UNIVERSIDADE FEDERAL DE ALFENAS
UNIFAL-MG

JOSÉ MÁRCIO MARTINS JÚNIOR

ANÁLISE DE ARQUÉTIPOS: INTRODUÇÃO A TEORIA E APLICAÇÕES

ALFENAS - MG
2015

JOSÉ MÁRCIO MARTINS JÚNIOR

ANÁLISE DE ARQUÉTIPOS: INTRODUÇÃO A TEORIA E APLICAÇÕES

Dissertação apresentada à Universidade Federal de Alfenas, como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria da Universidade Federal de Alfenas de Minas Gerais. Área de concentração: Estatística Aplicada e Biometria. Linha de pesquisa: Modelagem Estatística e Estatística Computacional.

Orientador: Dr. Eric Batista Ferreira
Coorientadores: Prof. Dr. Denismar Alves Nogueira
e Dr. Daniel Furtado Ferreira

ALFENAS - MG
2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
Biblioteca Central da Universidade Federal de Alfenas

Martins Júnior, José Márcio.
Análise de Arquétipos: introdução a teoria e aplicações / José Márcio Martins Júnior. -- Alfenas/MG, 2015.
61 f.

Orientador: Eric Batista Ferreira.
Dissertação (mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Alfenas, 2015.
Bibliografia.

1. Análise Multivariada. 2. Análise Sensorial. 3. Metodo de Monte Carlo. I. Ferreira, Eric Batista. II. Título.

CDD-519.53

JOSÉ MÁRCIO MARTINS JÚNIOR

"ANÁLISE DE ARQUÉTIPOS: INTRODUÇÃO À TEORIA E APLICAÇÕES".

A Banca Examinadora, abaixo assinada, aprova a Dissertação apresentada como parte dos requisitos para a obtenção do título de Mestre em Estatística Aplicada e Biometria pela Universidade Federal de Alfenas. Linha de Pesquisa: Modelagem Estatística e Estatística Computacional.

Aprovado em: 10 de abril de 2015.

Prof. Dr. Eric Batista Ferreira
Instituição: UNIFAL-MG
Assinatura: Eric Batista Ferreira

Prof. Dr. Júlio Silvio de Sousa Bueno Filho
Instituição: UFLA
Assinatura: Júl Bu

Prof.ª Dr.ª Adriana Dias
Instituição: UNIFAL-MG
Assinatura: Adriana Dias

Prof. Dr. Luiz Alberto Beijo
Instituição: UNIFAL-MG
Assinatura: Beijo

Aos avós da Heloisa
Aneci e José Márcio,
dedico este trabalho
a vocês.

AGRADECIMENTOS

Agradeço a Deus, em primeiro lugar, por me dar forças para concluir esta etapa da minha vida.

Aos meus pais José Márcio e Aneci pelo incentivo e apoio incondicional nas horas que mais precisei.

A minha esposa Regiane pela compreensão, amor e por ter me dado o meu maior presente, minha filha Heloisa.

Agradeço muito ao meu amigo e orientador Eric. O maior exemplo de trabalho e dedicação que já tive o prazer de conviver. Foi uma honra trabalhar com você.

Agradeço aos meus coorientadores, Daniel e Denismar.

Agradeço também minhas colegas, Mariana, Michelle, Lislaine e Bruna, desejo todo o sucesso do mundo para vocês.

Agradeço também ao Elcio, pela paciência, dedicação e seus sábios conselhos.

À Universidade Federal de Alfenas e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, principalmente a seus docentes e servidores.

Agradeço a CAPES, pela auxílio financeiro.

À todos que contribuíram de alguma forma para a realização deste trabalho, meu muito obrigado.

RESUMO

Existem diversas técnicas que auxiliam na interpretação e análise de dados multivariados. Uma das técnicas mais difundidas e utilizadas é a Análise de Componentes Principais, que tem como principal objetivo reduzir a dimensionalidade dos dados afim de facilitar a interpretação. A Análise de Arquétipos também é uma técnica multivariada que busca reduzir a dimensionalidade dos dados, mas por meio de combinações convexas dos próprios dados. Os arquétipos são selecionados pela minimização da soma de quadrados de resíduos que representa o erro cometido ao se reconstruir os dados originais utilizando os arquétipos. Este trabalho teve como objetivo geral explorar em mais detalhes a Análise de Arquétipos e traçar a sua história, além de verificar potencialidades de aplicações existentes e futuras em diversas áreas do conhecimento. De forma mais específica objetivou-se: descrever em detalhes a Análise de Arquétipos, identificar similaridades e diferenças entre a Análise de Arquétipos e a Análise de Componentes Principais, realizar estudos de simulação para avaliar qual a melhor métrica para medir a falta de ajuste dos dados recompostos pelos arquétipos, aplicar a técnica Análise de Arquétipos em dados sobre a movimentação de jogadores de futebol, realizar simulações Monte Carlo para avaliar se há algum ganho em executar a Análise de Arquétipos em conjunto com a Análise de Componentes Principais, e aplicar esta em dados sensométricos experimentais sobre hambúrgueres. As metodologias utilizadas foram uma extensa revisão bibliográfica e simulações Monte Carlo. Os resultados mostraram que a Análise de Arquétipos é uma técnica com ampla aplicabilidade e excelentes resultados práticos. O estudo das métricas concluiu que a soma de quadrados de resíduos deve ser utilizada por ser a mais simples, e que não há qualquer prejuízo em utilizar esta métrica em relação as outras estudadas. No contexto do futebol a Análise de Arquétipos foi capaz de verificar se o jogador ou grupo de jogadores atuam em seu espaço designado, se a apostura foi ofensiva ou defensiva entre outras, sendo uma nova abordagem para ser utilizada em conjunto com outras técnicas para este fim. No estudo de simulação que visava aplicar as duas técnicas, foi evidenciado a capacidade de reconstrução das técnicas em conjunto e a melhora na interpretabilidade que as técnicas apresentam quando utilizadas conjuntamente.

Palavras-chave: Análise Multivariada. Análise Sensorial. Método de Monte Carlo.

ABSTRACT

There are several techniques that aid in the interpretation and analysis of multivariate data. One of the most widespread techniques used and is the Principal Component Analysis, which aims to reduce the dimensionality of the data in order to facilitate interpretation. The Archetypal Analysis is also a multivariate technique that seeks to reduce the dimensionality of the data, but through convex combinations of the data itself. Archetypes are selected by minimizing the residual sum of squares that is the mistake to reconstruct the original data using the archetypes. This work aimed to explore in more detail the Archetypal Analysis and trace its history, and to identify capabilities of existing and future applications in various areas of knowledge. More specifically it aimed to: describe in detail the Archetypal analysis, identify similarities and differences between the Archetypal Analysis and Principal Component Analysis, conduct simulation studies to assess how best metric for measuring the lack of fit of the data recomposed by archetypes, apply the Archetypal Analysis on data of the movement of soccer players, perform Monte Carlo simulations to assess whether there is some gain in performing Archetypal Analysis in conjunction with the Principal Component Analysis, and apply this to experimental sensory data of burgers. The methodologies used were an extensive literature review and Monte Carlo simulations. The results showed that the Archetypal Analysis is a technique with wide applicability and excellent practical results. The study of metrics concluded that the residual sum of squares should be used because it is the simplest, and that there is no harm in using this metric compared to the others tha was studied. In the context of the soccer, Archetypal Analysis was able to verify that the player or group of players are in their designated space, the offensive or defensive behavior among others, being a new approach to be used in conjunction with other techniques for this purpose. About the study of simulation designed to implement both techniques, was evidenced a improvement of reconstruction capacity and in the interpretability.

Keywords: Monte Carlo method. Multivariate analysis. Sensory analysis.

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Existem diversas técnicas que auxiliam na interpretação e análise de dados multivariados. Uma técnica relativamente nova é a Análise de Arquétipos. Proposta por Cutler e Breiman (1994), esta técnica busca reduzir a dimensionalidade dos dados por meio de combinações convexas dos próprios dados, utilizando os elementos mais representativos chamados de arquétipos. Os arquétipos são selecionados pela minimização da soma de quadrados de resíduos que representa o erro cometido ao se reconstruir os dados originais utilizando os arquétipos.

Por ser uma técnica recente, a Análise de Arquétipos merece ser explorada em mais detalhes. No Brasil esta técnica não é difundida, pois não foi encontrado material científico em língua portuguesa que aborde o assunto.

Este trabalho apresenta algumas aplicações existentes e ainda propõe o uso de arquétipos na Análise Sensorial e na análise de dados sobre movimentação em esportes como o futebol.

Este trabalho teve como objetivo geral explorar em mais detalhes a Análise de Arquétipos e traçar a sua história, além de verificar potencialidades de aplicações existentes e futuras em diversas áreas do conhecimento. De forma mais específica objetivou-se: descrever em detalhes a Análise de Arquétipos, identificar similaridades e diferenças entre a Análise de Arquétipos e a Análise de Componentes Principais, realizar simulações Monte Carlo para avaliar se há algum ganho em executar a Análise de Arquétipos em conjunto com a Análise de Componentes Principais, realizar estudos de simulação para avaliar qual a melhor métrica para medir a falta de ajuste dos dados recompostos pelos arquétipos, aplicar as técnicas Análise de Arquétipos e Análise de Componentes Principais conjuntamente em dados sensométricos experimentais sobre hambúrgueres, aplicar a técnica Análise de Arquétipos em dados sobre a movimentação de jogadores de futebol;

Esta dissertação foi organizada como um conjunto de trabalhos sobre a Análise de Arquétipos, alguns destes já foram publicados como artigos. Nesta seção serão apresentadas as introduções referentes a cada um dos artigos, que são eles: “A Análise de Arquétipos: uma revisão bibliográfica” (MARTINS JÚNIOR et al., 2015b). “Avaliação Monte Carlo de métricas para falta de ajuste em Análise de Arquétipos”, será chamado apenas de “Métricas” (MARTINS JÚNIOR et al., 2014). “Análise de Arquétipos na avaliação da movimentação de jogadores de futebol”, será chamado de “Futebol” (MARTINS JÚNIOR et al., 2015a). O artigo provisoriamente denominado “Arquétipos e componentes principais concentrando a informa-

ção sensorial", que contempla a simulação proposta nesta dissertação, sua introdução também será apresentada nesta seção.

O artigo provisoriamente denominado "Arquétipos e componentes principais concentrando a informação sensorial", a partir de agora "Arquétipos e componentes principais concentrando informação", que contempla a simulação proposta nesta dissertação, sua introdução também é explicada nesta seção.

1.1 A ANÁLISE DE ARQUÉTIPOS: UMA REVISÃO BIBLIOGRÁFICA

Experimentos na prática coletam várias variáveis que se estudadas de forma simultânea, levam a resultados mais otimizados do que quando são analisadas de forma individual ou univariada. Porém, é comum experimentos multivariados serem analisados de forma univariada pela complexidade das análises envolvidas. Existem diversos métodos estatísticos para analisar, descrever e inferir sobre experimentos mensurados em mais de uma variável.

A Análise de Arquétipos é uma técnica multivariada que busca reduzir a dimensão de dados por meio de combinações convexas dos arquétipos, que são geralmente valores extremos dos dados. Os arquétipos são selecionados pela minimização da soma de quadrados de resíduos, que é o erro cometido ao se reconstruir os dados originais utilizando as combinações convexas. A Análise de Arquétipos não é uma técnica difundida como a Análise de Componentes Principais por exemplo, e é uma técnica relativamente nova e pouco estudada.

O Merriam-Webster online dictionary (2015) define arquétipo como: "padrão ou modelo no qual todas as coisas do mesmo tipo são representações ou cópias". Já o Dicionário online Michaelis (2015) define arquétipo como: "modelo dos seres criados. O que serve de modelo ou protótipo".

A Análise de Arquétipos (AA) é uma técnica multivariada introduzida por Cutler e Breiman (1994) e tem como propósito simplificar a estrutura de covariâncias, sendo utilizada para reduzir a dimensão de dados por meio de combinações convexas dos seus elementos mais representativos facilitando assim a interpretação. Uma combinação convexa é um caso particular da combinação linear e ocorre quando os coeficientes da combinação são valores positivos ou nulos e o somatório dos coeficientes resulta em um.

Os arquétipos são selecionados pela minimização da soma de quadrados de resíduos

(SQR) de representar cada observação x_i dos dados originais $\boldsymbol{X}$ como uma combinação convexa dos arquétipos $\boldsymbol{Z}$. Se uma determinada observação for um dos arquétipos, neste caso, será denominado arquétipo puro, e estará geralmente na fronteira do fecho convexo dos dados. Um fecho convexo é obtido quando dado dois pontos dentro de um conjunto, o menor caminho entre os dois pontos também reside dentro do conjunto. Os arquétipos são geralmente valores extremos que melhor representam os dados.

Sabe-se que quanto maior o número de arquétipos selecionados menor é a SQR, pois menos informação é perdida, e melhor é a captura da forma dos dados. Mas como consequência, menor é a redução da dimensão dos dados. Então, fica a cargo do pesquisador decidir quantos arquétipos (K) deve-se utilizar em um determinado conjunto de dados, desde que $1 \leq K \leq N$, em que N é o número de elementos na fronteira do fecho convexo. Recomenda-se o uso de um gráfico *screeplot* para ajudar na decisão de quantos arquétipos usar (CUTLER; BREIMAN, 1994). Um gráfico *screeplot* apresenta a quantidade de variação explicada pelo número de arquétipos para facilitar a definição da quantidade de arquétipos que deverá ser utilizada. Este tipo de gráfico é muito comum na ACP para definir o número de componentes a ser retidos e também na análise de fatores.

1.2 O MODELO MATEMÁTICO

Cada elemento $X_i \in \boldsymbol{X}$, em que $\boldsymbol{X}$ são os dados originais de tamanho n, é reescrito como uma combinação convexa dos arquétipos $\boldsymbol{z}_k \in \boldsymbol{Z}$ em que $\boldsymbol{Z}$ é a matriz $K \times p$ de todos os arquétipos, K é o número de arquétipos e p o número de variáveis, para recompor as informações originais, da forma

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \alpha_{ik} \boldsymbol{z}_k \tag{1.1}$$

em que $\alpha_{ik} \in \mathbb{R}$ é um escalar, $\boldsymbol{z}_k \in \mathbb{R}^p$ é um vetor denominado arquétipo.

Usando a soma de quadrados dos resíduos (SQR) para selecionar qual será o melhor conjunto de arquétipos, a minimização da Equação 1.2 é que gera este conjunto de arquétipos.

$$\sum_{i=1}^{n} ||\boldsymbol{X}_i - \sum_{k=1}^{K} \alpha_{ik} \boldsymbol{z}_k||^2 \tag{1.2}$$

em que $\boldsymbol{X}_i$ são as observações e n o número de observações.

Para dados multivariados $(\boldsymbol{X}_i, i = 1, \ldots, n)$ em que cada $\boldsymbol{X}_i$ é um vetor p-dimensional $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{ip})'$, o padrão arquétipo de uma massa de dados caracteriza o problema de encontrar vetores p-dimensionais $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$ com $1 \leq K \leq N$, sendo N o número de elementos na fronteira do fecho convexo (BAUCKHAGE; THURAU, 2009)

$$\boldsymbol{Z} = \sum_{i=1}^{n} \sum_{k=1}^{K} \beta_{ik} \boldsymbol{X}_i \tag{1.3}$$

em que $k = 1, \ldots, K$ e os coeficientes $\beta_{ik} \in \mathbb{R}$, com $\beta_{ik} \geq 0$ e $\sum_{i=1}^{n} \beta_{ik} = 1$. Essas duas restrições dos coeficientes caracterizam a combinação convexa.

Segundo Stone e Olson (1999), os próprios arquétipos aparecem como pontos no conjunto dos dados (em alguns casos podem ser alguns dos pontos dos dados), e os pontos reconstruídos nunca estarão fora do fecho convexo do conjunto original dos dados.

Para selecionar qual será o conjunto de arquétipos, a minimização da Equação 1.4 é que gera este conjunto.

$$SQR = \sum_{i=1}^{n} ||\boldsymbol{X}_i - \sum_{k=1}^{K} \alpha_{ik} \boldsymbol{z}_k||^2 = \sum_{i=1}^{n} ||\mathbf{X}_i - \sum_{k=1}^{K} \alpha_{ik} \sum_{l=1}^{n} \beta_{lk} \mathbf{X}_l||^2 \tag{1.4}$$

A resolução desta equação é feita por meio de um algoritmo descrito na seção 1.3, de forma iterativa e alternada em minimizar a SQR alterando o conjunto de coeficientes ($\boldsymbol{\alpha}$) de um dado conjunto de arquétipos fixos ($\boldsymbol{Z}$), e em seguida minimizar a SQR alterando o conjunto de arquétipos fixados anteriormente. Estes coeficientes podem ser encontrados utilizando algum método que resolva problemas de quadrados mínimos não-negativos, mais detalhes podem ser obtidos em (BRO; JONG, 1997). Desta forma a SQR sempre diminui a cada iteração, porém o resultado deste algoritmo pode "convergir" para um mínimo local. Cutler e Breiman (1994) recomendam inicializar diversas vezes o algoritmo, utilizando diferentes arquétipos iniciais. O critério de parada ocorre quando a SQR for um valor suficientemente pequeno, que é definido de maneira subjetiva pelo pesquisador no contexto de sua pesquisa. Se o objetivo for apenas descritivo uma SQR maior pode ser aceito. Se o objetivo for reconstrução dos dados, deve-se então optar pela menor SQR possível, porém, maior será o esforço computacional necessário para alcança-la.

O problema de encontrar o conjunto de arquétipos e seus respectivos coeficientes ($\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$) é de ordem quadrática ($O(n^2)$), por isso muitas vezes é inviável utilizar a Análise de

Arquétipos dependendo do tamanho da massa de dados a ser analisada. Bauckhage e Thurau (2009) propuseram uma restrição que diminui consideravelmente o tempo de execução deste algoritmo. A proposta é não incluir os pontos que estão fora da fronteira do fecho convexo dos dados para seleção dos arquétipos. Como os arquétipos são usualmente valores extremos, estes valores estão geralmente localizados na fronteira do fecho convexo.

O melhor número de arquétipos para um conjunto de dados é desconhecido na prática, por isso deve ser empiricamente computado (CORSARO; MARINO, 2010). Quando se retém apenas um arquétipo $(K = 1)$, este será o valor médio dos dados, pois é fácil notar de acordo com 1.2 que o valor que minimiza a SQR para encontrar o arquétipo único será o valor médio. Para $K = 2$, os dados serão reescritos em forma de uma reta que terá cada um dos arquétipos em uma das extremidades. Para $K = 3$, obtém-se as novas coordenadas dos dados como um triangulo, para $K = 4$ um quadrilátero e assim sucessivamente até o limite dos números de arquétipos que são o número de elementos no fecho convexo dos dados.

