Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Evolução de mapas de intensidade do FC Barcelona: uma análise via Séries Temporais Funcionais

Guilherme Rodrigues Boff

Porto Alegre
Maio de 2021

Guilherme Rodrigues Boff

Evolução de mapas de intensidade do FC Barcelona: uma análise via Séries Temporais Funcionais

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Eduardo de Oliveira Horta

Porto Alegre
Maio de 2021

Guilherme Rodrigues Boff

Evolução de mapas de intensidade do FC Barcelona: uma análise via Séries Temporais Funcionais

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pelo Orientador e pela Banca Examinadora.

Orientador:_____
Prof. Dr. Eduardo de Oliveira Horta, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Flávio Augusto Ziegelmann, UFRGS
Doutor pela University of Kent at Canterbury – UK

Porto Alegre
Maio de 2021

"Para realizar sua estratégia, você precisa acertar sua tática; e sua tática deve sempre se adequar ao seu time e ao seu rival".

Chris Anderson e David Sally

Agradecimentos

Sempre ouvi em vários lugares e em diferentes épocas que escrever um TCC era uma tarefa complicada e desafiadora. Hoje, após concluí-lo, posso afirmar que essa tarefa se torna muito mais complexa se realizada durante uma pandemia, vivendo 24 horas por dia no mesmo ambiente há mais de um ano. Por isso, meus primeiros agradecimentos vão à minha família, que teve toda a paciência comigo nesse período, e que sempre foi meu apoio nos momentos mais difíceis, nos quais eu acreditava que meus esforços não seriam suficientes. Acima de tudo à minha mãe, Teresinha, que sempre foi quem mais me entendeu e lutou por mim. Ao meu pai, Juares, por sempre se fazer presente do seu jeito e transmitir a paixão pelo futebol para mim. E à minha irmã, Camila, por sempre ser a minha inspiração para a vida. Só vocês sabem de toda a minha dedicação.

Na UFRGS, várias pessoas contribuíram para a minha trajetória no curso de Estatística. Primeiramente, agradeço ao Professor Eduardo por todas as ajudas, conselhos e ensinamentos durante esses mais de quatro anos de aulas, orientação de iniciação científica e, agora, do TCC; aprendi muito contigo. Ao Professor Flávio por aceitar o convite para fazer parte da banca avaliadora deste trabalho e pelas aulas de séries temporais que fizeram eu me interessar mais pela área. À Professora Márcia Barbian por todos apoios, incentivos (e muitos) e, na reta final, pela parceria e orientação em uma nova pesquisa. À Professora Patrícia por ser meu norte no início do curso e me despertar o verdadeiro interesse na Estatística durante as aulas de Introdução à Inferência. Aos Professores Álvaro e Gabriela e todos os demais que tive oportunidade de ser aluno.

Sem as amizades (verdadeiras), acredito que não conseguiria passar por esse período. Agradecimentos aos amigos que a vida me deu: Matheus, Renata; e aos que o curso me proporcionou: Filipe, Lucas, Juliana, Maicon, Carol, Franciele, Guilherme, Victor, Rosana, e tantas outras pessoas que inevitavelmente não conseguirei citar. Esse período foi mais leve com a ajuda de vocês.

Não poderia me esquecer de agradecer aos Professores das escolas Araguaia e Padre Reus, que, apesar de todas as dificuldades enfrentadas pelo ensino público, sempre empenharam-se em oferecer um ensino de qualidade para seus estudantes;

tenho imenso orgulho de ter feito parte deles.

A todos que acreditaram em mim.

Resumo

Dados funcionais são comuns em várias áreas de aplicação e, embora sejam frequentes no âmbito desportivo, geralmente não são analisados de acordo com sua natureza nessa área. Particularmente no futebol, um dos esportes mais atrativos e populares a nível mundial, percebe-se um aumento do volume de disponibilização de dados com alto grau de informações sobre partidas, devido ao advento de novas tecnologias de coleta observado nas últimas décadas. A partir desses dados, uma ferramenta comumente usada por analistas dos clubes e da imprensa são mapas de intensidade, os quais indicam como um jogador ou uma equipe se distribuiu em campo sob posse da bola, sendo este um tipo de dado funcional. Séries Temporais Funcionais abrangem um conjunto de técnicas para trabalhar com dados funcionais quando estes apresentam alguma estrutura temporal. Assim, neste trabalho propõe-se uma aplicação de Séries Temporais Funcionais para modelagem e previsão de mapas de intensidade de equipes de futebol. Os dados utilizados na análise referem-se a uma sequência de partidas do Fútbol Club Barcelona pelo Campeonato Espanhol. Os resultados demonstram uma boa qualidade preditiva de posturas de ofensividade/defensividade pelo modelo, o qual pode ser empregado por equipes de futebol em um contexto de planejamento tático para jogos futuros. Além disso, destaca-se que este é um dos primeiros trabalhos a analisar dados de futebol sob um ponto de vista de análise de Séries Temporais Funcionais.

Palavras-Chave: Séries Temporais Funcionais, Dados Funcionais, Mapas de intensidade, Futebol, Estatística nos Esportes.

Abstract

Functional data are common in several areas of application and, although they are frequent in sports, they are generally not analyzed according to their nature in this area. Particularly in football, one of the most attractive and popular sports worldwide, there is an increase in the volume of data available with a high degree of information about matches, due to the advent of new data collection technologies observed in recent decades. From these data, a tool commonly used by clubs and press analysts are intensity maps (heatmaps), which indicate how a player or team was distributed on the field in possession of the ball, this being a type of functional data. Functional Time Series covers a set of techniques for working with functional data when they have some time structure. Thus, this work proposes an application of Functional Time Series for modeling and forecasting intensity maps of football (soccer) teams. The data used in the analysis refer to a sequence of matches by Fútbol Club Barcelona for the Spanish League. The results demonstrate good predictive quality of offensive/defensive postures by the model, which can be used by football teams in the context of tactical planning for future games. In addition, it is noteworthy that this is one of the first studies to analyze football data from the Functional Time Series analysis' point of view.

Keywords: Functional Time Series, Functional Data, Football, Sports Statistics.

Sumário

1 Introdução **13**

2 Análise espaço-temporal no futebol **16**
2.1 Dados espaço-temporais no futebol 16
2.2 Posse de bola e mapas de intensidade no futebol 17

3 Análise de Dados Funcionais e Séries Temporais Funcionais **19**
3.1 Análise de Dados Funcionais nos esportes 20
3.2 Notação e convenções 21
3.3 Análise de Séries Temporais Funcionais: Representações espectrais baseadas no operador de autocovariância defasado . . 22
3.4 Estimadores e resultados amostrais 26

4 Aplicação: Análise de mapas de intensidade do FC Barcelona **30**
4.1 Fonte de dados 30
4.2 Organização dos dados e geração dos mapas de intensidade . . 31
4.3 Modelagem 33
4.4 Previsão 38
4.5 Previsão de ofensividade 46

5 Conclusão **49**

Referências Bibliográficas **51**

APÊNDICE A **55**
A.1 Lista de ações consideradas 55
A.2 Covariáveis selecionadas para o modelo final 56

APÊNDICE B **59**
B.1 Modelo $VARX$ 59
B.2 Metodologia $VARX$-L 60

APÊNDICE C **61**
C.1 Ajuste global com 428 partidas 61

APÊNDICE D **64**
D.1 Previsões finais 64

Lista de Figuras

4.1 Mapa de intensidade do FC Barcelona alusivo à 2ª rodada da temporada 2008/2009 do Campeonato Espanhol diante do Real Racing Club de Santander. Todas as coordenadas do campo são padronizadas para que o FC Barcelona ataque da esquerda para a direita. 32
4.2 Função média $\widehat{\mu}$ após $n = 374$ partidas do FC Barcelona no Campeonato Espanhol entre as temporadas 2008/2009 e 2018/2019. . 34
4.3 Dez primeiros autovalores obtidos via decomposição espectral da matriz $\boldsymbol{K}^*$. 34
4.4 Autofunções $\widehat{\psi}_j$, $j = 1, \ldots, 4$. 35
4.5 Séries temporais $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$, formadas pelos coeficientes das projeções dos mapas de intensidade observados nas autofunções. . 36
4.6 Funções de autocorrelação das séries $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$. 37
4.7 Funções de autocorrelação parcial das séries $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$. . . . 37
4.8 Gráficos de dispersão entre $\widehat{\eta}_{tj}$ e $\widehat{\eta}_{t-1,j}$, $j = 1, \ldots, 4$. 38
4.9 Funções de correlação cruzada entre as séries temporais $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$. 39
4.10 Gráficos de dispersão entre $\widehat{\eta}_{tj}$ e $\widehat{\eta}_{t-1,i}$, $j \neq i$. 40
4.11 Erro quadrático integrado de cada previsão feita pelo modelo final. 43
4.12 Melhor previsão e variação observada. $T + 1 = 358$. 43
4.13 Segunda melhor previsão e variação observada. $T + 1 = 339$. . . . 44
4.14 Pior previsão e variação observada. $T + 1 = 357$. 45
4.15 Segunda pior previsão e variação observada. $T + 1 = 345$. 46
4.16 Melhor e pior curvas de intensidade preditas e respectivas observações. 47
4.17 Erros de previsão $\widehat{y}_{T+1|T} - \widehat{y}_{T+1}, T = 324, \ldots, 373$, e erro médio (curva rosa). 48

C.1 Função média $\widehat{\mu}$ após ajuste com 428 partidas. 61
C.2 Dez primeiros autovalores da matriz $\boldsymbol{K}^*$ após ajuste com 428 partidas. 62
C.3 Autofunções $\widehat{\psi}_j$, $j = 1, 2$, obtidas via ajuste com 428 partidas. . . 62
C.4 Séries temporais $\widehat{\eta}_{tj}$, $j = 1, 2$, resultantes do ajuste com 428 partidas. 63

D.1 Variação predita e observada, $T + 1 = 325$. 64
D.2 Variação predita e observada, $T + 1 = 326$. 64
D.3 Variação predita e observada, $T + 1 = 327$. 65

D.4 Variação predita e observada, $T+1=328$. 65
D.5 Variação predita e observada, $T+1=329$. 65
D.6 Variação predita e observada, $T+1=330$. 66
D.7 Variação predita e observada, $T+1=331$. 66
D.8 Variação predita e observada, $T+1=332$. 66
D.9 Variação predita e observada, $T+1=333$. 67
D.10 Variação predita e observada, $T+1=334$. 67
D.11 Variação predita e observada, $T+1=335$. 67
D.12 Variação predita e observada, $T+1=336$. 68
D.13 Variação predita e observada, $T+1=337$. 68
D.14 Variação predita e observada, $T+1=338$. 68
D.15 Variação predita e observada, $T+1=340$. 69
D.16 Variação predita e observada, $T+1=341$. 69
D.17 Variação predita e observada, $T+1=342$. 69
D.18 Variação predita e observada, $T+1=343$. 70
D.19 Variação predita e observada, $T+1=344$. 70
D.20 Variação predita e observada, $T+1=346$. 70
D.21 Variação predita e observada, $T+1=347$. 71
D.22 Variação predita e observada, $T+1=348$. 71
D.23 Variação predita e observada, $T+1=349$. 71
D.24 Variação predita e observada, $T+1=350$. 72
D.25 Variação predita e observada, $T+1=351$. 72
D.26 Variação predita e observada, $T+1=352$. 72
D.27 Variação predita e observada, $T+1=353$. 73
D.28 Variação predita e observada, $T+1=354$. 73
D.29 Variação predita e observada, $T+1=355$. 73
D.30 Variação predita e observada, $T+1=356$. 74
D.31 Variação predita e observada, $T+1=359$. 74
D.32 Variação predita e observada, $T+1=360$. 74
D.33 Variação predita e observada, $T+1=361$. 75
D.34 Variação predita e observada, $T+1=362$. 75
D.35 Variação predita e observada, $T+1=363$. 75
D.36 Variação predita e observada, $T+1=364$. 76
D.37 Variação predita e observada, $T+1=365$. 76
D.38 Variação predita e observada, $T+1=366$. 76
D.39 Variação predita e observada, $T+1=367$. 77
D.40 Variação predita e observada, $T+1=368$. 77
D.41 Variação predita e observada, $T+1=369$. 77
D.42 Variação predita e observada, $T+1=370$. 78
D.43 Variação predita e observada, $T+1=371$. 78
D.44 Variação predita e observada, $T+1=372$. 78
D.45 Variação predita e observada, $T+1=373$. 79
D.46 Variação predita e observada, $T+1=374$. 79

Lista de Tabelas

4.1 Estrutura do banco de dados final. 31
4.2 Resultado do teste de raiz unitária aumentado de Dickey-Fuller. . 36
4.3 EQMI de previsão um passo à frente dadas combinações de d_0 e q. 42

A.1 Ações filtradas na análise e suas respectivas descrições. 55
A.2 Covariáveis exógenas selecionadas para a equação de $\widehat{\eta}_{t1}$ 57
A.3 Covariáveis exógenas selecionadas para a equação de $\widehat{\eta}_{t2}$ 58

1 Introdução

O futebol é um dos esportes mais populares do mundo, contando com um grande número de espectadores e praticantes. Segundo a Federação Internacional das Associações de Futebol (FIFA, 2007), estima-se que 265 milhões de pessoas praticam o esporte. Além disso, de acordo com a mesma organização, 3,5 bilhões de pessoas assistiram à Copa do Mundo de 2018 (FIFA, 2018). Dessa forma, o futebol profissional atrai grande atenção da imprensa e de diversos investidores e patrocinadores mundialmente, constituindo-se em um mercado que envolve uma alta quantidade de movimentações financeiras, desde negociações referentes a compra e venda de jogadores (RIVEIRA, 2020), até casos de empresários que adquirem clubes por valores bilionários (GRAFIETTI, 2020). Nesse sentido, do ponto de vista de um mercado, não é razoável que as decisões efetuadas no futebol sejam definidas apenas pelas tradições e opiniões pessoais de dirigentes e técnicos. No entanto, diferentemente de outros esportes, como o beisebol, o basquete e o futebol americano, o futebol tem realizado um movimento mais lento e tardio no que diz respeito ao uso de ferramentas analíticas (ANDERSON; SALLY, 2013; SCHULTZE; WELLBROCK, 2018). Desse modo, é um esporte que ainda não dispõe de uma área de análise estatística amplamente desenvolvida.

Não obstante, os clubes profissionais estão gradativamente incorporando a percepção de que suas performances nos campeonatos podem ser melhoradas através de tomadas de decisão baseadas em dados, de forma que montar uma equipe mais competitiva passa pelo investimento de recursos econômicos em departamentos de análise de dados e desempenho. Assim, nas últimas décadas, percebeu-se um considerável avanço de aplicações de modelagem estatística ao futebol. Concomitantemente a essa inicial evolução, de acordo com Lucey et al. (2013), é cada vez maior o volume de dados coletados em uma partida de futebol, tanto pela necessidade enxergada pelos clubes quanto pelo crescimento tecnológico observado nas últimas décadas. Tais dados podem servir como uma ferramenta para a tomada de decisão em questões de planejamento tático e técnico da equipe, além de serem muito úteis em pontos envolvendo a preparação física dos jogadores. Contudo, segundo os mesmos autores,

embora haja essa evolução, ainda torna-se necessário o desenvolvimento de técnicas mais apropriadas para lidar com a complexidade inerente a esses dados.

Nas mais diversas áreas da Ciência, como Biologia, Economia, Medicina, etc., dados funcionais estão surgindo de maneira mais frequente (RAMSAY; SILVERMAN, 2005). Dados funcionais são aqueles em que cada observação é uma função, a qual usualmente é uma curva ou superfície. Ao conjunto de técnicas desenvolvidas para lidar com esse tipo de dado se dá o nome de Análise de Dados Funcionais (daqui em diante, ADF). No futebol, um típico dado funcional ao qual as equipes têm acesso são os mapas de intensidade (*heatmaps*), que descrevem as regiões do campo nas quais os jogadores mais tocaram na bola, sendo capazes de indicar o comportamento coletivo do time com a posse da bola. No âmbito de ADF, Séries Temporais Funcionais é uma subárea que estuda dados funcionais indexados no tempo, e que está em crescente expansão nos últimos anos na Estatística. Contudo, ainda são poucos os trabalhos envolvendo dados esportivos cujo enfoque seja utilizar a ADF e, em particular, Séries Temporais Funcionais.

Assim, o presente trabalho pretende demonstrar o potencial de aplicação de técnicas de Séries Temporais Funcionais a dados advindos dos esportes, sendo seu objetivo fazer modelagem e previsão de mapas de intensidade de equipes de futebol sob um ponto de vista de Séries Temporais Funcionais, especificamente utilizando a metodologia de Bathia, Yao e Ziegelmann (2010). Como contribuição original, será considerada a inclusão de covariáveis exógenas ao modelo de séries temporais funcionais. Nesse sentido, uma abordagem de seleção de variáveis é empregada no contexto de modelos *VARX – Vector Autoregressions with Exogenous Variables*. Para tanto, se utilizará de um conjunto de dados referente a partidas do Fútbol Club (FC) Barcelona em um período de 11 temporadas do Campeonato Espanhol. Tal abordagem poderá servir de auxílio a comissões técnicas no planejamento tático da equipe, uma vez que possibilita fazer previsões da taxa de ocupação do campo com a bola dos adversários com base no padrão observado em jogos anteriores. Além disso, também servirá como mais uma colaboração em busca de uma maior consolidação da análise de dados no futebol.

Este estudo está estruturado da seguinte maneira: no Capítulo 2 é apresentada uma breve introdução sobre a análise de dados espaço-temporais no futebol, destacando-se alguns trabalhos cujo objeto está relacionado a esta pesquisa. No Capítulo 3, primeiramente é feita uma pequena contextualização geral sobre ADF e Séries Temporais Funcionais. Na Seção 3.1, são evidenciadas aplicações de ADF a dados esportivos, enquanto que nas Seções 3.3 e 3.4 descreve-se a metodologia de Bathia, Yao e Ziegelmann (2010). Nesta última propõe-se, ademais, uma discussão à luz do contexto de previsão de séries temporais funcionais e disponibiliza-se um algoritmo com o passo-a-passo do procedimento de modelagem e previsão, baseado

no trabalho de Aue, Norinho e Hörmann (2015). Já no Capítulo 4, realiza-se a análise de dados considerando o conjunto de mapas de intensidade de partidas do FC Barcelona no Campeonato Espanhol de futebol. Por fim, as conclusões gerais sobre os resultados e acerca de possíveis investigações futuras estão no Capítulo 5.

2 Análise espaço-temporal no futebol

2.1 Dados espaço-temporais no futebol

Nas últimas décadas, constatou-se o surgimento de várias empresas especializadas na coleta e armazenamento de dados de partidas de futebol, como *OptaSports*,[1] *Prozone*, *Amisco* (todas adquiridas pela *STATS*[2]) e, mais recentemente, a *StatsBomb*.[3] Devido ao investimento financeiro e surgimento de tecnologias mais sofisticadas, essas empresas têm disponibilizado informações cada vez mais detalhadas, como dados de rastreamento de bola e jogadores (*tracking data*), e dados que descrevem cada ação envolvendo a bola em uma partida (*data events* ou *log events*), aos quais a literatura no âmbito das Ciências Esportivas tem denominado de dados espaço-temporais (GUDMUNDSSON; HORTON, 2017). De acordo com Gudmundsson e Horton (2017) e Lucey et al. (2013), tais dados têm sido utilizados de maneira frequente para a visualização de ações ocorridas em partidas, principalmente de futebol e basquete.

Diferentemente dos dados presentes de forma majoritária nas análises estatísticas no futebol, os quais apenas informam estatísticas descritivas das equipes em uma partida (por exemplo, número de chutes, número de passes, percentual de passes corretos, percentual de posse de bola, etc.), dados espaço-temporais têm o potencial de informar não apenas o que aconteceu na partida, mas também onde e como aconteceu (LUCEY et al., 2013). No entanto, são poucos os métodos desenvolvidos ou adaptados para trabalhar com esse tipo de informação (BIALKOWSKI et al., 2014a) e, geralmente, os clubes recebem esse grande volume de dados sem saber ao certo como tratá-los e analisá-los, de forma que recursos vêm sendo naturalmente investidos em departamentos de análise estatística. Todavia, embora haja uma evolução da análise estatística no futebol, esse crescimento é mais lento do que o observado em outros esportes (SCHULTZE; WELLBROCK, 2018).

[1] <https://www.optasports.com/>
[2] <https://www.statsperform.com/>
[3] <https://statsbomb.com/>

2.2 Posse de bola e mapas de intensidade no futebol

Segundo Gudmundsson e Horton (2017), o futebol é um esporte baseado em invasão, no qual duas equipes se enfrentam em um período de tempo pré-determinado e os jogadores adversários disputam a posse da bola com o objetivo de marcar gols. A posse de bola é, portanto, um aspecto importante no futebol, uma vez que uma equipe deve tê-la para atingir o objetivo principal do esporte: o gol. Mapas de intensidade (ou de calor) – ver Figura 4.1 – estão intimamente ligados à posse de bola pois, como relatado por Gudmundsson e Horton (2017), são uma ferramenta gráfica que descreve em quais partes do campo os jogadores de uma equipe mais tocaram na bola (ou seja, mais mantiveram a posse dela) durante uma partida. Desse modo, os mapas de intensidade têm a capacidade de informar onde as ações ocorreram no jogo, podendo revelar (ao menos em parte) o padrão comportamental do time com a bola, e são utilizados com tal finalidade em Lucey et al. (2013), Bialkowski et al. (2014a) e Bialkowski et al. (2014b).

Lucey et al. (2013) propõem uma representação do estilo de jogo de equipes de futebol por meio de uma repartição discreta do campo em vários retângulos, sendo avaliada a quantidade de vezes que o time manteve a bola em cada um deles. Alternativamente, os autores chamam o gráfico resultante de *mapa de ocupação*. Na análise dos dados, os mapas de ocupação de equipes visitantes são subtraídos dos mapas de ocupação de equipes mandantes com o objetivo de verificar se seus comportamentos mudam conforme o mando de campo. A hipótese levantada é de que a vantagem do mandante nesse esporte deve-se, entre outros fatores, à postura conservadora (defensiva) de times visitantes. Os resultados mostram que os mandantes ocupam mais o campo ofensivo do que visitantes. No contexto do presente trabalho, esse achado justifica a inclusão de uma covariável indicadora de mando de campo no modelo de séries temporais funcionais.

Além disso, os mapas de intensidade são usados de maneira a verificar questões semelhantes nos estudos de Bialkowski et al. (2014a) e Bialkowski et al. (2014b), nos quais os autores sugerem que as equipes de futebol visam ganhar seus jogos em casa e empatar quando visitantes. Dessa forma, em ambos os trabalhos é implementada uma técnica baseada no algoritmo EM (*Expectation-Maximization*), em que os mapas de intensidade da equipe são modelados como uma combinação linear dos mapas de intensidade individuais dos jogadores, para determinar de maneira automática a formação das equipes, com vistas a avaliar se os diferentes comportamentos conforme o mando de campo devem-se ou ao fato de as formações serem diferentes, ou à postura distinta dos jogadores no campo.

Assim, embora os mapas de intensidade não descrevam a qualidade das ações

realizadas na partida, os mapas da própria equipe podem mostrar, em uma análise pós-jogo, se os jogadores estão mantendo a posse da bola em regiões específicas do campo conforme planejado taticamente. Além disso, os mapas de calor de times oponentes podem ser usados como mecanismo para o planejamento estratégico para o próximo jogo (pré-jogo), por revelarem o comportamento global do adversário quando este detém a posse da bola.

3 Análise de Dados Funcionais e Séries Temporais Funcionais

A Análise de Dados Funcionais é um ramo recente da Estatística que vem se desenvolvendo amplamente nas últimas décadas. Devido ao avanço das capacidades de processamento e armazenamento computacional, o tratamento de unidades observacionais (que tradicionalmente se restringia a considerá-las escalares, na estatística univariada, ou vetores, na estatística multivariada) pôde ser ampliado para o caso em que estas são funções. Nessa perspectiva, uma observação funcional pode ser definida como uma função $(y(u)\colon u \in S)$. Em termos probabilísticos, a função $(y(u)\colon u \in S)$ pode ser interpretada como a realização de uma função aleatória $(Y(u)\colon u \in S)$, ou, em outras palavras, como a realização de um processo estocástico com conjunto de índices S, de forma que, para cada $u \in S$, $Y(u)$ é uma variável (ou vetor) aleatória(o). Se n realizações $(y_i(u)\colon u \in S)$ de n funções aleatórias $(Y_i(u)\colon u \in S)$ são observadas, então tem-se acesso a uma amostra de dados funcionais (FERRATY; VIEU, 2006). A terminologia *funcional* implica que o domínio S tem ao menos a cardinalidade do contínuo. Assim, S pode representar o tempo ou o espaço, por exemplo, de forma que, em ADF, as unidades observacionais podem ser curvas (no cenário de funções de uma variável), superfícies (funções de duas variáveis), etc.

