COMPARAÇÃO DE MÉTODOS DE DECISÃO MULTICRITÉRIO PARA FINS DE RANQUEAMENTO UTILIZANDO DADOS TÉCNICOS DE JOGADORES DE FUTEBOL

Gabriel Daiha Alves

Projeto de Graduação apresentado ao Curso de Engenharia de Controle e Automação da Escola Politécnica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Engenheiro.

Orientador: Amit Bhaya

Rio de Janeiro

Fevereiro de 2019

COMPARAÇÃO DE MÉTODOS DE DECISÃO MULTICRITÉRIO PARA FINS DE RANQUEAMENTO UTILIZANDO DADOS TÉCNICOS DE JOGADORES DE FUTEBOL

Gabriel Daiha Alves

PROJETO DE GRADUAÇÃO SUBMETIDO AO CORPO DOCENTE DO CURSO DE ENGENHARIA DE CONTROLE E AUTOMAÇÃO DA ESCOLA POLITÉCNICA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE ENGENHEIRO DE CONTROLE E AUTOMAÇÃO

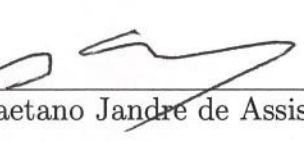
Autor:

Gabriel Daiha Alves

Orientador:

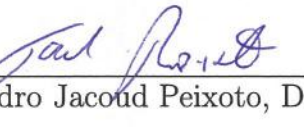
Prof. Amit Bhaya, Ph. D.

Examinador:



Prof. Frederico Caetano Jandre de Assis Tavares

Examinador:



Prof. Alessandro Jacoud Peixoto, D. Sc.

Rio de Janeiro

Fevereiro de 2019

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Escola Politécnica - Departamento de Eletrônica e de Computação
Centro de Tecnologia, bloco H, sala H-217, Cidade Universitária
Rio de Janeiro - RJ CEP 21949-900

AGRADECIMENTO

Agradeço primeiramente a minha família: minha mãe, Eliana, meu pai, Paulo e meu irmão Felipe, pelo apoio incondicional em todos os momentos, principalmente nos mais difíceis. Se não fossem vocês, eu provavelmente não estaria escrevendo esse parágrafo aqui nesse momento.

Também agradeço a minha querida amiga Júlia Queiroz, por todo o carinho e companheirismo nos quatro anos e meio dos meus cinco anos de faculdade, nos quais, desde o primeiro período, estivemos juntos em todas as situações de felicidade e também de dificuldade. Negar sua importância no processo sem essa devida homenagem, independente de nosso estágio atual de relacionamento, seria uma atitude injusta e você merece esse parágrafo especial.

Um agradecimento muito válido ao Club de Regatas Vasco da Gama, na figura do Centro de Inteligência e Análise do Vasco, que me deu a oportunidade de ter minha primeira experiência de emprego, me deu suporte para a realização desse trabalho com a disponibilidade dos dados e que, durante esses três anos e meio de permanência correntes, me ofereceu todo o apoio necessário para poder concluir a faculdade. A Pedro Monteiro e Alberto Tenan, meu eterno agradecimento por todo o ensinamento passado e por abrirem as portas para mim nesse ramo.

Um total agradecimento a Universidade Federal do Rio de Janeiro por me permitir diversas experiências que contribuíram para a minha formação como engenheiro, em especial ao professor Amit Bhaya, por todo o ensinamento passado e bom contato nesse tempo de construção desse trabalho. Não custa agradecer também ao Colégio Santo Inácio, ao CEFET e ao Colegio Marista São José por terem tido real importância na minha formação como pessoa e como cidadão.

Gostaria também de agradecer a todos meus amigos, com menção especial a Patusco, Bels, Elê, Fernanda, Doia, Raquel, Lu, May, Matheus, Cury, Fernando, Fidalgo, Queiroz, Novello, Caco, Padilha, Baiano, Raphael, Tio Gui, Elias, Leon,

Hilst, Jean, Oreiro, Ananias, Bruce e Xiaolin. Vocês, e todos os outros cujos quais cometi grande injustiça em não citar, me fazem seguir em frente e sempre estiveram ao meu lado para me manter firme nessa luta árdua pela conquista desse diploma.

E não poderia deixar de citar a minha turma T18, obrigado por cada momento de convivência nesse parto que foi a faculdade. Com menção especial a Amanda de Oliveira, querida Amy, por toda a contribuição dada durante minha estadia no curso, seja ajudando com dúvidas, seja fornecendo os cadernos mais completos e bem escritos da história do curso de Engenharia de Controle e Automação.

RESUMO

O futebol é um ambiente de decisões com um número grande de alternativas dependentes de muitos critérios. E uma das tarefas decisórias mais complexas envolve a comparação de um conjunto grande de jogadores, como por exemplo para a realização de contratação de atletas. Dentro desse contexto, esse trabalho se propõe a discutir estratégias para o ranqueamento do rendimento em campo de jogadores a partir de dados de eventos técnico-táticos ocorridos nas partidas. Serão avaliadas diversas combinações para pré-processamento dos dados, geração dos scores individuais das habilidades, aplicações de pesos para as ações analisadas e usos de métodos de decisão multicritério. O método heurístico de ponderação de critérios mostrou melhor desempenho em relação a não adoção e ao método por entropia. O método Contagem de Borda desempenhou melhor quando aplicado aos dados originais, enquanto TOPSIS performou melhor quando aplicado PCA.

Palavras-Chave: tomada de decisão multicritério, ranqueamento, futebol, ciência de dados, aprendizado não-supervisionado.

ABSTRACT

Football is a decision-making environment with a large number of alternatives depending on many criteria. And one of the most complex decision-making tasks involves comparing a large number of players, for example for hiring athletes. Within this context, this work proposes to discuss strategies for the ranking of players' in-game performance from technical-tactical events data occurred in football matches. Various combinations will be evaluated for data preprocessing, generation of individual skills scores, criteria weighting methods and multicriteria decision methods. The heuristic criteria weighting method showed better performance in relation to non-adoption and to the entropy method. The Borda Count method performed better when applied to the original data, while TOPSIS performed better when applied at PCA.

Key-words: multi-criteria decision making, ranking, soccer, data science, unsupervised learning.

SIGLAS

UFRJ - Universidade Federal do Rio de Janeiro

PCA - Principal Component Analysis

TOPSIS - Technique for Order Preference by Similarity to Ideal Solution

FIFA - Federation International Football Association

CPP - Composition of Probabilistic Preferences

LASSO - Least Absolute Shrinkage and Selection Operator

API - Application Programming Interface

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Tomar decisões com muitas alternativas e critérios é um processo que ocorre diariamente no cotidiano do ser humano. Sendo o futebol o esporte mais popular do mundo, com interesse de mais de 40% do público dos 18 maiores mercados mundiais (NIELSEN, 2018) [1], ele acaba se tornando mais visado por veículos midiáticos, fazendo com que qualquer movimento de relativa importância, como a contratação de um jogador, acabe gerando grande impacto nos seus adeptos.

Nesse contexto, tornam-se inevitáveis as comparações entre diversos atletas, seja por parte dos diretores dos clubes, seja por parte da mídia, seja por parte dos torcedores. Tendo em vista que jogadores de futebol possuem diversos tipos de características e que os mesmos realizam ações distintas durante as partidas, o ato de compará-los se torna um problema de decisão multi-critério.

Para o estudo de fenômenos desse cunho e suas respectivas variações, a ciência desenvolveu a área da Pesquisa Operacional, que lida com métodos avançados, especialmente da Matemática Aplicada, para dar base científica a gestores para tomar decisões ponderadas e eficazes. (RAMA MURTHY, 2007) [2] Tomada de decisão multi-critério é a sub-disciplina correspondente desta área que aborda estudos no âmbito do tema deste trabalho, onde alguns métodos já consolidados, como o TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) serão aplicados em diversas situações.

1.1 Motivação

O esporte vem sendo progressivamente reconhecido como um tipo de indústria, fazendo com que gestões baseadas em modernos modelos de negócio sejam cada vez mais necessárias para o seu bom funcionamento. A indústria do esporte é a esfera responsável por unir partes específicas da cadeia produtiva de um país para atender ao mercado no qual os produtos oferecidos aos compradores integram-se às mais diversas modalidades – fitness, recreação e lazer, assim como bens, serviços, pessoas, lugares ou ideias estão inseridos nesse processo (PITTS E STOTLAR, 2002) [3].

E o futebol não se difere disso. Este impacto pode ser mensurado pelo valor de mercado dos campeonatos das cinco maiores ligas nacionais do futebol internacional, apelidados de Big Five (Alemanha, Espanha, França, Inglaterra e Itália). Estes valores se aproximam a 25 bilhões de euros. A receita total das 20 equipes de maior destaque também atingiu um recorde, crescendo em 6% em relação a temporada de 2017 atingindo 7.9 bilhões de euros. (DELOITTE, 2018) [4]

No Brasil, o mercado do futebol sofre impacto das cifras bilionárias europeias. Segundo a FIFA [5], nosso país foi o que mais fez o mercado de transferências se mexer nos últimos anos. Em 2016, ocorreu a exportação de 806 jogadores e importação de 678 atletas brasileiros, girando recursos estimados em R$ 1.88 bilhão. Em 2017, os brasileiros participaram de 1755 transferências, sendo mais de um décimo do total mundial, gerando U$ 1.06 bilhão. (MANSUR, 2017) [6]

A partir deste contexto, presume-se a necessidade dos agremiações de futebol aplicarem uma gestão profissional com pessoas capacitadas e ferramentas compatíveis com essas exigências que o mercado atual mostra (KELLY, 2017) [7]. Além disso, a situação financeira das equipes sul-americanas exige novas estratégias para a montagem de seu elenco, levando em conta a dificuldade de cobrir as propostas de mercados com maiores poderios financeiros, como o europeu e o asiático. (MANZENREITER; HORNE, 2004) [8].

O principal motivo da existência de equipes esportivas profissionais está diretamente relacionado às conquistas das principais competições disputadas. Com isso, melhores premiações, mais exposição e consequentemente melhores contratos de patrocínio acabam se firmando. Portanto, a formação de uma equipe competitiva tende a ter relação direta com a captação de mais recursos, como por exemplo a ampliação de programas de sócios torcedores. (CURLEY; ROEDER, 2016) [9] . Hoje em dia, grandes clubes se transformaram em verdadeiras marcas mundiais, gerando receitas tanto no futebol quanto por fontes indiretas associadas à imagem construída pelo sucesso esportivo da equipe. (ROHDE; BREUER, 2016) [10].

Sob a ótica da gestão, uma equipe esportiva não se difere muito de outras organizações empresariais, no sentido de juntar os meios adequados para atender aos fins planejados (EDEN; ACKERMANN, 2004) [11]. Portanto, uma agremiação esportiva deve contar com os melhores recursos possíveis para alcançar suas conquistas, incluindo nisso atletas e todos os outros funcionários do clube. Em relação aos jogadores, essa lógica se evidencia pelos investimentos que os clubes campeões fazem em suas equipes, e isto tem proporcionado inúmeras conquistas (ROHDE; BREUER, 2016) [10].

A partir deste contexto, é relevante relacionar os insumos utilizados na produção esportiva em equipe e suas contribuições relativas à produção. Segundo (CARMICHAEL; THOMAS; WARD, 2000) [12], as equipes esportivas são semelhantes a outras empresas sob esta ótica, tentando fornecer um produto (vitórias e bom desempenho em campo), para isso empregando e combinando, como diferentes insumos, as habilidades dos jogadores e outros atributos dos membros do time.

Em relação à gestão estratégica das equipes, a seleção de jogadores também pode ser vista como uma atividade rotineira em qualquer modalidade esportiva. Diversos fatores quantitativos e qualitativos podem influenciar a escolha dos atletas (ALAMAR, 2013) [13]. No futebol, por exemplo, o desempenho de cada time depende de aspectos individuais e coletivos (BRADLEY et al., 2011; DRUST; ATKINSON; REILLY, 2007) [14], [15]. Desta forma, a análise dos resultados relacionado a esses aspectos do futebol tornou-se um elemento fundamental para melhorar o desempe-

nho das equipes (ARRIAZA; ZUNIGA, 2016) [16].

É sob os atributos técnicos e táticos que geralmente se concentra a análise de desempenho empregada pelos clubes (HUGHES; BARTLETT, 2002) [17], nos quais os dados de desempenho dos atletas nas partidas podem ser coletados e analisados para diferenciar padrões de jogo dos times e sobre os fundamentos da modalidade. Enquanto isso, as decisões de aquisição/venda de atletas geralmente são tomadas no âmbito político dos clubes, com a inclusão natural de fatores de negociação. como por exemplo o aspecto financeiro.

Levando em conta a premissa de que os melhores jogadores são relevantes às conquistas esportivas, é coerente afirmar que o aspecto técnico esteja entre os principais critérios na composição para avaliação do desempenho de uma equipe ou atleta (KATZENBACH; SMITH, 1993) [18]. Quando dois jogadores estão plenamente aptos para atuar em uma partida e taticamente ocupam a mesma posição no campo, faz sentido a escalação do mais eficiente (REZENDE; B. DE, 2006) [19], relacionado diretamente ao rendimento técnico do esporte. Estudos recentes mostram a preponderância da eficiência técnica sobre o desempenho físico dos jogadores (FILETTI et al., 2017) [20].

O crescimento do uso das ferramentas analíticas nos esportes vem de avanços da ciência e da disponibilidade de grandes quantidades de dados para as equipes e para o público em geral, o que gera uma oportunidade de vantagem competitiva. As organizações esportivas são capazes de identificar e aprimorar o desempenho do atleta através da aplicação de análise de dados. O conhecimento valioso é adquirido através do emprego de recursos da ciência da computação, estatística e matemática numa coleção de grandes conjuntos de dados complexos. Times como o Oakland A's, Tampa Bay Rays e San Antonio Spurs aplicaram o uso das análises, e todos os três clubes, apesar de serem considerados de pequena dimensão em seus mercados e com recursos limitados, têm obtido grande sucesso, em parte por causa dessa informação obtida (ALAMAR, 2013) [13].

Esse novo nicho de mercado também contribuiu como fonte de aprendizado e aprimoramento profissional de comentaristas, jornalistas e até dos próprios fãs do esporte (ROWE, 2011) [21]. No contexto empresarial, um conjunto de organizações surgiram prestando serviços em suporte à análise estatística ao mercado esportivo, muito devido ao multimilionário mercado de transferência de jogadores (HUTCHINS, 2016) [22]. Ainda que a contratação de um atleta não seja unicamente fundamentada nos dados de rendimento das partidas, a análise estatística pode se mostrar um atalho para reduzir a lista de potenciais atletas a serem recrutados.

Em situação semelhante nos anos 90, surgiu o conceito chamado "Moneyball", onde o foco era a abordagem analítica, totalmente baseada em evidências, denominadas de Sabermetrics, para a formação de uma equipe competitiva de beisebol. Sabermetrics é a análise empírica do beisebol, especialmente com base em estatísticas que medem a atividade no jogo (LEWIS, 2004) [23].

Levando essa ideia para os dias de hoje, ela se torna atrativa em mercados ineficientes, como o brasileiro. No caso do futebol, diversos autores defendem que o valor de mercado dos jogadores não está positivamente correlacionado à produtividade (VILAIN; KOLKOVSKY, 2016) [24], existindo uma distorção entre salário e desempenho (DEUTSCHER; BÜSCHEMANN, 2016; WEIMAR; WICKER, 2017) [25] [26]. Essa premissa age em favor do modelo Moneyball (LEWIS, 2004) [23], onde o responsável por sua aplicação no Oakland Athletics, o diretor Billy Beane, assumiu que o mercado de trabalho no beisebol subestima algumas habilidades dos atletas, no sentido de confirmar a possibilidade de identificação de jogadores com alto desempenho e reduzido valor de mercado.

Nas Ciências do Esporte, especificamente no caso do futebol, o processo de tomada de decisão apresenta basicamente duas esferas: (a) a primeira está relacionada com o processo de tomada de decisão por parte dos atletas durante os jogos e (b) a segunda sendo por parte da comissão técnica durante o planejamento da equipe na escolha de atletas para o elenco, na escolha dos jogadores para otimizar as escalações, na definição de melhores forma de treinamentos, tratamentos de lesões. (PRINCIPE, 2018) [27].

Portanto, o processo de tomada de decisão enquadra-se no segundo caso, onde a comissão técnica é responsável pela busca de soluções satisfatórias ao problema com base no compromisso, equilíbrio e variedade de pontos de vista. Segundo (POMEROL; BARBA-ROMERO, 2012a) [28], esses aspectos estão presentes em qualquer tomada de decisão, pois representam um compromisso entre necessidades que não podem ser realizadas por completo. Essas necessidades correspondem a critérios de escolha, responsáveis por juntar avaliações das diferentes alternativas. Neste contexto, a gestão da informação dentro dos clubes de futebol mostra-se útil para aplicação à ciência do esporte, mais especificamente à seleção de jogadores com base na análise de desempenho sobre o componente técnico (MACKENZIE; CUSHION, 2013) [29].

1.2 Problema da avaliação de atletas para seleção dos melhores

Mesmo em estágio primário, a criação de infraestrutura de análise de desempenho e o recrutamento de analistas nessa área têm sido observadas no futebol. Com a estrutura montada, os profissionais responsáveis por esse processo precisam de dados e modelos matemáticos eficientes. Atualmente, diversas empresas oferecem os mais variados dados de desempenho individual, mesmo que esses resultados estejam consolidados sob a forma de uma "caixa preta", ou seja muitas vezes não sabemos como os mesmos são adquiridos e processados (ANDERSON; SALLY, 2013) [30].

Com a grande competitividade no contexto do esporte profissional de alto rendimento, impõe-se que ocorra uma proteção institucional sobre o valioso conhecimento gerado com as análises criadas por esses departamentos. Os métodos eventualmente utilizados nos processos de escolha dos jogadores adquirem, assim, também este aspecto de "caixa preta".

Desta forma, a reserva de mercado também se reflete no reduzido volume de artigos científicos aplicados à montagem das equipes. Isto mostra o importante papel da academia, no que se refere à integração entre a teoria e a prática da ciência

do esporte, permitindo que o conhecimento seja promovido e aperfeiçoado em novos estudos.

A realidade financeira da maioria das equipes do futebol brasileiro é bem diferente daquelas que integram as ligas de maior valor na Europa (ANDERSON; SALLY, 2013) [30]. Os jogadores brasileiros que se destacam nos torneios pré-profissionais são rapidamente negociados ao exterior, em face da dificuldade dos clubes em cobrir as ofertas dessas ligas com maior capacidade de investimento (CAVALCANTI; CAPRARO, 2015) [31]. Dessa forma, os clubes brasileiros necessitam remontar seus elencos para as novas temporadas com pouco recurso, mostrando-se assim um problema similar ao do Oakland Athletics, retratado por (LEWIS, 2004) [23].

No nosso país, o uso da análise quantitativa para a gestão do elenco de jogadores ainda é incipiente. Os clubes de maior investimento iniciam, aos poucos, a criação de departamentos de análise de mercado. Em síntese, o problema de pesquisa está relacionado ao mercado do futebol brasileiro, que enfrenta limitações financeiras e sofre o assédio de praças esportivas com maior capacidade de investimento, necessitando, assim, aplicar novas abordagens para a escolha de jogadores para recompor seus elencos.

Conceitualmente, o Moneyball requer uma adaptação metodológica ao mercado do futebol, uma adaptaçao cultural ao futebol brasileiro e tecnológica por esta abordagem ter sido feita no contexto da MLB ao final da década de 1990. A proposta descrita neste trabalho busca apresentar uma nova abordagem para seleção de atletas de futebol.

1.3 Formalização do problema de *ranking*

O termo *sistema de rating*, no contexto de competições esportivas, refere-se a um método para atribuir uma avaliação da habilidade dos jogadores/equipes baseado no rendimento dos(as) mesmos(as) nas partidas. [32] Um dos exemplos mais conhecidos é a classificação das habilidades dos jogadores de xadrez, denominada *Rating ELO* [33].

Os sistemas de *rating* são utilizados para uma grande variedade de razões: classificação de jogadores em jogos *online*, previsão de resultados de competição para uso em serviços de apostas, ou para monetizar os jogadores em termos de suas habilidades.

Nos últimos tempos, *rating* e *ranking* foram conceitos que se tornaram extremamente importantes na área de desenvolvimento de mecanismos de pesquisa e outros tipos de serviços online, como o *Netflix*. O *Netflix Prize*, por exemplo, foi uma competição aberta, estimulada pelo serviço de *streaming* pelo melhor algoritmo de filtragem colaborativa para prever classificações (*ratings*) de usuários para filmes, com base em avaliações anteriores, sem qualquer outra informação sobre os usuários ou filmes.

Formalmente, deve-se distinguir entre as palavras *rating* e *ranking*, mesmo que sejam usadas de forma intercambiável. Um *ranking* se refere a uma lista ordenada, enquanto um *rating* se refere a uma lista de pontuações, uma para cada item. [34]

No contexto esportivo, embora o ranking de diferentes itens dê uma sensação de relação com as habilidades dos jogadores classificados, ele não dá nenhuma indicação do quão mais hábil um jogador com melhor classificação é em relação a um jogador com menor classificação.

A partir dessas definições de terminologia, fica claro que se um *rating* é atribuído a cada jogador em um determinado conjunto, então a organização dos jogadores na ordem de seus valores leva imediatamente a um *ranking*. [34]

Finalmente, se cada jogador é avaliado em várias habilidades diferentes, claramente cada um destes atributos leva a um *ranking* separado do conjunto de jogadores. Este, por sua vez, leva ao problema da agregação de *ranks*, que se refere à geração de uma única lista de classificação, o que melhor concorda com todas as diferentes listas classificadas do mesmo conjunto de jogadores.

Resumindo, os três principais conceitos são:

- Atribuir um valor, uma pontuação para cada habilidade do jogador (*rate*)
- Usar esses valores para classificar os jogadores por ordem (*rank*)
- Produzir um único *ranking* a partir das classificações de cada habilidade (*aggregate*)

1.4 Objetivos

O objetivo deste trabalho é estudar e comparar técnicas de agregação de ranqueamentos. Para atingi-lo, levando em conta o contexto dos dados técnicos de jogadores de futebol é necessário:

1. seleção dos dados (habilidades) relevantes para o trabalho proposto
2. desenvolver métodos de gerar *scores/ratings* através dos dados brutos escolhidos
3. aplicar técnicas de agregação de *scores* para produzir uma lista ordenada de atletas
4. comparar as listas agregadas por cada caminho desenvolvido
5. meta-agregar as listas no final, gerando um *ranking* geral

1.5 Organização do documento

O trabalho será dividido em 5 partes principais. No capítulo presente, foi realizada uma contextualização do problema, mostrando a necessidade da discussão do assunto, potenciais interessados no desenvolvimento de estudos desse entorno e os objetivos deste trabalho.

No segundo capítulo, será realizada uma revisão bibliográfica, mostrando trabalhos anteriores tanto na área de Pesquisa Operacional, com aplicações de métodos de decisão multi-critério em outras situações, como da área de Esporte, relacionados a indicadores de desempenho usados em comparações de rendimento de atletas

de futebol. O mesmo capítulo também aprofundará e fundamentará cada um dos conceitos matemáticos utilizados durante o trabalho. Serão explicados:

- os métodos para geração de *scores/ratings* para cada habilidade, onde se explicará os conceitos de suavização aditiva e média bayesiana.
- os processos a serem realizados previamente a agregação de *scores*, onde serão fundamentados os métodos de atribuição de pesos (por entropia e por heurística), os métodos de normalização de variáveis (Min-Max, por desvio padrão, por normalização vetorial) e o método de redução de dimensionalidade (PCA)
- as técnicas efetivas de agregação e meta-agregação de scores (TOPSIS, Contagem de Borda, Maior Média...)
- os métodos de comparação das agregações, onde se fundamentarão as diversas métricas de similaridade utilizadas, dentre elas a correlações de Spearman e a distância de Jaccard.

No terceiro capítulo, será apresentada de forma geral toda a metodologia que norteou o desenvolvimento do material, mostrando os recursos utilizados, a sequência de pesquisa adotada, os fluxos de trabalho desenvolvidos, as decisões de projeto adotadas, uma apresentação breve dos conceitos utilizados nas diversas fases do projeto e a escolha das métricas de avaliação dos resultados.

No capítulo 4, os resultados relacionados aos ranqueamentos serão apresentados e os seus respectivos desempenhos discutidos, realizando comparações entre os métodos adotados com índices de referência. Será avaliada também a sensibilidade de cada caminho desenvolvido aos parâmetros de pré-processamento e as métricas de distâncias adotadas.

No quinto e último capítulo, serão tiradas as conclusões gerais dos resultados e dos temas abordados, e também serão mostradas as limitações deste trabalho e pontos de desenvolvimento futuro.

Capítulo 2

Revisão da literatura

2.1 Estado da arte

Para analisar aplicações semelhantes anteriores, dividiu-se a revisão dos trabalhos em duas partes: a área relacionada ao estudo de indicadores de desempenho técnico no futebol e a área vinculada ao uso de métodos de ranqueamento e de tomada de decisão multi-critério no mercado.

2.1.1 Indicadores de desempenho no futebol

Segundo MAXCY & DRAYER (2014) [35] a análise estatística no esporte, que posteriormente ficou conhecida por Sabermetrics, foi pioneira e inovadora quando implantada por Bill James no beisebol que, na época, se encontrava em condições econômicas desafiadoras, onde clubes de baixo orçamento eram naturalmente desclassificados antes de chegarem às fases finais. LEWIS (2004) [23] reforçou que a situação se manteve até que Bill Beane, um ex-jogador e gerente geral do Oakland Athletics, juntamente com Paul Depodesta, um economista que nunca havia se envolvido em nenhum esporte, decidiu mudar as regras desse jogo injusto e progrediu para o uso de das técnicas do Sabermetrics.

Como o crescimento do uso das ferramentas analíticas nos esportes tem relação direta com os avanços científicos, a disponibilidade de grandes quantidades de dados para as equipes e para o público em geral proporciona uma grande quantidade de amostras que podem ser testadas. ALAMAR (2013) [13] cita que times de beisebol

como Oakland A's, Tampa Bay Rays e San Antonio Spurs passaram a adotar o uso da análise, e todos os três clubes, apesar de considerados pequenos em seus mercados e com recursos limitados, obtiveram grande sucesso, em parte devido à informação obtida pela as análises de dados.

LEWIS (2004) [23] também afirmou que embora a análise estatística tenha sido amplamente adotada por importantes tomadores de decisão e jornalistas no futebol americano, basquete e beisebol, sua utilização no futebol permanece um pouco limitada. TENGA et al (2009) [36] citam que um fator que contribui para isso é a natureza do esporte em si. Enquanto um jogo de beisebol é constituído de centenas de interações discretas e algumas dezenas de jogadas defensivas, um jogo de futebol pode conter milhares de eventos diferentes.

DUCH, WAITZMAN & NUNES AMARAL (2010) [37] afirmam que o futebol é um dos esportes mais difíceis de analisar quantitativamente devido à sua complexidade e fluxo quase ininterrupto da bola, comparado ao basquete ou ao beisebol. Estatísticas simples, como número de passes corretos, número de chutes, número de assistências ou número de gols dificilmente fornecem sozinhas uma medida confiável que permita diferenciar um jogador de futebol de outro.

A função da análise de desempenho como a ferramenta de investigação no esporte serve para desenvolver uma compreensão para um esporte em questão, melhorar o desempenho e também auxiliar o processo de tomada de decisão (HUGHES & BARTLETT, 2002; O'DONOGHUE, 2010) [17] [38].

No início, o foco da análise no futebol era sobre a performance física, enquanto outras pesquisas menos frequentemente se concentravam em analisar o desempenho técnico e tático dos times (ERMANNO RAMPININI et al., 2009; RUSSEL, REES & KINGSLEY, 2013) [39] [40]. Num momento seguinte, o foco principal passou a ser o desenvolvimento e criação de indicadores para medir desempenho das equipes. (CARLING et al., 2009, 2005; HUGHES & BARTLETT, 2002) [41] [17].

Útil do ponto de vista estratégico e tático, a análise estatística do desempenho fornece indicadores que se referem às variáveis de ação que definem um aspecto de

desempenho (PRZEDNOWEK et al.,2017) [42]. Atualmente, existem várias técnicas utilizadas para realizar a captura dos indicadores de desempenho de cada atleta. O Scout notacional é uma das mais adotadas, onde diversas ações técnicas e táticas da partida são estratificadas, como finalizações, passes longos, dribles, desarmes. Para cada um desses lances, aspectos relativos ao evento são classificados, tais como: quem realiza, local de realização da ação, consequência final, dentre outras. (SILVA, 2007) [43]. A empresa russa InStat Football e a italiana WyScout, são as principais do ramo que prestam este serviço. (VAN HOEVE, 2017)

Segundo ROBERTSON, BACK & BARTLETT (2016) [44], estas ferramentas possibilitam a coleta de dados de forma substancial para que a equipe de analistas de desempenho possam conduzir as análises relevantes e dar suporte ao processo de tomada de decisão do treinador no jogo com base nos indicadores de desempenho. Portanto, o número de variáveis que podem ser coletadas dos atletas usando as tecnologias contemporâneas é suficientemente extenso, sendo assim importante entender as principais variáveis de desempenho que contribuem para o sucesso de um jogador.

Dentro deste contexto, como descrito por WEINECK (1997) [45], em esportes competitivos, o nível é determinado por um conjunto complexo de variáveis (local da ação, momento da partida) que influenciam os indicadores de desempenho individuais (passes, dribles, chutes) em termos de aspectos técnicos e táticos. Estes indicadores podem ser entendidos como um conjunto de habilidades dos atletas (técnica individual) por cada posição no campo de jogo.

Dentro dessa ideia, KUMAR (2013) [46] realizou um estudo onde, a partir de ranqueamentos de jogadores de futebol feitos por índices especialistas, extraiu quais variáveis foram, por estes índices, consideradas as mais importantes para cada uma das posições analisadas, utilizando-se de métodos de aprendizado de máquina como Regressão Linear Múltipla, Regressão por discretização utilizando J48, Perceptron Multi-camadas, dentre outras. Para zagueiros, por exemplo, o número de falhas que causam gols contra a própria equipe foi a variável de maior relevância.

Nesse caminho, KASAP (1997) [47] propôs uma modelagem de banco de dados com um conjunto de ações especificado para cada posição, com objetivo de poder avaliar o rendimento técnico de jogadores e de realizar rankeamentos de atletas para contratação.

STANOJEVIC & GYARMATI (2016) [48] apresentaram em seu estudo uma proposta de metodologia para estimar o valor de mercado atual de um atleta a partir de seus dados de performance recentes, utilizando-se de Regressão Linear Múltipla. Assim como o trabalho vigente, ele extraiu dados técnicos de jogadores da plataforma InStat Football, como podemos ver na Tabela 2.1.

Fatores analisados (tipo de dado)
Jogos realizados (total)
Minutos jogados (total)
Gols (total)
Assistências (média por jogo e aproveitamento)
Passes (média por jogo e aproveitamento)
Confrontos (média por jogo e aproveitamento)
Duelos aéreos (média por jogo e aproveitamento)
Passes chaves (média por jogo e aproveitamento)
Finalizações (média por jogo e aproveitamento)
Dribles (média por jogo e aproveitamento)
Desarmes (média por jogo e aproveitamento)
Cruzamentos (média por jogo e aproveitamento)

Tabela 2.1: Variáveis utilizadas por STANOJEVIC & GYARMATI (2016)

Nessa mesma abordagem, HE, CACHUCHO E KNOBBE (2015) [49], procuraram estimar o valor de mercado de atacantes, dessa vez utilizando Regressão LASSO. Eles concluíram que um bom atacante deve ter bons índices em chutes e gols na pequena área, chutes na direção da baliza, faltas e pênaltis sofridos, dribles concluídos com sucesso e total de assistências.

2.1.2 Tomada de decisão multi-critério

2.1.2.1 Aplicações no futebol

Alguns estudos foram produzidos na área de tomada de decisão multi-critério aplicado ao futebol. PRÍNCIPE et al. (2017) [50] comparou o desempenho dos métodos CPP (Composition of Probabilistic Preferences), Fuzzy Multimoora e Fuzzy Vikor, utilizando-se de dados coletivos das equipes da Premier League da temporada 2015/2016, para rankeá-las e tentar prever a colocação final das mesmas. SANT'ANNA et al. (2010) [51] se utilizou do método CPP com mesmo propósito, porém se utilizando apenas dos resultados dos confrontos entre as equipes.

De forma semelhante, KIANI MAVI, KIANI MAVI & KIANI (2012) [52] estudaram a eficiência do uso dos métodos AHP e TOPSIS para analisar a performance das equipes que disputaram a Bundesliga (Liga alemã) na temporada de 1999/2000. PAPPALARDO & CINTIA (2017) [53] também fizeram estudo semelhante, avaliando as variáveis relevantes para se caracterizar o desempenho de uma equipe, rankeando as mesmas a partir desses dados pegando a média do valor das features utilizadas e vendo a equipe de maior média.

Porém, como podemos ver, a abordagem adotada nesses estudos citados foi totalmente voltada para dados de cunho coletivo, fugindo do âmbito individual, proposta deste trabalho.Foram poucos os estudos para o desenvolvimento de técnicas de apoio a decisão voltados para a avaliação técnica e seleção de atletas.

PAPPALARDO et al. (2018) [54] realizou um estudo próximo ao que será feito nesse trabalho, extraindo um conjunto de dados complexos sobre jogadores das ligas inglesa, alemã, espanhola, francesa e italiana, modelando a performance dos atletas na temporada 2015/2016 e rankeando o desempenho dos mesmos através do método PlayeRank. As features receberam pesos a partir do método supervisionado Support Vector Machines, conforme o quão mais elas contribuem para gols da equipe. Os atletas foram divididos em funções a partir do quão mais presentes eles ficam em uma determinada área do campo. Os resultados mostraram-se consistentes com rankings de referência, como por exemplo, o ranking PSV.

Já PINHO DE SÁ (2016) [55] propôs um sistema, baseado em lógica fuzzy, utilizando-se de dados coletados a partir do site WhoScored, para avaliaçao técnica de jogadores de futebol. ARAÚJO (2017) [56] continuou o trabalho, aprimorando a modelagem fuzzy proposta por PINHO DE SÁ (2016) [55] e avaliando seus resultados. Ambos os estudos apresentaram resultados relevantes, apesar do baixo número de variáveis utilizadas, como podemos ver na Tabela 2.2.

Dimensão técnica	**Variáveis consideradas em cada dimensão**
Finalização	gol / certas / erradas / bloqueadas
Passe Curto	assistencias / passes chaves / certos / errados/
Passe Longo	passes chaves longos / certos / errados
Controle de Bola	dribles certos / dribles errados / perdas de bola
Marcação	faltas cometidas / interceptacoes / bloqueios / desarmes
Disciplina	cartões amarelos / cartões vermelhos / faltas cometidas
Jogo Aéreo	duelos aéreos vencidos / perdidos / gols de cabeça / finalizações de cabeça
Bola Parada	gol de bola parada / passe chave de escanteio / passe chave de falta/ passe chave de lateral

Tabela 2.2: Variáveis utilizadas tanto por PINHO DE SÁ (2016) como por ARAÚJO (2017) no sistema fuzzy proposto.
Fonte: http://monografias.poli.ufrj.br/monografias/monopoli10022054.pdf

BROOKS, KERR & GUTLAG (2016) [57] propuseram o rankeamento de jogadores de futebol a partir do valor de seus passes completados. A partir de um algoritmo supervisionado de aprendizado de máquina (Support Vector Machines), eles atribuiram pesos a cada passe realizado por cada jogador da Liga Espanhola na temporada 2012/2013 conforme o quão mais importante esse passe foi para se gerar uma finalização posterior. Eles utilizaram variáveis como local e destino do passe, se foi concretizada de forma positiva a transferência de bola, dentre outras features.

MU (2016) questionou a premiação da bola de ouro ao melhor jogador da Copa do Mundo de 2014, propondo o uso do método AHP para rankear atletas a partir de variáveis de cunho técnico como critérios, sugerindo o prêmio a James Rodriguez, meia colombiano ao invés de Lionel Messi, meia-atacante argentino.

QADER (2017) [58] se utilizou do método TOPSIS para rankear jogadores de futebol, porém se utilizando de variáveis estritamente de cunho físico. Foi a única aplicação encontrada do método de ranqueamento a ser utilizado no trabalho vigente quanto a classificação de jogadores de futebol, o que evidencia a necessidade de um estudo do mesmo nessa área.

2.1.2.2 Aplicações em esportes em geral

Em outros esportes, muitos estudos já foram realizados se utilizando de métodos de decisão multi-critério, seja para ranquear equipes, seja para ranquear jogadores. DADELO et al. (2014) [59] sugeriram o uso do método TOPSIS integrado a opiniões de especialistas para atribuir pesos aos critérios a serem levados em conta na seleção de jogadores de basquete, num framework próximo ao que será adotado neste trabalho.

NIKJO, REZAEIAN & JAVADIAN (2015) [60] propuseram abordagem simular, com o uso de uma composição dos métodos AHP (para atribuir pesos aos critérios) e Extended-TOPSIS para aplicar pesos aos tomadores de decisão e ao ranking de alternativas para a contratação de atletas, generalizando para qualquer esporte. Também utilizando-se de AHP e TOPSIS, CHEN, LEE & TSAI (2014) [61] propuseram um framework para escalar jogadores de baseball antes de um jogo.

BALLI & KORUKOGLU (2012) [62] adotaram uma estratégia semelhante, porém utilizando-se de critérios fuzzy. Da mesma maneira, eles adotaram o método AHP para definição de pesos aos critérios e do método TOPSIS para ranqueamento final das decisões. Também utilizando-se de variáveis fuzzy, TAVANA et al. (2013) implementou um sistema de inferências para definir a escalação e a formação de uma equipe antes de uma partida, tendo abordagem de selecionar os melhores jogadores para um fim diferente, no caso, pré-jogo. Ao final eles avaliaram a correlação dos resultados com a opinião de especialistas na área.

No mesmo tema, AGILONU & BALLI (2009) [63] propuseram um método de decisão multi-critério com atributos fuzzy, misturando a abordagem do método AHP com o método MinMax para selecionar atletas de badminton.

BHARATAAN & ABHIJEET (2015) [64] propuseram um modelo auto-adaptável para selecionar atletas de cricket, aonde eles reduziam o conjunto de variáveis a partir de PCA (Principal Component Analysis), aplicaram regressão logística para avaliar quais delas impactaram mais para os possíveis resultados finais da partida e aplicaram programação inteira para ranquear os jogadores segundo sua utilidade (o quão ele impacta em vitórias de sua equipe).

FRY, LUNDBERG & OHLMANN (2007) [65] modelaram o processo de draft de atletas, comumente utilizado nos esportes americanos, levando como princípio que uma equipe contrata jogadores levando em conta o preço do atleta, o preço dos atletas concorrentes e a necessidade da mesma naquela posição e propuseram um método heurístico utilizando-se tanto de programação dinâmica estocástica, como determinística para auxiliar a tomada de decisão de contratações, simulando cenários onde a equipe possui pleno conhecimento das características dos atletas alvos e onde há desconhecimento das mesmas.

Para a definição de critérios relevantes para escolher jogadores de basquete, o estudo de COOPER, RUIZ & SIRVENT (2009) [66] teve foco em utilizar-se do método Data Envelopment Analysis para conseguir chegar a features com pesos não-nulos na comparação. Eles usaram-se de dados técnicos da liga espanhola na temporada 2008.

2.1.3 Aplicações em outros mercados

Métodos de tomada de decisão multi-critério inicialmente foram elaborados para atuar na área de gestão, tendo em vista a natureza desse trabalho em ter que decidir por alternativas de forma constante e frequente no dia-a-dia. SOLTANI, HEWAGE, HEZA & SADIQ (2015) [67] por exemplo, compararam o desempenho dos métodos ELECTRE, PROMETHEÉ, TOPSIS, AHP, MAUT, e uma combinação dos 5 métodos em diversas características para auxiliar na decisão de gerir os resíduos produzidos por uma cidade.

Já TONG, WANG & CHEN (2005) [68] aplicaram o método PCA-TOPSIS utilizando-se de pesos estabelecidos por especialistas na escolha dos componentes químicos

certos para otimizar o polimento químico-mecânico de filmes finos de cobre de uma indústria produtora de circuitos integrados em Taiwan.