De acordo com o aumento do número de arquétipos, as estruturas geométricas tornam-se cada vez mais difíceis de serem interpretadas (PORZIO; RAGOZINI; VISTOCCO, 2008).

Bauckhage e Thurau (2009) descrevem (1.4) em forma matricial. Assim, o conjunto de dados $\boldsymbol{X}_i \in \mathbb{R}^p$ compõe uma matriz $\boldsymbol{X}_{(p\times n)}$ e os arquétipos $\boldsymbol{z}_k \in \mathbb{R}^p$ compõem uma matriz $\boldsymbol{Z}_{(p\times k)}$.

$$SQR = ||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{A}||^2 = ||\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}||^2 \tag{1.5}$$

em que $\boldsymbol{A} \in \mathbb{R}^{K\times n}$ e $\boldsymbol{B} \in \mathbb{R}^{n\times K}$.

A Equação 1.5 utiliza a métrica quadrática para calcular o tamanho do erro. Outros autores como Corsaro e Marino (2010), utilizam a Norma de Frobenius. E afirmam que a utilização desta métrica é motivada pelo fato de que esta permite manter sob controle tanto a distância entre os centros quanto a localização, visando controlar a acurácia.

Bauckhage e Thurau (2009) também afirmam que dentre os métodos de redução de dimensionalidade, ou técnicas de agrupamento, a AA produz resultados facilmente interpretáveis por especialistas. E além disso, permite uma leve classificação e agrupamento dos dados, pois os coeficientes α_{ik} dos pontos $\boldsymbol{X}_i$ podem ser interpretados como probabilidades $p(\boldsymbol{X}_i|\boldsymbol{z}_k)$ indicando qual a classe (arquétipo $\boldsymbol{z}_k$) mais provável de representar o ponto.

1.3 O ALGORITMO

Para realizar a AA os dados devem estar na mesma escala de medida dentro de cada variável.

Segundo Chan, Mitchell e Cram (2003), a Análise de Arquétipos é sensível a *outliers* e por isso deve-se certificar que não haja *outliers* nos dados ou o resultado pode ser prejudicado. O algoritmo consiste dos seguintes passos:

(i) Sortear os coeficientes $\boldsymbol{\beta}$ de forma arbitrária seguindo as restrições da combinação convexa, $\beta_i > 0$ e $\sum_{i=1}^{n} \beta_i = 1$;

(ii) Encontrar o melhor conjunto de coeficientes $\boldsymbol{\alpha}$ para reescrever $\boldsymbol{X}$ com os arquétipos $\boldsymbol{Z}$ por meio de resolver n problemas de quadrados mínimos não-negativos; (i = 1, . . . , n)

$$min(||\boldsymbol{X}_i - \boldsymbol{Z}\boldsymbol{\alpha}_i||^2) \tag{1.6}$$

com a restrição de combinação convexa, $\alpha_i > 0$ e $\sum_{i=1}^{n} \alpha_i = 1$;

(iii) Recalcular os novos arquétipos ($\tilde{\boldsymbol{Z}}$) por meio de resolver o sistema linear de equações $\boldsymbol{X} = \tilde{\boldsymbol{Z}}\boldsymbol{\alpha}$

(iv) Encontrar o melhor conjunto $\boldsymbol{\alpha}$ para os arquétipos ($\tilde{\boldsymbol{Z}}$): resolver K problemas de quadrados mínimos não-negativos ($k = 1, \ldots, K$);

$$min_{\beta}||\tilde{\boldsymbol{Z}}_k - \boldsymbol{X}\beta_k||^2 \tag{1.7}$$

(v) Calcular a SQR;

(vi) Repetir os passos de (i) a (v) até que a SQR seja pequena o suficiente, o que varia de tamanho de acordo com o critério do pesquisador e ao contexto da pesquisa, por isso recomenda-se o uso de um gráfico *Screeplot*;

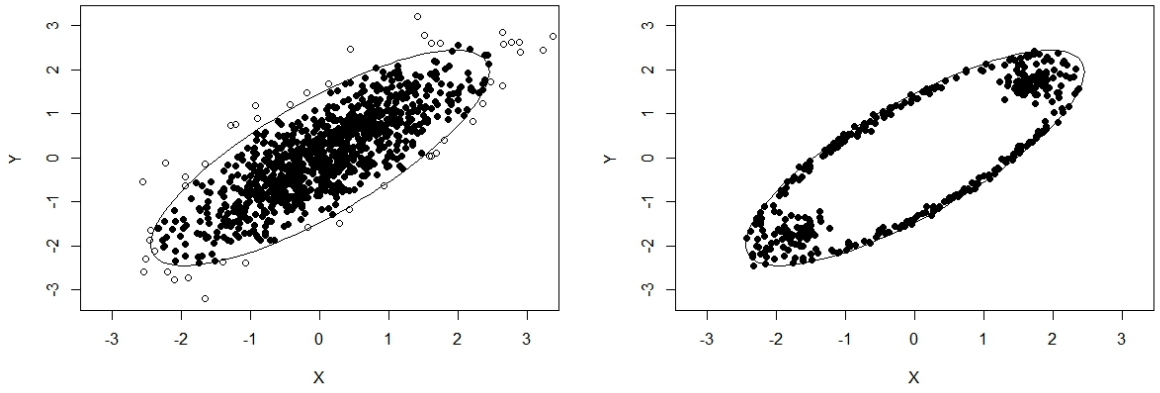
Os passos (ii) e (iv) demandam mais esforços computacionais por conta das restrições da combinação convexa, por isso deve-se optar por um algoritmo eficiente para tal cálculo. Na literatura é comum utilizar o método chamado Quadrados Mínimos Não-Negativos para encontrar estes coeficientes.

O passo (iii) exige a resolução de um conjunto de equações lineares. Tal método pode ser implementado utilizando a técnica de inversa generalizada de Moore-Penrose. $\tilde{\boldsymbol{Z}} = \alpha^{+}\boldsymbol{X}$ ou utilizando decomposição $\boldsymbol{QR}$, em que $\boldsymbol{Q}$ é uma matriz ortogonal e $\boldsymbol{R}$ uma matriz triangular superior, $\tilde{\boldsymbol{Z}} = Q^{T}\boldsymbol{X}\boldsymbol{R}^{-1}$.

1.4 INTERPRETAÇÃO GEOMÉTRICA

Com o objetivo de ilustrar geometricamente os arquétipos, uma simulação proposta por Cutler e Breiman (1994), foi reproduzida da seguinte forma:

- Foi gerada uma amostra de tamanho 1000 de uma distribuição Normal bivariada com vetor de médias $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ e Matriz de correlação $\Sigma = \begin{bmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{bmatrix}$.
- Foram descartados todos os pontos com distância de Mahalanobis $\geq \chi^2_2(0{,}95)$, que representa uma elipse de confiança de 95%.
- Foram ajustados quatro arquétipos e armazenados. O processo foi repetido cem vezes.



(a) (b)

Figura 1 – (a) Elipse de confiança com 95% dos dados sorteados.
(b) Arquétipos selecionados em 100 iterações da simulação.
Fonte: Do autor.

Pode-se observar pelo resultado representado nas Figura 1 (a) e Figura 1 (b) que os arquétipos se agruparam no fim dos eixos. O que era esperado, pois os arquétipos são geralmente valores extremos.

Na Figura 1 (a) está representado uma iteração do processo do sorteio e descarte dos pontos fora da elipse de confiança de 95%. Pontos sólidos foram mantidos, enquanto pontos vazios foram descartados.

Observa-se na Fgura 1 (b) que para uma amostra normal bivariada com correlação, os arquétipos estão localizados na fronteira do fecho convexo.

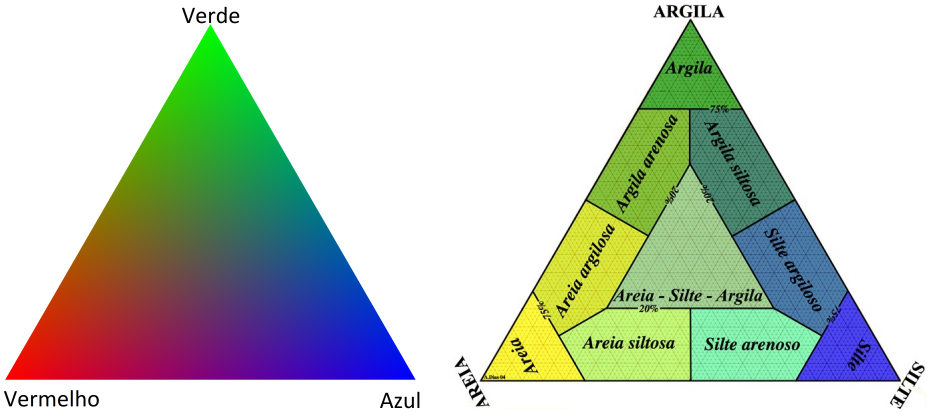
1.5 COMPREENSÃO DO RESULTADO DA ANÁLISE DE ARQUÉTIPOS

A Análise de Arquétipos permite reduzir a dimensionalidade dos dados utilizando combinações convexas dos elementos mais representativos, que geralmente são valores extremos (grandes ou pequenos). Por isso deixa de ser necessário conhecer todas as observações, bastando conhecer os mais representativos da base de dados, os arquétipos. Então, após realizar a Análise de Arquétipos, basta conhecer as matrizes de coeficientes ($\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$), lembrando que $\boldsymbol{\alpha}$ é a matriz que contém os coeficientes que descrevem a combinação de cada arquétipo ($\boldsymbol{z}_i$) para reconstrução dos dados originais ($\boldsymbol{X}$) com, geralmente, um pequeno erro associado, e $\boldsymbol{\beta}$ é a matriz que contém os coeficientes utilizados para construir a matriz dos arquétipos $\boldsymbol{Z}$, a partir dos dados originais. Isto permite uma nova abordagem na interpretação dos dados, pois observando os coeficientes $\boldsymbol{\alpha}$ é possível avaliar com quais arquétipos as observações são mais semelhantes, possibilitando assim a interpretação, comparação e até agrupamento dos dados.

Um bom exemplo da aplicação prática de arquétipos é quando utiliza-se o sistema vermelho verde azul, mais conhecido pela sigla em inglês para *red, green, blue*, RGB, que pode ser visto na Figura 2 (a). Nota-se que utilizando apenas três cores, é possível representar todas as outras cores do sistema. Ou seja, basta saber quais são as cores capazes de reescrever as outras, as cores "arquétipicas" das demais, e seus respectivos coeficientes na combinação da formação de uma determinada cor. Existe também o sistema Ciano (*Cyan*), Magenta (*Magenta*), Amarelo (*Yellow*) e Preto (*Black*), chamado de sistema CMYK de cores. Este ultimo é mais usado para impressões e na indústria gráfica. Vale notar que não existe apenas os sistemas RGB e CMYK para este tipo de problema, exemplo RYB (*red, yellow, blue* - vermelho, amarelo e azul), assim como não existe um número exato de combinações possíveis de arquétipos, pois várias combinações podem chegar a resultados similares.

Outro exemplo prático do uso de arquétipos pode ser visto na Figura 2 (b), o diagrama de

Shepard. Utilizando os elementos básicos (arquétipos) areia, silte e argila é possível descrever as possíveis composições da textura do solo (DIAS, 2004).



(a) (b)

Figura 2 – (a) Exemplo do sistema conhecido pela sigla em inglês para *red, green, blue*, RGB. (b) Diagrama de Shepard na textura do solo.

Fonte: (a) Adaptado de (GRAVE, 2014). (b) Adaptado de (DIAS, 2004).

1.6 CRONOLOGIA

O estado da arte da Análise de Arquétipos será resumido neste tópico em uma ordem cronológica de como diferentes autores adaptaram a ideia original proposta por Cutler e Breiman (1994).

Stone e Cutler (1997) estudaram o comportamento da técnica para analisar dados de estruturas móveis, como dados sobre propagação de ondas e sólitons da física. A técnica mostrou-se hábil em separar os pontos estacionários dos pontos que representam movimento, separando os dados em três estruturas: tempo, espaço e diferença de espaço. Porém os autores advertiram que está técnica é descritiva e existem ferramentas que utilizam fundamentos matemáticos mais elegantes, como a Análise de Componentes Principais.

A ideia da AA foi reformulada por D'Esposito, Palumbo e Ragozini (2006) que estenderam o método para utilizar dados intervalares ao invés de valores reais. Utilizaram simulações para comprovar que o método é tão bom quanto o original, pois mantém a mesma precisão e acurácia. Neste mesmo segmento de trabalho, mas com diferentes aplicações, pode-se citar também Corsaro e Marino (2010) e D'Esposito, Palumbo e Ragozini (2011), D'Esposito, Palumbo e Ragozini (2012).

Bauckhage e Thurau (2009) propuseram uma restrição que diminui consideravelmente o tempo de execução do algoritmo. A intensão é não incluir os pontos que estão fora da fronteira do fecho convexo dos dados para seleção dos arquétipos. Como os arquétipos são geralmente valores extremos, estes valores estão quase sempre na fronteira do fecho convexo. Desta forma o número de restrições do algoritmo diminui consideravelmente e a convergência é atingida de forma bem mais rápida, permitindo o uso de bases de dados bem mais extensas do que a abordagem original. O problema de encontrar o conjunto de arquétipos deixa de ser $O(n^2)$ e passa a ser $O(n'^2)$, em que n' é tão menor que n quanto o número de elementos fora da fronteira do fecho convexo.

Eugster e Leisch (2011) desenvolveram uma modificação que denominaram Arquétipos Ponderados e Arquétipos Robustos. Arquétipos Ponderados é a utilização de pesos dentro da Equação (1.2) em que a SQR é calculada, dando maior e menor importância a algumas observações de acordo com o contexto da pesquisa. Arquétipos Robustos é o termo utilizado para selecionar os arquétipos utilizando uma métrica visando não punir *outliers* tão severamente como a Norma Quadrática por exemplo, utilizando uma função com crescimento menor em paralelo com a métrica estabelecida.

Seth e Eugster (2014) cunham o termo Arquétipos Probabilísticos, que utilizam a AA observando que as matrizes $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ são matrizes estocásticas, e então os arquétipos passam a descrever os dados como probabilidades de serem gerados a partir de cada arquétipo. Os autores afirmam que os dados não são reduzidos a partir do espaço de observações, mas sim do espaço paramétrico dos arquétipos.

Sifa, Bauckhage e Drachen (2014a) utilizam a AA para um sistema de recomendação de jogos para jogadores, similares aos jogos que o jogador geralmente joga, com base no tempo jogado em horas e nas características dos jogos. A diferença de aplicação da AA convencional foi que os arquétipos foram selecionados de um subconjunto dos dados, os jogos que um certo jogador possui. E então os arquétipos selecionados deste conjunto foram utilizados para reconstruir o conjunto completo de todos os jogos da loja, de forma que os jogos que fossem melhor reconstruídos, eram recomendados para o jogador. Foram considerados quinhentos mil jogadores e três mil jogos. Esta abordagem obteve a maior taxa de aceitação em suas recomendações no estudo.

Vinuè, Epifanio e Alemany () propuseram e implementaram no pacote *Anthropometry* do R CORE TEAM (2014), uma mudança no algoritmo original da AA para encontrar apenas

arquétipos puros, ou seja, os arquétipos selecionados devem ser valores que foram realmente observados. O que foi uma solução elegante para o problema da AA selecionar arquétipos fora do espaço possível de observações. O método foi denominado de Arquetipóides e será descrito de forma detalhada na Seção (1.8).

1.7 APLICAÇÕES DA ANÁLISE DE ARQUÉTIPOS

A Análise de Arquétipos vem sendo utilizada em diversas áreas do conhecimento, como marketing, economia, aprendizado de máquinas, reconhecimento de padrões, astrofísica e análises esportivas. Alguns estudos são destacados nesta seção.

Stone e Cutler (1996) fizeram uma comparação entre a Análise de Arquétipos e a Análise de Componentes Principais, apontando características, vantagens e desvantagens de cada uma dessas técnicas.

Stone e Olson (1999) propuseram um método híbrido de Arquétipos e Componentes Principais para análise de sistemas dinâmicos, que foi testado com dados originados de espaços de dimensões acima de 500. Primeiro utilizaram ACP para reduzir a dimensão, e com o resultado obtido, aplicaram a AA. Os autores afirmaram que o processo conseguiu diferenciar situações de movimentação para situações de inércia. Este método foi chamado de Arquétipos Móveis.

Em astrofísica pode-se citar Chan, Mitchell e Cram (2003), que utilizaram a AA para estudar as formações de estrelas em galáxias e demonstraram que o método pode ser uma forma eficaz e eficiente para classificar os espectros das galáxias. Os autores mostraram que o método é robusto na presença de vários tipos de observações discrepantes. Neste trabalho os autores estudaram como a AA se comporta quando ruídos estão presentes nos dados, visto que a AA é altamente sensível a *outliers*. Foi utilizada AA em conjunto com ACP para uma melhor interpretação dos dados. Este foi um dos primeiros trabalhos a utilizar AA e ACP em conjunto.

Em *marketing*, D'Esposito, Palumbo e Ragozini (2006) utilizaram AA para segmentação de mercado ao invés da Análise de Agrupamentos pois esta técnica consiste em identificar em médias, e para o problema de segmentação de mercado espera-se conhecer os consumidores extremos e consumidores médios isolados.

Ainda no contexto de segmentação de mercado, Riedesel (2014) também aplicou a AA

ao invés da Análise de *Clusters* que é comumente utilizado para este fim, advertindo que a AA não foi utilizada como uma alternativa ao outro método, e sim para dar uma nova perspectiva de segmentação e heterogeneidade de consumidores. O autor destacou as vantagens de se conhecer os arquétipos puros principalmente para a área de comunicação e propaganda, enfatizando que o consumidor alvo deve "existir", não podendo ser uma combinação "teórica" de consumidores. Os autores Li et al. (2003) descreveram as vantagens em utilizar a AA para segmentação de mercado. E afirmaram que o fato da AA utilizar valores extremos facilita a interpretação dos resultados, possibilitando estudos para produção de novos produtos e renovação de produtos que já estão no mercado, bem como a mudança de contextos dos produtos.