Além de dados que naturalmente são coletados como funções, aqueles cuja estrutura não seja aparentemente funcional também podem ser tratados como tal, como, por exemplo, uma única longa realização de um processo estocástico a tempo contínuo $(Y(u)\colon u \in [0, T])$, que pode ser visto como a concatenação de curvas $(Y_k(\tau)\colon \tau \in [0, 1])$, $k = 0, \ldots, T-1$. Ainda, dados longitudinais também são um tipo de dado que comumente pode ser considerado como funcional mediante à suavização das observações repetidas. Ver, por exemplo, Rice (2004), Yao, Müller e Wang (2005) e Hall, Müller e Wang (2006) para discussões acerca de metodologias e aplicações de ADF para dados longitudinais. Nesse contexto, há de se lembrar que as observações funcionais são, por definição, funções que devem ser computá-

veis para todo o ponto do seu domínio. Contudo, no âmbito empírico, os dados de cada observação individual se apresentam como um conjunto finito de valores $\{\delta_{ij}\}$, $j = 1, \ldots, N_i$, em que δ_{ij} é um valor da função de interesse y_i no ponto do domínio u_{ij} e N_i é o número de medições avaliadas da i-ésima observação, $i = 1, \ldots, n$ (RAMSAY; SILVERMAN, 2005). Portanto, um primeiro passo na ADF envolve a aplicação de processos de suavização, tais como *splines*, estimação via *kernel*, *wavelets*, etc., aos dados brutos a fim de se obter funções suaves. Os capítulos de 3 a 6 de Ramsay e Silverman (2005) descrevem tais métodos.

Inicialmente, a teoria de ADF foi desenvolvida sobretudo no cenário em que as unidades observacionais são independentes e identicamente distribuídas (i.i.d.). Metodologias nessa situação envolvem, entre outros tópicos, generalizações de técnicas estatísticas clássicas, como modelos lineares (ver capítulos de 12 a 17 de Ramsay e Silverman, 2005), métodos não-paramétricos (FERRATY; VIEU, 2002; FERRATY; VIEU, 2006) e análise de componentes principais (capítulo 8 de Ramsay e Silverman, 2005). Em particular, em ADF, a redução de dimensionalidade é de grande interesse, pois dados funcionais são, por definição, de alta dimensão (dimensão infinita) e, ao se trabalhar com tal tipo de dado, espaços infinito-dimensionais surgem de maneira natural. Na Seção 3.3, há uma breve discussão sobre o método de Análise de Componentes Principais Funcionais na perspectiva de Séries Temporais Funcionais.

Séries Temporais Funcionais, como o nome sugere, são dados funcionais cuja estrutura de coleta possui alguma ordenação temporal. Segundo Bathia, Yao e Ziegelmann (2010), a vantagem de tratar uma típica série temporal pelo ponto de vista funcional está na viabilidade de acomodar de forma satisfatória possíveis características de não-estacionariedade dos dados. Bosq (2000) é uma importante referência para a teoria de Séries Temporais Funcionais, onde estuda-se o modelo autorregressivo funcional (FAR – *Functional Autoregressions*), uma extensão de modelos autorregressivos para espaços funcionais, que são aplicáveis a séries temporais funcionais lineares. A teoria de Séries Temporais Funcionais está em crescente expansão e desenvolvimentos recentes nessa área são diversos. Alguns deles podem ser encontrados em Bathia, Yao e Ziegelmann (2010) e Aue, Norinho e Hörmann (2015).

3.1 Análise de Dados Funcionais nos esportes

Dados funcionais são recorrentes em diversas áreas do conhecimento. Assim, aplicações de ADF e Séries Temporais Funcionais são encontradas, por exemplo, em estudos longitudinais (YAO; MÜLLER; WANG, 2005), estudos de variação climática (ANTONIADIS; SAPATINAS, 2003; BESSE; CARDOT; STEPHENSON, 2000), estudos de crescimento humano (RAMSAY; SILVERMAN, 2005), entre vários outros

campos de aplicação (ver Ramsay e Silverman, 2007, para exemplos aplicados à Economia, Criminologia, Medicina, etc.). Contudo, nos esportes as aplicações de ADF são limitadas, restringindo-se, em especial, a problemas da biomecânica dos esportes, como em Harrison, Ryan e Hayes (2007), Dona et al. (2009) e Harrison (2014).

Em contrapartida, Wakim e Jin (2014) apresentam uma aplicação de ADF a curvas de envelhecimento de jogadores de basquete e beisebol, com vistas a avaliar como a performance dos atletas varia com o passar dos anos. Por outro lado, Vinué e Epifanio (2017) usam ADF para avaliar o desempenho de jogadores de basquete e de times de futebol, destacando sua potencialidade de aplicação a dados dos esportes, e relatando a baixa quantidade de trabalhos que aplicam a ADF nessa área. Ademais, Vinué e Epifanio (2019) fazem previsão do desempenho de jogadores de basquete através de ADF, considerando inclusive que esses dados são possivelmente esparsos (poucos pontos avaliados para cada função) e irregulares (o domínio de cada função pode ser diferente). Quanto a Séries Temporais Funcionais, até o presente momento não há conhecimento de aplicações aos esportes.

Assim sendo, percebe-se que os esportes, predominantemente o futebol, constituem uma área pouco explorada pela ótica de ADF e, especialmente, de Séries Temporais Funcionais, o que, conjuntamente ao fato de o futebol estar substancialmente mais atrasado no uso de modelagem estatística, corresponde a uma lacuna que deve ser preenchida mediante esforços de pesquisadores e de agentes tomadores de decisão no futebol. Além disso, Lucey et al. (2013) citam que dados espaço-temporais são dados de alta dimensionalidade e suscetíveis a uma grande quantidade de ruídos. Portanto, são um tipo de dado propenso ao uso de metodologias de Séries Temporais Funcionais para modelagem e previsão. Em particular, a teoria proposta por Bathia, Yao e Ziegelmann (2010), juntamente com a generalização de Horta e Ziegelmann (2016), são metodologias potencialmente adequadas para trabalhar com tal tipo de dado.

3.2 Notação e convenções

Previamente à exposição da metodologia utilizada, é importante destacar algumas convenções e notações empregadas no texto daqui em diante. A série temporal funcional $(\lambda_t \colon t \in \mathbb{Z}_+)$ é formada por superfícies aleatórias $\lambda_t(\boldsymbol{u})$, $\boldsymbol{u} \in S := [0,1] \times [0,1]$.[4] As superfícies (λ_t) são definidas em um mesmo espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$, tal que $\lambda_t \colon \Omega \to L^2(S)$ para todo t, em que $L^2(S)$ é o espaço de Hilbert de funções quadrado-integráveis em S. O produto interno entre duas funções

[4] Posteriormente, S será interpretado como uma reparametrização geográfica de um campo de futebol.

reais $f, g \in L^2(S)$ é dado por

$$\langle f, g \rangle := \int_S f(\boldsymbol{u}) g(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u},$$

onde $\int_S f(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} := \int_0^1 \int_0^1 f(u_1, u_2) \, \mathrm{d}u_1 \, \mathrm{d}u_2$. A respectiva norma induzida pelo produto interno acima (norma L^2) é definida por

$$\|f\| := \left(\int_S \left\{ f(\boldsymbol{u}) \right\}^2 \mathrm{d}\boldsymbol{u} \right)^{1/2}.$$

Cabe ressaltar que os mapas de intensidade podem ser interpretados como as superfícies (λ_t), onde as diferentes colorações (intensidades) correspondem às diferentes alturas da superfície. Logo, no que segue, ao decorrer do texto, os termos *série temporal funcional*, *superfícies* e *mapas de intensidade* serão intercambiados livremente, representando o mesmo conceito. Além disso, do ponto de vista de ADF, os mapas de intensidade *são* as unidades observacionais. Assim, não se está interessado em modelar a dinâmica intrapartida. Para noções de tipos de modelos com tal grau de detalhamento, consultar Narayanan (2019).

A metodologia introduzida por Bathia, Yao e Ziegelmann (2010), a qual tem como objetivo determinar a dimensionalidade finita de séries temporais formadas por curvas, é apresentada a seguir, devidamente adaptada para o caso em que as unidades amostrais são superfícies. Como destacado anteriormente, a validade dessa adaptação baseia-se no desenvolvimento de Horta e Ziegelmann (2016).

3.3 Análise de Séries Temporais Funcionais: Representações espectrais baseadas no operador de autocovariância defasado

De acordo com Bathia, Yao e Ziegelmann (2010), uma característica prevalente em dados funcionais é a presença de ruídos de observação, os quais são decorrentes, entre outros motivos, de erros de medida (no caso de mapas de intensidade no futebol, esse erro de medida surge devido ao fato de que os dados relativos às ações da partida são coletados por anotadores via vídeo) e, além disso, como a observação funcional é obtida após a suavização de pontos discretos, espera-se que haja a inclusão de algum erro de aproximação advindo deste processo. Ainda, no contexto de mapas de intensidade, os ruídos de observação são resultantes, também, do fato de que os mapas observados são estimativas de mapas de intensidade latentes. Ou seja, não se tem acesso observacional à série temporal funcional de interesse (λ_t) – na aplicação, λ_t representará o mapa de intensidade latente da t-ésima partida. Observa-se unicamente a série temporal $(\widehat{\lambda}_t)$, conforme

$$\widehat{\lambda}_t(\boldsymbol{u}) = \lambda_t(\boldsymbol{u}) + \varepsilon_t(\boldsymbol{u}), \qquad \boldsymbol{u} \in S,$$

em que (ε_t) é ruído branco.

A hipótese feita acima sobre (ε_t) é de extrema importância e traz intrinsecamente as seguintes suposições:

(i) $\mathbb{E}\Big\{\varepsilon_t(\boldsymbol{u})\Big\} = 0, \ \forall\, \boldsymbol{u} \in S, \ \forall\, t;$

(ii) $\operatorname{Cov}\Big\{\varepsilon_t(\boldsymbol{u}), \varepsilon_s(\boldsymbol{v})\Big\} = 0, \ \forall\, \boldsymbol{u}, \boldsymbol{v} \in S, \ t \neq s;$

(iii) $\operatorname{Cov}\Big\{\lambda_t(\boldsymbol{u}), \varepsilon_{t+\ell}(\boldsymbol{v})\Big\} = 0, \ \forall\, \boldsymbol{u}, \boldsymbol{v} \in S, \ \forall\, t, \ \forall\, \ell \in \mathbb{Z}.$

Além disso, assume-se que

(iv) (λ_t) é estritamente estacionária.

Sob a suposição (*iv*), tanto a função média

$$\mu(\boldsymbol{u}) := \mathbb{E}\Big\{\lambda_t(\boldsymbol{u})\Big\}, \qquad \boldsymbol{u} \in S,$$

quanto as funções de autocovariância na ℓ-ésima defasagem

$$c_\ell(\boldsymbol{u}, \boldsymbol{v}) := \operatorname{Cov}\Big\{\lambda_t(\boldsymbol{u}), \lambda_{t+\ell}(\boldsymbol{v})\Big\}, \qquad \boldsymbol{u}, \boldsymbol{v} \in S, \quad \ell \in \mathbb{Z}, \tag{3.1}$$

de (λ_t) não dependem de t. A função de autocovariância acima pode ser vista como um *kernel* positivo-definido $c_\ell \colon S \times S \to \mathbb{R}$, no sentido de que c_ℓ satisfaz a condição

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \cdot c_\ell\Big(\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(j)}\Big) \geq 0,$$

para todos os conjuntos de elementos $\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(m)} \in S$ e números reais $a_1, \ldots, a_m$, $m \in \mathbb{Z}_+$ (MINH; NIYOGI; YAO, 2006). O operador linear de autocovariância $C_\ell \colon L^2(S) \to L^2(S)$ associado ao *kernel* c_ℓ é definido por

$$(C_\ell f)(\boldsymbol{u}) := \int_S c_\ell(\boldsymbol{u}, \boldsymbol{v}) f(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u} \in S, \quad f \in L^2(S).$$

Segundo Wang, Chiou e Müller (2016), a abordagem usual em ADF é a Análise de Componentes Principais Funcionais, a qual considera que, sob a suposição de que $\int_S \mathbb{E}\{(\lambda_t(\boldsymbol{u}))^2 + (\varepsilon_t(\boldsymbol{u}))^2\} \, \mathrm{d}\boldsymbol{u} < \infty$, os termos λ_t da série temporal (λ_t) podem ser representados pela decomposição de Karhunen-Loève (Karhunen, 1946; Loève, 1946)

$$\lambda_t(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \sum_{j=1}^{\infty} \xi_{tj} \varphi_j(\boldsymbol{u}), \qquad \boldsymbol{u} \in S, \, t \in \mathbb{Z}_+,$$

onde $\xi_{tj} := \langle \lambda_t - \mu, \varphi_j \rangle$ são variáveis aleatórias tais que $\mathbb{E}(\xi_{tj}) = 0$ e $\mathbb{V}\mathrm{ar}(\xi_{tj}) =: \pi_j$, e φ_j são autofunções do operador de autocovariância no *lag* zero C_0. Ou seja, as superfícies (λ_t) podem ser expressas em termos das autofunções do seu operador de autocovariância C_ℓ, $\ell = 0$. A igualdade acima vale no sentido de que $\lim_{\kappa\to\infty} \|\lambda_t - \mu - \sum_{j=1}^{\kappa} \xi_{tj}\varphi_j\| = 0$ quase certamente (HORTA; ZIEGELMANN, 2016).

Os autovalores de C_0 são dados pela sequência $(\pi_1, \pi_2, \dots)$, com $\pi_1 \geq \pi_2 \geq \dots$ Para assegurar que essa sequência está bem definida, segue-se a convenção de que ela só possui zeros se o operador C_0 tem posto finito (HORTA; ZIEGELMANN, 2018). Ademais, sendo $\{\varphi_1, \varphi_2, \dots\}$ um conjunto ortonormal em $L^2(S)$, cujos elementos são as respectivas autofunções associadas a esses autovalores, então, pelo Teorema de Mercer (MERCER, 1909), a função de autocovariância não-defasada c_0 admite uma representação espectral da forma

$$c_0(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{\infty} \pi_j \varphi_j(\boldsymbol{u}) \varphi_j(\boldsymbol{v}), \qquad \boldsymbol{u}, \boldsymbol{v} \in S.$$

As autofunções φ_j são obtidas a partir da equação integral $\int_S c_0(\boldsymbol{u}, \boldsymbol{v})\varphi_j(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v} = \pi_j \varphi_j(\boldsymbol{u})$. No entanto, tal procedimento não é o mais adequado quando o interesse é modelar variáveis funcionais sujeitas a ruídos de observação tal qual as superfícies (λ_t), pois, sob estacionariedade, $\mathrm{Cov}\{\widehat{\lambda}_t(\boldsymbol{u}), \widehat{\lambda}_t(\boldsymbol{v})\} = c_0(\boldsymbol{u}, \boldsymbol{v}) + \mathrm{Cov}\{\varepsilon_0(\boldsymbol{u}), \varepsilon_0(\boldsymbol{v})\} \neq c_0(\boldsymbol{u}, \boldsymbol{v})$, $\boldsymbol{u}, \boldsymbol{v} \in S$, exceto no caso em que ε_0 é ruído branco quando visto como um processo com espaço de índices S. Sendo assim, Bathia, Yao e Ziegelmann (2010) introduzem o *kernel* não-negativo k em $L^2(S)$, definido por

$$k(\boldsymbol{u}, \boldsymbol{v}) := \sum_{\ell=1}^{p} \int_S c_\ell(\boldsymbol{u}, \boldsymbol{w}) c_\ell(\boldsymbol{v}, \boldsymbol{w}) d\boldsymbol{w}, \tag{3.2}$$

com $\boldsymbol{u}, \boldsymbol{v} \in S$ e $p \in \mathbb{Z}_+$, baseando-se no fato de que $\mathrm{Cov}\{\widehat{\lambda}_t(\boldsymbol{u}), \widehat{\lambda}_{t+\ell}(\boldsymbol{v})\} = c_\ell(\boldsymbol{u}, \boldsymbol{v})$, $\ell \neq 0$. Conforme argumento exposto em Bathia, Yao e Ziegelmann (2010, p. 3359), o inteiro p expresso na construção do *kernel* k em (3.2) é utilizado com o objetivo de facilitar o cômputo das autofunções do operador integral $K\colon L^2(S) \to L^2(S)$, o qual é definido por

$$(Kf)(\boldsymbol{u}) := \int_S k(\boldsymbol{u}, \boldsymbol{v}) f(\boldsymbol{v})\, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u} \in S, \quad f \in L^2(S).$$

Sob a Suposição A1 em Horta e Ziegelmann (2016), a série temporal funcional (λ_t) pode ser expandida em termos das autofunções do operador K, o que, pela representação de Karhunen-Loève, equivale a escrever

$$\lambda_t(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \sum_{j=1}^{\infty} \eta_{tj} \psi_j(\boldsymbol{u}), \qquad \boldsymbol{u} \in S, \tag{3.3}$$

onde $\psi_j \in L^2(S)$ denotam as autofunções de K. As variáveis aleatórias η_{tj} representam os coeficientes das projeções das superfícies (λ_t) no subespaço gerado pelas autofunções ψ_j e, portanto, são obtidas através do produto interno

$$\eta_{tj} := \langle \lambda_t - \mu, \psi_j \rangle = \int_S \big\{\lambda_t(\boldsymbol{u}) - \mu(\boldsymbol{u})\big\}\psi_j(\boldsymbol{u})\ \mathrm{d}\boldsymbol{u}.$$

Além disso, segue que $\mathbb{E}(\eta_{tj}) = 0$ e $\mathbb{V}\mathrm{ar}(\eta_{tj}) =: \theta_j$.

Empiricamente, assume-se que (λ_t) possui dimensão finita, denotada por d, a qual é considerada um parâmetro do modelo. Isto é, o operador linear K possui exatamente d autovalores não-nulos (BATHIA; YAO; ZIEGELMANN, 2010). Logo, denotando por $(\theta_1, \theta_2, \ldots, \theta_d)$ o vetor de autovalores não-nulos de K, em que $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_d > 0$, com conjunto de autofunções correspondentes $\{\psi_1, \psi_2, \ldots, \psi_d\}$, vale a decomposição espectral:

$$k(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{d} \theta_j \psi_j(\boldsymbol{u})\psi_j(\boldsymbol{v}), \qquad \boldsymbol{u}, \boldsymbol{v} \in S,$$

e, consequentemente, a equação integral

$$\int_S k(\boldsymbol{u}, \boldsymbol{v})\psi_j(\boldsymbol{v})\ \mathrm{d}\boldsymbol{v} = \theta_j \psi_j(\boldsymbol{u}), \qquad \boldsymbol{u} \in S.$$

Implicitamente, assume-se que o conjunto $\{\psi_1, \psi_2, \ldots, \psi_d\}$ é ortonormal e adiciona-se, por conseguinte, as condições $\langle \psi_i, \psi_j \rangle = 0$, $i \neq j$, e $\|\psi_j\| = 1$. De acordo com a observação feita acima sobre a dimensão de (λ_t), a decomposição (3.3) pode ser alternativamente expressa através de uma representação finita:

$$\lambda_t(\boldsymbol{u}) = \mu(\boldsymbol{u}) + \sum_{j=1}^{d} \eta_{tj}\psi_j(\boldsymbol{u}), \qquad \boldsymbol{u} \in S. \tag{3.4}$$

Um fator chave sob essa metodologia é que, como os mapas de intensidade λ_t são expandidos como combinação linear das autofunções $\psi_1, \ldots, \psi_d$, cujos coeficientes são dados pelos componentes do vetor aleatório $\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{td})'$, a dinâmica das superfícies λ_t é diretamente capturada pela dinâmica da série temporal vetorial

$$\begin{pmatrix} \eta_{11} \\ \vdots \\ \eta_{1d} \end{pmatrix}, \begin{pmatrix} \eta_{21} \\ \vdots \\ \eta_{2d} \end{pmatrix}, \ldots, \begin{pmatrix} \eta_{n1} \\ \vdots \\ \eta_{nd} \end{pmatrix}, \ldots \tag{3.5}$$

Assim, o problema de modelar a série temporal funcional (λ_t) se resume à modelagem da série temporal multivariada (3.5), que pode ser realizada através de métodos clássicos, tais como modelos da família VARMA – *Vector Autoregression Moving-Average* – (BATHIA; YAO; ZIEGELMANN, 2010).

Embora Bathia, Yao e Ziegelmann (2010) tenham proposto essa abordagem

alternativa à modelagem FAR de Bosq (2000) e suas extensões, Aue, Norinho e Hörmann (2015) reportaram tal metodologia descrevendo-a explicitamente e justificando seu emprego sob o ponto de vista teórico. No entanto, os autores utilizam uma redução de dimensionalidade da série temporal funcional de interesse baseada na convencional Análise de Componentes Principais Funcionais. Além disso, também propõem um algoritmo para a previsão de séries temporais funcionais, inclusive considerando o caso em que há a presença de covariáveis exógenas.

O Teorema 3.1 de Aue, Norinho e Hörmann (2015) afirma que se a série temporal funcional (λ_t) segue um modelo FAR(1), então os coeficientes $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{td})'$ das projeções de (λ_t) nas primeiras d componentes principais seguem um modelo VAR(1), e as previsões h passos à frente da série (λ_t), obtidas via modelagem VAR dos coeficientes ξ_{tj}, são assintoticamente equivalentes à previsão FAR. A partir de uma representação de um espaço de Hilbert mais geral, o Corolário 3.1 do mesmo trabalho indica que a equivalência assintótica também é válida para ordens superiores dos modelos FAR e VAR. Todos esses resultados são derivados sob a suposição de estacionariedade (*iv*).

3.4 Estimadores e resultados amostrais

Em aplicações práticas, considera-se ter acesso a uma amostra aleatória de mapas de intensidade $\widehat{\lambda}_1(\cdot), \ldots, \widehat{\lambda}_n(\cdot)$. Assim, a função de autocovariância amostral na ℓ-ésima defasagem de $(\widehat{\lambda}_t)$ é definida por

$$\widehat{c}_\ell(\boldsymbol{u}, \boldsymbol{v}) := \frac{1}{n-p} \sum_{t=1}^{n-p} \Big\{\widehat{\lambda}_t(\boldsymbol{u}) - \widehat{\mu}(\boldsymbol{u})\Big\}\Big\{\widehat{\lambda}_{t+\ell}(\boldsymbol{v}) - \widehat{\mu}(\boldsymbol{v})\Big\}, \qquad \boldsymbol{u}, \boldsymbol{v} \in S,$$

onde

$$\widehat{\mu}(\boldsymbol{u}) := \frac{1}{n} \sum_{t=1}^{n} \widehat{\lambda}_t(\boldsymbol{u}), \qquad \boldsymbol{u} \in S,$$

é a média amostral das superfícies $(\widehat{\lambda}_t)$. Subsequentemente, pode-se definir um estimador para o *kernel* k introduzido na Equação (3.2):

$$\widehat{k}(\boldsymbol{u}, \boldsymbol{v}) := \sum_{\ell=1}^{p} \int_S \widehat{c}_\ell(\boldsymbol{u}, \boldsymbol{w}) \widehat{c}_\ell(\boldsymbol{v}, \boldsymbol{w}) \, d\boldsymbol{w}, \qquad \boldsymbol{u}, \boldsymbol{v} \in S. \tag{3.6}$$

Agora, conforme Bathia, Yao e Ziegelmann (2010), sendo $\widehat{d}$ o número de autovalores não-nulos do operador integral $\widehat{K}\colon L^2(S) \to L^2(S)$, dado por

$$(\widehat{K}f)(\boldsymbol{u}) := \int_S \widehat{k}(\boldsymbol{u}, \boldsymbol{v}) f(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v}, \qquad \boldsymbol{u} \in S, \quad f \in L^2(S),$$

os quais satisfazem $\widehat{\theta}_1 \geq \widehat{\theta}_2 \geq \cdots \geq \widehat{\theta}_{\widehat{d}} > 0$, e sendo $\{\widehat{\psi}_1, \widehat{\psi}_2, \ldots, \widehat{\psi}_{\widehat{d}}\}$ o conjunto formado pelas autofunções correspondentes a esses autovalores, então o *kernel* $\widehat{k}$ definido em (3.6) pode ser expresso através da decomposição espectral

$$\widehat{k}(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{\widehat{d}} \widehat{\theta}_j \widehat{\psi}_j(\boldsymbol{u}) \widehat{\psi}_j(\boldsymbol{v}), \qquad \boldsymbol{u}, \boldsymbol{v} \in S,$$

em que $\widehat{\theta}_j$ e $\widehat{\psi}_j$ satisfazem a equação integral

$$\int_S \widehat{K}(\boldsymbol{u}, \boldsymbol{v}) \widehat{\psi}_j(\boldsymbol{v}) \, \mathrm{d}\boldsymbol{v} = \widehat{\theta}_j \widehat{\psi}_j(\boldsymbol{u}), \qquad \boldsymbol{u} \in S,$$

e as condições de ortornormalidade $\langle \widehat{\psi}_i, \widehat{\psi}_j \rangle = 0$, $i \neq j$, e $\|\widehat{\psi}_j\| = 1$.