DONDAPATI (2016) [69] utilizou-se dos métodos TOPSIS e PCA-TOPSIS para escolher a melhor opção entre materiais para serem utilizados em ferramentas de corte na indústria. Eles compararam o desempenho dos dois métodos, percebendo que o último teve melhores resultados. Em nosso trabalho, também realizaremos essa comparação.

LI et al. (2011) [70] utilizou-se do método TOPSIS adotando pesos pelo método da entropia para avaliar minas de carvão. Ele mostrou que apesar de existirem métodos mais complexos e robustos que envolvam a adoção de redes neurais, por exemplo, a facilidade e a simplicidade do método TOPSIS em conjunto com a independência de subjetividade oriunda do método de pesos por entropia se mostra mais viável para adoção na prática.

LIANG, LIU & LI (2017) [71] utilizaram do método PCA-TOPSIS com adoção dos pesos por entropia para comparar a capacidade de sustentabilidade de paisagens naturais chinesas. Eles afirmaram o bom desempenho do método, principalmente no ponto da redução de complexidade das variáveis somado ao baixo tempo de computação dos resultados e precisão dos mesmos.

DUAN, ZANG & NI (2015) [72] aplicaram a abordagem TOPSIS com uso de pesos por entropia em comparação com a pura soma das componentes principais originadas do PCA para comparar e ranquear revistas acadêmicas. Eles concluíram que apesar da redução da subjetividade originada a partir dos pesos por entropia, nem sempre eles podem estar refletindo a realidade. A comparação com rankings de referência foi o principal argumento para essa conclusão.

2.2 Fundamentação teórica

2.2.1 Definição dos termos a serem utilizados no trabalho

Para facilitar a compreensão do trabalho, alguns termos serão definidos anteriormente nessa subseção para não haver confusões. São eles:

- Variável/dado: coluna original no conjunto de dados extraído da API. Ex: passes curtos para a direita - certos.
- Habilidade: critério utilizado na hora de ranquear os jogadores. Será sempre representado pela letra h nas equações. A agregação de scores gerará um valor escalar único para cada habilidade. Ex: Chance de gol: 0.756
- Jogador: é a unidade de alternativa do processo decisório. Será representado sempre pela letra j nas equações.
- Posição: é a unidade de conjunto de alternativas do processo decisório. Será sempre representado pela letra p Jogadores de uma posição só podem ser comparados com jogadores da mesma posição.
- Matriz de decisão: matriz que possui alternativas (jogadores) nas linhas e critérios (habilidades) nas colunas, com cada célula correspondendo a um score de um jogador j na habilidade h
- Matriz de pesos: matriz que possui as posições disponíveis nas linhas e critérios (habilidades) nas colunas, com cada célula correspondendo a o peso de da habilidade h na comparação dos jogadores da posição p

Nessa seção, cada um dos métodos matemáticos utilizados no decorrer do trabalho terão suas teorias fundamentadas, com desenvolvimento de intuição e cálculos matemáticos consequentes.

Eles foram divididos conforme a aplicação no trabalho:

1. Métodos de geração de ratings para as habilidades: onde se pega os dados brutos coletados de acerto/erro de cada habilidade e agrega-se em um único escalar para cada jogador em cada critério.

2. Métodos de ponderação de critérios: onde, a partir dos ratings individuais gerados, gera-se um peso para cada um dos critérios/habilidades para cada grupo de jogadores comparado

3. Métodos de normalização de variáveis: processo de tratamento dos valores de forma a deixá-los com um determinado padrão mais regular.

4. Método de redução de dimensionalidade (PCA): onde, a partir da tabela de ratings individuais gerados, extrai-se uma nova tabela com um novo conjunto de colunas, menor que o anterior, que permita ter o mínimo de perda de informação possível.

5. Métodos de agregação de scores: onde, a partir da tabela que contenha os jogadores e os respectivos critérios a serem analisados para se ranquear (seja reduzida ou não), procura-se atribuir um valor a cada atleta, de forma a poder compará-los numa lista ordenada.

6. Métricas de avaliação de resultados: para poder comparar com um ranqueamento de referência, de forma a validar o trabalho.

2.2.2 Métodos de geração de ratings para as habilidades

Dois métodos de geração de ratings serão utilizados:

- Suavização aditiva
- Média bayesiana

2.2.2.1 Suavização aditiva

Motivação do uso

Uma proposta pensada para a geração dos índices individuais para as habilidades seria a de se obter um valor para cada critério que permitisse predizer qual dentre os atletas comparados teria maior probabilidade de acerto num futuro evento correspondente a aquela habilidade, caso esse atleta fosse estimulado pelo jogo para tal.

Grande parte das habilidades que foram selecionadas possuem tanto dados correspondentes ao número de acertos, como dados correspondentes ao número de tentativas, sendo assim de natureza semelhante a da variável aleatória *Bernoulli*. Quando se fala em aproveitamento desse tipo de variável, naturalmente se remete ao valor de porcentagem, calculada pela razão entre o número de acertos e o número de tentativas:

$$\%_{jh} = \frac{\sum_{m=1}^{n_j} I^{+}_{mjh}}{\sum_{m=1}^{n_j} I^{+}_{mjh} + \sum_{m=1}^{n_j} I^{-}_{mjh}}, \tag{2.1}$$

sendo:

- $\%_{jh}$ = percentagem de acerto da habilidade h do jogador j
- n_j = número de jogos do jogador j em questão,
- I^{+}_{mjh} = acertos da habilidade h do atleta j na partida m,
- I^{-}_{mjh} = erros da habilidade h do atleta j na partida m.

Porém, como reforçado no parágrafo anterior, nem todas as variáveis são de natureza bernoulli, como por exemplo gols feitos, assistências, cartões amarelos, etc. Essas últimas são ações que possuem cunho estritamente positivo ou negativo, sendo assim, "impossível"de se obter uma relação entre acertos e tentativas.

Para que elas pudessem entrar na comparação, calculamos a probabilidade de, dado uma partida ocorrida, o atleta conseguir sair dela concretizando pelo menos uma vez a ação em questão. Para tal, os seguintes procedimentos foram utilizados:

- Duplicou-se a coluna da variável em questão
- Na coluna duplicada, substituiu-se os valores de zero (correspondentes a partidas que o jogador não realizou a determinada ação) por um.
- Somou-se o valor de cada coluna
- Dividiu-se o valor da primeira coluna pela duplicada modificada.

Para variáveis de cunho positivo, esse valor foi colocado como o rating do critério. Para variáveis de cunho negativo, o valor complementar foi colocado, para se ter a probabilidade do jogador sair de uma partida sem realizar a ação de cunho negativo em questão. Logo:

$$\%^{+}_{jh} = \frac{\sum_{m=1}^{n_j} I_{hjm}}{\sum_{m=1}^{n_j} I_{hjm} + \sum_{m=1}^{n_j} (I_{mjh} = 0)} \tag{2.2}$$

$$\%^{-}_{jh} = 1 - \%^{+}_{jh} \tag{2.3}$$

sendo $(I_{mjh} = 0)$ igual a 1 quando a afirmação for verdadeira e zero quando for falsa.

O problema é que a porcentagem por si só é uma métrica com alta sensibilidade a valores muito baixos de tentativas. Vejamos os casos apresentados na tabela 2.3, numa eventual simulação de um conjunto de 5 jogadores sendo comparados em um critério:

player_name	**Acertos**	**Erros**	**Total**	**% (sem suavização)**	**Número de jogos**
Jogador 1	1	0	1	100 %	1
Jogador 2	12	26	38	31.5 %	10
Jogador 3	15	0	15	100 %	9
Jogador 4	29	1	30	96.6 %	15
Jogador 5	2	1	3	66.6 %	2

Tabela 2.3: Simulação de uma comparação de 5 jogadores em uma habilidade

Podemos ver que o Jogador 1 realizou uma ação e acertou a mesma, tendo um índice de 100%, enquanto o Jogador 3 realizou 15 ações e acertou as 15, tendo o mesmo índice de 100%. Já o jogador 4 realizou 30 ações e acertou 29, tendo percentagem de 96.6%. Segundo o critério da porcentagem, o Jogador 1 e jogador 3 tiveram rendimentos equivalentes, e o jogador 4 esteve um pouco abaixo dos dois.

Só que o jogador 1 realizou muito menos tentativas que o 3, que realizou ainda bem menos tentativas que o 4. Levando em conta que quanto mais experimentos forem realizados, melhor o conhecimento sobre o sistema se tem, a estatística de

porcentagem, devido a sensibilidade evidenciada para valores baixos de experimentos, é problemática para a proposta de se ter um indicador que meça a probabilidade do atleta acertar/realizar determinada ação caso ele seja estimulado novamente a fazê-la pelo jogo.

Para evitar essa situação, a suavização aditiva, ou suavização de Laplace, entra como uma proposta de solução. Ela adiciona uma quantia pequena α de observações extras ao experimento de forma a evitar a existência de eventos com probabilidade zero e suavizar valores muito baixos de tentativas, fazendo com que o valor esperado da probabilidade a posteriori fique igual a:

$$Psmooth_{jh} = \frac{x_{jh} + \alpha}{N_{jh} + \alpha K} \tag{2.4}$$

onde:

- x_{jh} = número de acertos do jogador j na habilidade h
- α = número de observações adicionais a serem colocadas para suavizar
- N_{jh} = número de tentativas do jogador j na habilidade h
- K = número de possibilidades

Avaliando a fórmula 2.4, pode-se perceber que:

- para $x_{jh} >> \alpha$ e $N_{jh} >> \alpha K$;

$$Psmooth_{jh} \to \frac{x_{jh}}{N_{jh}} \tag{2.5}$$

- para $x_{jh}, N \to 0$;

$$Psmooth_{jh} \to \frac{1}{K} \tag{2.6}$$

Logo, para valores muito altos de acertos (x_jh) e de tentativas (N), o valor esperado da equação 2.4 tenderá ao resultado original da razão, enquanto para valores baixos de ambos, o mesmo valor esperado tenderá ao valor da função densidade uniforme ($\frac{1}{K}$), reduzindo-se assim, teoricamente, os efeitos negativos da estatística da porcentagem.

Na tabela 2.4, pode-se ver o efeito da suavização aditiva aplicado nos mesmos valores da tabela 2.3, com a respectiva habilidade analisada tendo seu valor de alpha = 2:

player_name	**Acertos**	**Erros**	**Total**	**%** **(sem suavização)**	**% com suavização** **(α = 2)**
Jogador 1	1	0	1	1	0.6
Jogador 2	12	26	38	0.315	0.333
Jogador 3	15	0	15	1	0.894
Jogador 4	29	1	30	0.966	0.912
Jogador 5	2	1	3	0.666	0.571

Tabela 2.4: Comparação dos valores da tabela 2.3 com os valores suavizados (última coluna da tabela acima)

Observando-se o resultado, percebe-se que os resultados suavizados correspondem mais a intuição. O jogador 1, por exemplo, que realizou um acerto em uma tentativa feita, teve seu índice bastante suavizado, atingindo o valor de 0.6. Comparando o mesmo ao Jogador 3, que possui mesma porcentagem, pode-se ver que ele ficou com índice bem abaixo, mostrando a necessidade de se obter mais observações do mesmo para se registrar bom rendimento nessa ação.

Quanto a comparação entre os jogadores 3 e 4, pode-se perceber que, mesmo com uma porcentagem original menor, o jogador 4, por ter um número duas vezes maior de tentativas e um índice quase tão alto quanto o jogador 3, obteve melhor índice suavizado. A suavização aditiva, além de resolver probabilidades nulas e de evitar que um número de tentativas baixo seja magnificado, ela potencializa altas relações de acerto para valores maiores de tentativas, sendo assim, justo por premiar regularidade em alto nível.

Fundamentação matemática

Realizando a leitura em BROWN (1986) [73]; CASELLA & BERGER (2001) [74]; DIACONIS & YVILSAKER (1979) [75], pôde-se encontrar uma prova para a

equação 2.4. Do ponto de vista bayesiano, ela é equivalente ao valor esperado da função de distribuição a posteriori, utilizando-se como função conjugada a priori da distirbuição multinomial, a distribuição de Dirichlet simétrica com parâmetro α.

No caso do trabalho em vigor, como o número de opções possíveis para cada ação é 2 (certo/errado,realizar/não realizar...), esse processo é equivalente a se utilizar uma distribuição Beta como uma prior conjugada de uma distribuição Binomial. O parâmetro α da distribuição (também chamado de pseudoconta) pode ter qualquer valor finito não negativo. A magnitude de α depende de um conhecimento prévio acerca do problema.

Levando em conta um conjunto de dados x, correspondente a um conjunto de variáveis independentes e identicamente distribuídas relativos a frequência de ocorrência de K possibilidades distintas. Logo, o número de tentativas se equivale a $N = \sum_{i=1}^{K} x_i$. Assume-se que x é extraído de uma distribuição desconhecida π, na qual se considerará, por conhecimento prévio, uma função de distribuição de Dirichlet (Dir(α) definida no espaço K-simplex ($\sum_{k=1}^{K} \pi_k = 1$).

A probabilidade a posteriori de π dado α e o conjunto de dados x é:

$$p(\pi|x, \alpha) = p(x|\pi)p(\pi|\alpha) \tag{2.7}$$

A função de likelihood, p(x|π) é de distribuição multinomial. Logo, sua função densidade de probabilidade é calculada por:

$$p(x|\pi) = \frac{N!}{x_1! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k} \tag{2.8}$$

Já a função $p(\pi|\alpha)$ é dada por:

$$p(\pi|\alpha) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} \pi_i^{\alpha-1} \tag{2.9}$$

onde:

$$\mathrm{B}(\alpha) = \frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)}. \tag{2.10}$$

Considerando-se que são K i.i.d possíveis observações de caráter multinomial, pode-se multiplicar as equações 2.8 e 2.9, obtendo:

$$p(\pi|\alpha, x) = p(x|\pi)p(\pi|\alpha) \propto \prod_{i=1}^{K} \pi_i^{x_i+\alpha-1}. \tag{2.11}$$

Como pôde-se ver na equação 2.11, a probabilidade a posteriori também tem a forma de uma distribuição de Dirichlet. Logo, podemos calcular a média posterior por:

$$E[\pi_i|\alpha, x] = \frac{x_i + \alpha}{N + K\alpha}, \tag{2.12}$$

que é equivalente a fórmula da suavização aditiva, encontrada na equação 2.4.

2.2.2.2 Média bayesiana

Uma outra interpretação matemática para a equação 2.4 está relacionada ao conceito de média bayesiana. Média bayesiana é um método de estimativa da média de uma população usando informação externa, especialmente uma crença pré-existente, que é levada em conta e pesada no cálculo em conjunto com a média amostral. Pode-se visualizar isso na equação 2.13, encontrada em YANG (2013) [76]:

$$\overline{x_{jh}} = \frac{x_{jh} + CAv_{hk}}{N_{jh} + C}, \tag{2.13}$$

onde:

- x_{jh} = acertos do jogador j na habilidade h
- Av_{hk} = média de acerto a priori da posição k na habilidade h
- C = constante de peso para a média a priori
- N_{jh} = número de tentativas do jogador j na habilidade h

O cálculo da média bayesiana usa a média a priori Av_{hk} e uma constante C. A essa constante é atribuído um valor que é proporcional ao tamanho do conjunto de dados. O valor é maior quando a variância esperada entre conjuntos de dados

(dentro da população maior) é baixa, e menor quando se espera que os conjuntos de dados variem substancialmente um do outro.

Fazendo um paralelo com a suavização aditiva, é o caso especial onde se tem algum conhecimento sobre a média dos dados em questão. Para tal, considera-se $C = \alpha K$, e no numerador, substitui-se o valor da média da função densidade uniforme $\frac{1}{K}$ pela média a priori m. Logo:

$$\frac{x_{jh} + \frac{1}{K}\alpha K}{N_{jh} + K\alpha} \equiv \frac{x_{jh} + Cm}{N_{jh} + C}, \tag{2.14}$$

2.2.3 Métodos de ponderação de critérios

Dois métodos de ponderação de critérios, além da não-adoção, serão utilizados durante o trabalho:

- Pesos por entropia
- Pesos heurísticos

2.2.3.1 Pesos por entropia

O método de entropia foi desenvolvido como um método objetivo de alocar pesos a partir dos dados da matriz de decisão sem afetar a preferência do tomador de decisão (ZELENY; 1982) [77]. A importância relativa (w_j) do critério j (habilidade) em uma situação de decisão está diretamente relacionada à quantidade de informação fornecida pelo conjunto intrínseco de alternativas em relação àquele critério (POMEROL; BARBA ROMERO, 1997) [28].

Quanto maior a diversidade nas avaliações das alternativas (no nosso caso os jogadores), maior a importância que o critério (habilidade) deve ter. Esta diversidade é conceitualmente baseada no conceito sólido e aceito de entropia em um canal de informação colocado por Claude Shannon (SHANNON; WEAVER, 1949) [78].

O método é aplicado da seguinte forma:

- Considera-se um score de um jogador j de uma determinada posição p em uma habilidade h como x_{hj}^{p}

- Estabelecer a matriz de habilidades h dos jogadores j de uma posição p avaliada de forma que ela fique com todos os seus elementos positivos. Logo:

$$\overline{x_{hj}^{p}} = x_{hj}^{p} - (\min_{j} x_{hj}^{p} - 1)(1 < h < c) \tag{2.15}$$

- Normalização da matriz pelo método da soma, onde a é o número de alternativas (jogadores j da posição) e c número de critérios (habilidades h):

$$X_{hj}^{p} = \frac{\overline{x_{hj}^{p}}}{\sum_{j=1}^{a} \overline{x_{hj}^{p}}}(1 < h < c) \tag{2.16}$$

- Calcula-se a entropia da matriz normalizada:

$$e_{h}^{p} = -\frac{1}{\ln a} \sum_{j=1}^{a} X_{hj}^{p} \ln X_{hj}^{p} (1 < h < c) \tag{2.17}$$

- A partir da entropia calculada, extrai-se então os pesos w_h^p de cada habilidade h para cada posição p:

$$g_{h}^{p} = 1 - eh^{p} (1 < h < c) \tag{2.18}$$

$$w_{h}^{p} = \frac{g_{h}^{p}}{\sum_{h=1}^{c} g_{h}^{p}}(1 < h < c) \tag{2.19}$$

2.2.3.2 Pesos heurísticos

Motivação

A ideia de se desenvolver uma heurística para a definição de pesos para os critérios (habilidades) surgiu durante o processo da pesquisa. Além da já citada necessidade de se desenvolver métodos independentes de subjetividade, o fato de que o método de pesos por entropia explora principalmente a característica de desordem nos critérios mostra a necessidade de procurar outros métodos que explorem outros aspectos dos dados.

Lógica adotada

O peso de cada critério h (habilidades) será aplicado com valores diferentes para cada posição p, tendo em vista que posições diferentes são estimuladas a realizar ações distintas durante os jogos. O valor final do peso de uma habilidade h para uma posição p é proporcional ao produto de três fatores diferentes:

$$w_{hp} \propto D_{nat}^{hp}.F_{nat}^{hp}.F_{esp}^{hp} \tag{2.20}$$

1. Peso de acordo a dificuldade natural da habilidade h para a posição p (D_{nat}^{hp})
2. Peso de acordo a frequência natural da habilidade h para a posição p (F_{nat}^{hp})
3. Peso de acordo a frequência da habilidade h para a posição p em comparação as outras posições (F_{esp}^{hp})

onde as seguintes regras, oriundas da prática, foram seguidas:

1. Ações de maior dificuldade natural para a posição possuem maior peso. Jogadores que possuem bons índices em ações que os atletas de sua posição em geral têm índices baixos de acerto/realização devem ser valorizados. Ex: chance de gol, passes chaves...
2. Ações de menor frequência natural na partida possuem maior peso. Tendo em vista que existem situações que ocorrem mais do que outras porque o jogo exige, por sua natureza, que elas venham a ocorrer, logo, elas devem ter um menor peso em relação as mais raras. Ex: passe para o lado é um evento muito mais comum numa partida de futebol do que uma finalização.
3. Ações de maior frequência comparada às outras posições possuem maior peso. Se a partida exige mais uma determinada habilidade técnica de um jogador de uma posição em relação a um atleta de outra posição, logo essa ação deve ser mais pesada para a primeira comparado a segunda. Ex: duelos aéreos é uma ação muito mais exigida para zagueiros do que para pontas.

Cálculo de D_{nat}^{hp}

O cálculo do índice de dificuldade natural da habilidade h para a posição p procedeu-se da seguinte maneira:

1. Agregou-se os valores das habilidades h por posição p pelo método da suavização aditiva, sendo:

$$P_{smooth}^{hp} = \frac{x_{hp} + \alpha}{N_{hp} + \alpha K} \tag{2.21}$$

onde:

- P_{smooth}^{hp} = probabilidade de um atleta da posição p acertar um evento correspondente da habilidade h numa nova tentativa
- x_{hp} = número de acertos do evento correspondente da habilidade h feito por jogadores da posição p
- N_{hp} = número de tentativas do evento correspondente da habilidade h feito por jogadores da posição p
- K = número de possibilidades para o evento (2)
- α = parâmetro do ajuste da distribuição beta das probabilidades das posições p na habilidade h

position_name	**Cobranças de escanteio**	**Cobranças de falta - cruzamento**	**Cobranças de falta - finalização**	**Criação de oportunidades**	**Cruzamentos**
Atacante	0.538	0.5	0.25	0.394	0.199
Goleiro	0.5	0.484	0.467	0.012	0.5
Lateral Direito	0.637	0.491	0.387	0.344	0.282
Lateral Esquerdo	0.571	0.506	0.378	0.355	0.253
Meia	0.606	0.486	0.371	0.525	0.289
Ponta	0.593	0.451	0.265	0.434	0.241
Segundo Volante	0.578	0.421	0.390	0.311	0.239
Volante	0.553	0.525	0.3	0.155	0.284
Zagueiro	0.5	0.534	0.385	0.09	0.316

Tabela 2.5: Amostra da tabela de índices de probabilidade de acerto suavizada de cada posição p para cada habilidade h

2. Multiplicou-se os valores agregados por -1, para que os valores de maior módulo tenham valor mais negativo.

3. Aplicou-se a normalização Min-Max com intervalo (0.1,1) para o conjunto com os scores de cada habilidade h em cada posição p (linha-a-linha da tabela 2.5)

$$D_{nat}^{hp} = \frac{P_{smooth}^{hp} - \min_{p} P_{smooth}^{hp}}{\max_{p} P_{smooth}^{hp} - \min_{p} P_{smooth}^{hp}} \tag{2.22}$$

4. O valor normalizado de cada habilidade h para cada posição p corresponde ao fator D_{nat}^{hp} que comporá os pesos.

Cálculo de F_{nat}^{hp}

O cálculo do índice de frequência natural na partida para a posição p em uma habilidade h procedeu-se da seguinte maneira:

Posição/Habilidade	**Número jogadores**	**Finalização**	**Drible**	**Cruzamento**	**Desarme**	**Duelo aéreo ofensivo**
		Tentativas	**Tentativas**	**Tentativas**	**Tentativas**	**Tentativas**
Goleiro	**40**	5	120	8	250	3
Lateral Direito	**55**	350	1700	1900	2100	1000
Lateral Esquerdo	**78**	350	1500	1800	2200	1000
Zagueiro	**90**	350	400	120	2500	1600
Primeiro volante	**71**	300	800	300	2700	1300
Segundo volante	**73**	500	950	700	2200	1500
Meia	**75**	1000	1400	1400	1100	2200
Ponta	**92**	1100	1800	2000	1500	2000
Atacante	**56**	1500	1100	1200	800	3000

Tabela 2.6: Amostra da tabela de volume de eventos realizados (tentativas) de cada habilidade h para cada posição p

1. Calculou-se a quantidade de eventos (tentativas) realizados da respectiva habilidade h pelos jogadores da posição p e dividiu-se pelo número A de jogadores da posição p (valor da coluna "Número jogadores"da tabela 2.6) para se calcular a média de ações realizadas por jogador (ver tabela 2.7)

$$F_{hp} = \frac{\sum_{j=1}^{A} F_{jhp}}{A} \tag{2.23}$$

Posição/Habilidade	Finalização	Drible	Cruzamento	Desarme	Duelo aéreo ofensivo
Goleiro	0,125	3	0,2	6,25	0,075
Lateral Direito	8,75	42,5	47,5	52,5	25
Lateral Esquerdo	8,75	37,5	45	55	25
Zagueiro	8,75	10	3	62,5	40
Primeiro volante	7,5	20	7,5	67,5	32,5
Segundo volante	12,5	23,75	17,5	55	37,5
Meia	25	35	35	27,5	55
Ponta	27,5	45	50	37,5	50
Atacante	37,5	27,5	30	20	75

Tabela 2.7: Tabela 2.6 normalizada pelo número de jogadores

2. Tratou-se os valores considerados outliers, saturando os valores acima do valor de quantil de 0.95 e abaixo do valor de quantil de 0.05 no conjunto de frequencias das habilidades por posição p.

3. Multiplicou-se os valores agregados por -1, para que os valores de maior módulo tenham valor mais negativo.

4. Aplicou-se a normalização Min-Max com intervalo (0.1,1) para o conjunto de valores de frequências das habilidades h para cada posição p (linha-a-linha da tabela 2.7:

$$F_{nat}^{hp} = \frac{F_{hp} - \min\limits_{h} F_{hp}}{\max\limits_{h} F_{hp} - \min\limits_{h} F_{hp}} \tag{2.24}$$

5. O valor normalizado de cada habilidade h para cada posição p corresponde ao fator F_{nat}^{hp} que comporá os pesos.

Cálculo de F_{esp}^{hp}

O cálculo do índice de frequência comparada entre posições procedeu-se da seguinte maneira:

1. Calculou-se a quantidade de eventos realizados da respectiva habilidade h pelos A jogadores j da posição p e dividiu-se pelo número A de jogadores j da posição

p para se calcular a média de ações realizadas por jogador (ver tabela 2.6 e 2.7:

$$F_{hp} = \frac{\sum_{j=1}^{A} F_{hjp}}{A} \tag{2.25}$$

2. Tratou-se os valores considerados outliers, saturando os valores acima do valor de quantil de 0.95 e abaixo do valor de quantil de 0.05 no conjunto de frequencias das posições p por habilidade h.

3. Aplicou-se a normalização Min-Max com intervalo (0.1,1) para o conjunto de valores de frequências das posições p para cada habilidade h (coluna-a-coluna da tabela 2.7):

$$F_{esp}^{hp} = \frac{F_{hp} - \min\limits_{p} F_{hp}}{\max\limits_{p} F_{hp} - \min\limits_{p} F_{hp}} \tag{2.26}$$

Cálculo de w_{hp}

O cálculo do peso da habilidade h para a posição p procedeu-se da seguinte maneira:

1. Multiplicou-se os valores dos três pesos calculados anteriormente para um mesmo critério h de uma mesma posição p.

$$W_{prev}^{hp} = D_{nat}^{hp}.F_{nat}^{hp}.F_{esp}^{hp} \tag{2.27}$$

2. Aplicou-se a normalização Min-Max com intervalo (0.1,1) para o conjunto de pesos para os critérios h de cada posição p:

$$w_{hp} = \frac{W_{prev}^{hp} - \min\limits_{h} W_{prev}^{hp}}{\max\limits_{h} W_{prev}^{hp} - \min\limits_{h} W_{prev}^{hp}} \tag{2.28}$$

2.2.4 Métodos de normalização de variáveis

Usando MILLIGAN & COOPER (1988) [79] como referência, três métodos de normalização de variáveis serão fundamentados:

- Min-Max (compressão dos valores em intervalo definido)
- Std (normalização por desvio padrão)
- Vector-Norm (normalização pela raiz quadrada da soma dos quadrados)

2.2.4.1 Min-Max

O método Min-Max de normalização, que já foi utilizado no trabalho na fase de pré-processamento dos pesos e será testado na fase de pré-processamento dos dados para a aplicação do TOPSIS, procura redefinir o intervalo no qual os dados se encontrarão dispersos, compactando os mesmos de forma a ficarem em um novo range definido. Para o intervalo padrão (0,1), a fórmula aplicada aos dados se encontra na equação 2.29:

$$X_k^{new} = \frac{X_k - \min\limits_k X_k}{\max\limits_k X_k - \min\limits_k X_k} \tag{2.29}$$

2.2.4.2 Z-Score

A normalização por Z-Score é uma das mais aplicadas nos estudos científicos, muito por sua característica de centralizar os dados em zero, "removendo"o valor de média, e de padronizar a variância dos dados no valor unitário. Para que isso seja possível, aplica-se a equação 2.30:

$$X_k^{new} = \frac{X_k - \mu_k}{\sigma_k} \tag{2.30}$$

onde

$$\mu_k = \frac{\sum_{k=1}^{N} X_k}{N} \tag{2.31}$$

e

$$\sigma_k = \sqrt{\frac{\sum_{k=1}^{N} (X_k - \mu)^2}{N}} \tag{2.32}$$

2.2.4.3 Vector-norm

O método de normalização vetorial é aplicado de forma que os dados são reescalados usando como referência a raiz quadrada da soma dos quadrados de cada valor do conjunto. Para que isso seja possível, aplica-se a equação Y:

$$X_k^{new} = \frac{X_k}{\sqrt{\sum_{k=1}^{N} X_k^2}} \tag{2.33}$$

2.2.5 Método de redução de dimensionalidade (PCA)

Segundo SHLENS (2014) [80], a análise de componentes principais (PCA) é um método simples e não paramétrico para extrair informações relevantes de conjuntos de dados muito extenso. Com o mínimo de esforço, o PCA fornece um roteiro de como reduzir um conjunto de dados complexo para uma dimensão mais baixa, revelando as estruturas mais simplificadas e ocultas.

A análise de componentes principais é um procedimento de redução de variáveis. É útil quando se tem dados obtidos sobre um grande número de variáveis, e acredita-se que existe alguma redundância nas mesmas. Por causa dessa redundância, o método permite reduzir o conjunto de variáveis observadas em um número menor de componentes principais (variáveis artificiais) que são responsáveis pela maior parte da variância nas features observadas.

O desenvolvimento matemático e a intuição em relação ao método será desenvolvido nas próximas subseções, inspirada pela literatura de SHLENS (2014) [80].

2.2.5.1 Intuição

O objetivo da análise de componentes principais é identificar a base mais significativa para re-expressar um conjunto de dados. A esperança é que esta nova base irá filtrar o ruído e revelar estruturas ocultas na nuvem de dados.

O ruído de medição em qualquer conjunto de dados deve ser baixo, não importa a técnica de análise, ou então pouca ou nenhuma informação sobre um sinal pode ser extraído. Não existe escala absoluta para o ruído, mas em vez disso, todo o ruído é quantificado em relação à intensidade do sinal. Uma medida comum é a relação sinal-ruído (SNR), representado pela equação 2.34:

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}} \tag{2.34}$$

A variação devido ao sinal e o ruído são indicados pelas linha da figura 2.1. A razão dos dois comprimentos mede o quão fina é a nuvem, podendo ser uma linha fina ($SNR >> 1$), um círculo ($SNR = 1$) ou até mesmo pior. Ao postular

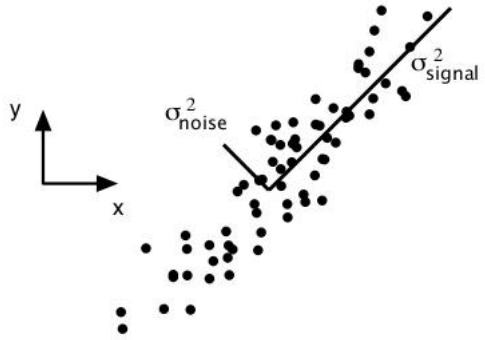


Figura 2.1: Representação gráfica do conceito de variância de sinal e ruído. Fonte: SHLENS (2014) [80]

medições razoavelmente boas, quantitativamente assumimos que as direções com maiores variações (logo as com valores de SNR altos) no espaço de medição contém a dinâmica de interesse. No caso da Figura 2.1, a direção com a maior variância não é nenhuma das canônicas ((1,0) ou (0,1)), mas sim a direção ao longo do longo eixo do nuvem.

Um outro fator a ser observado na figura 2.1 está relacionado a redundância. Se observarmos ela novamente, será fácil identificar que não era necessário utilizar 2 variáveis para representar aquele conjunto de dados, tendo em vista a baixa variância na direção de "ruído". A figura 2.2 reflete um conjunto de possíveis gráficos entre dois tipos de medições arbitrárias r_1 e r_2.

Enquanto o painel mais à esquerda mostra duas features sem relação aparente (e eventual não correlação), o gráfico mais à direita evidencia duas variáveis muito correlacionadas. Para este caso, gravar apenas uma das variáveis expressaria os dados de forma mais concisa e reduziria o número de gravações do medidor (de 2 para 1 variável). De fato, esta é a ideia chave por trás da redução de dimensionalidade.

Figura 2.2: Três casos possíveis de relações entre variáveis. Fonte: SHLENS (2014) [80]

2.2.5.2 Fundamentação matemática

Uma medida importante de se definir para avaliar o grau do relacionamento linear entre duas variáveis é a covariância. A covariância entre duas variáveis A e B se dá por:

$$\sigma_{AB}^2 = \frac{\sum_i a_i b_i}{n} \tag{2.35}$$

Um grande valor positivo indica dados correlacionados positivamente. Da mesma forma, um grande valor negativo indica dados negativamente correlacionados. Podemos definir a matriz de covariância C_X de um conjunto de dados X (m x n) por:

$$C_X = \frac{1}{n} X.X^T \tag{2.36}$$

onde C_X é uma matriz quadrada simétrica m x m, os termos diagonais de C_X são as variâncias dos determinados tipos de medição e os termos fora da diagonal de C_X são as covariâncias entre os tipos de medição. Os valores de covariância refletem o ruído e a redundância nas medições realizadas, logo valores grandes nos termos diagonais correspondem a estruturas interessantes, enquanto nos termos fora da diagonal, grandes magnitudes correspondem a alta redundância.

Com isso definido, os objetivos principais por trás da redução da dimensionalidade de um conjunto de dados pode ser resumido por: (i) minimizar a redundância, me-

dida pela magnitude da covariância, e (ii) maximizar o sinal, medido pela variância. A partir dessas duas metas, a matriz de covariância otimizada C_Y deveria ter a seguinte cara:

1. Todos os termos fora da diagonal em C_Y devem ser zero. Assim, C_Y deve ser uma matriz diagonal, com Y sendo uma matriz não correlata.

2. Cada dimensão sucessiva em Y deve ser ordenada por classificação de acordo com a magnitude da variância.

Para que a matriz C_Y atinja essa afeição, uma solução algébrica pode ser extraída com base em uma importante propriedade de decomposição de autovetores. O objetivo é resumido da seguinte forma: "encontre alguma matriz ortonormal P na qual $Y = PX$ tal que $C_Y \equiv \frac{1}{n}YY^T$ seja uma matriz diagonal. Para tal, uma boa ideia seria reescrever C_Y em termos da variável desconhecida:

$$C_Y = \frac{1}{n}YY^T$$
$$C_Y = \frac{1}{n}(PX)(PX)^T$$
$$C_Y = \frac{1}{n}PXX^TP^T$$
$$C_Y = P(\frac{1}{n}XX^T)P^T$$
$$C_Y = PC_XP^T$$

Note que na última sub-equação se encontra a matriz de covariância de X. Levando em conta os dois seguintes teoremas advindos da álgebra linear:

1. Uma matriz é simétrica se e somente se ela é ortogonalmente diagonalizável

2. Uma matriz simétrica é diagonalizável por uma matriz de seus autovetores ortonormais.

Logo, para uma matriz simétrica A, $A = EDE^T$, onde D é uma matriz diagonal e E é uma matriz de autovetores de A organizados como colunas, o truque é selecionar a matriz P para ser uma matriz onde cada linha p_i seja um autovetor de $\frac{1}{n}XX^T$. Com isso, $P = ET$. Sendo $P^{-1} = P^T$, pode-se calcular C_Y:

$$C_Y = PC_XP^T$$
$$C_Y = P(E^TDE)P^T$$
$$C_Y = P(P^TDP)P^T$$
$$C_Y = (PP^T)D(PP^T)$$
$$C_Y = (PP^{-1})D(PP^{-1})$$
$$C_Y = D$$

Podemos resumir os resultados do PCA nas matrizes P e C_Y:

- Os componentes principais de X são os autovetores de $C_X = \frac{1}{n}XX^T$.
- O i-ésimo valor diagonal de C_Y é a variância de X ao longo p_i.

Na prática, a computação de PCA de um conjunto de dados X implica em dois passos:

1. subtrair a média de cada tipo de medição
2. a computação dos autovetores de C_X.

2.2.6 Métodos de agregação de scores

Três métodos de agregação de scores serão utilizados:

- TOPSIS
- Contagem de Borda
- Maior média

2.2.6.1 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

Introdução

TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) é uma técnica útil para lidar com problemas de tomada de decisão por atributos ou critérios múltiplos (MADM / MCDM) no mundo real (HWANG & YOON (1981)) [81]. Ajuda o(s) tomador(es) de decisão (DMs) a organizarem os problemas a serem resolvidos e realizam análises, comparações e classificações das alternativas. Consequentemente, a seleção de uma (ou várias) alternativa(s) adequada(s) será(ão) feita(s). (SHIH (2006)) [82]

A ideia básica do TOPSIS é bastante direta. Origina-se do conceito de um ponto ideal deslocado a partir do qual a solução relevante tem a menor distância [83] [84]. HWANG & YOON [81] propõem ainda que o ranking de alternativas deve ser baseado na distância mais curta da solução ideal (positiva) (PIS) e o mais distante da solução ideal negativa (NIS). A TOPSIS considera simultaneamente as distâncias para o PIS e NIS, e uma ordem de preferência é classificada de acordo com sua proximidade relativa, e uma combinação dessas duas distâncias medidas. (SHIH (2006)) [82]

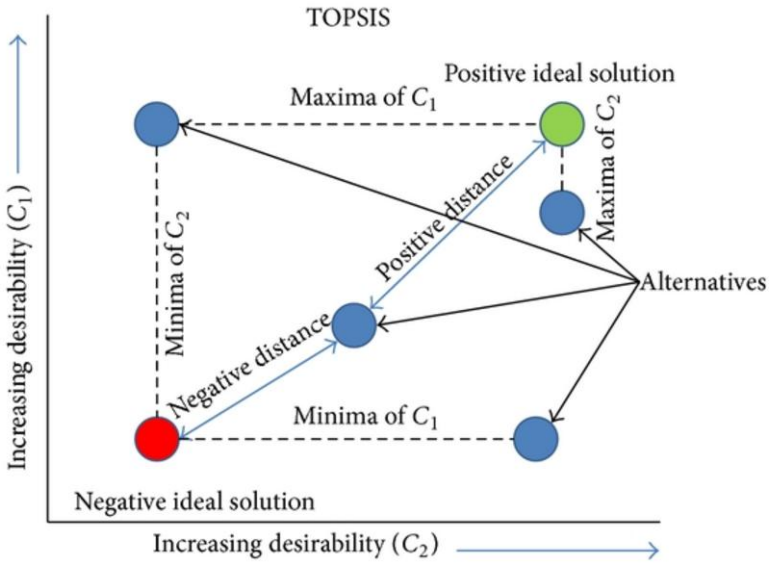


Figura 2.3: Explicação gráfica do TOPSIS. Fonte: CHAUHAN & VALSH (2013) [85]

Vantagens do uso

Segundo KIM ET AL.(1997) [86], quatro vantagens do TOPSIS são abordadas:

1. o método possui uma lógica que representa de forma bem próxima a lógica da escolha humana;

2. o rating que define a classificação é um valor escalar que representa tanto a melhor quanto a pior alternativa de forma simultânea;

3. possui um processo de computação simples que pode ser facilmente programado;

4. as medidas de desempenho de todas as alternativas nos atributos podem ser visualizadas em um poliedro, pelo menos para quaisquer duas dimensões.

Estas vantagens tornam o TOPSIS uma importante técnica de MADM em comparação com outras técnicas como o processo hierárquico analítico (AHP) e o ELECTRE (HWANG & YOON (1981)) [81]. De fato, o TOPSIS é um método que compara cada alternativa diretamente dependendo dos dados nas matrizes de avaliação e pesos (CHENG (2002)) [87]. Além disso, segundo a comparação de simulação de ZANAKIS ET AL. (1998) [88], TOPSIS tem o menor número de reversões entre os oito métodos na categoria. Assim, o TOPSIS é escolhido como o principal corpo de desenvolvimento.

