0,819**	0,450	**0,887**	0,9884
Nc	**0,995**	0,043	**0,944**	0,319	**0,861**	0,502	**0,786**	0,612	0,9924
epr	**0,958**	-0,068	**0,939**	0,202	**0,879**	0,387	**0,820**	0,500	0,9226
Ps30	**0,956**	0,285	**0,839**	0,540	**0,713**	0,698	0,614	**0,786**	0,9952
phi	**0,943**	-0,327	**0,997**	-0,051	**0,987**	0,151	**0,958**	0,281	0,9969
NRf	0,601	**0,797**	0,355	**0,933**	0,160	**0,985**	0,027	**0,998**	0,9959
NRc	**0,762**	0,614	0,561	**0,802**	0,388	**0,898**	0,265	**0,942**	0,9576
BPR	**-0,884**	-0,448	**-0,724**	-0,676	-0,573	**-0,808**	-0,460	**-0,877**	0,9811
farB	**0,836**	-0,252	**0,873**	-0,008	**0,857**	0,168	**0,827**	0,280	0,7624
htBleed	**0,996**	0,020	**0,951**	0,297	**0,872**	0,482	**0,800**	0,593	0,9914
Nf_dmd	**0,880**	0,462	**0,716**	0,689	0,563	**0,819**	0,449	**0,887**	0,9884
PCNfR_dmd	0,601	**0,797**	0,354	**0,933**	0,160	**0,985**	0,027	**0,998**	0,9959
W31	**0,932**	-0,359	**0,995**	-0,085	**0,991**	0,117	**0,967**	0,248	0,9966
W32	**0,932**	-0,359	**0,995**	-0,085	**0,991**	0,117	**0,967**	0,248	0,9966
VTE	16,908	3,578	15,871	4,615	14,009	6,477	12,428	8,058	20,4861
% Var	0,805	0,170	0,756	0,220	0,667	0,308	0,592	0,384	0,9755
EQM	266,5493		190,0294		85,0943		28,6362		

Negrito: maiores valores de cargas para cada fator

Comparado, em sequência, uma das variações de γ (com valor de 0,5), tem-se um aprimoramento na discriminação dos sensores, em que o F_1 explica 15 sensores e o F_2 explica 6. Esse aprimoramento é visto nos valores de VTE e também em um valor de 85,09 para o EQM_{VTE}. Por fim, ao analisar todas as variações com o valor encontrado na otimização (γ igual a 1), tem-se o maior poder discriminatório entre os sensores, em que o primeiro fator explicou 14 sensores e o segundo fator explicou 7. Essa divisão apresentou o melhor balanceamento possível para as cargas fatoriais, através da VTE. Em consequência disso, ao utilizar esse valor otimizado, tem-se um EQM_{VTE} de 28,636. É importante ressaltar que o valor γ igual a 1 indica

na aplicação da rotação *varimax*, em que, para esse conjunto de dados de degradação do motor *turbofan*, apresentou o melhor comportamento para a explicação das variáveis latentes. Além disso, tem-se que as cargas fatoriais, para o valor otimizado, apresentam altas cargas para os sensores que eles representam e, consequentemente, obtém uma quantidade mais baixa de cargas para o fator remanescente. Tal resultado pode ser ilustrado através de um dendrograma, conforme a Figura D.5, ilustrando o nível de similaridade entre fatores e seus respectivos sensores, além da discriminação existente entre os clusters dos fatores.

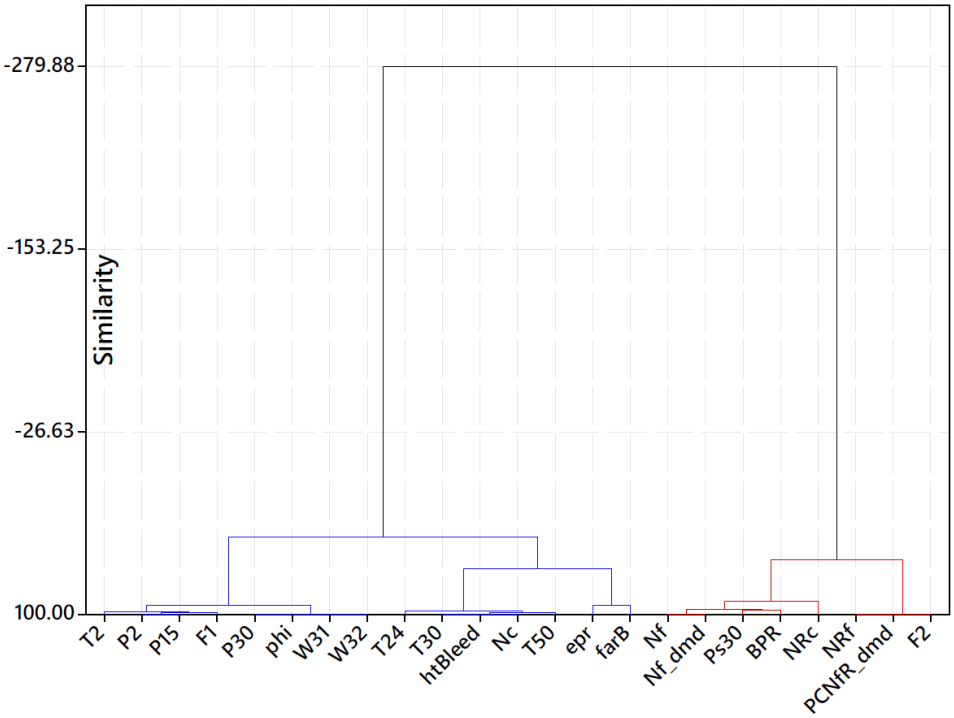


Figura D.5. Agrupamentos da relação "*Sensores* × *Escores de fator*"

A adequação e representação individual desse comportamento pode ser conferida também através da realização de regressões dos fatores de γ otimizado (F_1 e F_2) para cada um dos sensores de seu respectivo grupo. As Figuras D.6 e D.7 ilustram o comportamento dos gráficos de regressão F_1 para os sensores que o mesmo representa, em que é possível verificar que todos valores de ajustes (R^2_{adj}) foram maiores que 92%, com exceção das leituras dos sensores "*epr*" e "*farB*", os quais apresentaram pouca variabilidade durante os ensaios, levando a modelos preditivos com menor ajuste. De maneira análoga, os gráficos de regressão para F_2 apresentaram altos valores de ajustes, com valores de R^2_{adj} maiores que 99%. Tal comportamento pode ser verificado na Figura D.8. Diante deste resultado com alto nível de explicação e de ajustes estatísticos, pode-se concluir que o método proposto proporcionou uma grande redução de dimensionalidade, com uma adequação elevada para possíveis predições de

resultados no monitoramento de vida e outras aplicações, voltadas à análise de degradação de motores *turbofan*.

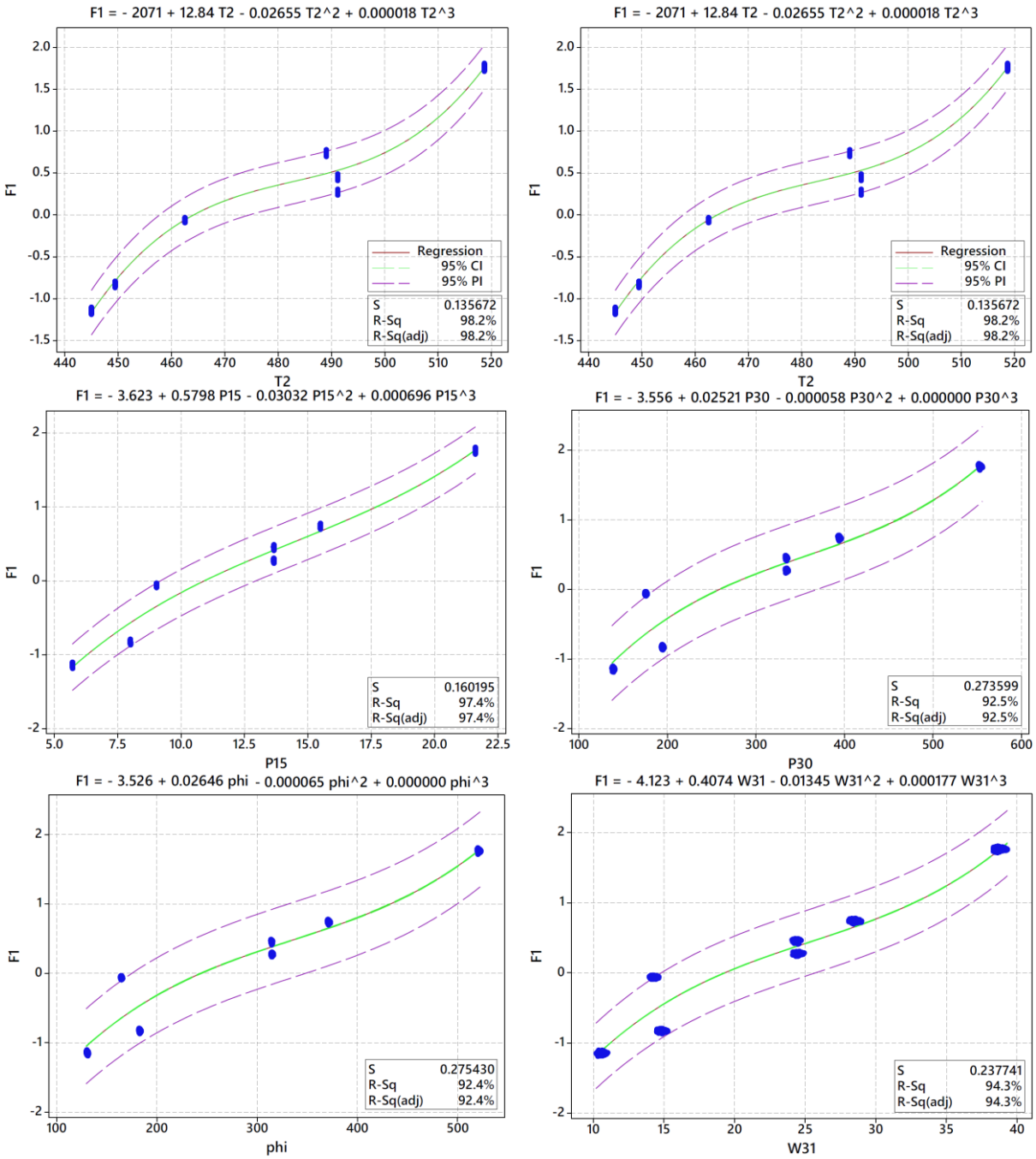


Figura D.6. Gráficos de regressão dos sensores para F1 (Parte I)

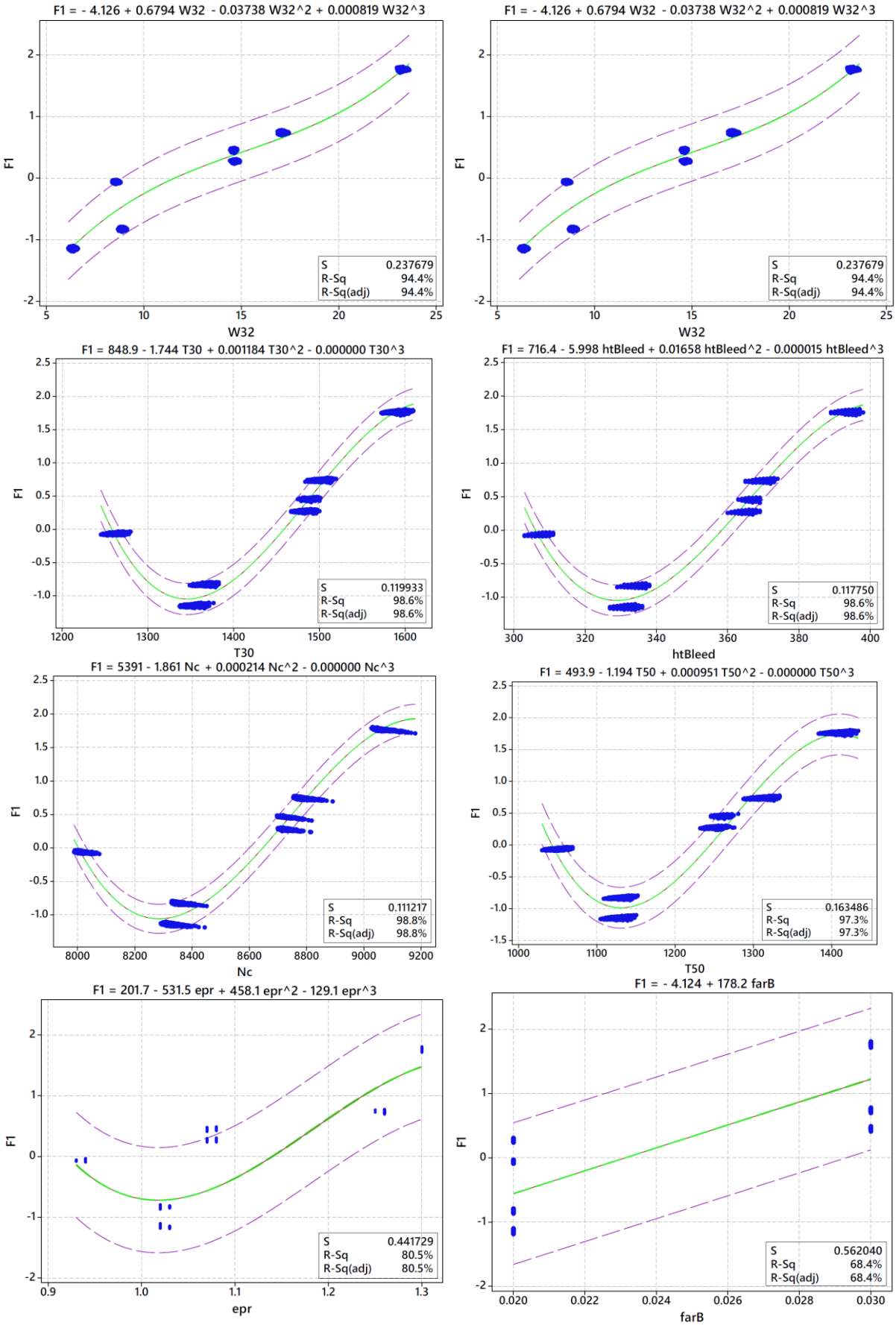


Figura D.7. Gráficos de regressão dos sensores para F1 (Parte II)

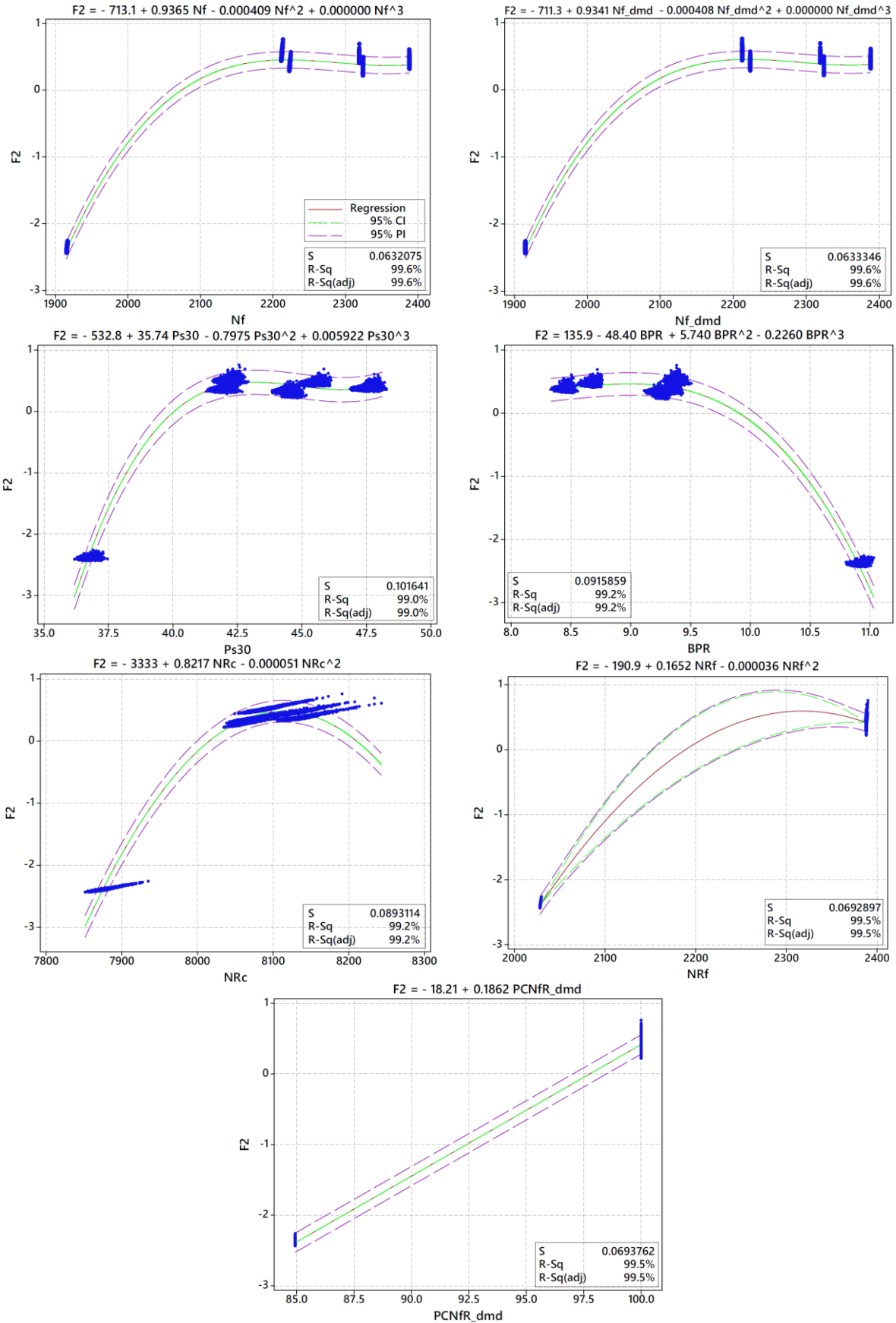


Figura D.8. Gráficos de regressão dos sensores para F2 (Parte I)

REFERÊNCIAS BIBLIOGRÁFICAS

[1] RAGSDALE, C. T. **Spreadsheet Modeling and Decision Analysis: A Practical Introduction to Business Analytics**. 7th. ed. [s.l.] Cengage Learning, 2014.

[2] CAI, L.; ZHU, Y. The challenges of data quality and data quality assessment in the big data era. **Data Science Journal**, v. 14, p. 1–10, 2015.

[3] GHASEMAGHAEI, M. Are firms ready to use big data analytics to create value? The role of structural and psychological readiness. **Enterprise Information Systems**, v. 13, n. 5, p. 650–674, 2019.

[4] FERREIRA, D. F. (FEDERAL U. OF L. **Estatística Multivariada**. 3th. ed. Lavras: UFLA, 2018.

[5] JOHNSON, R.A., WICHERN, D. **Applied Multivariate Statistical Analysis**. 6. ed. New Jersey: Prentice-Hall, 2007.

[6] PERUCHI, R. S. et al. A new multivariate gage R&R method for correlated characteristics. **International Journal of Production Economics**, v. 144, n. 1, p. 301–315, 1 Jul. 2013.

[7] ALMEIDA, F. A. DE et al. A Gage Study Through the Weighting of Latent Variables Under Orthogonal Rotation. **IEEE Access**, v. 8, p. 183557–183570, 2020.

[8] ALMEIDA, F. A. et al. Multivariate data quality assessment based on rotated factor scores and confidence ellipsoids. **Decision Support Systems**, v. 129, p. 113173, Oct. 2020.

[9] PEARSON, K. **On Lines and Planes of Closest Fit to Systems of Points in Space**. [s.l.] University College, 1901.

[10] HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 7, p. 498–520, 1933.

[11] JOLLIFFE, I. T. **Principal Component Analysis**. 2nd. ed. New York: Springer Series in Statistics, 2010.

[12] GAUDÊNCIO, J. H. D. et al. A multiobjective optimization model for machining quality in the AISI 12L14 steel turning process using fuzzy multivariate mean square error. **Precision Engineering**, v. 56, n. December 2018, p. 303–320, 2019.

[13] MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate Analysis**. 5th. ed. London: Academic Press Limited, 1995.

[14] SONG, J.; LI, B. Nonlinear and additive principal component analysis for functional data. **Journal of Multivariate Analysis**, v. 181, p. 104675, 2021.

[15] WANG, R. et al. PCA-assisted reproduction for continuous multi-objective optimization with complicated Pareto optimal set. **Swarm and Evolutionary Computation**, v. 60, p. 100795, 2021.

[16] BOUNOUA, W.; BAKDI, A. Fault detection and diagnosis of nonlinear dynamical processes through correlation dimension and fractal analysis based dynamic kernel PCA. **Chemical Engineering Science**, v. 229, p. 116099, 2021.

[17] JOSHUA, V. et al. Spatial mapping of COVID-19 for Indian states using Principal Component Analysis. **Clinical Epidemiology and Global Health**, v. 10, p. 100690, 2021.

[18] MAHMOUDI, M. R. et al. Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. **Alexandria Engineering Journal**, v. 60, n. 1, p. 457–464, 2021.

[19] FARRUGIA, J. et al. Principal Component Analysis of Hyperspectral Data for Early Detection of Mould in Cheeselets. **Current Research in Food Science**, 2021.

[20] NHU, V.-H. et al. Advanced soft computing techniques for predicting soil compression coefficient in engineering project: a comparative study. **Engineering with Computers**, v. 36, n. 4, p. 1405–1416, 2020.

[21] ASANTE-OKYERE, S. et al. Principal component analysis (PCA) based hybrid models for the accurate estimation of reservoir water saturation. **Computers & Geosciences**, v. 145, p. 104555, 2020.

[22] YU, Y. et al. Improved PCA model for multiple fault detection, isolation and reconstruction of sensors in nuclear power plant. **Annals of Nuclear Energy**, v. 148, p. 107662, 2020.