Por intermédio de uma análise de autovalores de uma matriz finita, é possível encontrar $\widehat{\theta}_j$ e $\widehat{\psi}_j$ que, segundo o Teorema 1 em Bathia, Yao e Ziegelmann (2010), são estimadores consistentes para θ_j e ψ_j, respectivamente, $j = 1, \ldots, \widehat{d}$. Assim, segue que o operador $\widehat{K}$ pode ser representado por uma matriz de dimensões infinitas

$$\widehat{\boldsymbol{K}} := \frac{1}{(n-p)^2} \boldsymbol{\gamma}_0 \sum_{\ell=1}^{p} \boldsymbol{\gamma}_\ell' \cdot \boldsymbol{\gamma}_\ell \cdot \boldsymbol{\gamma}_0'.$$

Bathia, Yao e Ziegelmann (2010) fazem uso da seguinte propriedade entre duas matrizes de mesma dimensão $\boldsymbol{A}$ e $\boldsymbol{B}$: $\boldsymbol{AB}'$ possui os mesmos autovalores não-nulos que $\boldsymbol{B}'\boldsymbol{A}$. Logo, o problema de fazer uma análise de autovalores da matriz de dimensões infinitas $\widehat{\boldsymbol{K}}$ se reduz a uma análise de autovalores da matriz $\boldsymbol{K}^*$, de ordem $(n-p) \times (n-p)$, dada por

$$\boldsymbol{K}^* := \frac{1}{(n-p)^2} \sum_{\ell=1}^{p} \boldsymbol{\gamma}_\ell' \cdot \boldsymbol{\gamma}_\ell \cdot \boldsymbol{\gamma}_0' \boldsymbol{\gamma}_0. \tag{3.7}$$

em que

(*M1*) $\boldsymbol{A} := \boldsymbol{\gamma}_0 = (\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_{n-p})$;

(*M2*) $\boldsymbol{B}' := \sum_{\ell=1}^{p} \boldsymbol{\gamma}_\ell' \cdot \boldsymbol{\gamma}_\ell \cdot \boldsymbol{\gamma}_0'$.

Em (*M1*), $\boldsymbol{\Lambda}_t$ é um vetor coluna de infinitas linhas, cujas entradas são dadas pela diferença $\widehat{\lambda}_t(\boldsymbol{u}) - \widehat{\mu}(\boldsymbol{u})$, $\forall\ \boldsymbol{u} \in S$. Já em (*M2*), a entrada (ts) da matriz $\boldsymbol{\gamma}_\ell' \cdot \boldsymbol{\gamma}_\ell$ é dada pelo produto interno $\boldsymbol{\Lambda}_{t+\ell}' \cdot \boldsymbol{\Lambda}_{s+\ell} = \langle \widehat{\lambda}_{t+\ell} - \widehat{\mu}, \widehat{\lambda}_{s+\ell} - \widehat{\mu} \rangle$. Portanto, se $\widehat{\theta}_1, \ldots, \widehat{\theta}_{\widehat{d}}$ denotam os $\widehat{d}$ maiores autovalores de $\boldsymbol{K}^*$ e $\widehat{\boldsymbol{\Psi}}_1, \ldots, \widehat{\boldsymbol{\Psi}}_{\widehat{d}}$, os $\widehat{d}$ autovetores correspondentes, com $\widehat{\boldsymbol{\Psi}}_j := (\Psi_{1j}, \ldots, \Psi_{n-p,j})$, então, $\widetilde{\psi}_1, \ldots, \widetilde{\psi}_{\widehat{d}}$ são as $\widehat{d}$ autofunções do operador $\widehat{K}$, onde

$$\widetilde{\psi}_j(\boldsymbol{u}) = \sum_{t=1}^{n-p} \Psi_{tj} \Big\{ \widehat{\lambda}_t(\boldsymbol{u}) - \widehat{\mu}(\boldsymbol{u}) \Big\}, \qquad \boldsymbol{u} \in S.$$

As autofunções $\widetilde{\psi}_1, \ldots, \widetilde{\psi}_{\widehat{d}}$ encontradas através da análise matricial acima podem não ser ortogonais entre si e de norma unitária. Dessa forma, o processo de ortonormalização de Gram-Schmidt pode ser aplicado a essas autofunções a fim de se obter um conjunto ortonormal em $L^2(S)$, cujos elementos serão as autofunções $\widehat{\psi}_1, \ldots, \widehat{\psi}_{\widehat{d}}$ (BATHIA; YAO; ZIEGELMANN, 2010).

Finalmente, introduz-se um estimador para os mapas de intensidade observados $\widehat{\lambda}_t$:

$$\widehat{\lambda}_t^*(\boldsymbol{u}) = \widehat{\mu}(\boldsymbol{u}) + \sum_{j=1}^{d_0} \widehat{\eta}_{tj}\widehat{\psi}_j(\boldsymbol{u}), \qquad \boldsymbol{u} \in S, \tag{3.8}$$

onde

$$\widehat{\eta}_{tj} := \langle \widehat{\lambda}_t - \widehat{\mu}, \widehat{\psi} \rangle = \int_S \left\{ \widehat{\lambda}_t(\boldsymbol{u}) - \widehat{\mu}(\boldsymbol{u}) \right\} \widehat{\psi}_j(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u},$$

são os coeficientes das projeções das superfícies $\widehat{\lambda}_t$ no subespaço (finito) gerado pelas autofunções $\widehat{\psi}_j$.

O inteiro d_0 utilizado na soma em (3.8) é um estimador de d, uma vez que pode ocorrer de $\widehat{d} \gg d$. Bathia, Yao e Ziegelmann (2010) implementam um teste baseado em *bootstrap* para determinar d_0, onde a hipótese nula a ser testada é $H_0 \colon \theta_{d_0+1} = 0$. Além disso, os autores também sugerem uma abordagem empírica, na qual seleciona-se a dimensão d_0 como sendo o número significativo de maiores autovalores de $\widehat{K}$, no sentido de que acontece uma queda abrupta do autovalor $\widehat{\theta}_{d_0+1}$ em relação ao autovalor $\widehat{\theta}_{d_0}$. Como o interesse deste trabalho se concentra em um viés preditivo, na aplicação optou-se por uma escolha baseada na minimização do Erro Quadrático Médio Integrado (EQMI) de previsão, complementando a segunda alternativa destacada acima.

Obtendo-se as quantidades apresentadas nesta seção, modela-se a série temporal multivariada $\widehat{\boldsymbol{\eta}}_t = (\widehat{\eta}_{t,1}, \ldots, \widehat{\eta}_{t,d_0})'$, $t = 1, \ldots, n$, a partir de alguma técnica adequada para esse tipo de problema. Ademais, na presença de um vetor observado de covariáveis exógenas $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tr})'$, $r \in \mathbb{N}$, modelos que abordam essa caracterização adicional devem ser considerados – na aplicação, modelos *VARX* foram ajustados. Dessa forma, a previsão um passo à frente para a série temporal $(\widehat{\boldsymbol{\eta}}_t)$ é denotada por $\widehat{\boldsymbol{\eta}}_{n+1|n} = (\widehat{\eta}_{n+1,1|n}, \ldots, \widehat{\eta}_{n+1,d_0|n})'$ e o mapa de intensidade predito $\widehat{\lambda}_{n+1|n}$ pode ser recuperado via representação de Karhunen-Loève, isto é,

$$\widehat{\lambda}_{n+1|n}(\boldsymbol{u}) = \widehat{\mu}_{|n}(\boldsymbol{u}) + \sum_{j=1}^{d_0} \widehat{\eta}_{n+1,j|n}\widehat{\psi}_{j|n}(\boldsymbol{u}), \qquad \boldsymbol{u} \in S,$$

em que as notações $\widehat{\mu}_{|n}, \widehat{\psi}_{j|n}$ são usadas para indicar, respectivamente, a função média e as autofunções estimadas através da amostra até o tempo n.

Um algoritmo contendo cada passo para o procedimento de modelagem e pre-

visão da série temporal funcional $(\widehat{\lambda}_t)$ com o acréscimo de covariáveis exógenas, cuja construção foi baseada no Algoritmo 2 do trabalho de Aue, Norinho e Hörmann (2015), é descrito abaixo:

1. Determine d_0. Decomponha a série temporal funcional observada $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n$, utilizando a representação de Karhunen-Loève, para obter a série temporal vetorial $\widehat{\boldsymbol{\eta}}_t = (\widehat{\eta}_{t,1}, \ldots, \widehat{\eta}_{t,d_0})'$;

2. Modele a série temporal $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_n$ com a adição do vetor de covariáveis exógenas $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nr})'$, $r \in \mathbb{N}$, através do modelo *VARX*, por exemplo, concebendo assim a previsão um passo à frente

$$\widehat{\boldsymbol{\eta}}_{n+1|n} = (\widehat{\eta}_{n+1,1|n}, \ldots, \widehat{\eta}_{n+1,d_0|n})',$$

   para $\widehat{\boldsymbol{\eta}}_{n+1}$;

3. Use novamente a representação de Karhunen-Loève para gerar a previsão um passo à frente

$$\widehat{\lambda}_{n+1|n}(\boldsymbol{u}) = \widehat{\mu}_{|n}(\boldsymbol{u}) + \sum_{j=1}^{d_0} \widehat{\eta}_{n+1,j|n} \widehat{\psi}_{j|n}(\boldsymbol{u}), \qquad \boldsymbol{u} \in S,$$

   para o mapa de intensidade $\widehat{\lambda}_{n+1}$.

A aplicação da metodologia descrita neste capítulo será realizada na sequência, em um estudo de mapas de intensidade da equipe espanhola Fútbol Club Barcelona.

4 Aplicação: Análise de mapas de intensidade do FC Barcelona

4.1 Fonte de dados

Os dados que serão usados para aplicação da metodologia de séries temporais funcionais descrita no Capítulo 3 foram obtidos pela empresa de coleta e análise de dados de futebol *StatsBomb*. A disponibilização desse material de maneira gratuita para o público em geral (algo raro no meio dos dados futebolísticos) parte de uma iniciativa da empresa de facilitar e propiciar novas pesquisas e descobertas para a análise estatística no futebol (STATSBOMB, 2019a). Tais dados registram informações de todas as partidas da carreira do jogador argentino Lionel Messi pelo FC Barcelona no Campeonato Espanhol entre as temporadas 2004/2005 e 2018/2019, totalizando mais de 450 partidas. Os dados são do tipo *event logs*, isto é, contêm todas as informações de cada ação envolvendo a bola realizada nas partidas. Os eventos são dos mais variados tipos, totalizando 34 categorias diferentes, das quais foram selecionadas somente aquelas feitas sob posse do FC Barcelona, partindo-se do princípio de que o padrão de jogo de uma equipe é caracterizado sobretudo por seu comportamento quando detém a posse da bola. A listagem e descrição dos eventos escolhidos podem ser consultadas na Seção A.1 do Apêndice A.

Após seleção dos eventos, considerando-se todas as partidas, os mais comuns são *passe* (45,98% do total) e *recebimento de passe* (44,65%). Cada ação ocorrida no jogo é caracterizada por diversas variáveis, entre elas *tipo do evento*, *time*, *jogador*, *minuto*, *segundo* e *coordenadas* v_1 *e* v_2 do campo onde a ação aconteceu, sendo estas últimas as mais importantes para a construção de mapas de intensidade. Para ter acesso de forma completa aos dados e a suas documentações, ver StatsBomb (2019a) e StatsBomb (2019b).

4.2 Organização dos dados e geração dos mapas de intensidade

Após *download* do conjunto de dados, o processamento, organização e filtragem foram realizados em linguagem `R` (R Core Team, 2021) por intermédio do *software* RStudio (RStudio Team, 2021), versão 1.4.1106. Primeiramente, as informações de cada partida, que estavam contidas em bancos de dados separados no formato `.JSON`, foram agregadas por temporadas em objetos do tipo *list*. Na sequência, devido ao fato de que apenas sete de 38 partidas da temporada 2004/2005 e 17 de 38 da temporada 2005/2006 haviam sido contempladas no registro dos dados, decidiu-se por descartá-las da análise. Nas demais temporadas, em ao menos 25 dos 38 jogos do FC Barcelona, Lionel Messi esteve em campo, o que configura uma representação quase completa das partidas do clube na competição durante o período de agosto de 2005 a maio de 2019 (428 de 494 partidas), embora o mais adequado fosse ter acesso às informações de todas as 38 partidas da equipe em cada temporada.

Como destacado na Seção 4.1, o passo seguinte compreendeu a filtragem das ações realizadas exclusivamente pelo FC Barcelona em situações nas quais a equipe detinha a posse da bola. Após esse processo, os dados foram armazenados em um único banco ordenado de acordo com a ocorrência temporal de cada partida. Assim, o primeiro jogo da temporada 2005/2006 foi identificado com o valor 1, ao passo que o valor 428 definiu a última partida da temporada 2018/2019. A Tabela 4.1 mostra o resultado dessa etapa.

Tabela 4.1: Estrutura do banco de dados final.

Time	Evento	Coord. v_1	Coord. v_2	Oponente	Temporada	Partida
Barcelona	*Pass*	50,9	61,8	Celta	2006/2007	1
Barcelona	*Ball Receipt*	63,7	70,0	Celta	2006/2007	1
Barcelona	*Miscontrol*	65,8	68,9	Celta	2006/2007	1
Barcelona	*Pass*	39,8	80,0	Celta	2006/2007	1
Barcelona	*Ball Receipt*	29,4	60,7	Celta	2006/2007	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Barcelona	*Pass*	7,0	45,0	Eibar	2018/2019	428

Fonte: *StatsBomb*.

Os mapas de intensidade observados $\widehat{\lambda}_t$, $t = 1, \ldots, 428$, foram gerados com base na suavização dos dados brutos apresentados na Tabela 4.1. Para tanto, utilizou-se o método de estimação de densidades via *kernel*. Dessa forma, define-se n_t como o número total de ações do FC Barcelona na partida t e, por conseguinte, $\boldsymbol{v}^{(ti)} :=$

$(v_1^{(ti)}, v_2^{(ti)})$, $i = 1, \ldots, n_t$, denota o par de coordenadas v_1 e v_2 [5] do campo onde a ação i ocorreu. Assim, para uma grade $G \subset S$, a superfície $\widehat{\lambda}_t$ pode ser estimada via *kernel* por

$$\widehat{\lambda}_t(\boldsymbol{u}) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_{\boldsymbol{h}_t}\Big(\boldsymbol{u} - \boldsymbol{v}^{(ti)}\Big), \qquad \boldsymbol{u} \in G.^{6}$$

$K_{\boldsymbol{h}_t}$ é uma função *kernel* que neste caso é a distribuição normal padrão bivariada, ou seja,

$$K_{\boldsymbol{h}_t}(\boldsymbol{u}) := \frac{1}{2\pi \boldsymbol{h}_t^2} \exp\Big\{ -\frac{1}{2}\Big(\frac{\boldsymbol{u}}{\boldsymbol{h}_t}\Big)'\Big(\frac{\boldsymbol{u}}{\boldsymbol{h}_t}\Big)\Big\},$$

em que $\boldsymbol{h}_t := \Big(h_t^{(1)}, h_t^{(2)}\Big)'$ é o vetor que indica o grau de suavidade do mapa de intensidade $\widehat{\lambda}_t$ na direção de cada uma das coordenadas do campo. A determinação do valor desse parâmetro foi baseada na simples ideia do estimador "*rule-of-thumb*" de Silverman, que sugere $\widehat{\boldsymbol{h}}_t = \Big(\widehat{\sigma}_t^{(1)}, \widehat{\sigma}_t^{(2)}\Big)'\, 1,06\; n_t^{-1/5}$, com $\widehat{\sigma}_t^{(1)}$ e $\widehat{\sigma}_t^{(2)}$ representando o desvio padrão estimado de $v_1^{(ti)}$ e $v_2^{(ti)}$, respectivamente (SILVERMAN, 1986; VENABLES; RIPLEY, 2002). A Figura 4.1 ilustra um mapa de intensidade resultante da suavização apresentada acima.

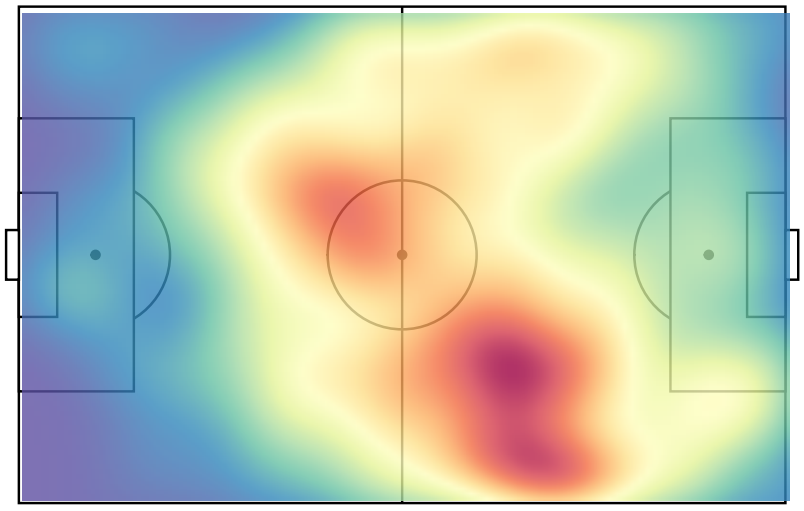


Figura 4.1: Mapa de intensidade do FC Barcelona alusivo à 2ª rodada da temporada 2008/2009 do Campeonato Espanhol diante do Real Racing Club de Santander. Todas as coordenadas do campo são padronizadas para que o FC Barcelona ataque da esquerda para a direita.

[5] Padronizadas para o quadrado unitário S.

[6] Na realidade, $\widehat{\lambda}_t$ obtida desta forma é uma função densidade de probabilidade, devendo ser interpretada como um mapa de intensidade padronizado.

4.3 Modelagem

Nesta seção, serão apresentados os resultados da aplicação da metodologia de Bathia, Yao e Ziegelmann (2010), detalhada nas Seções 3.3 e 3.4, aos mapas de intensidade obtidos após a suavização descrita anteriormente. O código construído em linguagem `R` foi baseado em uma implementação prévia da metodologia. Para o cálculo do estimador do *kernel* da Equação (3.2), utilizou-se, da mesma forma que nas simulações de Bathia, Yao e Ziegelmann (2010), $p = 5$. Entretanto, em primeira instância, fez-se necessário excluir as primeiras 54 observações (referentes às partidas somadas das temporadas 2006/2007 e 2007/2008) da análise devido ao fato de que uma das séries observadas $\widehat{\eta}_{tj}$ – a saber, $\widehat{\eta}_{t1}$ – demonstrava um comportamento em dois regimes. Mais detalhes sobre essa questão podem ser consultados no Apêndice B. Assim, após a exclusão das observações $t = 1, \ldots, 54$, readequando-se a notação para as partidas restantes, tem-se que a amostra final de mapas de intensidade do FC Barcelona considerada na análise é denotada por $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{374}$, ou seja, na nova notação $t = 1, \ldots, 374$, compreendendo as partidas entre as temporadas 2008/2009 e 2018/2019.

Finalmente, um ajuste global abrangendo todas as $n = 374$ observações foi avaliado para fins de análise exploratória e interpretação dos parâmetros funcionais estimados. Contudo, a determinação dos valores de d_0, da ordem do modelo $VARX$ aplicado à série temporal $\widehat{\boldsymbol{\eta}}_t$, bem como das covariáveis exógenas, foi feita com base em uma combinação destes parâmetros cujo valor de EQMI de previsão um passo à frente fosse minimizado, o que será realizado na Seção 4.4.

A Figura 4.2 mostra o mapa de intensidade médio do FC Barcelona, $\widehat{\mu}$, ao longo das $n = 374$ partidas consideradas. É importante salientar que as colorações mais quentes (tons em vermelho) indicam as regiões com maior intensidade de ocupação da equipe com a bola, enquanto que áreas cuja cor associada é mais fria (tons em azul) são aquelas com menor taxa de ocupação. Portanto, percebe-se que, em média, o clube espanhol manteve a bola em quase toda a extensão do meio de campo, o que é algo esperado para um time que baseia seu estilo de jogo em uma característica de retenção de posse de bola. Também, vê-se que a equipe manteve um padrão médio levemente mais ofensivo, pois a coloração avermelhada é um tanto mais presente à direita da linha central.

Já na Figura 4.3 encontram-se os dez primeiros autovalores $\widehat{\theta}_j$ após a decomposição espectral da matriz $\boldsymbol{K}^*$, apresentada na Equação (3.7). Neste gráfico, pode-se notar que os autovalores $\widehat{\theta}_1$ e $\widehat{\theta}_2$ se destacam em relação aos demais quanto a suas magnitudes, sendo que, além disso, $\widehat{\theta}_1$ é consideravelmente maior que $\widehat{\theta}_2$. A partir desta inspeção visual, é possível concluir que duas dimensões sejam suficientes para representar os mapas de intensidade observados $(\widehat{\lambda}_t)$ através da expansão de

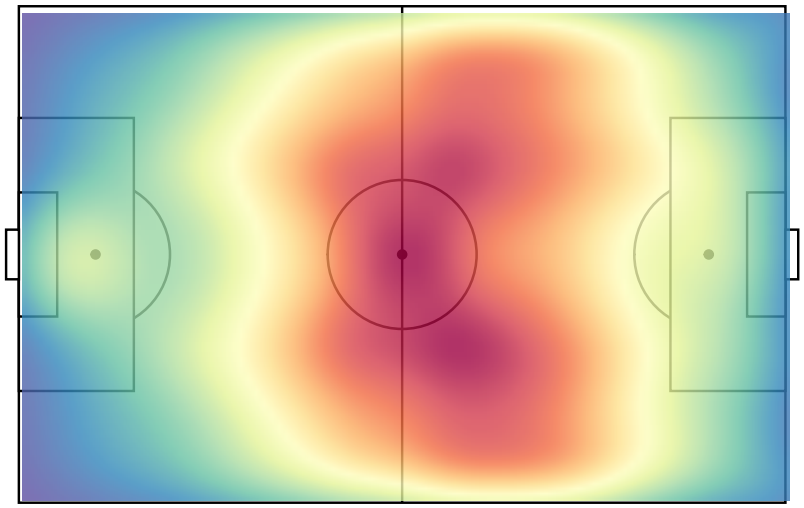


Figura 4.2: Função média $\widehat{\mu}$ após $n = 374$ partidas do FC Barcelona no Campeonato Espanhol entre as temporadas 2008/2009 e 2018/2019.

Karhunen-Loève, isto é, tal análise conduz a uma escolha de $d_0 = 2$. Todavia, na Seção 4.4 foram aplicados alguns cenários de modelos com $d_0 = 2, \ldots, 6$.