Porzio, Ragozini e Vistocco (2008) utilizaram os arquétipos como pontos de referência (*benchmark*). Como exemplo utilizaram uma base de dados formada por variáveis sobre o desempenho das 200 melhores universidades do mundo.

D'Esposito, Palumbo e Ragozini (2011) expandiram o uso da AA para utilização em dados intervalares, a técnica foi aplicada em um experimento sensorial de queijos e concluíram que foi possível destacar as características sensoriais dos queijos analisados. D'Esposito, Palumbo e Ragozini (2012) utilizaram a mesma técnica para provar a utilidade em análises exploratórias, com um conjunto de dados sobre medidas de partes de morcegos, a fim de identificar as espécies. Como os dados já haviam sido estudados por outro autor utilizando ACP, foi possível fazer uma breve comparação entre os métodos e os autores concluíram que a AA foi capaz de identificar as espécies de morcegos que podem representar as outras.

Eugster (2011) sugere em seu artigo uma nova abordagem para analisar os dados sobre atletas esportivos, e exemplifica com dados sobre jogadores de basquete e de futebol. O autor destaca que como os arquétipos são geralmente valores extremos, pode-se identificar as características dos arquétipos e assim descobrir qual o "melhor" em determinado aspecto, e comparar os coeficientes α de cada observação (no caso jogadores). Assim, o jogador que tiver o maior coeficiente para o arquétipo, é possivelmente o melhor jogador deste aspecto.

Em Matemática, pode-se citar Costantini et al. (2012), que utilizaram a AA para encontrar funções básicas que melhor descrevessem as funções de uma base de dados sobre funções matemáticas.

Thogersen et al. (2013) aplicaram a AA na análise de dados de expressão de genes. Os resultados mostraram que AA foi promissora em agrupamentos de dados de grupos biologicamente significativos. Os arquétipos selecionados foram capazes de extrair as características

principais do conjunto de dados.

Seiler e Wohlrabe (2013) aplicaram a AA para encontrar arquétipos de cientistas da área de economia em uma base de dados com quase trinta mil economistas. As variáveis entre outras eram: número de artigos (ponderados pela qualidade dos artigos), número de citações, downloads dos trabalhos etc. Concluíram que a AA foi capaz de identificar os arquétipos dos cientistas, e em seguida os autores analisaram cada arquétipo de forma individual.

Há também aplicações em Inteligência artificial e aprendizado de máquinas, como Sifa e Bauckhage (2013), que utilizaram a AA em dados sobre a movimentação de jogadores profissionais de um jogo de tiro em primeira pessoa, para utilizar o mesmo padrão de movimentação em um *Bot*, que é um jogador controlado pelo computador. Os resultados mostraram que o novo *Bot* foi capaz de movimentar-se como um humano e com custo computacional mais baixo que outras técnicas anteriores.

Sifa, Bauckhage e Drachen (2014b) ainda no contexto de jogos eletrônicos, apresentam um modelo utilizando a AA para descrever o interesse de jogadores por jogos eletrônicos em função do tipo de jogo e do tempo jogado. Os resultados mostram que a maioria dos jogadores, geralmente, perdem o interesse em um jogo após 10 horas e pouquíssimos jogadores jogam o mesmo jogo por mais de 35 horas.

Morup e Hansen (2012) mostraram a utilidade da AA principalmente para *Data Mining* e Aprendizado de Máquinas ao utilizarem cinco aplicações, que são: Visão Computacional, *NeuroImaging*, Química, Mineração de Texto e *Collaborative Filtering* (sugerir produtos a clientes com base em suas preferencias).

Epifanio, Vinuè e Alemany (2013) utilizaram a AA em um conjunto de dados antropométricos sobre tamanhos de roupas, e evidenciaram as vantagens em utilizar AA ao invés de ACP para este tipo de análise. Os autores propuseram como trabalho futuro a mesma aplicação porém ao invés de dados sobre o tamanho de roupas, utilizar medidas 3D das mesmas.

Bauckhage (2014) fez uma revisão sobre a AA e estudou o uso de heurísticas eficientes para a aproximação dos dados aos arquétipos, além de ter apresentado propriedades geométricas dos resultados obtidos.

1.8 A ANÁLISE DE ARQUETIPÓIDES

Nesta seção é feita a apresentação de uma proposta recente de um trabalho com o tema Arquétipos, em que os autores sugerem somente o uso de arquétipos puros.

Vinuè, Epifanio e Alemany () introduziram um novo conceito sobre a Análise de Arquétipos que visa representar cada objeto do conjunto de dados como apenas misturas de observações reais. Estes dados que serão utilizados para realizar tal mistura são denominados arquetipóides. Diferente da AA convencional, a Análise de Arquetipóides sempre encontra valores reais observados, e não combinações das observações. A importância deste método é reforçada quando os arquétipos são necessariamente reais, e não permitem arquétipos teóricos (combinações de observações para construir um arquétipo). Os autores propõe um algoritmo capaz de encontrar de maneira eficiente tal resultado, e fazem uma comparação com a abordagem original. O método foi exemplificado utilizando dados sobre jogadores de basquete, formas de *cockpits* de aeronaves e tamanhos de roupas femininas.

1.9 ARQUÉTIPOS *VERSUS* ARQUETIPÓIDES

Num contexto geral, a AA busca reescrever cada observação do conjunto de dados, como uma combinação convexa dos arquétipos que são os dados mais extremos e característicos de um conjunto de dados, mas não necessariamente valores observados. Na prática nem sempre isso é possível ou faz sentido para a interpretabilidade. Por exemplo, ao se analisar dados sobre jogadores profissionais de futebol cujas variáveis em estudo são velocidade, potência do chute, força física, peso, altura etc., informar que os arquétipos destes dados são combinações de vários jogadores, não faz sentido prático pois implicaria em ter que encontrar um jogador com exatamente as mesmas características combinadas desses jogadores, o que seria impossível. Assim, nota-se neste caso a necessidade de utilizar somente arquétipos puros, ou seja, a Análise de Arquetipóides.

Porém, a grande desvantagem da Análise de Arquetipóides é observada quando a reconstrução dos dados originais se faz necessária, pois espera-se que a SQR do resultado seja no mínimo igual mas geralmente será maior que ao utilizar arquétipos teóricos, que são menos restritos. Como o método utiliza o critério da SQR para selecionar os arquétipos, neste caso

deve-se abster um pouco da qualidade de ajuste dos dados reconstruídos por uma melhor interpretação. Com base nessas informações conclui-se que se o domínio do problema permitir, deve-se utilizar a AA comum, e caso não faça sentido a possibilidade de encontrar arquétipos teóricos, deve-se utilizar os arquetipóides.

1.10 MÉTRICAS

Para calcular a SQR, diversos autores discordam sobre qual métrica deve ser utilizada, como por exemplo pode-se citar Eugster e Leisch (2009), que utilizaram a Norma Espectral (NE), D'Esposito, Palumbo e Ragozini (2012) que utilizaram a Norma de Frobenius (NF) e Cutler e Breiman (1994) que utilizaram a Norma Quadrática, como pode ser visto na Equação 1.4.

Assim, esta trabalho também tem como objetivo, utilizando estudos de simulação Monte Carlo, avaliar a qualidade de ajuste da análise de arquétipos com as diferentes métricas citadas na literatura: Norma quadrática, Norma de Frobenius e Norma Espectral. Objetivou-se também, propor a utilização de outras duas métricas: Determinante (DET) e Soma de Quadrados e Produtos de Resíduos (SQPR), pois também são medidas sumarizantes de matrizes, como variâncias generalizadas.

1.11 FUTEBOL

Segundo Souza (2013), o futebol, como esporte moderno, foi criado na Inglaterra em 1863. Entretanto, há indícios que esta prática esportiva já era exercida em outras localidades do mundo, em formatos similares. O *Tsu-Chu*, que traduzindo significa "lançar com o pé" (*tsu*) uma "bola de couro" (*chu*), foi criado na China Antiga, por volta de 3000 a.C, para fins de treinamento militar. A bola era chutada pelos soldados chineses por entre duas traves cravadas no chão. Em contrapartida, o *Kemari* foi trazido da China ao Japão sendo um jogo mais cerimonial, não tinha um vencedor ou um perdedor. O jogo começava com 6 a 8 jogadores que formavam uma roda passando a bola um para o outro com as mãos.

Jogos semelhantes que utilizam um campo retangular, dividido por linhas e equipes,

destacaram-se em locais como Roma, Grécia, Ilhas Britânicas (*Rúgby*) e durante a Época dos Descobrimentos, a Cultura Maia apresentava o *Pok ta pok*, que se assemelha ao futebol. Posteriormente, com as devidas alterações, surge o futebol como se vê hoje. Assim como no decorrer dos tempos foram também criados inúmeros clubes, times, associações, grupos desportivos, campeonatos, torneios, etc.

De acordo com o SUAPESQUISA (2014), Charles Miller é considerado o precursor do futebol no Brasil. Nascido no bairro paulistano do Brás, viajou para Inglaterra aos nove anos de idade para estudar. Em 1894, ao retornar ao Brasil, trouxe na bagagem a primeira bola de futebol e um conjunto de regras. O primeiro jogo de futebol no Brasil foi realizado em 15 de abril de 1895 entre funcionários de empresas inglesas que atuavam em São Paulo.

No ano de 1904 foi criada a FIFA (*Fédération Internationale de Football Association*), órgão que rege o futebol ao redor do mundo e tem como sua principal competição internacional a Copa do Mundo. Suas regras, num número total de dezessete, não sofrem alterações independente da localidade, cultura ou crença. O jogo se estabelece na divisão do campo (gramado), em linhas que fundamentarão suas normas e as equipes adversárias devem conter no máximo onze e no mínimo sete jogadores, durante uma partida de noventa minutos, dividida em dois tempos iguais podendo ser acrescido o segundo tempo de períodos extras a depender da fase do campeonato e alterações realizadas durante a partida (SOUZA, 2013).

Em jogos de primeira divisão profissional e onde forem feitos os principais jogos internacionais e nacionais, os campos devem medir 105 m de comprimento por 68 m de largura. Estas dimensões são obrigatórias para a Copa do Mundo e para as competições finais nos campeonatos de confederações. Estas dimensões são definidas por corresponderem as maiores possíveis para um campo que se situe no interior de uma pista oficial de atletismo. Os projetistas frequentemente sofrem pressão para aumentar o tamanho do campo ou incluir pistas de corridas nos estádios. Às vezes, tais exigências são inevitáveis, entretanto, estas instalações não serão tão boas quanto um estádio construído especialmente em função das dimensões de um campo de futebol (FIFA, 2011).

1.12 FUTEBOL E ESTATÍSTICA

O esporte mais popular do planeta, o futebol, percebeu que a intuição e o improviso não são suficientes para determinar suas estratégias. Os times estão usando análise estatística para comprar jogadores, escalar equipes e substituir atletas durante as partidas. Como exemplo, pode-se citar a empresa de estatísticas de futebol, a Footstats (2014), fundada em 2004 e que acompanha mais de 50 campeonatos de futebol por ano, incluindo os latino americanos e os europeus. A empresa oferece um banco de dados bastante amplo com estatísticas de times e jogadores, controle de tabelas e placares eficazes para que os usuários estejam por dentro de todas as novidades do time. Além de disponibilizar, conteúdo detalhado com dados estatísticos, tabelas e acompanhamento de resultados de vários esportes tais como, Artes Marciais Mistas, mais conhecidas pela sigla MMA (do inglês: *Mixed Martial Arts*), Automobilismo, Basquete e Vôlei.

Além disso, possui um sistema, *Client*, que concentra toda a cobertura de campeonatos regionais e nacionais. OS técnicos e os clubes de futebol têm acesso a diversos tipos de relatórios de partidas, separados por performance coletiva e individual, que trazem resultados importantes para a tomada de decisões no momento dos jogos. Com essa base de dados acumulada ao longo dos anos, pode-se traçar o perfil dos jogadores que interessam aos clubes. Assim, cada vez mais o futebol se torna um esporte com métodos amparados nas estatísticas, e pode-se, por exemplo, contratar jogadores a partir de seus índices de desempenho. Ainda, de acordo com o *Footstats*, constatou-se que durante a Copa de 2014, a Seleção Brasileira foi o time que mais desarmou de forma correta na competição e a Alemanha foi a equipe que mais tocava a bola. Uma possível explicação para estes fatos é o esquema tático utilizado por essas equipes, como por exemplo o 4-4-2, utilizado pela Seleção Brasileira, representado na Figura 3.

Há diversos trabalhos na literatura que por meio de procedimentos matemáticos ou estatísticos, produzem valores com o intuito de ordenar equipes de uma determinada modalidade de acordo com rankings, que são classificações de equipes e consideram seus resultados históricos independentes de sua situação atual, e também prever resultados. Dentre estes, pode-se citar:

Arruda (2000) abordou o problema de previsões probabilísticas para eventos tricotômicos, além de comparar a qualidade das previsões por meio das curvas de calibração e da medida de *DeFinetti* fazendo uma aplicação para previsão de resultados no futebol. Os métodos utilizados no trabalho, apresentaram bom desempenho no que se refere a verificação da qualidade

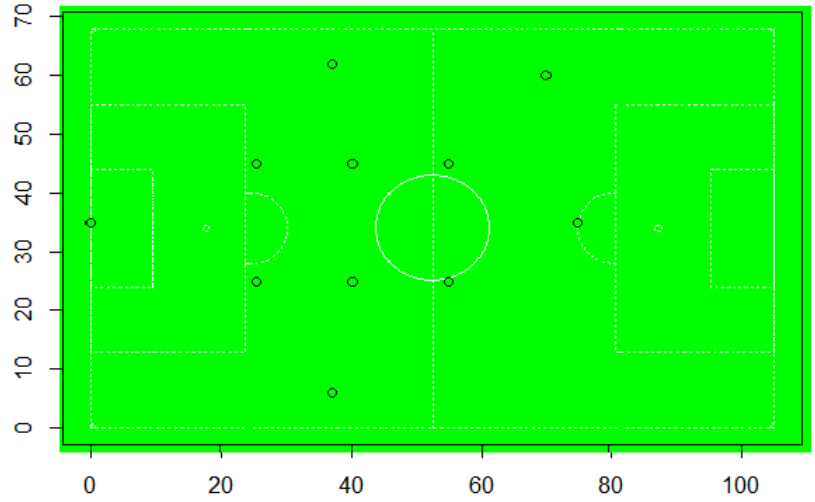


Figura 3 – Esquema tático semelhante ao utilizado pela Seleção Brasileira na Copa do Mundo FIFA de 2014 em que o lado esquerdo do campo é o defensivo e o lado direito é o ofensivo.
Fonte: Do autor.

de previsões probabilísticas.

Santos (2013) propôs um modelo de espaço de estados *Poisson* com verossimilhança exata para a modelagem dos confrontos entre Brasil e Argentina em partidas de futebol, de fácil implementação. O modelo permite o cálculo da função de verossimilhança exata, bem como das distribuições preditivas e distribuições suavizadas e de filtragem da variável latente. Com a metodologia apresentada, foi possível responder a todas as perguntas de interesse que foram levantadas a *priori*.

Nos esportes, a AA é comumente utilizada para dados multivariados de atletas. A ideia básica é a de olhar para um grande número de variáveis para cada jogador, selecionando-se os atletas com performances extremas ou as características multivariadas de cada jogador sendo reescritas como combinações de outros jogadores. Estes extremos são os arquétipos dos atletas. Os indicadores de desempenho e, em sua base, perfis de desempenho são um dos fundamentos da análise de desempenho no esporte. O ponto crucial é o desenvolvimento de perfis de desempenho que permitam avaliar os assuntos de interesse com precisão. Tais perfis não são baseados em performances médias, mas em performances geralmente extremas (atletas com desempenho muito acima ou muito abaixo da média).

Uma aplicação da análise de arquétipos foi feita por Eugster (2012) utilizando estatísticas de basquete e avaliações de habilidades no futebol para interpretar os atletas arquétipos e as suas características, além da composição de todos os atletas.

1.13 ARQUÉTIPOS E COMPONENTES PRINCIPAIS CONCENTRANDO INFORMAÇÃO SENSORIAL

Na prática experimentos coletam várias variáveis que se estudadas de forma simultânea, levam a resultados mais otimizados do que quando são analisadas de forma individual ou univariada. Porém, é comum experimentos multivariados serem analisados de forma univariada pela complexidade das análises envolvidas. Existem diversos métodos estatísticos para analisar, descrever e inferir sobre experimentos mensurados em mais de uma variável. Dentre estes métodos, destaca-se a Análise de Componentes Principais.

A Análise de Componentes Principais é uma técnica multivariada que tem por finalidade básica a análise dos dados visando reduzir sua dimensionalidade. É amplamente utilizada em diversas áreas do conheciment. Tem como objetivo principal a modelagem da estrutura de variância e covariância de um vetor aleatório composto de p–variáveis aleatórias denominadas variáveis latentes, por meio da construção de combinações lineares das variáveis originais que são denominadas componentes principais.

A Análise de Arquétipos também é uma técnica multivariada que busca reduzir a dimensão de dados por meio de combinações convexas dos arquétipos, que são geralmente valores extremos dos dados. Os arquétipos são selecionados pela minimização da soma de quadrados de resíduos, que é o erro cometido ao se reconstruir os dados originais utilizando as combinações convexas dos arquétipos. A Análise de Arquétipos não é uma técnica difundida como a Análise de Componentes Principais e é uma técnica relativamente nova e pouco estudada. Não foram encontrados trabalhos em português que trata sobre o tema.

O objetivo deste estudo é verificar por meio de simulação e de dados reais, a capacidade das técnicas concentrarem a informação trazida por dados sensoriais sem que haja prejuízo para a interpretação dos resultados.

1.14 SENSOMETRIA

A Sensometria é o conjunto de métodos estatísticos aplicados à análise sensorial. A Análise Sensorial é uma disciplina científica usada para evocar, medir, analisar e interpretar reações das características dos alimentos e materiais como são percebidas pelos sentidos da

visão, olfato, gosto, tato e audição (ABNT, 2014).

A Análise Sensorial pode ser realizada por potenciais consumidores de um produto, ou por uma equipe de juízes treinados e com sentidos aguçados para analisar e detectar as características sensoriais de um produto para um determinado fim. Esta equipe é denominada painel. Uma das características mais desejadas de um painel bem treinado é a repetibilidade, que é a capacidade de reproduzir fielmente dados de experimentos anteriores. Outra característica também intimamente ligada com a qualidade do painel é a concordância entre os juízes.