Funcionamento geral do TOPSIS

Para a adoção do método TOPSIS, os seguintes passos abaixo devem ser desenvolvidos:

1. Construção da matriz de decisão D^k com m jogadores/alternativas (A), n habilidades/critérios (C):

$$D^k = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \end{array}$$

Jogador/Habilidade	Finalização	Drible	Cruzamento	Desarme	Duelo aéreo ofensivo
Peso da habilidade para a comparação atual	**1.0**	**0.6**	**0.4**	**0.2**	**0.8**
Jogador 1	0.9	0.8	0.5	0.5	1
Jogador 2	0.8	0.9	0.7	0.3	0.7
Jogador 3	0.7	0.6	0.6	0.4	0.6
Jogador 4	0.6	0.5	0.8	0.2	0.9
Jogador 5	0.5	0.7	0.9	0.6	0.8

Tabela 2.8: Exemplo de uma matriz de decisão, com jogadores nas linhas e habilidades nas colunas

2. Definição da direção de otimização critério a critério (coluna a coluna).

   Se critério/habilidade h for da natureza "quanto mais negativo melhor":

   $$x_{jh} = -x_{jh}$$

   Se critério/habilidade h for da natureza "quanto mais positivo melhor":

   $$x_{jh} = x_{jh}$$

3. Normalização critério a critério (habilidade a habilidade) da matriz de decisão, onde três métodos de escalamento linear serão testados (ver subseção 3.5.4)

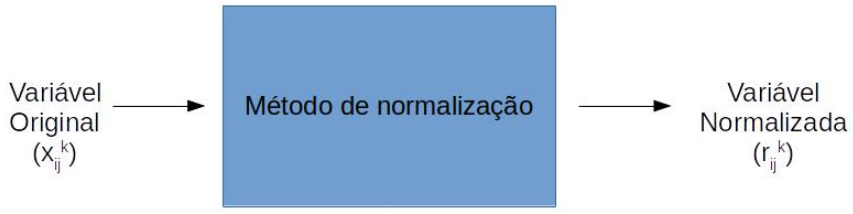


Figura 2.4: Normalização das variáveis

4. Extração das melhores (PIS) e piores (NIS) jogadores/alternativas em cada critério/habilidade, onde:

$$PIS = r_h^+ = \{r_1^+, r_2^+ ... r_n^+\} = \{\underset{j}{max} r_{jh}\}$$
$$NIS = r_h^- = \{r_1^-, r_2^- ... r_n^-\} = \{\underset{j}{min} r_{jh}\}$$

Jogador/Habilidade	**Finalização**	**Drible**	**Cruzamento**	**Desarme**	**Duelo aéreo ofensivo**
Peso da habilidade para a comparação atual	**1.0**	**0.6**	**0.4**	**0.2**	**0.8**
PIS	0.9	0.9	0.9	0.6	1
NIS	0.5	0.5	0.5	0.2	0.6

Tabela 2.9: Tabela de valores dos PIS e NIS de cada critério para o exemplo da tabela 2.8 (sem normalização prévia aplicada)

5. Cálculo da soma das distâncias (com norma p, a ser testado conforme seção 2.2.6.1) dos scores de cada alternativa/jogador em cada habilidade ao PIS e ao NIS de cada habilidade, aplicando os devidos pesos:

$$S_j^+ = [\sum_{h=1}^{N} w_h (r_{jh} - r_h^+)^p]^{\frac{1}{p}}$$
$$S_j^- = [\sum_{h=1}^{N} w_h (r_{jh} - r_h^-)^p]^{\frac{1}{p}}$$

Jogador/Habilidade	**Tipo**	**Finalização**	**Drible**	**Cruzamento**	**Desarme**	**Duelo aéreo ofensivo**
Peso da habilidade para a comparação		**1.0**	**0.6**	**0.4**	**0.2**	**0.8**
Jogador 1	**Distância ao NIS**	0.4	0.3	0.0	0.3	0.4
	Distância ao PIS	0.0	0.1	0.4	0.1	0.0
Jogador 2	**Distância ao NIS**	0.3	0.4	0.2	0.1	0.1
	Distância ao PIS	0.1	0.0	0.2	0.3	0.3
Jogador 3	**Distância ao NIS**	0.2	0.1	0.1	0.2	0.0
	Distância ao PIS	0.2	0.3	0.3	0.2	0.4
Jogador 4	**Distância ao NIS**	0.1	0.0	0.3	0.4	0.3
	Distância ao PIS	0.3	0.4	0.1	0.0	0.1
Jogador 5	**Distância ao NIS**	0.0	0.2	0.4	0.4	0.2
	Distância ao PIS	0.4	0.2	0.0	0.0	0.2

Tabela 2.10: Cálculo da distância ao PIS e ao NIS de cada jogador em cada habilidade do exemplo da tabela 2.8

Logo, o score S_j^- do jogador 1 é calculado por:

$$S_j^- = (1(0.4^p) + 0.6(0.3^p) + 0.4(0.0^p) + 0.2(0.3^p) + 0.8(0.4^p))^{\frac{1}{p}} \tag{2.41}$$

Enquanto o score S_j^+ do jogador 1 é calculado por:

$$S_j^+ = (1(0.0^p) + 0.6(0.1^p) + 0.4(0.4^p) + 0.2(0.1^p) + 0.8(0.0^p))^{\frac{1}{p}} \tag{2.42}$$

6. Para gerar os ratings finais de cada jogador, se calculará a razão da distância para o NIS do atleta sobre a soma das distâncias para o NIS e o PIS do mesmo:

$$C_j = \frac{S_j^-}{S_j^+ + S_j^-} \tag{2.43}$$

7. Ordena-se os jogadores conforme o maior valor do rating calculado.

Medidas de distância adotadas

Como o método TOPSIS é inteiramente baseado na ideia de distância para o perfeito/imperfeito, urge-se a necessidade de se definir uma métrica que permita trabalhar com esse conceito. Uma das métricas mais utilizadas na ciência é a distância euclidiana, também chamada de norma L_2, definida na equação:

$$D_j^2 = [\sum_{h=1}^{N}(r_{jh} - r_h)^2]^{\frac{1}{2}} \tag{2.44}$$

onde r_{jh} é o score de um jogador j em uma habilidade h e r_h o melhor/pior valor de score entre os atletas comparados na habilidade h

Outra métrica de distância bastante conhecida é chamada de distância de Manhattan, também chamada de norma L_1. Ela é chamada desse jeito por simular a distância entre dois pontos na cidade de Manhattan percorrendo o caminho pelas ruas da cidade:

$$D_j^1 = \sum_{h=1}^{N} |r_{jh} - r_h| \tag{2.45}$$

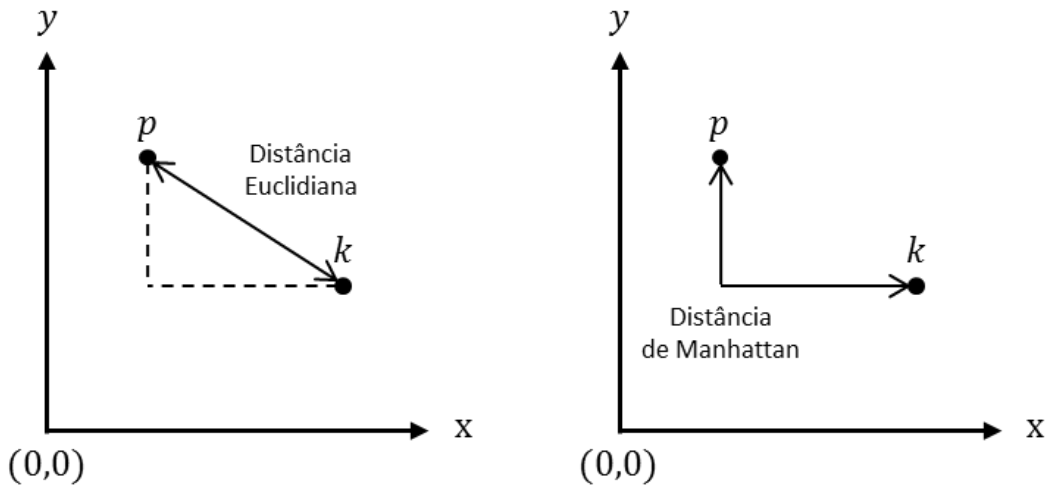


Figura 2.5: Comparação em duas dimensões das distâncias de Manhattan e Euclidiana. Fonte: https://pt.stackoverflow.com/questions/163899/geometria-computacional-determinar-vizinho-mais-pr%C3%B3ximo

Uma generalização das distâncias anteriores foi proposta por Minkowski, onde os valores dos expoentes é substituído por uma constante p, onde p é qualquer número > 0:

$$D_j^p = [\sum_{h=1}^{N}(r_{jh} - r_h)^p]^{\frac{1}{p}} \tag{2.46}$$

A escolha por testar métricas de distâncias diferentes se deu pelo fato de que os resultados com a distância euclidiana não estavam satisfatórios. Dado a isso, uma busca na literatura chegou a um estudo, realizado por AGGARWAL, HINNEBURG & KEIM (2001) [89], falando sobre como métricas de distância se comportam de forma diferente em conjuntos de dados de altas dimensões. Esse fenômeno, conhecido por *mal da dimensionalidade*, mostra que pontos no espaço R^n, para n $>>$ 3 tendem a ficar distribuidos de forma esparsa, onde cada ponto fica praticamente equidistante em relação a origem, de modo que o conceito de vizinho mais próximo perca o sentido.

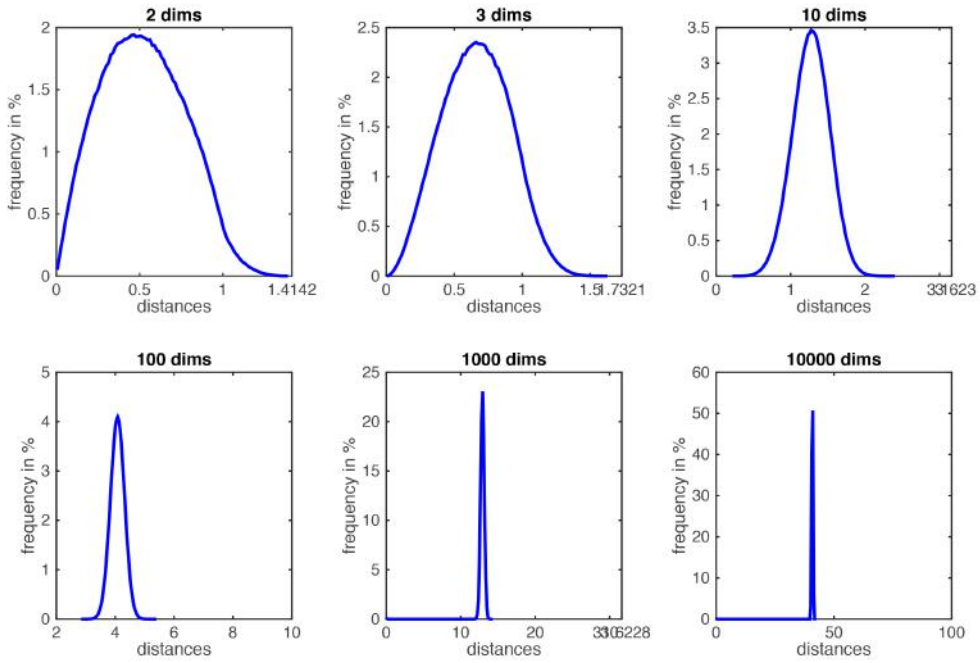


Figura 2.6: Concentração de dados em relação a distância para a origem em diversas dimensões. Fonte: http://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02_kNN.html

2.2.6.2 Contagem de Borda

A Contagem de Borda, adotada em alguns sistemas eleitorais espalhados pelo mundo, é um dos métodos da famĺia Condorcet, que elege o candidato que obtiver a maioria dos votos em todas as eleições frente a frente contra cada um dos outros candidatos (FRAENKEL (2014)) [90]. A ideia de adotar esse método vem da ideia de que o jogador ideal é aquele que consegue superar os outros em qualquer critério que eles estejam sendo comparados.

O método é aplicado da seguinte maneira:

1. Ordena-se os jogadores critério a critério (habilidade a habilidade) segundo o maior valor de x_{jh} (score do jogador j na habilidade h), atribuindo o ranking correspondente.

$$p_{jh} = \underset{j}{rankx}_{jh} \tag{2.47}$$

Jogador/Habilidade	**Finalização**	**Drible**	**Cruzamento**	**Desarme**	**Duelo aéreo ofensivo**
Peso da habilidade para a comparação atual	**1.0**	**0.6**	**0.4**	**0.2**	**0.8**
Jogador 1	1	2	5	2	1
Jogador 2	2	1	3	4	4
Jogador 3	3	4	4	3	5
Jogador 4	4	5	2	5	2
Jogador 5	5	3	1	1	3

Tabela 2.11: Aplicação do item 1 ao exemplo da tabela 2.8

2. Subtrai-se do número total de atletas comparados (M) o valor encontrado no primeiro item (p_{jh}), de forma que o jogador j de maior score em uma habilidade h tenha o maior valor.

$$r_{jh} = M - p_{jh} \tag{2.48}$$

Jogador/Habilidade	**Finalização**	**Drible**	**Cruzamento**	**Desarme**	**Duelo aéreo ofensivo**
Peso da habilidade para a comparação atual	**1.0**	**0.6**	**0.4**	**0.2**	**0.8**
Jogador 1	4	3	0	3	4
Jogador 2	3	4	2	1	1
Jogador 3	2	1	1	2	0
Jogador 4	1	0	3	0	3
Jogador 5	0	2	4	4	2

Tabela 2.12: Aplicação do item 2 ao exemplo da tabela 2.11

3. Normaliza-se por Min-Max cada conjunto de scores de jogadores correspondentes a um critério/habilidade h, no intervalo entre (0,1).

$$r_{jh}^{norm} = \frac{r_{jh} - \min\limits_{j} r_{jh}}{\max\limits_{j} r_{jh} - \min\limits_{j} r_{jh}} \tag{2.49}$$

Jogador/Habilidade	**Finalização**	**Drible**	**Cruzamento**	**Desarme**	**Duelo aéreo ofensivo**
Peso da habilidade para a comparação atual	**1.0**	**0.6**	**0.4**	**0.2**	**0.8**
Jogador 1	1.00	0.75	0.00	0.75	1.00
Jogador 2	0.75	1.00	0.50	0.25	0.25
Jogador 3	0.50	0.25	0.25	0.50	0.00
Jogador 4	0.25	0.00	0.75	0.00	0.75
Jogador 5	0.00	0.50	1.00	1.00	0.50

Tabela 2.13: Aplicação do item 3 ao exemplo da tabela 2.11

4. Somam-se, ponderando os pesos de cada critério/habilidade, os valores atribuidos a cada habilidade para o jogador. O atleta que possuir a maior soma terá o melhor ranking, e por tanto será o melhor na comparação.

$$r_j = \sum_{h=1}^{N} w_h r_{jh}^{norm} \tag{2.50}$$

2.2.6.3 Maior média

Com objetivo de ver o jogador com maior regularidade, propôs-se o ranqueamento pelo valor da maior média dos valores dos critérios.

O método é aplicado da seguinte maneira:

1. Calcula-se a média dos ratings dos jogadores. O atleta que possuir o maior valor de média terá o melhor ranking.

$$r_i = \frac{\sum_{j=1}^{N} r_{ij}}{N} \tag{2.51}$$

Jogador	Score final	Rank
Jogador 1	2,4	1
Jogador 2	1,8	2
Jogador 3	0,85	5
Jogador 4	1,15	4
Jogador 5	1,3	3

Tabela 2.14: Resultado final do ranqueamento exemplo da tabela 2.11

2.2.7 Métricas de avaliação de resultados

Duas métricas de avaliação de resultados serão adotadas. São elas:

- Correlação de Spearman
- Distância de Jaccard

2.2.7.1 Correlação de Spearman

Na estatística, o coeficiente de correlação de Spearman, ou rho de Spearman, nomeado por Charles Spearman e frequentemente denotado pela letra grega ρ ou r_s, é uma medida não-paramétrica de correlação de posição (dependência estatística entre as classificações de duas variáveis). Ele avalia o quão bem a relação entre duas variáveis pode ser descrita usando uma função monotônica.

O coeficiente possui um valor entre +1 e -1, onde 1 é correlação linear positiva total, 0 é equivalente a não correlação linear e -1 é correlação linear negativa total.

Esse coeficiente, segundo SPEARMAN (1904) [91] é calculado pela equação 2.52:

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_X \sigma_Y} \tag{2.52}$$

onde:

- ρ_{rg_X, rg_Y} = Coeficiente de correlação de Spearman
- $cov(rg_X, rg_Y)$ = Covariância entre os rankings das variáveis X e Y
- $\sigma_X \sigma_Y$ = desvios padrões das variáveis X e Y, respectivamente

Uma outra aproximação para o cálculo desse coeficiente se dá pela equação 2.53, para o caso em que os rankings estabelecidos são todos únicos (não-repetidos):

$$r_s = 1 - \frac{6 - \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2.53}$$

onde:

- r_s = Coeficiente de correlação de Spearman
- $d_i = rg_{X_i} - rg_{Y_i}$ = Distancia entre os rankings de cada variável X e Y.
- n = número de elementos comparados

2.2.7.2 Distância de Jaccard

Tendo em vista que a meta final do trabalho é realizar um filtro dos n melhores jogadores num conjunto de m atletas (com $m >> n$), precisa-se de uma métrica para poder avaliar o quão semelhante um filtro foi em relação ao outro.

Para esse caso, uma métrica, criada por Paul Jaccard, avalia o quão um determinado conjunto de variáveis discretas é similar ao outro. O cálculo se baseia na razão entre a interseção dos conjuntos e a uniao dos mesmos (JACCARD (1901)) [92]:

$$\tau = \frac{A \cap B}{A \cup B} \tag{2.54}$$

onde valores próximos de 1 apresentam conjuntos mais similares e valores próximos de 0 = mostram conjuntos mais distintos.

Figura 2.7: Exemplo de três casos para a métrica de Jaccard. Fonte: https://en.wikipedia.org/wiki/Jaccard_index

Capítulo 3

Metodologia

3.1 Fluxo geral de trabalho

3.1.1 Definição do escopo de projeto

Com objetivo de poder desenvolver e estudar as nuances relacionadas ao ranqueamento e a uma posterior seleção de atletas de forma que se pudesse ter uma boa realimentação em cima dos resultados gerados, foi decidido que o escopo do projeto se restringirá aos jogadores que atuaram no Campeonato Brasileiro de 2018. Além do fator de serem jogos recentes, a escolha foi dada por ser um campeonato de natureza de pontos corridos, que tende a mostrar uma maior continuidade de jogos em relação a outros modelos de torneio.

Uma outra decisão capital de projeto foi a divisão dos atletas por posição. Além do fato de que a comparação e o ranqueamento somente será realizado entre atletas de mesma posição, os pesos atribuidos a cada habilidade serão diferentes conforme o grupo de jogadores selecionado, tendo em vista que o jogo de futebol exige uma maior acurácia de ações diferentes para atletas de posições distintas. Elas foram divididas em:

- Goleiro
- Lateral Direito
- Lateral Esquerdo

- Zagueiro
- Volante
- Segundo Volante
- Meia
- Ponta
- Atacante

Uma terceira escolha fundamental feita para o trabalho foi a de que um mesmo evento realizado num determinado momento de uma partida somente poderia vir a ser classificada por **no máximo três** habilidades diferentes. Além disso, habilidades com o escopo de eventos possíveis totalmente incluído no escopo de outras serão removidas. O processo de seleção será mais detalhado na seção 3.3.

Quanto aos métodos de decisão multi-critério a serem testados, a escolha pelo TOPSIS como principal modelo se deu por ser um método amplamente utilizado em diversas áreas, além de ser um método bem intuitivo em sua concepção. Diversos estudos comparando as várias técnicas existentes dessa área, como feito por exemplo por LEONETI (2016) [93], já comprovaram seu bom desempenho e sua alta correlação com opiniões de especialistas.

A decisão de testar se vale a pena reduzir ou não a tabela de habilidades pela Análise de Componentes Principais (PCA) se dá pela clara natureza do problema ser de alta dimensionalidade (devido ao grande número de critérios). Como o TOPSIS se baseia na ideia de distância menor/maior para o ideal perfeito/imperfeito, o mal da dimensionalidade, problema muito discutido na Computação e que foi debatido na seção 2.2.6.1, pode vir a trazer problemas para os resultados.

3.1.2 Workflow desenvolvido

Como forma de cumprir o objetivo do trabalho, citado na seção 1.4, foi proposto o fluxo de tarefas da figura 3.1.

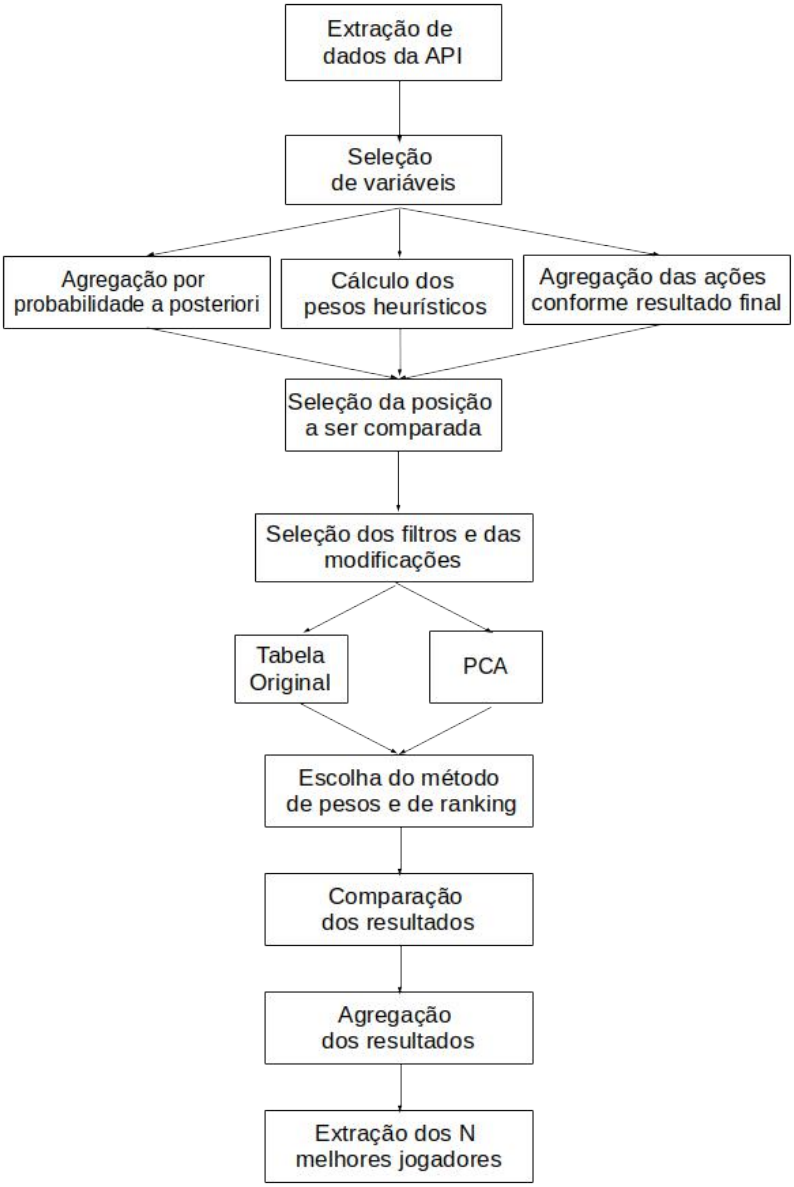


Figura 3.1: Fluxo de tarefas desenvolvidas no trabalho

O primeiro passo foi a extração dos dados técnicos da API da plataforma Instat Scout, que será detalhado na seção 3.2. Posteriormente, foi realizada uma seleção das variáveis relevantes a partir do critério citado anteriormente, melhor explicado

na seção 3.3. Feito isso, os dados serão agregados para que se possa extrair os *ratings* de cada habilidade de cada jogador, a ser detalhado na seção 3.4 e os pesos dos critérios por posição, o que será aprofundado na seção 3.5.1.

Após a seleção da posição a ser analisada, um conjunto de filtros e modificações a serem aplicadas a tabela de habilidades serão testadas, de forma a simular considerações feitas por analistas de mercado na hora de observar atletas. Elas serão melhor estudadas nas seções 3.5.2, 3.5.3 e 3.5.4.

Os métodos de ranking a serem comparados serão testados tanto com uma aplicação prévia do PCA (que será detalhado na seção 3.5.5), como diretamente nos dados originais agregados, e terão seus resultados avaliados conforme as métricas definidas em 2.2.7. Esse processo será detalhado na seção 3.6.

3.1.3 Recursos utilizados

Para extração dos dados técnicos de desempenho, foi utilizada a plataforma InStat Sports Performance Analysis and Scouting (http://instatscout.com), por meio de sua API (Application Programming Interface).

O InStat é uma plataforma baseada na web para análise de desempenho que fornece acesso a uma ampla variedade de estatísticas individuais e das equipes, todas com suporte de vídeo. Desta forma, foi extraído da plataforma informações de desempenho de 380 jogos disputados durante o Brasileiro da série A do ano de 2018. Os dados são coletados ao vivo por uma equipe de analistas altamente treinados que usam um sistema baseado em vídeo para coletar informações sobre o que acontece toda vez que um jogador toca a bola durante cada partida.

Para o processamento dos dados coletados, elaboração e implementação dos algoritmos e métodos a serem utilizados no trabalho, a linguagem de programação Python 3.6, por meio das bibliotecas de manipulação de dados Pandas e SciKit Learn e das bibliotecas de computação científica Numpy e Scipy foram utilizadas. O ambiente Jupyter Notebook, devido a facilidade de manuseio e de visualização dos resultados, foi utilizado para o desenvolvimento do processo.

Figura 3.2: Algumas das bibliotecas e plataformas utilizadas no trabalho que usam Python 3.6 . Fonte: houseofbots.com

3.2 Extração dos dados da API

O processo de extração dos dados de desempenho técnico dos jogadores foi realizado em sua forma geral da seguinte maneira:

1. cada *template* de dados que podem ser requisitados na API possui um link
2. usou-se a biblioteca *requests* para fazer a aquisição dos dados do link em formato JSON
3. converteu-se o conteúdo requisitado para o formato CSV, para melhor manipulação.

Amostras das tabelas com suas respectivas colunas podem ser vistas no apêndice A

Esse processo dividido em três partes, de forma a se organizar as aquisições da maior para a menor camada de dados:

1. Num primeiro momento, foram requisitados os jogos que serão levados em conta para a comparação e guardados em uma tabela.
2. Com as ids de cada partida já em mãos, as variáveis de cada jogador em cada partida foram extraídas e armazenadas em uma outra tabela.

3. Com as ids dos atletas que atuaram em pelo menos uma partida, pôde-se requisitar os dados básicos dos atletas, como nome, posição e idade.

3.2.1 Extração dos jogos a serem levados em conta

Na tabela A.1, contida no apêndice A, podemos ver uma amostra das características coletadas dos jogos, contendo as partidas do Campeonato Brasileiro de 2018. Foi um total de 380 partidas, com cada equipe analisada atuando num total de 38 jogos. Para o trabalho vigente, como uma primeira publicação do estudo, foi decidido que as partidas nao iriam possuir pesos diferenciados na agregação.

3.2.2 Extração do rendimento dos jogadores por partida

Na tabela A.2, contida no apêndice A, podemos ver uma amostra das variáveis que são coletadas partida-a-partida dos jogadores. Cada linha dessa tabela corresponde ao rendimento técnico de um atleta em um jogo.

Nela, existem dados de diversas naturezas (positivo, negativo, %, total). A classificação, tradução e o agrupamento de cada uma das colunas por habilidade será detalhado na seção 3.3.

3.2.3 Extração dos dados gerais dos jogadores

Na tabela A.3, contida no apêndice A, podemos ver uma amostra dos dados pessoais dos jogadores, contendo informações como nome, data de nascimento, clube, pé dominante, posição. Foi um total de 936 atletas analisados, divididos em 9 posições. Como, por inspeção, pôde-se perceber alguns erros na classificação de posição dos jogadores, a base de dados Transfermarkt (www.transfermarkt.pt) foi utilizada como referência.

3.3 Seleção dos critérios

O processo de seleção dos critérios (habilidades) foi o que mais durou no caminho da pesquisa. Tendo em vista que é a base de toda a pirâmide e que afeta todo o caminho posterior, essa foi a parte que sofreu mais modificações no percurso.

Figura 3.3: Página do atleta Diego Souza no site transfermarkt.pt. Fonte: https://www.transfermarkt.pt/diego-souza/profil/spieler/33315

3.3.1 Classificação das variáveis

As variáveis, quando são extraídas diretamente da API, chegam na forma de quantidade bruta pura (ou de percentagem). Uma habilidade de passe curto para a direita, por exemplo, seria apresentada nas formas de dados possíveis relacionadas a ela:

- Passe curto para a direita - acertos
- Passe curto para a direita - erros
- Passe curto para a direita - total
- Passe curto para a direita - %

O problema que se pôde notar é que nem todas as habilidades possuíam exatamente esses quatro tipos de dados. Algumas possuíam três dessas, enquanto outras

apenas duas, o que exigiu a criação de um conjunto de dados de classificação e agrupamento das variáveis. Podemos ver uma amostra desse conjunto na tabela 3.1:

Label	**Grupo**	**Dado**	**Caso**	**Tipo**
Assists	Assistências	POS	POS	Individual
Clearance	Bolas afastadas	POS	AVG	Individual
YC, including short-data information	Cartões amarelos	NEG	AVG	Individual
RC, including short-data information	Cartões vermelhos	NEG	AVG	Individual
Chance was not converted by	Chance de gol	NEG	POS+TOT	Individual

Tabela 3.1: Amostra da tabela de classificação das variáveis

Podemos ver que:

- Label: Nome da variável que sai diretamente da API. Ex: "Chance was not converted by"
- Grupo: Habilidade a qual a variável pertencerá. Ex: Chance de gol. As habilidades serão os critérios a serem levados em conta na hora de aplicar o método de decisão.
- Dado: Natureza da variável que sai da API. Ela pode ser uma contagem bruta de cunho positivo (POS), negativo (NEG), total (TOT) ou uma percentagem (%)
- Caso: Caso que a respectiva habilidade se enquadra no aspecto de quais dados disponíveis relativas a ela. Se a habilidade possui somente dado de cunho positivo, se ela só possui um dado de cunho negativo, se possui dados positivos e de total, etc.
- Tipo: Se é uma variável de natureza individual (que pertence a um jogador em questão) ou se é uma variável coletiva (que pertence a equipe do jogador analisado).

3.3.2 Critérios de seleção

Dessa maneira, algumas decisões de projeto foram tomadas:

1. Somente variáveis do tipo individual serão levadas em conta para essa pesquisa. O objetivo é tentar focar ao máximo no rendimento do atleta em questão. Variáveis como "finalizações da equipe enquanto jogador em campo", "gols sofridos pela equipe enquanto jogador em campo"foram removidas, por acreditar-se que a influência de um atleta para essas consequências sejam muito baixas.

2. Uma ação realizada num determinado momento de uma partida somente poderia vir a ser classificada por no máximo três categorias de ação diferentes. Um lance de cruzamento para gol, por exemplo, se enquadra em: cruzamento, passe para finalização e assistência

3. Habilidades com o escopo de possibilidades totalmente incluído no escopo de outras serão removidas. Um exemplo disso: passes curtos para a direita se encontra totalmente incluída dentro do escopo de passes curtos. Logo, a opção ficou por trabalhar com a camada mais detalhada, com a segunda variável (mais geral) sendo removida.

Com isso, chegamos a um número de 52 habilidades, que serão levadas em conta na hora de analisar e decidir pelo grupo seleto de atletas de melhor rendimento por posição no grupo de jogos analisados. Na tabela 3.2 pode-se ver a lista das habilidades selecionadas que serão utilizadas como critérios para se tomar a decisão no processo de ranqueamento.

3.4 Geração dos scores/ratings das habilidades

Para se gerar os índices de cada jogador em cada critério (habilidade) que será levado em conta na hora de realizar o ranqueamento dos atletas, duas propostas serão sugeridas e eventualmente comparadas:

- Utilizando a suavização aditiva pura (que foi fundamentada na seção 2.2.2.1, com α calculado a partir do ajuste em uma distribuição beta do conjunto das porcentagens por posição, feito pelo método da máxima verossimilhança.

Chance de gol	Duelos no chão defensivos	Passes extra-ofensivos
Cobranças de escanteio	Duelos no chão ofensivos	Passes longos para a direita
Cobranças de falta - cruzamento	Finalizações	Passes curtos para trás
Cobranças de falta - finalização	Goleiro - defesas	Passes de primeira
Cruzamentos	Interceptações do goleiro	Passes longos para a esquerda
Desarmes	Passes chaves	Passes longos para frente
Dribles	Passes curtos para a direita	Passes médios para a direita
Duelos aéreos defensivos	Passes curtos para a esquerda	Passes médios para a esquerda
Duelos aéreos ofensivos	Passes curtos para frente	Passes médios para frente
Recuperação de bola no campo do adversário	Penaltis sofridos	Faltas sofridas
Assistencias	Manutenção de posse	Interceptações
Criação de oportunidades	Bolas afastadas	Interceptações no campo do adversário
Gols feitos	Cartoes amarelos	Manutenção de posse (campo de defesa)
Defesas do goleiro	Cartões vermelhos	Passes no campo do adversário
Dominio de bola	Faltas cometidas	Passes ofensivos
Erros graves	Faltas cometidas - penalti	Penaltis cobrados (incluindo disp. penaltis)
		Penaltis defendidos (incluindo disp. penaltis)

Tabela 3.2: Lista dos critérios selecionados para as comparações

- Utilizando a média bayesiana (que foi fundamentada na seção 2.2.2.2, agregando os valores de cada critério por posição e encontrando a média a priori m da mesma, com $C = \alpha d$.

3.5 Processos realizados pré-agregação de scores

3.5.1 Atribuição de pesos

A grande maioria dos trabalhos envolvendo métodos de decisão multi-critério possuem um momento onde há a necessidade de se atribuir pesos aos aspectos que serão levados em conta. Os pesquisadores da área, em sua grande maioria, procuram um conjunto de especialistas do tema que envolve a decisão em questão e perguntam a eles quais ponderações eles aplicam/utilizam nos critérios na hora de realizar a escolha final.

Para tal fim, o método AHP (Analytic Hierarchic Process) é comumente utilizado, muito devido a sua praticidade e facilidade de entendimento. NIKJO, REZAEIAN & JAVADIAN (2015) [60]; BALLI & KORUKOGLU (2012) [62], AGILONU & BALLI (2009) [63] propuseram a adoção dessa técnica para definição de pesos a critérios em contexto de seleção de atletas no esporte.

Porém o método possui algumas limitações. Por se tratar da atribuição de notas a comparações 2 a 2 entre critérios, ele se torna inviável para um número muito alto de aspectos a serem levados em conta, como é o caso do estudo vigente. Além disso, a atribuição de notas é definida por critérios subjetivos, o que pode mostrar certo enviesamento conforme as opiniões pessoais do decisor.

Para evitar a subjetividade, dois métodos de atribuição de pesos, que procuram defini-los a partir dos próprios dados, sem interferência externa de humanos, serão propostos para esse trabalho:

- Método de pesos por entropia, fundamentado na seção 2.2.3.1
- Método heurístico, proposto pelo autor, fundamentado na seção 2.2.3.2
- Sem adoção de pesos

3.5.2 Filtros pré-agregação

Os primeiros testes, aplicando puramente os métodos de ranking nos índices calculados por jogador (conforme detalhado na seção 3.4), mostraram resultados inconsistentes (serão apresentados e discutidos no capítulo 4). A partir desses outputs, algumas ideias foram pensadas, de forma que se alterasse o número de critérios e de alternativas a serem levadas em conta:

- Levar em conta (ou não) somente as ações de maior relevância para a posição.
- Excluir (ou não) critérios com ações que foram pouco realizadas pelos atletas da posição.
- Excluir (ou não) jogadores com baixo número de ações realizadas no total.

3.5.2.1 Filtro das ações de maior relevância

O primeiro filtro pensado, naturalmente, seria o de realizar uma segunda seleção sobre as n variáveis, de forma que somente um subconjunto com m critérios (com $m < n$) fosse ser colocado como entrada para os métodos de tomada de decisão multi-critério.

Para tal, foi pensado o seguinte filtro: somente as m ações de maior peso (estabelecidas pelo método heurístico) seriam levadas em conta na comparação. Vários números de subconjuntos das features mais relevantes foram testadas e tiveram seus resultados avaliados (pelas métricas detalhadas na seção 2.2.7).

Isso simularia um tomador de decisão com visão menos holística, que ignora o rendimento do atleta em habilidades de menor importância para o mesmo, porém mais objetiva, que foca no que realmente é necessário para o tal jogador. Além disso, é um bom método para avaliar quais variáveis os rankings de referência tendem a pesar mais nas análises.

Um outro fator para a realização do filtro seria a de diminuir a dimensionalidade do conjunto de dados, de forma que problemas nas métricas de distância relacionados a altos números de features fossem evitados.

3.5.2.2 Filtro de ações de baixa frequência

O segundo filtro a ser testado seria o de se excluir ações que foram pouco realizadas pelos jogadores da comparação em questão. Ele foi planejado devido ao fato de que ações de baixa frequência tinham grande impacto nos resultados finais dos ranqueamentos, tendo em vista que o método TOPSIS, utilizado no trabalho, usa distâncias como referência.

Uma feature com poucas amostras no total para os jogadores da comparação evidencia a baixa necessidade de se colocá-la como um critério para comparar, além de que evita que jogadores com altos valores nessas ações se destaquem em relação aos outros.

Para a definição do limiar mínimo para as ações, foi estabelecido que o número de tentativas mínimas que os jogadores de uma posição devem realizar é de metade do número de jogadores incluídos na comparação.

3.5.2.3 Filtro de jogadores pouco participativos

O terceiro e último filtro a ser testado seria o de se excluir jogadores com baixo número de ações por partida realizado no conjunto de jogos analisado. Alguns motivos levaram a esse teste:

- Jogadores pouco participativos possuem menos chance de errar do que os mais participativos.
- Avaliação da característica de independência de alternativas irrelevantes por parte dos métodos de ranking testados.

Para tal, definiu-se o limiar mínimo de ações como o valor do primeiro quartil (0.25) do conjunto de tentativas de ações por partida dos atletas comparados.

Ao invés de aplicar o filtro diretamente no número de jogos, o que é comumente usado, aplicá-los no número de ações evita casos de rendimentos excepcionais em jogos específicos.

Além de testar filtros, que implicam em mudanças no volume de dados (tanto no número de critérios como no número de alternativas), modificações nos dados já existentes serão avaliadas no quesito de impacto final nos resultados.

3.5.3 Bonificação critério-a-critério a jogadores que realizaram mais ações

Um fator que, por diversas vezes, gera discussões quando o tema é comparação de atletas é quanto a questão de jogadores com número de ações muito discrepantes. Um jogador menos participativo na partida, que arrisca menos situações, tende a ter menos chance de errar que um que procura realizar mais as ações.

Logo, para avaliar esse efeito, uma proposta de bonificação a jogadores com maior número de eventos realizados no respectivo tipo de critério (habilidade) foi feita, sendo aplicada da seguinte forma:

1. Extrai-se uma tabela com o número de ações médio realizado por cada jogador j em cada critério (habilidade) h num total de N partidas.

$$F_{tot}^{hj} = \frac{\sum_{m=1}^{N} F_{tot}^{mhj}}{N} \tag{3.1}$$

2. Aplica-se uma normalização Min-Max, semelhante a realizada na equação 2.26.

$$W_{freq}^{hj} = \frac{F_{tot}^{hj} - \min_{j} F_{tot}^{hj}}{\max_{j} F_{tot}^{hj} - \min_{j} F_{tot}^{hj}} \tag{3.2}$$

3. Multiplica-se os valores dos ratings (P_{smooth}^{hj}) por essa bonificação (W_{freq}^{hj}) calculada.

$$P_{up}^{hj} = W_{freq}^{hj}.P_{smooth}^{hj} \tag{3.3}$$

3.5.4 Normalização das variáveis

Em diversos procedimentos no ramo de Data Science, é recomendado um passo prévio de tratamento dos dados relativo a redefinição do intervalo dos mesmos. Em tabelas que possuem features com unidades de medida e escalas diferentes, normalizar variáveis é um processo comum.