[23] MENG, T. et al. Application of principal component analysis in measurement of flow fluctuation. **Measurement**, p. 108503, 2020.

[24] ALVIN C. RENCHER. **Methods of Multivariate Analysis**. 2. ed. New York: John Wiley & Sons, Inc., 2002.

[25] COSTELLO, A. B.; OSBORNE, J. W. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. **Practical Assessment Research & Evaluation**, p. 1–9, 2005.

[26] YIN, M.; HUANG, H.; OLDENBURG, T. B. P. An application of exploratory factor analysis in the deconvolution of heavy oil biodegradation, charging and mixing history in southeastern Mexico. **Organic Geochemistry**, v. 151, p. 104161, 2021.

[27] ASADA, T. et al. What factor induces stress in patients with AIS under brace treatment? Analysis of a specific factor using exploratory factor analysis. **Journal of Orthopaedic Science**, 2020.

[28] ASGHARNEZHAD, M. et al. Exploratory factor analysis of gender-based metabolic syndrome components: Results from the PERSIAN Guilan cohort study (PGCS). **Clinical Nutrition ESPEN**, v. 40, p. 252–256, 2020.

[29] LAURETT, R.; PAÇO, A.; MAINARDES, E. W. Measuring sustainable development, its antecedents, barriers and consequences in agriculture: An exploratory factor analysis. **Environmental Development**, p. 100583, 2020.

[30] BACCI, L. A. et al. Optimization of combined time series methods to forecast the demand for coffee in Brazil: A new approach using Normal Boundary Intersection coupled with mixture designs of experiments and rotated factor scores. **International Journal of Production Economics**, v. 212, p. 186–211, 2019.

[31] NAVES, F. L. et al. Multivariate Normal Boundary Intersection based on rotated factor scores: A multiobjective optimization method for methyl orange treatment. **Journal of Cleaner Production**, v. 143, p. 413–439, 2017.

[32] HAYTON, J. C.; ALLEN, D. G.; SCARPELLO, V. Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. **Organizational Research Methods**, v. 7, n. 2, p. 191–205, 1 Apr. 2004.

[33] BROWNE, M. W. An Overview of Analytic Rotation in Exploratory Factor Analysis. **Multivariate Behavioral Research**, v. 36, n. 1, p. 111–150, 2001.

[34] CARROLL, J. B. An analytical solution for approximating simple structure in factor analysis. **Psychometrika**, v. 18, n. 1, p. 23–38, 1953.

[35] SAUNDERS, D. R. An Analytic Method for Rotation to Orthogonal Simple Structure. **ETS Research Bulletin Series**, v. 1953, n. 1, p. i-28, 1 Jun. 1953.

[36] NEUHAUS, J. O.; WRIGLEY, C. The Quartimax Method. **British Journal of Statistical Psychology**, v. 7, n. 2, p. 81–91, 1 Nov. 1954.

[37] FERGUSON, G. A. The concept of parsimony in factor analysis. **Psychometrika**, v. 19, n. 4, p. 281–290, 1954.

[38] KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, v. 23, n. 3, p. 187–200, 1958.

[39] HARMAN, H. H. **Modern factor analysis**. 3th. ed. Chicago: University of Chicago Press., 1976.

[40] HAIR, J. F. et al. **Multivariate Data Analysis**. 5th. ed. [s.l.] Pearson College Div, 1998.

[41] TABACHNICK, B. G.; FIDELL, L. S. **Using Multivariate Statistics**. 6th. ed. [s.l.] Pearson, 2012.

[42] TRIPATHI, M.; SINGAL, S. K. Allocation of weights using factor analysis for development of a novel water quality index. **Ecotoxicology and Environmental Safety**, v. 183, p. 109510, 2019.

[43] BELLMANN, M. Factor analysis of geometric figures with four attributes: A comparison of PCA, varimax and varimin. **Personality and Individual Differences**, v. 90, p. 326–331, 2016.

[44] DINÇ, E. et al. A new approach to the characterization of the nano-surface structure by using factor analysis. **Communications in Nonlinear Science and Numerical Simulation**, v. 17, n. 2, p. 1012–1020, 2012.

[45] ERTEL, S. Exploratory factor analysis revealing complex structure. **Personality and Individual Differences**, v. 50, n. 2, p. 196–200, 2011.

[46] LAVINE, B. K.; RITTER, J. P.; VOIGTMAN, E. Multivariate curve resolution in liquid chromatography—resolving two-way multi-component data using a Varimax extended rotation. **Microchemical Journal**, v. 72, n. 2, p. 163–178, 2002.

[47] SOARES, P. K.; BRUNS, R. E.; SCARMINIO, I. S. Statistical mixture design—Varimax factor optimization for selective compound extraction from plant material. **Analytica Chimica Acta**, v. 613, n. 1, p. 48–55, 2008.

[48] SHARMA, S. **Applied multivariate techniques**. [s.l.] John Wiley & Sons, Inc., 1996.

[49] PUGGINA BIANCHESI, N. M. et al. A design of experiments comparative study on clustering methods. **IEEE Access**, v. 7, p. 167726–167738, 2019.

[50] FÁVERO, L. P. **Análise de Dados: Técnicas Multivariadas Exploratórias com SPSS e STATA**. [s.l.] Elsevier, Grupo Gen, 2015.

[51] BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Simpósio Brasileiro de Probabilidade e Estatística, 1990

[52] INGUANZO, A. et al. Hierarchical cluster analysis of multimodal imaging data identifies brain atrophy and cognitive patterns in Parkinson's disease. **Parkinsonism & Related Disorders**, v. 82, p. 16–23, 2021.

[53] LINGG, N. et al. Proteomics analysis of host cell proteins after immobilized metal affinity

chromatography: Influence of ligand and metal ions. **Journal of Chromatography A**, v. 1633, p. 461649, 2020.

[54] AHMAD, Z. et al. Pollution indicandum and marble waste polluted ecosystem; role of selected indicator plants in phytoremediation and determination of pollution zones. **Journal of Cleaner Production**, v. 236, p. 117709, 2019.

[55] DARAND, M.; PAZHOH, F. Synoptic analysis of sea level pressure patterns and Vertically Integrated Moisture Flux Convergence VIMFC during the occurrence of durable and pervasive rainfall in Iran. **Dynamics of Atmospheres and Oceans**, v. 86, p. 10–17, 2019.

[56] WIRASUTA, I. M. A. G. et al. A rapid method for screening and determination test of methanol content in ethanol-based products using portable Raman spectroscopy. **Forensic Chemistry**, v. 16, p. 100190, 2019.

[57] LANCASTER, M. C. et al. Phenotypic Clustering of Left Ventricular Diastolic Function Parameters: Patterns and Prognostic Relevance. **JACC: Cardiovascular Imaging**, v. 12, n. 7, Part 1, p. 1149–1161, 2019.

[58] SÁFADI, T. Using independent component for clustering of time series data. **Applied Mathematics and Computation**, v. 243, p. 522–527, 2014.

[59] JAIN, N. C.; INDRAYAN, A.; GOEL, L. R. Monte Carlo comparison of six hierarchical clustering methods on random data. **Pattern Recognition**, v. 19, n. 1, p. 95–99, 1986.

[60] PINEL, D. Clustering methods assessment for investment in zero emission neighborhoods' energy system. **International Journal of Electrical Power & Energy Systems**, v. 121, p. 106088, 2020.

[61] MIRANDA, J. et al. A PCA-based approach for substation clustering for voltage sag studies in the Brazilian new energy context. **Electric Power Systems Research**, v. 136, p. 31–42, 2016.

[62] ALMEIDA, F. A. DE et al. Enhancement of discriminatory power by ellipsoidal functions for substation clustering in voltage sag studies. **Electric Power Systems Research**, v. 185, p. 106368, 2020.

[63] MONTGOMERY, D. C.; RUNGER, G. C. **Applied Statistics and Probability for Engineers**. 7th. ed. New York: Wiley, 2019.

[64] SANTOS, A. S. et al. Development of reference material from powdered milk: Uncertainties and interlaboratory evaluation through confidence ellipses. **Microchemical Journal**, v. 159, p. 105330, 2020.

[65] MATSUO, M. et al. MON-P133: Analysis of the Body Composition Distribution by Confidence Ellipse of RXC Graph for Japanese Diabetes Mellitus Patients. **Clinical Nutrition**, v. 35, p. S202, 2016.

[66] SARLIS, N. V; CHRISTOPOULOS, S.-R. G. Visualization of the significance of Receiver Operating Characteristics based on confidence ellipses. **Computer Physics Communications**, v. 185, n. 3, p. 1172–1176, 2014.

[67] CADORET, M.; HUSSON, F. Construction and evaluation of confidence ellipses applied at sensory data. **Food Quality and Preference**, v. 28, n. 1, p. 106–115, 2013.

[68] HUSSON, F.; LÊ, S.; PAGÈS, J. Confidence ellipse for the sensory profiles obtained by principal component analysis. **Food Quality and Preference**, v. 16, n. 3, p. 245–250, 2005.

[69] NAESER, K.; GUO, S. Precision of autokeratometry expressed as confidence ellipses in Euclidian 2-space. **Ophthalmic and Physiological Optics**, v. 20, n. 2, p. 160–168, 2000.

[70] MIRANDA FILHO, J. **Agrupamento de Subestações para Estudos de Afundamentos de Tensão por Análise de Componentes Principais.** 2016. Universidade Federal de Itajubá. 2016.

[71] ANEEL. AGENCIA NACIONAL DE ENERGIA ELÉTRICA (ANEEL). Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST Módulo 8 – Qualidade da Energia Elétrica. In: [s.l: s.n.]. p. 76.

[72] SANTIS, M. DE et al. Analysis of the origin of measured voltage sags in interconnected networks. **Electric Power Systems Research**, v. 154, p. 391–400, 2018.

[73] LIAO, H.; ANANI, N. Fault Identification-based Voltage Sag State Estimation Using Artificial Neural Network. **Energy Procedia**, v. 134, p. 40–47, 2017.

[74] BRANCO, H. M. G. C. et al. Multiobjective optimization for power quality monitoring allocation considering voltage sags in distribution systems. **International Journal of Electrical Power & Energy Systems**, v. 97, p. 1–10, 2018.

[75] SANTOS, A. DOS; CORREIA DE BARROS, M. T. Voltage sag prediction for network planning. **Electric Power Systems Research**, v. 140, p. 976–983, 2016.

[76] MAJUMDER, S. et al. Allocation of Common-Pool Resources in an Unmonitored Open System. **IEEE Transactions on Power Systems**, v. 34, n. 5, p. 3912–3920, 2019.

[77] MORADI, M. H.; MOHAMMADI, Y. Voltage sag source location: A review with introduction of a new method. **International Journal of Electrical Power & Energy Systems**, v. 43, n. 1, p. 29–39, 2012.

[78] MOKHLIS, H.; LI, H. Non-linear representation of voltage sag profiles for fault location in distribution networks. **International Journal of Electrical Power & Energy Systems**, v. 33, n. 1, p. 124–130, 2011.

[79] GENCER, Ö.; ÖZTÜRK, S.; ERFIDAN, T. A new approach to voltage sag detection based on wavelet transform. **International Journal of Electrical Power & Energy Systems**, v. 32, n. 2, p. 133–140, 2010.

[80] ERIŞTI, H. et al. Optimal feature selection for classification of the power quality events using wavelet transform and least squares support vector machines. **International Journal of Electrical Power & Energy Systems**, v. 49, p. 95–103, 2013.

[81] SUN, H. et al. Voltage Sag Source Identification Based on Few-Shot Learning. **IEEE Access**, v. 7, p. 164398–164406, 2019.

[82] NAGATA, E. A. et al. Voltage sag and swell detection and segmentation based on Independent Component Analysis. **Electric Power Systems Research**, v. 155, p. 274–280, 2018.

[83] SADEGHI, M. H.; DASTFAN, A.; DAMCHI, Y. Optimal coordination of directional overcurrent relays in distribution systems with DGs and FCLs considering voltage sag energy index. **Electric Power Systems Research**, v. 191, p. 106884, 2021.

[84] REY-BOUÉ, A. B. et al. Frequency- adaptive control of a three-phase single-stage grid-connected photovoltaic system under grid voltage sags. **International Journal of Electrical Power & Energy Systems**, v. 125, p. 106416, 2021.

[85] SUDHARANI, S.; GODWIN IMMANUEL, D. Mitigation of voltage sag/swell by dynamic voltage restorer using fuzzy based particle swarm controller. **Materials Today: Proceedings**, 2021.

[86] HAN, Y. et al. Cause, Classification of Voltage Sag, and Voltage Sag Emulators and

Applications: A Comprehensive Overview. **IEEE Access**, v. 8, p. 1922–1934, 2020.

[87] HASAN, S. et al. Calculation of the Voltage Sag Recovery Point-on-Wave and Sag Duration Using System Parameters. **IEEE Transactions on Industry Applications**, v. 56, n. 4, p. 4588–4601, 2020.

[88] LIU, Y. et al. Multi-Objective Optimal STATCOM Allocation for Voltage Sag Mitigation. **IEEE Transactions on Power Delivery**, v. 35, n. 3, p. 1410–1422, 2020.

[89] SHAREEF, H.; MOHAMED, A.; IBRAHIM, A. A. Identification of voltage sag source location using S and TT transformed disturbance power. **Journal of Central South University**, v. 20, n. 1, p. 83–97, 2013.

[90] THAKUR, P.; SINGH, A. K. A novel way to quantify the magnitude of voltage sag. **Electrical Engineering**, v. 95, n. 4, p. 331–340, 2013.

[91] LATTIN, J. M.; CARROLL, J. D.; GREEN, P. E. **Análise de dados multivariados**. [s.l.] Cengage Learning, 2011.

[92] QIN, Y. J. et al. Reduction of non-linear many objectives for coordinated operation of integrated energy systems. **International Journal of Electrical Power & Energy Systems**, v. 117, p. 105657, 2020.

[93] CHRÉTIEN, S.; CLARKSON, P.; GARCIA, M. S. Application of Robust PCA with a structured outlier matrix to topology estimation in power grids. **International Journal of Electrical Power & Energy Systems**, v. 100, p. 559–564, 2018.

[94] MOHAMMADI, H.; DEHGHANI, M. PMU based voltage security assessment of power systems exploiting principal component analysis and decision trees. **International Journal of Electrical Power & Energy Systems**, v. 64, p. 655–663, 2015.

[95] KHOSRAVI, A. et al. Classification of Voltage Sags based on Multiway Principal Component Analysis and Case Based Reasoning. **IFAC Proceedings Volumes**, v. 41, n. 2, p. 5529–5534, 2008.

[96] MOHAMMADI, Y.; MORADI, M. H.; CHOUHY LEBORGNE, R. A novel method for voltage-sag source location using a robust machine learning approach. **Electric Power Systems Research**, v. 145, p. 122–136, 2017.

[97] BAKDI, A. et al. Nonparametric Kullback-divergence-PCA for intelligent mismatch detection and power quality monitoring in grid-connected rooftop PV. **Energy**, v. 189, p. 116366, 2019.

[98] PIRES, V. F.; AMARAL, T. G.; MARTINS, J. F. Power quality disturbances classification using the 3-D space representation and PCA based neuro-fuzzy approach. **Expert Systems with Applications**, v. 38, n. 9, p. 11911–11917, 2011.

[99] KHOSRAVI, A.; MELÉNDEZ, J.; COLOMER, J. Classification of sags gathered in distribution substations based on multiway principal component analysis. **Electric Power Systems Research**, v. 79, n. 1, p. 144–151, 2009.

[100] WANG, N.; WANG, S.; JIA, Q. **The method to reduce identification feature of different voltage sag disturbance source based on principal component analysis**. 2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC Asia-Pacific). **Anais**...2014

[101] JIAYU, L. et al. **Principal component analysis and identification of power quality disturbance signal phase space reconstructed images**. Proceedings of the 31st Chinese Control Conference. **Anais**...2012

[102] JASIŃSKI, M.; SIKORSKI, T.; BORKOWSKI, K. Clustering as a tool to support the assessment of power quality in electrical power networks with distributed generation in the mining industry. **Electric Power Systems Research**, v. 166, p. 52–60, 2019.

[103] LÓPEZ, J. J. et al. Hopfield–K-Means clustering algorithm: A proposal for the segmentation of electricity customers. **Electric Power Systems Research**, v. 81, n. 2, p. 716–724, 2011.

[104] LI, P. et al. Dynamic equivalent modeling of two-staged photovoltaic power station clusters based on dynamic affinity propagation clustering algorithm. **International Journal of Electrical Power and Energy Systems**, v. 95, p. 463–475, 2018.

[105] VINOTHKUMAR, K.; SELVAN, M. P. Hierarchical Agglomerative Clustering Algorithm method for distributed generation planning. **International Journal of Electrical Power and Energy Systems**, v. 56, p. 259–269, 2014.

[106] BARANWAL, M.; SALAPAKA, S. Clustering and supervisory voltage control in power systems. **International Journal of Electrical Power and Energy Systems**, v. 109, n. October 2018, p. 641–651, 2019.

[107] FERREIRA, A. M. S. et al. Pattern recognition as a tool to support decision making in the management of the electric sector. Part II: A new method based on clustering of multivariate time series. **International Journal of Electrical Power and Energy Systems**, v. 67, p. 613–626, 2015.

[108] ROMERO, M.; GALLEGO, L.; PAVAS, A. Fault Zones Location on Distribution Systems Based on Clustering of Voltage Sags Patterns. **2012 IEEE 15th International Conference on Harmonics and Quality of Power**, p. 486–493, 2012.

[109] HARIYANTO, N.; NOEGROHO, R. New Probabilistic Approach for Identification Event Severity Index Due To Short Circuit Fault. **2014 International Conference on Electrical Engineering and Computer Science (ICEECS)**, n. November, p. 1–5, 2014.

[110] BALOUJI, E.; SALOR, O. Eigen-Analysis Based Power Quality Event Data Clustering and Classification. **IEEE PES Innovative Smart Grid Technologies, Europe**, p. 1–5, 2014.

[111] DUAN, R. C.; WANG, F. H.; ZHANG, J. Data Mining & Pattern Recognition of Voltage Sag Based on K-means Clustering Algorithm. **2015 IEEE Power & Energy Society General Meeting**, n. 51307106, p. 1–5, 2015.

[112] INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **IEEE Std 1564 - Guide for Voltage Sag Indices**. [s.l.] IEEE, 2014.

[113] GOMES, G. F.; ALMEIDA, F. A. DE. Tuning metaheuristic algorithms using mixture design: Application of sunflower optimization for structural damage identification. **Advances in Engineering Software**, v. 149, p. 102877, 2020.

[114] PAULA, T. I. DE et al. **A Mixture Design of Experiments Approach for Genetic Algorithm Tuning Applied to Multi-objective Optimization BT - Optimization of Complex Systems: Theory, Models, Algorithms and Applications**. (H. A. Le Thi, H. M. Le, T. Pham Dinh, Eds.)Cham: Springer International Publishing, 2020

[115] STURGES, H. A. The Choice of a Class Interval. **Journal of the American Statistical Association**, v. 21, n. 153, p. 65–66, 1 Mar. 1926.

[116] MONTGOMERY, D. C. **Design and Analysis of Experiments**. 9th. ed. New York: John Wiley & Sons, Inc., 2017.

[117] FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological bulletin**, v. 76, n. 5, p. 378–382, 1971.

[118] FLEISS, J. L.; LEVIN, B.; PAIK, M. C. **Statistical Methods for Rates and Proportions**. Third Edit ed. [s.l.] John Wiley & Sons, Inc., 2003.

[119] AIAG. **Measurement systems analysis: reference manual**. 4th. ed. Detroid, MI, USA: Automotive Industry Action Group, 2010.

[120] GISEV, N.; BELL, J. S.; CHEN, T. F. Interrater agreement and interrater reliability: key concepts, approaches, and applications. **Research in social & administrative pharmacy : RSAP**, v. 9, n. 3, p. 330–338, 2013.

[121] LEWIS, G. H.; JOHNSON, R. G. Kendall's Coefficient of Concordance for Sociometric Rankings with Self Excluded. **Sociometry**, v. 34, n. 4, p. 496–503, 24 Aug. 1971.

[122] LEGENDRE, P. Species associations: the Kendall coefficient of concordance revisited. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 10, n. 2, p. 226, 2005.

[123] SIDNEY, S. Nonparametric Statistics for the Behavioral Sciences. **The Journal of Nervous and Mental Disease**, v. 125, n. 3, 1957.

[124] AGRESTI, A. **Analysis of Ordinal Categorical Data**. 2nd. ed. [s.l.] John Wiley & Sons, Inc., 2010.

[125] HINKLE, D. E.; WIERSMA, W.; JURS, S. G. **Applied Statistics for the Behavioral Sciences**. Boston: Houghton Mifflin, 2002.

[126] GOMES, J. H. F. **Método dos Polinômios Canônicos e misturas para otimização multi-objetivo**. 2013. Universidade Federal de Itajubá. 2013.

[127] JOHN A. CORNELL. **Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data**. 3th. ed. New York: John Wiley & Sons, Inc., 2002.

[128] MYERS, R. H.; MONTGOMERY, D. C.; ANDERSON-COOK, C. M. **Response Surface Methodology: Process and Product Optimization Using Designed Experiments**. 4th. ed. New York: John Wiley & Sons, 2016.

[129] TAGUCHI G, ELSAYED EA, T. H. **Quality engineering in production systems**. [s.l.] McGraw-Hill, 1989.

[130] CIRILLO, M. A. **Otimização na Experimentação: Aplicações nas Engenharias e Ciências Agrárias**. Lavras: Editora UFLA, 2015.

[131] RAMACHANDER, J. et al. Performance and emission predictions of a CRDI engine powered with diesel fuel: A combined study of injection parameters variation and Box-Behnken response surface methodology based optimization. **Fuel**, v. 290, p. 120069, 2021.

[132] REJEB, O. et al. Investigation of a solar still behaviour using response surface methodology. **Case Studies in Thermal Engineering**, v. 24, p. 100816, 2021.