A ferramenta de medida da Análise Sensorial é composta pelos sistemas sensoriais humano: olfativo, gustativo, tátil, auditivo e visual. Estes sistemas avaliam as características e atributos dos produtos. Esta avaliação está intimamente associada com a aceitação ou rejeição de um determinado produto no mercado, bem como a fidelidade de consumidores para com o produto. A Análise Sensorial também é amplamente utilizada em estudo de vida de prateleira (*shelf life*), na identificação de diferenças e similaridades entre produtos concorrentes, determinação das preferências dos consumidores por um determinado produto, para a otimização e melhoria da qualidade (FERREIRA; OLIVEIRA, 2007).

Atualmente, os segmentos que mais utilizam a análise sensorial para avaliar o lançamento de novos produtos ou mesmo mudanças em produtos que já estão no mercado são as indústrias: alimentícia, automobilística, cosméticos e telefonia móvel (FERREIRA; OLIVEIRA, 2007).

1.15 ANÁLISE DE DADOS MULTIVARIADOS

Em experimentos de quase todas as áreas de pesquisas, várias variáveis são mensuradas e, em geral, essas devem ser analisadas simultaneamente. Quando um pesquisador deseja analisar seu experimento, dificilmente o foco é somente em uma variável. Comumente, as variáveis em estudo possuem relações entre si, que, se forem exploradas de forma conjunta, conduzirão a análises mais robustas e informativas (FERREIRA, 2011).

Diante desse fato pode-se notar a importância do estudo da análise de dados multivariados. Existem diversos métodos estatísticos para analisar, descrever e inferir sobre experimentos de múltiplas variáveis. Dentre estes métodos, destaca-se a Análise de Componentes Principais (ACP). A ACP é uma técnica de redução da dimensionalidade de dados, bastante difundida e

aplicada em diversas áreas do conhecimento.

Outro método que também tem como propósito reduzir a dimensionalidade dos dados para, entre outras vantagens, facilitar sua interpretação é a Análise de Arquétipos (AA), proposta por Cutler e Breiman (1994). Este método tem sido utilizado em diferentes áreas do conhecimento e apesar de não ser tão difundido quanto a ACP, há um aumento considerável de trabalhos relacionados à esta técnica nos últimos tempos. Os métodos ACP e AA serão detalhados nas seções a seguir.

1.16 ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (ACP) é uma técnica multivariada, introduzida por Pearson (1901), consolidada por Hotelling (1933) e tem por finalidade básica a análise dos dados visando reduzir sua dimensionalidade. Seu objetivo principal é a modelagem da estrutura de variância e covariância de um vetor aleatório composto de p–variáveis aleatórias chamadas variáveis latentes, por meio da construção de combinações lineares das variáveis originais. Essas combinações lineares são denominadas componentes principais, são ortogonais e não correlacionadas entre si. Por meio dessa análise, a informação contida nas p–variáveis originais é substituída pela informação contida em $q\,(q \leq p)$ componentes principais mas em geral, deseja-se obter uma redução do número de variáveis a serem avaliadas e interpretadas $(q < p)$.

Assim, pode-se obter a redução do número de variáveis a serem avaliadas, bem como a interpretação das combinações lineares construídas, em que a informação contida nas variáveis originais é de fato substituída pela informação contida em um número reduzido de componentes principais. Entretanto a qualidade dessa aproximação depende do número de componentes utilizados e pode ser mensurada, por exemplo, por meio da avaliação da proporção da variância total explicada pelos mesmos, uma vez que a variância dos dados é o principal critério a ser maximizado na ACP. Quando a distribuição de probabilidades do vetor aleatório em estudo é normal, os componentes principais são independentes e têm distribuição normal. Entretanto, a normalidade não é requisito necessário para que a técnica possa ser utilizada (MINGOTI, 2007).

Os componentes principais são encontrados por meio da diagonalização de matrizes simétricas positivas semi-definidas que facilita os cálculos principalmente pela quantidade de

programas de computadores capazes de realizar cálculos matriciais. Pesquisadores têm utilizado esta técnica para resolver problemas diversos, como: multicolinearidade em regressão, estimação de fatores, modelagem de interação de fatores em experimentos sem repetição, estudo de divergências, agrupamento entre genótipos em estudos de genética e melhoramento de plantas e animais, entre outras aplicações (FERREIRA, 2011).

Diversos critérios foram propostos sobre a determinação de uma dimensionalidade mínima admissível para representar os dados originais, ou seja, a identificação do número q de componentes principais a se reter de forma que uma aproximação q-dimensional dos dados possa ser considerada aceitável. Mais detalhes encontram-se em (JOLLIFFE, 2002).

1.16.1 Componentes principais exatos extraídos da matriz de covariâncias

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^{'}$ um vetor aleatório com vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)^{'}$ e matriz de covariâncias $\boldsymbol{\Sigma}_{p\times p}$ com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ e respectivos autovetores $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\ldots$, $\boldsymbol{e}_p$. Assim o i-ésimo componente principal é dado por

$$Y_i = \boldsymbol{e}_i^{'}\boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \ldots + e_{ip}X_p \quad i = 1, 2, \ldots, p \tag{1.8}$$

Para encontrar as variáveis latentes Y_i, é encontrar um plano que maximize a distância entre os pontos originais, que é equivalente a maximizar a variabilidade. A variância de Y_i é dada por

$$Var(Y_i) = Var(\boldsymbol{e}_i^{'}\boldsymbol{X}) = \boldsymbol{e}_i^{'}Var(\boldsymbol{X})\boldsymbol{e}_i = \boldsymbol{e}_i^{'}\boldsymbol{\Sigma}\boldsymbol{e}_i \tag{1.9}$$

e a covariância entre Y_i e Y_k por

$$Cov(Y_i, Y_k) = \boldsymbol{e}_i^{'}Var(\boldsymbol{X})\boldsymbol{e}_k = \boldsymbol{e}_i^{'}\boldsymbol{\Sigma}\boldsymbol{e}_k \text{ para } i \neq k \tag{1.10}$$

A variância máxima do componente principal acompanha o crescimento do vetor de coeficientes $\boldsymbol{e}_i$, logo a variância tenderá ao infinito. Por isso restringe-se $\boldsymbol{e}_i^{'}\boldsymbol{e}_i = 1$, para maximizar a variância na Equação (1.9) (FERREIRA, 2011).

Utilizando a maximação de formas quadráticas, tem-se que (JOHNSON; WICHERN,

2007, p. 80)

$$\lambda_i = \max_{e_i} \frac{\boldsymbol{e}_i^{'} \boldsymbol{\Sigma} \boldsymbol{e}_i}{\boldsymbol{e}_i^{'} \boldsymbol{e}_i} \tag{1.11}$$

Derivando-se (1.11) em relação a $\boldsymbol{e}_i$ e igualando a zero, obtém-se o máximo a partir da resolução da equação (1.12)

$$\left(\boldsymbol{\Sigma} - \lambda_i \boldsymbol{I}\right) \boldsymbol{e}_i = \mathbf{0} \Rightarrow \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i \tag{1.12}$$

Tem-se de (1.9) e (1.10) que a variância e a covariância de Y_i são dadas por

$$Var(Y_i) = \boldsymbol{e}_i^{'} \boldsymbol{\Sigma} \boldsymbol{e}_i = \boldsymbol{e}_i^{'} \lambda_i \boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i^{'} \boldsymbol{e}_i = \lambda_i \tag{1.13}$$

e

$$Cov(Y_i, Y_k) = \boldsymbol{e}_i^{'} \boldsymbol{\Sigma} \boldsymbol{e}_k = \boldsymbol{e}_i^{'} \lambda_k \boldsymbol{e}_k = \lambda_k \boldsymbol{e}_i^{'} \boldsymbol{e}_k = 0 \; i \neq k \tag{1.14}$$

uma vez que $\boldsymbol{e}_i$ e $\boldsymbol{e}_k$ são ortogonais, ou seja, $\boldsymbol{e}_i^{'} \boldsymbol{e}_k = 0$

Cada autovalor λ_i representa a variância de um componente principal Y_i. Como os autovalores estão ordenados em ordem decrescente, o primeiro componente tem maior variabilidade e o p-ésimo a menor.

Pelo teorema da decomposição espectral de $\boldsymbol{\Sigma}$, descrita por $\boldsymbol{\Sigma} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}^{'}$, em que $\boldsymbol{P}$ é a matriz composta pelos respectivos autovetores de $\boldsymbol{\Sigma}$ e $\boldsymbol{\Lambda}$ é a matriz de autovalores de $\boldsymbol{\Sigma}$ observa-se que (JOHNSON; WICHERN, 2007, p. 65)

$$tr(\boldsymbol{\Sigma}) = tr(P \boldsymbol{\Lambda} \boldsymbol{P}^{'}) = tr(\boldsymbol{\Lambda} \boldsymbol{P}^{'} \boldsymbol{P}) = tr(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} Var(Y_i) = \sum_{i=i}^{p} \lambda_i \tag{1.15}$$

Sabe-se ainda que

$$tr(\boldsymbol{\Sigma}) = \sum_{i=i}^{p} \sigma_{ii} = \sum_{i=i}^{p} \lambda_i \tag{1.16}$$

De onde conclui-se que a variância total do vetor aleatório $\boldsymbol{X}$ pode ser descrita pela variância total do vetor aleatório $\boldsymbol{Y}$

A proporção da variância total de $\boldsymbol{X}$ que é explicada pelo i-ésimo componente principal é dada por

$$P_i = \frac{Var(Y_i)}{tr(\boldsymbol{\Sigma})} = \frac{\lambda_i}{\sum_{i=i}^{p} \lambda_i} \tag{1.17}$$

É comum padronizar os dados antes de executar a ACP. Geralmente as variáveis são centralizadas em suas médias e divididas pelas respectivas variâncias. Um dos métodos para encontrar os componentes principais é decompor a matriz $\boldsymbol{X}$ geralmente padronizada, no produto de duas matrizes distintas, $\boldsymbol{T}$ e $\boldsymbol{L}$, da forma

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{L}^{'}$$

em que $\boldsymbol{T}$ é a matriz denominada *scores* e $\boldsymbol{L}$ é a matriz denominada *loadings*. A matriz de *scores*, $\boldsymbol{T}$, traz a projeção dos pontos definidos pelas linhas de $\boldsymbol{X}$ no novo espaço definido pelos eixos do componente principal enquanto que a matriz de *loadings*, $\boldsymbol{L}$, descreve os pesos de cada variável original para a construção dos componentes principais. A representação gráfica da matriz de *scores* $\boldsymbol{T}$ e da matriz de *loadings* $\boldsymbol{L}$ também são chamadas de espaço de observações e espaço de variáveis, respectivamente.

A decomposição em valores singulares pode ser utilizada para encontrar as matrizes $\boldsymbol{T}$ e $\boldsymbol{L}$, pois pode-se reescrever uma matriz $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ no produto de três matrizes, $\boldsymbol{U}, \boldsymbol{S}$ e $\boldsymbol{V}$ da forma

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{'}$$

em que $\boldsymbol{U} \in \mathbb{R}^{n \times n}$, $\boldsymbol{S} \in \mathbb{R}^{n \times k}$ e $\boldsymbol{V} \in \mathbb{R}^{k \times k}$

As matrizes $\boldsymbol{U}$ e $\boldsymbol{V}$ são ortogonais. Suas colunas trazem respectivamente, os autovetores dos produtos $\boldsymbol{X}\boldsymbol{X}^{'}$ e $\boldsymbol{X}^{'}\boldsymbol{X}$ que são proporcionais às matrizes de covariâncias dos objetos e das variáveis, respectivamente. $\boldsymbol{S}$ é diagonal e traz os valores singulares destas matrizes, ordenados em ordem decrescente, e estes são relacionados com os autovalores ($\boldsymbol{\lambda}$) da matriz $\boldsymbol{X}$. Utilizando deste artificio, pode-se obter as matrizes $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{S}$ e $\boldsymbol{L} = \boldsymbol{V}$. Um componente principal Y_i é então definido pelo produto vetorial de uma dada coluna de $\boldsymbol{T}$ pela coluna de $\boldsymbol{L}$ correspondente

$$Y = \sum_{i=1}^{min(n,q)} \boldsymbol{T}_i \boldsymbol{L}_i^{'}$$

A redução da dimensionalidade ocorre quando se usa as matrizes $\boldsymbol{T}$ e $\boldsymbol{L}$ truncadas na q-ésima coluna com $q < p$. Obtém-se assim uma aproximação dos dados originais, pois um pouco da informação contida nas variáveis originais é perdida ao substituí-las pelos componentes principais (PEDRO, 2009).

1.17 COMPONENTES PRINCIPAIS *VERSUS* ARQUÉTIPOS

Os Componentes Principais são ordenados de forma que o primeiro é o que explica a maior parte da variância, e o segundo é o próximo que mais explica a variância dos dados até o p-ésimo que é o que menos explica a variação dos dados. Já na AA não há diferença entre os arquétipos selecionados, e não é possível afirmar se algum arquétipo é mais importante que outro, pois não há ordem de importância. De acordo com Cutler e Breiman (1994), os arquétipos não se aninham, ou seja, ao selecionar três arquétipos e posteriormente quatro arquétipos não é possível garantir que entre os quatro arquétipos haverá os três ou algum dos três selecionados anteriormente. E ainda há uma natureza aleatória na escolha dos arquétipos, por isso inclusive é indicado executar a análise de arquétipos mais de uma vez, para observar o comportamento dos arquétipos selecionados múltiplas vezes até que seja possível notar um certo padrão.

Os arquétipos não são ortogonais uns aos outros, e também não indicam a direção de maior variação dos dados como os componentes principais. Além disso o método da AA visa minimizar a SQR a cada iteração para selecionar os arquétipos, enquanto o método da ACP prioriza a maximização da variância para rotacionar os novos eixos no sentido de maior variância dos dados.

Métodos como a ACP que lidam com a representação de dados comprimidos, tem a interpretação dos dados tediosa. O benefício principal da Análise de Arquétipos é que os arquétipos são conotações dos próprios elementos dos dados, por isso levam a uma fácil interpretação (PRABHAKARAN et al., 2012).

2 METODOLOGIA

Nesta seção serão apresentadas as metodologias referentes a cada um dos artigos, que são eles: “Avaliação Monte Carlo de métricas para falta de ajuste em Análise de Arquétipos”, será chamado apenas de “Métricas” (MARTINS JÚNIOR et al., 2014). “Análise de Arquétipos na avaliação da movimentação de jogadores de futebol”, será chamado de “Futebol” (MARTINS JÚNIOR et al., 2015a). O artigo provisoriamente denominado “Arquétipos e componentes principais concentrando a informação sensorial”, que contempla a simulação proposta nesta dissertação e sua metodologia também é explicada nesta seção.

2.1 MÉTRICAS

Para o cálculo das faltas de ajuste utilizando as diferentes métricas e avaliação do desempenho dos diferentes métodos, desenvolveu-se uma rotina de simulação utilizando o software R (R CORE TEAM, 2014).

Uma matriz $\mathbf{M}$ foi composta por dois vetores linearmente independentes ($\mathbf{a_1}$ e $\mathbf{a_2}$) e por quatro vetores ($\mathbf{a_3}, \mathbf{a_4}, \mathbf{a_5}$ e $\mathbf{a_6}$) definidos como uma combinação linear dos dois primeiros vetores, com coeficientes p, q, r e s, respectivamente, pertencentes ao intervalo (0,1) e com a restrição de que o somatório fosse igual a um.

Os vetores $\mathbf{a_1}$ e $\mathbf{a_2}$ foram fixados na forma $\mathbf{a_1} = (1, 2, 3)$ e $\mathbf{a_2} = (7, 7, 8)$. Assumindo os coeficientes $p = 0{,}4$; $q = 0{,}1$; $r = 0{,}9$ e $s = 0{,}5$; foram obtidos os vetores $\mathbf{a_3} = (0{,}4 \times \mathbf{a_1} + 0{,}6 \times \mathbf{a_2})$, $\mathbf{a_4} = (0{,}1 \times \mathbf{a_1} + 0{,}9 \times \mathbf{a_2})$, $\mathbf{a_5} = (0{,}9 \times \mathbf{a_1} + 0{,}1 \times \mathbf{a_2})$ e $\mathbf{a_6} = (0{,}5 \times \mathbf{a_1} + 0{,}5 \times \mathbf{a_2})$, resultando em $\mathbf{M} = (\mathbf{a_1}\ \mathbf{a_2}\ \ldots\ \mathbf{a_6})'$.

Os dados foram simulados de uma distribuição normal tri-variada com vetor de médias nulo (sem perda de generalidade) e matriz de covariâncias equicorrelacionada, dada por

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \tag{2.1}$$

em que as variâncias (σ^2) foram fixadas em $0{,}1$ (pequena); $0{,}5$ (média); $1{,}0$ (grande) , as corre-

lações (ρ) fixadas em $0; 0{,}25; 0{,}5; 0{,}75$ e $0{,}95$. As diferentes variâncias, representam diferentes graus de perturbação no ajuste. Por outro lado, estudar covariâncias crescentes tem como objetivo verificar se as métricas são influenciadas por multicolinearidade. O número de repetições Monte Carlo foi fixado em 1000, totalizando 15000 cenários simulados.

Depois de agregado o erro/perturbação nos vetores, o algoritmo capaz de encontrar os arquétipos foi executado utilizando a função `archetypes()` do pacote de mesmo nome do software R (R CORE TEAM, 2014). Uma das informações que é recuperada após a execução do algoritmo é a matriz de resíduos ($\mathbf{E}$), que nada mais é que a diferença entre os dados originais e os dados reconstruídos pelos arquétipos, ou seja, $\mathbf{E} = \mathbf{X} - \mathbf{XBA}$. Então é aplicada uma função sumarizante nessa matriz de resíduos. As métricas estudadas neste trabalho são apresentadas na Tabela 1.

Tabela 1 – Métricas utilizadas no cálculo da falta de ajuste da análise de Arquétipos

Métricas	Notação	
Soma de Quadrados de Resíduos	$\|\|.\|\|^2$	$tr(\mathbf{EE}^{'})$
Norma de Frobenius	$\|\|.\|\|_F$	$\sqrt{tr(\mathbf{EE}^{'})}$
Norma Espectral	$\|\|.\|\|_2$	$\sqrt{\lambda_{max}(\mathbf{EE}^{'})}$
Soma de Quadrados e Produtos de Resíduos	$\|\|.\|\|^{QP}$	$\mathbf{1}^{'}\text{abs}(\mathbf{EE}^{'})\mathbf{1}$
Determinante	$\|.\|$	$\|\mathbf{EE}^{'}\|$

Fonte: Do autor.
Em que: $tr(.)$ é o traço, $\lambda_{max}(.)$ é o maior autovalor e $abs(.)$ é o valor absoluto dos elementos de uma matriz.