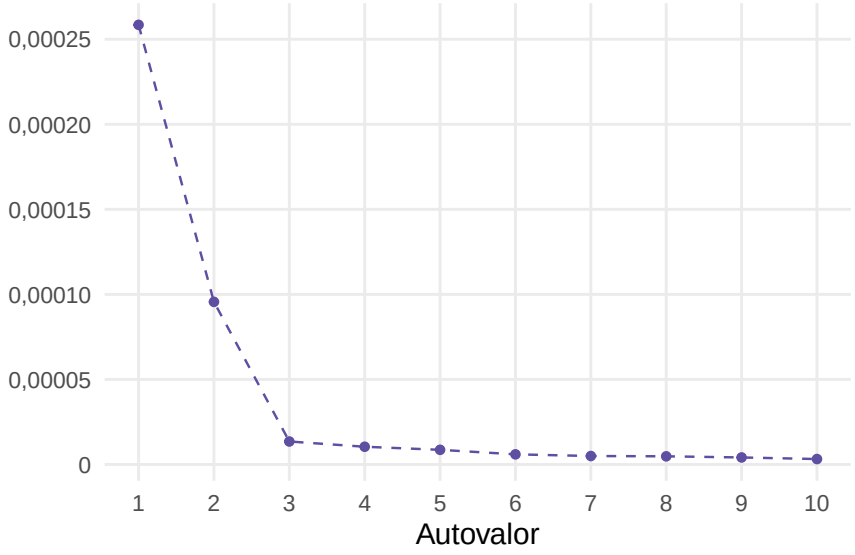
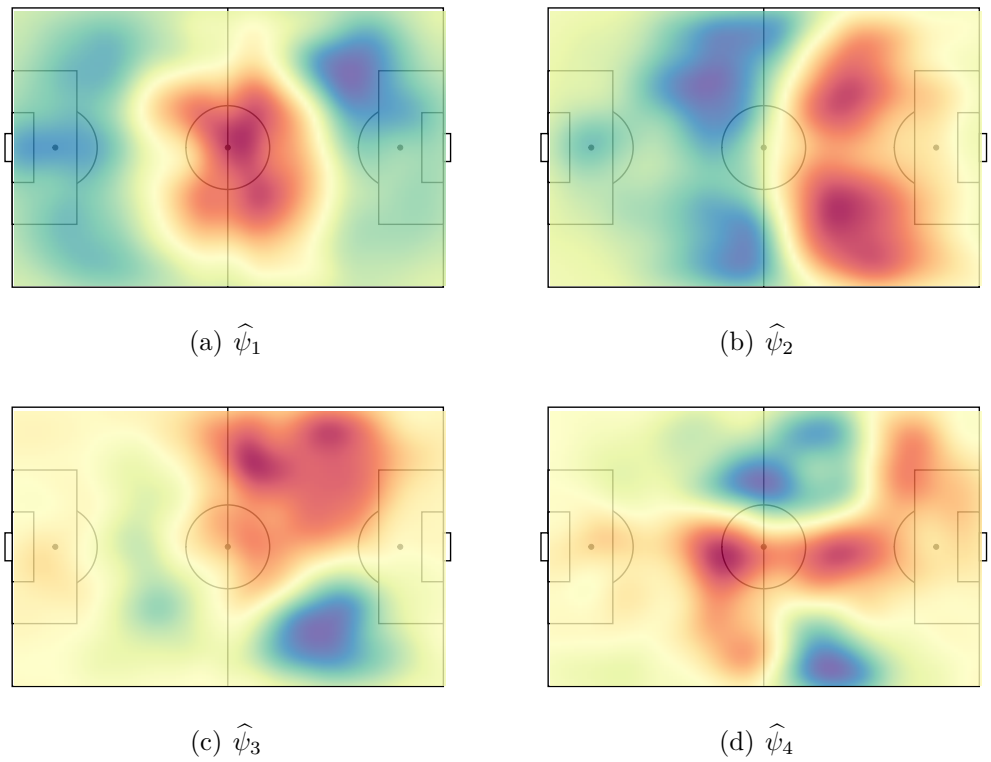


Figura 4.3: Dez primeiros autovalores obtidos via decomposição espectral da matriz $\boldsymbol{K}^*$.

As primeiras quatro autofunções resultantes podem ser observadas na Figura 4.4. Desse modo, em 4.4(a) tem-se que a autofunção $\widehat{\psi}_1$ captura se a equipe portou-se de uma forma mais centralizada no campo de jogo, ao passo que as autofunções $\widehat{\psi}_2$ – painel 4.4(b) – e $\widehat{\psi}_3$ – painel 4.4(c) – indicam ofensividade e o lado do campo em que o time predominou seus ataques, respectivamente. A interpretação prática acerca da aparência de $\widehat{\psi}_4$, autofunção exposta em 4.4(d), já parece ser mais complicada,

assim como das demais autofunções não reportadas no trabalho. Nesse sentido, valores positivos de $\widehat{\eta}_{t1}$ estão associados a jogos em que o time permaneceu mais tempo com a posse de bola no meio de campo, valores positivos de $\widehat{\eta}_{t2}$ caracterizam uma postura ofensiva do FC Barcelona, enquanto que partidas cujo comportamento é defensivo são tais que $\widehat{\eta}_{t2} < 0$. Ademais, se $\widehat{\eta}_{t3} > 0$ (resp. $\widehat{\eta}_{t3} < 0$), então na partida t a equipe ocupou de forma mais expressiva o lado esquerdo (resp. direito) do campo.



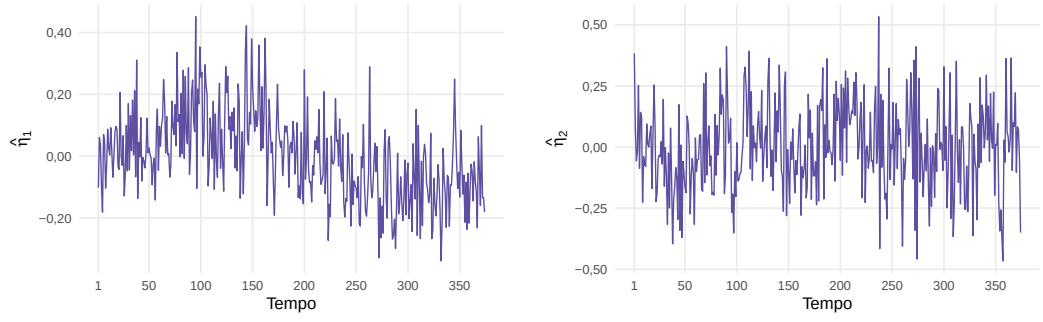
(a) $\widehat{\psi}_1$ (b) $\widehat{\psi}_2$

(c) $\widehat{\psi}_3$ (d) $\widehat{\psi}_4$

Figura 4.4: Autofunções $\widehat{\psi}_j$, $j = 1, \ldots, 4$.

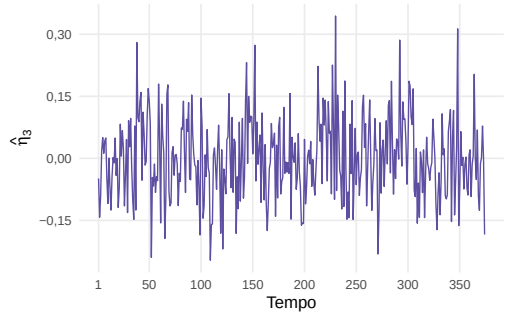
A Figura 4.5 mostra as séries temporais dos coeficientes $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$, a partir da qual não há um indício visual de possível falta de estacionariedade para essas séries. Tal averiguação vai ao encontro dos resultados do teste aumentado de Dickey-Fuller (FULLER, 2009), cujos resultados são listados na Tabela 4.2, os quais sugerem a rejeição da hipótese nula de presença de raiz unitária para as quatro séries (p-valores $< 0,01$).

Ao se analisar a estrutura de correlação das séries temporais $\widehat{\eta}_{tj}$, infere-se que somente $\widehat{\eta}_{t1}$ apresenta uma dependência temporal linear significativa. A Figura 4.6(a) mostra que a função de autocorrelação de $\widehat{\eta}_{t1}$ é significativamente diferente de zero por vários *lags*, enquanto que sua função de autocorrelação parcial, exposta na Figura 4.7(a), apresenta valores significativos para os primeiros *lags*. Dessa forma, o gráfico de dispersão entre a série $\widehat{\eta}_{t1}$ e suas defasagens – Figura 4.8(a) – exibe uma relação linear crescente. De maneira oposta, as funções de autocorrelação nas Figuras

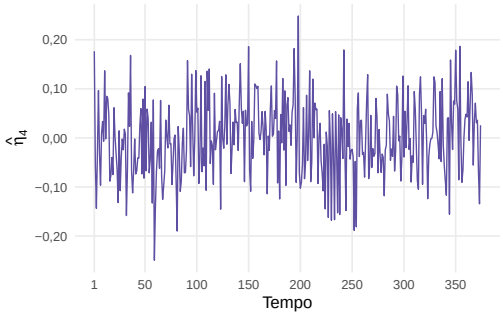


(a) $\widehat{\eta}_{t1}$

(b) $\widehat{\eta}_{t2}$



(c) $\widehat{\eta}_{t3}$



(d) $\widehat{\eta}_{t4}$

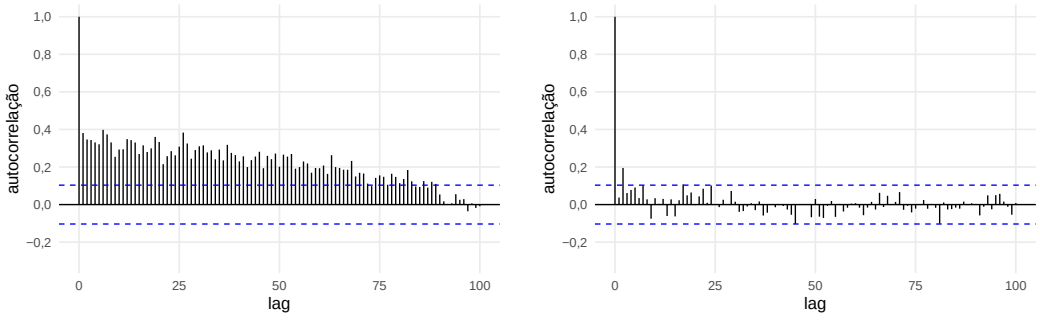
Figura 4.5: Séries temporais $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$, formadas pelos coeficientes das projeções dos mapas de intensidade observados nas autofunções.

Tabela 4.2: Resultado do teste de raiz unitária aumentado de Dickey-Fuller.

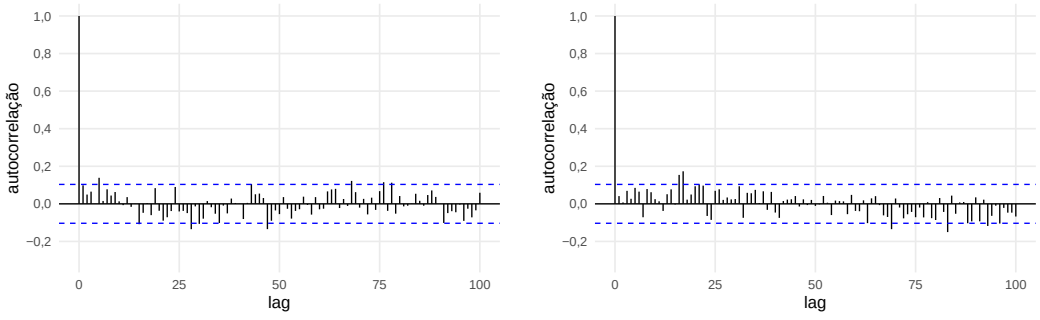
Série Temporal	Estatística de Teste	p-valor
$\widehat{\eta}_{t1}$	-4,2572	<0,01
$\widehat{\eta}_{t2}$	-5,3807	<0,01
$\widehat{\eta}_{t3}$	-5,3701	<0,01
$\widehat{\eta}_{t4}$	-5,8669	<0,01

4.6(b), 4.6(c) e 4.6(d), referentes às séries temporais $\widehat{\eta}_{t2}$, $\widehat{\eta}_{t3}$ e $\widehat{\eta}_{t4}$, respectivamente, quase não apresentam *lags* significativos, assim como as funções de autocorrelação parcial presentes nas Figuras 4.7(b), 4.7(c) e 4.7(d). Por esses motivos, os gráficos de dispersão das Figuras 4.8(b), 4.8(c) e 4.8(d) não aparentam seguir algum padrão.

De fato, o teste de Ljung-Box (LJUNG; BOX, 1978) aplicado à série $\widehat{\eta}_{t1}$ leva à rejeição da hipótese nula de ausência de autocorrelação para $lags = 1, \ldots, 10$. No entanto, a hipótese nula não é rejeitada para o primeiro *lag* ao se considerar $\widehat{\eta}_{t2}$, embora não se tenha rejeitado essa hipótese para $lags = 2, \ldots, 10$, sugerindo haver alguma dependência temporal (fraca) na série $\widehat{\eta}_{t2}$. Para $\widehat{\eta}_{t3}$ e, principalmente, $\widehat{\eta}_{t4}$, a um nível de 5% de significância, a hipótese de ausência de autocorrelação serial não é rejeitada para a maioria das defasagens.

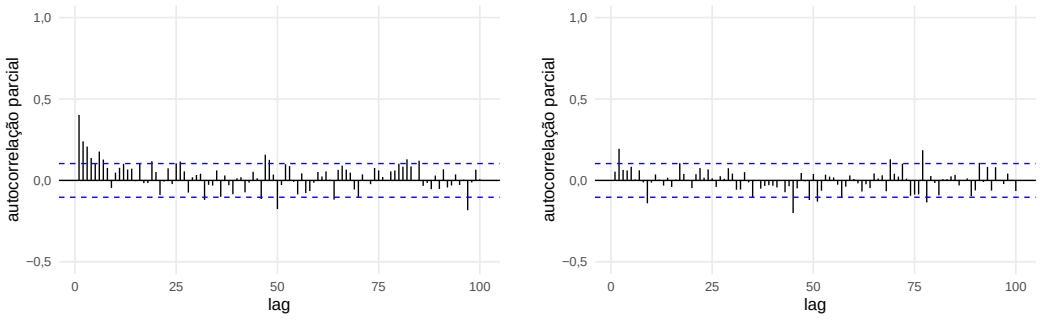


(a) $\widehat{\eta}_{t1}$ (b) $\widehat{\eta}_{t2}$

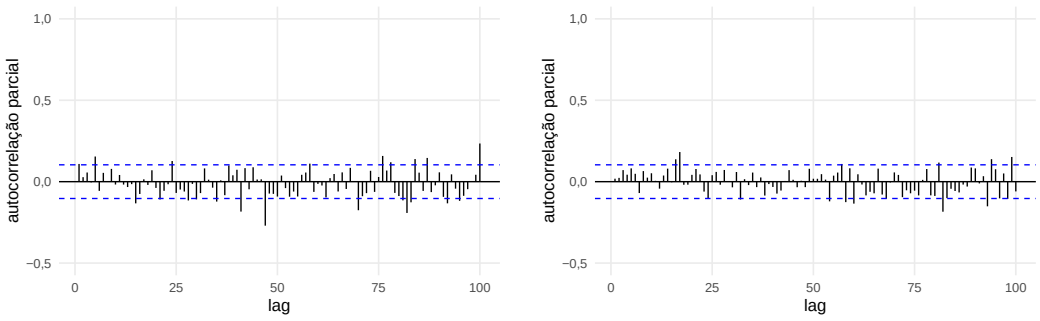


(c) $\widehat{\eta}_{t3}$ (d) $\widehat{\eta}_{t4}$

Figura 4.6: Funções de autocorrelação das séries $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$.

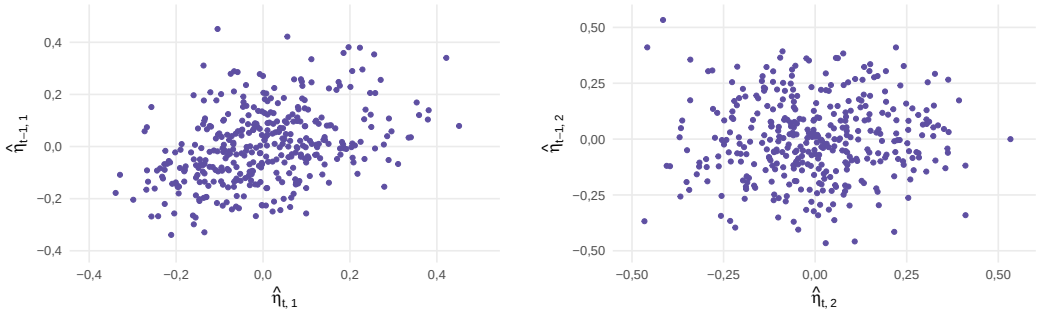


(a) $\widehat{\eta}_{t1}$ (b) $\widehat{\eta}_{t2}$

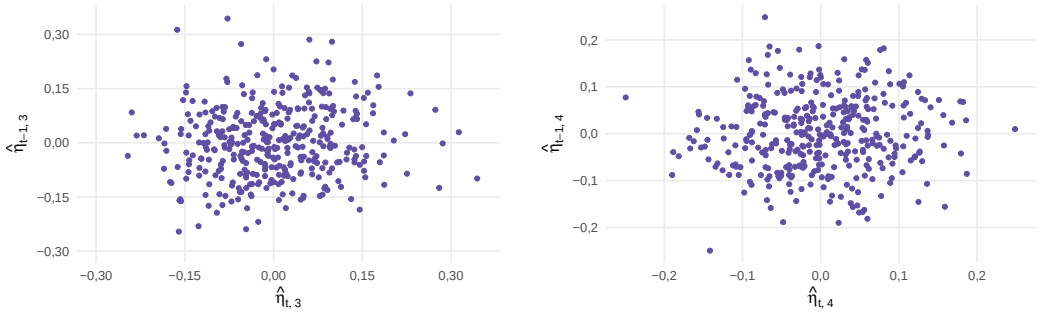


(c) $\widehat{\eta}_{t3}$ (d) $\widehat{\eta}_{t4}$

Figura 4.7: Funções de autocorrelação parcial das séries $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$.



(a) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t-1,1}$ (b) $\widehat{\eta}_{t2}$ vs $\widehat{\eta}_{t-1,2}$



(c) $\widehat{\eta}_{t3}$ vs $\widehat{\eta}_{t-1,3}$ (d) $\widehat{\eta}_{t4}$ vs $\widehat{\eta}_{t-1,4}$

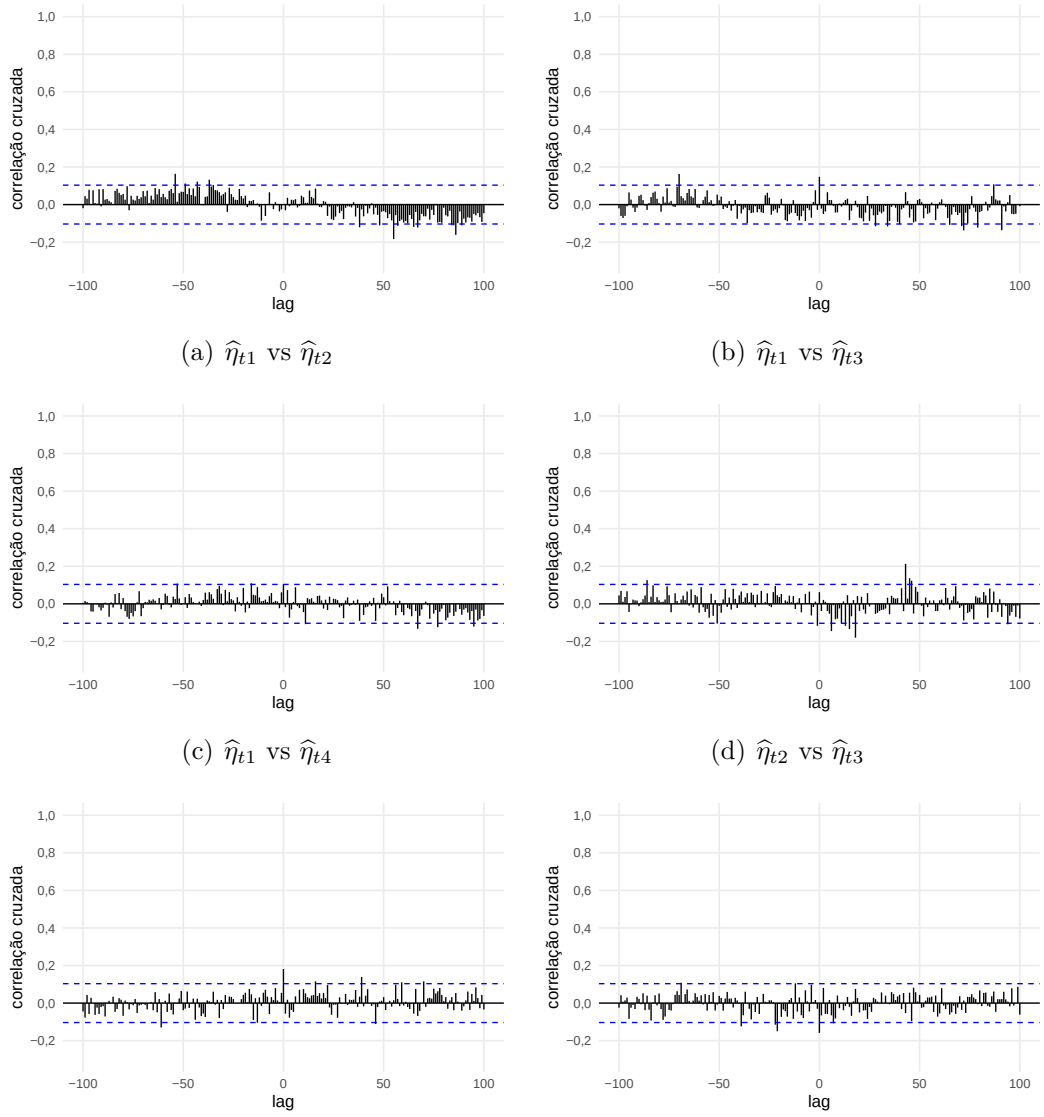
Figura 4.8: Gráficos de dispersão entre $\widehat{\eta}_{tj}$ e $\widehat{\eta}_{t-1,j}$, $j = 1, \ldots, 4$.

Para entender a dependência entre as séries temporais $\widehat{\eta}_{tj}$, é comum expressar essa relação através das funções de correlação cruzada, que podem ser vistas na Figura 4.9. É possível notar nesses gráficos que parece haver pouca dependência entre as séries consideradas, sendo destacável algumas correlações significativas entre $\widehat{\eta}_{t2}$ e $\widehat{\eta}_{t-1,3}$ na Figura 4.9(d) (*lags* positivos). Devido a essa característica de pouca associação entre as séries, os gráficos de dispersão de cada uma delas contra as defasagens das demais não apresentam qualquer tendência, conforme a Figura 4.10.

Modelos *VAR – Vector Autoregressions* – aplicados a estas séries, considerando variações tanto da dimensão d_0 quanto da ordem do modelo, não conseguiram gerar previsões com boa qualidade. Assim sendo, incentiva-se a inclusão de covariáveis exógenas no processo de modelagem e, posteriormente, previsão. Nesse sentido, o modelo *VARX – Vector Autoregressions with Exogenous Variables* – mostra-se uma alternativa relevante. A próxima seção descreve a forma pela qual as séries temporais exógenas foram selecionadas, além de detalhar o procedimento de determinação do melhor modelo que foi baseado em um ponto de vista preditivo.

4.4 Previsão

Conforme apontado ao longo do texto, o objetivo maior deste trabalho é fazer previsão dos mapas de intensidade do FC Barcelona. Assim, 50 previsões um passo



(a) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t2}$ (b) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t3}$

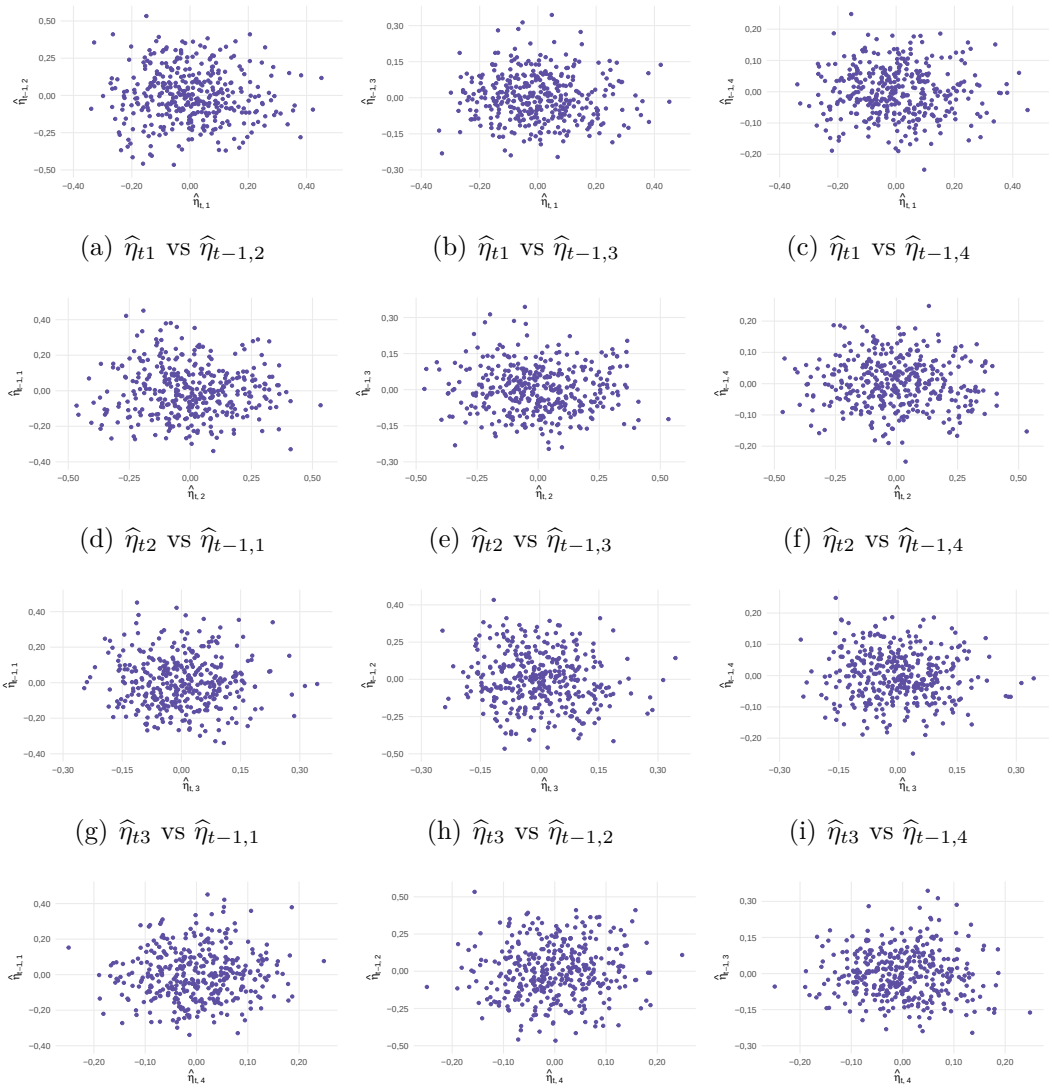
(c) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t4}$ (d) $\widehat{\eta}_{t2}$ vs $\widehat{\eta}_{t3}$

(e) $\widehat{\eta}_{t2}$ vs $\widehat{\eta}_{t4}$ (f) $\widehat{\eta}_{t3}$ vs $\widehat{\eta}_{t4}$

Figura 4.9: Funções de correlação cruzada entre as séries temporais $\widehat{\eta}_{tj}$, $j = 1, \ldots, 4$.