[133] TARIQ, R.; MAQBOOL, F.; ABBAS, S. Z. Small-scale production of hydrogen via auto-thermal reforming in an adiabatic packed bed reactor: Parametric study and reactor's optimization through response surface methodology. **Computers & Chemical Engineering**, v. 145, p. 107192, 2021.

[134] DURÁN, I. R. et al. Response surface methodology as a predictive tool for the fabrication of coatings with optimal anti-fogging performance. **Thin Solid Films**, v. 718, p. 138482, 2021.

[135] CAO, X. et al. Modeling and optimization of resistance spot welded aluminum to Al-Si coated boron steel using response surface methodology and genetic algorithm. **Measurement**, v. 171, p. 108766, 2021.

[136] KARAOGLAN, A. D. et al. Design Optimization of Magnetic Flux Distribution for PMG

by Using Response Surface Methodology. **IEEE Transactions on Magnetics**, v. 56, n. 6, p. 1–9, 2020.

[137] CHEN, X. et al. Empirical Passive Intermodulation Multiphysics Modeling Using Design of Experiment Method. **IEEE Transactions on Instrumentation and Measurement**, v. 69, n. 12, p. 9371–9373, 2020.

[138] ZAMAN, S. A.; ROY, D.; GHOSH, S. Process modeling and optimization for biomass steam-gasification employing response surface methodology. **Biomass and Bioenergy**, v. 143, p. 105847, 2020.

[139] LI, Q. et al. Response surface methodology to optimize the conditions for Enterococcus faecium YA002 producing H2 from xylose. **International Journal of Hydrogen Energy**, 2020.

[140] GONZÁLEZ-VEGA, R. I. et al. Optimization of growing conditions for pigments production from microalga Navicula incerta using response surface methodology and its antioxidant capacity. **Saudi Journal of Biological Sciences**, 2020.

[141] ALMEIDA, F. A. DE et al. A weighted mean square error approach to the robust optimization of the surface roughness in an AISI 12L14 free-machining steel-Turning process. **Strojniski Vestnik/Journal of Mechanical Engineering**, v. 64, n. 3, p. 147–156, 2018.

[142] ALMEIDA, F. A. DE et al. Multivariate Taguchi loss function optimization based on principal components analysis and normal boundary intersection. **Engineering with Computers**, 2020.

[143] GAUDÊNCIO, J. H. D. et al. Fuzzy multivariate mean square error in equispaced pareto frontiers considering manufacturing process optimization problems. **Engineering with Computers**, v. 35, n. 4, p. 1213–1236, 2019.

[144] VELASCO, J. A.; AMARIS, H.; ALONSO, M. Deep Learning loss model for large-scale low voltage smart grids. **International Journal of Electrical Power & Energy Systems**, v. 121, p. 106054, 2020.

[145] ALMEIDA, F. A. **Análise Multivariada do Sistema de Medição de um Processo de Solda a Ponto por Resistência Elétrica utilizando Componentes Principais Ponderados.** 2017. Universidade Federal de Itajubá. 2017.

[146] ALMEIDA, F. A. et al. A multivariate GR&R approach to variability evaluation of measuring instruments in resistance spot welding process. **Journal of Manufacturing Processes**, v. 36, n. July, p. 465–479, 2018.

[147] ARANHA, F.; ZAMBALDI, F. **Análise Fatorial em Administração**. São Paulo: Cengage Learning, 2008.

[148] REIS, E. **Estatística multivariada aplicada**. 2th. ed. Lisboa: Edições Sílabo, 2001.

[149] BARTLETT, M. S. A Note on the Multiplying Factors for Various χ^2 Approximations. **Journal of the Royal Statistical Society. Series B (Methodological)**, v. 16, n. 2, p. 296–298, 11 Jan. 1954.

[150] KAISER, H. F. A second generation little jiffy. **Psychometrika**, v. 35, n. 4, p. 401–415, 1970.

[151] LEITE, R. R. **Método de Interseção Normal à fronteira para Modelos Quadráticos de Escores Fatoriais Rotacionados.** 2019. Universidade Federal de Itajubá. 2019.

[152] VISINESCU, L. L.; EVANGELOPOULOS, N. Orthogonal rotations in latent semantic analysis: An empirical study. **Decision Support Systems**, v. 62, p. 131–143, 1 Jun. 2014.

[153] DARTON, R. A. Rotation in Factor Analysis. **Journal of the Royal Statistical Society. Series D (The Statistician)**, v. 29, n. 3, p. 167–194, 12 Jun. 1980.

[154] THURSTONE, L. L. Multiple-factor analysis. **Chicago: University of Chicago Press,** p. 535., 1947.

[155] MULAIK, S. A. **Foundations of Factor Analysis**. 2th. ed. [s.l.] Chapman & Hall/CRC, 2010.

[156] RUMMEL, R. J. **Applied Factor Analysis**. [s.l.] Northwestern University Press, 1970.

[157] ALGULIYEV, R. M.; ALIGULIYEV, R. M.; SUKHOSTAT, L. V. Parallel batch k-means for Big data clustering. **Computers & Industrial Engineering**, v. 152, p. 107023, 2021.

[158] AFSHOON, I.; MIRI, M.; MOUSAVI, S. R. Combining Kriging meta models with U-function and K-Means clustering for prediction of fracture energy of concrete. **Journal of Building Engineering**, v. 35, p. 102050, 2021.

[159] SIPKENS, T. A.; ROGAK, S. N. Technical note: Using k-means to identify soot aggregates in transmission electron microscopy images. **Journal of Aerosol Science**, v. 152, p. 105699, 2021.

[160] PRAMOD, C. P.; PILLAI, G. N. K-Means clustering based Extreme Learning ANFIS with improved interpretability for regression problems. **Knowledge-Based Systems**, v. 215, p. 106750, 2021.

[161] LONG, Z.-Z. et al. Flexible Subspace Clustering: A Joint Feature Selection and K-Means Clustering Framework. **Big Data Research**, v. 23, p. 100170, 2021.

[162] DONG, W. et al. Short-Term Wind-Speed Forecasting Based on Multiscale Mathematical Morphological Decomposition, K-Means Clustering, and Stacked Denoising Autoencoders. **IEEE Access**, v. 8, p. 146901–146914, 2020.

[163] MEHDIPOUR, S. H.; MACHADO, J. A. T. Cluster analysis of the large natural satellites in the solar system. **Applied Mathematical Modelling**, v. 89, p. 1268–1278, 2021.

[164] ALGULIYEV, R. M.; ALIGULIYEV, R. M.; SUKHOSTAT, L. V. Weighted consensus clustering and its application to Big data. **Expert Systems with Applications**, v. 150, p. 113294, 2020.

[165] PERALTA, D.; SAEYS, Y. Robust unsupervised dimensionality reduction based on feature clustering for single-cell imaging data. **Applied Soft Computing**, v. 93, p. 106421, 2020.

[166] GUEORGUIEVA, N.; VALOVA, I.; GEORGIEV, G. M&MFCM: Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics. **Procedia Computer Science**, v. 114, p. 224–233, 2017.

[167] GROENEN, P. J. F.; JAJUGA, K. Fuzzy clustering with squared Minkowski distances. **Fuzzy Sets and Systems**, v. 120, n. 2, p. 227–237, 2001.

[168] JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. 1. ed. [s.l.] Pearson College Div, 1988.

[169] ROMESBURG, H. C. **Cluster Analysis for Researchers**. [s.l.] Lulu Press, 2004.

[170] MCQUITTY, L. L. FOR DISCRETE AND CONTINUOUS DATA analysis (McQuitty , 1955). Equation ij. p. 825–831, 1966.

[171] WARD, J. H. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association**, v. 58, n. 301, p. 236–244, 1 Mar. 1963.

[172] BABNIK, T.; AGGARWAL, R. K.; MOORE, P. J. Principal Component and Hierarchical Cluster Analyses as Applied to Transformer Partial Discharge Data With Particular Reference to Transformer Condition Monitoring. **IEEE Transactions on Power Delivery**, v. 23, n. 4, p. 2008–2016, 2008.

[173] GOMES, J. H. F. et al. Weighted Multivariate Mean Square Error for processes optimization: A case study on flux-cored arc welding for stainless steel claddings. **European Journal of Operational Research**, v. 226, n. 3, p. 522–535, 2013.

[174] DUARTE COSTA, D. M. et al. A normal boundary intersection with multivariate mean square error approach for dry end milling process optimization of the AISI 1045 steel. **Journal of Cleaner Production**, v. 135, p. 1658–1672, 2016.

[175] PAIVA, A. P. et al. A multivariate mean square error optimization of AISI 52100 hardened steel turning. **International Journal of Advanced Manufacturing Technology**, v. 43, n. 7–8, p. 631–643, 2009.

[176] KÖKSOY, O. Multiresponse robust design: Mean square error (MSE) criterion. **Applied Mathematics and Computation**, v. 175, n. 2, p. 1716–1729, 2006.

[177] BRITO, T. G. et al. A normal boundary intersection approach to multiresponse robust optimization of the surface roughness in end milling process with combined arrays. **Precision Engineering**, v. 38, n. 3, p. 628–638, 2014.

[178] LIN, D. K. J.; TU, W. Dual Response Surface Optimization. **Journal of Quality Technology**, v. 27, n. 1, p. 34–39, 1 Jan. 1995.

[179] RAO, S. S. **Engineering Optimization: Theory and Practice**. 5th. ed. Coral Gables, Florida, USA: Wiley, 2020.

[180] ELSAYED, S. K.; ELATTAR, E. E. Hybrid Harris hawks optimization with sequential quadratic programming for optimal coordination of directional overcurrent relays incorporating distributed generation. **Alexandria Engineering Journal**, v. 60, n. 2, p. 2421–2433, 2021.

[181] HMIDA, J. BEN; CHAMBERS, T.; LEE, J. Solving constrained optimal power flow with renewables using hybrid modified imperialist competitive algorithm and sequential quadratic programming. **Electric Power Systems Research**, v. 177, p. 105989, 2019.

[182] PATEL, A. R.; PATEL, M. A.; VYAS, D. R. Variational Analysis and Sequential Quadratic Programming Approach for Robotics. **Procedia Technology**, v. 4, p. 636–640, 2012.

[183] SIVASUBRAMANI, S.; SHANTI SWARUP, K. Hybrid DE–SQP algorithm for non-convex short term hydrothermal scheduling problem. **Energy Conversion and Management**, v. 52, n. 1, p. 757–761, 2011.

[184] HAN, X.; QUAN, L.; XIONG, X. A modified gravitational search algorithm based on sequential quadratic programming and chaotic map for ELD optimization. **Knowledge and Information Systems**, v. 42, n. 3, p. 689–708, 2015.

[185] ZHANG, J.-T. et al. Optimal design of a rod shape ultrasonic motor using sequential quadratic programming and finite element method. **Finite Elements in Analysis and Design**, v. 59, p. 11–17, 2012.

[186] BOGGS, P. T.; TOLLE, J. W. Sequential Quadratic Programming. **Acta Numerica**, v. 4, p. 1–51, 1995.

[187] MICHAEL, B. B. **Nonlinear Optimization with Engineering Applications**. New York: Springer, 2008.

[188] ALMEIDA, F. A. DE et al. A new multivariate approach based on weighted factor scores

and confidence ellipses to precision evaluation of textured fiber bobbins measurement system. **Precision Engineering**, v. 60, p. 520–534, Nov. 2019.

[189] BELINATO, G. et al. A multivariate normal boundary intersection PCA-based approach to reduce dimensionality in optimization problems for LBM process. **Engineering with Computers**, v. 35, n. 4, p. 1533–1544, 2019.

[190] PERUCHI, R. S. et al. Weighted approach for multivariate analysis of variance in measurement system analysis. **Precision Engineering**, v. 38, n. 3, p. 651–658, 1 Jul. 2014.

[191] MIGUEL, P. A. C. et al. **Metodologia de pesquisa em engenharia de produção e gestão de operações**Elsevier, , 2014.

[192] APPOLINÁRIO, F. **Metodologia da ciência – filosofia e prática da pesquisa**. São Paulo: Editora Pioneira Thomson Learning, 2006.

[193] WILL M. BERTRAND, J.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v. 22, n. 2, p. 241–264, 1 Jan. 2002.

[194] CHOUHY, R. et al. New approach for power lines performance estimation based on load variation due to voltage sags. **Electric Power Systems Research**, v. 83, n. 1, p. 35–40, 2012.

[195] SAXENA, A. et al. **Damage propagation modeling for aircraft engine run-to-failure simulation**. 2008 International Conference on Prognostics and Health Management. **Anais**...2008

[196] KUROSAKI, M. et al. Fault Detection and Identification in an IM270 Gas Turbine Using Measurements for Engine Control . **Journal of Engineering for Gas Turbines and Power**, v. 126, n. 4, p. 726–732, 24 Nov. 2004.

[197] CHATTERJEE, S.; LITT, J. Online Model Parameter Estimation of Jet Engine Degradation for Autonomous Propulsion Control. In: **AIAA Guidance, Navigation, and Control Conference and Exhibit**. Guidance, Navigation, and Control and Co-located Conferences. [s.l.] American Institute of Aeronautics and Astronautics, 2003.

[198] GOEBEL, K. et al. **Modeling Propagation of Gas Path Damage**. 2007 IEEE Aerospace Conference. **Anais**...2007

[199] LISTOU ELLEFSEN, A. et al. Remaining useful life predictions for turbofan engine degradation using semi-supervised deep architecture. **Reliability Engineering & System Safety**, v. 183, p. 240–251, 2019.

[200] DENG, Y.; BUCCHIANICO, A. DI; PECHENIZKIY, M. Controlling the accuracy and uncertainty trade-off in RUL prediction with a surrogate Wiener propagation model. **Reliability Engineering & System Safety**, v. 196, p. 106727, 2020.

[201] XU, H.; FARD, N.; FANG, Y. Time series chain graph for modeling reliability covariates in degradation process. **Reliability Engineering & System Safety**, v. 204, p. 107207, 2020.

[202] SAXENA, A.; GOEBEL, K. **Turbofan Engine Degradation Simulation Data Set**. Disponível em: <http://ti.arc.nasa.gov/project/prognostic-data-repository>. Acesso em: 17 oct. 2020.

[203] FREDERICK, D.; DECASTRO, J.; LITT, J. **User's Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS)**. Technical ed. [s.l.] NASA/ARL, 2007.

[204] FERREIRA GOMES, G.; SOUZA CHAVES, J. A.; ALMEIDA, F. A. DE. An inverse damage location problem applied to AS-350 rotor blades using bat optimization algorithm and

multiaxial vibration data. **Mechanical Systems and Signal Processing**, v. 145, p. 106932, 2020.

[205] GOMES, G. F. et al. Optimized damage identification in CFRP plates by reduced mode shapes and GA-ANN methods. **Engineering Structures**, v. 181, p. 111–123, Feb. 2019.

[206] MAIA, P. R. et al. Multivariate steepest ascent method based on latent variables. **Applied Mathematical Modelling**, v. 90, 2021.

[207] GOMES, G. F. et al. Inverse structural damage identification problem in CFRP laminated plates using SFO algorithm based on strain fields. **Engineering with Computers**, 2020.

[208] GOMES, G. F. et al. A multiobjective sensor placement optimization for SHM systems considering Fisher information matrix and mode shape interpolation. **Engineering with Computers**, v. 35, n. 2, p. 519–535, 28 Apr. 2019.

[209] TORRES, A. F. et al. Multivariate Stochastic Optimization Approach Applied in a Flux-Cored Arc Welding Process. **IEEE Access**, v. 8, p. 61267–61276, 2020.

[210] RIBEIRO JUNIOR, R. F.; ALMEIDA, F. A. DE; GOMES, G. F. Fault classification in three-phase motors based on vibration signal analysis and artificial neural networks. **Neural Computing and Applications**, 2020.

[211] TORRES, A. F. et al. Impact of stochastic industrial variables on the cost optimization of AISI 52100 hardened-steel turning process. **The International Journal of Advanced Manufacturing Technology**, v. 104, n. 9–12, p. 4331–4340, 16 Oct. 2019.

[212] GOMES, G. F. et al. An estimate of the location of multiple delaminations on aeronautical CFRP plates using modal data inverse problem. **International Journal of Advanced Manufacturing Technology**, v. 99, n. 5–8, p. 1155–1174, 16 Nov. 2018.

[213] ALMEIDA, F. A. et al. A Gage Study Applied in Shear Test to Identify Variation Causes from a Resistance Spot Welding Measurement System. **Strojniski Vestnik Journal of Mechanical Engineering**, v. 64, p. 621–631, 2018.

[214] DINIZ, C. A. et al. Optimum design of composite structures with ply drop-offs using response surface methodology. **Engineering Computations (Swansea, Wales)**, 2021.

[215] SANTOS, A. C. O. et al. Customer value in lean product development: Conceptual model for incremental innovations. **Systems Engineering**, p. sys.21514, 6 Oct. 2019.

[216] CORRÊA, J. É. et al. Development of a System Measurement Model of the Brazilian Hospital Accreditation System. **International journal of environmental research and public health**, v. 15, n. 11, 2018.

[217] GONÇALVES, W. L. et al. A numerical-experimental evaluation of beams composed of a steel frame with welded and conventional stirrups. **Computers and Concrete**, v. 22, n. 1, p. 27–37, 2018.

[218] ALMEIDA, F. A. DE et al. Variation causes analysis attributed to different metrological instruments to verify the geometric characteristics of a spot welding process. **Soldagem e Inspecao**, v. 23, n. 4, p. 485–504, 2018.

[219] AMARAL, F. F. et al. Application of the response surface methodology for optimization of the resistance spot welding process in AISI 1006 galvanized steel. **Soldagem & Inspeção**, v. 23, n. 2, p. 129–142, 28 Jun. 2018.

[220] ALMEIDA, F. A. et al. Measurement data from bobbins of Partially Oriented Yarns: Univariate and multivariate aspects. **Data in Brief**, v. 27, p. 104637, Dec. 2019.

[221] ALMEIDA, F. A. DE et al. A linear programming optimization model applied to the

decision-making process of a Brazilian e-commerce company. **Exacta**, v. 17, n. 3, p. 149–157, 30 Sep. 2019.

[222] ALMEIDA, F. A.; SILVA, A. S. B.; ALMEIDA, F. A. Analysis of Concentration Measures Applied to Manufacture Industry in a State of Brazilian Northeast. **International Journal of Engineering Applied Sciences and Technology**, v. 2, n. 7, p. 9–13, 2017.

ANEXO A – Artigos publicados em periódicos

Artigo "*Enhancement of discriminatory power by ellipsoidal functions for substation clustering in voltage sag studies*", publicado na "***Electric Power Systems Research***" [62].

Electric Power Systems Research 185 (2020) 106368

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Enhancement of discriminatory power by ellipsoidal functions for substation clustering in voltage sag studies

Fabricio Alves de Almeida[a,*], Jacques Miranda Filho[b], Leandro Framil Amorim[a], José Henrique de Freitas Gomes[a], Anderson Paulo de Paiva[a]

[a] Federal University of Itajubá, Av. BPS, 1303, Pinheirinho, Itajubá, MG, Brazil
[b] IFES Federal Institute of Espírito Santo, Estrada da Tartaruga, s/n – Muquiçaba, Guarapari, ES, Brazil

ARTICLE INFO

Keywords:
Voltage sag
Substation cluster
Factor analysis
ANCOVA
Confidence ellipsoid
Power quality regulation

ABSTRACT

Herein, an innovative methodology is proposed to improve the discriminatory power in the estimation of voltage sag patterns and substation groupings attributed to the power quality distribution. Regulation for the Brazilian energy context was considered for the analysis of real data. For this, factor analysis with varimax rotation was applied, which favors the explanation of latent variables. On the basis of these considerations, substation clusters were formed according to the level of similarity of the voltage sag based on the Ward's method. After that, the ellipses of confidence for the clusters were proposed and it was possible to estimate clusters for voltage sag regulatory purposes at three levels: low, medium, and high numbers of events. To prove the efficiency of this approach, the design of the experiments considered different multivariate configurations, linkage methods, and analysis. Then, simultaneous optimization was performed to verify the optimal parameterization that reduced the variance in the clustering estimation. The method was applied in different scenarios to verify the robustness for estimating cluster patterns. The method promoted better discriminatory power for the ellipsoidal functions to estimate the voltage sag patterns and substation groupings, providing results that are more reliable, precise, and stable.

1. Introduction

Research on quality improvements is ongoing in several sectors of power quality, where the quality of the electric supply directly influences industrial processes [1]. The quality of electricity distribution is of great importance, and the voltage sag is a parameter of considerable concern in power quality improvement. Voltage sag has a direct impact on losses in manufacturing processes, in addition to other effects on power distribution equipment in industrial systems with sensitive loads [2]. The importance of the topic was verified in studies such as that of Santis *et al.* [3], which assessed the origin of voltage sag resulting from failures in real and interconnected networks. In the same way, Liao and Anani [4] used artificial neural networks to identify faults in the estimation of the voltage sag state. Nagata *et al.* [5] discussed the overvoltage sag, fault detection, and segmentation based on independent component analysis. In other works, such as that of Santos and Barros [6], an attempt was made to predict the amplitude and duration of voltage sag in network planning. Costa *et al.* [7] presented a method to determine the site index considering the sensitivity of the equipment to voltage sags, in addition to varying the degree of sensitivity and influence of unbalanced sags (in three-phase loads). In the work of Wang *et al.* [8], they analyzed nonsimultaneous trips of protections on two sides of the line and the protection failure (in short-circuit faults), and they proposed an approach for stochastic assessment of voltage sag based on kernel density estimation with a fault position method. In other words, voltage sag stands out as an important parameter in energy distribution systems, because it makes it possible to know the quality level of the substations in the power distribution system, and consequently, to classify the different suppliers of the system.

Analyzing this performance is an important task because power quality regulatory agencies, as in the Brazilian context, seek to map substations based on voltage sags. Many researchers have used clustering techniques to determine the clusters of lower quality, and thus, to apply quality control and regulation measures, as verified in the study of Miranda Filho *et al.* [9].