Para o trabalho em vigor, utilizaremos normalizações tanto no procedimento de aplicação de pesos como no pré-processamento para a adoção do TOPSIS. Para tal, compararemos o uso de três métodos de normalização, avaliando o impacto dos mesmos nos resultados finais. São eles:

1. Min-Max
2. Z-Score
3. Vector-Norm

3.5.5 Aplicação da redução de dimensionalidade (PCA)

Para reduzir o número de dimensões do conjunto de dados de jogadores com suas respectivas habilidades, a adoção da função PCA (implementada na biblioteca Scikit-Learn de Python 3.6) será testada após a execução dos filtros e das modificações propostas nas seções 3.5.2, 3.5.3 e 3.5.4 e antes da aplicação dos métodos de tomada de decisão multi-critério.

Para tal, a matriz de dados agregada, com cada jogador sendo uma linha e cada habilidade sendo uma coluna, já filtrada e modificada será colocada como entrada da função PCA. Ela retorna os seguintes resultados:

- Os componentes principais ordenados por variância explicada (C).
- A variância explicada acumulada por cada componente gerada (VE) .
- Os autovalores associados a cada componente gerada (AV)

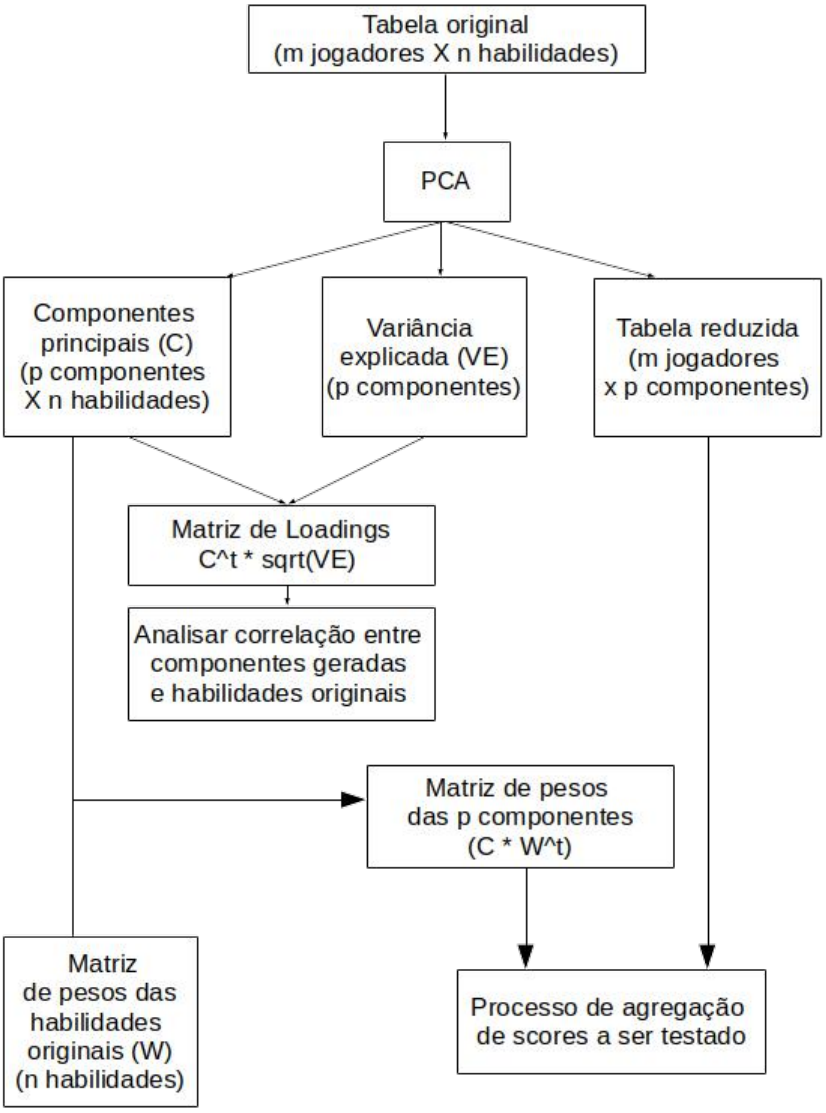


Figura 3.4: Diagrama de blocos explicativo do processo pré e pós PCA

Para se avaliar a correlação entre as componentes geradas e as habilidades originais (para poder se interpretar melhor os resultados do PCA), a matriz de loadings foi computada da seguinte forma:

$$M_{load} = C^T * \sqrt{VE} \tag{3.4}$$

O número de componentes retidos foi definido de forma que se permita reter uma determinada porcentagem da variância explicada acumulada dos dados (serão testados vários valores).

Além disso, para que se possa adotar pesos a um sistema de coordenadas diferente, como o gerado pelas componentes, os mesmos serão aplicados ao valor de cada associação entre componente e habilidade original (C_i^c), correspondente a matriz C e posteriormente somados:

$$W^c = \sum_{i=1}^{n} C_i^c * W_i \tag{3.5}$$

O sinal do valor encontrado em W^c corresponde ao sentido de otimização da componente e a magnitude do valor é diretamente relacionada com a importância da componente.

3.6 Comparação dos resultados

Como forma de avaliar o impacto de cada teste realizado nos resultados finais, foi utilizado o seguinte método:

- Para avaliar a sensibilidade dos resultados em relação aos testes realizados, a média das métricas de comparação (correlação de Spearman e distância de Jaccard) de todos os métodos de agregação de scores foi utilizada como referência para análise.
- Para analisar o desempenho de cada método de agregação, as métricas de comparação de cada método serão extraídas e comparadas.

Os seguintes métodos de agregação serão testados:

- TOPSIS (com peso heurístico, por entropia e sem peso)
- Contagem de borda
- Maior média (com peso heurístico, por entropia e sem peso)

O ranking estabelecido pela plataforma InStat Scout será usado como referência (padrão ouro) tanto para a análise de correlação como para a análise de filtro (distância de Jaccard). Um bom desempenho dos mesmos é representado por maiores valores nessas métricas.

A escolha pelo mesmo se deu por ser uma plataforma de dados amplamente utilizada por clubes brasileiros e pela mídia esportiva, e considerada uma das grandes referências mundiais no âmbito de dados aplicado ao futebol.

Em cada um dos testes realizados, dois contextos serão levados em conta:

- Aplicação direta no conjunto de dados originais, variando o número de critérios (habilidades) levados em conta. As habilidades foram ordenadas conforme o peso estabelecido pelo método heurístico e o número de variáveis n foi selecionado levado em conta as n variáveis mais importantes segundo a heurística.
- Aplicação sob o conjunto de dados reduzido por PCA, variando a quantidade de variância explicada desejada.

Capítulo 4

Resultados

4.1 Avaliando os tipos de geração de scores

O primeiro passo realizado é o de gerar os scores a partir dos dados coletados. Duas propostas foram comparadas:

- Suavização aditiva, que leva em conta uma prior uniforme de 0.5.
- Média bayesiana, que leva em conta uma prior calculada a partir de todos os jogadores envolvidos na comparação.

4.1.1 Aplicação direta nos dados originais

O gráfico da figura 4.1 permite observar que a diferença da correlação média dos rankings entre os dois métodos de geração de scores (bayesiana e aditiva) utilizando o conjunto de dados originais é bem pequena para todas as posições, com o maior valor de diferença sendo na faixa de 0.01.

Além disso, pode-se ver que a escolha do tipo de agregação impacta pouco nos valores da distância de Jaccard, como retratado na figura 4.2.

Logo, independente do uso de suavização aditiva ou de média bayesiana, os resultados se mostraram semelhantes para ambos os casos.

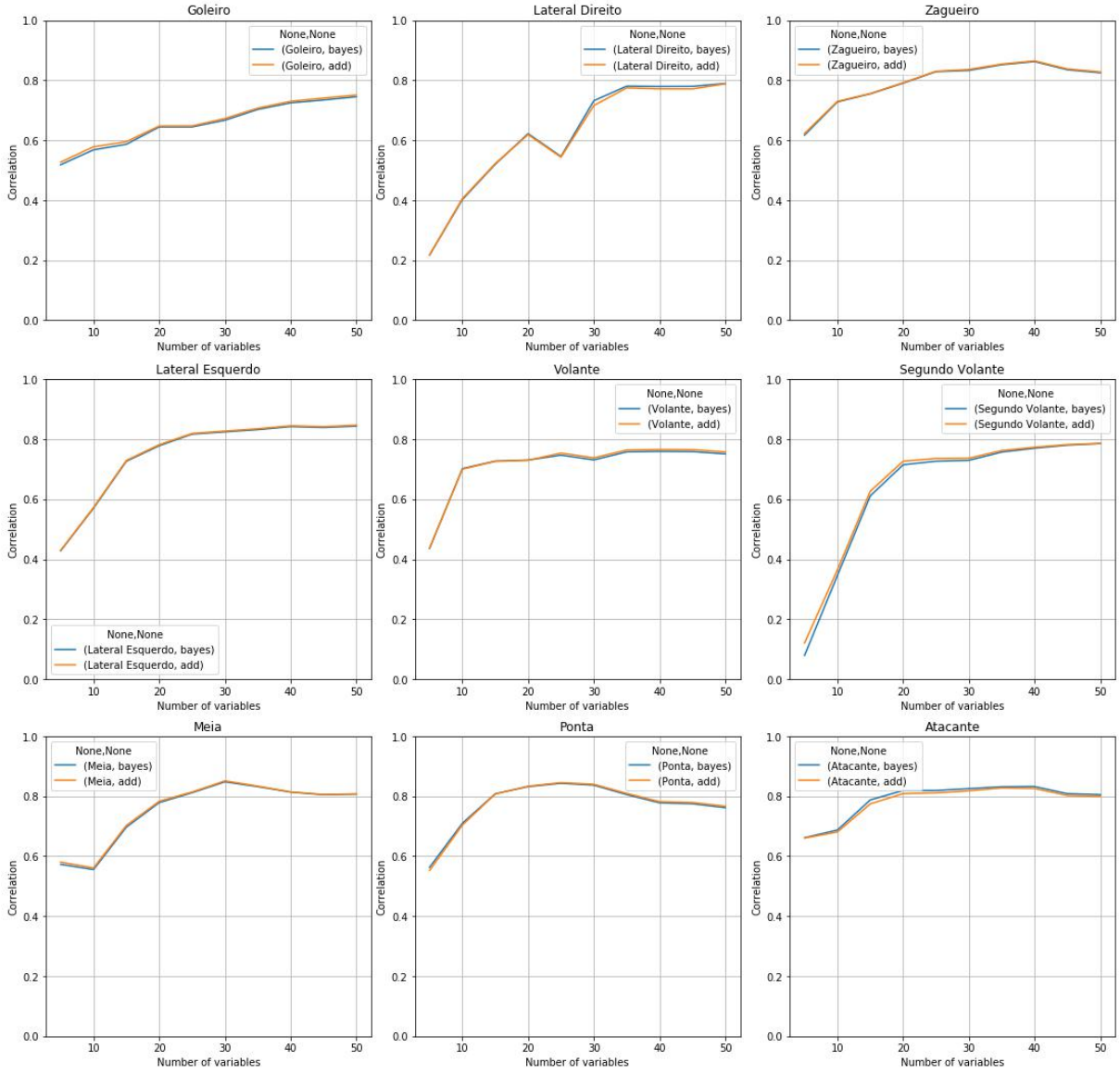


Figura 4.1: Comparação entre os tipos de geração de scores usados a partir das correlações médias entre os rankings gerados e o da plataforma InStat

4.1.2 Aplicação com PCA

Se aplicar o PCA antes para reduzir dimensionalidade, a diferença da correlação média dos rankings entre os dois métodos de geração de scores (bayesiana e aditiva) já fica maior.

A figura 4.3 mostra que para os laterais direitos e para os atacantes, a abordagem com média bayesiana deu melhores resultados, enquanto que para os segundos

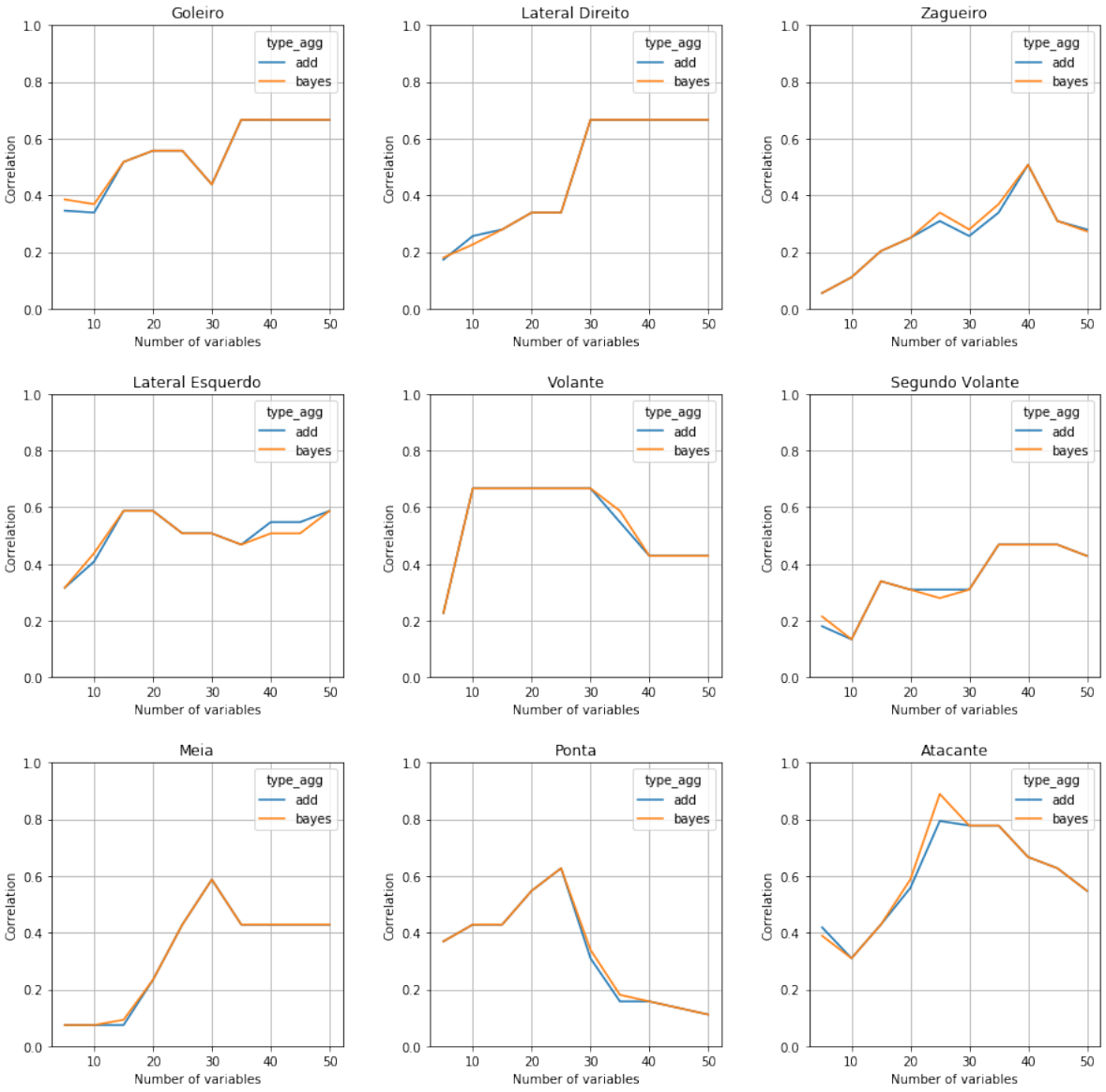


Figura 4.2: Comparação entre os tipos de geração de scores usados a partir das distâncias de Jaccard médias entre os rankings gerados e o da plataforma InStat

volantes e para os pontas a abordagem com suavização aditiva obteve maior correlação.

A figura 4.4 mostra o impacto da redução de dimensionalidade na distância de Jaccard entre o top-5 dos rankings gerados e o da plataforma InStat. Pode-se perceber que os valores são bem inferiores ao caso sem aplicação do PCA. Comparando entre os métodos de geração de scores, o método de suavização aditiva mostrou

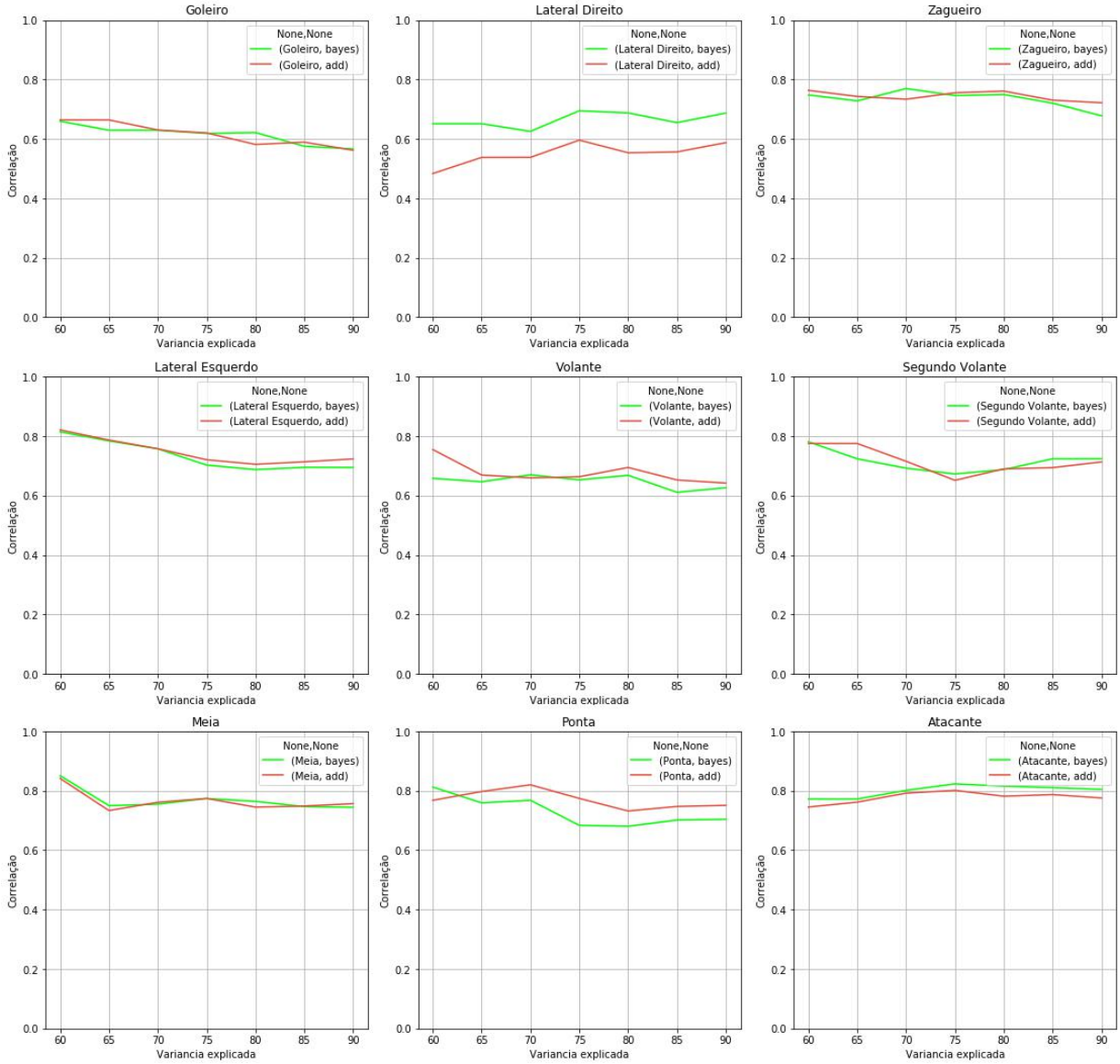


Figura 4.3: Comparação entre os tipos de geração de scores usados a partir das correlações de Spearman médias entre os rankings gerados e o da plataforma InStat

melhor rendimento nos goleiros e nos laterais direitos, enquanto a média bayesiana mostrou melhor desempenho nos atacantes.

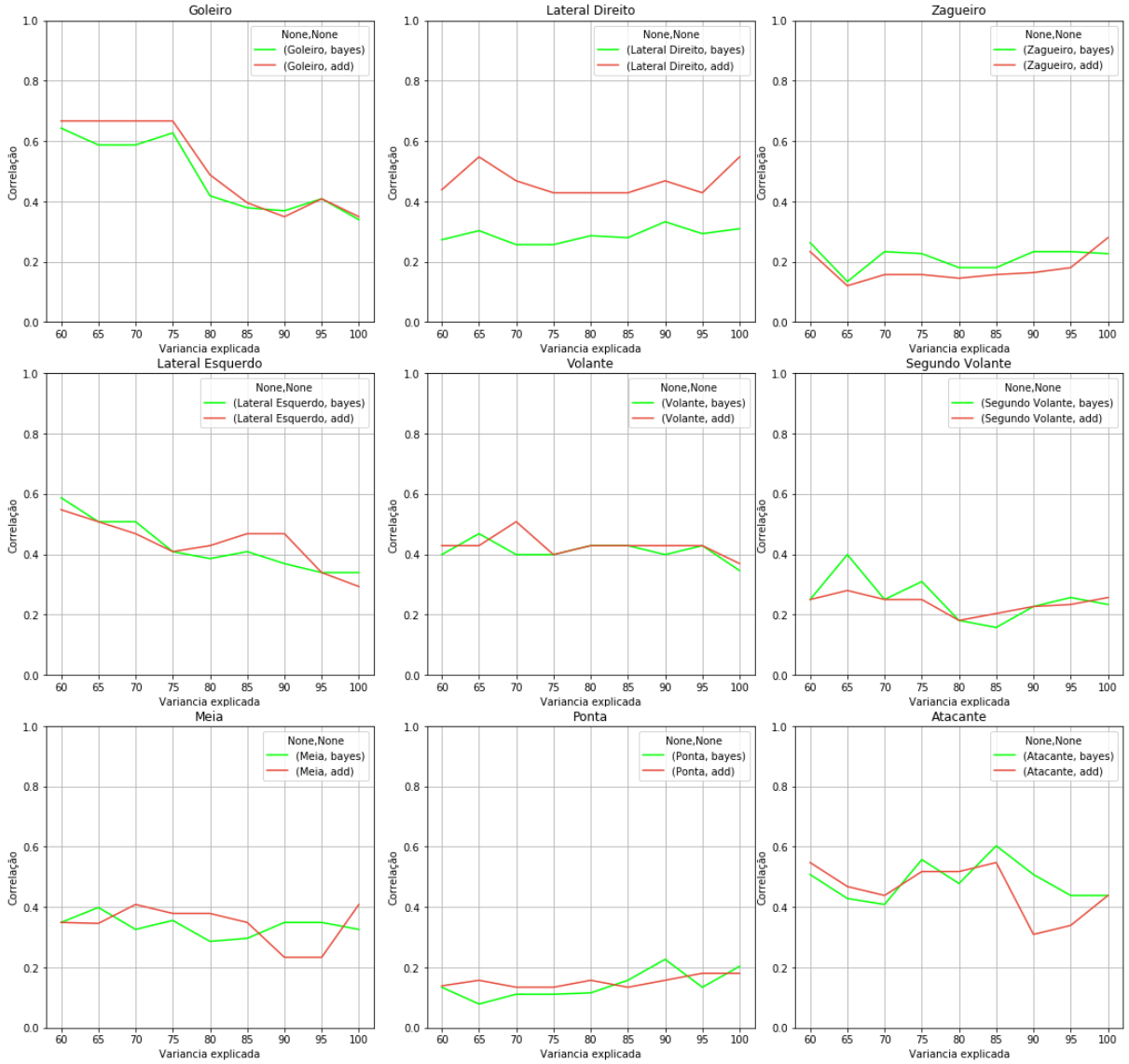


Figura 4.4: Comparação entre os tipos de geração de scores usados a partir das distâncias de Jaccard médias entre os rankings gerados e o da plataforma InStat

4.2 Comparando os métodos de ponderação

Três métodos de atribuição de pesos foram propostos e aplicados aos métodos de ranqueamento testados. São eles:

- Pesos por entropia (ent_pure)
- Pesos por heurística definida (pure_wei)
- Sem aplicação de pesos (pure)

Como forma de comparação, observaremos o respectivo impacto dos dois métodos de ponderação nas seguintes agregações:

- TOPSIS
- Maior média (hmean)

4.2.1 Aplicação direta

A figura 4.5, mostra como cada método de ponderação impactou para melhorar a correlação da técnica TOPSIS em relação ao ranking da plataforma InStat:

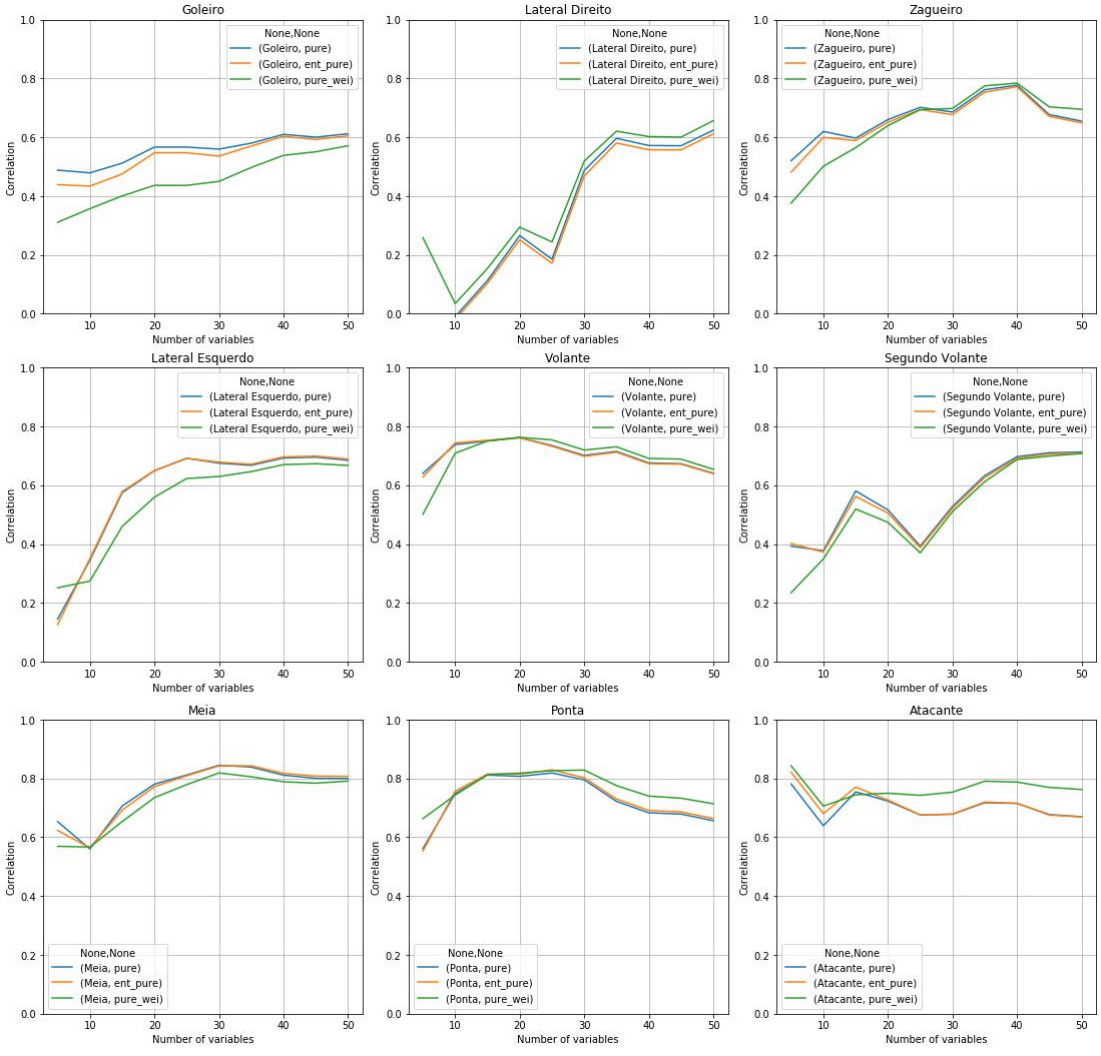


Figura 4.5: Comparação entre os métodos de ponderação utilizados a partir das correlações de Spearman entre o ranking TOPSIS e o da plataforma InStat

Pode-se observar que para Lateral Direito, Zagueiro, Volante, Ponta e Atacante, o método de pesos heurísticos foi o que obteve os maiores índices de correlação com o ranking da plataforma InStat. Para os Goleiros, sem ponderação foi o método que melhor performou nessa métrica. Para Meias, Segundos Volantes e Laterais Esquerdos, a técnica heurística mostrou um desempenho abaixo das outras duas.

A figura 4.6 mostra como que cada método de ponderação impactou na correlação entre a técnica Maior Média e o ranking da plataforma InStat:

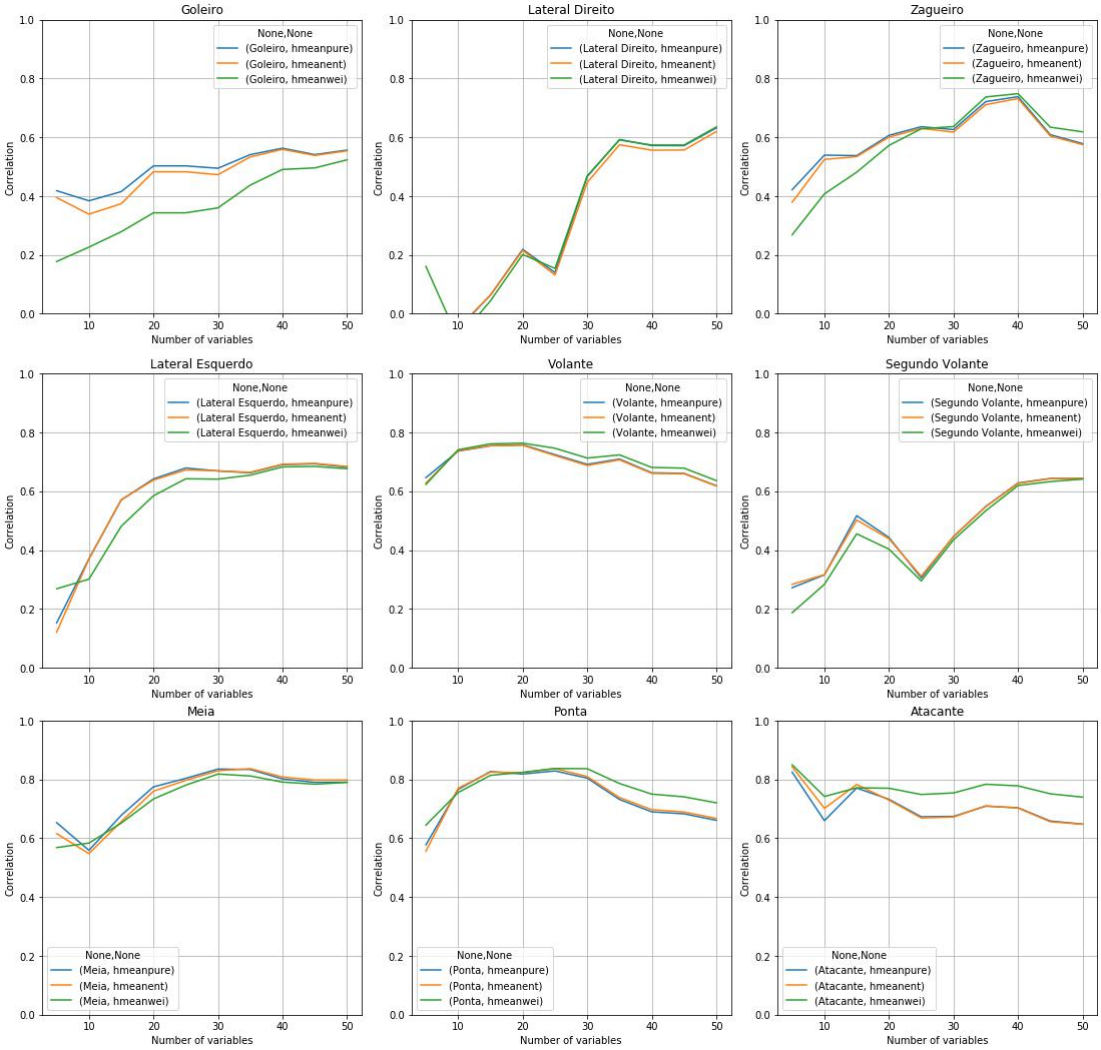


Figura 4.6: Comparação entre os métodos de ponderação utilizados a partir das correlações de Spearman entre o ranking Maior Média e o da plataforma InStat

Pode-se observar que a performance foi similar ao do método TOPSIS. De uma forma geral, o método heurístico (verde) teve o desempenho mais interessante em relação aos outros, já que foi o que mais teve posições em que teve o melhor desempenho e não ficou muito abaixo (exceto para o caso dos goleiros) nas posições em que foi superado.

A figura 4.7 nos permite observar a evolução da distância de Jaccard do top-5 entre os métodos de ponderação aplicados ao TOPSIS e o ranking da plataforma InStat:

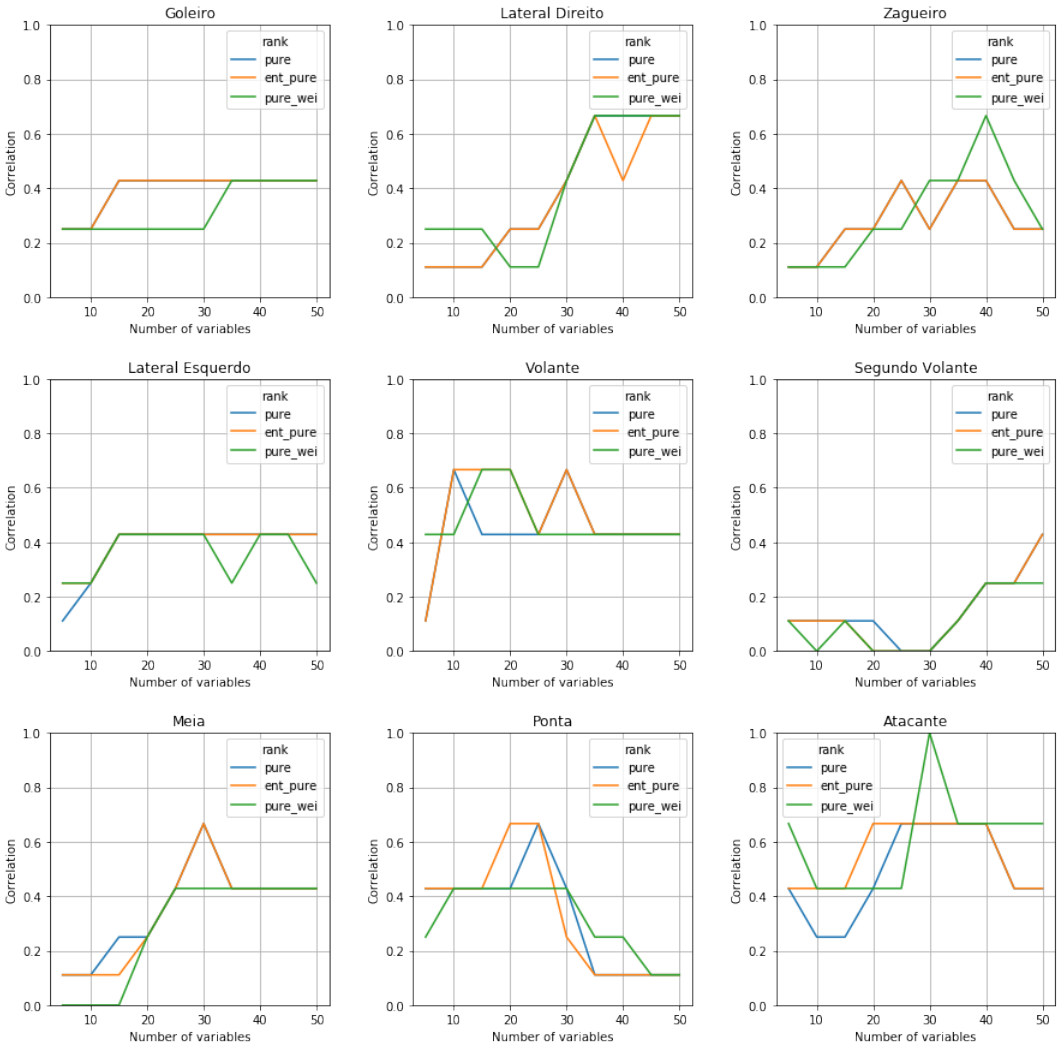


Figura 4.7: Comparação entre os métodos de ponderação utilizados a partir da distância de Jaccard entre o top-5 do ranking TOPSIS e o da plataforma InStat

Os resultados mostram que, para um filtro de atletas (top-5) levando em conta o método TOPSIS, os três métodos de ponderação atingem os picos de valor na grande maioria das posições. Nas que os três não atingem, o método heurístico (verde) teve os maiores valores no primeiro e no último, enquanto o método por entropia (amarelo) dominou no segundo e no terceiro.

Já a figura 4.8 nos permite observar a evolução da distância de Jaccard do top-5 entre os métodos de ponderação aplicados a técnica de Maior Média e o ranking de referência da plataforma InStat:

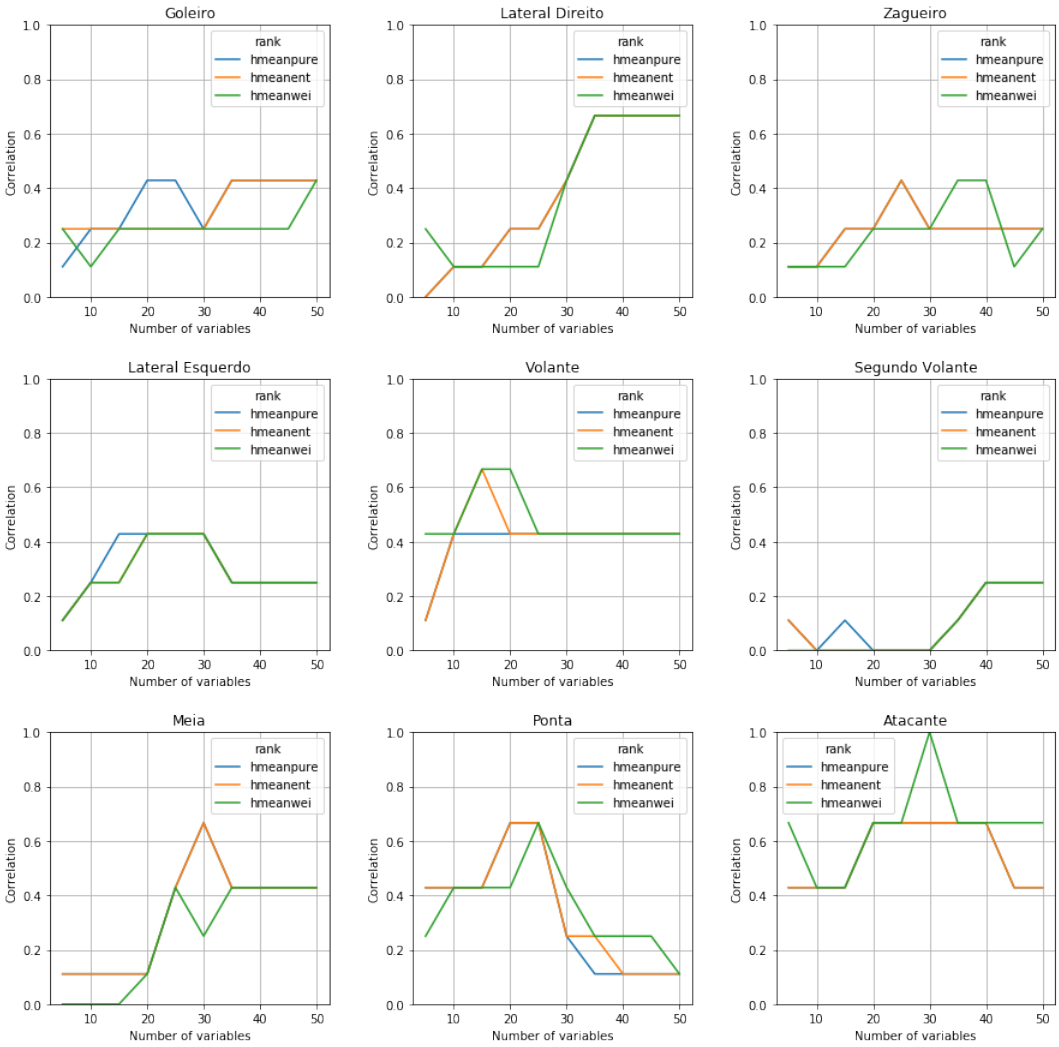


Figura 4.8: Comparação entre os métodos de ponderação utilizados a partir da distância de Jaccard entre o top-5 do ranking Maior Média e o da plataforma InStat

Os resultados mostram que, para um filtro mais apertado (top-5) levando em conta o método Maior Média de ranqueamento, os três métodos de ponderação tiveram desempenhos semelhantes ao ranqueamento por TOPSIS.