à frente foram consideradas. Para tanto, uma abordagem de janela móvel (*rolling window*) foi aplicada à amostra de $n = 374$ mapas de intensidade, a qual, portanto, foi separada em 50 subamostras tais que a i-ésima subamostra, com $i \in \{1, \ldots, 50\}$, é composta pelas observações $i, i+1, \ldots, T$, em que $T = T(i) = 324+i-1$. Ou seja, a cada passo do procedimento de previsão, a observação para a qual se fez previsão no passo anterior é incorporada à modelagem e a mais antiga é descartada. Então, a previsão um passo à frente para a série temporal de mapas de intensidade é expressa por

$$\widehat{\lambda}_{T+1|T}(\boldsymbol{u}) = \widehat{\mu}_{|T}(\boldsymbol{u}) + \sum_{j=1}^{d_0} \widehat{\eta}_{T+1,j|T}\widehat{\psi}_{j|T}(\boldsymbol{u}), \qquad \boldsymbol{u} \in S,$$



(a) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t-1,2}$ (b) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t-1,3}$ (c) $\widehat{\eta}_{t1}$ vs $\widehat{\eta}_{t-1,4}$

(d) $\widehat{\eta}_{t2}$ vs $\widehat{\eta}_{t-1,1}$ (e) $\widehat{\eta}_{t2}$ vs $\widehat{\eta}_{t-1,3}$ (f) $\widehat{\eta}_{t2}$ vs $\widehat{\eta}_{t-1,4}$

(g) $\widehat{\eta}_{t3}$ vs $\widehat{\eta}_{t-1,1}$ (h) $\widehat{\eta}_{t3}$ vs $\widehat{\eta}_{t-1,2}$ (i) $\widehat{\eta}_{t3}$ vs $\widehat{\eta}_{t-1,4}$

(j) $\widehat{\eta}_{t4}$ vs $\widehat{\eta}_{t-1,1}$ (k) $\widehat{\eta}_{t4}$ vs $\widehat{\eta}_{t-1,2}$ (l) $\widehat{\eta}_{t4}$ vs $\widehat{\eta}_{t-1,3}$

Figura 4.10: Gráficos de dispersão entre $\widehat{\eta}_{tj}$ e $\widehat{\eta}_{t-1,i}$, $j \neq i$.

onde $\widehat{\eta}_{T+1,j|T}$ denota a previsão um passo à frente para $\widehat{\eta}_{T+1,j}$, cuja obtenção é feita a partir de um modelo $VARX(q,1)$ ajustado à série temporal $\widehat{\boldsymbol{\eta}}_t$, $t = i, \ldots, T$. Do mesmo modo que no algoritmo apresentado na Seção 3.4, $\widehat{\mu}_{|T}$ e $\widehat{\psi}_{|T}$ representam a média e as autofunções calculadas a partir das observações até o tempo T, respectivamente.

Quanto aos valores de d_0 e da ordem q do modelo $VARX$, a determinação foi realizada de acordo com a combinação que minimizasse o EQMI de previsão um passo à frente,

$$\frac{1}{50}\sum_{i=1}^{50}\Big\{\int_S \Big(\widehat{\lambda}_{T+1}(\boldsymbol{u}) - \widehat{\lambda}_{T+1|T}(\boldsymbol{u})\Big)^2 \, \mathrm{d}\boldsymbol{u}\Big\}, \qquad T = 324 + i - 1. \tag{4.1}$$

Com base nos resultados discutidos na seção anterior, covariáveis exógenas que

pudessem estar relacionadas com a forma que a equipe ocupa o campo de jogo foram consideradas, resultando em um total de 43 preditores candidatos. Tais covariáveis representam informações que podem ser obtidas até momentos anteriores ao início da respectiva partida, como, por exemplo, se jogadores importantes jogarão, a formação do FC Barcelona, a formação do adversário, o mando de campo, o resultado da partida anterior, o treinador do FC Barcelona, etc. Dada a grande quantidade de parâmetros induzidos pela presença dessas covariáveis, é interessante que se utilize algum método de seleção de variáveis, de modo a manter no modelo apenas aquelas que sejam relevantes para explicar a série temporal vetorial $\widehat{\boldsymbol{\eta}}_t$.

Métodos de seleção de variáveis são largamente empregados na prática, sobretudo no contexto de observações independentes, como, por exemplo, os métodos Lasso – *Least Absolute Shrinkage and Selection Operator* – (TIBSHIRANI, 1996) e *Elastic Net* (ZOU; HASTIE, 2005). No cenário de observações dependentes no tempo, o método *VARX-L* proposto por Nicholson, Matteson e Bien (2017b) introduz penalizações, inclusive estruturadas, aos modelos *VAR* e *VARX*. Neste trabalho, considerou-se uma penalização que consiste em uma adaptação da regularização Lasso padrão (ou seja, uma restrição não-estruturada). Ademais, semelhantemente ao procedimento de previsão fora da amostra, a determinação dos parâmetros de penalização é feita via validação cruzada *rolling window*, cujo mecanismo é o mais adequado para séries temporais, uma vez que preserva a ordenação temporal dos dados. Em `R`, esses métodos encontram-se implementados no pacote *BigVAR* (NICHOLSON; MATTESON; BIEN, 2017a). O detalhamento e descrição da modelagem *VARX* e da metodologia *VARX-L* são feitos no Apêndice B.

Foram consideradas combinações com $d_0 = 2, \ldots, 6$ e $q = 1, \ldots, 4$. Dessa maneira, dada cada combinação e a primeira subamostra (observações $t = 1, \ldots, 324$), selecionou-se as covariáveis exógenas e as defasagens importantes para cada equação a partir do método *VARX-L*, obtendo-se, assim, uma matriz de coeficientes esparsa. Na sequência, um modelo $VARX(q, 1)$ restrito foi estimado via Mínimos Quadrados Ordinários (MQO), com matriz de restrição fornecida pela modelagem com penalização. Nesse ponto, é importante citar que, embora tenha-se usado MQO, a reestimação do modelo com a presença de restrições é assintoticamente mais eficiente quando realizada através de Mínimos Quadrados Generalizados (NICHOLSON; MATTESON; BIEN, 2017a). Geradas as 50 previsões por modelo ajustado, foi avaliado o EQMI definido na Equação (4.1). Os resultados estão presentes na Tabela 4.3.

Analisando-se a Tabela 4.3, pode-se perceber que a combinação $d_0 = 2$ e $q = 2$ é aquela que minimiza o EQMI de previsão um passo à frente dentre todas as combinações avaliadas. Esse resultado vai ao encontro das características presentes na Figura 4.3, na qual se viu que os dois primeiros autovalores da matriz $\boldsymbol{K}^*$ eram

Tabela 4.3: EQMI de previsão um passo à frente dadas combinações de d_0 e q.

	q			
d_0	1	2	3	4
2	0,11001	**0,10972**	0,11002	0,10977
3	0,11131	0,11177	0,11172	0,11147
4	0,11164	0,11313	0,11378	0,11392
5	0,11249	0,11397	0,11786	0,11994
6	0,11275	0,11387	0,11456	0,12244

muito maiores que os demais. Portanto, o modelo final escolhido para fazer previsão dos mapas de intensidade de partidas futuras do FC Barcelona é um modelo com a série temporal bivariada $\widehat{\boldsymbol{\eta}}_t = (\widehat{\eta}_{t1}, \widehat{\eta}_{t2})'$, considerando duas defasagens e um conjunto de covariáveis exógenas selecionadas a partir do método *VARX-L*, que podem ser consultadas nas Tabelas A.2 e A.3 da Seção A.2 do Apêndice A. Ou seja, a previsão um passo à frente $\widehat{\boldsymbol{\eta}}_{T+1|T}$ é obtida através da equação de previsão do modelo $VARX(2,1)$ dada por

$$\widehat{\boldsymbol{\eta}}_{T+1|T} = \boldsymbol{\nu} + \sum_{\ell=1}^{2} \boldsymbol{\phi}_{T+1-\ell}\widehat{\boldsymbol{\eta}}_{T+1-\ell} + \boldsymbol{\beta}_0\boldsymbol{x}_{T+1}^{(0)} + \boldsymbol{\beta}_1\boldsymbol{x}_{T}^{(1)},$$

onde $\boldsymbol{x}_t^{(0)}$ é um conjunto de covariáveis selecionadas e avaliadas no tempo t e $\boldsymbol{x}_{t-1}^{(1)}$ são covariáveis exógenas selecionadas que são observadas no tempo $t-1$ Mais detalhes podem ser vistos nos Apêndices A e B.

As previsões obtidas pelo modelo final foram muito parecidas com a função média $\widehat{\mu}$. Tal constatação sugere que o padrão de jogo do FC Barcelona é estável ao longo do tempo, com algumas pequenas variações idiossincráticas. Dessa forma, ao se avaliar a diferença entre os mapas de intensidade predito, $\widehat{\lambda}_{T+1|T}$, e observado na partida T, $\widehat{\lambda}_T$ (isto é, $\widehat{\lambda}_{T+1|T}(\boldsymbol{u}) - \widehat{\lambda}_T(\boldsymbol{u})$, $\boldsymbol{u} \in S$) pode-se ter uma boa previsão da variação de posicionamento entre partidas subsequentes (como, por exemplo, se a equipe será mais ofensiva/defensiva no jogo seguinte, se ocupará mais as laterais do campo, etc.), que é dada por $\widehat{\lambda}_{T+1} - \widehat{\lambda}_T$.

A Figura 4.11 mostra o erro quadrático integrado associado a cada previsão, ou seja, reporta os valores $\int_S (\widehat{\lambda}_{T+1}(\boldsymbol{u}) - \widehat{\lambda}_{T+1|T}(\boldsymbol{u}))^2 \, \mathrm{d}\boldsymbol{u}$, $\forall\ T = 324, \ldots, 373$. O valor mínimo para essa medida foi obtido para a previsão $\widehat{\lambda}_{358|357}$, referente à partida válida pela 20ª rodada da temporada 2018/2019 contra o Club Deportivo Leganés.

Na Figura 4.12(a) pode ser visualizada a variação predita $\widehat{\lambda}_{358|357} - \widehat{\lambda}_{357}$. Em comparação à variação observada $\widehat{\lambda}_{358} - \widehat{\lambda}_{357}$, dada na Figura 4.12(b), percebe-se que foi possível prever que o FC Barcelona seria mais ofensivo na partida $T+1 = 358$,

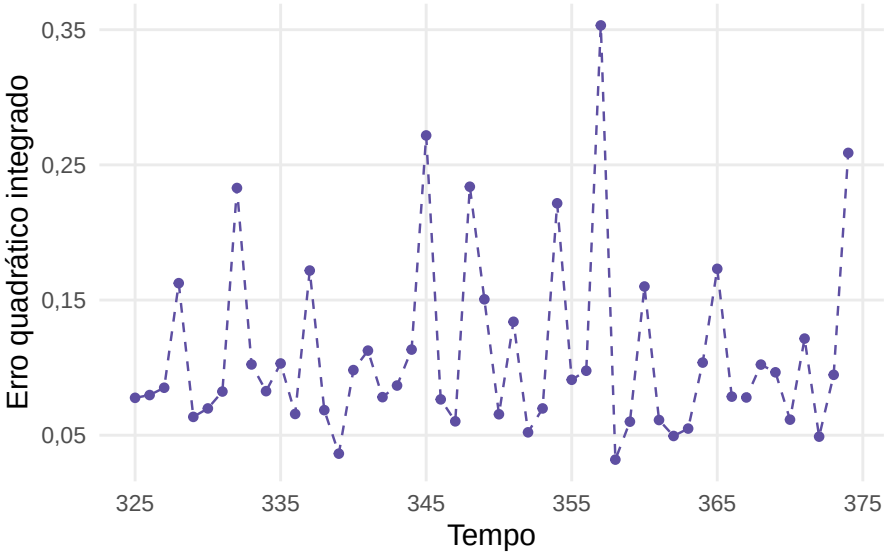
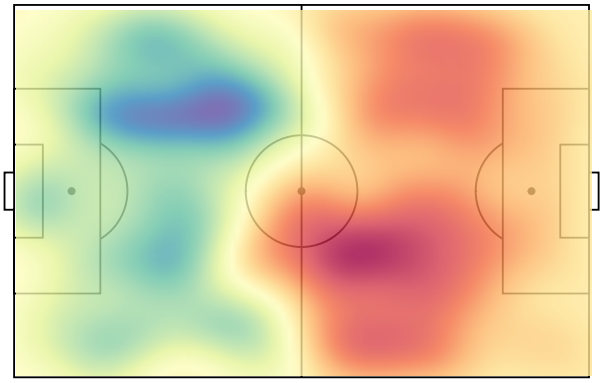
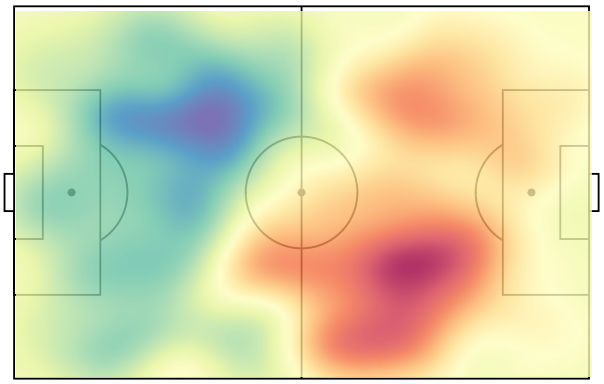


Figura 4.11: Erro quadrático integrado de cada previsão feita pelo modelo final.



(a) $\widehat{\lambda}_{358|357} - \widehat{\lambda}_{357}$



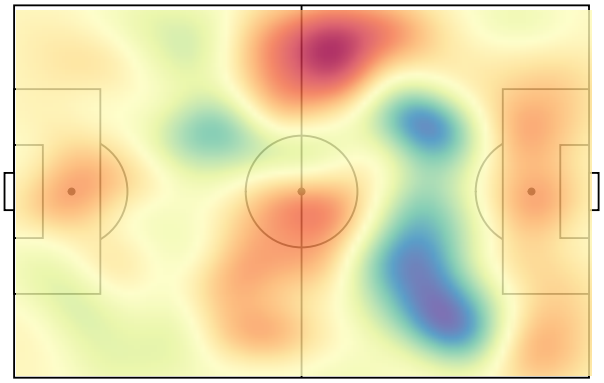
(b) $\widehat{\lambda}_{358} - \widehat{\lambda}_{357}$

Figura 4.12: Melhor previsão e variação observada. $T + 1 = 358$.

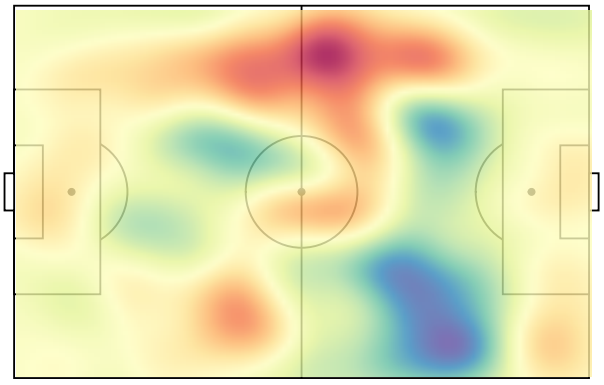
uma vez que as tonalidades mais avermelhadas encontram-se na parte ofensiva do

campo em ambos os mapas. Além disso, a previsão consegue capturar, também, que a equipe ocuparia com mais intensidade o lado direito do campo, embora tenha-se previsto uma taxa de ocupação levemente mais intensa no campo de ataque esquerdo, se comparada à variação observada.

Segundo o erro quadrático integrado, tem-se que a segunda melhor previsão foi para o mapa de intensidade da partida $T + 1 = 339$, que corresponde à 34ª rodada da temporada 2017/2018, cujo oponente foi o Villarreal Club de Fútbol. A Figura 4.13(a) contém a variação predita $\widehat{\lambda}_{339|338} - \widehat{\lambda}_{338}$. Nota-se que foi prevista uma intensidade maior de ocupação na lateral esquerda do campo em relação à partida $T = 338$ e uma variação negativa (cores azuis) no meio de campo ofensivo próximo à grande área do adversário. Tais características realmente estão presentes na variação observada $\widehat{\lambda}_{339} - \widehat{\lambda}_{338}$ expressa na Figura 4.13(b).



(a) $\widehat{\lambda}_{339|338} - \widehat{\lambda}_{338}$

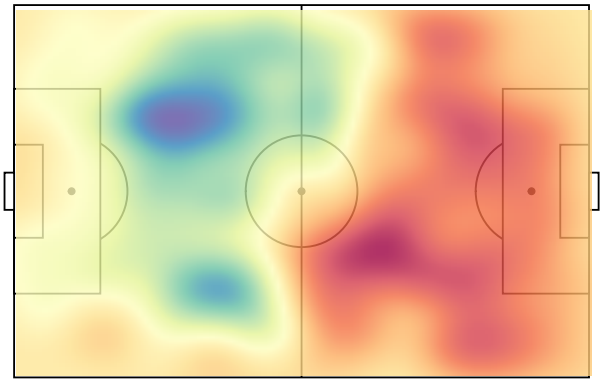


(b) $\widehat{\lambda}_{339} - \widehat{\lambda}_{338}$

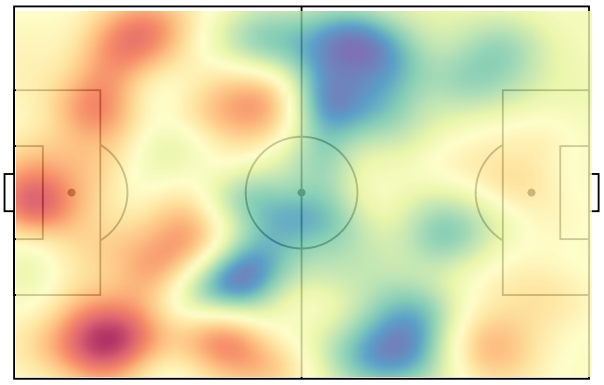
Figura 4.13: Segunda melhor previsão e variação observada. $T + 1 = 339$.

Além disso, na Figura 4.11 pode-se considerar que, de forma preponderante, a pior das previsões foi aquela relativa à partida $T+1 = 357$ (19ª rodada da temporada 2018/2019 contra a Sociedad Deportiva Eibar), cuja variação predita é exibida na

Figura 4.14(a) e a respectiva variação observada, na Figura 4.14(b). Analisando-se ambos os mapas, vê-se que foi prevista uma variação positiva mais intensa no campo ofensivo, enquanto que no campo de defesa esperava-se uma variação negativa em relação à partida $T = 356$. Em contrapartida, a variação observada apresentou um comportamento oposto: variação positiva no campo de defesa, e negativa no campo de ataque. Ou seja, previu-se uma postura mais ofensiva do FC Barcelona em comparação a seu jogo antecedente, ao passo que se observou uma postura relativa mais defensiva.



(a) $\widehat{\lambda}_{357|356} - \widehat{\lambda}_{356}$



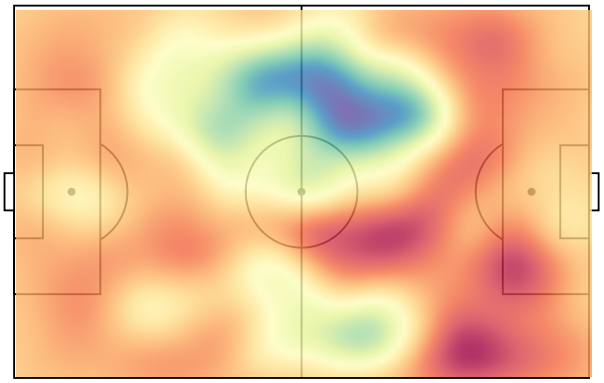
(b) $\widehat{\lambda}_{357} - \widehat{\lambda}_{356}$

Figura 4.14: Pior previsão e variação observada. $T + 1 = 357$.

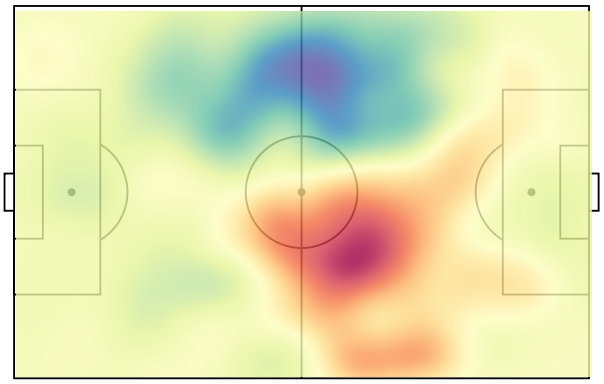
Ainda, observando-se os resultados, o segundo maior erro quadrático integrado está associado à previsão $\widehat{\lambda}_{345}$, relativa à partida diante do Getafe Club de Fútbol pela 5ª rodada da temporada 2018/2019. A Figura 4.15(b) retrata a variação observada, a partir da qual nota-se que a equipe de Barcelona concentrou mais predominantemente sua posse de bola no lado direito do campo, demonstrando, outrossim, um comportamento levemente mais ofensivo, comparando-se com o mapa de intensidade da partida $T = 344$. Apesar de a variação prevista, expressa na Fi-

gura 4.15(a), conseguir captar que o FC Barcelona habitaria com menos intensidade a zona à esquerda do campo em relação ao jogo anterior, não foi possível prever as demais características, tendo sido indicado que haveria uma alta variação positiva em todas as áreas restantes do campo, fato que não se observou.

As variações preditas e as respectivas variações observadas associadas para as demais partidas estão presentes no Apêndice D.



(a) $\widehat{\lambda}_{345|344} - \widehat{\lambda}_{344}$



(b) $\widehat{\lambda}_{345} - \widehat{\lambda}_{344}$

Figura 4.15: Segunda pior previsão e variação observada. $T + 1 = 345$.