Methods for estimating patterns and groupings are widely used in the electrical sector, such as clustering techniques [10–13] that use nonhierarchical k-means analysis techniques. Some methods may perform better than others, such as that in a study [9] in which fault simulations and network modeling at the distribution and transmission

* Corresponding author.
E-mail address: fabricio.alvesdealmeida@gmail.com (F.A.d. Almeida).

https://doi.org/10.1016/j.epsr.2020.106368
Received 3 July 2019; Received in revised form 9 March 2020; Accepted 2 April 2020
Available online 16 May 2020
0378-7796/

Artigo "*Multivariate data quality assessment based on rotated factor scores and confidence ellipsoids*", publicado na "***Decision Support Systems***" [8].

Decision Support Systems 129 (2020) 113173

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Multivariate data quality assessment based on rotated factor scores and confidence ellipsoids

Fabrício Alves de Almeida[a,*], Rodrigo Reis Leite[a], Guilherme Ferreira Gomes[b], José Henrique de Freitas Gomes[a], Anderson Paulo de Paiva[a]

[a] Institute of Industrial Engineering and Management, Federal University of Itajubá, Brazil
[b] Mechanical Engineering Institute, Federal University of Itajubá, Brazil

ARTICLE INFO

Keywords:
Data quality assessment
Decision-making
Multivariate measurement system
Factor analysis
Varimax rotation
Confidence ellipsoid

ABSTRACT

This study explores the nature of the correlation in data to estimate the data quality to be used in decision-making processes. The main contribution of this research is the introduction of a new multivariate method based on rotated factor scores by varimax strategy for the repeatability and reproducibility study to effectively identify possible data of poor quality leading to measurement errors. In addition, a new confidence ellipsoid-based decision support method is developed. The efficiency of the proposed method was demonstrated using the metallographic measurements of the geometric characteristics of the resistance spot welding process. To prove the efficiency of the proposed method, it was compared with other consolidated techniques such as the analysis of variance, weighted principal components method, and factor analysis without rotation. Thus, we verified that the proposed method performed better interpretation of the latent information, minimizing the dimensionality of the data, and separating the quality attributes analyzed by clusters. One response group was classified as acceptable, and the other as marginal. These results were verified by the confidence ellipsoids, in which the proposed method obeyed the Bonferroni bilateral limits, outlining the factors which demonstrated superior discriminatory power with non-overlapping ellipsoids avoiding the confounding and favoring the better data quality analysis for multicriteria decision-making. When compared with the other approaches, the proposed method demonstrated more reliable and robust results without such deficiencies as inversion of the groupings, neglection of the variance-covariance structure, and the variability attributed to the data within the measurement system.

1. Introduction

Improvements in industrial processes aimed at cost reduction and quality improvement [1] are widely discussed. The researchers seek to introduce innovative methodologies based on mathematical modeling to maximize the efficiency and to improve the decision-making in these processes. Among these proposals is the study conducted by McHaney and Douglas [2], in which they developed a multivariate regression metamodel of a decision support system (DSS) for the task of daily resource allocation in an industry. Gomes et al. [3] used an approach based on artificial neural network (ANN) modeling together with a genetic algorithm for damage detection in carbon fiber reinforced polymer (CFRP) aeronautical plates aiming to create a DSS to provide more precise decision-making for the coupling of sensors in commercial aircrafts. We can also highlight here the work of Gaudencio et al. [4], in which they used the fuzzy decision-making strategy together with the mean square error multivariate approach for the identification of optimal parameters in robust estimators applied to AISI 12L14 free-machining steel-turning.

However, focusing all efforts on the exclusive improvement of the process may not yield a satisfactory result, as variability can often be attributed to the measurement process [5], which may compromise the quality of the data to be analyzed by the decision maker. According to Moges et al. [6], among the many factors that can affect the decision-making process, data quality is the most critical. According to these authors, the poor-quality data may lead to poor decision-making. Thus, they highlight the data quality issue as one of the most crucial problems in many industries. Heinrich and Klier [7] state that data quality assessment has been extensively discussed in the literature related to fields in which high-quality data is required for various business or decision-making processes. Furthermore, Timmerman and Bronselaer [8] infer that data quality is of great interest for the scientific research.

* Corresponding author.
E-mail address: fabricio-almeida@unifei.edu.br (F.A. Almeida).

https://doi.org/10.1016/j.dss.2019.113173
Received 11 April 2019; Received in revised form 30 September 2019; Accepted 26 October 2019
Available online 31 October 2019

Artigo "*A new multivariate approach based on weighted factor scores and confidence ellipses to precision evaluation of textured fiber bobbins measurement system*", publicado na "***Precision Engineering***" [188].

Precision Engineering 60 (2019) 520–534

Contents lists available at ScienceDirect

Precision Engineering

journal homepage: http://www.elsevier.com/locate/precision

A new multivariate approach based on weighted factor scores and confidence ellipses to precision evaluation of textured fiber bobbins measurement system

Fabrício Alves de Almeida [a,*], Guilherme Ferreira Gomes [b], Juliana Helena Daroz Gaudêncio [a], José Henrique de Freitas Gomes [a], Anderson Paulo de Paiva [a]

[a] Institute of Industrial Engineering and Management, Federal University of Itajubá, Brazil
[b] Institute of Mechanical Engineering, Federal University of Itajubá, Brazil

ARTICLE INFO

Keywords:
Textured fiber bobbins
Multivariate measurement system analysis
Repeatability and reproducibility
Principal component factor analysis
Confidence ellipse

ABSTRACT

In a process that has multiple correlated characteristics, it is more appropriate to use multivariate techniques. This study proposes a multivariate gage repeatability and reproducibility ($GR\&R_m$) approach based on weighted principal component factor analysis and confidence ellipses for the number of distinct categories (ndc). The method was applied to the textured fiber bobbins process to analyze the measurement system of a leading textile company in Brazil, at its factory responsible for 15% of its total production. Among quality characteristics, measurements of mass, density, and diameter of the bobbins were analyzed. The analysis is divided into two main groups: the initial situation of the measurement system, and the verification measurements taken after calibration adjustments. The adjustments were made to the calibration of the equipment and significantly reduced the variability of the measurements, with the proposed method providing adequate identification of the variability, describing the confidence ellipses of the ndc in a non-overlapping manner, and classifying the measurement system as acceptable. In this study, it was possible to verify that the proposed strategy yields satisfactory results, minimized the dimensionality of the data evaluated, reduced the time and isolated analyses considering the variance-covariance structure. Consequently, it was possible to validate the company's measurement system, bringing greater reliability to actions performed by its process control division.

1. Introduction

The textile industry is one of the base industries that has evolved into manufacturing due to market needs [1], where ensuring control and optimization of processes is of paramount importance to ensuring product quality. According to Atilgan [2], tolerance limits for product quality are increasingly critical, necessitating statistical process control in the textile & apparel sector.

Texturing is characterized by modification of synthetic and natural polymer fibers (which takes advantage of their thermoplasticity), whereby the fibers undergo thermofixation and draw. According to Bernard [3], the thermofixation process (or heat treatment) is characterized by heating where the upper limit is given by the melting temperature and the lower limit by the glass transition temperature of the fiber, which is essential for breakage of its secondary bonds. Consequently, thermofixation seeks to improve dimension stability, suppress the internal tension of the fiber (reduce the shrinkage thereof), and make the fiber homogeneous. Stability of dyebaths is a crucial factor in avoiding color variation of the yarns [4], directly influencing the final quality of the product.

Equally important, according to Salerno-Kochan [5], the quality of textile products has a significant impact on consumer perception. Several techniques of quality improvement have been applied to manufacturing processes, including texturization, as shown by Silva et al. [6], who carried out parameter modeling and optimization for the false-twist texturing of the polyester process. Manufacturing companies aim at quality and cost as their primary objectives [7,8], however, targeting quality improvements only for capacity may not significantly favor the process. When process capability is already high, the variability of the measurement system and the manufacturing system is decisive for decision making [9,10].

In quality improvement projects, such as Six Sigma, before analyzing

* Corresponding author.
E-mail address: fabricio-almeida@unifei.edu.br (F.A. Almeida).

https://doi.org/10.1016/j.precisioneng.2019.09.010
Received 14 May 2019; Received in revised form 6 August 2019; Accepted 13 September 2019
Available online 14 September 2019

Artigo "*A Gage Study Through the Weighting of Latent Variables Under Orthogonal Rotation*", publicado na "***IEEE Access***" [7].

IEEE Access
Multidisciplinary | Rapid Reviews | Open Access Journal

Received August 4, 2020, accepted August 18, 2020, date of publication August 24, 2020, date of current version October 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019031

A Gage Study Through the Weighting of Latent Variables Under Orthogonal Rotation

FABRICIO ALVES DE ALMEIDA[1,2], SIMONE CARNEIRO STREITENBERGER[1], ALEXANDRE FONSECA TORRES[1], ANDERSON PAULO DE PAIVA[1], AND JOSÉ HENRIQUE DE FREITAS GOMES[1]

[1]Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá 37500-903, Brazil
[2]Department of Economics, Faculty of Economic Sciences Southern Minas Gerais, Itajubá 37504-066, Brazil

Corresponding author: Fabricio Alves de Almeida (fabricio.alvesdealmeida@gmail.com)

This work was supported in part by FAPEMIG (Fundação de Amparo a Pesquisa do Estado de Minas Gerais), in part by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), in part by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), and in part by FAPEPE (Fundação de Apoio ao Ensino, Pesquisa e Extensão de Itajubá).

ABSTRACT A new approach to identify and diagnose the quality of extensive and multivariate data is presented, using the gage repeatability and reproducibility (GR&R) study through the weighting of rotated factor scores. The proposal uses axis rotation to improve the explanation and interpretations of latent information, providing a statistically appropriate alternative when dealing with two or more correlated data sets. To analyze data with a significant variance-covariance structure, factor analysis (FA) is applied for calculating the eigenvalues and extracting of the rotated scores. Once obtained, these scores are then weighted with their respective eigenvalue for each factor. This procedure results in a single response vector, which is capable of properly interpreting all of the quality responses analyzed. To illustrate an application of the method, a real data set from a resistance spot welding process is selected, and two different types of rotation are compared. The proposed method provided an output that contemplated all of the significant variability of the data in a unique and significant way. In addition, the method enabled a reduction in the data dimensionality, thus minimizing the time for analysis and computational effort.

INDEX TERMS Multivariate measurement system, repeatability and reproducibility, orthogonal rotation, weighted factor analysis, resistance spot welding.

I. INTRODUCTION

Multivariate statistical techniques are widely used to analyze data that has a significant variance-covariance structure [1]. Such methods have been applied in many engineering problems to improve the interpretation of extensive and correlated data. In fact, several studies already use multivariate strategies in a handful of applications, such as flux-cored arc welding process [2], moving average control chart [3], design of experiments on clustering methods [4] and applications in process monitoring [5], [6]. Such approaches are also used in the energy [7], healthcare [8] and economy [9] sectors. Among several methods, some of them stand out in view of their characteristics. The principal component analysis (PCA), for instance, is a multivariate strategy that reduces the data dimensionality and promotes uncorrelated vectors, considering its variance-covariance structure [10], [11]. PCA has been used in several applications focused on quality improvement, such as the studies of [12]–[16].

Another widely used approach is the factor analysis (FA), which promotes the grouping of characteristics based on their explanation level [17]. FA has some advantages over the PCA technique. FA provides a better interpretation and explanation of the data with a simpler structure [1]. FA also enables the reduction of repetitive information between variables, using a smaller amount of latent variables [18]. Another advantage is that FA allows the grouping of the variables observed in relation to the factor loads. For example, in a suitable application, one factor would have a high factor load value, while the other factors would have small or moderate loads [1]. Such a characteristic would favor the simplicity of the structure and, consequently, the explanation of the data. However, this structure is not always obtained [17], so it is often recommended to use methods to rotate the axes of the factors to improve the explanation of the variables. The purpose of this rotation approach is to acquire a simple data structure, with easy interpretation of the observed variables [19].

The associate editor coordinating the review of this manuscript and approving it for publication was Qingchao Jiang.

Artigo "*An inverse damage location problem applied to AS-350 rotor blades using bat optimization algorithm and multiaxial vibration data*", publicado na "***Mechanical Systems and Signal Processing***" [204].

Mechanical Systems and Signal Processing 145 (2020) 106932

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

An inverse damage location problem applied to AS-350 rotor blades using bat optimization algorithm and multiaxial vibration data

Guilherme Ferreira Gomes [a,*], João Artur Souza Chaves [a], Fabricio Alves de Almeida [b]

[a] Mechanical Engineering Institute, Federal University of Itajubá – UNIFEI, Brazil
[b] Institute of Industrial Engineering and Management, Federal University of Itajubá – UNIFEI, Brazil

ARTICLE INFO

Article history:
Received 17 December 2019
Received in revised form 2 April 2020
Accepted 22 April 2020

Keywords:
Damage identification
Inverse problem
Bat optimization algorithm
Helicopter rotor blade
RSM
Structural health monitoring

ABSTRACT

In this study, a damage identification method is proposed using both the finite element method and the bat optimization algorithm applied to the AS-350 helicopter main rotor blade. First, the structure is numerically modeled and evaluated with and without the presence of induced damages. In a second approach, an inverse problem of optimization is constructed in order to identify certain damages in terms of its position and severity level. Three different objective functions are evaluated according to the modal parameters of the rotor blade (vibrations in *x*, *y* and *z* directions). Numerical results, through analysis of variance, showed that local damage significantly modifies the modal response into a non-linear aspect. The modal response used was able to identify, with great efficiency, the actual (noise simulated) damages induced in terms of location and severity. Accordingly, a damage identification method is developed in order to better handle any measurement data (to find/regarding) structural changes (or damages) in complex aerospace structures. The obtained results from these numerical examples indicate that the proposed approach can detect true damage locations and estimate damage magnitudes with satisfactory accuracy, even under high measurement noise.

1. Introduction

The structural condition depends on the operating regime of the aircraft, namely its use (operational load) that conditions the consumption of its useful life. The knowledge of the current state of structural components (structural health) is important as it allows to diagnose its condition, and the knowledge of the consumption of useful life allows to make a forecast of the remaining time of life. The introduction of Structural Health Monitoring (SHM) systems in aircraft has been increasing, given its recognized benefit, however older aircraft do not always have the necessary instrumentation [35].

Equally important, there are the global and local methods, which use different tools such as ultrasound, radiograph, eddy-current and thermal fields [9]. Although these methods are effective, they need a previous knowledge of the damaged region, but in the practical application, such information is almost never known. Finally, the global methods are mentioned, where

* Corresponding author.
E-mail addresses: guilhermefergom@unifei.edu.br (G. Ferreira Gomes), artur.joao@yahoo.com.br (J.A. Souza Chaves), fabricio.alvesdealmeida@gmail.com (F.A. de Almeida).

https://doi.org/10.1016/j.ymssp.2020.106932
0888-3270/

Artigo "*Tuning Metaheuristic Algorithms using Mixture Design: Application of Sunflower Optimization for Structural Damage Identification*", publicado na "***Advances in Engineering Software***" [113].

Advances in Engineering Software 149 (2020) 102877

Contents lists available at ScienceDirect

Advances in Engineering Software

journal homepage: www.elsevier.com/locate/advengsoft

Research paper

Tuning metaheuristic algorithms using mixture design: Application of sunflower optimization for structural damage identification

Guilherme Ferreira Gomes[a,*], Fabricio Alves de Almeida[b]

[a] Mechanical Engineering Institute, Federal University of Itajubá - UNIFEI, Brazil
[b] Institute of Industrial Engineering and Management, Federal University of Itajubá - UNIFEI, Brazil

ARTICLE INFO

Keywords:
Structural health monitoring
Inverse problem
Sunflower optimization
Mixture design
Genetic algorithm

ABSTRACT

This paper presents an efficient inverse global optimization approach for damage identification of plate-like structures. In this approach, the damage identification process is performed by minimizing an objective function based on modal parameters of CFRP laminated structures. The identification process entails two steps: *i)* the direct problem is modeled using the finite element method. Damage is induced into the two different situations, first as a variation in physical properties, i.e., delamination, as a variation in stiffness and also as a variation in the grommet properties, for example small circular holes; *ii)* For solving the optimization problem, an enhanced SunFlower Optimization (SFO) algorithm is applied in the inverse problem methodology. The SFO metaheuristic algorithm has its biological operators optimized by mixture design method. The efficiency of the proposed identification is investigated through two numerical examples for laminated composite plates where Genetic Algorithm, SFO and an improved SFO algorithm are compared. The obtained results indicate that the proposed Structural Health Monitoring method can successfully identify the location and the severity of small induced damage cases in the laminated composite plate. In addition, the improved algorithm was shown to be more efficient and accurate than the widely known and applied Genetic Algorithm.

1. Introduction

Nowadays, there has been a huge and increasing volume of research on structural health monitoring since the 1970s with 17,000 papers published just in the last decade [1]. However, this enormous research effort has yielded only a small number of routine industrial applications.

It is well known that composite structures offer numerous advantages over conventional structural systems in the form of higher specific stiffness and strength, lower life-cycle costs, and benefits such as easy installation and improved safety [2]. However, composite materials can deteriorate due to the excessive use, cyclic loading, climatic conditions and deficiencies in inspection methods. In particular, visual inspections are usually time-consuming, costly and require components readily accessible. Other conventional methods (acoustic emission, ultrasonic methods, thermography, x-rays and others) for detecting damage in composites are often costly and depend heavily on the skill and experience of the operator. Structural Health Monitoring (SHM) techniques applied to composite structures offer a promising alternative and involve continuous monitoring of a structure using a non-destructive inspection (NDI) together with integrated sensors and advanced computational techniques [3].

The basic idea of an SHM system is to provide a structure of interest with detection and analysis capabilities, and to allow monitoring and evaluation to be performed periodically or continuously autonomously. The SHM method in principle offers higher security, since failures do not evolve to an alarming level.

As the potential benefits of this incorporation of SHM are enormous, a great deal is underway worldwide to incorporate some degree of 'self-diagnosis' ability into man-made structures [4].

In other words, the process of detecting or identifying damage is essentially an inverse problem, where input(s) and output(s) are known, and, through these, the location of a present damage in a given structure can be determined. There are several recent works that address the most diverse types of computational methodologies aimed at detecting damage in structures, such as Artificial Neural Networks (ANN) [5–7], Fuzzy logic [8–10], intelligent signal processing [11–13], optimization algorithms such as Genetic Algorithm (GA) [14–16], Particle Swarm Optimization [17–19] and others non traditional heuristics [20–22].

In essence, damage detection is a system identification problem, where for a given set of input-output parameters the location and extent

* Corresponding author.
E-mail address: guilhermefergom@unifei.edu.br (G.F. Gomes).

https://doi.org/10.1016/j.advengsoft.2020.102877
Received 24 June 2019; Received in revised form 22 June 2020; Accepted 28 June 2020
Available online 17 August 2020

Artigo "*Optimized damage identification in CFRP plates by reduced mode shapes and GA-ANN methods*", publicado na "***Engineering Structures***" [205].

Engineering Structures 181 (2019) 111–123

Contents lists available at ScienceDirect

Engineering Structures

journal homepage: www.elsevier.com/locate/engstruct

Optimized damage identification in CFRP plates by reduced mode shapes and GA-ANN methods

Guilherme Ferreira Gomes[a,*], Fabricio Alves de Almeida[b], Diego Morais Junqueira[a], Sebastiao Simões da Cunha Jr.[a], Antonio Carlos Ancelotti Jr.[a]

[a] Mechanical Engineering Institute, Federal University of Itajubá, Brazil
[b] Institute of Industrial Engineering and Management, Federal University of Itajubá, Brazil

ARTICLE INFO

Keywords:
Damage identification
Sensor placement optimization
Structural health monitoring
Inverse problem
Artificial neural networks
Composite plates

ABSTRACT

Delamination is one of the most common failure mode in laminated composites that leads the separation along the interfaces of the layers. The structural performance can be significantly affected by this degradation. Such damages are not always visible on the surface and could potentially lead to catastrophic structural failures. The existence of delamination alters the vibration characteristics of the laminated structures, so if they are detected and measured previously, they can be used as indicator for quantifying health and the potential risk of catastrophic failures. To ensure structural performance and integrity, accurate Structural Health Monitoring (SHM) is crucial. In this study, an optimized methodology for delamination identification on laminated composite plates involving the use of reduced mode shapes and computational tools, i.e., Genetic Algorithm (GA) and Artificial Neural Networks (ANN) is performed. In a first step, the sensor distribution on the surface of the structure was optimized using Fisher Information Matrix (FIM) criteria. After, GA and ANN were applied in order to identify and predict delamination location. A feed-forward based neural network is used in order to detect damage on the laminated plate using data obtained from Finite Element Analysis (FEA). The present methodology identifies damage localization in structures and also quantifies damage severity. The applicability of the technique is demonstrated on laminated plates and results are compared with numerical algorithms. This paper shows the effectiveness of GA and ANN as tools for delamination damage identification problem. The algorithms in their inverse formulations are capable of predicting accurately delamination position in plates-like structures.

1. Introduction

The application of advanced materials such as composites in components and structures have envolved in last decades due to the need for improvements in terms of weight/strength ratio performance. In addition, good fatigue strength and high structural performance, have also been significant contributions to the rapid increase of the application of those materials. However, according to [29], composite materials are subjected to high structural load demands due to their good mechanical performance. Due to high loads, damage may occur internally to the material. Knowing the damage and Being able to detect, locate and identify the type of damage in a structure is crucial for engineers. However, the development of techniques capable of detecting, locating, identifying and characterizing composite damages are still considered a challenge. Actually, most of the non-destructive inspection techniques applied in composite structures require high levels of operator experience since the inspection procedure and the interpretation of the results are very complex steps.

A reliable and effective non-destructive damage identification method is crucial to maintain the safety and integrity of mechanical structures (aircrafts, ships, buildings, etc.). The most common non-destructive damage identification techniques include visual inspection and conventional nondestructive testing (NDT), such as tap coin, ultrasonic inspection, penetrating liquids and thermography. However, the visual inspection techniques are unable to detect damage which is embedded in a structure or invisible to human eyes while the conventional NDT requires that the vicinity of damage be known a priori and readily accessible for testing [12].