2.2 FUTEBOL

Foram simulados dados que representassem as posições ocupadas por jogadores em um campo de futebol com comprimento (X) e largura (Y) contínuos, limitadas entre 0 e 105 m para X, e 0 a 68 m para Y de acordo com o padrão da FIFA.

Para a simulação dos dados foi utilizado uma distribuição normal bivariada, com vetor de médias e matriz de covariâncias pertinente a cada posição do jogador simulado (atacante, zagueiro, lateral etc), descritas na Tabela 2. Os valores utilizados na matriz de covariâncias

foram definidos de forma arbitrária, com base na área do campo que espera-se que o jogador mais atue de acordo com sua posição e característica. As covariâncias foram fixadas em 0, pois em caso contrário, os dados sorteados estariam dispostos em uma diagonal, o que nem sempre ocorre com dados reais. Foram sorteadas amostras independentes de tamanho 200 de uma normal bivariada truncada dentro dos limites do campo para cada jogador analisado, sendo retidos 45 pontos representando cada minuto de um tempo normal de um jogo sem os acréscimos. Portanto não há dependência entre os pontos que representam as posições ocupadas em algum momento do jogo por um jogador.

Tabela 2 – Valores utilizados para o vetor de médias e a matriz de covariâncias definidos com base na área do campo e de acordo com a posição e característica que o jogador mais atua.

Posição	Centro (X, Y)	Covariâncias
Zagueiro 1	30, 45	$\begin{pmatrix} 450 & 0 \\ 0 & 75 \end{pmatrix}$
Zagueiro 2	30, 25	
Lateral 1	42, 6	$\begin{pmatrix} 1000 & 0 \\ 0 & 15 \end{pmatrix}$
Lateral 2	42, 62	
Meio-Campo 1	45, 20	
Meio-Campo 2	45, 40	$\begin{pmatrix} 950 & 0 \\ 0 & 75 \end{pmatrix}$
Meio-Campo 3	45, 60	
Meio-Camp 4	60, 45	
Atacante 1	60, 25	$\begin{pmatrix} 500 & 0 \\ 0 & 55 \end{pmatrix}$
Atacante 2	80, 35	

Fonte: Do autor.

A Análise de Arquétipos foi realizada considerando duas abordagens: as posições ocupados por um único jogador e as posições de vários jogadores simultaneamente. O goleiro não foi levado em consideração neste trabalho pois este, geralmente, atua em uma parte bem limitada e extrema do campo, porém nada impede que ele seja inserido na análise. Para gerar os dados fictícios, adotou-se uma tática semelhante a utilizada pela Seleção Brasileira de Futebol na Copa do Mundo de 2014, com dois zagueiros, dois laterais, quatro meio-campistas e dois atacantes.

O número de arquétipos selecionados foi definido a partir do gráfico *ScreePlot* (Figura 4), observando-se que a minimização da soma de quadrados de resíduos ocorre entre cinco e oito arquétipos. Assim, após a simulação e análise dos dados foram retidos seis arquétipos para análise de um único jogador e oito arquétipos para análise de um grupo de jogadores. Estes

números de arquétipos foram definidos com o intuito de capturar melhor a forma como os dados estão distribuídos, e não só pela redução da SQR.

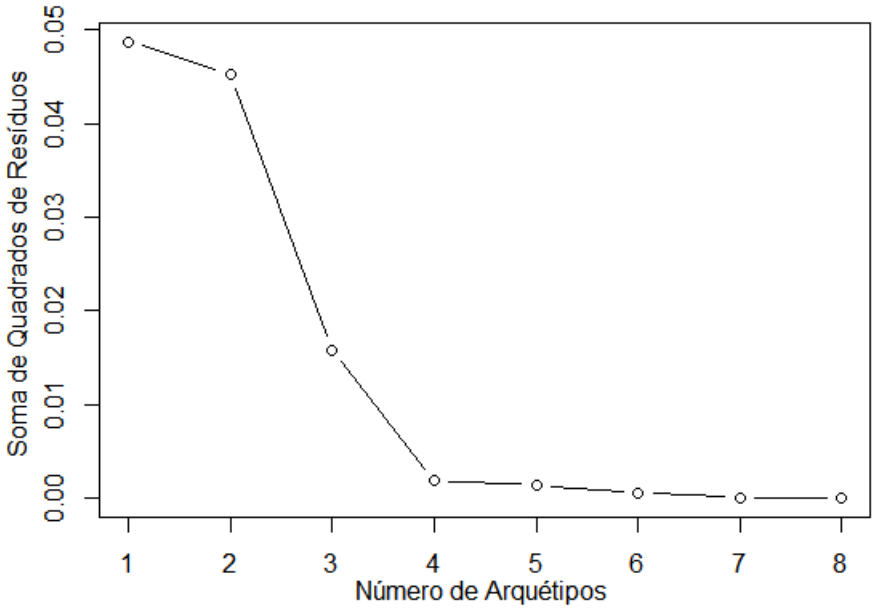


Figura 4 – *ScreePlot* da soma de quadrados de resíduos para os dados simulados.
Fonte: Do autor.

2.3 ARQUÉTIPOS E COMPONENTES PRINCIPAIS CONCENTRANDO INFORMAÇÃO SENSORIAL

O objetivo deste estudo foi, no tocante à interpretabilidade dos resultados e a capacidade de reconstruir dos dados originais, comparar três formas de aplicar a Análise de Componentes Principais, ilustradas na Figura 5, a saber: precedido de Análise de Arquétipos (AA_ACP), diretamente nos dados (D_ACP); precedido da média dos tratamentos (X_ACP). Em todos os casos serão utilizados dois componentes principais, afim de deixar os procedimentos comparáveis.

A capacidade dos métodos reconstruírem os dados foi avaliada por meio de simulações, descritos na seção 2.3.1. Já a interpretabilidade dos resultados foi avaliada em uma massa de dados reais descrita na seção 2.3.2.

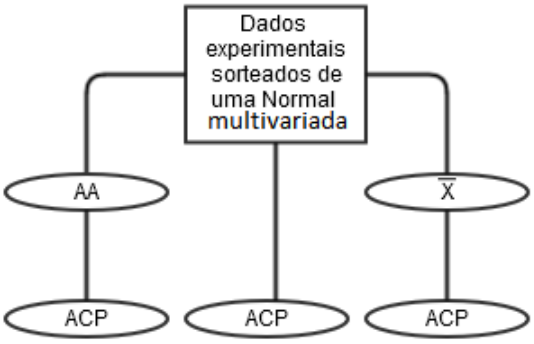


Figura 5 – Exemplo gráfico da ordem que serão aplicadas as análises na simulação.

Fonte: Do autor.

Para realização desta simulação foram desenvolvidas várias rotinas no programa R (R CORE TEAM, 2014). Foram desenvolvidas rotinas para realizar a ACP bem como os gráficos referentes a esta análise. Para realização da AA, foi utilizado o pacote *archetypes* do R (EUGSTER; LEISCH, 2009).

O método AA_ACP utiliza uma AA que retem dois arquétipos para cada tratamento. O objetivo deste primeiro passo é reduzir a dimensão-linha da matriz de dados para depois aplicar a ACP nos arquétipos selecionados. Portanto, o total de elementos será o dobro do número de tratamentos do experimento. A hipótese é que haja vantagem na associação das duas técnicas pois cada uma atua de uma forma na concentração da informação. Este procedimento é uma proposta deste trabalho.

O método X_ACP obtém primeiro a médias dos tratamentos e então aplica a ACP nas médias. Num contexto de experimentação o conceito de repetição é importante principalmente para estimar o erro experimental. Porém no contexto da Estatística Multivaria, o interesse principal é a similaridades e diferenças entre os tratamentos. Deste modo é normal que este método seja o mais utilizado na prática. Por isso o interesse em avaliar o seu desempenho tanto para reconstrução dos dados quanto para a interpretabilidade.

Já o método 3, a partir de agora D_ACP, não é tão utilizado na prática principalmente quando há repetições, pois este retorna um gráfico com o mesmo número de pontos da base de dados que dificulta a interpretabilidade da técnica.

2.3.1 Estudo de simulação

Os dados simulados foram sorteados de uma normal multivariada, com uma matriz $\boldsymbol{\Sigma}$ equicorrelacionada, com o parâmetro de correlação com os possíveis valores $\rho \in \{0;\ 0{,}25;\ 0{,}5;\ 0{,}75;\ 0{,}95\}$ e o vetor de médias foi tomado como o vetor nulo sem perda de generalidade. Foram considerados os números de tratamentos $t \in \{3, 10\}$, ambos os casos as médias dos tratamentos foram consideradas iguais. Assumiu-se para o número de variáveis os valores $p \in \{2, 5, 10e30\}$. As repetições de cada tratamento foram $r \in \{3, 5, 10, 15\}$. Com isso, foram avaliados 160 casos distintos, que é a combinação de cada situação possível descrita acima, e o número de repetições Monte Carlo foi fixado em 2000. Foram coletados os dados sobre a melhor estrutura de reconstrução para cada um dos métodos, bem como o erro de ajuste a cada iteração, para ao final calcular o total do erro cometido pelo método, o média dos erros e dos desvios-padrão.

A matriz de dados ${}_n\boldsymbol{X}_p$ pode se reescrita como ${}_{rt}\boldsymbol{X}p = \begin{bmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_i \\ \vdots \\ \boldsymbol{W}_t \end{bmatrix}$, em que $\boldsymbol{W}_i$ tem dimensão $r \times p$ e representa os dados do tratamento i.

A falta de ajuste dos métodos foi calculada com base na soma das soma de quadrados de resíduos (SQR). A soma das SQR de cada etapa resulta na SQR total, que foi posteriormente relativizada pela maior SQR de cada cenário estudado.

O erro cometido do método D_ACP foi apenas o erro ao reescrever os dados originais utilizando os componentes principais, por tanto, só o erro decorrente da ACP.

$$SQR_{D\_ACP} = SQR_{ACP} = tr[(\boldsymbol{X} - \boldsymbol{X}_{ACP})(\boldsymbol{X} - \boldsymbol{X}_{ACP})^{'}],$$

em que $\boldsymbol{X}_{ACP}$ são os dados recompostos utilizando os componentes principais.

O erro de ajuste para o método X_ACP foi calculado como a SQR de reconstruir os dados originais a partir dos componentes principais que foram extraídos das médias dos tratamentos. Ou seja, há o erro cometido pela ACP, e o erro cometido simplesmente por tomar a média dos dados, que foi calculado como

$$SQR_{X\_ACP} = SQR_{\bar{X}} + SQR_{ACP},$$

em que

$$SQR_{\bar{X}} = tr[(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}_i)(\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}_i)^{'}],$$

para $i = 1 \ldots t$ de forma que ao final a matriz resultante de $\mathbf{1}\bar{\boldsymbol{x}}_i$ tenha dimensão $rt \times p$ igual a de $\boldsymbol{X}$, tr é a função traço, $\mathbf{1}$ é um vetor unitário de dimensão $r \times 1$ e $\bar{\boldsymbol{x}}_i$ tem dimensão $1 \times p$ que contêm as médias do tratamento i. De forma mais detalhada, $\mathbf{1}\bar{\boldsymbol{x}}_i$ ao final de todas as iterações (cada tratamento), resultará em uma matriz da seguinte forma:

$$\mathbf{1}\bar{\boldsymbol{x}}_i = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} & \ldots & \bar{x}_{1p} \\ \bar{x}_{11} & \bar{x}_{12} & \ldots & \bar{x}_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{11} & \bar{x}_{12} & \ldots & \bar{x}_{1p} \\ \bar{x}_{21} & \bar{x}_{22} & \ldots & \bar{x}_{2p} \\ \bar{x}_{21} & \bar{x}_{22} & \ldots & \bar{x}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{21} & \bar{x}_{22} & \ldots & \bar{x}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \bar{x}_{i1} & \bar{x}_{i2} & \ldots & \bar{x}_{ip} \end{bmatrix},$$

que tem a mesma dimensão da matriz dos dados originais $\boldsymbol{X}$.

Ou seja, os dados são recompostos como o vetor média de cada tratamento repetido r vezes para cada tratamento, desta forma a matriz $\bar{\boldsymbol{x}}_i$ será os dados recompostos utilizando a função média ($\bar{X}$).

Logo este erro é calculado como a soma do erro de ter calculado a média mais o erro da ACP (SQR_{ACP}).

O erro de ajuste do método AA_ACP foi similar ao do X_ACP, porém foi calculado passo a passo para cada tratamento durante a AA na matriz que é entrada para a ACP, pois como os arquétipos são referentes a cada tratamento, não seria possível fazer tal calculo de outra maneira. Logo este erro é calculado como o somatório do erro pela realização da AA retendo dois arquétipos para cada tratamento, somado ao erro da realização da ACP.

$$SQR_{AA\_ACP} = SQR_{AA} + SQR_{ACP},$$

em que SQR_{AA} é a expressão 1.4 reescrita da seguinte forma:

$$SQR_{AA} = \sum_{i=1}^{n} ||\boldsymbol{W}_i - \boldsymbol{W}_i^*||^2$$

em que W_i^* é a matriz que representa os dados originais do tratamento i recomposta utilizando os dois arquétipos de cada tratamento i.

É importante notar que o método proposto neste trabalho, AA_ACP, é puramente descritivo, e não entrou-se no mérito de estudar suas propriedades para inferência.

2.3.2 Estudo com dados reais

Os mesmos métodos (D_ACP, X_ACP e AA_ACP) descritos em 2.3.1 foram utilizados em um conjunto de dados reais sobre características sensoriais de um alimento. Estes dados foram extraídos de um experimento sensorial sobre hambúrgueres realizado na UNIFAL-MG, departamento de nutrição, durante a disciplina "Análise Sensorial de alimentos e bebidas" ministrada em 2013. As cinco marcas de hambúrgueres estudadas foram: Sadia, Sadia-Frango, Pif-Paf, Perdigão e Friboi. Os hambúrgueres foram considerados os tratamentos e foram avaliadas as seguintes variáveis: cor marrom, oleosidade superficial, aroma característico, aroma de ervas, sabor característico, gosto salgado, oleosidade, dureza, fraturabilidade e suculência. Cada tratamento foi repetido com sete provadores que foram previamente treinados e notas de zero à dez foram atribuídas a cada uma das variáveis.

3 RESULTADOS

Nesta seção serão apresentadas os resultados referentes a cada um dos artigos, que são eles: “Avaliação Monte Carlo de métricas para falta de ajuste em Análise de Arquétipos”, será chamado apenas de “Métricas” (MARTINS JÚNIOR et al., 2014). “Análise de Arquétipos na avaliação da movimentação de jogadores de futebol”, será chamado de “Futebol” (MARTINS JÚNIOR et al., 2015a). O artigo provisoriamente denominado “Arquétipos e componentes principais concentrando a informação sensorial”, que contempla a simulação proposta nesta dissertação e seus resultados também são apresentados nesta seção.

3.1 MÉTRICAS

Tabela 3 – Valores médios da SQR e respectivos erros-padrão considerando variâncias (σ^2) 0,1 (pequena); 0,5 (média); 1,0 (grande) dos erros (ϵ), correlação (ρ) $0{,}0; 0{,}25; 0{,}5; 0{,}75; 0{,}95$ e número de repetições Monte Carlo fixado em 1000, para as diferentes métricas.

		Correlação				
	Variância	ρ =0,00	ρ =0,25	ρ =0,5	ρ =0,75	ρ =0,95
	Pequena	0,15 ± 0,002	-	-	-	-
NQ	Média	0,69 ± 0,011	0,36 ± 0,006	0,01 ± 0,000	-	-
	Grande	1,28 ± 0,021	0,97 ± 0,017	0,56 ± 0,009	0,36 ± 0,006	0,08 ± 0,001
	Pequena	1,24 ± 0,010	-	-	-	-
SQPR	Média	2,66 ± 0,021	1,91 ± 0,016	0,27 ± 0,003	-	-
	Grande	3,59 ± 0,030	3,11 ± 0,027	2,35 ± 0,020	1,88 ± 0,017	0,87 ± 0,008
	Pequena	0,38 ± 0,003	-	-	-	-
Frob	Média	0,81 ± 0,006	0,58 ± 0,005	0,08 ± 0,001	-	-
	Grande	1,10 ± 0,009	0,95 ± 0,008	0,72 ± 0,006	0,58 ± 0,005	0,27 ± 0,002
	Pequena	0,33 ± 0,003	-	-	-	-
Espec	Média	0,70 ± 0,006	0,50 ± 0,004	0,08 ± 0,001	-	-
	Grande	0,95 ± 0,008	0,82 ± 0,007	0,62 ± 0,006	0,50 ± 0,005	0,23 ± 0,002
	Pequena	$10^{-47} \pm 10^{-47}$	-	-	-	-
Det	Média	$10^{-42} \pm 10^{-43}$	$10^{-44} \pm 10^{-44}$	-	-	-
	Grande	$10^{-41} \pm 10^{-41}$	$10^{-42} \pm 10^{-42}$	$10^{-43} \pm 10^{-43}$	$10^{-44} \pm 10^{-44}$	$10^{-49} \pm 10^{-49}$

Fonte: Do autor.

Na Tabela 3 são apresentados os resultados obtidos no estudo de simulação. As células que contem o sinal “-” representam situações impossíveis de simulação, pois a covariância, nesses casos precisaria ser maior do que a

própria variância, o que não faz sentido.

Por meio dos resultados descritos, observa-se que todas as métricas estudadas apresentam o mesmo comportamento. À medida que a correlação ρ aumenta a qualidade de ajuste melhora, ou seja, o valor da diferença entre os dados originais e os dados recompostos diminui. Quando a perturbação σ^2 causada nos erros aumenta, a qualidade do ajuste piora, ou seja, o valor da diferença aumenta.

Usar o determinante como métrica não apresentou bons resultados, pois o resultado é sempre muito próximo de zero. Isso era esperado, pois ele é aplicado em uma matriz positiva semi-definida de soma de quadrados e produtos de resíduos. Uma alternativa para contornar este problema seria utilizar somente os autovalores não-nulos.

Uma possível explicação para a qualidade do ajuste melhorar à medida que a correlação aumenta, é que fica mais fácil obter os coeficientes da mistura dos arquétipos, implicando em uma melhor adaptação dos arquétipos aos dados e consequentemente na diminuição do erro da diferença entre os dados originais e dos dados recuperados a partir dos arquétipos.