4.5 Previsão de ofensividade

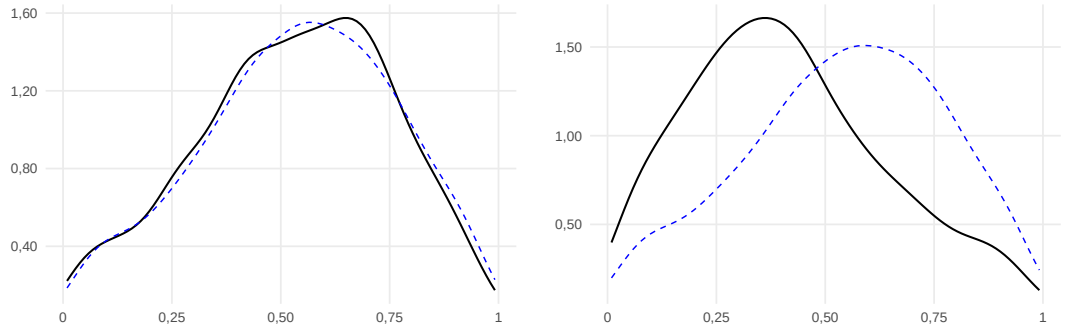
Ao se analisar os mapas de intensidade preditos $\widehat{\lambda}_{T+1|T}$, $T = 324, \ldots, 373$, pode-se perceber uma melhor previsão de características ofensivas ou defensivas do FC Barcelona. Dessa forma, uma possibilidade é investigar curvas de intensidade na direção da coordenada horizontal do campo. Assim, considere a curva de intensidade observada $\widehat{y}_t$, que pode ser recuperada a partir do mapa de intensidade observado $\widehat{\lambda}_t$ através da relação

$$\widehat{y}_t(u_1) := \int_0^1 \widehat{\lambda}_t(u_1, u_2)\, \mathrm{d}u_2, \qquad u_1 \in [0, 1].$$

Dada uma subamostra observada até o tempo T, como definido na Seção 4.4, a previsão um passo à frente para a curva de intensidade $\widehat{y}_{T+1}$, denotada por $\widehat{y}_{T+1|T}$, é gerada via

$$\widehat{y}_{T+1|T}(u_1) := \int_0^1 \widehat{\lambda}_{T+1|T}(u_1, u_2)\, \mathrm{d}u_2, \qquad u_1 \in [0, 1].$$

Calculando-se as curvas de intensidade observadas e preditas para as 50 partidas do FC Barcelona separadas para o procedimento de previsão anteriormente, tem-se que a melhor e pior previsões referem-se, respectivamente, às partidas $T+1 = 358$ e $T + 1 = 357$, de modo semelhante aos resultados obtidos para os mapas de intensidade. A Figura 4.16(a) (resp. Figura 4.16(b)) ilustra a melhor (resp. pior) curva de intensidade predita (linha azul pontilhada), juntamente com a curva observada (linha sólida). As interpretações são similares às realizadas para os mapas de intensidade das Figuras 4.12(a) e 4.14(a), considerando apenas a coordenada horizontal do campo, ou seja, tendências ofensivas/defensivas.



(a) $T + 1 = 358$ (b) $T + 1 = 357$

Figura 4.16: Melhor e pior curvas de intensidade preditas e respectivas observações.

Na Figura 4.17 estão contidos todos os 50 erros de previsão $\widehat{y}_{T+1|T}(\cdot) - \widehat{y}_{T+1}(\cdot)$, $T = 324, \ldots, 373$. A linha em rosa representa o erro de previsão médio, o qual é próximo de zero para toda a extensão da coordenada horizontal do campo, com apenas algum leve viés entre o intervalo $[0,5; 0,8]$. Tal percepção sugere que, em média, consegue-se prever os padrões de ofensividade do FC Barcelona.

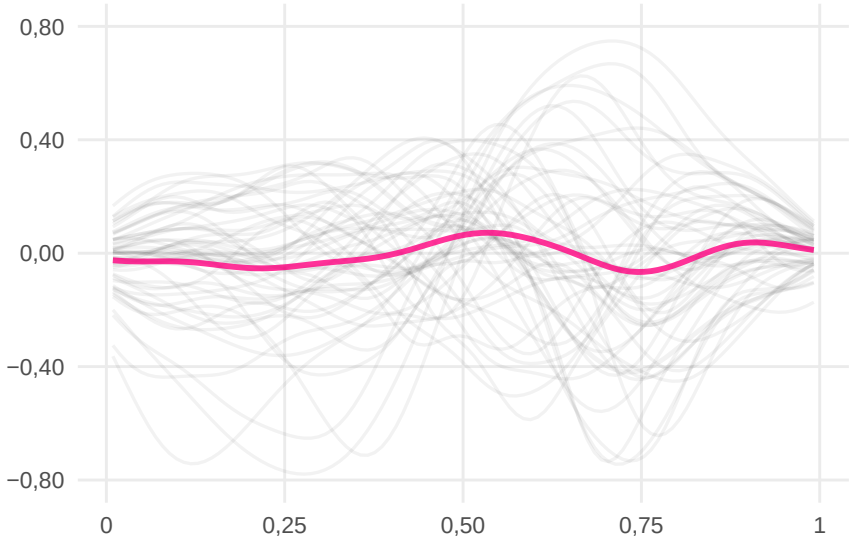


Figura 4.17: Erros de previsão $\widehat{y}_{T+1|T} - \widehat{y}_{T+1}, T = 324, \ldots, 373$, e erro médio (curva rosa).

5 Conclusão

Este trabalho teve como objetivo propor uma aplicação de modelos de Séries Temporais Funcionais a dados oriundos do futebol, utilizando-se da metodologia de Bathia, Yao e Ziegelmann (2010). Para tanto, considerou-se fazer modelagem e, sobretudo, previsão de mapas de intensidade de equipes de futebol. Destaca-se que a principal contribuição prática desta pesquisa está no fato de que a previsão de mapas de intensidade de jogos futuros de times adversários pode ser muito útil para uma equipe de futebol do ponto de vista tático, uma vez que a estratégia adotada pode ser desenvolvida com base, entre outros aspectos, nessas previsões. Além disso, dada a escassez de trabalhos na área das Ciências Esportivas que fazem uso de metodologias de Séries Temporais Funcionais, acredita-se que este trabalho traga uma contribuição para a área devido a sua proposta inovadora.

A aplicação da metodologia de Bathia, Yao e Ziegelmann (2010) se deu a partir de um conjunto de mapas de intensidade do clube espanhol FC Barcelona, que foram gerados por intermédio de dados coletados e disponibilizados ao público pela empresa *StatsBomb* referentes a partidas disputadas pela equipe no Campeonato Espanhol entre as temporadas de 2008/2009 e 2018/2019 em que o jogador argentino Lionel Messi atuou, correspondendo a 374 jogos. Além de ter como hipótese a existência de uma dinâmica entre partidas subsequentes, fato que se confirmou mais notoriamente apenas para uma das séries temporais de coeficientes observada, também foram incorporadas covariáveis exógenas ao modelo, e um método de seleção de variáveis foi empregado no contexto da modelagem *VARX*.

Os resultados foram interpretados considerando a variação entre mapas de intensidade de partidas consecutivas, abordagem que permite fazer previsões da forma como a equipe ocupará as extensões do campo em relação ao jogo anterior. No geral, a partir das previsões resultantes, é possível perceber que o modelo proposto consegue capturar de forma mais acurada posturas ofensivas/defensivas do FC Barcelona. Isto é, os mapas de intensidade preditos absorvem melhor variações na coordenada horizontal do campo em comparação àquelas observadas na coordenada vertical. Desse modo, curvas de intensidade que possam indicar tendências

ofensivas/defensivas também foram avaliadas, a partir das quais pôde-se perceber mais claramente tais características. Ao se considerar um ambiente naturalmente competitivo como aquele em que o futebol está inserido, no qual qualquer vantagem mínima sobre o adversário tem um peso muito relevante nos enfrentamentos diretos, acredita-se que essa informação parcial prévia do comportamento coletivo do oponente possa ser importante. Todavia, deve-se ponderar que a variabilidade intrapartida é um fator destacável no futebol e que não pode ser modelado de antemão pela metodologia tratada nesta pesquisa, motivo que pode justificar, ao menos parcialmente, algumas das limitações encontradas na aplicação, conjuntamente com o fato de que a imprevisibilidade tem um peso considerável nesse esporte, inclusive sendo um tema discutido em diversos estudos.

Em trabalhos futuros, espera-se aprimorar a aplicação, considerando dados mais informativos, como, por exemplo, os emergentes dados de rastreamento de bola e jogadores, ou ao menos dados completos de todas as partidas de uma equipe em uma temporada em vez de somente aquelas em que um jogador em específico esteve em campo. Ademais, com base nos resultados obtidos no presente trabalho, pode-se considerar uma análise mais profunda sob o contexto de curvas de intensidade, que indiquem apenas tendências ofensivas e defensivas dos times. Além disso, ressalta-se que os resultados encontrados estão restritos apenas ao FC Barcelona e, portanto, deseja-se ampliar o estudo para dados de outras equipes, que preferencialmente apresentem comportamentos de jogo distintos; assim, será possível avaliar se as previsões têm uma boa performance independentemente do estilo praticado pela equipe. Também é pauta de novas investigações utilizar outras abordagens na modelagem das séries temporais multivariadas, e estudar a inclusão de novas covariáveis exógenas, como, por exemplo, os próprios mapas de intensidade do adversário. No entanto, para expandir o alcance da metodologia, os clubes e empresas que coletam e armazenam os dados gerados nas partidas precisam torná-los, de alguma forma, mais acessíveis a pesquisadores e interessados na disseminação da Estatística no futebol, talvez por intermédio de parcerias entre universidades ou grupos de pesquisa e as equipes.

Referências Bibliográficas

ANDERSON, C.; SALLY, D. *Os números do jogo: por que tudo o que você sabe sobre futebol está errado*. São Paulo: Paralela, 2013.

ANTONIADIS, A.; SAPATINAS, T. Wavelet methods for continuous-time prediction using hilbert-valued autoregressive processes. *Journal of Multivariate Analysis*, Elsevier, v. 87, n. 1, p. 133–158, 2003.

AUE, A.; NORINHO, D. D.; HÖRMANN, S. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, Taylor & Francis, v. 110, n. 509, p. 378–392, 2015.

BATHIA, N.; YAO, Q.; ZIEGELMANN, F. A. Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 38, n. 6, p. 3352–3386, 2010.

BESSE, P. C.; CARDOT, H.; STEPHENSON, D. B. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, Wiley Online Library, v. 27, n. 4, p. 673–687, 2000.

BIALKOWSKI, A.; LUCEY, P.; CARR, P.; YUE, Y.; SRIDHARAN, S.; MATTHEWS, I. Large-scale analysis of soccer matches using spatiotemporal tracking data. In: IEEE. *2014 IEEE International Conference on Data Mining*. [S.l.], 2014. p. 725–730.

BIALKOWSKI, A.; LUCEY, P.; CARR, P.; YUE, Y.; MATTHEWS, I. Win at home and draw away: Automatic formation analysis highlighting the differences in home and away team behaviors. In: CITESEER. *Proceedings of 8th annual MIT sloan sports analytics conference*. [S.l.], 2014. p. 1–7.

BOSQ, D. *Linear Processes in Function Spaces: Theory and Applications*. 1. ed. [S.l.]: Springer-Verlag New York, 2000. v. 149. (Lecture Notes in Statistics, v. 149).

DONA, G.; PREATONI, E.; COBELLI, C.; RODANO, R.; HARRISON, A. J. Application of functional principal component analysis in race walking: an emerging methodology. *Sports Biomechanics*, Taylor & Francis, v. 8, n. 4, p. 284–301, 2009.

FERRATY, F.; VIEU, P. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, Springer Nature BV, v. 17, n. 4, p. 545–564, 2002.

FERRATY, F.; VIEU, P. *Nonparametric Functional Data Analysis: Theory and Practice*. [S.l.]: Springer-Verlag New York, 2006. (Springer Series in Statistics).

FIFA. Fifa Big Count 2006: 270 million people active in football. *FIFA*, 2007. Disponível em: <https://resources.fifa.com/image/upload/big-count-stats-package-520046.pdf?cloudid=mzid0qmguixkcmruvema>. Acesso em: 24 ago. 2020.

FIFA. More than half the world watched record-breaking 2018 World Cup. *FIFA*, 2018. Disponível em: <https://www.fifa.com/worldcup/news/more-than-half-the-world-watched-record-breaking-2018-world-cup#>. Acesso em: 17 jun. 2020.

FULLER, W. A. *Introduction to statistical time series*. [S.l.]: John Wiley & Sons, 2009. v. 428.

GRAFIETTI, C. Quanto vale um clube de futebol? Os cálculos e as idiossincrasias do negócio da bola. *InfoMoney*, 2020. Disponível em: <https://www.infomoney.com.br/colunistas/cesar-grafietti/quanto-vale-um-clube-de-futebol-os-calculos-e-as-idiossincrasias-do-negocio-da-bola/>. Acesso em: 31 ago. 2020.

GUDMUNDSSON, J.; HORTON, M. Spatio-temporal analysis of team sports. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 50, n. 2, p. 1–34, 2017.

HALL, P.; MÜLLER, H.-G.; WANG, J.-L. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, JSTOR, v. 34, n. 3, p. 1493–1517, 2006.

HARRISON, A.; RYAN, W.; HAYES, K. Functional data analysis of joint coordination in the development of vertical jump performance. *Sports Biomechanics*, Taylor & Francis, v. 6, n. 2, p. 199–214, 2007.

HARRISON, A. J. Applications of functional data analysis in sport biomechanics. In: *ISBS-Conference Proceedings Archive*. [S.l.: s.n.], 2014.

HORTA, E.; ZIEGELMANN, F. Identifying the spectral representation of Hilbertian time series. *Statistics & Probability Letters*, Elsevier, v. 118, p. 45–49, 2016.

HORTA, E.; ZIEGELMANN, F. Conjugate processes: Theory and application to risk forecasting. *Stochastic Processes and their Applications*, Elsevier, v. 128, n. 3, p. 727–755, 2018.

KARHUNEN, K. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, v. 34, 1946.

LJUNG, G. M.; BOX, G. E. On a measure of lack of fit in time series models. *Biometrika*, Oxford University Press, v. 65, n. 2, p. 297–303, 1978.

LOÈVE, M. Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, v. 84, p. 159–162, 1946.

LUCEY, P.; OLIVER, D.; CARR, P.; ROTH, J.; MATTHEWS, I. Assessing team strategy using spatiotemporal data. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2013. p. 1366–1374.

MERCER, J. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, The Royal Society London, v. 209, n. 441-458, p. 415–446, 1909.

MINH, H. Q.; NIYOGI, P.; YAO, Y. Mercer's theorem, feature maps, and smoothing. In: SPRINGER. *Learning Theory*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. (Lecture Notes in Computer Science, v. 4005), p. 154–168.

NARAYANAN, S. Flexible multivariate processes for modelling football matches. *MathSport International*, 2019. Disponível em: <http://www2.stat-athens.aueb.gr/~jbn/conferences/MathSport_presentations/TRACK%20B/B6%20-%20Understanding%20the%20game%20using%20inplay%20analysis/Santhosh_Multivariate%20processes%20Football.pdf>. Acesso em: 01 jul. 2020.

NICHOLSON, W.; MATTESON, D.; BIEN, J. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094*, 2017.

NICHOLSON, W. B.; MATTESON, D. S.; BIEN, J. Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting*, Elsevier, v. 33, n. 3, p. 627–651, 2017.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>.

RAMSAY, J. O.; SILVERMAN, B. W. *Functional Data Analysis*. 2. ed. New York: Springer-Verlag New York, 2005. (Springer Series in Statistics).

RAMSAY, J. O.; SILVERMAN, B. W. *Applied functional data analysis: methods and case studies*. [S.l.]: Springer, 2007.

RICE, J. A. Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, JSTOR, v. 14, n. 3, p. 631–647, 2004.

RIVEIRA, C. Venda de jogadores de futebol movimentou US$ 7 bilhões em 2019. *Revista Exame*, 2020. Disponível em: <https://exame.com/negocios/venda-de-jogadores-de-futebol-movimentou-us-8-bilhoes-em-2019/>. Acesso em: 24 ago. 2020.

RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2021. Disponível em: <http://www.rstudio.com/>.

SCHULTZE, S. R.; WELLBROCK, C.-M. A weighted plus/minus metric for individual soccer player performance. *Journal of Sports Analytics*, IOS Press, v. 4, n. 2, p. 121–131, 2018.

SILVERMAN, B. W. *Density estimation for statistics and data analysis*. Londres: Chapman & Hall Ltd, 1986.

STATSBOMB. *StatsBomb Academy*. [S.l.], 2019. Disponível em: <https://statsbomb.com/academy/>. Acesso em: 10 abr. 2020.

STATSBOMB. *StatsBomb data*. [S.l.], 2019. Disponível em: <https://github.com/statsbomb>. Acesso em: 10 abr. 2020.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.

TSAY, R. S. *Multivariate time series analysis: with R and financial applications*. [S.l.]: John Wiley & Sons, 2013.

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S*. 4. ed. New York: Springer, 2002. (Statistics and Computing).

VINUÉ, G.; EPIFANIO, I. Archetypoid analysis for sports analytics. *Data Mining and Knowledge Discovery*, Springer, v. 31, n. 6, p. 1643–1677, 2017.

VINUÉ, G.; EPIFANIO, I. Forecasting basketball players' performance using sparse functional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 12, n. 6, p. 534–547, 2019.

WAKIM, A.; JIN, J. Functional data analysis of aging curves in sports. *arXiv preprint arXiv:1403.7548*, 2014.

WANG, J.-L.; CHIOU, J.-M.; MÜLLER, H.-G. Functional data analysis. *Annual Review of Statistics and Its Application*, Annual Reviews, v. 3, p. 257–295, 2016.

YAO, F.; MÜLLER, H.-G.; WANG, J.-L. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, JSTOR, v. 33, n. 6, p. 2873–2903, 2005.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, Wiley Online Library, v. 67, n. 2, p. 301–320, 2005.

APÊNDICE A

A.1 Lista de ações consideradas

A Tabela A.1 contém as ações selecionadas para a construção dos mapas de intensidade do FC Barcelona.

Tabela A.1: Ações filtradas na análise e suas respectivas descrições.

Ação	Descrição
Ball Receipt	Bola recebida a partir de um passe de um companheiro de equipe
Ball Recovery	Posse de bola recuperada pela equipe após roubá-la do adversário
Dispossessed	Perda da posse de bola para um jogador adversário
Dribble	Deixar um adversário para trás carregando a bola
Duel	Disputa (ganha) pela bola contra um jogador adversário
Foul Won	Tiro livre (direto ou indireto) obtido após infração cometida pelo adversário
Miscontrol	Posse de bola não totalmente controlada pela equipe
Pass	Passar (tocar) a bola para um companheiro de equipe
Referee Ball-Drop	Posse de bola ganha pela equipe após o árbitro da partida recolocar a bola em jogo após alguma parada
Shield	Proteção de bola, impedindo que o adversário fique com a posse
Shot	Chute a gol

A.2 Covariáveis selecionadas para o modelo final

A metodologia *VARX-L* seleciona tanto as covariáveis exógenas quanto as defasagens das séries temporais $\widehat{\eta}_{tj}$. Assim, a Tabela A.2 (resp. Tabela A.3) lista os componentes da equação para $\widehat{\eta}_{t1}$ (resp. $\widehat{\eta}_{t2}$) ajustada a partir do modelo *VARX* após aplicação da penalização via *VARX-L*.

- Equação para $\widehat{\eta}_{t1}$:

  Defasagens: $\widehat{\eta}_{t-1,1}, \widehat{\eta}_{t-1,2}, \widehat{\eta}_{t-2,1}$ e $\widehat{\eta}_{t-2,2}$.

Tabela A.2: Covariáveis exógenas selecionadas para a equação de $\widehat{\eta}_{t1}$

Covariável	Descrição
home	Partida disputada sob mando de campo do FC Barcelona
play.xavi	O jogador Xavi Hernández é escalado
play.iniesta	O jogador Andrés Iniesta é escalado
play.suarez	O jogador Luis Suárez é escalado
play.dani	O jogador Daniel Alves é escalado
play.puyol	O jogador Carles Puyol é escalado
play.pique	O jogador Gerard Piqué é escalado
play.busquets	O jogador Sergio Busquets é escalado
ftto.adv	A formação tática do adversário é diferente de 4-2-3-1 (4 defensores, 2 meio-campistas defensivos, 3 meio-campistas ofensivos e 1 atacante)
fft.bar	A formação tática do FC Barcelona é diferente de 4-3-3 (4 defensores, 3 meio-campistas e 3 atacantes)
messi.pos	O jogador Lionel Messi não jogou pelo lado direito do campo
guard	Josep Guardiola é o treinador do FC Barcelona
tito	Tito Vilanova é o treinador do FC Barcelona
luis	Luis Enrique é o treinador do FC Barcelona
emp	O FC Barcelona empatou a partida $t-1$
dif.day	Diferença de dias em relação à partida $t-1$
cinq	Número de disputas 50/50 do FC Barcelona na partida $t-1$
drib	Número de dribles do FC Barcelona na partida $t-1$
duel	Número de duelos do FC Barcelona na partida $t-1$
error	Número de erros do FC Barcelona na partida $t-1$
foul.com	Número de faltas cometidas pelo FC Barcelona na partida $t-1$
foul.won	Número de faltas ganhas pelo FC Barcelona na partida $t-1$
pass	Número de passes trocados pelo FC Barcelona na partida $t-1$
pressure	Número de pressões exercidas pelo FC Barcelona na partida $t-1$
shield	Número de proteções de bola feitas pelo FC Barcelona na partida $t-1$
shot	Número de chutes ao gol efetuados pelo FC Barcelona na partida $t-1$

- Equação para $\widehat{\eta}_{t2}$:

Defasagens: $\widehat{\eta}_{t-2,2}$.

Tabela A.3: Covariáveis exógenas selecionadas para a equação de $\widehat{\eta}_{t2}$

Covariável	Descrição
home	Partida disputada sob mando de campo do FC Barcelona
play.iniesta	O jogador Andrés Iniesta é escalado
play.suarez	O jogador Luis Suárez é escalado
play.dani	O jogador Daniel Alves é escalado
play.puyol	O jogador Carles Puyol é escalado
play.busquets	O jogador Sergio Busquets é escalado
play.fab	O jogador Cesc Fàbregas é escalado
messi.pos	O jogador Lionel Messi não jogou pelo lado direito do campo
der	O FC Barcelona perdeu a partida $t-1$
cinq	Número de disputas 50/50 do FC Barcelona na partida $t-1$
clearance	Número de bolas rebatidas pelo FC Barcelona na partida $t-1$
disp	Número de posses perdidas pelo FC Barcelona na partida $t-1$
drib.past	Número de dribles sofridos pelo FC Barcelona na partida $t-1$
error	Número de erros do FC Barcelona na partida $t-1$
gk	Número de participações do goleiro do FC Barcelona na partida $t-1$
shield	Número de proteções de bola feitas pelo FC Barcelona na partida $t-1$

APÊNDICE B

B.1 Modelo *VARX*

De acordo com as explanações de Tsay (2013) e Nicholson, Matteson e Bien (2017b), seja $\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{t,d_0})'$ uma série temporal multivariada de dimensão $d_0 \in \mathbb{N}$, e $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tr})'$, $r \in \mathbb{N}$, um vetor de covariáveis exógenas. Um modelo *VARX* (*Vector Autoregressions with Exogenous Variables*) de ordens q e b, denotado por VARX(q, b), é definido pela seguinte equação

$$\boldsymbol{\eta}_t = \boldsymbol{\nu} + \sum_{\ell=1}^{q} \boldsymbol{\phi}_\ell \boldsymbol{\eta}_{t-\ell} + \sum_{i=0}^{b} \boldsymbol{\beta}_i \boldsymbol{x}_{t-i} + \boldsymbol{e}_t, \tag{B.1}$$

em que $\boldsymbol{\nu} := (\nu_1, \ldots, \nu_{d_0})'$ é um vetor de constantes, $\boldsymbol{\phi}_\ell$ é uma matriz $d_0 \times d_0$ de coeficientes da ℓ-ésima defasagem de $\boldsymbol{\eta}_t$, $\boldsymbol{\beta}_i$ é uma matriz $d_0 \times r$ de coeficientes associados à i-ésima defasagem das covariáveis exógenas $\boldsymbol{x}_t$, e $\boldsymbol{e}_t$ é ruído branco, com $\mathbb{E}(\boldsymbol{e}_t | \boldsymbol{\eta}_{t-\ell}, \boldsymbol{x}_{t-i-1}) = 0$, para $\ell = 1, \ldots, q$ e $i = 0, \ldots, b+1$. Ou seja, o valor de uma específica série η_{tj}, $j = 1, \ldots, d_0$, depende de suas próprias defasagens, além das defasagens das demais séries envolvidas e das covariáveis (defasadas ou não). Além do mais, não é feita nenhuma especificação sobre a dinâmica do vetor $\boldsymbol{x}_t$.

No contexto de um modelo VARX(q, b), os parâmetros $\boldsymbol{\nu}$, $\boldsymbol{\phi} := (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q)$ e $\boldsymbol{\beta} := (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_b)$ são estimados ao se resolver o problema de minimização

$$\underset{\boldsymbol{\nu}, \boldsymbol{\phi}, \boldsymbol{\beta}}{\arg\min} \sum_{t=1}^{T} \left\| \boldsymbol{\eta}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{q} \boldsymbol{\phi}_\ell \boldsymbol{\eta}_{t-\ell} - \sum_{i=0}^{b} \boldsymbol{\beta}_i \boldsymbol{x}_{t-i} \right\|_F^2, \tag{B.2}$$

em que $\| \cdot \|_F^2$ é a norma de Frobenius (norma L_2) de uma matriz. A solução de (B.2) pode ser encontrada via MQO. Discussões e exemplos podem ser vistos em Tsay (2013).