Vibration-based damage identification method has been widely used as NDT [23,24]. Many approaches based on this method have been proven to be effective in addressing problems in both basic and complex structures [20]. This approach explains that damage can affect both the physical and dynamic characteristics of the structural properties. Physical characteristics include the mass, stiffness and damping, while

* Corresponding author.
E-mail address: guilhermefergom@unifei.edu.br (G.F. Gomes).

https://doi.org/10.1016/j.engstruct.2018.11.081
Received 29 March 2018; Received in revised form 30 October 2018; Accepted 30 November 2018
Available online 11 December 2018

Artigo "*A multivariate GR&R approach to variability evaluation of measuring instruments in resistance spot welding process*", publicado na "***Journal of Manufacturing Process***" [146].

Journal of Manufacturing Processes 36 (2018) 465–479

Contents lists available at ScienceDirect

Journal of Manufacturing Processes

journal homepage: www.elsevier.com/locate/manpro

Technical Paper

A multivariate GR&R approach to variability evaluation of measuring instruments in resistance spot welding process

F.A. Almeida[a,*], T.I. De Paula[a], R.R. Leite[a], G.F. Gomes[b], J.H.F. Gomes[a], A.P. Paiva[a], P.P. Balestrassi[a]

[a] Institute of Industrial Engineering, Federal University of Itajubá, Itajubá, MG 37500-903, Brazil
[b] Mechanical Engineering Institute, Federal University of Itajubá, Itajubá, MG 37500-903, Brazil

ARTICLE INFO

Keywords:
Measuring instruments
Multivariate gage study
Repeatability and reproducibility
Resistance spot welding
Weighted principal components

ABSTRACT

The resistance spot welding is a promising and widely used method for joining mechanical structures, being applicable in various industrial sectors and constantly improved in order to guarantee a product with high reliability and mechanical quality. This paper aims to verify the variability attributed to different measuring instruments for this process, considering the multivariate nature of its critical-to-quality characteristics. In order to verify the degree of reliability of the results in this process, the multivariate repeatability and reproducibility study based on the weighted principal components method was used, considering two different responses. Design of experiments methodology was used for collecting the data, in order to obtain a real representation of the process amplitude. The measurements were evaluated by image analyzer and conventional metrology mechanical instruments. The results showed a better precision of the metrics performed by the image analyzer, where the measurements presented lower variability, while the mechanical instruments showed greater uncertainties for the measured values.

1. Introduction

Resistance spot welding (RSW) is a process that uses heat from an electric current for joining metallic structures. Although RSW is widely used in different segments of the industry, it is worth highlighting its use in the automotive industry, especially as a welding process for steel sheet products [1,2]. Among the main features of the RSW process are high operating speeds [3] and suitability for automation [4].

To ensure that the weld maintains constant quality during the production process two factors should be considered: (1) optimum welding parameters and (2) process control [5]. Due to the importance of weld quality, methodologies for defining optimal welding parameters are widely used for process improvement, contributing to the control and capability of the RSW process.

Several specific methods of quality assessment can be used to verify dimensionality measurements of the weld point, such as electrode displacement [6], X-ray [7] and ultrasound [8]. In addition to these approaches, the metallographic analysis is used to analyze the geometrical characteristics of the welding process [9–11]. The geometrical characteristics are directly related to weld strength, such as the weld button size characteristics [12]. Thus, this study evaluates the indentation depth and nugget width, which are the geometric quality characteristics that have major impact on products from RSW process.

The high quality of the products are one of the main targets of manufacturing companies [13]. To reach these conditions, it is critical to control the variability of manufacturing processes. Part of the total variability of this process is due to the variability associated with the measurement system (special cause) and the other part is due to the process itself (common cause) [14,15]. Thus, in order to monitor and improve processes such as RSW, it is extremely necessary to determine the capability of the measurement system [16,17].

In order to evaluate the capability of measurement systems, one can use gage repeatability and reproducibility (GR&R) studies [18]. The most used methods for performing GR&R studies is the Analysis of Variance (ANOVA) approach, that segregates the variance of the measurement system into two components: variance due to repeatability and variance due to reproducibility. The GR&R strategy is widely used for measurement system analysis (MSA) as in [14,19–21].

In manufacturing processes, several outcomes are taken into account for process monitoring and improvement. These outcomes are usually statistically correlated, and when the variance–covariance structure among these outcomes is not considered, it can lead to inaccurate analysis and incorrect conclusions. Hence, in order to overcome the aforementioned issue of correlated outcomes in multivariate

* Corresponding author.

https://doi.org/10.1016/j.jmapro.2018.10.030
Received 17 July 2018; Received in revised form 11 October 2018; Accepted 24 October 2018

Artigo "*A multiobjective optimization model for machining quality in the AISI 12L14 steel turning process using fuzzy multivariate mean square error*", publicado na "***Precision Engineering***" [12].

Precision Engineering 56 (2019) 303–320

Contents lists available at ScienceDirect

Precision Engineering

journal homepage: www.elsevier.com/locate/precision

A multiobjective optimization model for machining quality in the AISI 12L14 steel turning process using fuzzy multivariate mean square error

Juliana Helena Daroz Gaudêncio[a,*], Fabrício Alves de Almeida[a], João Batista Turrioni[a], Roberto da Costa Quinino[b], Pedro Paulo Balestrassi[a], Anderson Paulo de Paiva[a]

[a] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, MG, 37500-903, Brazil
[b] Department of Statistics, Federal University of Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil

ARTICLE INFO

Keywords:
Robust parameter design
Normal boundary intersection
Principal components analysis
Fuzzy decision maker

ABSTRACT

Organizations focus on determining optimal operating conditions to ensure quality; however, industrial processes exhibit a high degree of variability and the use of robust estimators is a suitable alternative to model experimental data. As a case study, the surface roughness (R_a) of an AISI 12L14 steel turning process is optimized to find a centrality measure close to its target with minimum dispersion and thus improve the quality of the machined surface by choosing the best values of the associated parameters. The main contribution of this research is the proposal of a multiobjective optimization method that uses principal components analysis to minimize the redundancy of objective functions in terms of multivariate mean square error, thus making optimization of the process possible with a better explanation of all centrality and dispersion estimators proposed herein. The method uses a fuzzy decision maker to show the surface roughness' optimum result with the most efficient production taken into consideration. To prove its efficiency, confirmation runs were conducted. At a confidence level of 95%, the optimal value falls within the multivariate confidence intervals only for Model B, in which the estimators' median and median absolute deviation are considered, thus affirming which pair of estimators achieves the most robust parameter design solution. Through the proposed research, the developed model can be used in industries for determining machining parameters to attain high quality with minimum power consumption and hence maximum productivity.

1. Introduction

Organizations that try to focus on optimum operating conditions must ensure quality and continuously search for improvements. They aim at minimizing the uncertain measurements that provide variation and affect accuracy. Measurement uncertainties affect the responses (y) and also the predictor variables (x); therefore, it is reasonable to use robust analysis tools to minimize these effects. In that scope, the design of experiments (DOE) approach is a particularly excellent tool for optimizing certain process quality characteristics [1]. Many researchers have applied the response surface methodology (RSM), in which one aims to discern answers that are influenced by some variables to optimize results. The principal objective is to obtain an adequate model from a sequence of designed experiments to obtain optimal operating conditions for the process [2,3].

Experimentation directs the development and improvement of existing processes and facilitates the development of robust processes that are minimally affected by variation of external agents. The processes that produce minimal variation in the presence of noise fall under the category of robust parameter design (RPD), which is an engineering methodology intended to be a cost-effective approach to improving product quality [4].

RPD was allied with the RSM approach by Myers and Carter [5]. They were aiming to solve problems in which the experimenter was able to identify a primary response to be optimized by being limited to a specific value of average or variance as the secondary response. This idea was popularized by Vining and Myers [6], who pointed out that the objective of optimizing the mean $\mu(x)$ and variance $\sigma^2(x)$ simultaneously could be achieved through the dual-response surface method.

Consequently, estimators of mean and variance are normally used. Industrial processes, however, often exhibit high variability in repeated observations. This approach raises a question pointed out by Boylan and Cho [7]: Which robust estimators achieve the best RPD solutions? Therefore, in this work, we compare eight dual-response surface models to define the best. The first model (Model A) comprises a sample mean

* Corresponding author.
E-mail addresses: juliana_hdg@yahoo.com.br, ju.gaudencio@gmail.com (J.H. Daroz Gaudêncio).

https://doi.org/10.1016/j.precisioneng.2019.01.001
Received 4 October 2018; Received in revised form 21 December 2018; Accepted 7 January 2019
Available online 08 January 2019
0141-6359/

Artigo "*Multivariate steepest ascent method based on latent variables*", publicado na "***Applied Mathematical Modelling***" [206].

Applied Mathematical Modelling 90 (2021) 30–45

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

Multivariate steepest ascent method based on latent variables

Paulo Roberto Maia, Fabrício Alves de Almeida*, Vinicius de Carvalho Paes, José Henrique de Freitas Gomes, Anderson Paulo de Paiva

Institute of Industrial Engineering and Management, Federal University of Itajubá, 1303 BPS Avenue, Itajubá, MG 37500-903, Brazil

ARTICLE INFO

Article history:
Received 8 May 2020
Revised 31 August 2020
Accepted 3 September 2020
Available online 22 September 2020

Keywords:
Path of steepest ascent
Principal component analysis
Design of experiments
Flux-cored arc welding (FCAW)

ABSTRACT

This paper presents a multivariate steepest ascent method based on the gradient of the first order principal component score model, with direction, step sizes and shifts driven by an integrated variance mapping. Using a random initial center point guess within regions of minimal prediction error, gradual improvements are done towards the curvature region where a response surface may be properly fitted. Experimentations carried out in such regions allow a large step size since coefficients standard error are very low. In order to illustrate this approach, a Flux-Cored arc welding cladding process of AISI 1020 carbon steel sheets with AISI 316L stainless steel tubular wires was studied considering a full factorial design with four input parameters for correlated pairs of responses. The case study and additional simulations highlights the suitable optimization results obtained with the method and its practical and successful implementation in a real-word manufacturing problem.

1. Introduction

Steepest Ascent Method (SAM) is a gradient-based search algorithm that considers the ratio between the coefficients of a first order model as a lead-off track to experimentally move the first center point vector toward the direction of maximum increase for the objective function [1]. Having been originally developed for just one response, SAM applicability for the multiresponse case remains underexplored. Among the papers treating multiresponse steepest ascent, two approaches are observed: (1) the weighted priority method (**WP**) [2,3] and (2), the Desirability-based (**D**) search direction [4,5]. **WP** consists of a path of steepest ascent based on the individual gradients of original responses and their respective relative priorities (π_i) computed from a first order model (generally estimated from a 2^k full factorial design or a 2^{k-p} fractional factorial design, where k is the number of factors used in the experiment) combined as a weighted sum. Similarly, Desirability-based approach employs the geometric mean of individual desirability indexes to serve as a common direction for the responses. As will be seen in the next sections, when the individual responses are correlated, **WP** and **D** tends to overestimate the projection of gradients onto the direction of maximum improvement, deviating the center point of the most reliable improvement direction considerably. In order to mitigate the correlationś influence over multiple objective functions coefficients, [3] considers the ridge path estimated by seemingly unrelated regression (SUR) model as well as standard multivariate regression (SMR) model. Applying such multivariate techniques, the formed models were combined using the Desirability index.

Thanks to its remarkable capacity to represent correlated variables as independent linear combinations, Principal Component scores (**PC**) [6] may be used as a suitable alternative to **D** or **WP** agglutinative methods. Using the eigenvectors of

* Correspondent author.
E-mail address: fabricio-almeida@unifei.edu.br (F.A.d. Almeida).

https://doi.org/10.1016/j.apm.2020.09.011
0307-904X/

Artigo "*Multivariate Taguchi Loss Function Optimization Based on Principal Components Analysis and Normal Boundary Intersection*", publicado na "***Engineering with Computers***" [142].

Engineering with Computers
https://doi.org/10.1007/s00366-020-01122-8

ORIGINAL ARTICLE

Multivariate Taguchi loss function optimization based on principal components analysis and normal boundary intersection

Fabrício Alves de Almeida[1,2] · Ana Carolina Oliveira Santos[3] · Anderson Paulo de Paiva[1] · Guilherme Ferreira Gomes[4] · José Henrique de Freitas Gomes[1]

Received: 3 June 2020 / Accepted: 20 July 2020

Abstract
Optimization methods are widely used to improve industrial processes and enhance the quality characteristics of product, where process costs are directly linked. Given this assumption, this study aims to present a multivariate proposal of the Taguchi loss function, to model and optimize manufacturing processes, searching to establish values that prioritize quality and provide the minimum loss in view of the process costs. For this, design of experiments techniques will be used to model the process and the calculated loss functions. The strategy of principal components analysis is used to minimize the data dimension, considering the structure of variance–covariance. Then, the normal boundary intersection method is used to find the Pareto frontier. Based on the values, the method also proposes a total loss function equation, which is characterized as an approach to choose the optimal point based on the sum of the loss functions for the Pareto frontier through the process cost. To demonstrate the behavior of the method, the flux-cored arc welding of stainless-steel cladding process was applied. In view of the results, the method provided an optimal value at the Pareto frontier, contemplating an appropriate balance between minimal loss and higher quality, which were compared with other studies in the literature. The method also provided a reduction in computational effort of approximately 90% (from 210 to 21 subproblems), obtaining the best solution and contemplating the multivariate nature of the data.

Keywords Response surface methodology · Taguchi loss function · Principal component analysis · Normal boundary intersection · Flux-cored arc welding process

1 Introduction

To survive in the competitive industry and meet customer demands, industries should constantly enhance their productive processes. Improvements in quality, cost, flexibility, speed and reliability are frequent challenges faced by the sector. It is possible to find in the literature several articles that present mathematical and computational strategies to improve quality in industrial processes, such as: [1–5]. On the other hand, the relationship between quality and price is a very important factor since price represents a loss for the consumer at the time of purchase and low quality represents an additional loss during the use of the product. This "loss" includes the cost of customer dissatisfaction that leads to the denigration of the company's reputation [6, 7]. This concept is very different from traditional guidelines for producers, and includes rework, waste, warranty and services costs as measures of quality.

For these reasons, Taguchi presented the quadratic quality loss function to redefine the quality of a product. All processes that have a decreasing value for the quality loss function (QLF) can assure that performance has been improved. QLF is a mathematical model that accounts for the quality loss in terms of monetary values resulting from the deviation in quality related to the target specification. When analyzing the QLF of a process, the existence of few variables facilitates the calculation. However, an industrial process with

✉ Fabrício Alves de Almeida
fabricio.alvesdealmeida@gmail.com

[1] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil

[2] Department of Economics, Faculty of Economic Sciences Southern Minas, Itajubá, Brazil

[3] Institute of Integrated Engineering, Federal University of Itajubá, Itabira, Brazil

[4] Mechanical Engineering Institute, Federal University of Itajubá, Itajubá, Brazil

Published online: 29 July 2020

Artigo "*Inverse structural damage identification problem in CFRP laminated plates using SFO algorithm based on strain fields*", publicado na "***Engineering with Computers***" [207].

Engineering with Computers
https://doi.org/10.1007/s00366-020-01027-6

ORIGINAL ARTICLE

Inverse structural damage identification problem in CFRP laminated plates using SFO algorithm based on strain fields

Guilherme Ferreira Gomes[1] · Fabricio Alves de Almeida[2] · Antonio Carlos Ancelotti Jr.[1] · Sebastião Simões da Cunha Jr.[1]

Received: 22 July 2019 / Accepted: 18 April 2020

Abstract
Damage detection methods are an important field of engineering and crucial in terms of structural safety. However, in many practical cases, the process of monitoring and identifying damage is extremely difficult or even impractical due to the conditions of access and operation of a given component/structure. In this study, an inverse algorithm based on strain fields for damage identification in composite plate structures is presented. The inverse analyses combine experimental tests and digital image correlation (DIC) with numerical models based on finite element update method with great advantage of being a non-contact method. The proposed technique identifies the location and dimension of damages in a CFRP plate using static strains formulated as an objective function to be minimized. By model updating, the discrepancies between the experimental and the numerical results are minimized. For the success of the model updating, the efficiency of the optimization algorithm is essential. A powerful new metaheuristic sunflower optimization (SFO) is employed to update the unknown model parameters. Experimental results showed the excellent efficiency in the combined use of DIC, numerical modeling and SFO optimization to accurately identify the location of damage in numerical and experimental tests. The obtained results indicate that the proposed method can be used to determine efficiently the location and dimension of structural damages in mechanical structures.

Keywords Structural health monitoring · Inverse problem · Sunflower optimization · Digital image correlation · Composite plates

1 Introduction

The detection of damages is a field of extreme importance in engineering, since through it corrective maintenance can be applied and in this way structural safety can be guaranteed. A prognosis of the structure can be made from the moment that a damage is correctly detected, thus being able to evaluate the integrity of the structure and determine its life time [1, 2].

In the same way, the application of composite materials has become increasingly constant in several areas of industry, but especially in the aerospace field. Its use is justified due to the fact that this type of material has good mechanical characteristics such as high stiffness, high mechanical strength and stiffness-to-mass ratio [3–5].

At the same time, despite these good mechanical characteristics, composite materials can present certain failures when subjected to extreme conditions such as static overload, impact, fatigue, design errors and overheating [6, 7]. These faults can be translated as matrix microcracking, interface delamination and then a strength redistribution followed by fiber rupture [8, 9]. Most of the methods used to detect damage are currently visual or experimental, such as acoustic or ultrasonic methods, thermography, radiographs, among others. These methods are in most cases time-consuming and costly, thus requiring structures to be located in accessible locations and heavily dependent on the skill and experience of the professional performing the inspection [10, 11].

In view of this scenario, there is a need for more viable structural monitoring methods. In view of this, structural

✉ Guilherme Ferreira Gomes
guilhermefergom@unifei.edu.br

[1] Mechanical Engineering Institute, Federal University of Itajubá, Itajubá, Brazil

[2] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil

Published online: 14 May 2020

Artigo "*A multiobjective sensor placement optimization for SHM systems considering Fisher information matrix and mode shape interpolation*", publicado na "***Engineering with Computers***" [208].

Engineering with Computers
https://doi.org/10.1007/s00366-018-0613-7

ORIGINAL ARTICLE

A multiobjective sensor placement optimization for SHM systems considering Fisher information matrix and mode shape interpolation

Guilherme Ferreira Gomes[1] · Fabricio Alves de Almeida[2] · Patricia da Silva Lopes Alexandrino[1] · Sebastiao Simões da Cunha Jr.[1] · Bruno Silva de Sousa[1] · Antonio Carlos Ancelotti Jr.[1]

Received: 23 March 2018 / Accepted: 30 April 2018

Abstract
Sensor placement optimization plays a key role in structural health monitoring (SHM) of large mechanical structures. Given the existence of an effective damage identification procedure, the problem arises as to how the acquisition points should be placed for optimal efficiency of the detection system. The global multiobjective optimization of sensor locations for structural health monitoring systems is studied in this paper. First, a laminated composite plate is modelled using Finite Element Method (FEM) and put into modal analysis. Then, multiobjective genetic algorithms (GAs) are adopted to search for the optimal locations of sensors. Numerical issues arising in the selection of the optimal sensor configuration in structural dynamics are addressed. A method of multiobjective sensor locations optimization using the collected information by Fisher Information Matrix (FIM) and mode shape interpolation is presented in this paper. The sensor locations are prioritized according to their ability to localize structural damage based on the eigenvector sensitivity method. The proposed method presented in this paper allows to distribute the points of acquisition on a structure in the best possible way so as to obtain both data of greater modal information and data for better modal reconstruction from a minimum point interpolation. Numerical example and test results show that the proposed method is effective to distribute a reduced number of sensors on a structure and at the same time guarantee the quality of information obtained. The results still indicate that the modal configuration obtained by multiobjective optimization does not become trivial when a set of modes is used in the construction of the objective function. This strategy is an advantage in experimental modal analysis tests, since it is only necessary to acquire signals in a limited number of points, saving time and operational costs.

Keywords Sensor placement optimization · Structural health monitoring · Multiobjective optimization · Genetic algorithm · Mode shape interpolation

1 Introduction

The increasingly development of complex mechanical structures with high structural responsibility, the need arises to develop ever more intelligent Structural Health Monitoring (SHM) systems. The basic problem of an SHM system is to get the structural response from a specific number of sensors and from there, from data processing, infer whether there is a damage or not. In this way, it is extremely important to optimize the sensors to distribute them in the best possible way., with which one can have a quantity and quality of information of the reliable structural response.

The acquisition points on a mechanical structure is a crucial problem in SHM systems. In aeronautical systems, it is impossible to install sensor on every part of the fuselage by various functional and economic factors. In this way, taking into account a finite (limited) number of sensor, two fundamental questions arise about the location optimization [15]: (1) which type of performance index should be evaluated and (2) which optimization technique can be used. Answering those questions, a placement optimization problem is formulated based on a multicriteria optimization and a non-sorted genetic algorithm (NSGA).

In aeronautical systems, it is impossible to install sensor on every part of the fuselage by various functional and

✉ Guilherme Ferreira Gomes
guilhermefergom@gmail.com

[1] Mechanical Engineering Institute, Federal University of Itajubá, Itajubá, Brazil

[2] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil

Published online: 28 May 2018

Artigo "*Fuzzy multivariate mean square error in equispaced pareto frontiers considering manufacturing process optimization problems*", publicado na "***Engineering with Computers***" [143].