3.2 FUTEBOL

O fluxograma da Figura 6, descreve os passos para realizar a análise apresentada neste trabalho.

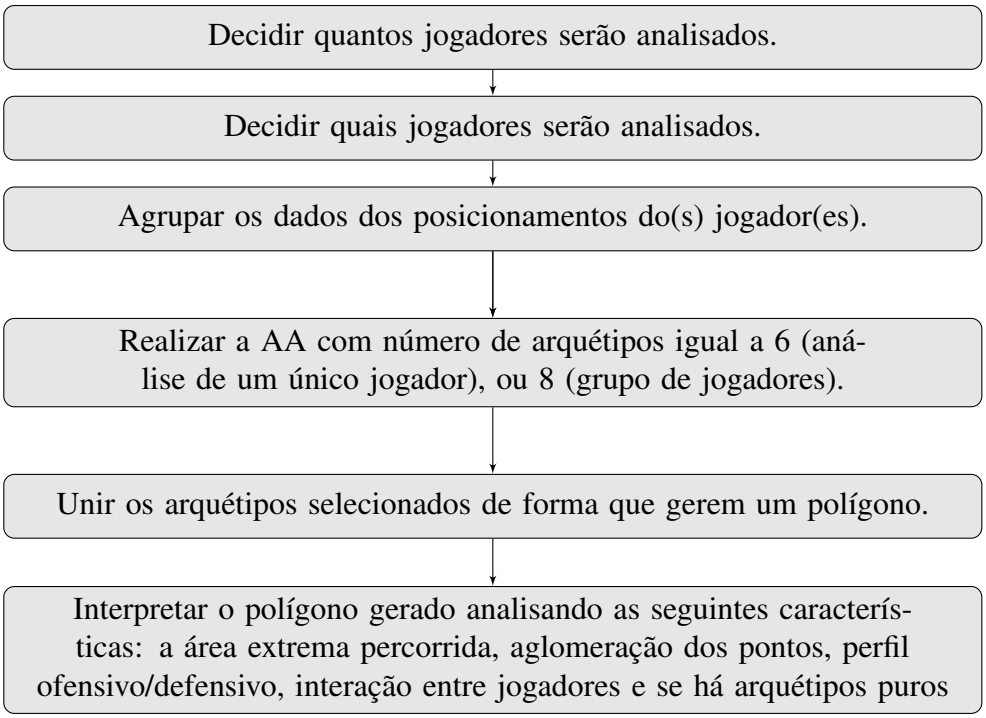


Figura 6 – Fluxograma contendo os passos para a análise sugerida neste trabalho.
Fonte: Do autor.

Para a análise e discussão dos resultados deste trabalho, levou-se em consideração a tendência de atuação mais frequente em um setor que em outro (aglomeração dos pontos); onde o(s) jogador(es) analisado(s) estava(m) mais concentrado(s) (perfil ofensivo/defensivo); a interação entre grupos de jogadores ou jogadores individuais foi

verificada com base na intersecção dos polígonos gerados pelos arquétipos; a área que o polígono dos arquétipos identifica como sendo as áreas mais frequentes utilizadas no tempo determinado (área extrema percorrida); na identificação dos arquétipos, se são jogadores designados para atuar na área estabelecida e, mais especificamente, verificar se houve arquétipos puros (coeficiente b igual a um para alguma observação), e também se o arquétipo puro é o jogador que deveria estar naquele setor.

Para este tipo de análise, a grande vantagem em se utilizar a AA, reside no fato de que se um ponto for arquétipo, este permite literalmente reescrever os dados originais, ou seja, o arquétipo é um dado que junto com outros arquétipos, representa bem o conjunto de dados originais.

A movimentação de todo o time em um dos tempos normais de jogo (sem acréscimos) de uma partida de futebol, está representada na Figura 7, destacando-se que os pontos azuis foram considerados arquétipos. Excluindo apenas as bordas mais extremas do campo, observa-se que praticamente todo campo foi percorrido. Pode-se afirmar ainda que o time agiu de forma balanceada em termos de ataque e defesa, já que o campo todo está representado por arquétipos não havendo grandes aglomerações em um único setor. Os resultados sugerem que um lado ofensivo do campo (inferior direito) ficou mais vago, o que já era esperado devido à escolha tática durante a simulação (Figura 3).

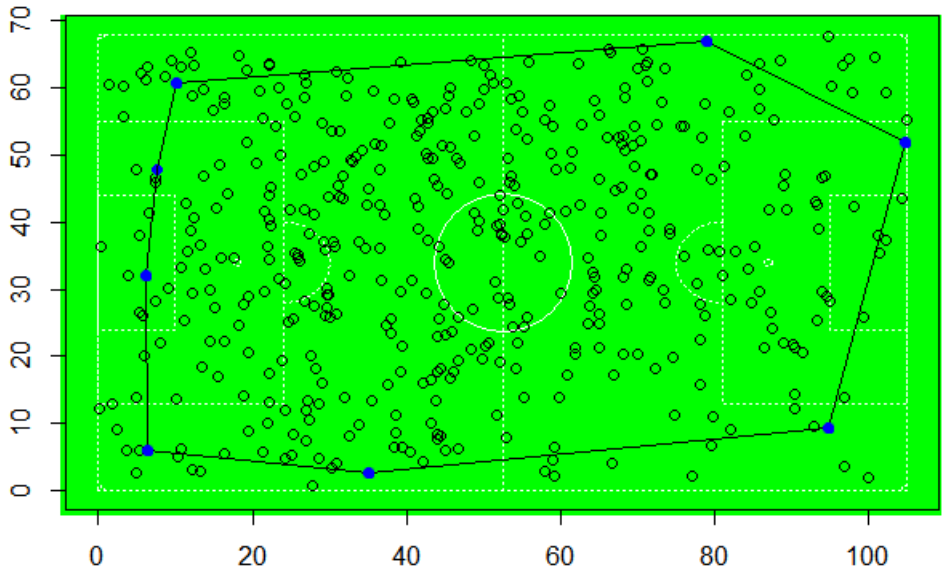


Figura 7 – Movimentação dos jogadores em um dos tempos normais da partida. Pontos com cores sólidas representam os arquétipos, enquanto os pontos sem preenchimento representam os dados. Todo o time (dez jogadores de linha) foram levados em consideração.
Fonte: Do autor.

Vale ressaltar que os arquétipos podem ou não ter sido observados pelo experimento. Neste caso não houve nenhum arquétipo puro, ou seja, nenhum ponto foi exatamente um arquétipo. Caso fosse encontrado um arquétipo puro, seria importante verificar se foi um ponto da movimentação do jogador designado para aquela área ou se foi de um jogador de outra área do campo. Para representar a movimentação de dois jogadores do time individualmente em um determinado tempo normal de jogo, sem os acréscimos de uma partida de futebol, foram selecionados seis arquétipos para a movimentação de cada um dos jogadores.

A Figura 8 representa os resultados da análise da movimentação dos dois atacantes. Observa-se que a

Análise de Arquétipos conseguiu destacar a área que foi percorrida pelos jogadores analisados. Os arquétipos selecionados mostraram que ambos jogadores atuaram, mesmo que pouco, fora de seus territórios de oficio. Nota-se também que o atacante mais avançado (arquétipos vermelhos) recuou no máximo até a intermediária defensiva, mostrando uma postura bem ofensiva, assim como o atacante mais próximo à lateral (arquétipos azuis). Este resultado já era esperado, dado que são os jogadores mais avançados de acordo com o esquema tático simulado para este time. Os resultados para este caso, reforçam ainda mais a postura ofensiva do jogador visto que houve poucos pontos do lado defensivo do campo.

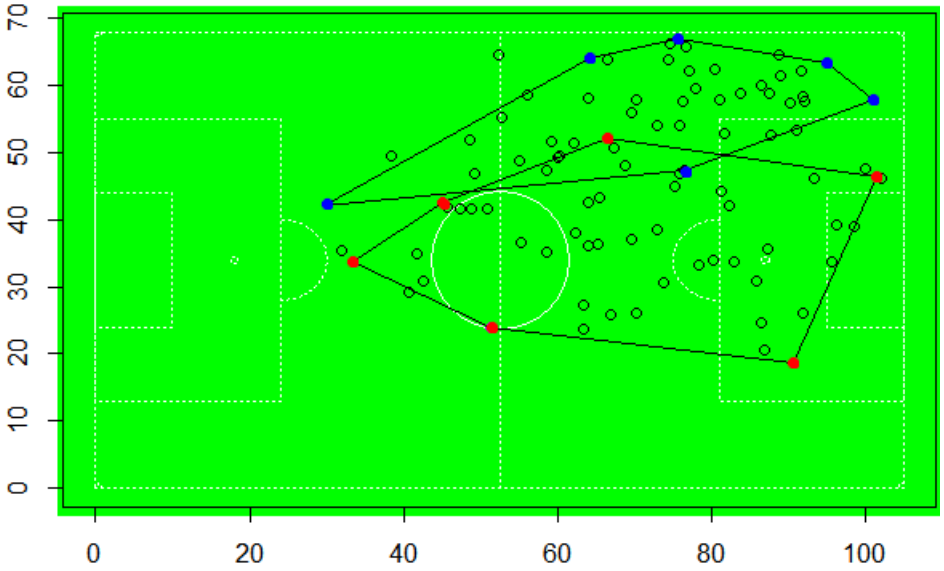


Figura 8 – Movimentação dos dois atacantes do time em um dos tempos normais da partida. Pontos com cores sólidas representam os arquétipos, enquanto os pontos sem preenchimento representam os dados. Os arquétipos do atacante mais próximo a lateral esta em azul, os arquétipos do atacante mais centralizado está em vermelho.
Fonte: Do autor.

Como a tática estudada, (Figura 3), apresentava apenas um jogador realmente avançado, nota-se que ele não percorreu toda a área do campo adversário, optando geralmente pelo lado superior (Figura 8). O atacante mais próximo à lateral (arquétipos azuis), jogou de forma mais compactada que o atacante central. Por isso, pode-se observar uma maior aglomeração de pontos em regiões menores, devido a restrição dos pontos estarem todos dentro do campo. Como este está mais próximo a lateral, seus pontos simulados que situarem fora dos limites do campo, serão descartados, logo os pontos serão mais aglomerados que de outro jogador mais centralizado que em geral não sofreu tanto com a restrição dos limites do campo.

Vale ressaltar que a área de atuação do jogador por si só, não é informativa, pois não necessariamente um jogador que ficou mais avançado marcou gols. Da mesma forma, um jogador que não esteve presente na área pode invariavelmente marcar algum gol.

3.3 ARQUÉTIPOS E COMPONENTES PRINCIPAIS CONCENTRANDO INFORMAÇÃO SENSORIAL

Neste tópico são apresentados os resultados referentes ao estudo de simulação e a aplicação a dados reais de hambúrgueres dos métodos propostos para concentrar a informação decorrente de um experimento sensorial.

3.3.1 Resultados do estudo de simulação

As Figuras de 10 a 13 representam a SQR média ao longo das repetições de Monte Carlo relativizada ao reconstruir os dados utilizando os três métodos estudados.

Como a SQR dos métodos X_ACP e AA_ACPé composta pela soma de duas SQR parciais, esta distinção foi feita nos gráficos. O método D_ACP não tem outra fonte de variação além da ACP, por isso não houve necessidade de tal distinção. Em todos os casos, a ACP foi representada como o valor mais acima como pode ser visto na Figura 9. A soma das SQR de cada etapa resulta na SQR total, que foi posteriormente relativizada pela maior SQR de cada cenário estudado.

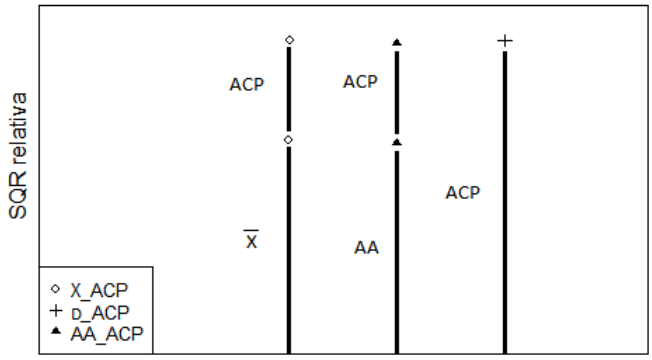


Figura 9 – Exemplo de como serão apresentados os erros de cada método.

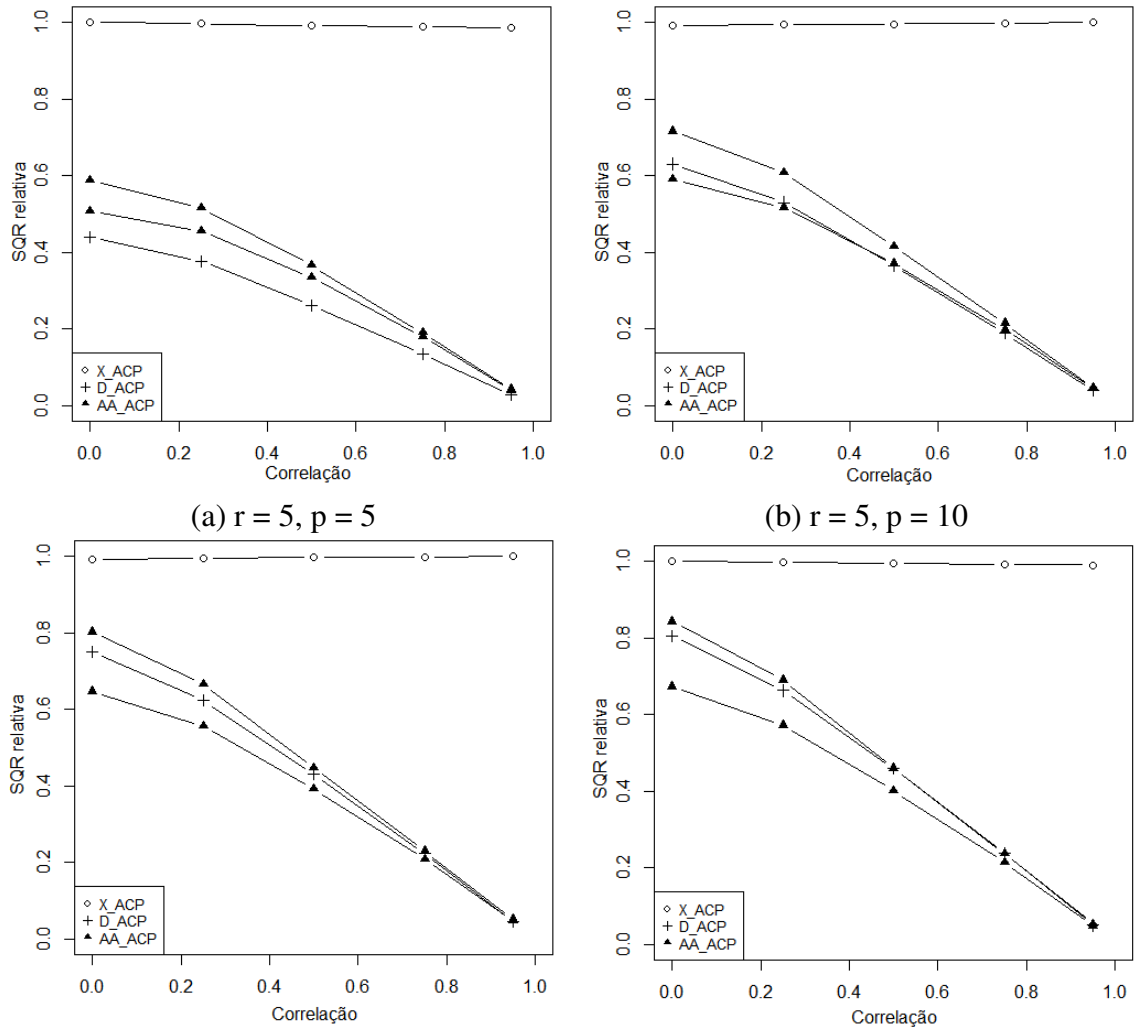
Fonte: Do autor.

Na Figura 10 pode-se observar que em todos os casos estudados a X_ACP teve o pior desempenho em reconstrução das informações, o que já era esperado pois a função média descarta muitas informações sobre os dados. A melhor função no tocante a reconstrução foi a D_ACP. O método AA_ACP obteve um desempenho muito semelhante ao D_ACP, porém levemente pior.

Na Figura 11 o comportamento foi o mesmo anterior. Porém neste cenário é mais evidente que AA_ACP reconstrói pior que D_ACP à medida que o numero de repetições aumentou.

Na Figura 12 o comportamento continua parecido com os anteriores mas no caso 12(d) o método D_ACP teve desempenho infimamente pior que o AA_ACP para correlação $0{,}25$ à $0{,}75$.

Na Figura 13 o comportamento das situações anteriores novamente aconteceu, X_ACP como pior reconstrução, seguido de AA_ACP e D_ACP. Quanto menor o número de variáveis, mais evidente que AA_ACP reconstrói pior que D_ACP.