B.2 Metodologia *VARX-L*

O modelo *VARX* pode se tornar altamente parametrizado na medida em que as dimensões da série temporal $\boldsymbol{\eta}_t$, d_0, e do vetor de covariáveis $\boldsymbol{x}_t$, r, aumentam. Tal característica também ocorre com o crescimento das ordens q e b. De fato, um modelo VARX(q, b) como definido na Equação (B.1) apresenta $d_0(1 + d_0q + rb + r)$ parâmetros. Assim, a metodologia *VARX-L* consiste em estimar modelos *VARX* com a adição de penalizações, de modo a obter matrizes $\boldsymbol{\phi}_\ell$ e $\boldsymbol{\beta}_i$ esparsas.

Dessa forma, no contexto *VARX-L*, o objetivo é resolver a minimização dada em (B.2) com a inclusão de um termo de penalização:

$$\underset{\boldsymbol{\nu},\boldsymbol{\phi},\boldsymbol{\beta}}{\arg\min} \sum_{t=1}^{T} \Big\| \boldsymbol{\eta}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{q} \boldsymbol{\phi}_\ell \boldsymbol{\eta}_{t-\ell} - \sum_{i=0}^{b} \boldsymbol{\beta}_i \boldsymbol{x}_{t-i} \Big\|_F^2 + \boldsymbol{\alpha}\Big\{\mathcal{P}(\boldsymbol{\phi}) + \mathcal{P}(\boldsymbol{\beta})\Big\},$$

onde $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_{d_0})$, com $\alpha_j \geq 0$, é um vetor de parâmetros de penalização, enquanto que $\mathcal{P}(\boldsymbol{\phi})$ e $\mathcal{P}(\boldsymbol{\beta})$ são funções de penalização atribuídas às matrizes $\boldsymbol{\phi}$ e $\boldsymbol{\beta}$, respectivamente (NICHOLSON; MATTESON; BIEN, 2017b).

Na aplicação aos dados deste trabalho, considerou-se uma versão da penalização Lasso para cada uma das matrizes de coeficientes. Nesse cenário, a estimação de $\boldsymbol{\nu}, \boldsymbol{\phi}$ e $\boldsymbol{\beta}$ é feita ao se resolver o problema de minimização

$$\underset{\boldsymbol{\nu},\boldsymbol{\phi},\boldsymbol{\beta}}{\arg\min} \sum_{t=1}^{T} \Big\| \boldsymbol{\eta}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{q} \boldsymbol{\phi}_\ell \boldsymbol{\eta}_{t-\ell} - \sum_{i=0}^{b} \boldsymbol{\beta}_i \boldsymbol{x}_{t-i} \Big\|_F^2 + \boldsymbol{\alpha}\Big\{\|\boldsymbol{\phi}\|_1 + \|\boldsymbol{\beta}\|_1\Big\}, \tag{B.3}$$

cuja solução é encontrada a partir de uma adaptação do algoritmo de um Lasso padrão para o contexto de modelos *VARX*. Para detalhes, consultar Nicholson, Matteson e Bien (2017a) e Nicholson, Matteson e Bien (2017b).

A determinação do parâmetro de penalização $\boldsymbol{\alpha}$ é obtida via validação cruzada *rolling window*, que preserva a estrutura temporal da série, escolhendo-se o valor $\hat{\boldsymbol{\alpha}}$ que minimiza o erro quadrático médio de previsão um passo à frente, por exemplo. Descrições mais detalhadas desse procedimento também podem ser consultadas em Nicholson, Matteson e Bien (2017b).

Após a imposição da penalização na Equação (B.3), algumas entradas das matrizes $\boldsymbol{\phi}$ e $\boldsymbol{\beta}$ serão zeradas e um reajuste pode ser feito ao se fixar de antemão que tais coeficientes são nulos, considerando a minimização expressa em (B.2).

APÊNDICE C

C.1 Ajuste global com 428 partidas

Anteriormente à modelagem final que considerou os mapas de intensidade das partidas do FC Barcelona entre as temporadas 2008/2009 e 2018/2009 (amostra de tamanho $n = 374$), um ajuste global incluindo, também, as partidas das temporadas 2006/2007 e 2007/2008 foi realizado (resultando em uma amostra de tamanho $n = 428$).

A Figura C.1 indica que a função média $\widehat{\mu}$ apresenta as mesmas características encontradas na média mostrada na Figura 4.2, calculada a partir do modelo que exclui as observações nos tempos $t = 1, \ldots, 54$. Da mesma maneira, assim como no ajuste com menos observações, os dois primeiros autovalores se destacaram em relação aos demais, como pode ser visto na Figura C.2.

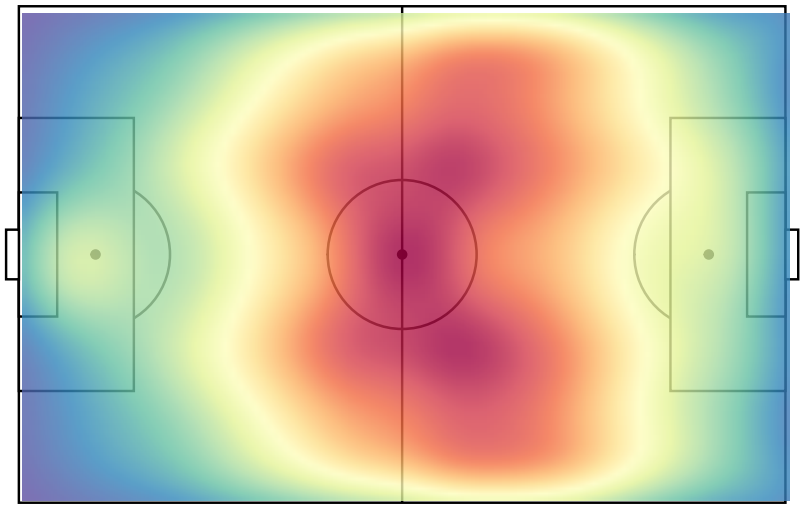


Figura C.1: Função média $\widehat{\mu}$ após ajuste com 428 partidas.

Como no modelo final ajustado na Seção 4.3 utilizou-se apenas as duas primeiras autofunções, por simplicidade e para comparações com relação ao que realmente

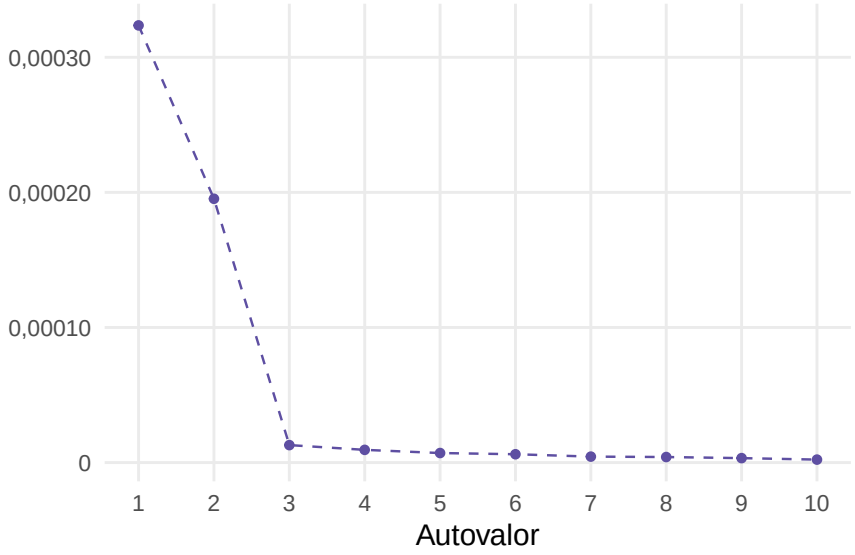
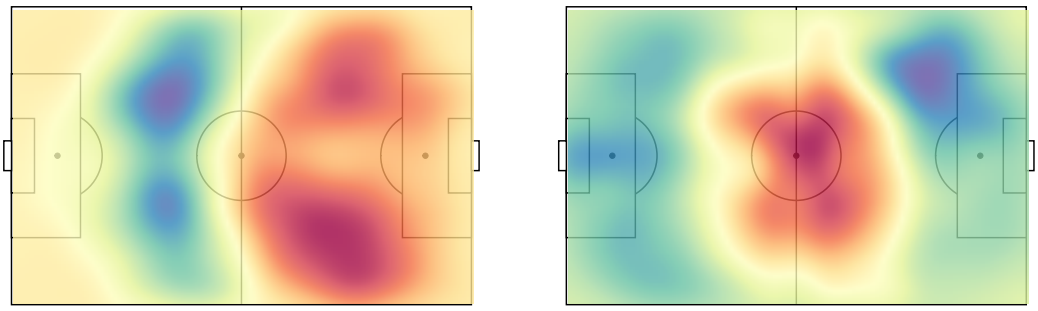


Figura C.2: Dez primeiros autovalores da matriz $\boldsymbol{K}^*$ após ajuste com 428 partidas.

é relevante para a modelagem final, a Figura C.3 contém somente essas autofunções. Contudo, em contraste às autofunções ilustradas na Figura 4.4, é possível notar que houve uma mudança de posição entre elas. Agora, $\widehat{\psi}_1$, apresentada na Figura C.3(a), captura se o FC Barcelona foi ofensivo ou defensivo em suas partidas, ao passo que $\widehat{\psi}_2$, exposta na Figura C.3(b), caracteriza se a equipe manteve mais a bola no setor de meio-campo ou não.

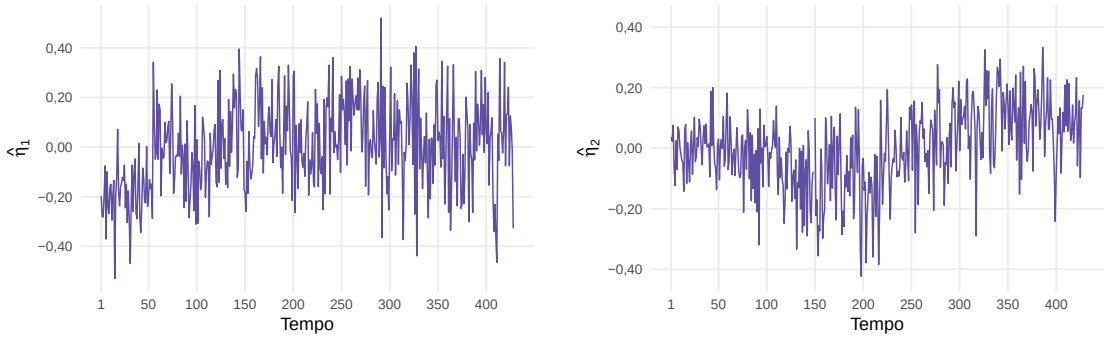


(a) $\widehat{\psi}_1$ (b) $\widehat{\psi}_2$

Figura C.3: Autofunções $\widehat{\psi}_j$, $j = 1, 2$, obtidas via ajuste com 428 partidas.

As séries temporais $\widehat{\eta}_{tj}$, $j = 1, 2$, associadas às autofunções acima podem ser visualizadas na Figura C.4. Neste ponto é de interesse particular analisar a série $\widehat{\eta}_{t1}$ (Figura C.4(a)), na qual observa-se valores majoritariamente negativos nos tempos $t = 1, \ldots, 54$, enquanto que para as demais partidas, $t = 55, \ldots, 428$, tem-se valores de $\widehat{\eta}_{t1}$ ao redor de zero. Tal constatação sugere que $\widehat{\eta}_{t1}$ apresenta um comportamento em dois regimes (isto é, com diferentes médias). Como viu-se que a autofunção $\widehat{\psi}_1$ indica ofensividade/defensividade, pode-se concluir que nas partidas

$t = 1, \ldots, 54$ a equipe espanhola apresentou uma postura mais defensiva, ocupando de forma mais predominante o campo de defesa. Além disso, destaca-se que essas partidas coincidem com o período final em que o time de Barcelona foi treinado pelo holandês Frank Rijkaard. Consequentemente, como uma das covariáveis adicionadas ao modelo *VARX* é uma indicadora do treinador na partida t, na equação de $\widehat{\eta}_{t1}$ apenas essa covariável acaba se tornando importante para explicar a série temporal. Por essas razões, optou-se por descartar as partidas $t = 1, \ldots, 54$ da análise e o procedimento de modelagem e previsão foi realizado com as demais observações conforme apresentado nas Seções 4.3, 4.4 e 4.5.

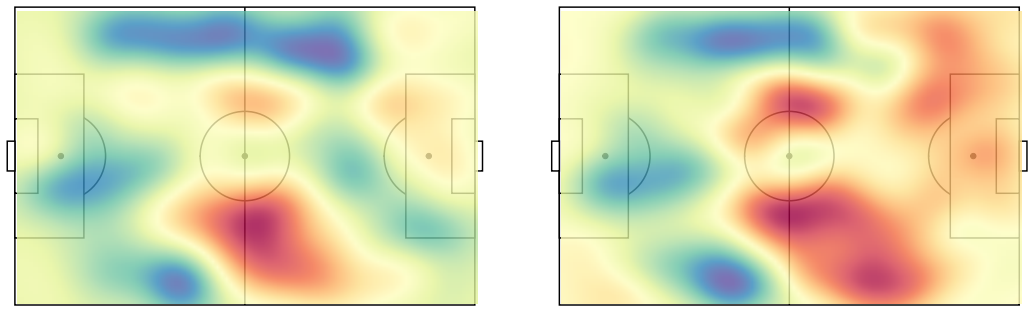


(a) $\widehat{\eta}_{t1}$ (b) $\widehat{\eta}_{t2}$

Figura C.4: Séries temporais $\widehat{\eta}_{tj}$, $j = 1, 2$, resultantes do ajuste com 428 partidas.

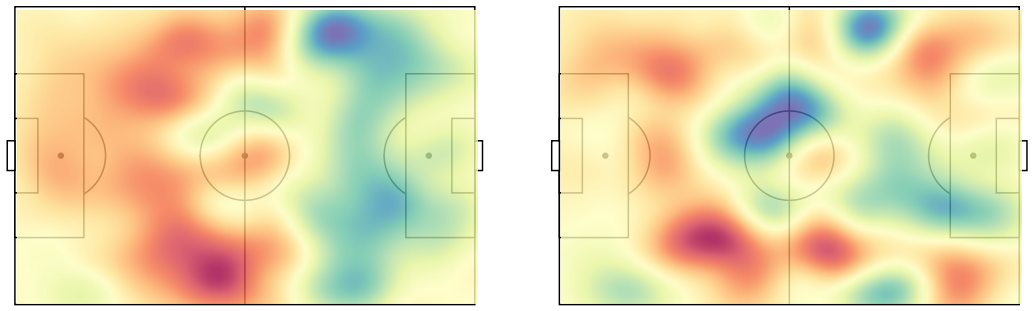
APÊNDICE D

D.1 Previsões finais



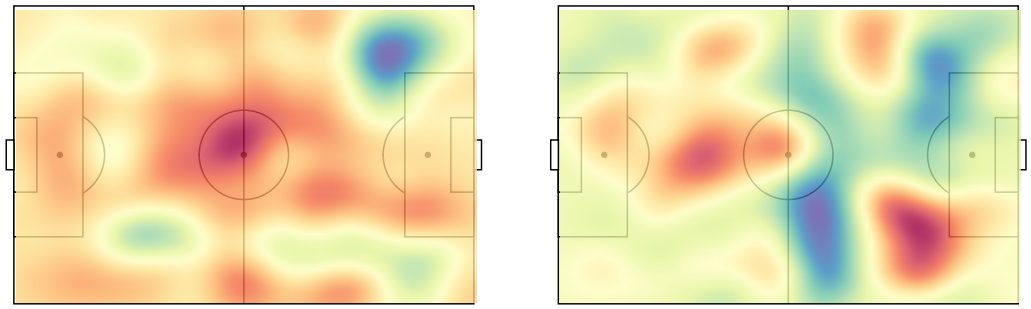
(a) $\widehat{\lambda}_{325|324} - \widehat{\lambda}_{324}$ (b) $\widehat{\lambda}_{325} - \widehat{\lambda}_{324}$

Figura D.1: Variação predita e observada, $T + 1 = 325$.



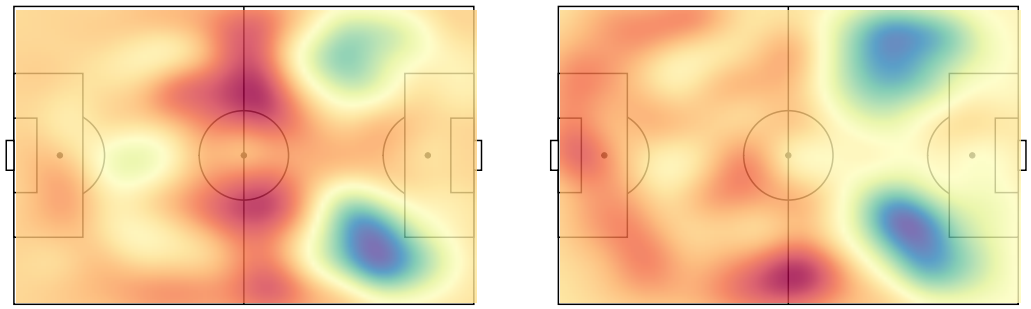
(a) $\widehat{\lambda}_{326|325} - \widehat{\lambda}_{325}$ (b) $\widehat{\lambda}_{326} - \widehat{\lambda}_{325}$

Figura D.2: Variação predita e observada, $T + 1 = 326$.



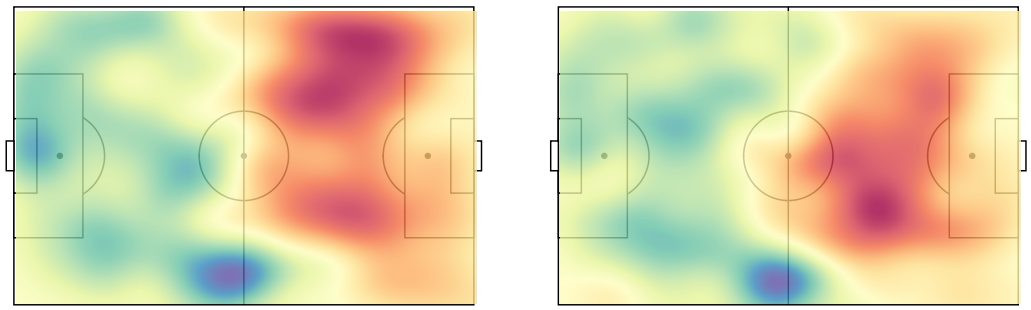
(a) $\widehat{\lambda}_{327|326} - \widehat{\lambda}_{326}$ (b) $\widehat{\lambda}_{327} - \widehat{\lambda}_{326}$

Figura D.3: Variação predita e observada, $T + 1 = 327$.



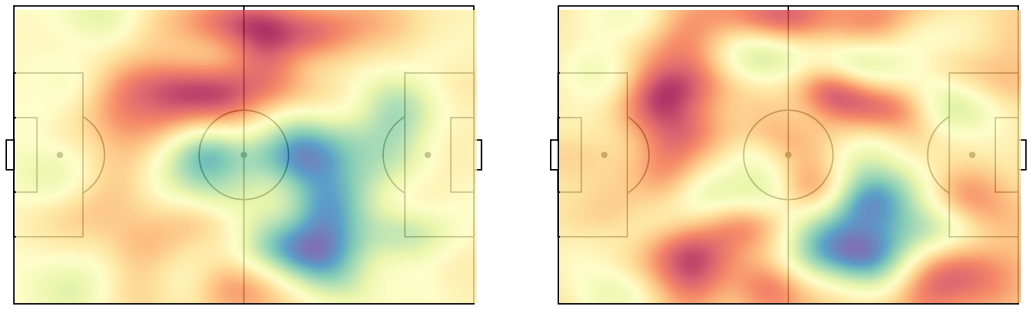
(a) $\widehat{\lambda}_{328|327} - \widehat{\lambda}_{327}$ (b) $\widehat{\lambda}_{328} - \widehat{\lambda}_{327}$

Figura D.4: Variação predita e observada, $T + 1 = 328$.



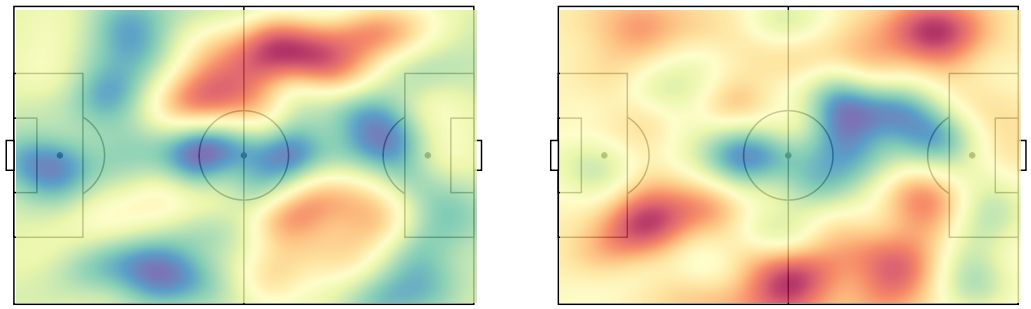
(a) $\widehat{\lambda}_{329|328} - \widehat{\lambda}_{328}$ (b) $\widehat{\lambda}_{329} - \widehat{\lambda}_{328}$

Figura D.5: Variação predita e observada, $T + 1 = 329$.



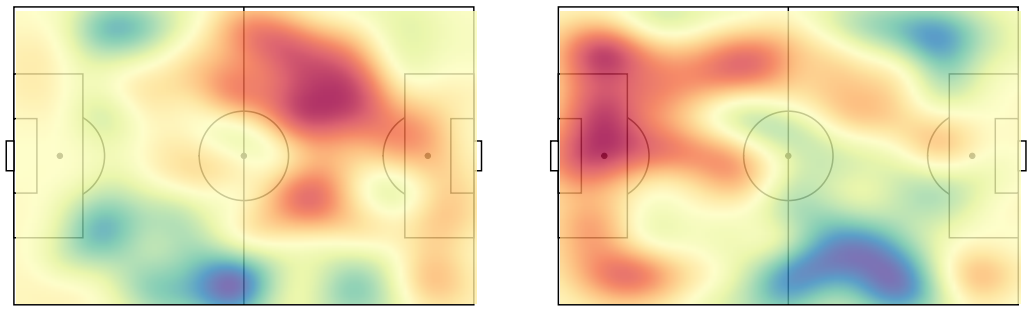
(a) $\widehat{\lambda}_{330|329} - \widehat{\lambda}_{329}$ (b) $\widehat{\lambda}_{330} - \widehat{\lambda}_{329}$

Figura D.6: Variação predita e observada, $T + 1 = 330$.



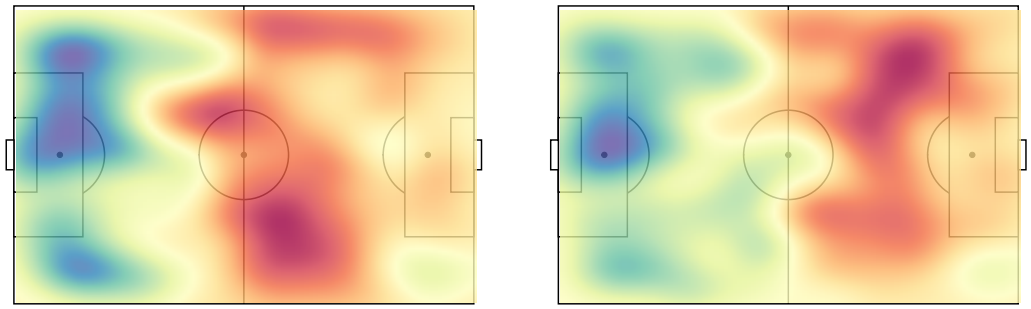
(a) $\widehat{\lambda}_{331|330} - \widehat{\lambda}_{330}$ (b) $\widehat{\lambda}_{331} - \widehat{\lambda}_{330}$

Figura D.7: Variação predita e observada, $T + 1 = 331$.