Engineering with Computers
https://doi.org/10.1007/s00366-018-0660-0

ORIGINAL ARTICLE

Fuzzy multivariate mean square error in equispaced pareto frontiers considering manufacturing process optimization problems

Juliana Helena Daroz Gaudêncio[1] · Fabrício Alves de Almeida[1] · Rachel Campos Sabioni[2] · João Batista Turrioni[1] · Anderson Paulo de Paiva[1] · Paulo Henrique da Silva Campos[1]

Received: 20 June 2018 / Accepted: 3 November 2018

Abstract
This paper proposes a combined approach using the normal boundary intersection (NBI) and multivariate mean square error (MMSE) that is an alternative approach to outperform the traditional NBI driving to an equispaced Pareto Frontier in a low-dimension space with a considerable reduction in the number of iterations. The method participating in the evolutionary stage of creating a uniformly spread Pareto Frontier for a nonlinear multi-objective problem is the NBI using normalized objective functions allied to MMSE. In sequence, the fuzzy MMSE approach is utilized to determine the optimal point of the multi-objective optimization. For sake of comparison, the performance of arc homotopy length, global criterion method, and weighted sums were explored. To illustrate this proposal, a multivariate case of AISI H13 hardened steel-turning process is used. Experimental results indicate that the solution found by NBI-MMSE approach is a more appropriate Pareto frontier that surpassed all the competitors and also provides the best-compromised solution to set the machine input parameters. Further, this algorithm was also tested in benchmark functions to confirm the NBI-MMSE efficiency.

Keywords Principal component analysis · Multivariate mean square error · Normal boundary intersection · Fuzzy decision maker · Hardened steel turing

1 Introduction

Engineering problems have been precisely solved with the advancement of mathematical modeling coupled with computer methods, once industrial processes have several input variables that influence the output variables. This way, it is possible to find in the literature studies, the mathematical and computational strategies applied to engineering, such as genetic algorithm [1, 2], particle swarm [3], ant colony [4], grey wolf optimizer [5], multilevel cross entropy optimizer [6], and sunflower optimization [7]. These applications are widely used in manufacturing processes, such as the machining process, where it is possible to find jobs that aim at the efficiency of the process, using techniques such as mean square error (MSE) [8], Taguchi method [9], and response surface methodology (RSM) [10].

According to Almeida et al. [8], RSM is a kind of design of experiment (DOE), a statistical technique capable of modeling, optimizing, and reducing experimental costs in manufacturing processes. Usually, these processes present several responses of interest that may present a certain level of correlation. In applications and mathematical models of correlated characteristics, the variance–covariance structure of these characteristics should be considered [11]. Thus, the modeling of correlated characteristics for application in heuristic strategies should be seen as a multivariate strategy to optimize all variables simultaneously. In a multi-objective optimization, efforts must be made to find the set of optimal solutions, by considering all objectives to be important. Within this approach, this work presents an algorithm focused in the engineering process optimization that combines the normal boundary intersection (NBI) method with multivariate mean square

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00366-018-0660-0) contains supplementary material, which is available to authorized users.

✉ Fabrício Alves de Almeida
fabricio.alvesdealmeida@gmail.com

[1] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil

[2] Department of Mechanical Engineering, Sorbonne University, University of Technology of Compiègne, Compiègne, France

Published online: 26 November 2018

Artigo "*A multivariate normal boundary intersection PCA-based approach to reduce dimensionality in optimization problems for LBM process*", publicado na "***Engineering with Computers***" [189].

Engineering with Computers
https://doi.org/10.1007/s00366-018-0678-3

ORIGINAL ARTICLE

A multivariate normal boundary intersection PCA-based approach to reduce dimensionality in optimization problems for LBM process

Gabriela Belinato[1,2] · Fabrício Alves de Almeida[1] · Anderson Paulo de Paiva[1] · José Henrique de Freitas Gomes[1] · Pedro Paulo Balestrassi[1] · Pedro Alexandre Rodrigues Carvalho Rosa[3]

Received: 25 July 2018 / Accepted: 7 December 2018

Abstract
Laser beam machining (LBM) is a promising manufacturing process that exhibits several desirable quality characteristics. Given a large number of objective functions, the level of complexity increases in an optimization problem. Therefore, this study presents a multivariate application of the normal boundary intersection (NBI) method to reduce dimensionality in optimization problems of the LBM process. Such an approach is capable of exploring the entire solution space with only a small number of Pareto points, and generating equispaced frontiers based on the objective functions written in terms of principal component scores. Hence, a design of experiment with three input parameters and six quality characteristics was undertaken to appropriately model the process requirements applied to AISI 314S steel. The results indicate that the proposed methodology is capable of achieving optimal values for interest characteristics. In addition, this approach shows a reduction in computational effort of approximately 91.89% (from 259 to 21 subproblems) in obtaining the best solution for rough operation.

Keywords Laser beam machining · Principal component analysis · Normal boundary intersection · Material removal rate · Roughness

1 Introduction

Manufacturing processes, such as machining processes, present many critical quality characteristics, such as various types of roughness, as well as performance and productivity characteristics. Among the machining processes, the laser beam machining (LBM) process is prevalent as a promising and non-consumable method. LBM is a non-conventional method that exhibits several industrial advantages [1], in addition to being widely applicable in automotive, civil, and nuclear sectors [2]. This process has high potential for applications in precision mechanics and micromechanics, but the expansion is being postponed due to a high initial investment cost and the high energy consumption involved in the process. This difficulty causes the number of published studies to be reduced, limiting the information on the appropriate parameters to allow optimization of the manufacturing process, either the rate of removal of material or the finish of the roughness surfaces. Many studies of multiobjective applications can be found in the literature, such as [3–8], however the LBM process is not very exploited.

Machining processes are, in general, considered as multiobjective problems once they involve more than one performance characteristic. LBM processes involve important input parameters, each of which is crucial in the process. Some of the main parameters are the following: laser frequency (f), cutting speed (S), laser power (I), pulse intensity, [9]. The main quality characteristics investigated in LBM are the material removal rate (MRR), roughness, and metallurgical and mechanical properties [2].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00366-018-0678-3) contains supplementary material, which is available to authorized users.

✉ Fabrício Alves de Almeida
fabricio.alvesdealmeida@gmail.com

[1] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil

[2] IFSULDEMINAS Federal Institute of South Minas Gerais, Pouso Alegre, Brazil

[3] IDMEC Department of Mechanical Engineering, Technician Superior Institute of Lisbon, University of Lisbon, Lisbon, Portugal

Published online: 12 December 2018

Artigo "*Multivariate Stochastic Optimization Approach Applied in a Flux-Cored Arc Welding Process*", publicado na "***IEEE Access***" [209].

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 2, 2020, accepted March 16, 2020, date of publication March 26, 2020, date of current version April 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2983566

Multivariate Stochastic Optimization Approach Applied in a Flux-Cored Arc Welding Process

ALEXANDRE F. TORRES[1], FRANCO B. ROCHA[2,3], FABRÍCIO A. ALMEIDA[1], JOSÉ H. F. GOMES[1], ANDERSON P. PAIVA[1], AND PEDRO PAULO BALESTRASSI[1,2]

[1]Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá 37500-903, Brazil
[2]Institute of Systems Engineering and Information Technology, Federal University of Itajubá, Itajubá 37500-903, Brazil
[3]Institute of Exact Sciences, Federal University of Alfenas, Alfenas 37130-001, Brazil

Corresponding author: Alexandre F. Torres (alexandrefonsecatorres@gmail.com)

This work was supported in part by FAPEMIG, in part by CNPq, in part by CAPES, and in part by FAPEPE.

ABSTRACT One of the main goals in flux-cored arc welding processes is the optimization of bead geometry, in which multiple geometric characteristics of the welding bead are important; therefore, multiobjective optimization programming is often applied. However, several optimization problems that use stochastic programming do not consider the impact of the correlation between the output variables on their probabilistic constraints. In this context, this paper aims to present a multiobjective optimization method based on multivariate stochastic programming. To demonstrate the applicability of the proposal, we conducted a design of experiments to optimize a flux-cored arc welding process for stainless-steel claddings. The weighting-sums method was applied to formulate the multiobjective optimization problem. It was possible to formulate a multivariate probability distribution for the penetration and dilution. In addition, a 95% probability to meet the predefined specification limits of the geometric characteristics was achieved.

INDEX TERMS Bead geometry, bivariate normal distribution, flux-cored arc welding, multiobjective optimization programming, stainless steel claddings, stochastic programming.

I. INTRODUCTION

In practical optimization problems of industrial processes, the assumption that the input data are deterministic is rarely sustained. In fact, certain key inputs that are clearly random are instead represented by their expected values. Such an approach may be justified under special conditions; however, in several applications, it is possible to demonstrate that such a formulation is inadequate [1].

For example, in the general linear programming formulation, $Min\ \mathbf{f(x)} = \mathbf{C'x}$ is a vector of deterministic objective functions that needs to be minimized and $\mathbf{Ax} \leq \mathbf{b}$ is a set of constraints [2]. In most approaches reported in the literature, matrices **C** and **A** and vector **b** are composed of deterministic values. Nevertheless, these vectors and matrices may have random inputs in actual problems.

Therefore, it is important to model the stochastic nature of the inputs in optimization problems. To achieve this, stochastic programming (SP) can be used as a technique to measure and analyze the impact on the variability of the responses [3]–[6] and has been used in different sectors. Dai *et al.* [7] presented a literature review of all different SP methods that have been applied only in unit commitment (UC), including multi-stage SP and chance-constrained programming (CCP). Reddy *et al.* [8] provided a literature review on stochastic programming methods applied in the optimization of smart grids (SG). The recourse method and CCP were cited as widely used within the SG context.

In addition to the random aspect, most optimization problems present multiple and, often, conflicting responses of interest [9]. In such cases, it is necessary to consider the multivariate nature of the data [10]. For instance, correlated data present a significant variance–covariance structure; to properly compute the probabilities involved in the problem, multivariate techniques should be used. These strategies have been widely employed in various segments [11]–[14]. Considering a random distribution of vectors that contain correlated variables, there is a multivariate normal distribution (MND), each element of which has been assigned a univariate normal distribution [15].

Within this context, the present authors propose a multivariate stochastic optimization method, referred to as

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

Artigo "*Fault classification in three-phase motors based on vibration signal analysis and artificial neural networks*", publicado na "***Neural Computing and Applications***" [210].

Neural Computing and Applications (2020) 32:15171–15189
https://doi.org/10.1007/s00521-020-04868-w

ORIGINAL ARTICLE

Fault classification in three-phase motors based on vibration signal analysis and artificial neural networks

Ronny Francis Ribeiro Junior[1,3] · Fabrício Alves de Almeida[2] · Guilherme Ferreira Gomes[1]

Received: 20 December 2019 / Accepted: 14 March 2020 / Published online: 27 March 2020

Abstract
Competition in the industrial environment is increasingly intense, so it is of utmost importance that organizations keep their assets in operation as much as possible (in order to produce more). In this context, there is a need for predictive maintenance, a technique that detects the health of assets in real time, allowing failures to be diagnosed before they can interrupt the operation of the assets, avoiding high financial losses. This study uses a sixteen-motor experimental setup with four different known operating conditions. The vibration signal of these motors, through signal analysis, both in time and frequency domains, is performed to evaluate the types and severities of the defects. An artificial neural network (ANN) is used to classify these defects. Considering the vibration analysis, mechanical faults can be identified quickly and conveniently. For the development of the ANN, it was necessary to perform a preprocessing of the vibration signal (response in time) due to the data size, which overwhelms the network. Thus, statistical data were used to extract key information from the vibration signal. Finally, the neural network created based on this study's methodology presents extremely reliable results, allowing a quick and robust diagnosis of the motor operating condition.

Keywords Predictive maintenance · Vibration analysis · FFT · Artificial neural networks · Damage classification

1 Introduction

Induction motors present high performance and reliability, playing a critical role in many industrial sectors. However, despite their reliability, they are subject to failure [26]. Being able to classify or predict failures (or operating condition) is a task of great importance and crucial for engineers, especially in the field of maintenance. One of the possible solutions is through the use of artificial neural networks (ANN) based on data from certain engines. An effective ANN is able to efficiently predict the assessed response saving time and maintenance costs.

Thus, the general industry's demand for predictive maintenance products and services is increasing. Predictive maintenance is one that indicates the actual operating conditions of equipment based on elements that report wear or degradation process. Therefore, long-term maintenance costs can be reduced with adequate predictive maintenance techniques [18].

In this context, the vibration analysis method is a mature and applicable alternative for predictive motor maintenance. There are two important steps to implement the fault diagnosis process: The first is signal processing and the second is signal classification based on the characteristics obtained in the previous step. The diagnosis is usually much more difficult than the detection because different failures may exhibit similar symptoms and multiple failures may occur at the same time [22].

Currently, artificial neural networks (ANN) techniques are attracting attention in studies given their ability to perform difficult tasks [5, 8, 12, 13], such as vibration signal diagnosis, quickly and efficiently [14, 23, 25, 26]. Therefore, a neural network can contribute to the speedy diagnosis of a failure by increasing the efficiency of predictive maintenance. Nevertheless, to the best of the authors' knowledge, very few efforts have been devoted to

✉ Guilherme Ferreira Gomes
guilhermefergom@unifei.edu.br

[1] Mechanical Engineering Institute, Federal University of Itajubá – UNIFEI, Itajubá, Brazil

[2] Institute of Industrial Engineering and Management, Federal University of Itajubá – NIFEI, Itajubá, Brazil

[3] PS Solutions, Itajubá, Brazil

Artigo "*Impact of stochastic industrial variables on the cost optimization of AISI 52100 hardened-steel turning process*", publicado na "***International Journal of Advanced Manufacturing Technology***" [211].

The International Journal of Advanced Manufacturing Technology
https://doi.org/10.1007/s00170-019-04273-1

ORIGINAL ARTICLE

Impact of stochastic industrial variables on the cost optimization of AISI 52100 hardened-steel turning process

Alexandre Fonseca Torres[1] · Fabrício Alves de Almeida[1] · Anderson Paulo de Paiva[1] · João Roberto Ferreira[1] · Pedro Paulo Balestrassi[1] · Paulo Henrique da Silva Campos[1]

Received: 13 March 2019 / Accepted: 5 August 2019

Abstract
An optimization problem of the AISI 52100 hard-steel turning process is examined. A new approach is presented in which not only the machine parameters (cutting speed, feed rate, and depth of cut) but also the stochastic industrial variables of setup time, insert changing time, batch size, machine and labor costs, tool holder price, tool holder life, and insert price are considered. By representing each of these variables by a given probability distribution, the goal was to analyze their impact on the total process cost per piece (K_p). Experiments were carried out following a central composite design to model tool life (T), average surface roughness (R_a), and peak-to-valley surface roughness (R_t) using a response surface methodology. Then, stochastic programming was used to model K_p's expected value and standard deviation. The approach to the optimization problem aimed to maximize the probability for the cost to be less than a target value, subject to the experimental space and to maximum values of both R_a and R_t. The results were optimal values for the cutting conditions that provide a suitable confidence interval for K_p. The most-significant industrial variables on K_p were ranked. In addition, it was found that, in the addressed case, cutting conditions for maximum tool life actually increase K_p.

Keywords Stochastic programming · Hardened-steel turning · Process cost optimization

1 Introduction

Recently, advances in the machining of hardened steels, such as hard turning, have significantly contributed to product quality in manufacturing industries [1–3]. In fact, hard turning is a manufacturing process widely applied in industry. Compared with grinding, hard turning can provide equal or even better surface finish [4] with higher material removal rates [5]. Other benefits provided by hard turning include coolant reduction or elimination, process cost reduction, productivity increase, improved material properties, and reduced power consumption [6–8].

Gears, shafts, bearings, bushes, dies, crushing cones, and jet engine mounting are some of the applications of hardened steels [9]. In particular, AISI 52100 hardened steel is frequently used to manufacture bearings, ball screws, gauges, axles, and joints because of its strength and corrosion resistance [10]. AISI 52100 is considered to be one of the hard-to-cut steel alloys [11] in terms of cutting tool materials and economical machining.

Nevertheless, only a few studies on hardened-steel turning optimization have considered the impact of industrial variables and their effect on the variability of the process cost. Industrial variables include setup time, insert changing time, batch size, and others [12]. Most of them are stochastic, and some are not controllable. There are already different stochastic-programming models available in the literature [13], and some researchers have already applied them in manufacturing systems to analyze setup times [14], batch size [15], and machines and labor [16]. Within this context, this study aimed to optimize the total process cost per piece for AISI 52100 hardened-steel turning by also taking into account the following stochastic variables: setup time, insert changing time, batch size, machine and labor cost, tool holder price, tool holder life, and insert price.

This work is structured as follows. In Section 2, a review of the literature about response surface methodology (RSM), total process cost per piece in turning, and stochastic programming is presented. In Section 2.3, an equation used to calculate

✉ Fabrício Alves de Almeida
fabricio.alvesdealmeida@gmail.com

[1] Institute of Industrial Engineering, Federal University of Itajubá, Itajubá, MG 37500-903, Brazil

Published online: 16 August 2019

Artigo "*An estimate of the location of multiple delaminations on aeronautical CFRP plates using modal data inverse problem*", publicado na "***International Journal of Advanced Manufacturing Technology***" [212].

The International Journal of Advanced Manufacturing Technology
https://doi.org/10.1007/s00170-018-2502-z

ORIGINAL ARTICLE

An estimate of the location of multiple delaminations on aeronautical CFRP plates using modal data inverse problem

Guilherme Ferreira Gomes [1] · Fabricio Alves de Almeida [2] · Sebastiao Simões da Cunha Jr [1] · Antonio Carlos Ancelotti Jr [1]

Received: 22 March 2018 / Accepted: 20 July 2018

Abstract
With the increase in the use of composite materials, especially in the aeronautical industry, it is essential that a complete evaluation of the mechanical performance of such structures be undertaken, especially with regard to structural integrity. To assist in this task, structural health monitoring methodologies are employed in order to minimize time and maintenance costs, and errors arising principally from human factors, and which can occasionally result from the failure to properly inspect the aircrafts. This study addresses the use of an inverse method for delamination identification in carbon fiber reinforced polymers plates. First, the direct problem was modeled via a finite element method in order to obtain a faithful model that represented the real case studied. The inverse problem was solved by minimizing an objective function through genetic algorithms. Modal responses of delaminated plates are able to identify the possible location of multiple delaminations in laminated plates since the structural matrices are changed as a function of the induced damage. Numerical and experimental results showed excellent identification of small delaminations, reducing the initial search area by up to 96%, which can lead to savings in time and costs for the aeronautical industry.

Keywords Damage identification · Structural health monitoring · Genetic algorithm · Aeronautics · CFRP plates

1 Introduction

Structural health monitoring (SHM) is an interdisciplinary field in engineering that deals with innovative methods of structural monitoring, integrity, and performance without affecting the structure itself or harming its operation. The SHM methodology uses several types of sensors to detect the presence, location, and severity of structural damage. Such technology integrates non-destructive evaluation (NDE) techniques using sensory and intelligent materials to create self-monitoring mechanisms characterized by greater reliability and longer structural life. The method is applied mainly to systems with critical requirements regarding structural performance, where the classical evaluation of localized inspection is costly, difficult, or even impossible in terms of operationality [1].

SHM methods that are able to find changes in structural characteristics due to damage or degradation can be defined as damage detection methods [2]. According to [3], damage is defined as an undesirable weakening of a structure that has a negative effect on its performance and affects the safety of the structural system. Damage can also be defined as any change in geometric characteristics or material properties of the structure in question, which may cause undesirable stresses, displacements, or vibrations. The effects of damage on a structure can be classified as linear or nonlinear. Linear damage is defined as a situation in which the initially linear-elastic structure remains linear after the damage. [4] *apud* [2] have defined nonlinear damage as a situation in which the initially linear-elastic structure behaves in a nonlinear fashion after the damage.

Given the extensive literature on the subject of structural monitoring published in the last 20 years, one can argue that this field has matured to the point of establishing several fundamental axioms or general principles [5]. Firstly, according

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00170-018-2502-z) contains supplementary material, which is available to authorized users.

✉ Guilherme Ferreira Gomes
guilhermefergom@unifei.edu.br

[1] Mechanical Engineering Institute, Federal University of Itajubá, Itajubá, Brazil

[2] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, Brazil

Published online: 16 August 2018

Artigo "*A Weighted Mean Square Error Approach to the Robust Optimization of the Surface Roughness in an AISI 12L14 Free-Machining Steel-Turning Process*", publicado na "***Strojniski Vestnik - Journal of Mechanical Engineering***" [141].

Strojniški vestnik - Journal of Mechanical Engineering 64(2018)3, 147-156
DOI:10.5545/sv-jme.2017.4901
Original Scientific Paper
Received for review: 2017-09-06
Received revised form: 2018-01-29
Accepted for publication: 2018-02-14

A Weighted Mean Square Error Approach to the Robust Optimization of the Surface Roughness in an AISI 12L14 Free-Machining Steel-Turning Process

Fabrício Alves de Almeida[1] – Guilherme Ferreira Gomes[2] – Vinícius Renó de Paula[1] – João Ederson Corrêa[1] – Anderson Paulo de Paiva[1] – José Henrique de Freitas Gomes[1] – João Batista Turrioni[1]
[1]Federal University of Itajubá, Institute of Industrial Engineering and Management, Brazil
[2]Federal University of Itajubá, Mechanical Engineering Institute, Brazil

The objective of this work is to determine an optimal setup for the 12L14 free-machining steel-turning process that will be able to neutralize the influence of tool wear in the workpiece's mean roughness R_a. Aiming this, equations for the mean and variance of the roughness were modelled using the response surface methodology. A crossed array with three input variables of the turning process (cutting speed, feed and depth of cut) and a noise variable (use of new and wear tools) is applied to the methodology. Subsequently, these same responses were optimized using the mean square error, which allows the response mean value to approach a predetermined target value by cancelling variations thereof through a weighted objective. Confirmation experiments were conducted to prove the suitability of the method and excellent results were obtained.

Keywords: robust parameter optimization, mean square error, 12L14 free-machining steel turning, response surface methodology

Highlights

- *A robust parameter design applied to the AISI 12L14 free machining steel turning process.*
- *Using techniques such as Response Surface Methodology (RSM) and crossed array to generate a model for process optimization.*
- *The Mean Square Error (MSE) method followed by a meta-modelling of the answers in order to neutralize the roughness of the answers for both a new and a wear tool through a weighted objective.*
- *Based on the presented methodology, the confirmation experiments proved the adequacy of the method, neutralizing the influence of the noise in the roughness response.*

0 INTRODUCTION

Manufacturing processes are aimed at transforming materials into goods, generating wealth [1]. Camposeco-Negrete [2] states that cost and quality are the main goals of manufacturing companies. To improve quality in this type of process, several authors have studied the turning process using mathematical strategies in order to contribute to the efficiency of these processes, such as: the Taguchi method [3], ant colony optimization [4], genetic algorithm [5] and response surface methodology (RSM) [6].