(a) r = 5, p = 5

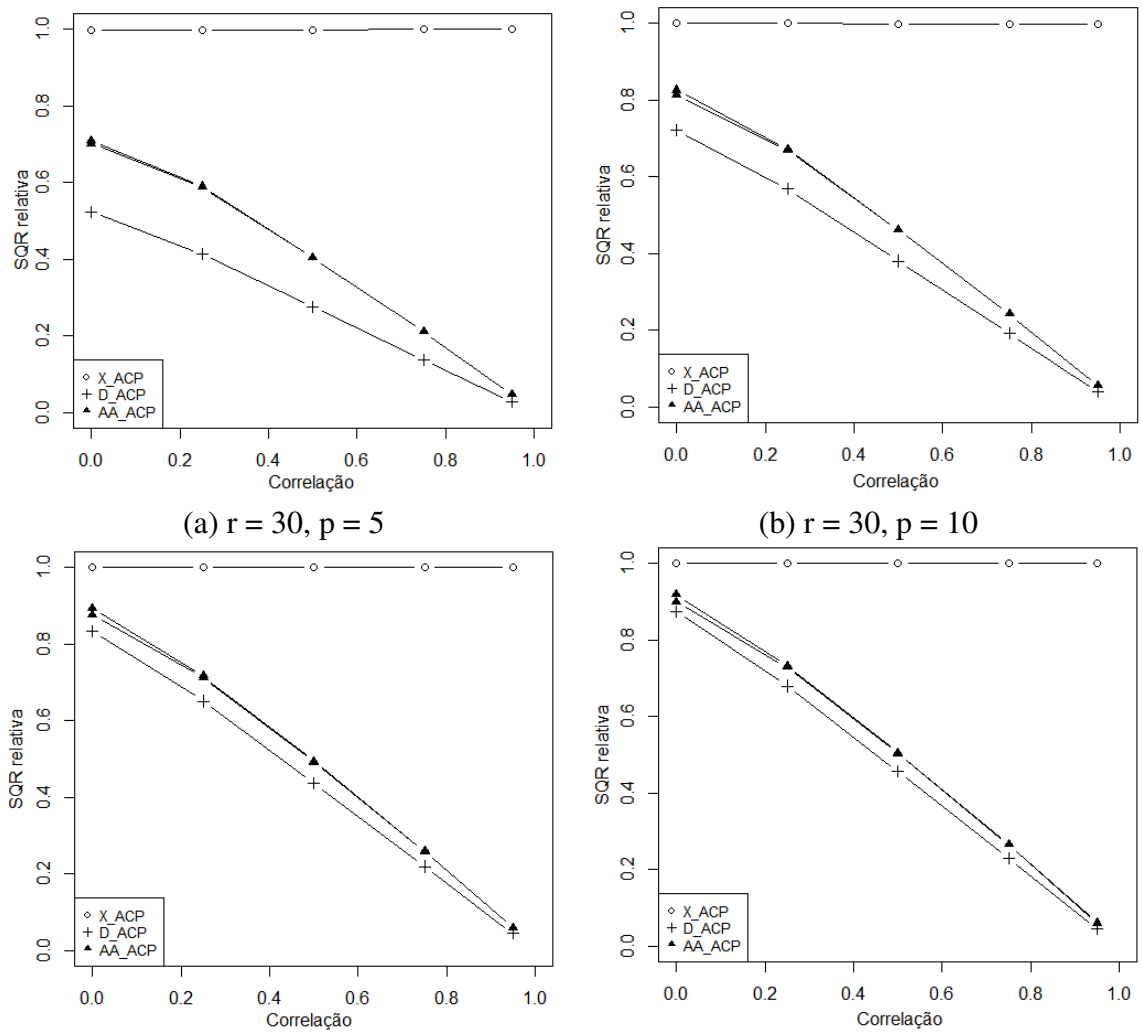
(b) r = 5, p = 10

(c) r = 5, p = 20

(d) r = 5, p = 30

Figura 10 – Soma de quadrado de resíduo relativa para os métodos X_ACP, D_ACP e AA_ACP ao longo de diferentes correlações para os cenários com $t = 3$ tratamentos e $r = 5$ repetições.

Fonte: Do autor.



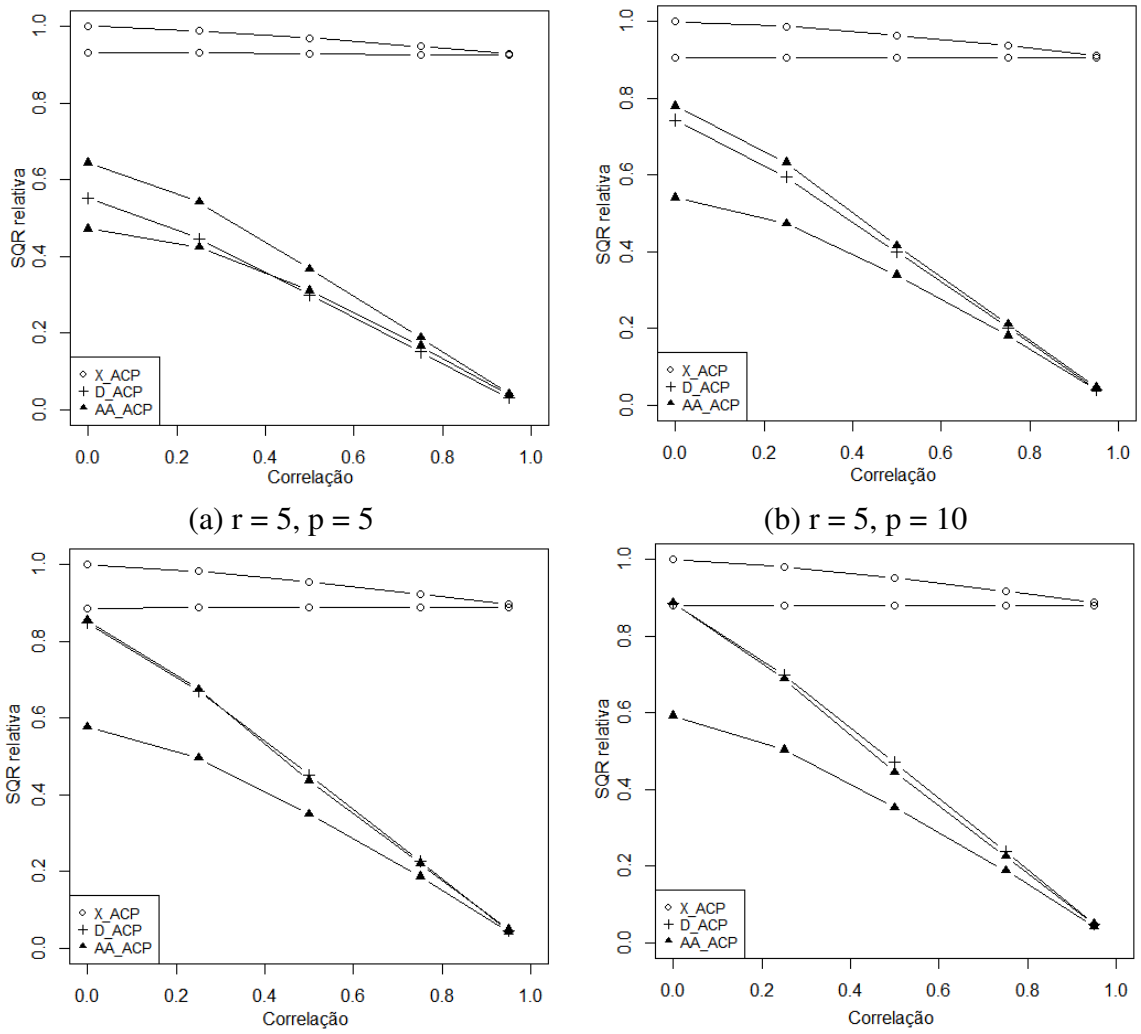
(a) r = 30, p = 5

(b) r = 30, p = 10

(c) r = 30, p = 20

(d) r = 30, p = 30

Figura 11 – Soma de quadrado de resíduo relativa para os métodos X_ACP, D_ACP e AA_ACP ao longo de diferentes correlações para os cenários com $t = 3$ tratamentos e $r = 30$ repetições.
Fonte: Do autor.



(a) r = 5, p = 5

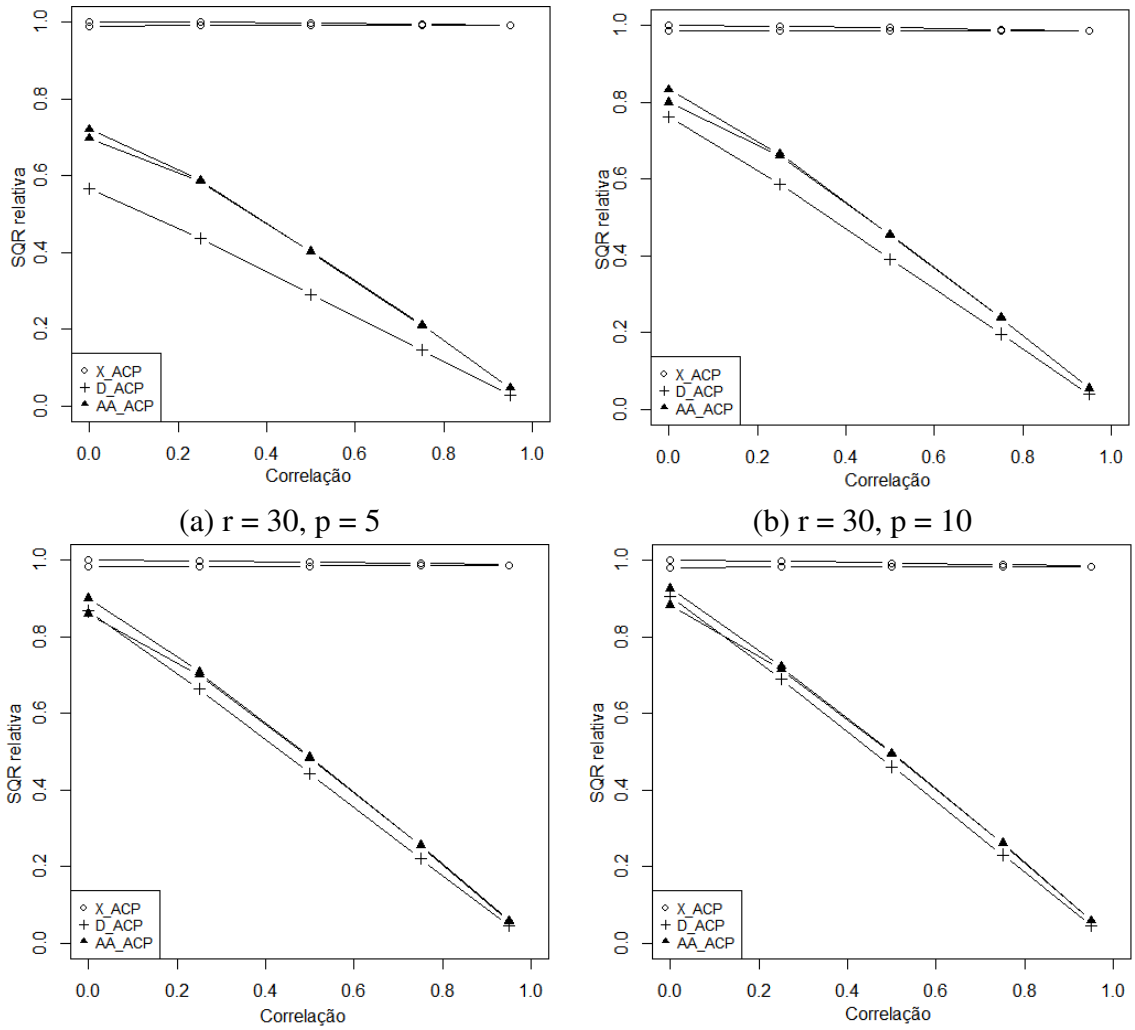
(b) r = 5, p = 10

(c) r = 5, p = 20

(d) r = 5, p = 30

Figura 12 – Soma de quadrado de resíduo relativa para os métodos X_ACP, D_ACP e AA_ACP ao longo de diferentes correlações para os cenários com $t = 10$ tratamentos e $r = 5$ repetições.

Fonte: Do autor.



(a) r = 30, p = 5

(b) r = 30, p = 10

(c) r = 30, p = 20

(d) r = 30, p = 30

Figura 13 – Soma de quadrado de resíduo relativa para os métodos X_ACP, D_ACP e AA_ACP ao longo de diferentes correlações para os cenários com $t = 10$ tratamentos e $r = 30$ repetições.

Fonte: Do autor.

As Figuras 10 a 13 apresentam o resultado da dos métodos X_ACP, D_ACP e AA_ACP utilizando dados simulados. A SQR foi expressa relativa ao maior valor em cada cenário com o objetivo de facilitar a comparação, visto que a medida que a quantidade de dados aumenta, é obvio que a SQR aumentará, causando uma possível falsa impressão que os métodos pioram.

3.3.2 A influência da correlação

No tocante à influência da correlação entre variáveis na capacidade dos métodos reconstruírem os dados originais, pode-se observar nas Figuras de 10 a 13 que o erro diminui a medida que a correlação aumenta. Uma possível explicação para esse fato é que quanto maior a correlação, mais informação é fornecida, resultando em um melhor ajuste, independente da técnica utilizada (ACP ou AA).

Quando se utiliza o método X_ACP, a SQR relativa diminui, porém muito lentamente, sendo quase impossível notar esse comportamento a olho nu. O interessante é notar que a $SQR_{\bar{X}}$ é constante, pois a média não é afetada pelo aumento da correlação. Por outro lado, a SQR_{ACP} se reduz, mas como já era pequena desde o início, sua redução mal é percebida.

3.3.3 A influência do número de variáveis

Pode-se observar também que à medida o número de variáveis aumenta, os métodos D_ACP e AA_ACP erram proporcionalmente mais. Destaque deve ser dado ao D_ACP que, pela primeira vez, supera o erro cometido pelo AA_ACP (Figura 12(d)).

O aumento do número de variáveis não é tão impactante quanto o aumento da correlação entre elas. Por exemplo, o erro cometido pelos três métodos em um cenário com poucas variáveis com correlação alta é menor que o erro cometido em um cenário com muitas variáveis com correlação baixa.

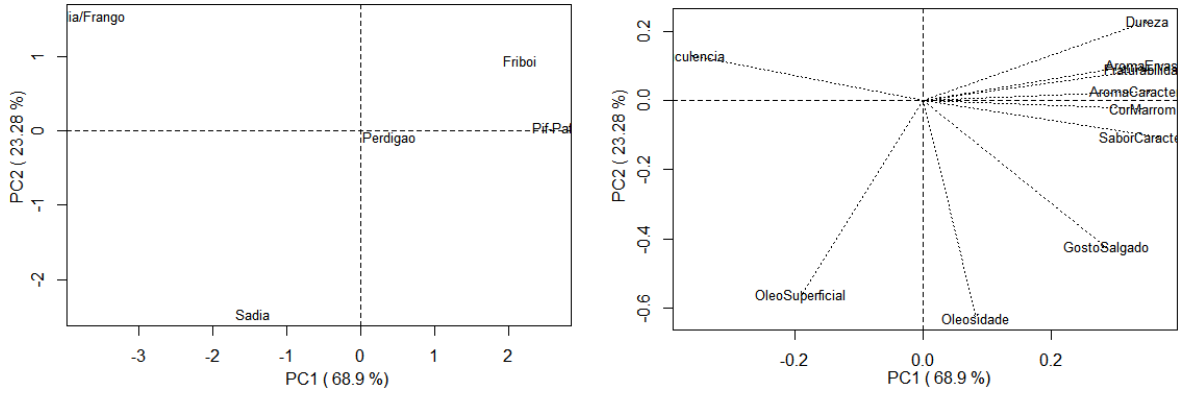
3.3.4 A influência do número de tratamentos e repetições

Submeter os métodos testados a diferentes números de tratamentos pouco afeta o comportamento relativo dos erros cometidos. Uma pequena diferença pode ser observada em cenários com muitos tratamentos e poucas repetições (Figuras 12(a), 12(b), 12(c) e 12(d)), onde se percebe, no método X_ACP uma participação mais expressiva do erro cometido pela sua fração SQR_{ACP}.

Aumentar o número de repetições também pouco afeta a proporcionalidade dos erros. De forma geral, aumentar o número de repetições implica em aumentar o tamanho da massa de dados, e isso aumenta o número de dados que precisam ser reescritos, gerando maiores SQR.

3.3.5 Resultados do experimento com dados reais

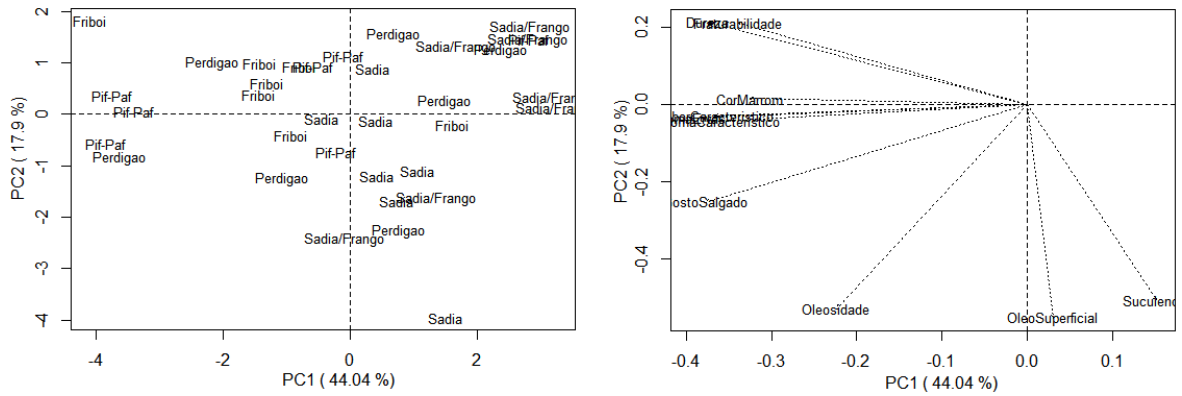
As Figuras de 14 a 16 representam o resultado dos três métodos avaliados neste trabalho aplicados a dados reais de um experimento sensorial demarcas de hambúrgueres de marcas comerciais da cidade de Alfenas-MG. As figuras com legenda a significam espaço de observações e as figuras com legenda b representam o espaço de variáveis.



(a) (b)

Figura 14 – Espaços de observações (a) e de variáveis (b) resultante do método X_ACP.
Fonte: Do autor.

O método X_ACP mostrou-se com uma excelente interpretabilidade. Porém esse método tem resultado pontual, não tendo noção sobre variabilidade, o que também não é um problema pois poderia ser resolvido, por exemplo, utilizando uma elipse ou região de confiança *bootstrap*. Com esse método é possível chegar em resultados como, quais tratamentos são similares ou diferentes, peso médio das variáveis em cada tratamento, por exemplo, o hambúrguer Friboi teve notas mais altas em média para a variável dureza; o hambúrguer de frango da Sadia em média é o mais suculento de todos.

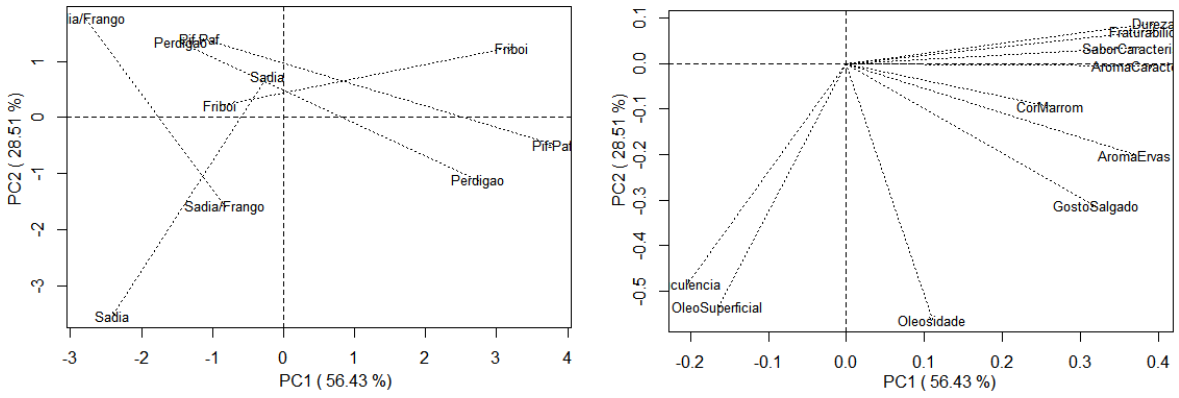


(a) (b)

Figura 15 – Espaços de observações (a) e de variáveis (b) resultante do método D_ACP.
Fonte: Do autor.