4.2.2 Aplicação com PCA

Com aplicação do PCA, a dominância do método heurístico (em verde no primeiro e em amarelo no segundo) fica mais clara na questão dos desempenhos das métricas de correlação. As figuras 4.9 e 4.10 mostram como que os métodos TOPSIS e Maior Média reagiram a escolha do tipo de atribuição de peso.

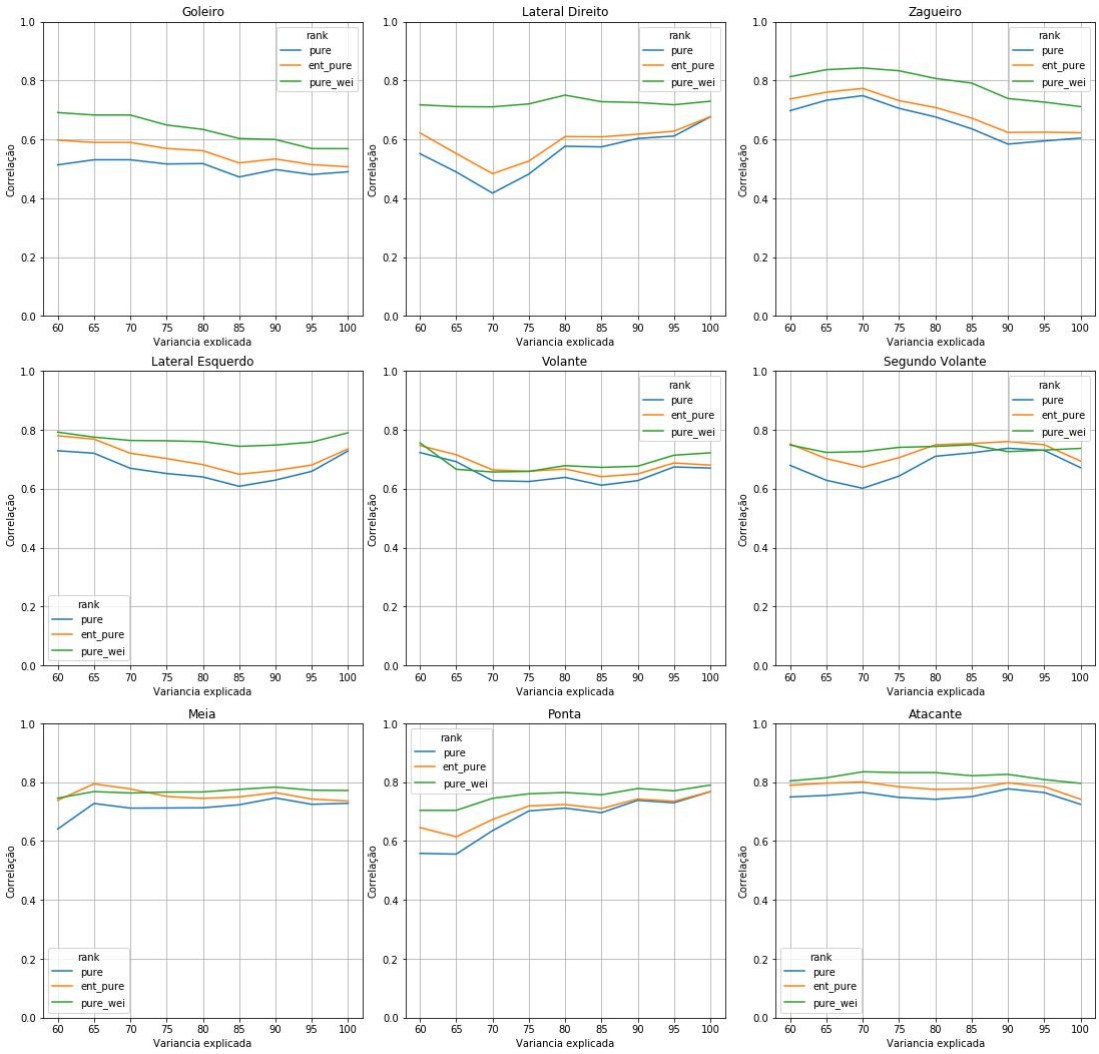


Figura 4.9: Comparação entre os métodos de ponderação utilizados a partir da correlação de Spearman entre o ranking TOPSIS e o da plataforma InStat

A figura 4.9 deixa bem claro o domínio da linha verde do peso heurístico em relação aos outros métodos no aspecto de correlação com o ranking de referência. O mesmo ocorre na figura 4.10 (em amarelo):

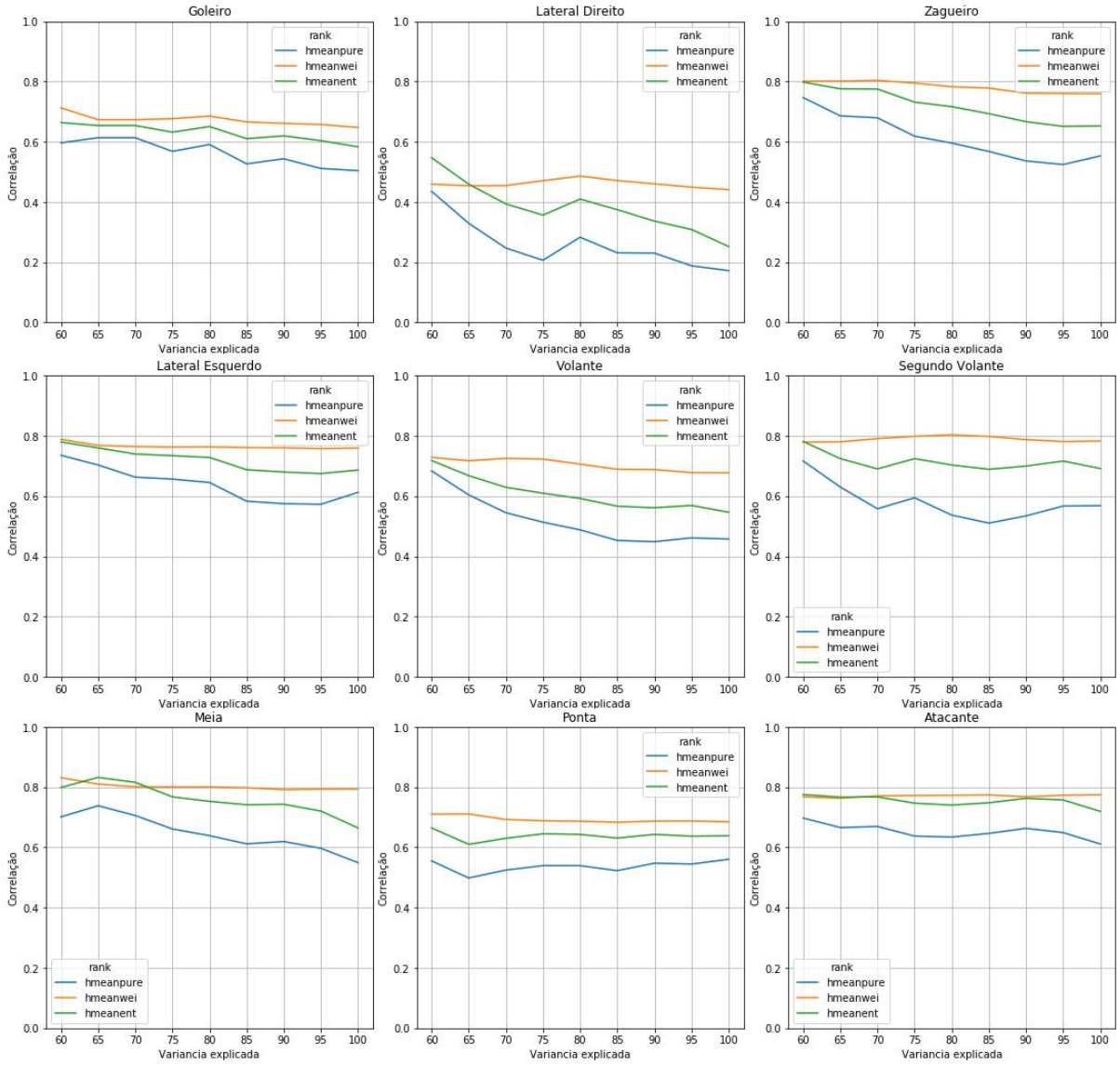


Figura 4.10: Comparação entre os métodos de ponderação utilizados a partir da correlação de Spearman entre o ranking Maior Média e o da plataforma InStat

Já no aspecto de selecionar os 5 melhores pelo método TOPSIS, os resultados variam mais, como podemos ver na figura 4.11. O método heurístico (verde) se mostra melhor para Laterais Direitos, Volantes e Atacantes nesse aspecto. O desempenho

para seleção dos melhores zagueiros, segundo volantes, meias e pontas se mostrou pobre para qualquer um dos três métodos.

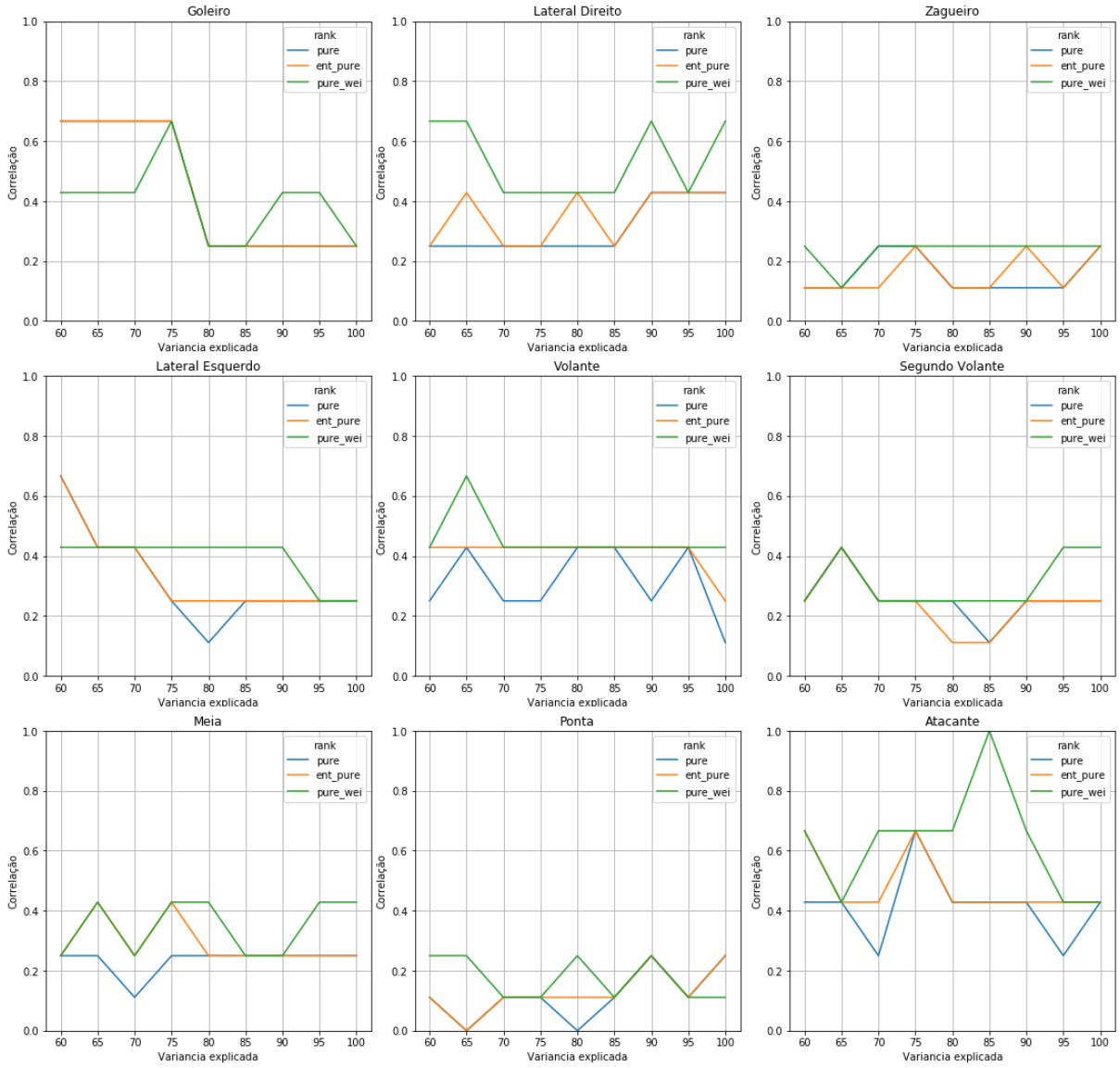


Figura 4.11: Comparação entre os métodos de ponderação utilizados a partir da distância de Jaccard entre o top-5 do ranking TOPSIS e o da plataforma InStat

Analisando a evolução da distância de Jaccard entre os rankings que usam agregação por Maior Média na figura 4.12, podemos perceber que o desempenho da seleção de meias e laterais esquerdos aumentam consideravelmente, indo para a casa dos 0.6. O método heurístico (amarelo) se mostra melhor para a seleção de meias, late-

rais esquerdos e atacantes, enquanto o entrópico se mostra melhor para zagueiros e segundos volantes.

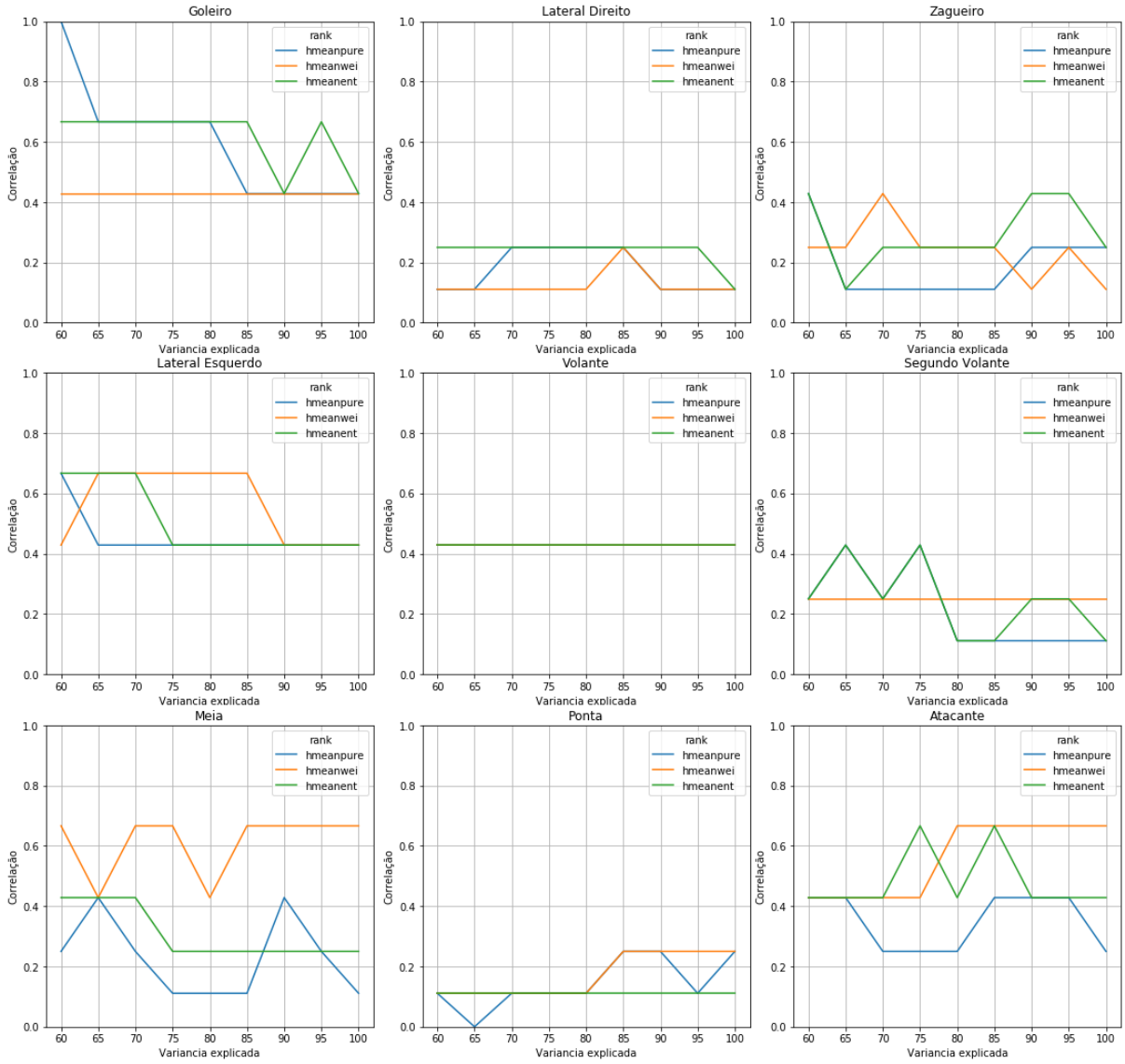


Figura 4.12: Comparação entre os métodos de ponderação utilizados a partir da distância de Jaccard entre o top-5 do ranking Maior Média e o da plataforma InStat

4.3 Teste do filtro de ações de baixa frequência

O filtro de ações de baixa frequência, detalhado na seção 3.5.2.2, tem como objetivo remover as ações que foram pouco realizadas pelos jogadores da comparação.

4.3.1 Aplicação direta

A figura 4.13 mostra como que cada posição reage a presença desse filtro na aplicação direta ao conjunto de dados originais, no aspecto da correlação média de Spearman entre os rankings gerados e o ranking de referência:

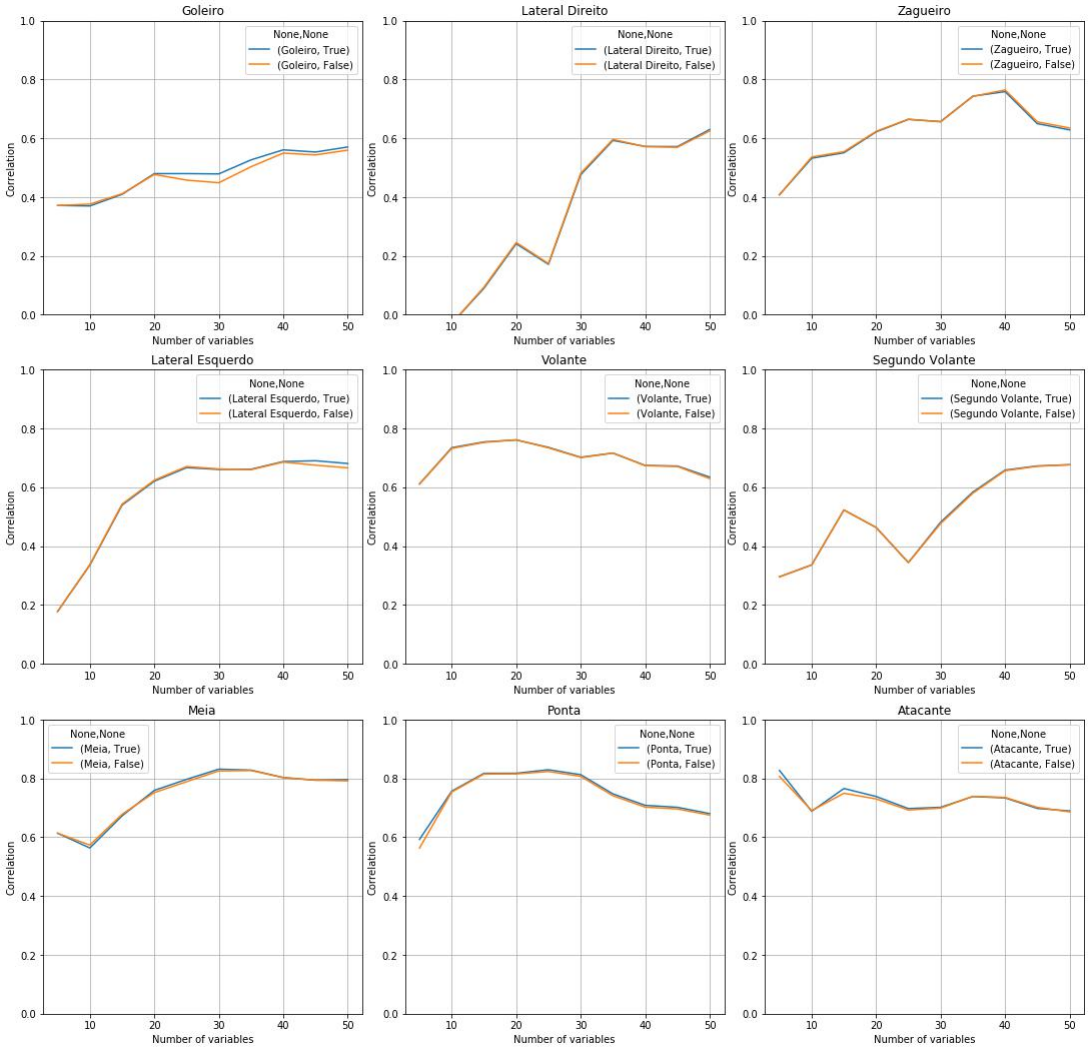


Figura 4.13: Comparação do uso ou não do filtro de ações de baixa frequência a partir da correlação média de Spearman entre os rankings gerados e o da plataforma InStat

Alguns casos chamaram atenção, como os goleiros, que tiveram uma diferença considerável, no entorno de 0.03. Isso pode se explicar por ações de jogadores de linha nas quais os goleiros foram submetidos (passe, drible). Por não serem frequentes para eles, atletas que as realizaram acabam se beneficiando bastante.

Já no aspecto de filtragem, a figura 4.14 mostra o impacto desse filtro nas distâncias jaccardianas médias entre o top-5 dos rankings gerados e o ranking de referência.

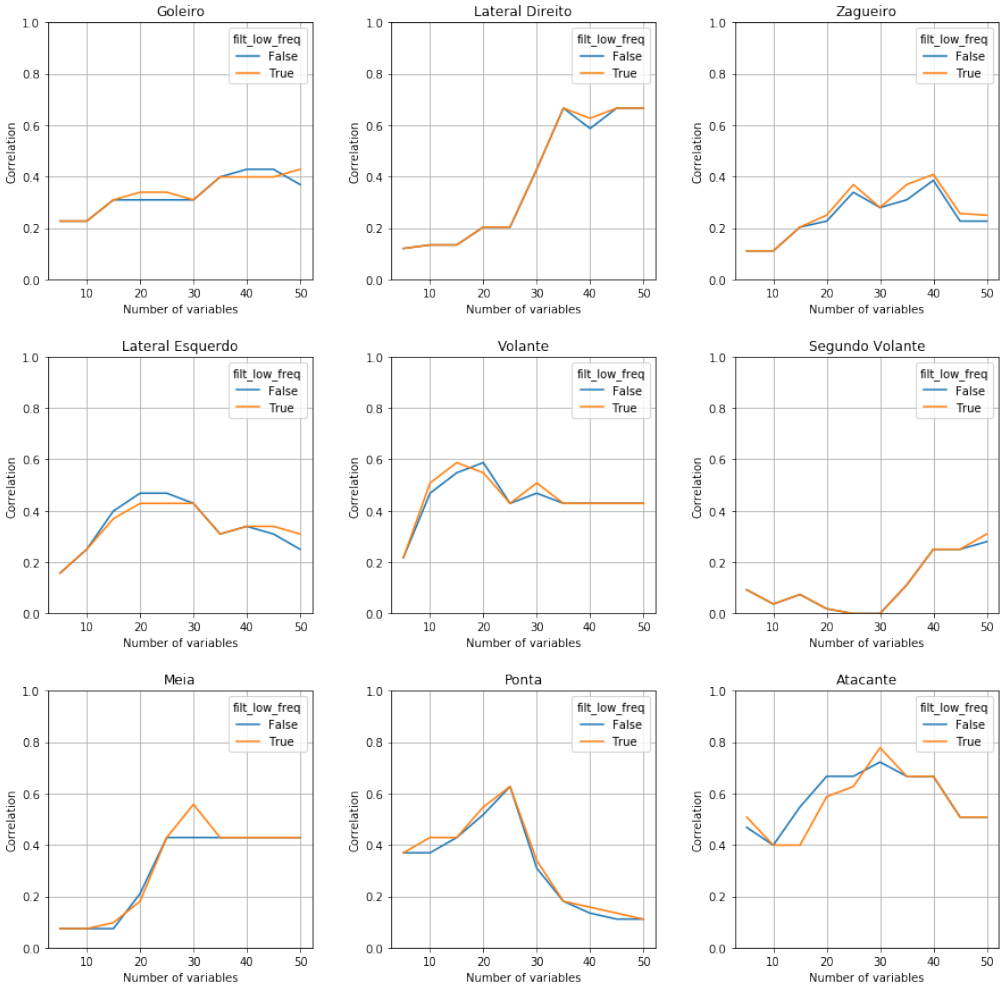


Figura 4.14: Comparação do uso ou não do filtro de ações de baixa frequência a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

O filtro de baixa frequência se mostrou útil (laranja), tendo resultados superiores nos picos de similaridade para cada posição em relação a não adoção (azul).

4.3.2 Aplicação com PCA

A figura 4.15 mostra a evolução da correlação de Spearman de acordo a quantidade de variância explicada acumulada para comparar a necessidade ou não da adoção do filtro de ações de baixa frequência.

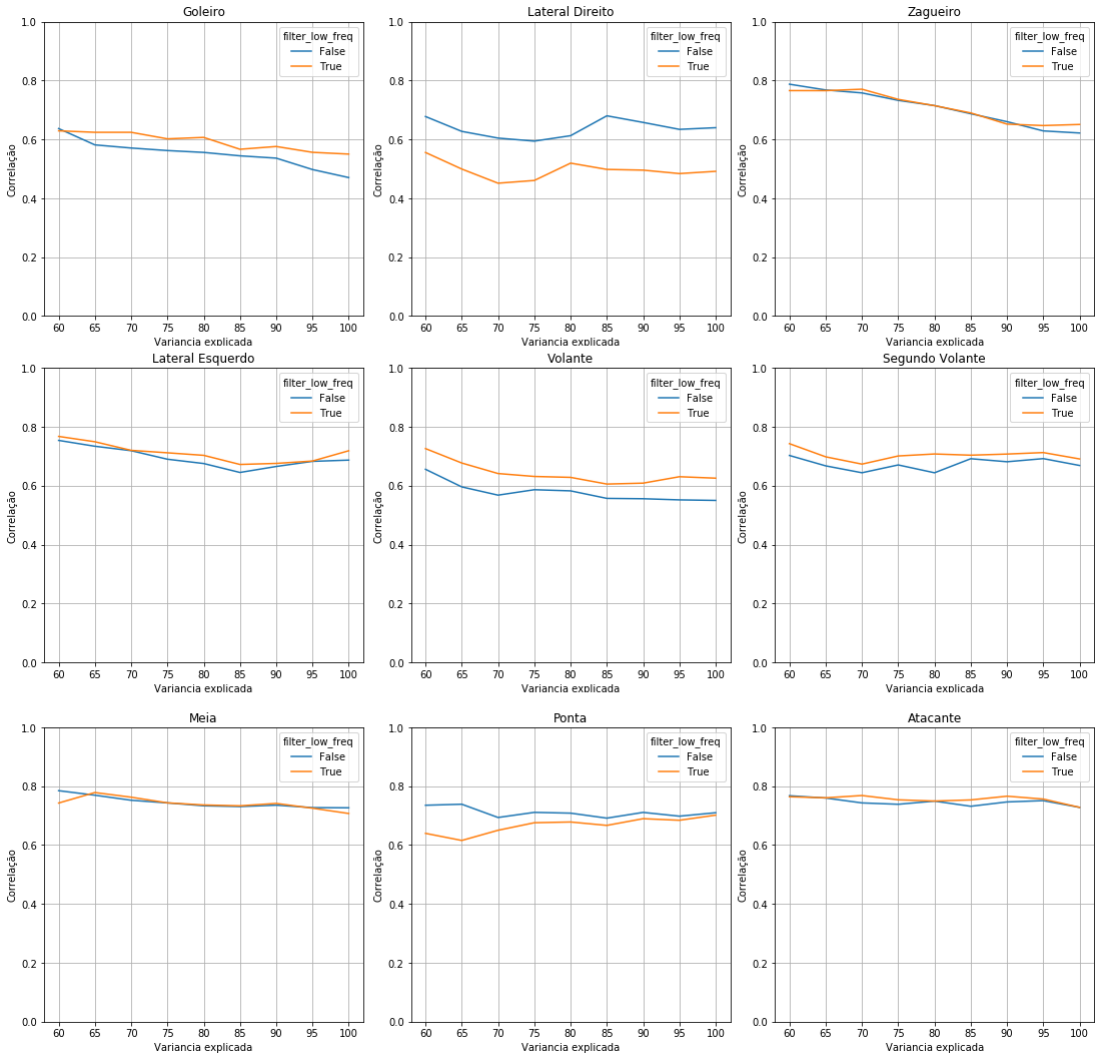


Figura 4.15: Comparação do uso ou não do filtro de ações de baixa frequência a partir da correlação média de Spearman entre os rankings gerados e o da plataforma InStat

Para os laterais direitos e para os pontas, não filtrar as ações de baixa frequência se mostrou uma melhor opção, com grande discrepância nos resultados. Para os goleiros, filtrar as ações se mostrou uma melhor opção, impactando em melhores resultados de correlação.

Sob o aspecto de filtragem, a figura 4.16 mostra a evolução da distância jaccardiana do top-5 de cada posição entre os rankings gerados e o da plataforma InStat. Nela, podemos reafirmar a necessidade de filtragem (laranja) para o caso dos goleiros, e o contrário pode ser dito para os laterais direitos, onde não filtrar (azul) se mostrou uma opção bem superior.

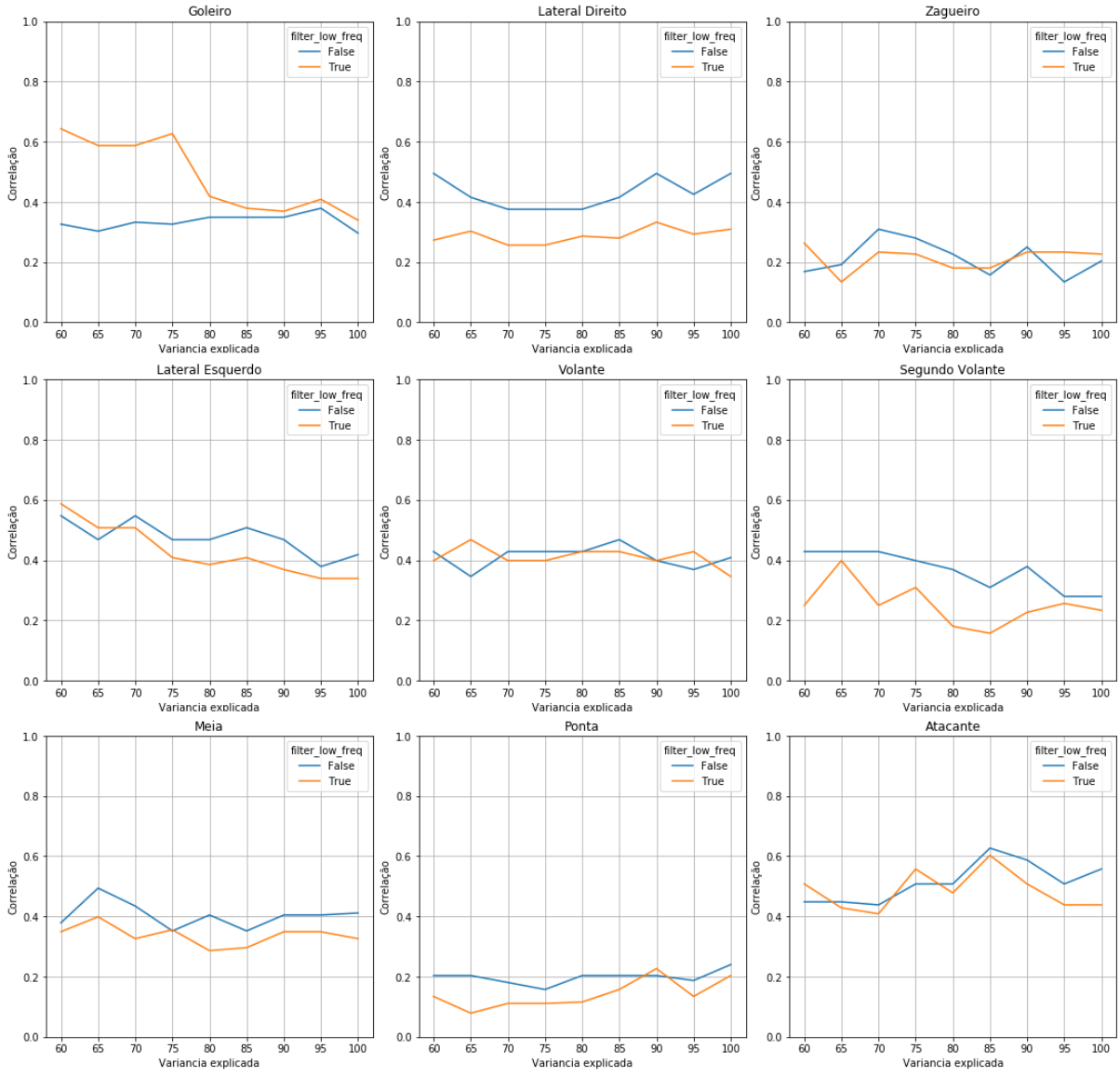


Figura 4.16: Comparação do uso ou não do filtro de ações de baixa frequência a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

4.4 Filtro de jogadores pouco participativos

4.4.1 Aplicação direta

O filtro de jogadores pouco participativos, detalhado na subsubseção 3.5.2.3, tem como objetivo testar o desempenho dos métodos quando se removem os jogadores com menores números de ações por partida. A figura 4.17 mostra como que cada posição reage a presença desse filtro, no aspecto da correlação média de Spearman entre os rankings gerados e o ranking de referência:

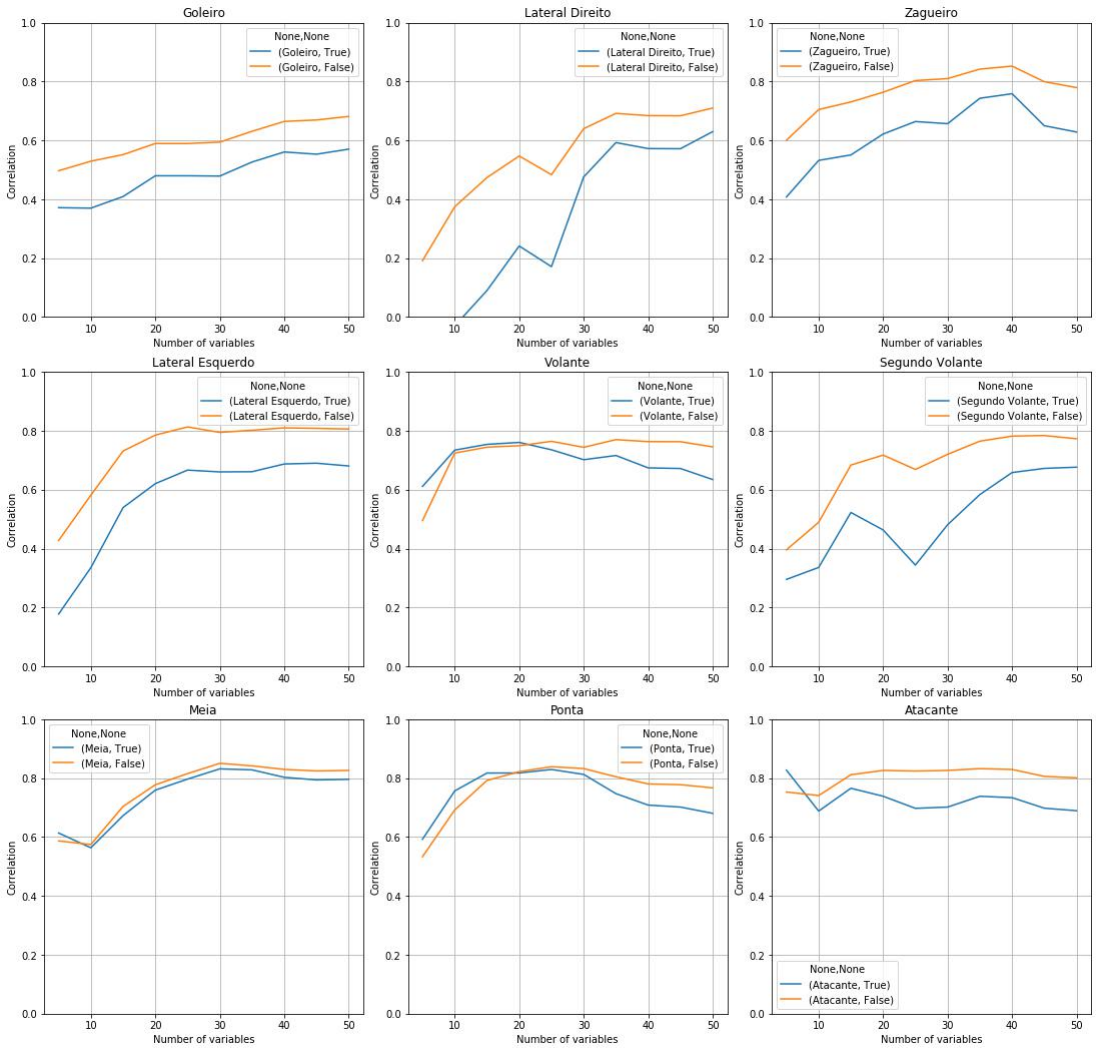


Figura 4.17: Comparação do uso do filtro de jogadores pouco participativos a partir da correlação média de Spearman entre os rankings gerados e o da plataforma InStat

O desempenho se mostrou muito superior sem o filtro de jogadores menos participativos (laranja), tendo diferenças de correlação maiores que 0.1.