The primary input parameters in the turning process, i.e., cutting speed, feed rate and depth of cut [7], are directly responsible for the quality and productivity characteristics of the process, such as the amount of material removed, tool wear, and finishing of the product [8].

Furthermore, the finishing of the machined parts can be evaluated according to the surface roughness, which are irregularities presented on the surface of the parts, characterized by grooves made by the tool during the machining process [9] of the cutting parameters (ranging from a single parameter per experiment) in the quality responses, such as tool life and surface roughness. This paper makes use of only one roughness parameter (arithmetic average roughness, R_a), considering the calculation of its metrics of the roughness characteristics (R_a, R_y, R_q, R_z and R_t). The average arithmetic roughness (R_a) is the arithmetic mean of the absolute values of the ordinates of the effective (measured) profile in relation to the midline in a sample length (Fig. 1). In addition, R_a is the most used parameter for general quality control [10].

The steel used in the turning process of this study was AISI 12L14 carbon steel (used in studies such as: Peruchi et al. [11], Kishawy et al. [12], Overcash and Cuttino [13], Milstein and Marschall, [14]). The surface roughness of the turned parts and how their roughness is affected by the wear of the cutting tool was studied. The interaction of the wear on the tool (notch wear, flank wear, crater wear, among others) and the cutting parameters used in the process can be critical to the machined work surface finish [15] and [16] and may give unsatisfactory results.

As a result, to minimize experimental costs, it is necessary to use strategies such as the design of

*Corr. Author's Address: Federal University of Itajuba, 1303 BPS Avenue, Itajubá, Brazil. fabricio.alvesdealmeida@gmail.com

Artigo "*A Gage Study Applied In Shear Test To Identify Variation Causes From A Resistance Spot Welding Measure System*", publicado na "***Strojniski Vestnik - Journal of Mechanical Engineering***" [213].

Strojniški vestnik - Journal of Mechanical Engineering 64(2018)10, 621-631
DOI:10.5545/sv-jme.2018.5235
Original Scientific Paper
Received for review: 2018-01-23
Received revised form: 2018-05-23
Accepted for publication: 2018-06-13

A Gage Study Applied in Shear Test to Identify Variation Causes from a Resistance Spot Welding Measurement System

Fabrício Alves de Almeida[1,*] – Guilherme Ferreira Gomes[2] – Rachel Campos Sabioni[3] – José Henrique de Freitas Gomes[1] – Vinícius Renó de Paula[1] – Anderson Paulo de Paiva[1] – Sebastião Carlos da Costa[1]

[1]Federal University of Itajubá – Institute of Industrial Engineering and Management, Brazil
[2]Federal University of Itajubá – Mechanical Engineering Institute, Brazil
[3]Sorbonne Universités, Université de Technologie de Compiègne - Department of Mechanical Engineering, France

Resistance welding processes, especially spot welding, have wide applicability in the industry, especially in the automotive sector, due to its fast execution and the non-use of consumables. In addition, the search for quality improvement of the final product is incessant and, in a capable process, there should be no error related to the measurements. In this study, the NGR&R was used by the ANOVA method to identify the variation components of the measurement system in the shear test for two quality characteristics: tensile-shear strength and ultimate strain. The experiments were conducted on a hot dip galvanized steel by using design of experiments to select the parts in order to represent the real amplitude of the process. From the results it was possible to verify that one of the destructive test machines used in this study has a strong variability, evidencing that some adjustments and improvements are necessary in the coupling of the specimens for steels with coating layers (such as galvanized steel, which has a layer of zinc).

Keywords: spot welding, measurement system analysis, shear test, NGR&R, ANOVA

Highlights

- *A nested GR&R study applied in shear test for a resistance spot welding process.*
- *Design of experiments to select the parts in order to represent the real amplitude of the process.*
- *The analysis of variance method to identify variation causes for two tensile machines and two quality characteristics: tensile-shear strength and ultimate strain.*
- *The results showed that Machine 1 presents greater contribution on the system variability, with measurement results outside the control, as well as a lower degree of repetitiveness than Machine 2.*

0 INTRODUCTION

Resistance spot welding (RSW) is a structural joint technique widely used in the automotive sector [1]. RSW is highlighted among welding processes due to its features that favor the industry such as agile operation, which is easily suitable for automatic processes, simple handling, diverse applications and low cost [2] to [4]. Because of its wide applicability and importance, new methodologies for parameter adjustment have been applied to RSW improvement, contributing to the process control and capability.

Among the available methods for verification of the weld point, there is the shear test, which is characterized by the application of opposing forces causing stress in a sliding movement for a given sample. Since this type of test allows to evaluate the quality of welded point, it is being increasingly used, as described by Feng et al. [5], Zhang et al. [6], Martín et al. [7], Shan et al. [8], Chen et al. [9], Manladan et al. [10].

The search for quality improvements has been leading industries to improve their efficiency. However, devoting improvements only to the process may not contribute to make it better, as the variability can also be caused by the measurement system. Therefore, it is necessary to verify the measurement system variability in industrial processes, such as RSW.

There are several methods for controlling and monitoring quality in the RSW process, such as: expulsion detection in materials [1] and [11]; strength estimation based on sonic emission [12]; welding current analysis on weld strength [13]; temperature measurement [14]; electrode displacement [15] and [16]; and other types of control (i.e. electrical variables, ultrasound transmission and acoustic emission) [17]. However, the control approaches must be verified through specific tests, which illustrate the mechanical characteristics necessary for their capability evaluation, such as the shear test.

The shear test is characterized as a destructive test that evaluates the mechanical strength of the weld point in relation to shear stresses. Destructive tests are performed from time to time, by sampling, being widely employed in the automotive sector.

*Corr. Author's Address: Federal University of Itajubá, 1303 BPS Avenue, Itajubá, Brazil, fabricio.alvesdealmeida@gmail.com

Artigo "*Optimum design of composite structures with ply drop-offs using response surface methodology*", publicado na "***Engineering Computations***" [214].

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0264-4401.htm

Optimum design of composite structures with ply drop-offs using response surface methodology

Camila Aparecida Diniz and Yohan Méndez
Mechanical Engineering Institute, Federal University of Itajubá, Itajuba, Brazil
Fabrício Alves de Almeida
Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajuba, Brazil, and
Sebastião Simões da Cunha Jr and G.F. Gomes
Mechanical Engineering Institute, Federal University of Itajubá, Itajuba, Brazil

Received 2 July 2020
Revised 2 October 2020
17 December 2020
18 December 2020
Accepted 19 December 2020

Abstract

Purpose – Many studies only take into account the ply stacking sequence as the design variable to determine the optimal ply drop-off location; however, it is necessary to optimize other parameters that have a direct influence on the ply drop-off site such as which plies should be dropped and in which longitudinal direction. That way, the purpose of this study is to find the most significant design variables relative to the drop-off location considering the transversal and longitudinal positions, seeking to achieve the optimal combination of ply drop-off locations that provides excellent performance for the laminate plate.

Design/methodology/approach – This study aims to determine the optimal drop-off location in a laminate plate using the finite element method and an approach statistical with design of experiments (DOE).

Findings – The optimization strategy using DOE revealed to be satisfactory for analyzing laminate structures with ply drop-offs, demonstrating that not all design factors influence the response variability. The failure criterion response variable revealed a poor fit, with an adjusted coefficient of determination lower than 60%, thus demonstrating that the response did not vary with the ply drop-off location. Already the strain and natural frequency response variables presented high significance. Finally, the optimization strategy revealed that the optimal drop-off location that minimizes the strain and maximizes the natural frequency is the ply drop-off located of the end plate.

Originality/value – It was also noted that many researchers prefer evolutionary algorithms for optimizing composite structures with ply drop-offs, being scarce to the literature studies involving optimization strategies using response surface methodology. In addition, many studies only take into account the ply stacking sequence as the design variable to determine the optimal ply drop-off location; however, in this study, the authors investigated other important parameters that have direct influence on the ply drop-off site such as which plies should be dropped and in which longitudinal direction.

Keywords Optimization, Failure, Composites, Response surface methodology, Ply drop-off

Paper type Research paper

Nomenclature

F = Strength parameters;
σ_1 = Longitudinal normal stress;

The authors would like to acknowledge the financial support from the Brazilian agency CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais – APQ-00385–18) and Tutorial Education Program (PET – Programa de Educação Tutorial).

Engineering Computations

0264-4401
DOI 10.1108/EC-07-2020-0354

Artigo "*Customer value in lean product development: Conceptual model for incremental innovations*", publicado na "***Systems Engineering***" [215].

Received: 22 February 2019 | Revised: 1 August 2019 | Accepted: 12 September 2019

DOI: 10.1002/sys.21514

REGULAR PAPER

WILEY

Customer value in lean product development: Conceptual model for incremental innovations

Ana Carolina Oliveira Santos[1] | Carlos Eduardo Sanches da Silva[2] | Rodrigo Aparecido da Silva Braga[3] | João Éderson Corrêa[2] | Fabricio Alves de Almeida[2]

[1] Institute of Integrated Engineering, Federal University of Itajuba, Itabira, MG, Brazil

[2] Institute of Industrial Engineering and Management, Federal University of Itajuba, Itajuba, MG, Brazil

[3] Institute of Science and Technology, Federal University of Itajuba, Itabira, MG, Brazil

Correspondence
Ana Carolina Oliveira Santos, Institute of Integrated Engineering, Federal University of Itajuba, Itabira, MG, 35903 087, Brazil.
Email: anasantos@unifei.edu.br

Funding information
Minas Gerais State Agency for Research and Development (FAPEMIG), Grant/Award Number: PPM00799-18; The Brazilian National Council for Scientific and Technological Development (CNPq), Grant/Award Number: Process 431596/2016-6

Abstract

As the lean product development (LPD) process is responsible for maximizing the value/waste relation by encouraging different types of innovation, it needs a clear understanding of end customer value. Moreover, lean systems engineering is related to the application of lean thinking (LT) in systems engineering to improve the delivery of value to all stakeholders of the system. Then, faced with this scenario, this research aims to propose a value-adding conceptual model for incremental product innovations in LPD. A systematic review of the literature from the LT perspective was performed and the conceptions of value were identified and then analyzed and contrasted with the concepts and characteristics of the areas of product development, marketing, and consumer psychology, which made possible the proposition of a conceptual model. The results show that customer value is multidimensional, and it was possible to identify the importance and the need for prioritization of certain dimensions of value for specific markets and customers. The conceptual model named VA2I presented in this study is generic and can be adapted and applied in any type of organization.

KEYWORDS

customer value, industrial application, lean thinking, lean systems engineering, manufactured products, product development

1 | INTRODUCTION

Lean manufacturing was initially developed in Japan by Toyota to compete with the mass-production system adopted by U.S. automakers, which until then stood out for their performance in terms of quality and cost. As the postwar scenario in Japan did not allow adoption of the American method of production, Toyota created a new management strategy focused on waste reduction in all aspects of its operations: the Toyota Production System (TPS). In the 1990s, the TPS principles were analyzed in more detail in the book "*Lean Thinking*,"[1] and from then on, its essence was transferred from production efficiency to a certain type of organizational intervention and management focused on the best practices and methodologies of process improvement. In this way, their efforts began to focus on increasing added value throughout the flow (from suppliers to final customers) and reducing waste from their processes. For this reason, the term lean thinking (LT) became as famous as lean production or lean manufacturing—especially in the western industry, where it is also known only by the term "*Lean*."[2–8]

Although specifying value is described as the first lean principle, there are few studies in the LT literature on how it is set.[9–12] In most of the work on LT, customer satisfaction is often used as a measure of performance to monitor how well a product or service is delivering value. However, consumer satisfaction or dissatisfaction is often a one-dimensional subjective construction based only on how well a product or service meets customer expectations, rather than its actual performance attributes.[5,9,13]

In this sense, if the concept of value is indefinite or intangible, the definition of waste will become even more incomprehensible—since the concept of waste is given as any activity that consumes resources but does not create value for the customer.[7,14–16]

Haque and James-Moore[4] state that from an implementation point of view, although there is a large amount of studies on LT, most of them

Artigo "*Development of a System Measurement Model of the Brazilian Hospital Accreditation System*", publicado na "***International Journal of Environmental Research and Public Health***" [216].

International Journal of *Environmental Research and Public Health*

MDPI

Article

Development of a System Measurement Model of the Brazilian Hospital Accreditation System

João Éderson Corrêa [1], João Batista Turrioni [1], Carlos Henrique Pereira Mello [1], Ana Carolina Oliveira Santos [2,*], Carlos Eduardo Sanches da Silva [1] and Fabrício Alves de Almeida [1]

1 Institute of Industrial Engineering and Management, Federal University of Itajubá, Av. BPS, 1303, Itajubá, Minas Gerais 37500-903, Brazil; edercorrea@unifei.edu.br (J.E.C.); joabatu@gmail.com (J.B.T.); carlos.mello@unifei.edu.br (C.H.P.M.); sanches@unifei.edu.br (C.E.S.d.S.); fabricio.alvesdealmeida@gmail.com (F.A.d.A.)

2 Institute of Integrated Engineering, Itabira Campus, Federal University of Itajubá, Rua Irmã Ivone Drumond, 200, Itabira, Minas Gerais 35903-087, Brazil

* Correspondence: anasantos@unifei.edu.br; Tel.: +55-031-3839-0847

Received: 31 August 2018; Accepted: 5 November 2018; Published: 11 November 2018

check for updates

Abstract: The purpose of this study is to develop and validate a measurement model that evaluates the Brazilian hospital accreditation methodology (ONA), based on a multivariate model using structural equation modeling (SEM). The information used to develop the model was obtained from a questionnaire sent to all organizations accredited by the ONA methodology. A model was built based on the data obtained and tested through a structural equation modeling (SEM) technique using the LISREL® software (Scientific Software International, Inc., Skokie, IL, USA). Four different tests were performed: Initial, calibrated, simulated, and cross-validation models. By analyzing and validating the proposed measurement model, it can be verified that the selected factors satisfy the required criteria for the development of a structural model. The results show that leadership action is one of the most important factors in the process of health services accredited by ONA. Although, leadership, staff management, quality management, organizational culture, process orientation, and safety are strongly linked to the development of health organizations, and directly influence the accreditation process.

Keywords: health services; accreditation; structural equation modeling

1. Introduction

1.1. Motivations

Accreditation is an approach established at international level and is a quality assurance system that adheres to a specific standard, approved by an accreditation body [1]. In the area of health, hospital accreditation is used as a management methodology that fosters strategic understanding involving all employees of the institution through a permanent educational process that shares principles, goals and objectives to be achieved. This multidisciplinary "productive" consensus aims to rationalize the use of resources and optimize results [2].

There has been an increase in the number of accreditation programs in the health area. Nevertheless, the way in which these methodologies are proposed by different accreditation models is one of the points that has been debated at the international level and also by the International Society for Quality in Health Care (ISQua) [3–6]. In this context, hospital accreditation is becoming recognized as an important approach in quality assurance processes, although few studies seek to analyze its performance and influence in organizations that adopt this type of methodology [7–14].

Artigo "*A numerical-experimental study of concrete beams with welded and convenional stirrups*", publicado na "***Computers and Concrete***" [217].

Computers and Concrete, *Vol. 22, No. 1 (2018) 27-37*
DOI: https://doi.org/10.12989/cac.2018.22.1.027

A numerical-experimental evaluation of beams composed of a steel frame with welded and conventional stirrups

Wagner L. Gonçalves[*2], Guilherme F. Gomes[1a], Yohan D. Mendéz[1b], Fabrício A. Almeida[3c], Valquíria C. Santos[4d] and Sebastião S. Cunha Jr.[1e]

[1]*Mechanical Engineering Institute, Federal University of Itajubá (UNIFEI), Avenue BPS, 1303, Itajubá, Brazil*
[2]*University Center of the Guaxupé Educational Foundation (UNIFEG), Avenue Dona Floriana, 463, Guaxupé, Brazil*
[3]*Institute of Industrial Engineering and Management, Federal University of Itajubá (UNIFEI), Avenue BPS, 1303, Itajubá, Brazil*
[4]*Institute of Natural Resources, Federal University of Itajubá (UNIFEI), Avenue BPS, 1303, Itajubá, Brazil*

(Received December 13, 2017, Revised April 11, 2018, Accepted April 12, 2018)

Abstract. Reinforced concrete structures are widely used in civil engineering projects around the world in different designs. Due to the great evolution in computational equipment and numerical methods, structural analysis has become more and more reliable, and in turn more closely approximates reality. Thus among the many numerical methods used to carry out these types of analyses, the finite element method has been highlighted as an optimized tool option, combined with the non-linear and linear analysis techniques of structures. In this paper, the behavior of reinforced concrete beams was analyzed in two different configurations: i) with welding and ii) conventionally lashed stirrups using annealed wire. The structures were subjected to normal and tangential forces up to the limit of their bending resistance capacities to observe the cracking process and growth of the concrete structure. This study was undertaken to evaluate the effectiveness of welded wire fabric as shear reinforcement in concrete prismatic beams under static loading conditions. Experimental analysis was carried out in order compare the maximum load of both configurations, the experimental load-time profile applied in the first configuration was used to reproduce the same loading conditions in the numerical simulations. Thus, comparisons between the numerical and experimental results of the welded frame beam show that the proposed model can estimate the concrete strength and failure behavior accurately.

Keywords: cracks; finite element method; nonlinear concrete; prismatic beam; steel frame; welded stirrups

1. Introduction

Reinforced concrete (RC) is one of the most widely used materials in civil and industrial engineering applications due to the fact that it is highly resistant and can be easily modeled for wide variety of different formats (Mosoarca and Victor 2013, Haifeng and Jianguo 2009). This variety has resulted in concrete being one of the most widely consumed materials in the world, according to Aïtcin (2000), and as such, its quality must be continually improved. Modern reinforced concrete structures need to be designed not only to withstand normal impact loads, such as weight, but also more significant impact loads, such as earthquakes and explosions (Haifeng and Jianguo 2009).

*Corresponding author, Professor
E-mail: wagner1709@mail.unifeg.edu.br
[a]Professor
E-mail: guilhermefergom@unifei.edu.br
[b]Ph.D. Student
E-mail: yohan.g8@unifei.edu.br
[c]Ph.D. Student
E-mail: fabricio-almeida@unifei.edu.br
[d]Professor
E-mail: valquiria@unifei.edu.br
[e]Professor
E-mail: sebas@unifei.edu.br

Several studies are focused on the research surrounding reinforced concrete, such as: Saw *et al.* (2017), Yelgin *et al.* (2014), Lin *et al.* (2013), Han *et al.* (2011), Haifeng and Jianguo (2009). In addition to this, companies are continually seeking to adapt to new technologies and methodologies in waste reduction, minimizing cost without sacrificing quality.

Different concrete structures may have specific characteristics represented by the more resistant parts of their construction, these parts which seek to absorb and transmit forces, if they seek to maximize safety and the integrity of a building.

Notably, according to Fan and Hu (2013), reinforced concrete is made up of diverse combinations of materials, presenting forms of traction reinforcement. In addition to this, Fan and Hu (2013) state that many authors have contributed several theoretical and experimental studies on the quality of reinforced concrete, such as Colajanni *et al.* (2014), Azad *et al.* (1989), Ruiz *et al.* (1998), Ferro *et al.* (2007), Shaowei *et al.* (2011), in which they used experimental techniques, linear elastic fracture mechanics, acoustic emissions, among others, to verify fracture problems in reinforced concrete.

Research on reinforced concrete structures is often directed towards the feasibility of the execution of a project, with the primary concerns being saving material and manpower. Reinforced concrete structures constitute civil works of great responsibility, which call for confidence in

ISSN: 1598-8198 (Print), 1598-818X (Online)

Artigo "*Análise das Causas de Variação Atribuídas a Diferentes Instrumentos Metrológicos para Verificação das Características Geométricas de um Processo de Soldagem por Pontos*", publicado na "***Soldagem & Inspeção***" [218].

Soldagem & Inspeção. 2018;23(4):485-504
https://doi.org/10.1590/0104-9224/SI2304.05
ISSN 1980-6973 (Online)
ISSN 0104-9224 (Print)

Artigos Técnicos

Análise das Causas de Variação Atribuídas a Diferentes Instrumentos Metrológicos para Verificação das Características Geométricas de um Processo de Soldagem por Pontos

Fabrício Alves de Almeida[1*], José Henrique de Freitas Gomes[1], Guilherme Ferreira Gomes[2], Estevão Luiz Romão[1], Pedro Paulo Balestrassi[1]

[1] Universidade Federal de Itajubá – UNIFEI, Instituto de Engenharia de Produção e Gestão, Itajubá, MG, Brasil.
[2] Universidade Federal de Itajubá – UNIFEI, Instituto de Engenharia Mecânica, Itajubá, MG, Brasil.

Recebido: 07 Nov., 2018
Aceito: 08 Fev., 2019

E-mails: fabricio.alvesdealmeida@gmail.com (FAA)

Resumo: Neste estudo, buscou-se analisar a variabilidade de instrumentos utilizados na indústria para avaliar as características geométricas de um ponto de solda. Para tal, inicialmente, utilizou-se da técnica de planejamento de experimentos para gerar um arranjo fatorial fracionado para as configurações dos parâmetros de soldagem em corpos de prova de aços galvanizados por imersão a quente, a fim de representar a amplitude real do processo de soldagem por pontos. Para as análises, utilizou-se o estudo de repetitividade e reprodutibilidade (GR&R), pelo método de análise de variância, para identificar os componentes de variação do sistema de medição avaliando três instrumentos distintos, comparando as medições realizadas por um analisador de imagens, a partir do ensaio metalográfico, e métricas de instrumentos convencionais de metrologia como o relógio apalpador e o paquímetro manual. Foram avaliadas duas características da qualidade, sendo elas: a profundidade de indentação e o diâmetro do ponto. A partir dos resultados foi possível verificar que o analisador de imagem apresentou uma menor variabilidade nas medições, caracterizando-se como a melhor escolha para as medições das respostas de qualidade do processo de solda por pontos apresentando um GR&R classificado como aceitável.