O método D_ACP teve o melhor desempenho em reconstrução na simulação anterior, porém no contexto

experimental, fica evidente que esse método não tem boa interpretabilidade devido a quantidade de observações apresentadas simultaneamente. É difícil chegar em conclusões gerais sobre as variáveis mas pode-se fazer um raciocínio análogo ao método anterior para uma determinada observação fixada.



(a) (b)

Figura 16 – Espaços de observações (a) e de variáveis (b) resultante do método AA_ACP.
Fonte: Do autor.

O método AA_ACP é capaz de fornecer as mesmas informações que podem ser observadas no X_ACP da Figura 14, além de já apresentar uma noção de variabilidade devido aos arquétipos. Ainda há uma vantagem pois o método selecionou observações de referência, caricatas dos dados originais e com capacidade de reconstruir os dados originais com um pequeno erro associado.

Quanto ao percentual dos dados explicado pelos componentes principais, pode-se afirmar que claramente o método X_ACP, que tem 68,8% de informação no primeiro componente e 23,28% no segundo, superestima seu percentual explicado, pois como foi comprovado que este método perde mais informações que os outros métodos, não importa explicar muito bem a informação restante. Já o método D_ACP que parece explicar menos que os outros métodos, com apenas 44,04% no primeiro componente e 17,9% no segundo componente, mostra exatamente quando a ACP é capaz de explicar desses dados em apenas dois componentes principais. Neste quesito, porcentagem de explicação, este método é a referencia honesta de resultado. O método AA_ACP explica 56,43% no primeiro componente e 28,51% no segundo componente, também é superestimado, porém este é mais honesto que o X_ACP pois perde menos informação na hora de reconstruir os dados.

Na Figura 14, é a maneira como os pesquisadores em geral utilizam a ACP. Este método fornece uma excelente maneira de interpretar os resultados e chegar em conclusões sobre as médias dos tratamentos que neste caso eram marcas de hambúrgueres. Algumas variáveis ficaram bem correlacionadas visualmente, por exemplo aroma de ervas e fraturabilidade, aroma característico e cor marrom. O hambúrguer de frango da marca Sadia é caracterizado pela variável suculência, e também pela ausência de cor marrom e sabor característico. O hambúrguer da Sadia feito com carne bovina foi caracterizado pela oleosidade superficial, suculência e ausência de aroma de ervas e dureza. O hambúrguer da Perdigão ficou na média das variáveis, por isso encontra-se próximo a origem dos eixos. Os hambúrgueres Pif-Paf e Friboi foram bem similares, destaca-se apenas que Friboi foi caracterizado pela dureza, enquanto Pif-Paf foi caracterizado principalmente por aroma característico e cor marrom. A interpretação deste método é bem simples, mas tem a desvantagem da reconstrução dos dados originais, pois como foi mostrado,

esse método desperdiça muito informações importantes sobre os dados.

Na Figura 15 a interpretação é análoga a anterior, porém como não foi realizada nenhum procedimento prévio nos dados (como tomar a média ou utilizar a AA em cada tratamento), todas as observações são apresentadas simultaneamente, tornando a interpretação árdua e tediosa, de forma que fica difícil chegar em alguma conclusão geral para algum tratamento específico.

Na Figura 16, pode-se chegar em diversas conclusões quanto as marcas estudadas. Os arquétipos dos hambúrgueres da marca Sadia de Frango, são caracterizados pela suculência e ausência de cor marrom, o que é esperado de um hambúrguer feito de carne mais branca como a de frango. Nota-se pela direção em que se encontra os arquétipos, que houve discordância entre a graduação dos níveis de ausência de gosto salgado, aromas de ervas e oleosidade. Os arquétipos dos hambúrgueres da marca Friboi foram caracterizados pelas variáveis dureza, fraturabilidade, sabor característico e aroma característico. Já o outro arquétipo foi caracterizados pela média das outras variáveis, ficando bem próximo a origem dos eixos. Os arquétipos dos hambúrgueres da marca Friboi foram caracterizados pelas variáveis dureza, fraturabilidade, sabor característico, aroma característico. Já o outro arquétipo foi caracterizados pela média das outras variáveis, ficando bem próximo a origem dos eixos. Em posse dessas informações o profissional da tecnologia de alimentos tem noção da faixa de variação daquele tratamento, e então pode controlar o processo de fabricação para manter o padrão estabelecido.

4 CONCLUSÕES

Nesta seção serão apresentadas as conclusões referentes a cada um dos artigos, que são eles: “A Análise de Arquétipos: uma revisão bibliográfica”, chamado de “Revisão” (MARTINS JÚNIOR et al., 2015b). “Avaliação Monte Carlo de métricas para falta de ajuste em Análise de Arquétipos”, será chamado apenas de “Métricas” (MARTINS JÚNIOR et al., 2014). “Análise de Arquétipos na avaliação da movimentação de jogadores de futebol”, será chamado de “Futebol” (MARTINS JÚNIOR et al., 2015a). O artigo provisoriamente denominado “Arquétipos e componentes principais concentrando a informação sensorial”, a partir de agora “Arquétipos e componentes principais concentrando informação”, sua conclusão também será explicada nesta seção.

4.1 REVISÃO

Portanto, a Análise de Arquétipos se apresenta como técnica multivariada relativamente recente e bastante promissora. Suas aplicações se estendem pelas diversas áreas do conhecimento. Ainda não é tão estudada quanto outras técnicas multivariadas como a Análise de Componentes Principais, mas tem potencial para mudar esse cenário no futuro.

4.2 MÉTRICAS

De acordo com os resultados apresentados, pode-se concluir que todas as métricas são equivalentes. Portanto, este trabalho indica o uso da mais simples, ou seja, a soma de quadrados de resíduos.

4.3 FUTEBOL

A Análise de Arquétipos permitiu uma nova abordagem para análise de dados sobre movimentação de jogadores nos esportes, auxiliando tanto para a análise com foco em um jogador como para um grupo de jogadores. A técnica pode ser utilizada como uma alternativa a outras análises já difundidas no cenário esportivo para auxiliar em tomadas de decisão pela comissão técnica de uma equipe.

Com os resultados obtidos da AA de um tempo normal do jogo sem acréscimos, é possível avaliar se o grupo de jogadores analisados realmente atuam na área designada ou se eles tendem a atuar em outros setores do campo, seja para surpreender o time adversário como também para auxiliar em jogadas em outras partes do campo.

Os resultados também permitiram avaliar se o time/jogador teve um caráter mais ofensivo ou defensivo durante o tempo analisado, de acordo com o polígono construído unindo os arquétipos.

Como motivação para a continuidade do trabalho, pode-se investir no uso de *heatmaps* e curvas de níveis para avaliar as áreas de atuação do(s) jogador(es).

4.4 ARQUÉTIPOS E COMPONENTES PRINCIPAIS CONCENTRANDO INFORMAÇÃO SENSORIAL

Em geral, X_ACP foi pior em termos de reconstrução dos dados, pois a função média descarta muita informação, mas por outro lado tem uma excelente interpretabilidade prática dos resultados, como pode ser visto na Figura 14. Já D_ACP obteve o melhor resultado em reconstrução, porém como pode ser visto nos na Figura 15, a interpretação dos dados fica prejudicada pela quantidade de pontos apresentados simultaneamente. 16 O método AA_ACP que foi proposto neste trabalho, obteve um desempenho em reconstrução suavemente pior que o D_ACP, porém sua interpretação é privilegiada pelo uso dos arquétipos, e deste modo é possível obter informações que não seria possível utilizando as funções D_ACP ou X_ACP.

Vale notar que o método AA_ACP é puramente descritivo e não é possível a realização de inferências em seu resultado.

REFERÊNCIAS

ARRUDA, M. L. **Poisson, Bayes, Futebol e DeFinetti**. 2000. 127 f. Dissertação (Mestrado em Estatística) — Universidade de São Paulo, São Paulo, 2000.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS - ABNT. **NBR ISO 5492:2014**: Análise sensorial — vocabulário. Rio de Janeiro, 2014. 25 p.

BAUCKHAGE, C. A note on archetypal analysis and the approximation of convex hulls. **arXiv preprint arXiv:1410.0642**, 2014.

BAUCKHAGE, C.; THURAU, C. Making archetypal analysis practical. In: **Pattern Recognition**. Springer, 2009. p. 272–281. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03798-6_28>. Acesso em: 21 ago. 2014.

BRO, R.; JONG, S. D. A fast non-negativity-constrained least squares algorithm. **Journal of chemometrics**, v. 11, n. 5, p. 393–401, 1997.

CHAN, B. H. P.; MITCHELL, D. A.; CRAM, L. E. Archetypal analysis of galaxy spectra. **Monthly Notice of the Royal Astronomical Society**, v. 338, n. 3, p. 790–795, 2003.

CORSARO, S.; MARINO, M. Archetypal analysis of interval data. **Reliable Computing**, v. 14, p. 105–116, 2010.

COSTANTINI, P.; PORZIO, G. C.; RAGOZINI, G.; ROMO, J. Archetypal functions. In: **Analysis and Modeling of Complex Data in Behavioural and Social Sciences**. Anacapri, Italy: Springer, 2012. p. 4.

CUTLER, A.; BREIMAN, L. Archetypal analysis. **Technometrics**, v. 36, n. 4, p. 338–347, 1994.

D'ESPOSITO, M. R.; PALUMBO, F.; RAGOZINI, G. Archetypal analysis for interval data in marketing research. **Statistica Applicata**, v. 18, n. 2, p. 343–358, 2006.

____. On the use of archetypes and interval coding in sensory analysis. In: FICHET, B.; PICCOLO, D.; VERDE, R.; VICHI, M. (Ed.). **Classification and Multivariate Analysis for Complex Data Structures**. Italy: Springer Berlin Heidelberg, 2011, (Studies in Classification, Data Analysis, and Knowledg Organization). p. 353–361.

____. Interval archetypes: a new tool for interval data analysis. **Statistical Analysis and Data Mining**, Wiley Online Library, v. 5, n. 4, p. 322–335, 2012.

DIAS, J. M. A. **A análise sedimentar e o conhecimento dos sistemas marinhos**. Universidade do Algarve, Faro, Portugal, 2004. v. 28, n. 01. Disponível em: <http://w3.ualg.pt/~jdias/JAD/eb_Sediment.html>. Acesso em: 22 jan. 2015.

Dicionário online Michaelis. **"arquétipo"**. 2015. Disponível em: <http://michaelis.uol.com.br>. Acesso em: 11 abr. 2015.

EPIFANIO, I.; VINUè, G.; ALEMANY, S. Archetypal analysis: contributions for estimating boundary cases in multivariate accommodation problem. **Computers & Industrial Engineering**, Elsevier, v. 64, n. 3, p. 757–765, 2013.

EUGSTER, M. J. A. Archetypal athletes. **arXiv preprint arXiv:1110.1972**, 2011.

____. Performance profiles based on archetypal athletes. **International Journal of Performance Analysis in Sport**, v. 12, n. 1, p. 166–187, 2012.

EUGSTER, M. J. A.; LEISCH, F. From spider-man to hero - archetypal analysis in r. **Journal of Statistical Software**, p. 1–23, 2009.

____. Weighted and robust archetypal analysis. **Computational Statistics & Data Analysis**, Elsevier, v. 55, n. 3, p. 1215–1225, 2011.

FERREIRA, D. F. **Estatística Multivariada**. 2. ed. Lavras: Ed. UFLA, 2011. 676 p. ISBN 978-85-87692-92-4.

FERREIRA, E. B.; OLIVEIRA, M. S. de. **Sensometria: uma abordagem com ênfase em Procrustes**. Santa Maria: UFSM, 2007. 71 p.

FOOTSTATS. **Estatísticas**. [S.l.], 2014. Disponível em: <http://footstats.net/campeonatos/copa-do-mundo-2014/estatisticas>. Acesso em: 28 jul. 2014.

FÉDÉRATION INTERNATIONALE DE FOOTBALL ASSOCIATION - FIFA. **Estádios de Futebol**: Recomendações e requisitos técnicos. [S.l.], 2011. Disponível em: <http://pt.fifa.com/mm/document/tournament/competition/01/37/17/76/p_sb2010_stadiumbook_ganz.pdf>. Acesso em: 07 jul. 2014.

GRAVE, G. **Les cartes graphiques**. 2014. Disponível em: <http://guy-grave.developpez.com/tutoriels/hardware/les-cartes-graphiques/pipeline-non-programmable/rasterization/>. Acesso em: 22 jan. 2015.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, v. 24, n. 6, p. 417–441, September 1933. ISSN 0022-0663.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis**. 6th. ed. New Jersey: Prentice Hall, 2007. 793 p.

JOLLIFFE, I. **Principal component analysis**. 2. ed. [S.l.]: Wiley Online Library, 2002. ISBN 0-387-95442-2.

LI, S.; WANG, P.; LOUVIERE, J.; CARSON, R. Archetypal analysis: a new way to segment markets based on extreme individuals. **ANZMAC 2003 Conference Proceedings**, p. 1674–1679, 2003.

MARTINS JÚNIOR, J. M.; CHAGAS, E. D. N.; NOGUEIRA, D. A.; FERREIRA, D. F.; FERREIRA, E. B. Avaliação monte carlo de métricas para falta de ajuste em análise de arquétipos. **Revista da Estatística da Universidade Federal de Ouro Preto**, v. 3, n. 2, p. 42–47, 2014.

____. Análise de arquétipos na avaliação da movimentação de jogadores de futebol. **Revista Brasileira de Biometria**, São Paulo, v. 33, n. 1, p. 30–41, 2015.

____. A análise de arquétipos: uma revisão bibliográfica. **Revista Brasileira de Biometria**, São Paulo, xx, n. x, p. x–xx, 2015.

Merriam-Webster online dictionary. **"archetype"**. 2015. Disponível em: <http://www.merriam-webster.com>. Acesso em: 22 jan. 2015.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada:**: uma abordagem aplicada. 1^a reimpressão. Belo Horizonte: Editora UFMG., 2007.

MORUP, M.; HANSEN, L. K. Archetypal analysis for machine learning and data mining. **Neurocomputing**, p. 54–63, 2012. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231211006060>. Acesso em: 24 fev. 2014.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 6, p. 559–572, 1901.

PEDRO, A. M. K. **Desenvolvimento do método multivariado acelerado para determinação do prazo de validade de produtos unindo Quimiometria e Cinética Química**. 2009. 169 f. Tese (Doutorado em Química) — Universidade Estadual de Campinas, Campinas, 2009.

PORZIO, G. C.; RAGOZINI, G.; VISTOCCO, D. On the use of archetypes as benchmarks. Wiley Online Library, v. 24, n. 5, p. 419–437, 2008.

PRABHAKARAN, S.; RAMAN, S.; VOGT, J. E.; ROTH, V. Automatic model selection in archetype analysis. In: **Pattern Recognition - Joint 34th DAGM and 36th OAGM Symposium, Graz, Austria, August 28-31, 2012. Proceedings**. [s.n.], 2012. p. 458–467. Disponível em: <http://dx.doi.org/10.1007/978-3-642-32717-9_46>. Acesso em: 24 fev. 2014.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Acesso em: 18 jun. 2014.

RIEDESEL, P. Archetypal analysis in marketing research: a new way of understanding consumer heterogeneity. **Action Marketing Research**, v. 24, 2014.

SANTOS, T. R. Um modelo de espaço de estados poisson para modelagem dos confrontos de futebol entre brasil e argentina. **Sigmae**, v. 2, n. 1, p. 42–47, 2013.

SEILER, C.; WOHLRABE, K. Archetypal scientists. **Journal of Informetrics**, Elsevier, v. 7, n. 2, p. 345–356, 2013.

SETH, S.; EUGSTER, M. J. A. Probabilistic archetypal analysis. **ArXiv e-prints**, 2014.

SIFA, R.; BAUCKHAGE, C. Archetypical motion: Supervised game behavior learning with archetypal analysis. In: **IEEE Conference on Computational Intelligence in Games (CIG)**. Dortmund: IEEE, 2013. p. 1–8.

SIFA, R.; BAUCKHAGE, C.; DRACHEN, A. Archetypal game recommender systems. In: **T. Seidl, M. Hassani, C. Beecks (Eds.): Proceedings of the LWA 2014 Workshops**. Aachen, Germany: KDML, IR, FGWM, 2014. Disponível em: <http://ceur-ws.org>. Acesso em: 18 out. 2014.

____. The playtime principle: Large-scale cross-games interest modeling. In: **IEEE Conference on Computational Intelligence and Games (CIG)**. Dortmund: IEEE, 2014. p. 1–8.

SOUZA, E. D. **Futebol: paixão, produto ou identidade cultural**. 2013. Trabalho de Conclusão de Curso (Lato Sensu em Mídia, Informação e Cultura) - Escola de Artes e Comunicações. Universidade de São Paulo, São Paulo.

STONE, E.; CUTLER, A. Archetypal analysis of spatio-temporal dynamics. **Physica D**, v. 90, p. 209–224, 1996.

____. Moving archetypes. **Physica D**, v. 107, p. 1–16, 1997.

STONE, E.; OLSON, B. **Archetypal Analysis of Cellular Flame Data**. Utah State University, 1999.

SUAPESQUISA. **"História do Futebol - Origens do futebol, Chegada do futebol no Brasil, Charles Miller, FIFA, Copa do Mundo bibliografia"**. [S.l.], 2014. Disponível em: <http://www.suapesquisa.com/futebol/>. Acesso em: 28 jul. 2014.

THOGERSEN, J. C.; MORUP, M.; DAMKIAER, S.; MOLIN, S.; JELSBAK, L. Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. **BMC Bioinformatics**, v. 279, n. 14, p. 1–15, 2013.

VINUè, G.; EPIFANIO, I.; ALEMANY, S. Archetypoids: a new approach to define representative archetypal data. **Computational Statistics & Data Analysis**.