Já no aspecto de filtragem, a figura 4.18 mostra o impacto desse filtro nas distâncias jaccardianas médias entre o top-5 dos rankings gerados e o ranking de referência:

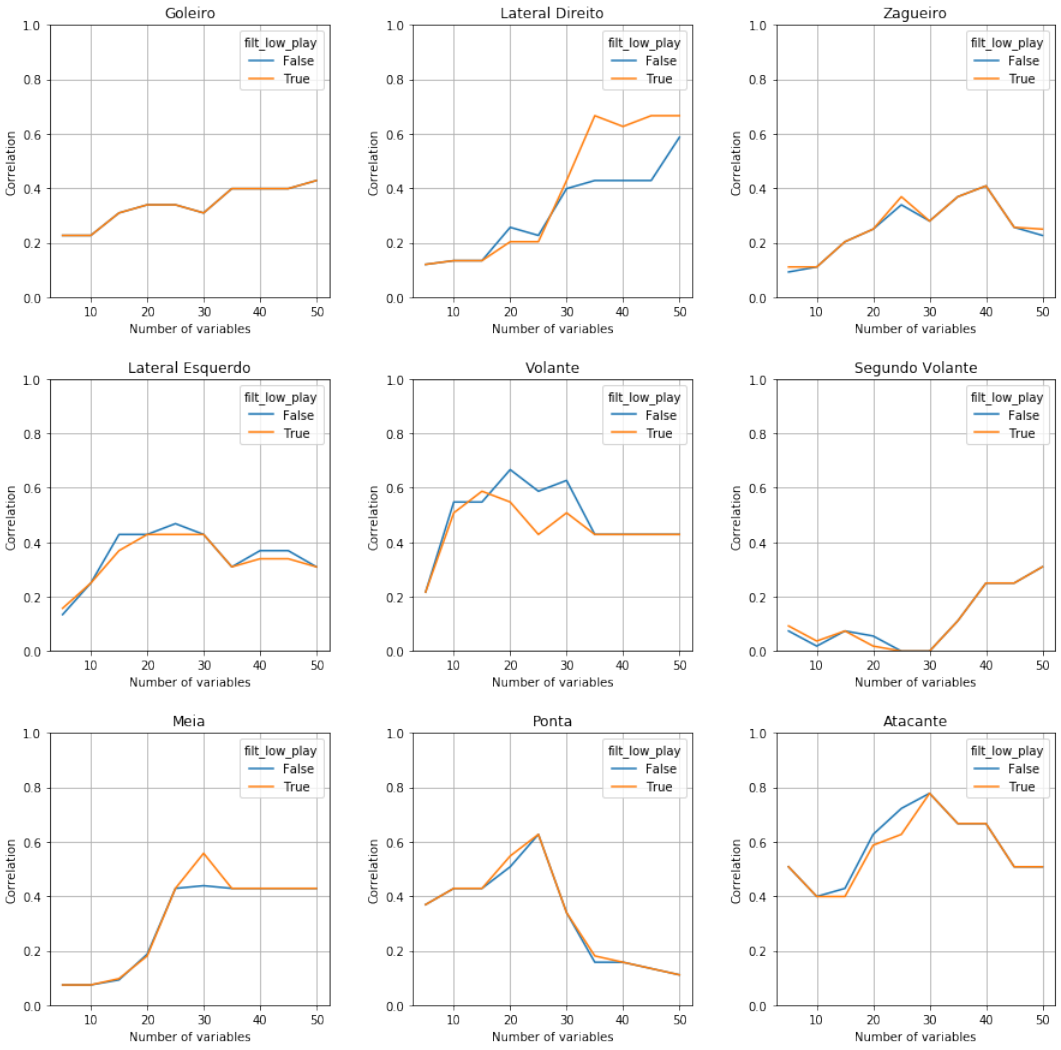


Figura 4.18: Comparação do uso do filtro de jogadores pouco participativos a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

Sob esta ótica, o filtro de participação mostrou ser irrelevante para quase todas as posições (exceto Lateral Direito), onde houve um impacto positivo (laranja).

4.4.2 Aplicação com PCA

A figura 4.19 mostra como que cada posição reage a presença ou não desse filtro nas variáveis geradas pelo PCA, no aspecto da correlação média de Spearman entre os rankings gerados e o ranking de referência:

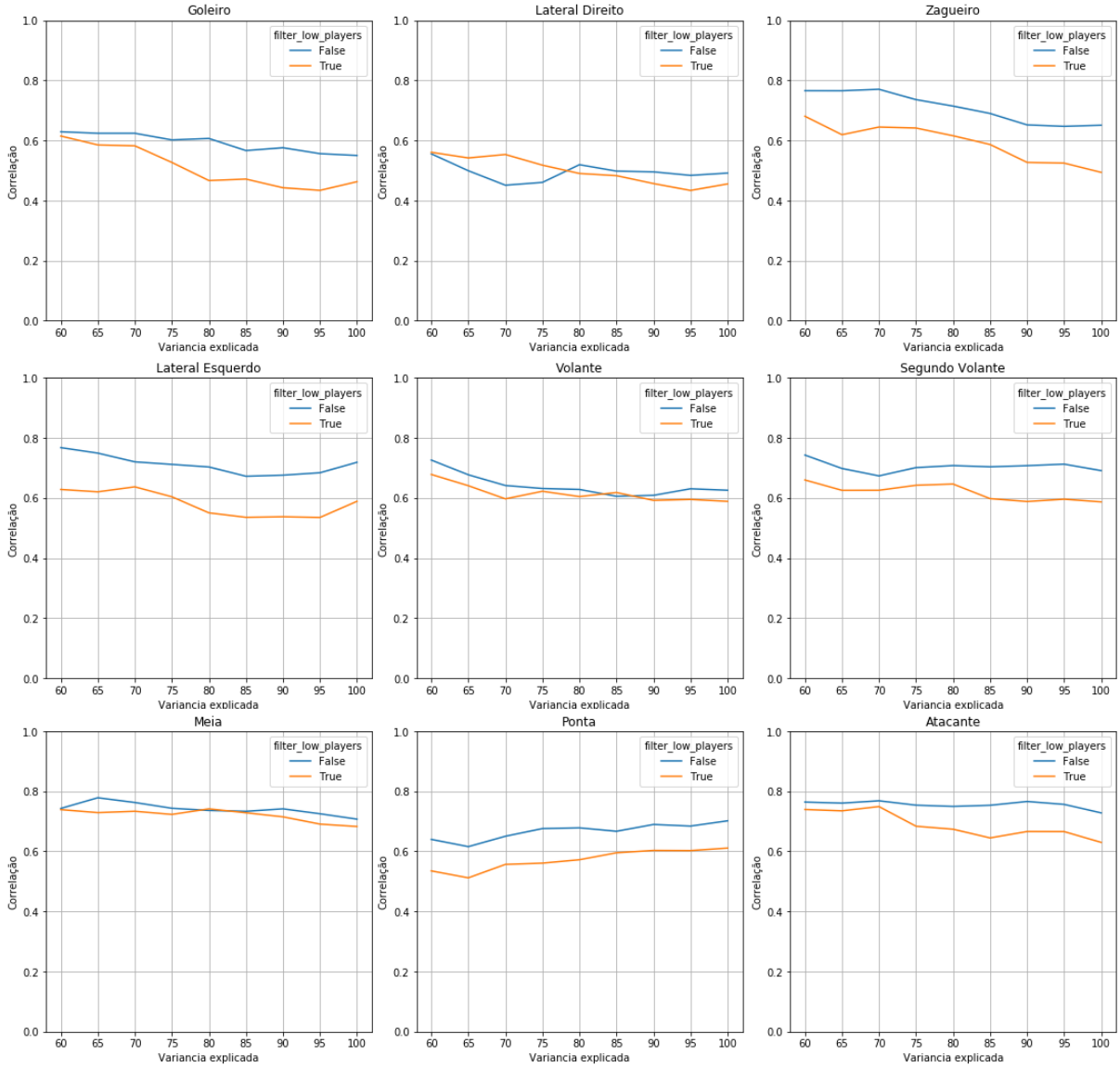


Figura 4.19: Comparação do uso do filtro de jogadores pouco participativos a partir da correlação média de Spearman entre os rankings gerados e o da plataforma InStat

O desempenho se mostrou muito superior sem o filtro de jogadores menos participativos (azul), tendo diferenças de correlação maiores que 0.1.

Já no aspecto de filtragem, a figura 4.20 mostra o impacto desse filtro nas distâncias jaccardianas médias entre o top-5 dos rankings gerados e o ranking de referência:

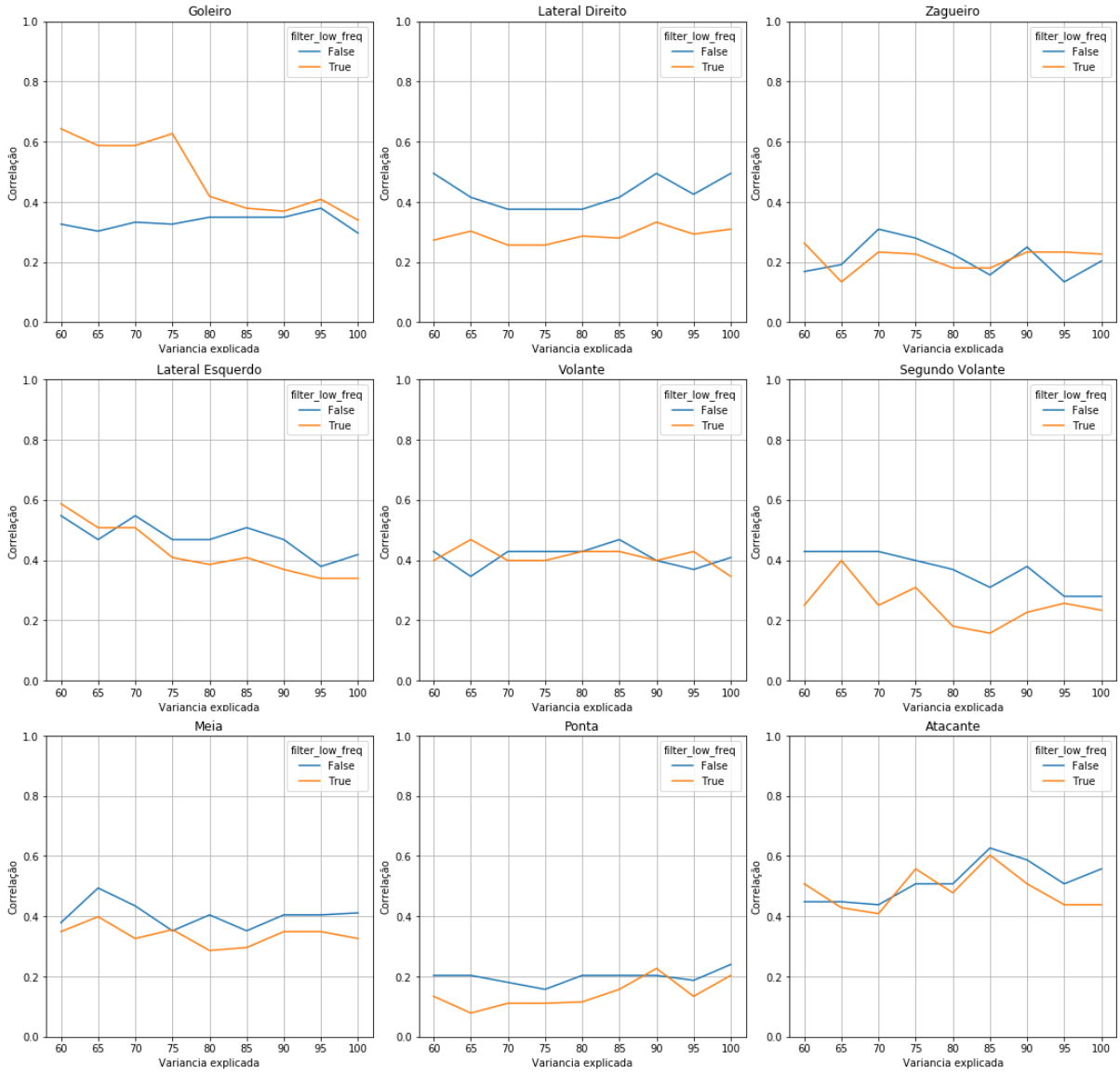


Figura 4.20: Comparação do uso do filtro de jogadores pouco participativos a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

Sob esta ótica, o filtro de participação mostrou impactar de forma negativa (azul) para quase todas as posições, exceto para os goleiros, onde houve um impacto positivo (laranja).

4.5 Definição da bonificação ótima para cada posição

4.5.1 Aplicação direta

Uma proposta é a de premiar jogadores mais participativos. Testa-se um fator a ser multiplicado ao valor de 1 (sem bonificação), para definir o limite da normalização Min-Max do conjunto de ações realizadas por jogador em cada habilidade. A figura 4.21 mostra a evolução dos valores de correlação para os quatro valores propostos (0.0 - azul, 0.5 - amarelo, 1.0 - verde, 1.5 - vermelho).

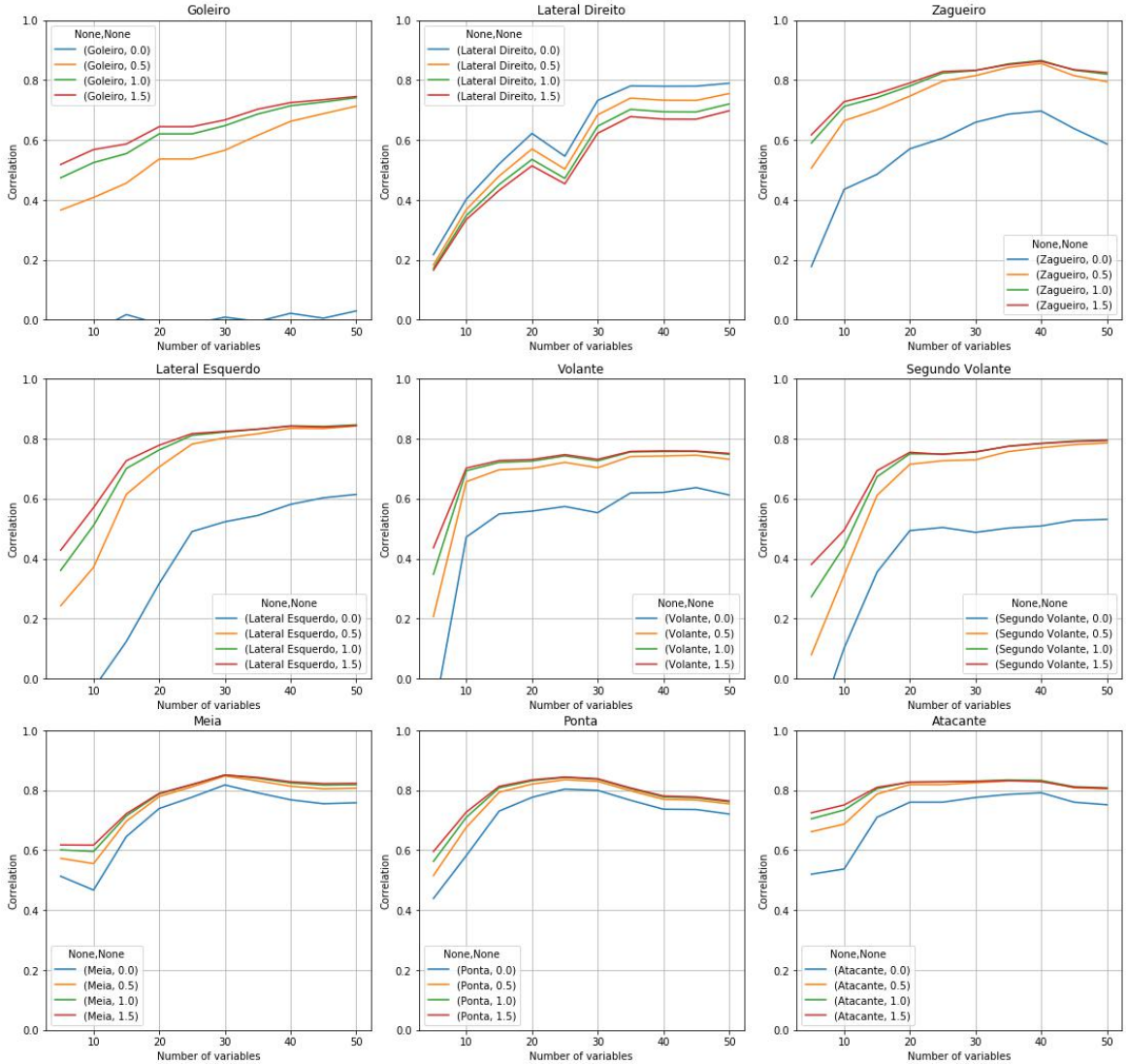


Figura 4.21: Comparação dos valores de bonificação a partir da correlação média entre os rankings gerados e o da plataforma InStat

Pode-se notar que para a maioria das posições, maiores bonificações implicaram em maiores correlações, exceto para os laterais direitos. Outro ponto a se destacar é que para um número menor de variáveis, o impacto da bonificação é maior do que para um número maior de habilidades (onde os valores são mais próximos).

Para analisar filtragem, a figura 4.22 mostra o impacto causado na distância entre o top-5 dos rankings gerados e o da referência. Observa-se que, exceto para os laterais direitos, a bonificação a esses jogadores se mostra necessária. Os picos de valores jaccardianos coincidem com os das maiores correlações.

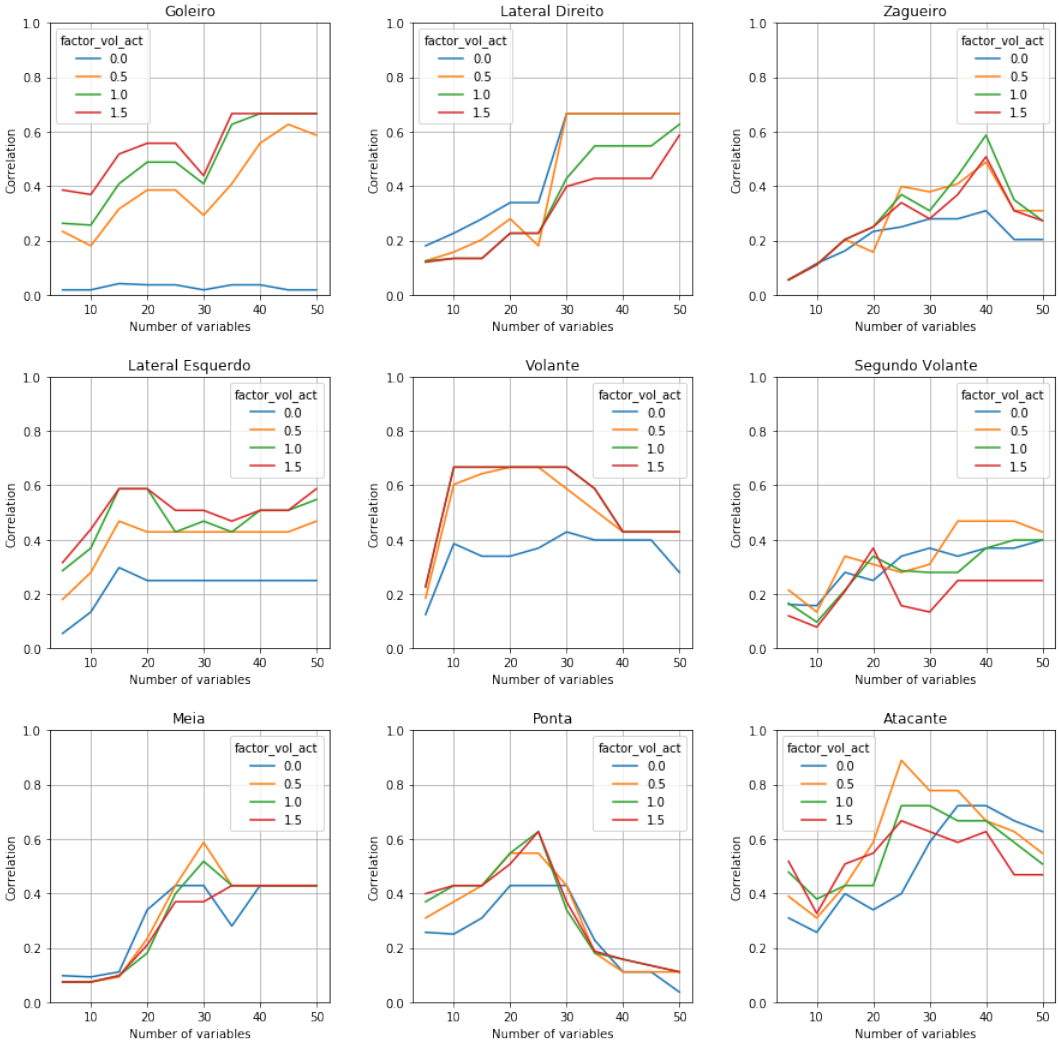


Figura 4.22: Comparação dos valores de bonificação a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

4.5.2 Aplicação com PCA

A figura 4.23 mostra a evolução dos valores de correlação para os quatro valores propostos para a bonificação (0.0 - azul, 0.5 - amarelo, 1.0 - verde, 1.5 - vermelho).

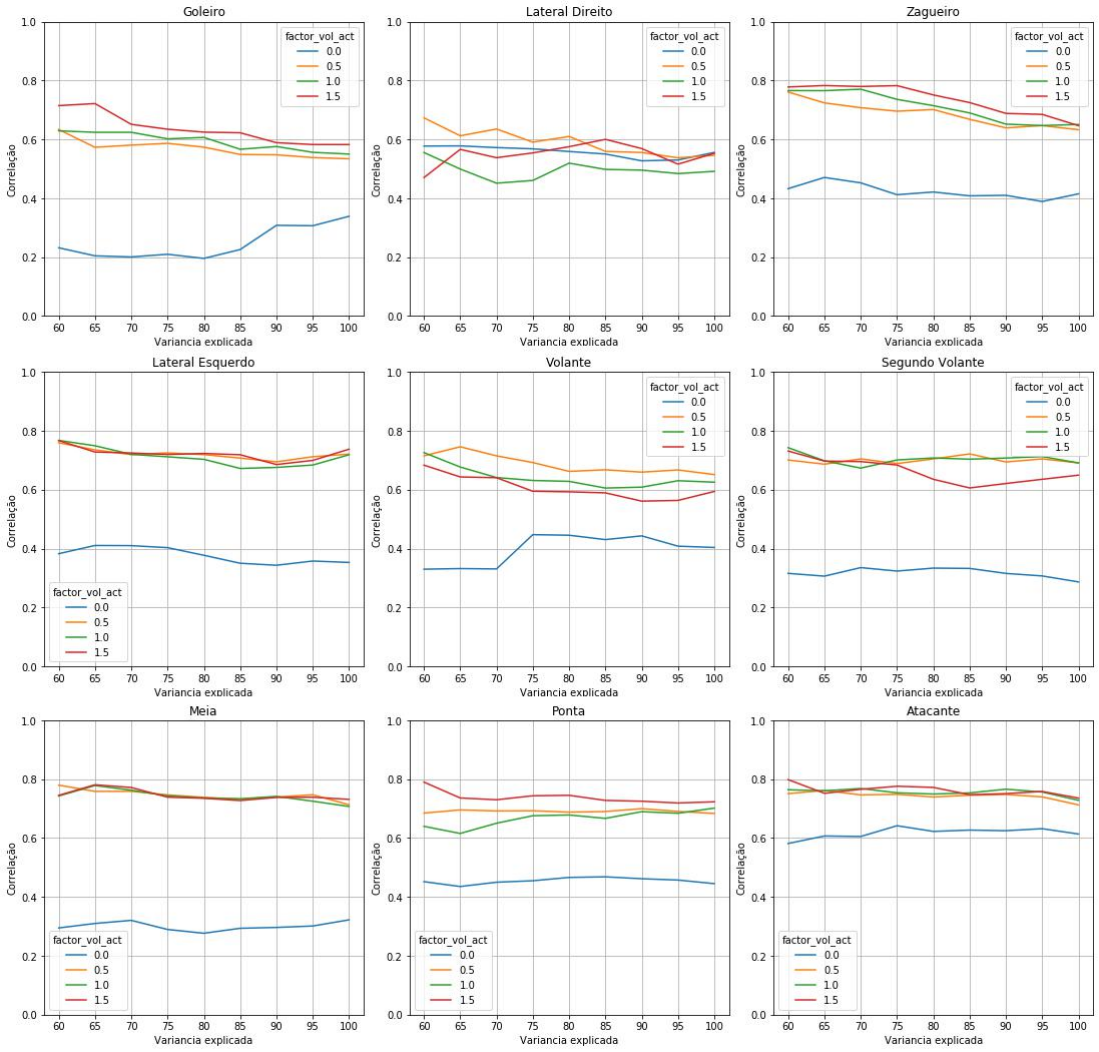


Figura 4.23: Comparação dos valores de bonificação a partir da correlação média entre os rankings gerados e o da plataforma InStat

Pode-se notar que para a maioria das posições, a bonificação se mostra necessária, tendo em vista o baixo desempenho para os casos onde ela não foi aplicada. Valores menores de variância explicada mostraram melhores rendimentos que os maiores, mostrando que nem todo o conjunto de dados é necessário.

Para analisar a filtragem, a figura 4.24 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência.

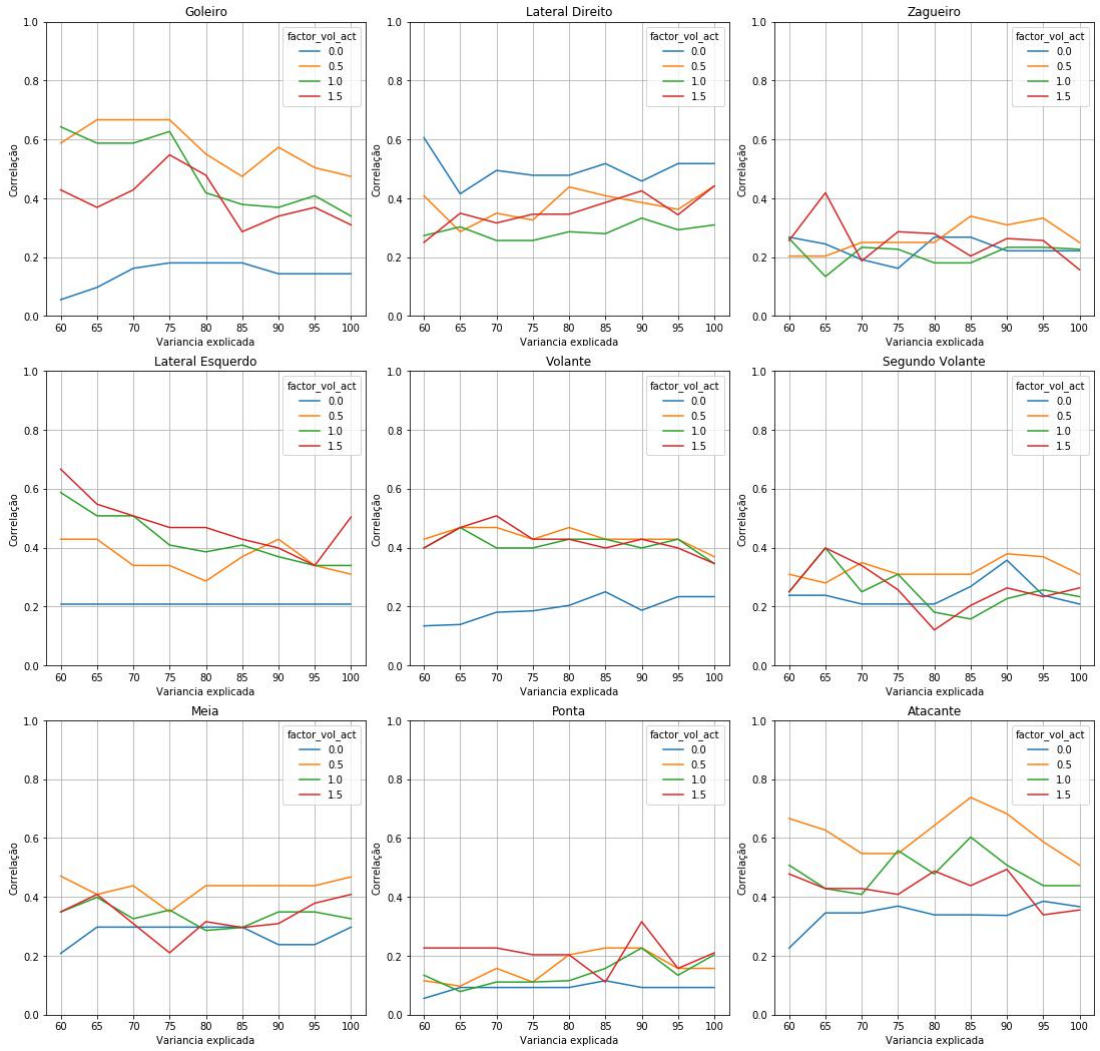


Figura 4.24: Comparação dos valores de bonificação a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

Observa-se que, exceto para o caso dos laterais direitos, a presença de uma bonificação a esses jogadores se torna necessária. O desempenho dessa métrica mostrou ter valores mais baixos com a aplicação do PCA do que sem o seu uso.

4.6 Comparando o efeito dos tipos de normalização

Outro teste é o do tipo de normalização prévio ao método de decisão multicritério. Foram testadas: Min-Max, por desvio padrão (Std) e a normalização vetorial.

4.6.1 Aplicação direta

A figura 4.25 mostra a evolução da correlação para os três tipos: Min-Max (verde), Std (azul) e Vnorm (vermelho). Nota-se que o desempenho de Min-Max e Std se mostrou semelhante para a grande maioria das posições, e superior ao Vnorm.

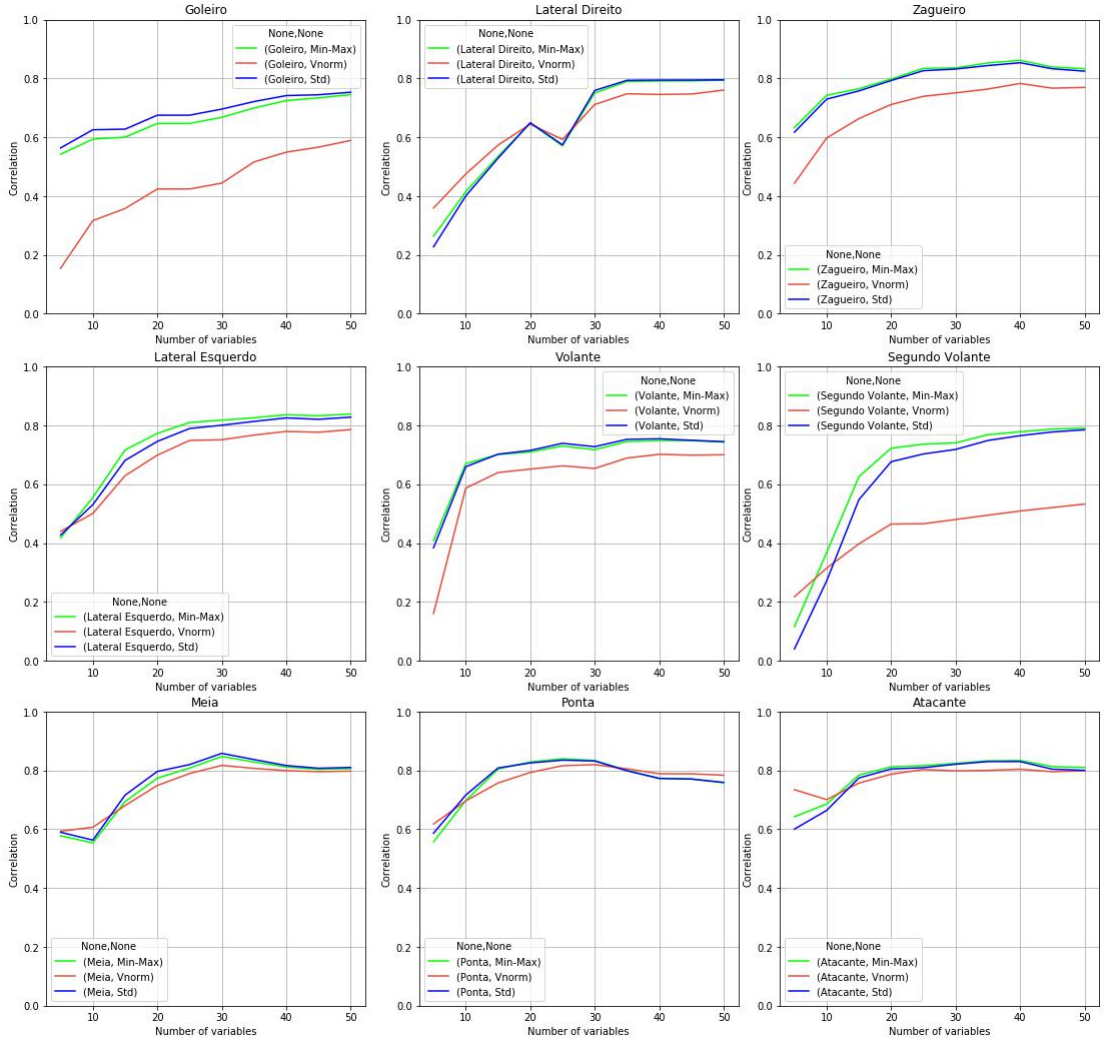


Figura 4.25: Comparação dos métodos de normalização prévia a partir da correlação média entre os rankings gerados e o da plataforma InStat

Para analisar a filtragem, a figura 4.26 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência. Nesse aspecto, os métodos Min-Max e Std mostraram rendimento ainda superior em relação ao VNorm. Para os atacantes e laterais esquerdos, o Min-Max mostrou impactar de forma mais positiva que o Std para selecionar os melhores atletas.

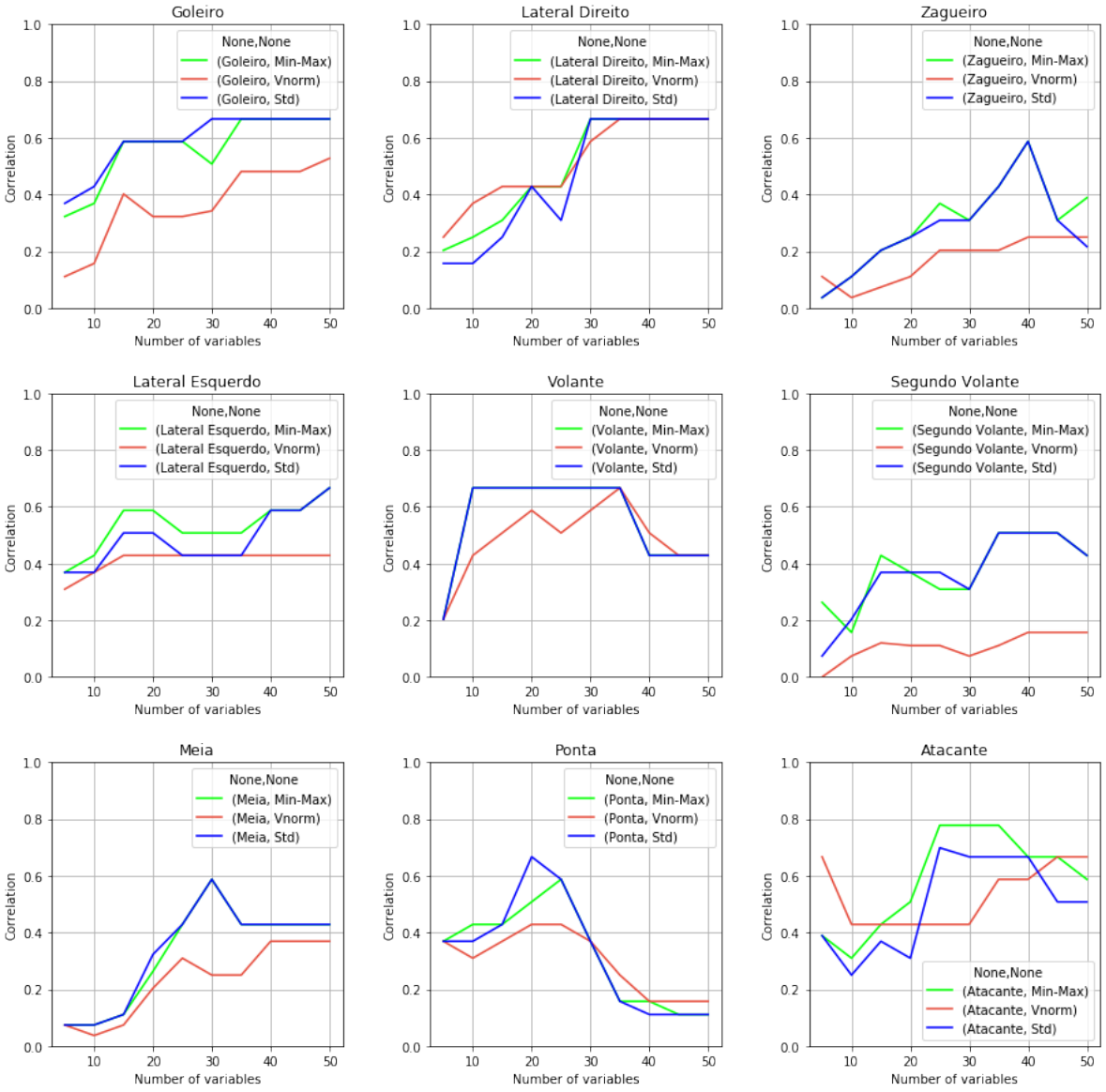


Figura 4.26: Comparação dos métodos de normalização prévia a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

4.6.2 Aplicação com PCA

A figura 4.27 mostra a evolução dos valores de correlação para as três tipos de normalização propostas (Min-Max - verde, Std - azul e Vnorm - vermelho) com a aplicação prévia do PCA.

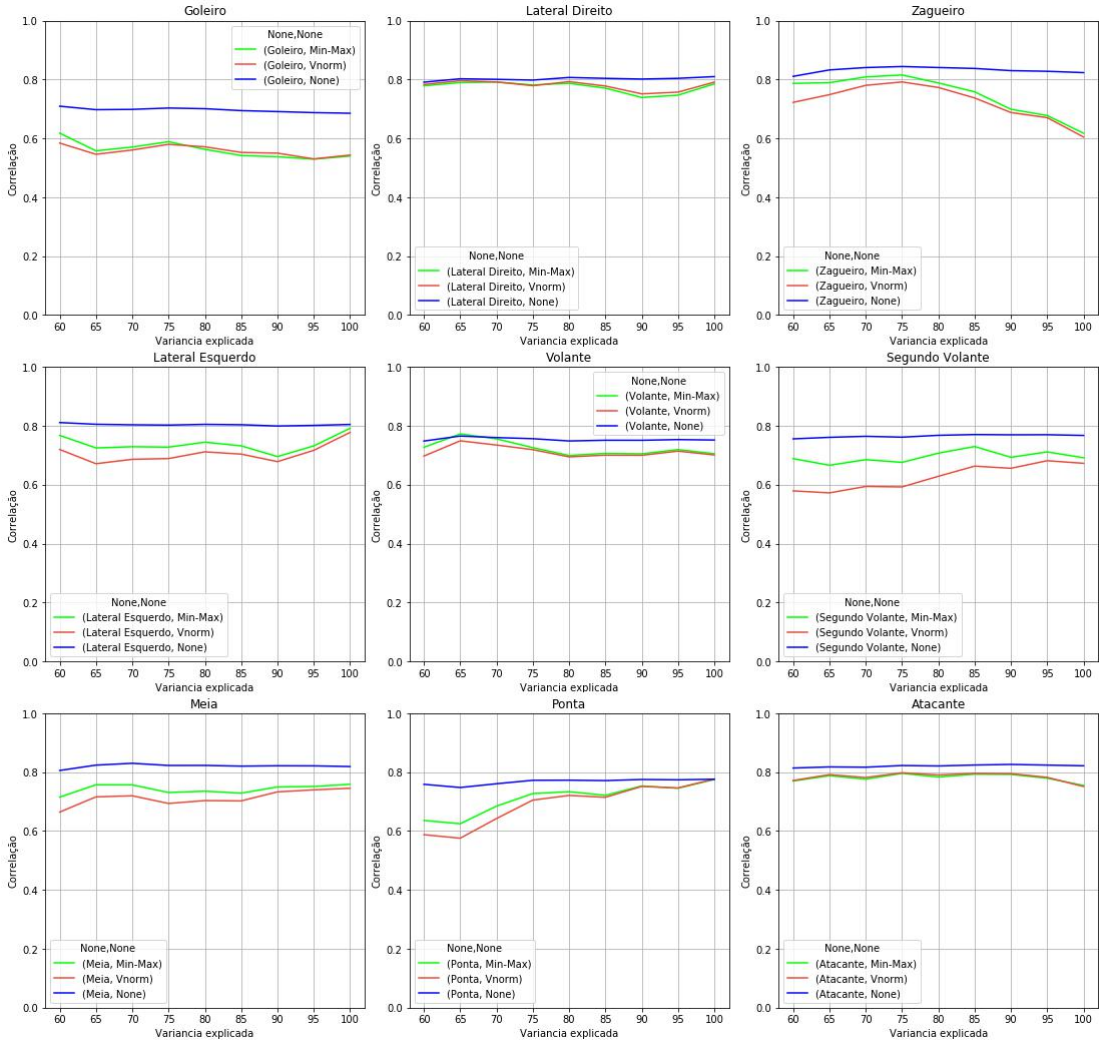


Figura 4.27: Comparação dos métodos de normalização prévia a partir da correlação média entre os rankings gerados e o da plataforma InStat

Pode-se notar que com a aplicação prévia do PCA, o método de normalização por desvio padrão (Std em azul) foi o que mostrou a maior correlação geral.