Palavras-chave: Profundidade de indentação; Diâmetro do ponto; Análise do sistema de medição; GR&R; Variabilidade.

Variation Causes Analysis Attributed to Different Metrological Instruments to Verify the Geometric Characteristics of a Spot Welding Process

Abstract: In this study, we sought to analyze the variability of instruments used in industry to evaluate the geometric characteristics of a welding point. For this purpose, the design of experiments technique was used to generate a fractional factorial design for the welding parameter configurations in test specimens of hot dip galvanized steels in order to represent the real amplitude of the resistance spot welding process. For the analyzes, the repeatability and reproducibility (GR&R) study was used by the analysis of variance method to identify the variation components of the measurement system by evaluating three different instruments, comparing the measurements performed by an image analyzer, from the metallography and measurements of conventional metrology instruments such as the dial-gauge and the caliper. Two quality characteristics were evaluated: indentation depth and nugget width. From the results it was possible to verify that the image analyzer presented a lower variability in the measurements, being the best choice for the measurements of the quality responses of the spot welding process, presenting a GR&R classified as acceptable.

Key-words: Indentation depth; Nugget width; Measurement system analysis; GR&R; Variability.

1. Introdução

O processo de soldagem por pontos, ou simplesmente RSW (*Resistance Spot Welding*) é um método de junção de estruturas amplamente utilizada no setor automobilístico [1], se destacando entre os processos de soldagem por apresentar características que favorecem a indústria com operação ágil, facilmente adequada para processos automáticos, manuseio simples, aplicações diversas e baixo custo [2-7]. Dada sua ampla aplicabilidade e importância na indústria, novas metodologias para ajuste de parâmetros são aplicadas para aperfeiçoar o RSW, favorecendo o controle e a capabilidade desse processo [8].

Artigo "*Aplicação da Metodologia de Superfície de Resposta para Otimização do Processo de Solda a Ponto no Aço Galvanizado AISI 1006*", publicado na "***Soldagem & Inspeção***" [219].

Soldagem & Inspeção. 2018;23(2):129-142
https://doi.org/10.1590/0104-9224/SI2302.02
ISSN 1980-6973 (Online)
ISSN 0104-9224 (Print)

Artigos Técnicos

Aplicação da Metodologia de Superfície de Resposta para Otimização do Processo de Solda a Ponto no Aço Galvanizado AISI 1006

Fernando Ferraz Amaral[1], Fabrício Alves de Almeida[1], Sebastião Carlos Costa[1], Rafael Coradi Leme[1], Anderson Paulo de Paiva[1]

[1] Universidade Federal de Itajubá – UNIFEI, Instituto de Engenharia de Produção e Gestão, Itajubá, MG, Brasil.

Recebido: 08 Mar., 2018
Aceito: 16 Maio, 2018

E-mail: fferrazamaral@yahoo.com.br (FFA)

Resumo: A soldagem a ponto por resistência elétrica possui um papel importante na fabricação de diversos produtos da indústria. Apesar de sua ampla aplicabilidade, este processo está sujeito a apresentar certa inconsistência na qualidade devido a variáveis controláveis e não controláveis. Desta forma, este estudo busca aplicar a metodologia de superfície de resposta para a obtenção de valores otimizados de fatores controláveis do processo, tais como corrente de soldagem, tempo de soldagem e força do eletrodo. Verificou-se o comportamento das curvas do deslocamento do eletrodo, a fim de observar o efeito da queima do revestimento das chapas de aço carbono galvanizadas. Avaliou-se também o teste de cisalhamento para verificar a deformação na força máxima de ruptura dos corpos de prova. Os resultados da aplicação mostraram que parâmetros elevados de corrente e parâmetros reduzidos dos tempos de ciclo proporcionaram maiores níveis de força e valores desejáveis de indentação. Além disso, foi possível verificar através dos experimentos que os critérios de qualidade do ponto de solda apresentam correlação com as características das curvas de deslocamento do eletrodo, proporcionando uma maneira eficaz para avaliar a qualidade do processo de solda a ponto a partir de testes não destrutivos.

Palavras-chave: Solda a ponto por resistência elétrica; Metodologia de superfície de resposta; Otimização; Aço galvanizado; Deslocamento do eletrodo.

Application of the Response Surface Methodology for Optimization of the Resistance Spot Welding Process in AISI 1006 Galvanized Steel

Abstract: The resistance spot welding has a great importance in the manufacture of several industrial products. Despite its wide applicability, this process is subject to certain quality inconsistency due to controllable and non-controllable variables. Thus, this study seeks to apply the response surface methodology to obtain optimized values of controllable process factors, such as welding current, welding time and electrode strength. The behavior of the electrode displacement curves has been verified in order to observe the burning effect of the coating of the galvanized carbon steel plates. It was also evaluated the shear test to verify the tensile-shear strength of the specimens. The results of the application showed that high current parameters and reduced cycle time parameters provided higher levels of strength and desirable indentation values. In addition, it was possible to verify through the experiments that the quality criteria of the weld point correlate with the characteristics of the displacement curves of the electrode, providing an efficient way to evaluate the quality of the process of welding to point from nondestructive tests.

Key-words: Resistance spot welding; Response surface methodology; Optimization; Galvanized steels; Electrode displacement.

1. Introdução

A soldagem a ponto por resistência elétrica (RSW– *Resistance Spot Welding*) possui grande importância na fabricação de diversos produtos da indústria, sendo um dos processos mais utilizados em produções seriadas e abrangendo muitos segmentos industriais como automobilísticas, eletrônicas, nucleares, tubulações, equipamentos ferroviários, aeroespacial, entre outras [1,2]. A facilidade de operação e de automação, rapidez, realização de vários pontos de solda sem que sejam necessários grandes ajustes dos parâmetros, e o baixo investimento, são os principais fatores que levam a grande utilização deste processo abrindo possibilidades para a obtenção de produtos de alta qualidade [3].

Artigo "*Measurement Data from bobbins of Partially Oriented Yarns: univariate and multivariate aspects*", publicado na "***Data in Brief***" [220].

Data in brief 27 (2019) 104637

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Measurement data from bobbins of Partially Oriented Yarns: Univariate and multivariate aspects

Fabrício A. Almeida [a, *], Daniel S. Cortez [a], Guilherme F. Gomes [b], Juliana H.D. Gaudêncio [a], Rachel C. Sabioni [c], José H.F. Gomes [a], Anderson P. Paiva [a]

[a] Institute of Industrial Engineering and Management, Federal University of Itajubá, Brazil
[b] Institute of Mechanical Engineering, Federal University of Itajubá, Brazil
[c] Department of Mechanical Systems Engineering, Sorbonne University, University of Technology of Compiègne, France

ARTICLE INFO

Article history:
Received 17 April 2019
Received in revised form 16 September 2019
Accepted 27 September 2019
Available online 7 October 2019

Keywords:
Partially oriented yarns
Measurement data
Automatic package analyser
Univariate and multivariate aspects
Gage repeatability and reproducibility

ABSTRACT

In this paper, we present data from measurements made in the textured fibers bobbins in two different conditions, presenting critical quality characteristics such as diameter, mass and density. In order to obtain a significant amount of information, in each of the two conditions, 270 measurements were obtained for each of the quality characteristics. Three different equipments (Automatic Package Analyzer - APA) were used in ten different parts, replicated three times for each of them. Considering the two measurement data collection, an amount of 540 bobbins measurements were obtained. Almeida et al., (2019) applied these measurement data in his study. Taking into account the multicorrelated nature of the information, we also have the representation of the principal components' scores for these measurements, besides the eigenvalues and eigenvectors of the data.

* Corresponding author.
E-mail address: fabricio-almeida@unifei.edu.br (F.A. Almeida).

https://doi.org/10.1016/j.dib.2019.104637

Artigo "*A linear programming optimization model applied to the decision-making process of a Brazilian e-commerce company*", publicado na "***Exacta***" [221].

EXACTA Engenharia de Produção eISSN 1983-9308

Artigos

https://doi.org/10.5585/ExactaEP.v17n3.8503

A linear programming optimization model applied to the decision-making process of a Brazilian e-commerce company

Um modelo de otimização de programação linear aplicado ao processo de tomada de decisão de uma empresa brasileira de comércio eletrônico

Prof. MSc. Fabrício Alves de Almeida[1]
Prof. MSc. João Paulo Barbieri[2]
Prof. PhD. José Arnaldo Barra Montevechi[3]
Prof. PhD. José Henrique de Freitas Gomes[4]
Prof. PhD. Alexandre Ferreira de Pinho[5]

[1]Professor at the Department of Economics of the Faculty of Applied Social Sciences of the South of Minas (FACESM), Master in Production Engineering at the Federal University of Itajubá (UNIFEI) and Bachelor in Economic Sciences at the FACESM. Itajubá, Minas Gerais, Brazil. fabricio.alvesdealmeida@gmail.com

[2]Professor at the Department of Management Science of the Federal Institute of South of Minas Gerais, Master in Production Engineering at the Federal University of Itajubá (UNIFEI) and Bachelor in Management Science at FACESM. Machado, Minas Gerais, Brazil. joao.barbieri@ifsuldeminas.edu.br

[3]Professor at the Institute of Industrial Engineering and Management of the Federal University of Itajubá (UNIFEI), PhD in Mechanical Engineering at the University of São Paulo (USP), Master in Mechanical Engineering at the Federal University of Santa Catarina (UFSC) and Bachelor in Mechanical Engineering at the UNIFEI. Itajubá, Minas Gerais, Brazil. montevechi@unifei.edu.br

[4]Professor at the Institute of Industrial Engineering and Management of the Federal University of Itajubá (UNIFEI), PhD in Production Engineering at the UNIFEI, Master in Production Engineering at the UNIFEI and Bachelor in Mechanical Engineering at the UNIFEI. Itajubá, Minas Gerais, Brazil. ze_henriquefg@unifei.edu.br

[5]Professor at the Institute of Industrial Engineering and Management of the Federal University of Itajubá (UNIFEI), PhD in Mechanical Engineering at the São Paulo State University (UNESP), Master in Production Engineering at the UNIFEI and Bachelor in Mechanical Engineering at the UNIFEI. Itajubá, Minas Gerais, Brazil. pinho@unifei.edu.br

Abstract
The decision-making process is not always simple and requires a more careful analysis to maximize the company's revenue. This paper proposes a linear programming model applied to the decision-making of the section of quality monitoring and packaging of a Brazilian company of e-commerce, in which the simplex method was used to maximize the company's revenue from historical time data of the activities for each type of product. From the results, it was verified which products should be prioritized, providing a revenue of US$ 74,681.50. In addition, a simulation was applied to include two employees in the process, which would provide a 32.76% increase in the company's profitability and a new revenue of US$ 99,145.00.

Key-words: Linear programming; Optimization; Simplex method; E-commerce; Decision-making.

Resumo
O processo de tomada de decisão nem sempre é simples e requer uma análise mais cuidadosa para maximizar a receita da empresa. Este trabalho propõe um modelo de programação linear aplicado à tomada de decisão do setor de monitoramento e embalagem da qualidade de uma empresa brasileira de e-commerce, no qual o método simplex foi utilizado para maximizar a receita da empresa a partir de dados históricos das atividades para cada tipo de produto. A partir dos resultados, verificou-se quais produtos deveriam ser priorizados, proporcionando uma receita de US$ 74.681,50. Além disso, uma simulação foi aplicada para incluir dois funcionários no processo, o que proporcionaria um aumento de 32,76% na lucratividade da empresa e uma nova receita de US$ 99.145,00.

Palavras-chave: Programação linear; Otimização; Método Simplex; E-commerce; Tomada de decisão.

Artigo "*Analysis of Concentration Measures Applied to Manufacture Industry in a State of Brazilian Northeast*", publicado na "***International Journal of Engineering Applied Sciences and Technology***" [222].

International Journal of Engineering Applied Sciences and Technology, 2017
Vol. 2, Issue 7, ISSN No. 2455-2143, Pages 9-13
Published Online in IJEAST (http://www.ijeast.com)

ANALYSIS OF CONCENTRATION MEASURES APPLIED TO MANUFACTURE INDUSTRY IN A STATE OF BRAZILIAN NORTHEAST

Fabrício Alves de Almeida
Institute of Industrial Engineering and Management
Federal University of Itajubá,
Av. BPS, 1303, Brazil

Antonio Suerlilton Barbosa da Silva
Faculty of Applied Social Sciences of the South of Minas,
Av. Pres. Tancredo de Almeida Neves, 45, Brazil

Fabianne Alves de Almeida
BSP Business School São Paulo
Anhembi Morumbi University,
Av. Casa do Ator, 275, Brazil

Abstract — **Obtain analytical knowledge of market structures of a given region is useful in managerial decision-making and establishing company pricing strategies and policies. In view of these attributions, this article presents the application and analysis of the concentration level to the industrial sector in a state of northeastern Brazil. For this, three concentration indicators were used: Concentration Ratio (CR_k), Hirchsman-Herfindahl (HHI) and Theil's entropy coefficient (TE). The results showed that the increase in the number of employees in all segments over the period 2002 to 2013 allows us to infer a possible economic growth for this region, which can be better seen in future research on the sector.**

Keywords— **Industrial concentration; Manufacturing industry; Concentration indices; Northeastern Brazil.**

I. INTRODUCTION

According to [1], the conduct and performance of firms result from their market structure, where this structure influences the actions to be taken by the firm. The conduct favors the attitudes used by the firms in order to adapt to the market.
Knowing the market structure of a given region is useful in managerial decision-making and, more specifically, in establishing company pricing strategies and policies. For [2], this set of strategies is defined as the conduct of the company, being characterized by virtue of the Structure-Conduct-Performance (SCP) model. The perception of the market structure, combined with the quantification of the structural component in terms of synthetic measures, is strategic in the SCP paradigm, as well as finding broad use in studies on Industrial Economics, for example, [3-8].

The empirical applications of concentration are supported in concentration measurements. These measures have the objective of capturing how economic agents show a "dominant behavior" in a given market. In this sense, the different indicators consider the market share of the agents, for example, the number of employees of each company for the total employment bonds of the sector (according to different weighting criteria). Industrial concentration measures are useful to indicate, preliminarily, the sectors for which "market power" is expected to be significant [9].

However, some empirical applications often rely on deficient concentration measurements, which has served as motivation for comparative presentations of the main measures used in studies in the area. The results of these motivations have given rise to studies that address different methodological lines with emphasis on industrial concentration, but finding results with a high degree of divergence [10-12].

In the process of regional migration of industries, new industrial agglomerations occur in peripheral regions, inducing research on industrial concentration and contributing to these studies gain more space in academic debate and development promotion agencies and institutions to support micro and small enterprises. This process draws the attention of the research support agencies to the valorization and, consequently, sponsorship to the development of investigations whose objective is the verification and explanation on the phenomenon of industrial concentration.

However, there are positive aspects, a priori, arising from this concentration. An example of this is the fact that high concentration does not reflect in oligopolistic practices, given the need for technological innovations and modernizations [8]. In view of these and other attributions, this article presents the application and analysis of the level of concentration of the industrial sector in a state of northeastern Brazil. For this, three concentration indicators were used: Concentration Ratio (CR_k), Hirchsman-Herfindahl (HHI) and Theil's entropy coefficient (TE).

This paper is organized as follows. The mathematical models to measure the level of concentration in section II. The case study in section III. The application and results are presented in section IV. Concluding remarks are given in section V.

ANEXO B – Artigos publicados em congressos

A lista abaixo indica as principais publicações realizadas em congressos nacionais e internacionais durante o período de doutorado do proponente:

- ALMEIDA, F. A., BARBIERI, J. P., ROMÃO, E. L., STREITENBERGER, S. C., GOMES, J. H. F. Modelagem e otimização linear baseada no método Simplex para gerenciar estoque em uma empresa de comércio eletrônico. **XXXX Encontro Nacional de Engenharia de Produção**, 2020.
- STREITENBERGER, S. C., ROMÃO, E. L., ALMEIDA, F. A., PAIVA, A. P. Metodologia de apoio à tomada de decisão sobre melhoria de processos industriais baseada em análise multivariada de dados e técnicas de otimização. **XXXX Encontro Nacional de Engenharia de Produção**, 2020.
- ROMÃO, E. L., STREITENBERGER, S. C., ALMEIDA, F. A., BALESTRASSI, P. P. Estudo comparativo entre modelos auto-regressivos integrados de médias móveis e redes neurais artificiais na modelagem e previsão de séries econométricas. **XXXX Encontro Nacional de Engenharia de Produção**, 2020.
- SILVA, A. S. R., ALMEIDA, F. A. Mensuração da concentração industrial baseado em análise hierárquica de cluster das empresas que compõem o índice IBrX-50 entre 2014 e 2018. **XXXX Encontro Nacional de Engenharia de Produção**, 2020.
- PACHECO, L. C., ALMEIDA, F. A. Análise das oscilações da taxa Selic e seus impactos cambiais no período de 2011 ao primeiro trimestre de 2020. **XII Encontro Científico Sul Mineiro de Administração, Contabilidade e Economia** (ECOSUL), 2020.
- PODDIS, N. S., ALMEIDA, F. A. Análise do perfil de usuários de instituições financeiras diante da inclusão digital: um estudo à luz da microrregião de Itajubá. **XII Encontro Científico Sul Mineiro de Administração, Contabilidade e Economia** (ECOSUL), 2020.
- SILVA, R. K. S., ALMEIDA, F. A. Análise de indicadores macroeconômicos pós plano real: um estudo à luz dos governos entre 1995 à 2019. **XII Encontro Científico Sul Mineiro de Administração, Contabilidade e Economia** (ECOSUL), 2020.
- SANTOS, D. J., ALMEIDA, F. A. Análise e mensuração da concentração industrial: um estudo à luz das empresas com mais representatividade no mercado de ações brasileiro. **XII Encontro Científico Sul Mineiro de Administração, Contabilidade e Economia** (ECOSUL), 2020.
- MONTEIRO, C. E. O., TORRES, A. F., ALMEIDA, F. A., BALESTRASSI, P. P. Comparative study of water inflow forecasts: a case study in Brazil. **XIII Latin-American Congress on Electric Power Generation, Transmission and Distribution**, CLAGTEE 2019, Santiago - Chile, 2019.
- ALMEIDA, F.A., GOMES, J. H. F., BELINATO, G., PAULA, T. I., PAULA, V. R. Um estudo de GR&R aninhado para identificar causas de variação no sistema de medição do processo de solda por pontos. **XXXVIII Encontro Nacional de Engenharia de Produção**, Maceió – AL, 2018.
- BIANCHESI, N. M. P., PAULA, V. R., ALMEIDA, F. A., BELINATO, G., BALESTRASSI, P. P. Aplicação de princípios e ferramentas da gestão da qualidade total em uma empresa francesa de manutenção em transporte ferroviário. **XXXVIII Encontro Nacional de Engenharia de Produção**, Maceió – AL, 2018.

- BELINATO, G., ALMEIDA, F.A., PAULA, V. R., BALESTRASSI, P. P., ROSA, P. A. R. Aplicação Multivariada do Método de Interseção Normal à Fronteira para Otimização do Processo de Usinagem a Laser do Aço AISI 314S. **XXXVIII Encontro Nacional de Engenharia de Produção**, Maceió – AL, 2018.
- GASPAR JUNIOR, F. C., ALMEIDA, F.A., LEITE, R. R., BELINATO, G., GOMES, J. H. F. Otimização de um processo de torneamento interno de buchas de ferro fundido nodular em uma empresa do setor de autopeças. **XXXVIII Encontro Nacional de Engenharia de Produção**, Maceió – AL, 2018.
- ALMEIDA, F. A., AMORIM, L. F., PAULA, T. I., DE PAULA, V. R., SABIONI, R. C., PAIVA, A. P., GOMES, J. H. F. Projeto de parâmetro robusto aplicado à otimização do processo de torneamento do aço ABNT 12L14. **XLIX Simpósio Brasileiro de Pesquisa Operacional**, Blumenau – Santa Catarina, 2017.
- ALMEIDA, F. A., SABIONI, R. C., DE PAULA, V. R., CORTEZ, D. S., GOMES, J. H. F. Análise e Validação do Sistema de Medição de um Processo de Texturização utilizando GR&R Multivariado. **XXXVII Encontro Nacional de Engenharia de Produção**, Joinville – Santa Catarina, 2017.
- RIBEIRO, G. F., DE PAULA, V. R., ALMEIDA, F. A., SABIONI, R. C., TURRIONI, J. B. Análise da criação e implantação de documentação pop (procedimento operacional padrão) em uma empresa do setor aeronáutico. **XXXVII Encontro Nacional de Engenharia de Produção**, Joinville – Santa Catarina, 2017.
- CORREA, J. E., DE PAULA, V. R., SABIONI, R. C., ALMEIDA, F. A., TURRIONI, J. B. Desenvolvimento de um modelo de medição do sistema de acreditação hospitalar brasileiro utilizando modelagem de equações estruturais. **XXXVII Encontro Nacional de Engenharia de Produção**, Joinville – Santa Catarina, 2017.
- LEITE, P. N., SABIONI, R. C., ALMEIDA, F. A., DE PAULA, V. R., GOMES, J. H. F. Estudo de repetitividade e reprodutibilidade para análise do sistema de medição de um processo de etiquetagem de bombas. **XXXVII Encontro Nacional de Engenharia de Produção**, Joinville – Santa Catarina, 2017.
- CORREA, J. E., DE PAULA, V. R., ALMEIDA, F. A., SABIONI, R. C., TURRIONI, J. B. Estudo de caso: uso da metodologia AHP para priorizar o *servqual* na avaliação da qualidade em hospitais. **XXXVII Encontro Nacional de Engenharia de Produção**, Joinville – Santa Catarina, 2017.