Para analisar a filtragem, a figura 4.28 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência.

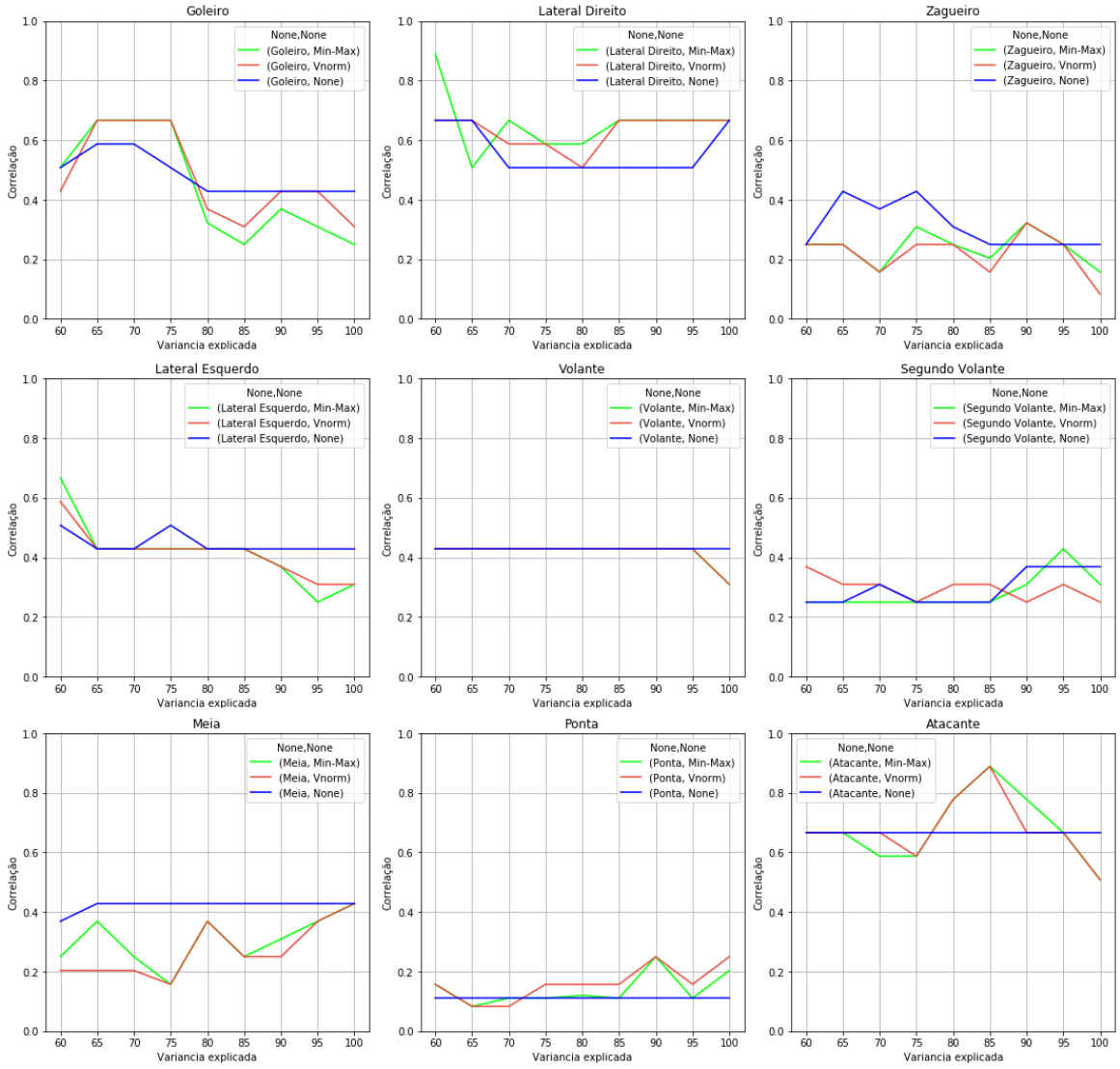


Figura 4.28: Comparação dos métodos de normalização prévia a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

Observa-se que, o resultado superior da normalização por desvio não se repete em todos os gráficos como no caso da correlação. Nesse aspecto, goleiros, laterais, segundo volantes, pontas e atacantes mostraram resultados superiores quando submetidos a Min-Max ou a VNorm.

4.7 Comparando os tipos de agregação de scores

Três tipos de agregação foram testados: TOPSIS, Maior Média e *Borda Count*. Para comparar a eficiência dos métodos, foi extraída a média das três ponderações utilizadas (ver seção 4.2).

4.7.1 Aplicação direta

A figura 4.29 mostra a evolução da correlação para os três tipos de agregação (TOPSIS - verde, Borda - azul e Maior média - vermelho). Nota-se a baixa diferença de desempenho dos métodos, sendo mais relevante para goleiros e laterais direitos.

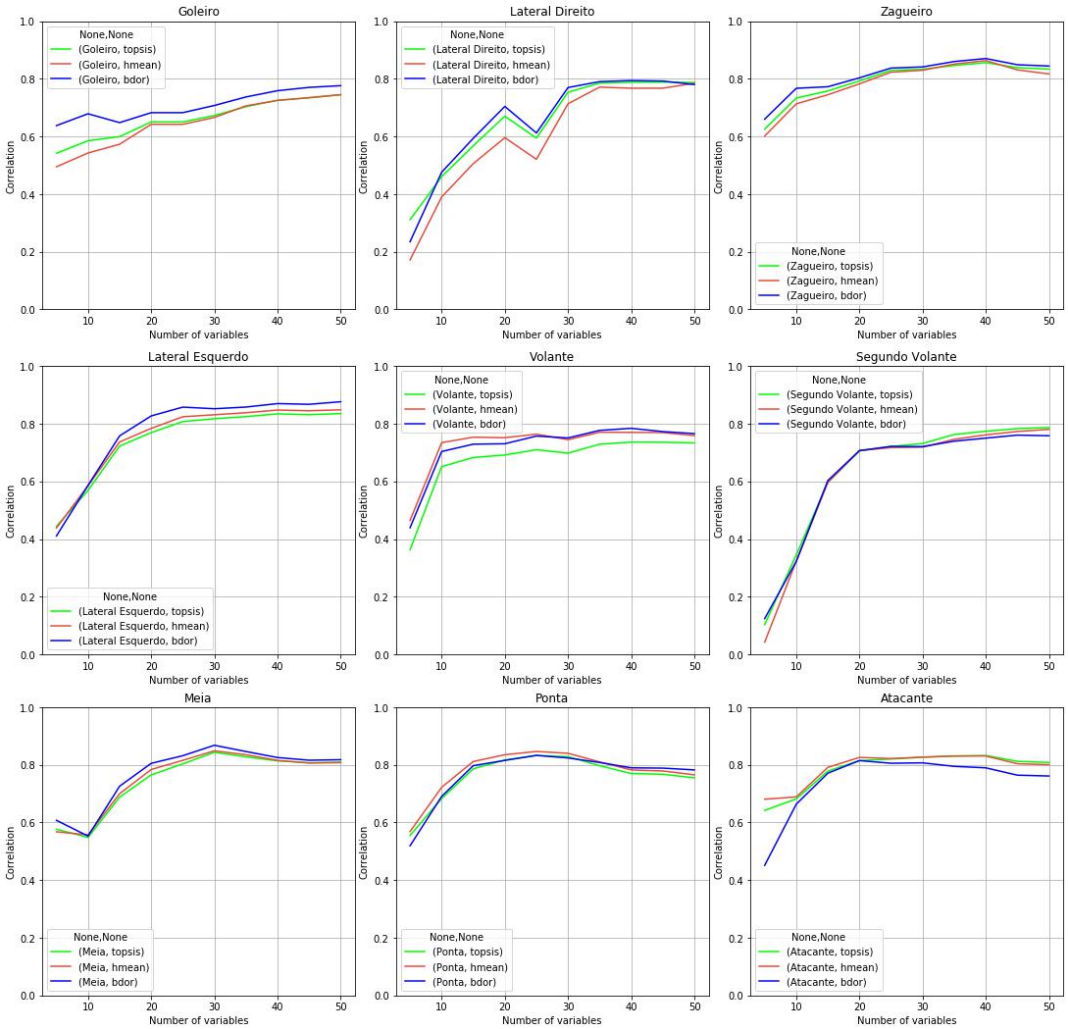


Figura 4.29: Comparação dos métodos de agregação de scores a partir da correlação média entre os rankings gerados e o da plataforma InStat

Para analisar a filtragem, a figura 4.30 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência. Nesse aspecto, o método de Contagem de Borda mostrou melhores resultados num geral, exceto para os Atacantes, onde se mostrou bem abaixo dos outros dois.

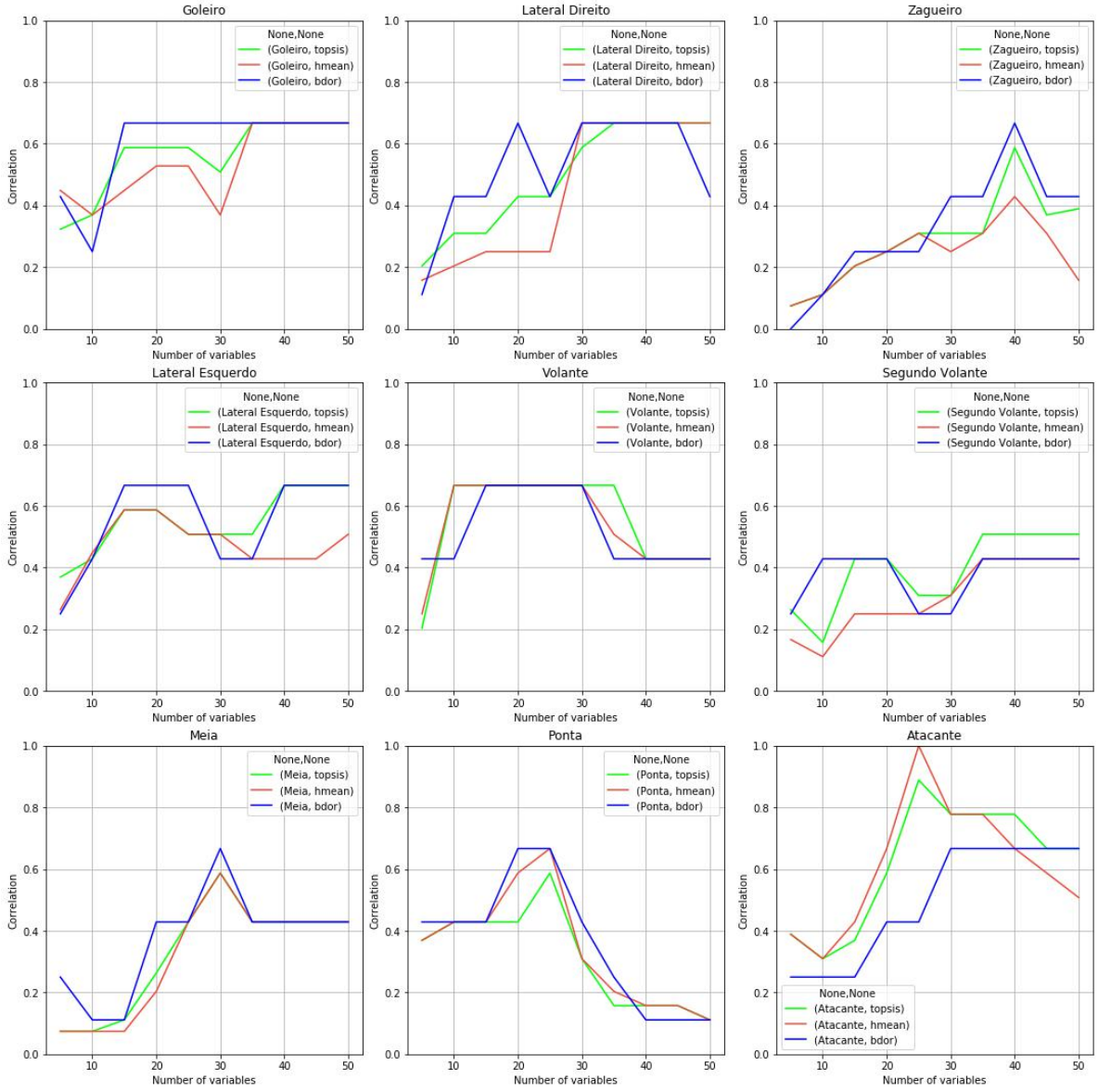


Figura 4.30: Comparação dos métodos de agregação de scores a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

4.7.2 Aplicação com PCA

A figura 4.31 mostra a evolução dos valores de correlação para os três tipos de agregação propostos (TOPSIS - verde, Borda - azul e Maior média - vermelho) com a aplicação do PCA.

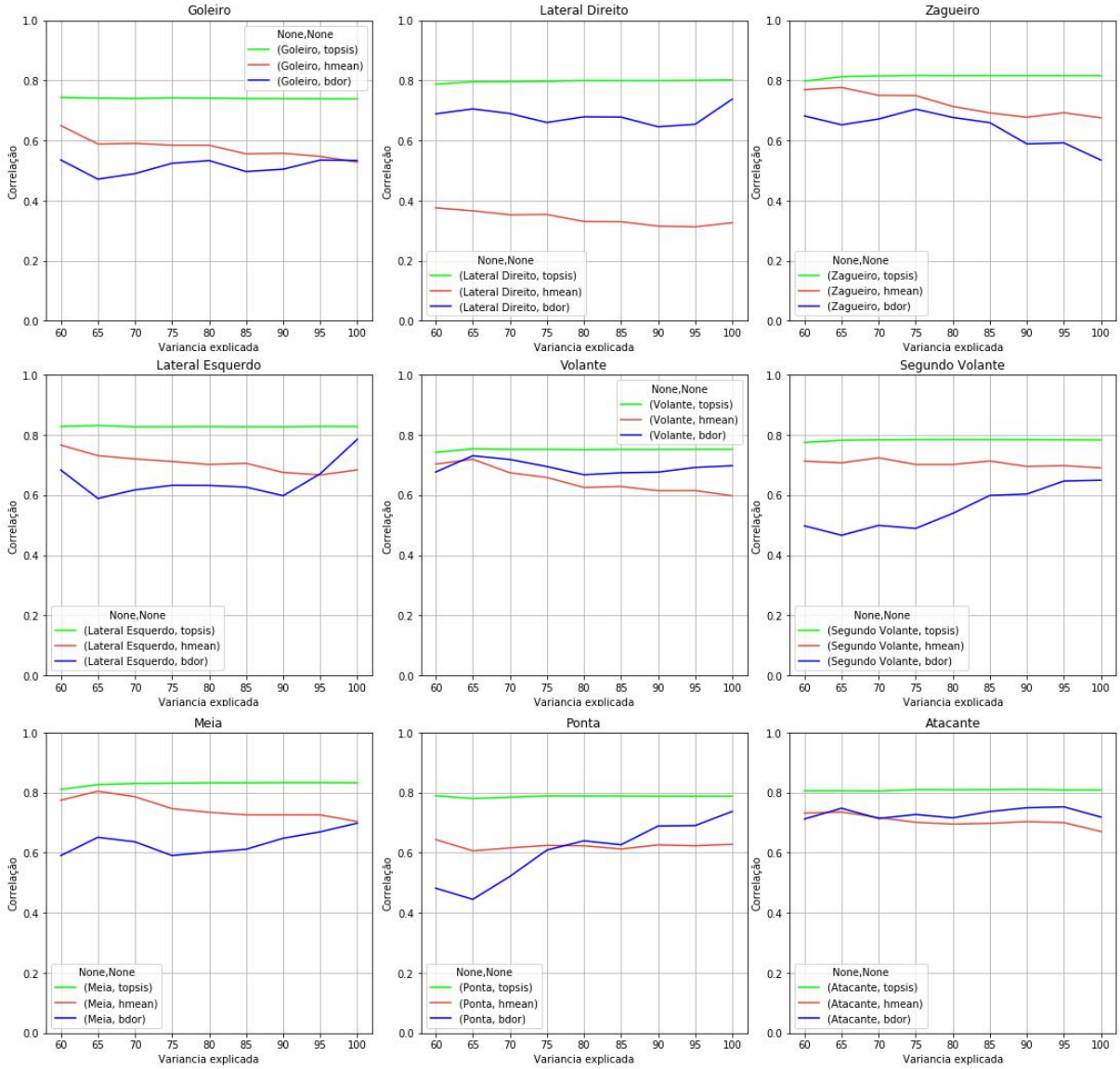


Figura 4.31: Comparação dos métodos de agregação de scores a partir da correlação média entre os rankings gerados e o da plataforma InStat

Pode-se notar que com a aplicação prévia do PCA, o método de agregação TOPSIS dominou em todas as posições, tendo larga discrepância em relação aos outros.

Para analisar a filtragem com o PCA, a figura 4.32 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência.

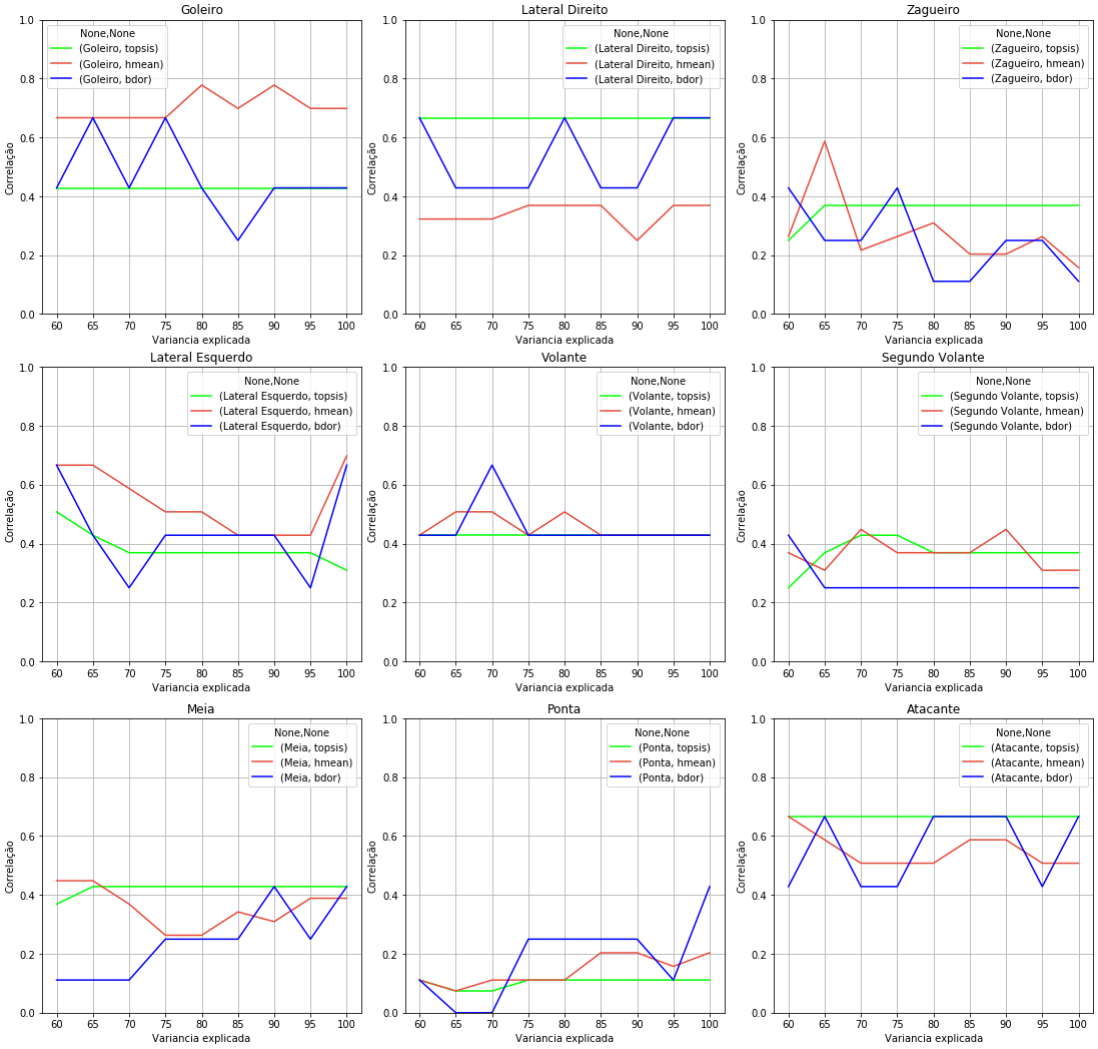


Figura 4.32: Comparação dos métodos de agregação de scores a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

Observa-se que o alto predomínio do TOPSIS não se repete, quando o aspecto é acertar os 5 melhores jogadores em relação ao ranking de referência. Os métodos de Maior Média e Contagem de Borda mostraram melhores resultados para tal. Para compensar os defeitos e potencializar as virtudes de cada método, uma meta-agregação dos métodos será realizada na seção 4.9

4.8 Comparando o efeito da norma da distância

Um teste também realizado é o da variação da norma p relativo a distância de Minkowski. Para tal, as normas 5, 4, 3, 2 e 1 serão testadas e avaliadas.

4.8.1 Aplicação direta

A figura 4.33 mostra a evolução dos valores de correlação para os cinco valores de norma p propostos (azul: 1, amarelo: 2, verde: 3, vermelho: 4 e roxo: 5)

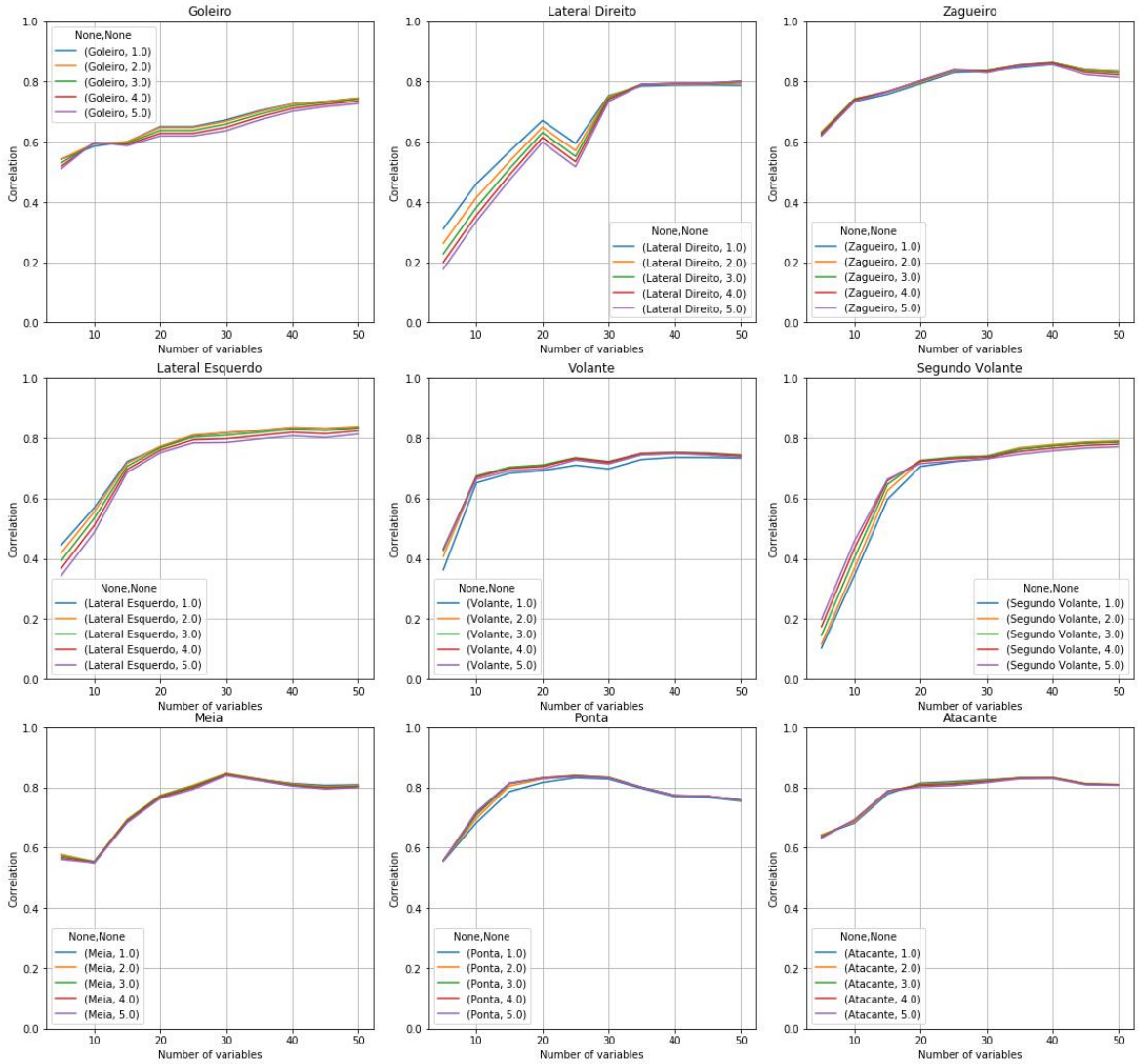


Figura 4.33: Comparação dos valores de norma p a partir da correlação média entre os rankings gerados e o da plataforma InStat

Percebe-se que o impacto do valor da norma é maior para dimensões menores de conjuntos de dados. Para os valores ótimos de correlação, a norma mostrou ter pouco impacto no resultado final.

Para analisar a filtragem, a figura 4.34 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência.

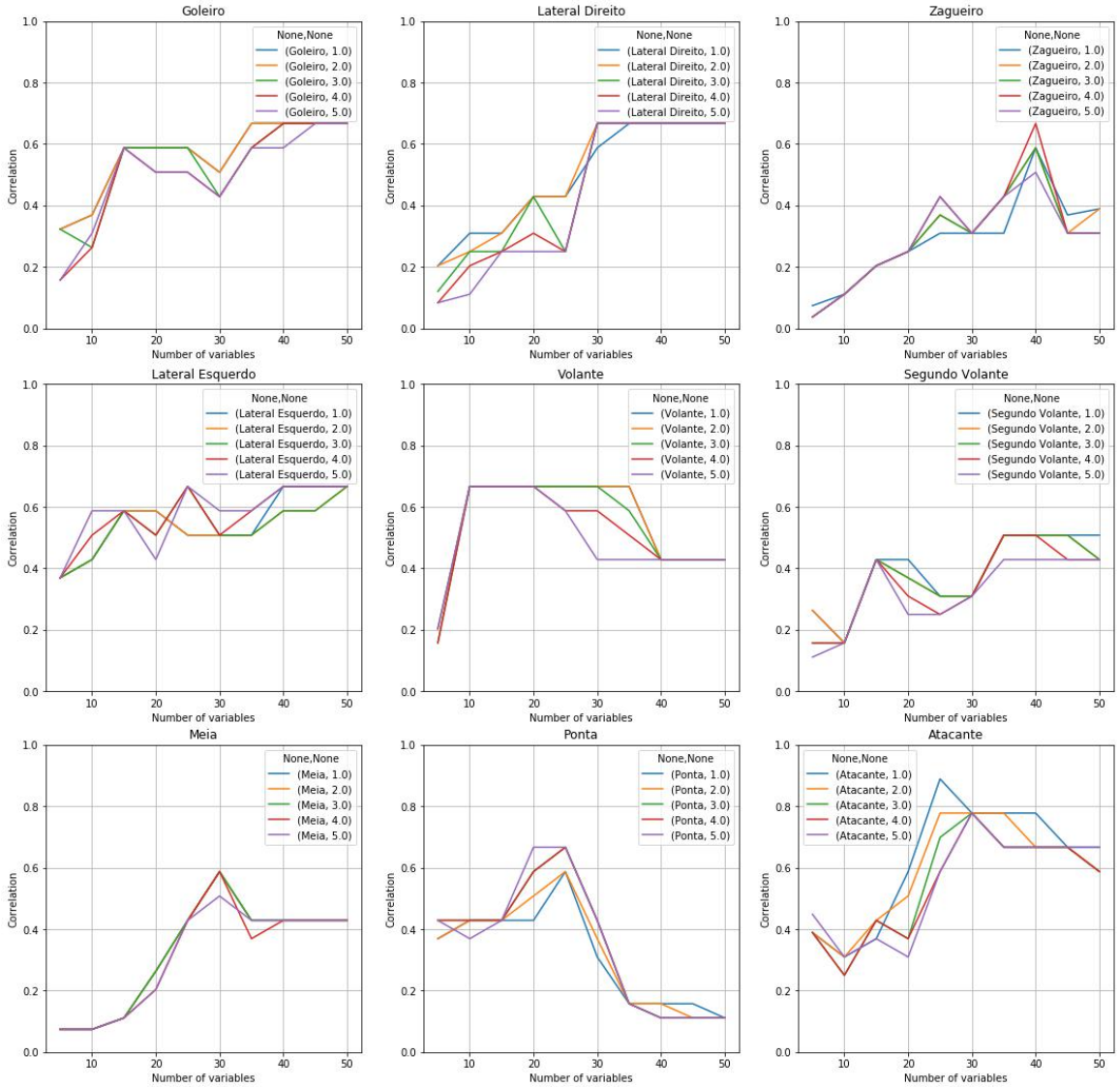


Figura 4.34: Comparação dos valores de norma p a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

Pode-se ver que o impacto do valor da norma nos valores ótimos é baixo, tendo em vista que na grande maioria das posições, todas as normas atingem o valor de pico.

4.8.2 Aplicação com PCA

A figura 4.35 mostra a evolução dos valores de correlação para os cinco valores de norma p propostos (azul: 1, amarelo: 2, verde: 3, vermelho: 4 e roxo: 5)

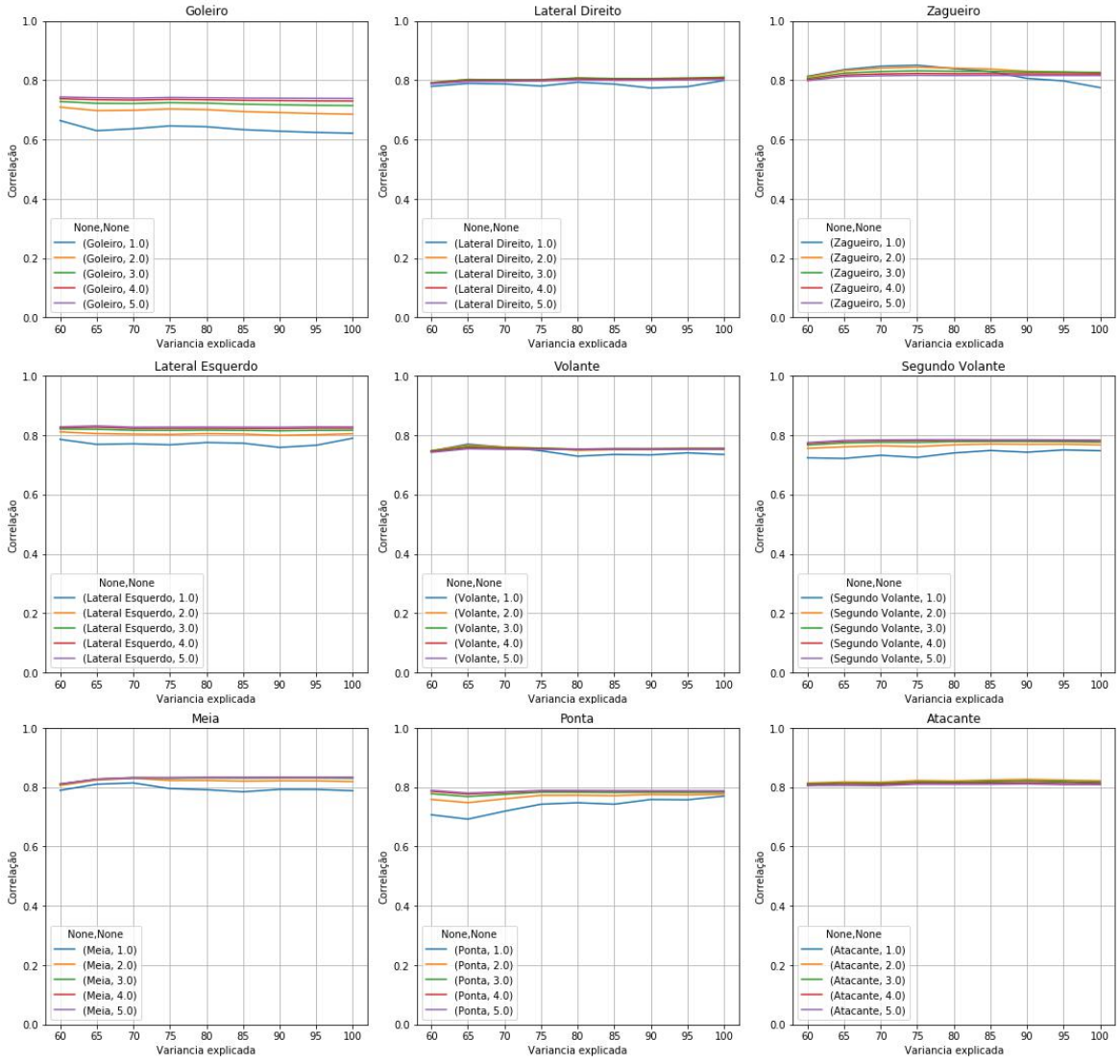


Figura 4.35: Comparação dos valores de norma p a partir da correlação média entre os rankings gerados e o da plataforma InStat

Observa-se que as normas 1 e 2 mostraram rendimento um pouco abaixo em relação as outras, que mostraram rendimento semelhante.

Para analisar a filtragem com o PCA, a figura 4.36 mostra o impacto causado na distância jaccardiana. O impacto causado pelo valor da norma se mostrou bem baixo, aplicando pouca variação nos valores jaccardianos.

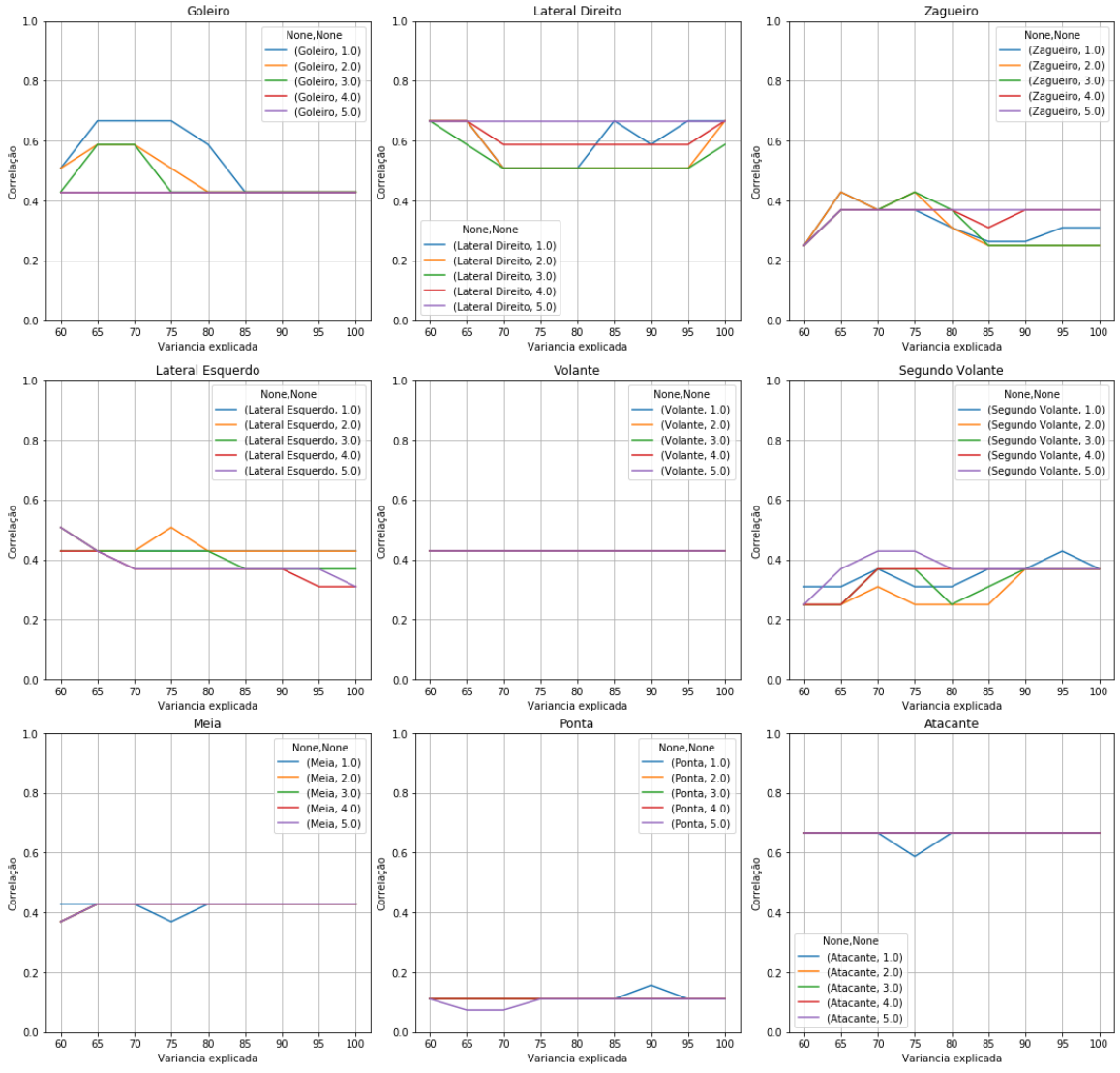


Figura 4.36: Comparação dos valores de norma p a partir da distância média de Jaccard entre os top-5 dos rankings gerados e o da plataforma InStat

4.9 Geração do ranking único

Para finalizar, testaremos os três métodos de agregação propostos para meta-agregar os scores gerados por cada agregação, gerando um ranking único.

4.9.1 Aplicação direta

A figura 4.37 mostra a evolução da correlação para os três tipos de meta-agregação propostos (azul: TOPSIS, laranja: Maior Média e verde: Contagem de Borda)

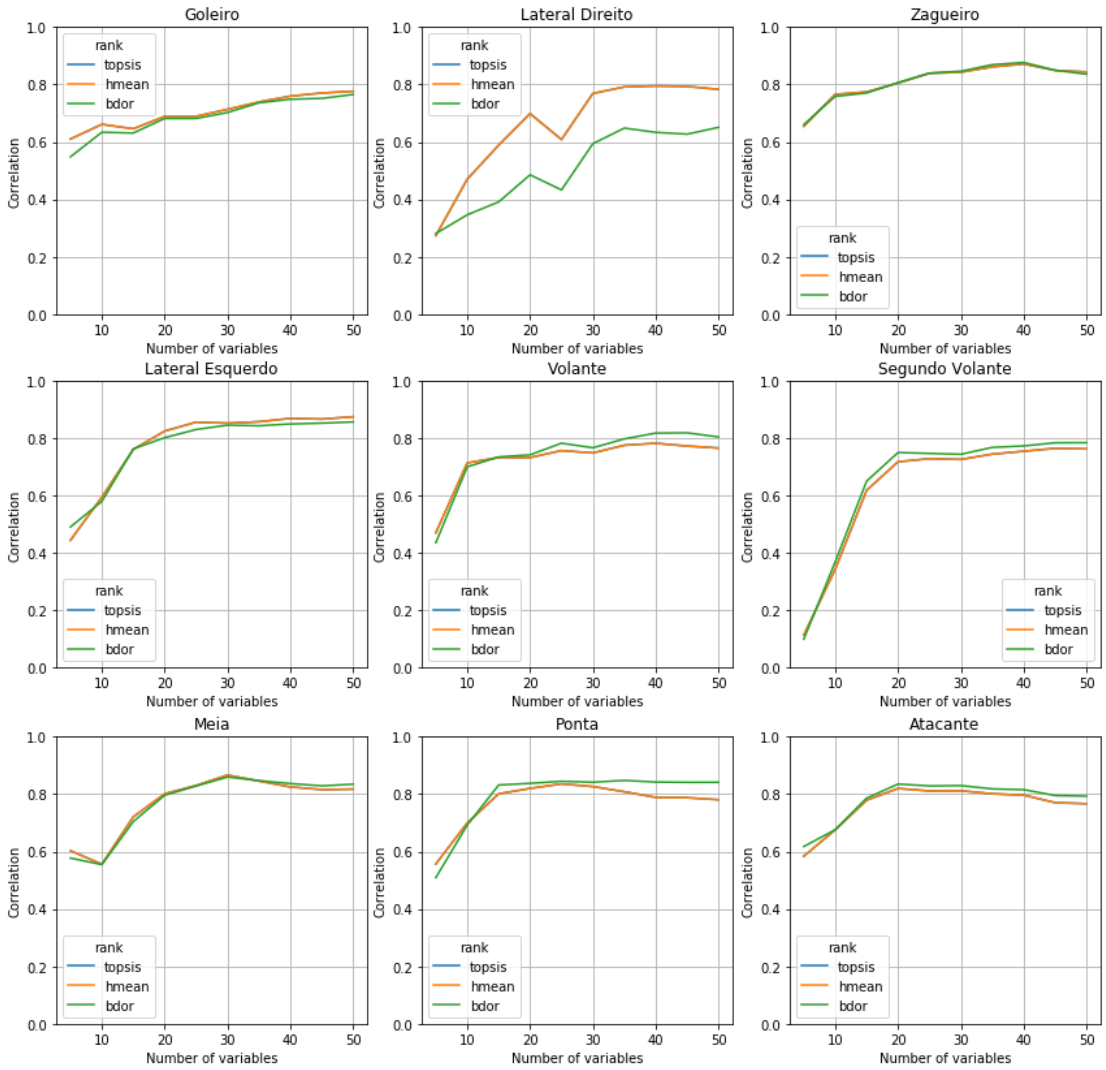


Figura 4.37: Comparação dos tipos de meta-agregação a partir da correlação entre o ranking meta-gerado e o da plataforma InStat

Percebe-se que TOPSIS e Maior média possuem valores exatamente iguais, indicando rankings iguais. Para jogadores de defesa, o Maior Média/TOPSIS mostrou melhores resultados, enquanto para jogadores de ataque, a Contagem de Borda mostrou melhor desempenho.

Para analisar a filtragem, a figura 4.38 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência.

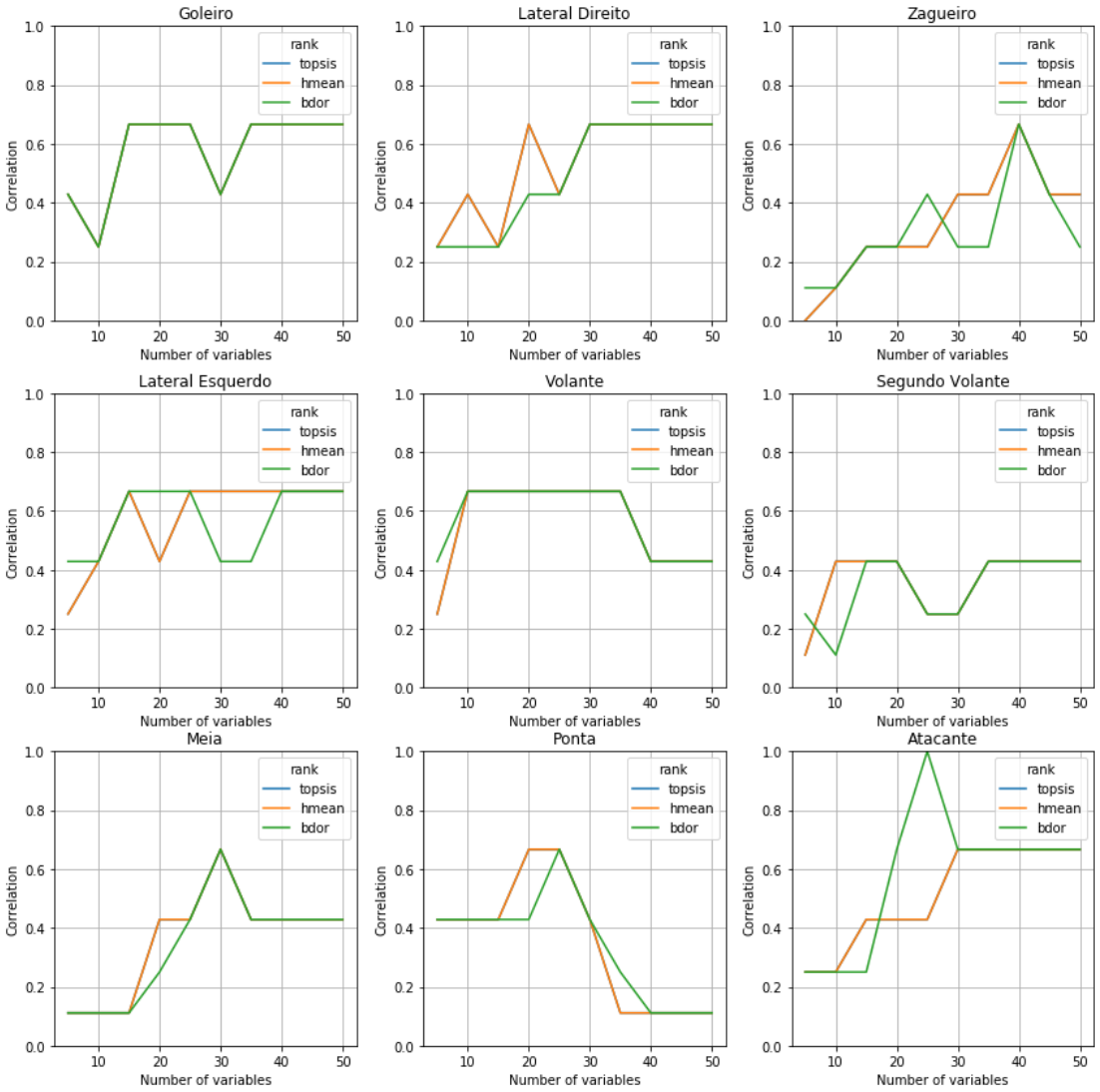


Figura 4.38: Comparação dos tipos de meta-agregação a partir da distância de Jaccard entre o top-5 do ranking meta-gerado e o da plataforma InStat

Pode-se ver que o impacto na filtragem de acordo o tipo de agregação é pequeno, com os três métodos tendo resultados similares. Nas poucas diferenças, seguiu-se a lógica da correlação. Jogadores de defesa tiveram melhor desempenho com Maior Média/TOPSIS, enquanto jogadores de ataque mostraram melhor filtragem com Contagem de Borda.

4.9.2 Aplicação com PCA

A figura 4.39 mostra a evolução da correlação para os três tipos de meta-agregação propostos (azul: TOPSIS, laranja: Maior Média e verde: Contagem de Borda)

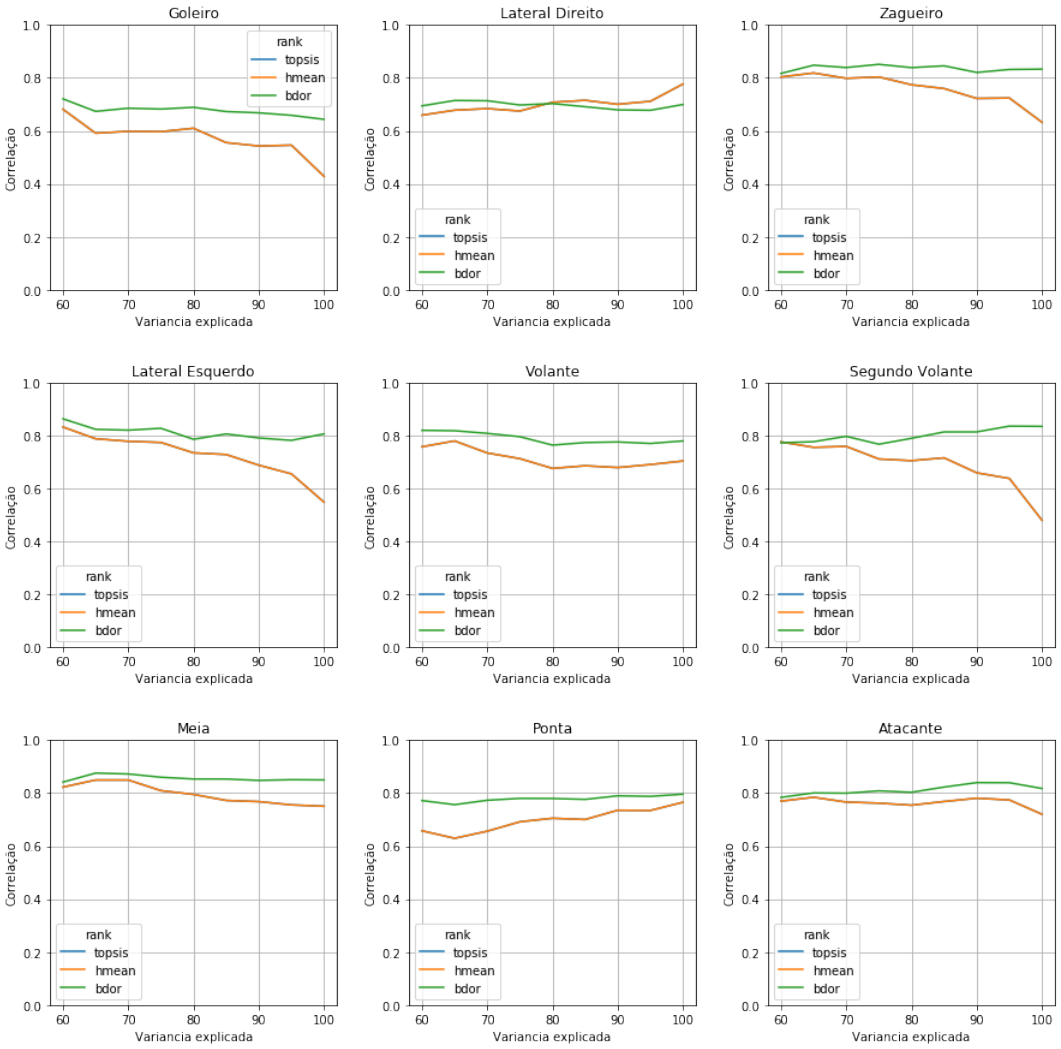


Figura 4.39: Comparação dos tipos de meta-agregação a partir da correlação entre o ranking meta-gerado e o da plataforma InStat

Observa-se que a meta-agregação por Contagem de Borda mostrou melhores resultados que Maior Média/TOPSIS em todas as posições.

Para analisar a filtragem com o PCA, a figura 4.40 mostra o impacto causado na distância jaccardiana entre o top-5 dos rankings gerados e o da referência. A contagem de Borda mostrou ter melhor desempenho na maioria das posições, com exceção de atacantes, pontas e volantes.

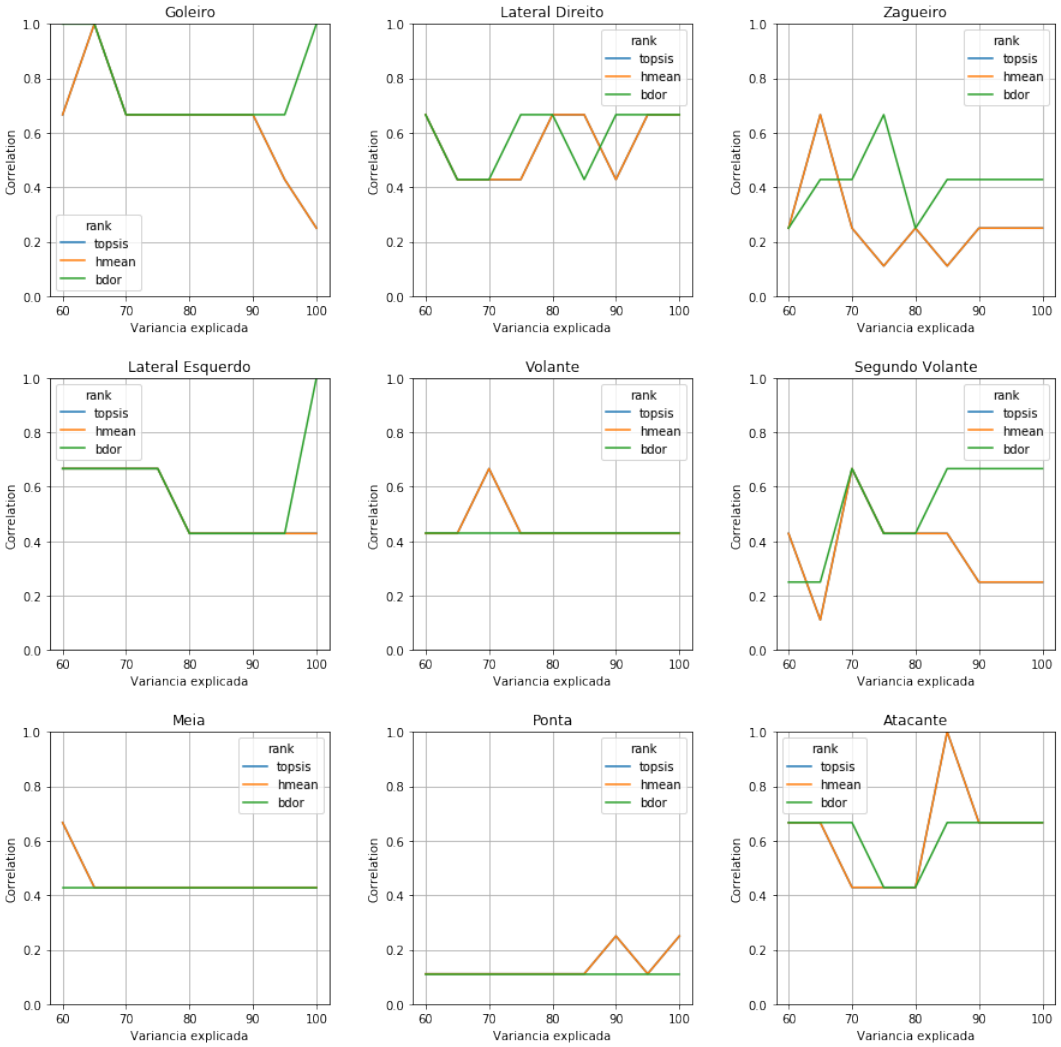


Figura 4.40: Comparação dos tipos de meta-agregação a partir da distância de Jaccard entre o top-5 do ranking meta-gerado e o da plataforma InStat

Capítulo 5

Conclusões

Algumas conclusões podem ser tiradas advindas desse trabalho. O processo de seleção das variáveis relevantes para definir os atletas é de alta complexidade e exige cuidado, pois tem grande impacto nos resultados finais. Tomando como referência o ranking da plataforma InStat, pôde-se perceber que jogadores de defesa se mostraram melhor definidos usando um número maior de variáveis que jogadores de ataque.

A prior utilizada para as habilidades em cada posição se mostrou irrelevante, impactando pouco nos resultados finais. Utilizar-se do método de suavização aditiva ou do método de média bayesiana se mostrou indiferente, independente do uso prévio ou não do PCA.

Dentre os métodos de ponderação testados, o método heurístico proposto pelo autor mostrou melhores resultados do que o método clássico por entropia e do que sem ponderação. Maiores estudos em cima dessa ideia para outros esportes/áreas que possuam eventos frequentes de natureza bernoulli (certo/errado) devem ser estimulados para uma melhor validação do método.

Filtrar habilidades pouco realizadas pelos jogadores da comparação impactou pouco nos resultados finais, apesar da diferença notável para o caso dos goleiros. Nesse caso, a inclusão de algumas ações raras como cobranças de falta e finalizações, influencia sobremaneira no ranking final.

Deixar de considerar jogadores pouco participativos não se mostrou uma boa ideia, mostrando que os métodos de decisão multi-critério testados possuem grande influência na presença ou não de alternativas irrelevantes.

A necessidade de uma bonificação para jogadores mais participativos ficou bem clara, tendo grande impacto no resultado final. Mesmo com o método de suavização reduzindo o efeito negativo relativo aos jogadores com poucas tentativas e alta porcentagem nas habilidades terem altos scores, uma magnificação extra se mostrou necessária para atingir melhores resultados.

O tipo de normalização prévia utilizada também mostrou forte impacto nos resultados finais, principalmente quando o PCA é aplicado previamente. A necessidade de se normalizar por desvio padrão para esse caso se mostrou de grande relevância para um melhor desempenho. Na aplicação direta às variáveis originais, o método Min-Max mostrou melhor desempenho.

No aspecto de comparar os métodos de agregação de scores, a contagem de Borda mostrou melhor performance em relação às outras quando se utilizado o conjunto original de dados. Com a aplicação prévia do PCA, o TOPSIS mostrou melhor performance, muito explicado pela redução de dimensionalidade aplicada, influenciando menos nas métricas de distância.

A meta-agregação melhorou gradativamente os resultados finais por posição, tendo melhor desempenho principalmente quando se aplicada com a Contagem de Borda.

Quanto a decisão de aplicar ou não o PCA previamente para reduzir a dimensionalidade do conjunto de dados, os resultados mostraram que, sem a aplicação prévia, precisa-se ter uma escolha mais precisa do conjunto de variáveis a ser levada em conta, pois os resultados variam bem mais.

Com a escolha precisa do número de variáveis, o método sem a aplicação do PCA se mostrou mais eficiente, tendo melhores resultados de correlação e de distância jaccardiana. Com a aplicação do PCA, o tempo de processamento dos resultados se

mostrou menor, porém um pouco menos preciso, apesar de uma maior estabilidade de valores conforme o quanto de variância se pretende reter dos dados.

Apesar da abrangência, o estudo realizado possui limitações. A não aplicação de um processo de validação cruzada para testar os valores ótimos dos parâmetros em conjuntos de dados diferentes é uma delas, sendo assim, objeto de estudo para novas investigações.

Um outro ponto limitante do estudo é a carência de dados abertos de atletas de futebol no país. A falta de rankings de referência para poder se comparar resultados atrapalha a definição de estudos nessa área.

Outro aspecto que pode ser considerado em futuras aplicações é a diferenciação das partidas por dificuldade das mesmas. Como forma de simplificar a pesquisa, foi decidido que não se diferenciaria os jogos e uma possível diferenciação pode vir a melhorar os resultados finais. Avaliar a relação do atleta com a sua equipe (poder de decisão, impacto no placar da partida...) pode ser um outro caminho para também aprimorar o estudo.

Para que isso possa ser aplicado para contratações (como no Moneyball, bastante citado na introdução), outras variáveis devem ser levadas em conta, como idade, preço de mercado estimado, número de jogos realizados, potencial de evolução, lesões, dentre outros fatores extra-técnicos.

Além de contratações, o uso dessas técnicas para comparar atletas em outras situações pode ser aplicada, como em rankings de campeonatos, dentre outras situações.

Com todos esses pontos citados, pode-se afirmar que o estudo conseguiu concluir seu objetivo de testar e definir estratégias para ranquear jogadores de futebol a partir de dados técnicos (eventos nas partidas). Que seja um ponto de partida para novas ideias e aplicações na área de ciência de dados aplicada ao esporte, tão pouco explorada no futebol brasileiro.

Referências Bibliográficas

[1] SPORTS, N., "Fan Favorite: The Global Popularity of Football is Rising", https://www.nielsen.com/pk/en/insights/news/2018/fan-favorite-the-global-popularity-of-football-is-rising.html, 2018, (Acesso em 26 Fevereiro 2018).

[2] MURTHY, P. R., *Operations Research (Second Edition)*. Anantapur, New Age, 2007.

[3] PITTS, B. G., STOTLAR, D. K., *Fundamentos do marketing esportivo*. São Paulo, Phorte, 2002.

[4] DELOITTE, "Rising stars Football Money League", https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance.html, 2018.

[5] FIFA, "Global Transfer Market Report 2018 - A review of all international football transfers in 2017", https://www.fifatms.com/wp-content/uploads/dlm_uploads/2018/01/GTM_2018.pdf, 2018, (Acesso em 18 Novembro 2018).

[6] MANSUR, C. E., "Brasil é o recordista em exportação e importação de jogadores", https://oglobo.globo.com/esportes/brasil-o-recordista-em-exportacao-importacao-de-jogadores-20866699, 2017, (Acesso em 18 Novembro 2018).

[7] KELLY, S., "The role of the professional football manager", https://books.google.co.uk/books?id=LCAlDwAAQBAJ, 2017, (Acesso em 18 Novembro 2018).

[8] MANZENREITER, W., HORNE, J., "Football goes east: Business, culture and the people's game in China, Japan and South Korea", https://doi.org/10.4324/9780203619216, 2004, (Acesso em 18 Novembro 2018).

[9] CURLEY, J. P., ROEDER, O., *English soccer's mysterious worldwide popularity*, New York, Contexts, pp. 78–81, 2016.

[10] ROHDE, M., BREUER, C., "Europe's Elite Football: Financial Growth, Sporting Success, Transfer Investment, and Private Majority Investors", *International Journal of Financial Studies*, v. 4(2), pp. 12, 2016.

[11] ACKERMANN, F., EDEN, C., *Making Strategy: The Journey od Strategic Management*. London, SAGE Publications, 2004.

[12] CARMICHAEL, F., THOMAS, D., WARD, R., "Team performance: the case of English premiership football", *Managerial and decision Economics*, v. 21, pp. 31–45, 2000.

[13] ALAMAR, B. C., *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers*. New York, Columbia University Press, 2013.

[14] BRADLEY, P. S., CARLING, C., ARCHER, D., *et al.*, "The effect of playing formation on high-intensity running and technical profiles in English FA Premier League soccer matches", *Journal of Sports Sciences*, v. 29(8), pp. 821–830, 2011.

[15] DRUST, B., REILLY, T., ATIKINSON, G., "Future perspectives in the evaluation of the physiological demands of soccer", *Sports Medicine*, v. 37(9), pp. 783–805, 2007.

[16] ARRIAZA, E. J., ZUNIGA, M. D., "Soccer as a Study Case for Analytic Trends in Collective Sports Training: A Survey", *International Journal of Performance Analysis in Sport*, v. 16, n. 1, pp. 171–190, April 2016. ISI.

[17] HUGHES, M., BARTLETT, R., "The use of performance indicators in performance analysis", *Journal of Sports Sciences*, v. 20, n. 10, pp. 739–754, 2002.

[18] KATZENBACH, J. R., SMITH, D. K., *The Wisdom of Teams: Creating the High-Performance Organization*. Boston, Harvard Business Review Press, 1993.

[19] REZENDE, B. D., *Transformando Suor em Ouro*. Rio de Janeiro, Sextante, 2006.

[20] FILETTI, C., RUSCELLO, B., D'OTTAVIO, S., *et al.*, "A Study of Relationships among Technical, Tactical, Physical Parameters and Final Outcomes in Elite Soccer Matches as Analyzed by a Semiautomatic Video Tracking System", *Perceptual and Motor Skills*, v. 124, n. 3, pp. 601–620, 2017.

[21] ROWE, D., *Global Media Sport: Flows, Forms and Futures*. London, Bloomsbury, 2011.

[22] HUTCHINS, B., "Tales of the digital sublime: Tracing the relationship between big data and professional sport", *Convergence*, v. 22, n. 5, pp. 494–509, 2016.

[23] LEWIS, M., *Moneyball: The Art of Winning an Unfair Game*. Norton & Company, 2004.

[24] VILAIN, J., KOLKOVSKY, R. L., "Estimating individual productivity in football", *Sciences Po Internationl University*, pp. 1–24, 2016.

[25] DEUTSCHER, C., BUSCHEMANN, A., "Does Performance Consistency Pay Off Financially for Players? Evidence From the Bundesliga", *Journal of Sports Economics*, v. 17, n. 1, pp. 27–43, 2016.

[26] WEIMAR, D., WICKER, P., "Moneyball Revisited: Effort and Team Performance in Professional Soccer", *Journal of Sports Economics*, v. 18, n. 2, pp. 140–161, 2017.

[27] PRINCIPE, V. A., *DADOS/FC: A Gestão da Informação aplicada ao Futebol*. Natal, Editora Primeiro Lugar, 2018.

[28] POMEROL, J. C., BARBA-ROMERO, S., *Multicriterion Decision in Management: Principles and Practice*. New York, Springer, 2012.

[29] MACKENZIE, R., CUSHION, C. J., "Performance analysis in football: a critical review and implications for future research", *Journal of Sports Sciences*, v. 31, n. 6, pp. 639–676, 2013.

[30] ANDERSON, C., SALLY, D., *Os Números do Jogo: Por que tudo o que você sabe sobre futebol está errado*. São Paulo, Editora Paralela, 2013.

[31] CAVALCANTI, E. A., CAPRARO, A. M., "Transferências internacionais no futebol: um estudo de caso comparativo entre os maiores clubes europeus e brasileiros", *Revista Brasileira de Futsal e Futebol*, v. 7, n. 23, pp. 3–15, 2015.

[32] WENG, R. C., LIN, C.-J., "A Bayesian Approximation Method for Online Ranking", *Journal of Machine Learning Research*, v. 12, pp. 267–300, 2011.

[33] HOOPER, D., WHYLD, K., *The Oxford Companion to Chess*. Oxford, Oxford University Press, 1992.

[34] SAATY, T. L., "Rank from comparisons and from ratings in the analytic hierarchy/network processes", *European Journal of Operational Research*, v. 168, n. 2, pp. 557–570, 2006.

[35] MAXCY, J., DRAYER, J., *Sports Analytics: Advancing Decision Making Through Technology and Data.*, Report TR-97/ONR-EPIC-08, Fox School of Business, 2014.

[36] TENGA, A., KANSTAD, D., RONGLAN, L. T., *et al.*, "Developing a New Method for Team Match Performance Analysis in Professional Soccer and Testing its Reliability.", *International Journal of Performance Analysis of Sport*, v. 9, pp. 8–25, 2009.

[37] DUCH, J., WAITZMAN, J. S., AMARAL, L. A. N., "Quantifying the performance of individual players in a team activity.", *PLoS ONE*, v. 5, n. 6, pp. 1–7, 2010.

[38] O'DONOGHUE, P., *Research Methods for Sports*. London, Routledge, 2010.

[39] RAMPININI, E., IMPELLIZZERI, F., CASTAGNA, C., *et al.*, "Technical performance during soccer matches of the Italian Serie A league: Effect of fatigue

and competitive level.", *Journal of Science and Medicine in Sport*, v. 12, n. 1, pp. 227–233, 2009.

[40] RUSSEL, M., REES, G., KINGSLEY, M. I., "Technical demands of soccer match play in the English championship.", *The Journal of Strength & Conditioning Research*, v. 27, n. 10, pp. 2869–2873, 2013.

[41] CARLING, C., REILLY, T., WILLIAMS, M. A., "Performance assessment. Preparing for Inclusive Teaching: Meeting The Challenges of Teacher Education Reform.", http://www.scopus.com/inward/record.url?eid=2-s2.0-84895637177&partnerID=tZOtx3y1, 2009.

[42] PRZEDNOWEK, K., ISKRA, J., WIKTOROWICZ, K., *et al.*, "Planning Training Loads for the 400 M Hurdles in Three-Month Mesocycles Using Artificial Neural Networks.", *Journal of Human Kinetics*, v. 60, n. 1, pp. 175–189, 2017.

[43] SILVA, E. J. D. O., *Análise do jogo de futebol: características do processo de transição defesaataque das sequências ofensivas com finalização*. UTAD, 2007.

[44] ROBERTSON, S., BACK, N., BARTLETT, J. D., "Explaining match outcome in elite Australian Rules football using team performance indicators.", *Journal of Sports Sciences*, v. 34, n. 7, pp. 637–644, 2016.

[45] WEINECK, J., *Coaching soccer - Conditioning*. Thessaloniki, Salto, 1997.

[46] KUMAR, G., *Machine Learning for Soccer Analytics*. M.Sc. dissertation, University College Dublin, Setembro 2013.

[47] KASAP, S., KASAP, N., "Development of a database and decision support system for performance evaluation of soccer players", https://www.researchgate.net/publication/228558144_Development_of_a_database_and_decision_support_system_for_performance_evaluation_of_soccer_players, 1997.

[48] STANOJEVIC, R., GYARMATI, L., "Towards data-driven football player assessment", *2016 IEEE 16th International Conference on Data Mining Workshops*, , 2016.

[49] HE, M., CACHUCHO, R., KNOBBE, A., "Football player's performance and market value", https://www.researchgate.net/publication/321623604_Football_player%27s_performance_and_market_value, 2017.

[50] PRINCIPE, V., GAVIAO, L. O., HENRIQUES, R., *et al.*, "Multicriteria analysis of football match performances: Composition of Probabilistc Preferences applied to the English Premier League 2015/2016", *Pesquisa Operacional*, v. 37, n. 2, 2017.

[51] SANT'ANNA, A. P., BARBOZA, E. U., MELLO, J. C. C. B. S. D., "Classification of the teams in the Brazilian Soccer Championship by probabilistic criteria composition", *Soccer & Society*, v. 11, n. 3, pp. 261–276, 2010.

[52] MAVI, R. K., MAVI, N. K., KIANI, L., "Ranking football teams with AHP and TOPSIS methods", *International Journal of Decision Sciences, Risk and Management*, v. 4, n. 1-2, 2012.

[53] PAPPALARDO, L., CINTIA, P., "Quantifying the relation between performance and success in soccer", https://www.researchgate.net/publication/316642937_Quantifying_the_relation_between_performance_and_success_in_soccer, 2017, Acesso em 5 dezembro 2018.

[54] PAPPALARDO, L., CINTIA, P., FERRAGINA, P., *et al.*, "PlayeRank: Multi-dimensional and role-aware rating of soccer player performance", https://arxiv.org/pdf/1802.04987, 2018, Acesso em 5 dezembro 2018.

[55] SA, L. H. P. D., *Sistema de apoio à decisão para avaliação técnica de jogadores de futebol: implementação de ferramenta de ETL e modelagem conceitual baseada em lógica Fuzzy.* M.Sc. dissertation, Universidade Federal do Rio de Janeiro, Março 2016.

[56] ARAUJO, A. S. D., *Sistema de apoio à decisão para futebol baseado em lógica Fuzzy.* M.Sc. dissertation, Universidade Federal do Rio de Janeiro, Setembro 2017.

[57] BROOKS, J., KERR, M., GUTTAG, J., "Developing a Data-Driven Player Ranking in Soccer Using Predictive Model Weights", *ACM SIGKDD International Conference*, v. 22, 2016.

[58] QADER, M., ZAIDAN, B., ZAIDAN, A., *et al.*, "A methodology for football players selection problem based on multi-measurements criteria analysis", *Journal of the International Measurement Confederation*, v. 111, pp. 38–50, 2017.

[59] DADELO, S., TURSKIS, Z., ZAVADSKAS, E. K., *et al.*, "Multi-criteria assessment and ranking system of sport team formation based on objective-measured values of criteria set", *Expert Systems with Applications*, v. 41, n. 14, pp. 6106–6113, 2014.

[60] NIKJO, B., REZAEIAN, J., JAVADIAN, N., "Decision Making in Best Player Selection: An Integrated Approach with AHP and Extended TOPSIS Methods Based on Wefa Freamwork in MAGDM Problems", *International Journal of Research in Industrial Engineering*, v. 4, n. 1-4, pp. 1–14, 2015.

[61] CHEN, C. C., LEE, Y.-T., TSAI, C.-M., "Professional Baseball Team Starting Pitcher Selection Using AHP and TOPSIS Methods.", *International Journal of Performance Analysis in Sport*, v. 14, n. 2, pp. 545–563, 2014.

[62] BALLI, S., KORUKOGLU, S., "Development of a fuzzy decision support framework for complex multi-attribute decision problems: A case study for the selection of skilful basketball players.", *Expert Systems*, v. 31, n. 1, pp. 56–69, 2012.

[63] AGILONU, A., BALLI, S., "Developing computer aided model for selecting players in badminton", *International Journal of Human Sciences*, v. 6, n. 2, 2009.

[64] BHARATHAN, S., RP, R. P. S., ABHIJEET, B., *et al.*, "Adapting Intelligent Optimized Analytical Model for team selection using player performance utility in Cricket", http://www.sloansportsconference.com/wp-content/uploads/2015/06/A-Self-Adapting-Intelligent-Optimized-Analytical-Model.pdf, 2015.

[65] FRY, M. J., LUNDBERG, A. W., OHLMANN, J. W., "A Player Selection Heuristic for a Sports League Draft", *Journal of Quantitative Analysis in Sports*, v. 3, n. 2, 2007.

[66] COOPER, W. W., RUIZ, J. L., SIRVENT, I., "Selecting non-zero weights to evaluate effectiveness of basketball players with DEA.", *European Journal of Operational Research*, v. 195, n. 2, pp. 563–574, 2009.

[67] SOLTANI, A., HEWAGE, K., REZA, B., *et al.*, "Multiple stakeholders in multicriteria decision-making in the context of Municipal Solid Waste Management: A review", *International Journal of Waste Management*, v. 35, pp. 318–328, 2005.

[68] TONG, L.-I., WANG, C.-H., CHEN, H.-C., "Optimization of multiple responses using principal component analysis and technique for order preference by similarity to ideal solution", *International Journal of Advanced Manufactured Technologies*, v. 3, n. 2, pp. 407–414, 2005.

[69] DONDAPATI, S., SATYAKIRAN, K., GEETHA, J., *et al.*, *Multi-Response Optimization using Grey Relational Analysis, TOPSIS and PCA-TOPSIS*. M.Sc. dissertation, Department of Mechanical Engineering R.V.R. & J.C. College of Engineering, April 2016.

[70] XIANGXIN, L., KONGSEN, W., LIWEN, L., *et al.*, "Application of the Entropy Weight and TOPSIS Method in Safety Evaluation of Coal Mines", *First International Symposium on Mine Safety Science and Engineering*, v. 26, pp. 2085 – 2091, 2011.

[71] LIANG, X., LIU, C., LI, Z., "Measurement of Scenic Spots Sustainable Capacity Based on PCA-Entropy TOPSIS: A Case Study from 30 Provinces of China", *International Journal of Environmental Research and Public Health*, v. 12, 2017.

[72] DUAN, W.-T., ZHANG, Y.-B., NIE, H., "Journals Evaluation and the Application Based on Entropy-TOPSIS", *Engineering Management Research*, v. 4, n. 1, 2015.

[73] BROWN, L. D., *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, 1986.

[74] CASELLA, G., BERGER, R. L., *Statistical Inference*. North Scituate, MA, Duxbury Press, 2001.

[75] DIACONIS, P., YLVISAKER, D., "Conjugate priors for exponential families", *Annals of Statistics*, v. 7, pp. 269–281, 1979.

[76] YANG, X., ZHANG, Z., "Combining prestige and relevance ranking for personalized recommendation", *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management*, v. 13, 2013.

[77] ZELENY, M., *Multiple Criteria Decision Making*. New York, Mc Graw Hill, 1982.

[78] SHANNON, C. E., WEAVER, W. W., *The Mathematical Theory of Communication*. Urbana, 1949.

[79] MILLIGAN, G. W., COOPER, M. C., "A study of standardization of variables in cluster analysis", *Journal of Classification*, v. 5, n. 2, pp. 181–204, 1988.

[80] SHLENS, J., "A Tutorial on Principal Component Analysis", *International Journal of Remote Sensing*, v. 51, n. 2, 2014.

[81] HWANG, C., YOON, K., *Multiple Attribute Decision Making*. Berlin, Springer-Verlag, 1981.

[82] SHIH, H.-S., SHYUR, H.-J., LEE, E. S., "An extension of TOPSIS for group decision making", *Mathematical and Computer Modelling*, v. 45, pp. 801–813, 2006.

[83] BELENSON, S., KAPUR, K., "An algorithm for solving multicriterion linear programming problems with examples", *Operational Research Quarterly*, v. 24, n. 1, pp. 65–77, 1973.

[84] ZELENY, M., "A concept of compromise solutions and the method of the displaced ideal", *Computers and Operations Research*, v. 1, pp. 479–496, 1973.

[85] CHAUHAN, A., VAISH, R., "Fluid Selection of Organic Rankine Cycle Using Decision Making Approach", *Journal of Computational Engineering*, v. 2013, 2013.

[86] KIM, G., PARK, C., YOON, K., "Identifying investment opportunities for advanced manufacturing systems with comparative-integrated performance measurement", *International Journal of Production Economics*, v. 50, pp. 23–33, 1997.

[87] CHENG, S., CHAN, C., HUANG, G., "Using multiple criteria decision analysis for supporting decision of solid waste management", *Journal of Environmental Science and Healths*, v. 37, n. 6, pp. 975–990, 2002.

[88] ZANAKIS, S., SOLOMON, A., WISHART, N., *et al.*, "Multi-attribute decision making: A simulation comparison of selection methods", *European Journal of Operational Research*, v. 107, pp. 507–529, 1998.

[89] AGGARWAL, C. C., HINNEBURG, A., KEIM, D. A., "On the Surprising Behavior of Distance Metrics in High Dimensional Space", *International Conference on Database Theory*, v. 1973, pp. 420–434, 2001.

[90] FRAENKEL, J., GROFMAN, B., "The Borda Count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia.", *Australian Journal of Political Science*, v. 49, n. 2, pp. 186–205, 2014.

[91] SPEARMAN, C., "The Proof and Measurement of Association between Two Things", *The American Journal of Psychology*, v. 15, n. 1, pp. 72–101, 1904.

[92] JACCARD, P., "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bulletin de la Société Vaudoise des Sciences Naturelles*, v. 37, n. 1, pp. 547–579, 1901.

[93] LEONETI, A. B., "Considerations regarding the choice od ranking multiple criteria decision making methods", *Jornal da Sociedade Brasileira de Pesquisa Operacional*, v. 36, n. 2, pp. 259–277, 2016.

Apêndice A

Amostras das tabelas extraídas

id	match_date	match_name	round_name	season_name	team1_name	team1_score	team2_name	team2_score	tournament_name	home_pos	away_pos
1143146	2018-11-18 23:00:00	Botafogo RJ - Internacional	Week 35	2018	Botafogo RJ	0	Internacional	0	Brazil. Serie A	11.0	2.0
1143140	2018-11-16 03:00:00	Internacional - America Mineiro	Week 34	2018	Internacional	2	America Mineiro	0	Brazil. Serie A	2.0	19.0
1143127	2018-11-11 22:00:00	Ceara - Internacional	Week 33	2018	Ceara	1	Internacional	1	Brazil. Serie A	15.0	2.0
1143120	2018-11-05 00:00:00	Internacional - Atletico Paranaense	Week 32	2018	Internacional	2	Atletico Paranaense	1	Brazil. Serie A	3.0	8.0
1143105	2018-10-27 03:30:00	Vasco da Gama - Internacional	Week 31	2018	Vasco da Gama	1	Internacional	1	Brazil. Serie A	15.0	3.0

Tabela A.1: Amostra da tabela de jogos

player_id	match_id	Tackles successful	Tackles	Shots	Shots / on target	Short passes	Short passes - accurate
13001	**1143090.0**	1.0	2.0	3.0	2.0	4.0	3.0
13264	**1143090.0**	2.0	3.0	4.0	1.0	11.0	10.0
13404	**1143090.0**	0.0	0.0				
13777	**1143090.0**	3.0	5.0	1.0		6.0	5.0
22052	**1143090.0**	4.0	4.0			3.0	3.0
29163	**1143090.0**	0.0	0.0				
81462	**1143090.0**	3.0	4.0	2.0	1.0	12.0	9.0
82464	**1143090.0**	0.0	4.0	2.0	1.0	8.0	6.0
86780	**1143090.0**	5.0	5.0	2.0		11.0	9.0
91448	**1143090.0**	0.0	0.0	1.0		7.0	6.0

Tabela A.2: Amostra da tabela do rendimento técnico dos jogadores nas partidas

birthday	club_number	club_team_name	contract_ending	country1_name	firstname	foot_name	gender_name	height	**id**	lastname	position1_name	weight
1989-01-26	10.0	America Mineiro	2018-05-21	Brazil	Ruy Franco de Almeida	Left	Male	171.0	**36858**	Junior	Attacking midfielder - Left	74.0
1998-11-07	27.0	America Mineiro	2019-12-31	Brazil	Lincoln Henrique	Left	Male	178.0	**287846**	Oliveira dos Santos	Attacking midfielder - Left	65.0
1988-09-21	11.0	America Mineiro	2018-12-31	Brazil	Luan	Left	Male	186.0	**13279**	Michel Louza	Attacking midfielder - Left	73.0

Tabela A.3: Amostra da tabela de jogadores