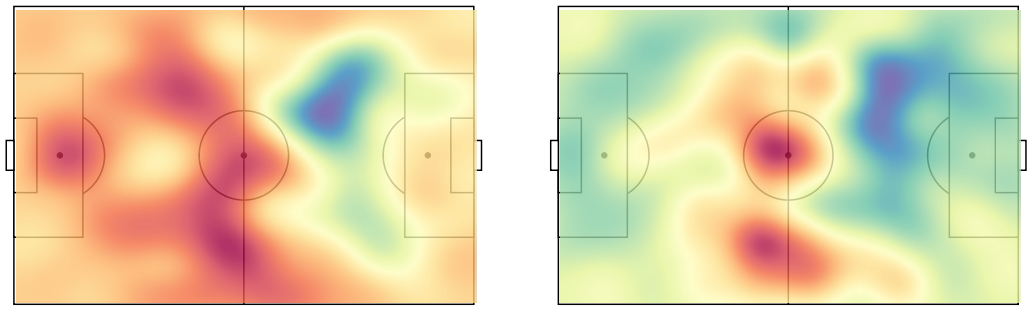
(a) $\widehat{\lambda}_{332|331} - \widehat{\lambda}_{331}$ (b) $\widehat{\lambda}_{332} - \widehat{\lambda}_{331}$

Figura D.8: Variação predita e observada, $T + 1 = 332$.



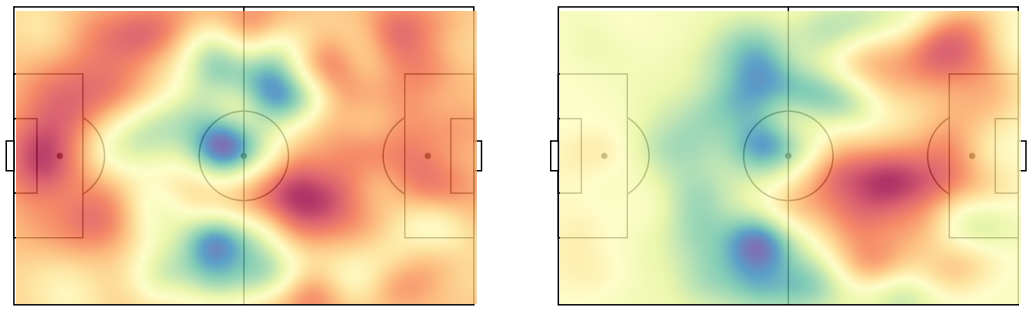
(a) $\widehat{\lambda}_{333|332} - \widehat{\lambda}_{332}$ (b) $\widehat{\lambda}_{333} - \widehat{\lambda}_{332}$

Figura D.9: Variação predita e observada, $T + 1 = 333$.



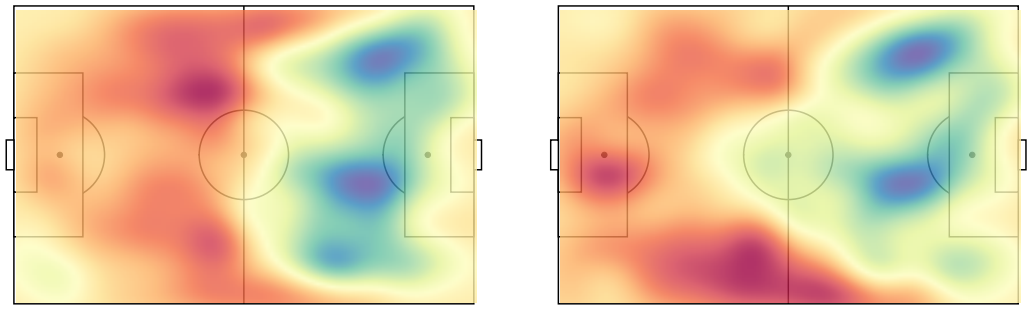
(a) $\widehat{\lambda}_{334|333} - \widehat{\lambda}_{333}$ (b) $\widehat{\lambda}_{334} - \widehat{\lambda}_{333}$

Figura D.10: Variação predita e observada, $T + 1 = 334$.



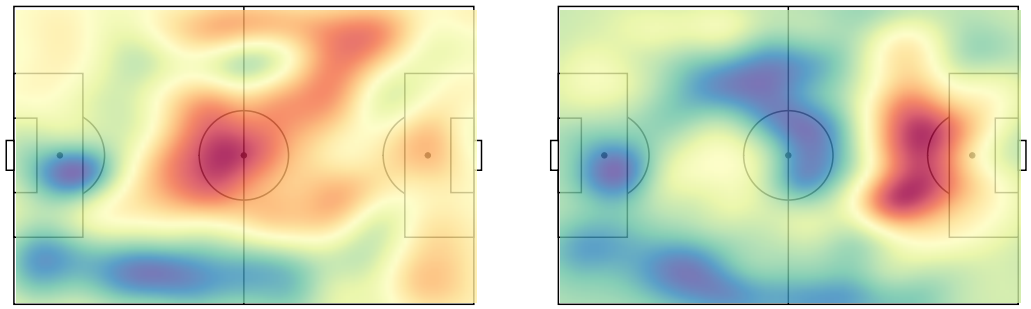
(a) $\widehat{\lambda}_{335|334} - \widehat{\lambda}_{334}$ (b) $\widehat{\lambda}_{335} - \widehat{\lambda}_{334}$

Figura D.11: Variação predita e observada, $T + 1 = 335$.



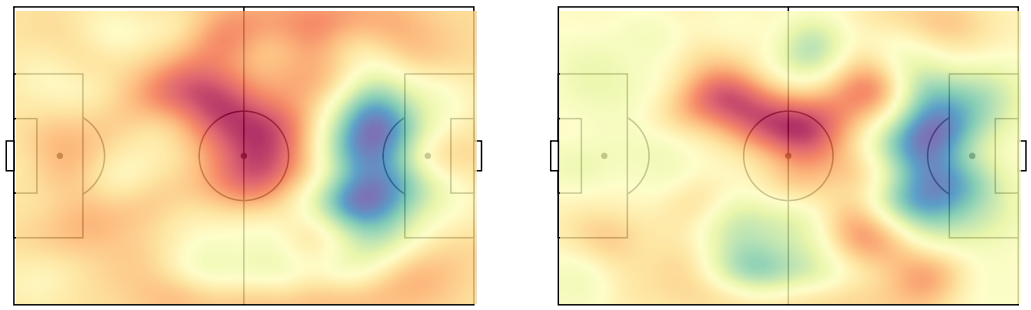
(a) $\widehat{\lambda}_{336|335} - \widehat{\lambda}_{335}$ (b) $\widehat{\lambda}_{336} - \widehat{\lambda}_{335}$

Figura D.12: Variação predita e observada, $T + 1 = 336$.



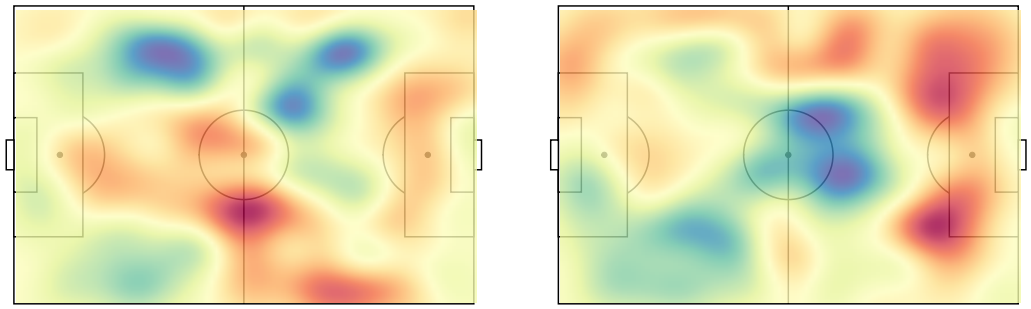
(a) $\widehat{\lambda}_{337|336} - \widehat{\lambda}_{336}$ (b) $\widehat{\lambda}_{337} - \widehat{\lambda}_{336}$

Figura D.13: Variação predita e observada, $T + 1 = 337$.



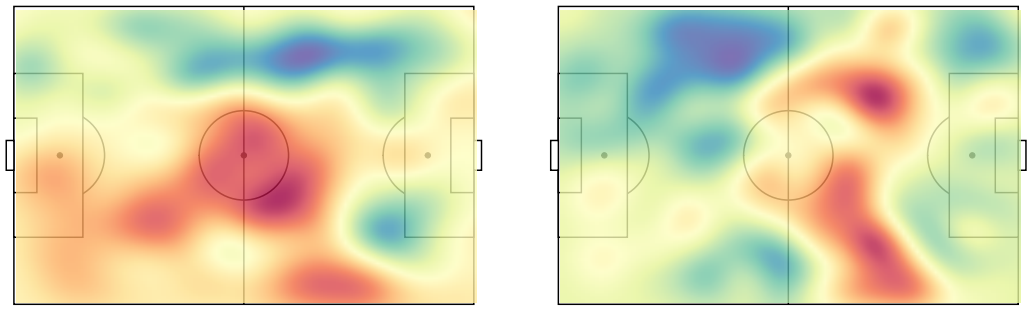
(a) $\widehat{\lambda}_{338|337} - \widehat{\lambda}_{337}$ (b) $\widehat{\lambda}_{338} - \widehat{\lambda}_{337}$

Figura D.14: Variação predita e observada, $T + 1 = 338$.



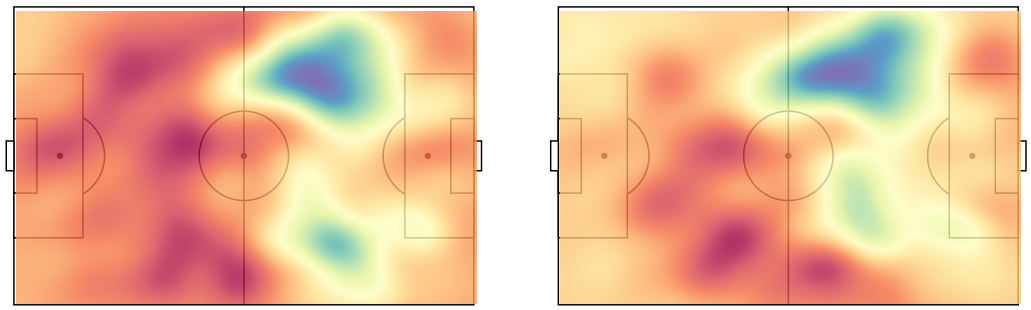
(a) $\widehat{\lambda}_{340|339} - \widehat{\lambda}_{339}$ (b) $\widehat{\lambda}_{340} - \widehat{\lambda}_{339}$

Figura D.15: Variação predita e observada, $T + 1 = 340$.



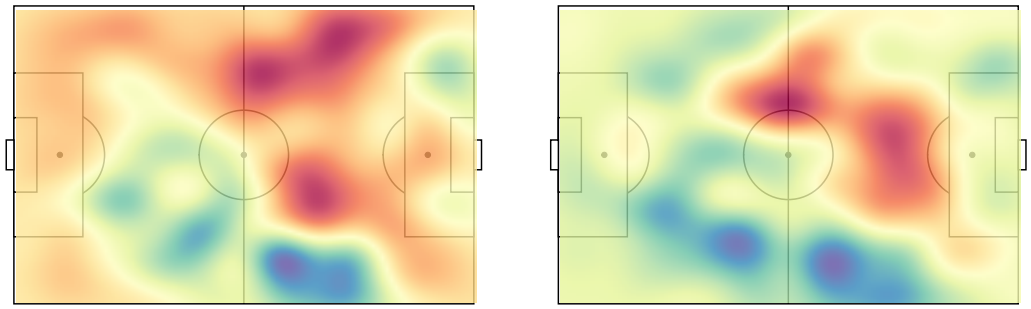
(a) $\widehat{\lambda}_{341|340} - \widehat{\lambda}_{340}$ (b) $\widehat{\lambda}_{341} - \widehat{\lambda}_{340}$

Figura D.16: Variação predita e observada, $T + 1 = 341$.



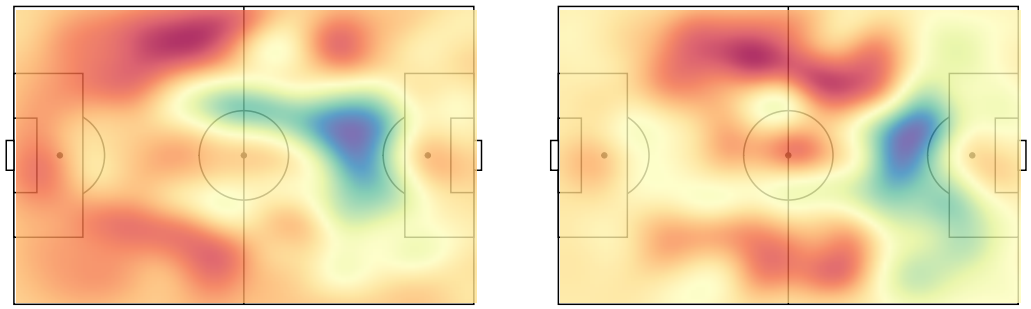
(a) $\widehat{\lambda}_{342|341} - \widehat{\lambda}_{341}$ (b) $\widehat{\lambda}_{342} - \widehat{\lambda}_{341}$

Figura D.17: Variação predita e observada, $T + 1 = 342$.



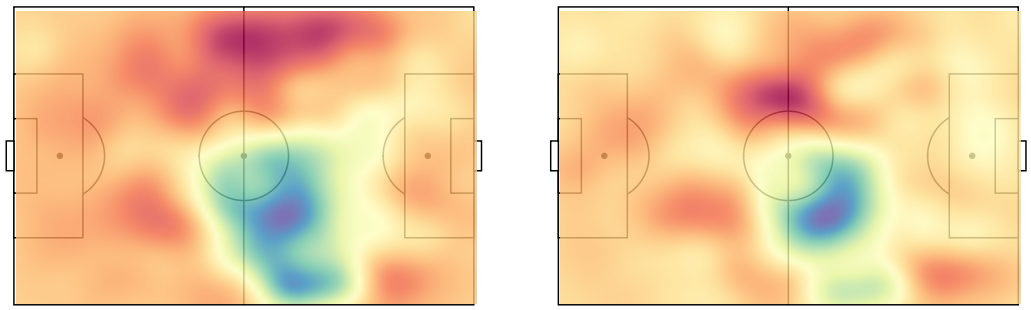
(a) $\widehat{\lambda}_{343|342} - \widehat{\lambda}_{342}$ (b) $\widehat{\lambda}_{343} - \widehat{\lambda}_{342}$

Figura D.18: Variação predita e observada, $T + 1 = 343$.



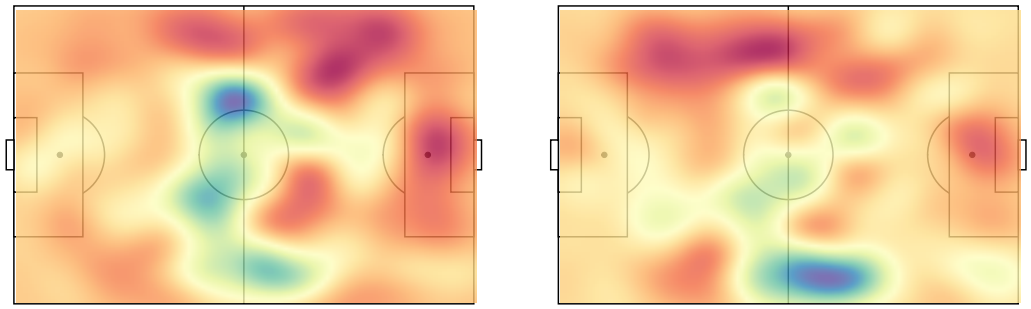
(a) $\widehat{\lambda}_{344|343} - \widehat{\lambda}_{343}$ (b) $\widehat{\lambda}_{344} - \widehat{\lambda}_{343}$

Figura D.19: Variação predita e observada, $T + 1 = 344$.



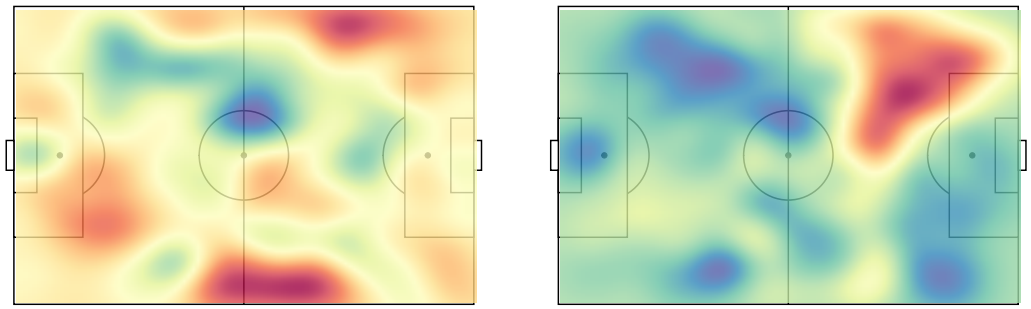
(a) $\widehat{\lambda}_{346|345} - \widehat{\lambda}_{345}$ (b) $\widehat{\lambda}_{346} - \widehat{\lambda}_{345}$

Figura D.20: Variação predita e observada, $T + 1 = 346$.



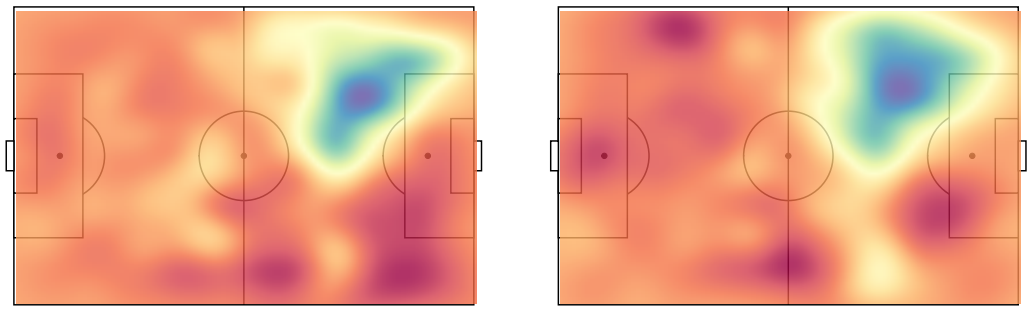
(a) $\widehat{\lambda}_{347|346} - \widehat{\lambda}_{346}$ (b) $\widehat{\lambda}_{347} - \widehat{\lambda}_{346}$

Figura D.21: Variação predita e observada, $T + 1 = 347$.



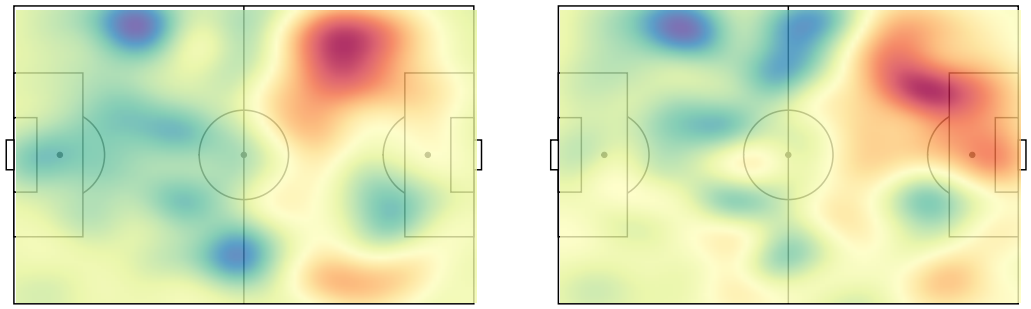
(a) $\widehat{\lambda}_{348|347} - \widehat{\lambda}_{347}$ (b) $\widehat{\lambda}_{348} - \widehat{\lambda}_{347}$

Figura D.22: Variação predita e observada, $T + 1 = 348$.



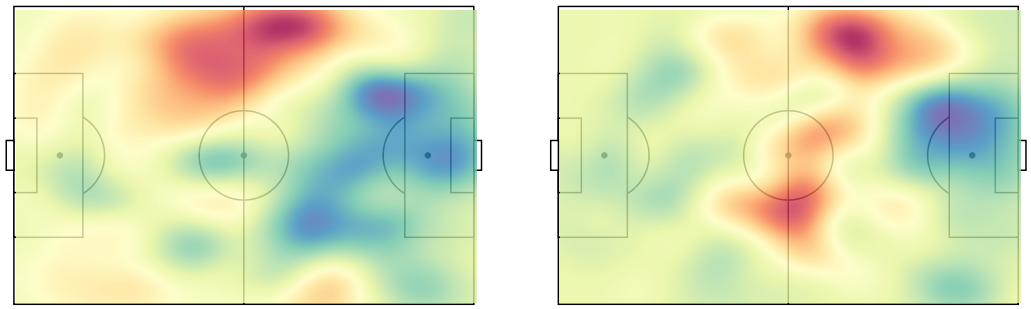
(a) $\widehat{\lambda}_{349|348} - \widehat{\lambda}_{348}$ (b) $\widehat{\lambda}_{349} - \widehat{\lambda}_{348}$

Figura D.23: Variação predita e observada, $T + 1 = 349$.



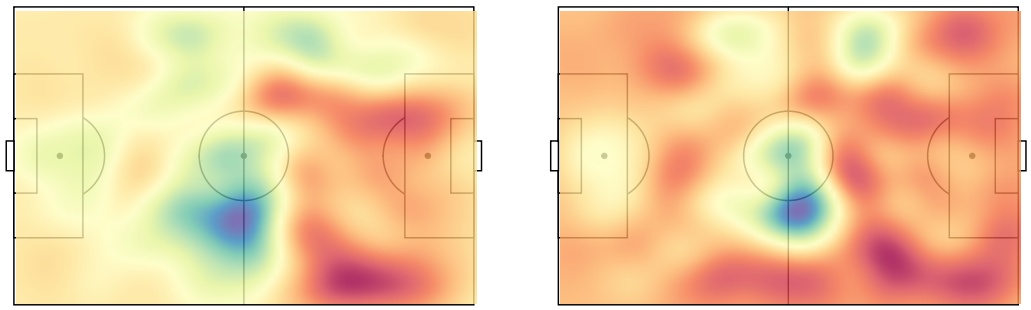
(a) $\widehat{\lambda}_{350|349} - \widehat{\lambda}_{349}$ (b) $\widehat{\lambda}_{350} - \widehat{\lambda}_{349}$

Figura D.24: Variação predita e observada, $T + 1 = 350$.



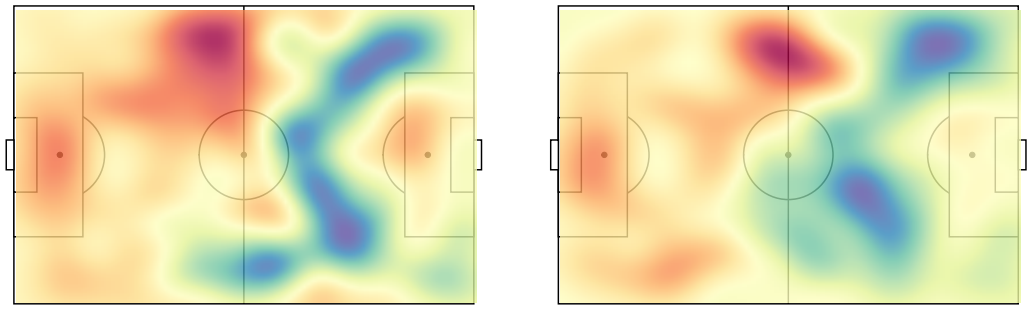
(a) $\widehat{\lambda}_{351|350} - \widehat{\lambda}_{350}$ (b) $\widehat{\lambda}_{351} - \widehat{\lambda}_{350}$

Figura D.25: Variação predita e observada, $T + 1 = 351$.



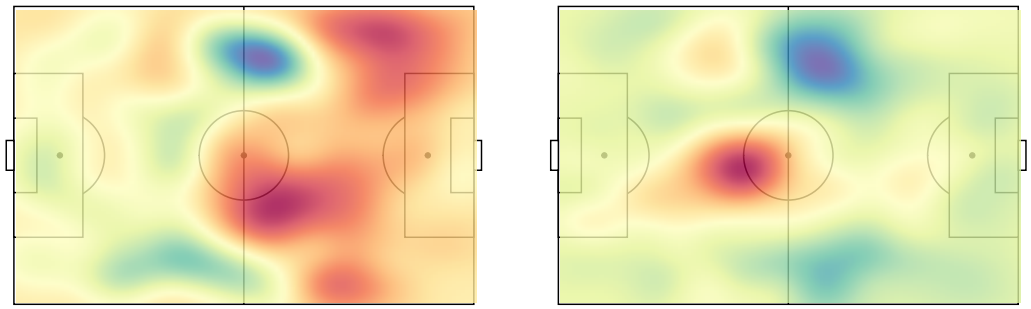
(a) $\widehat{\lambda}_{352|351} - \widehat{\lambda}_{351}$ (b) $\widehat{\lambda}_{352} - \widehat{\lambda}_{351}$

Figura D.26: Variação predita e observada, $T + 1 = 352$.



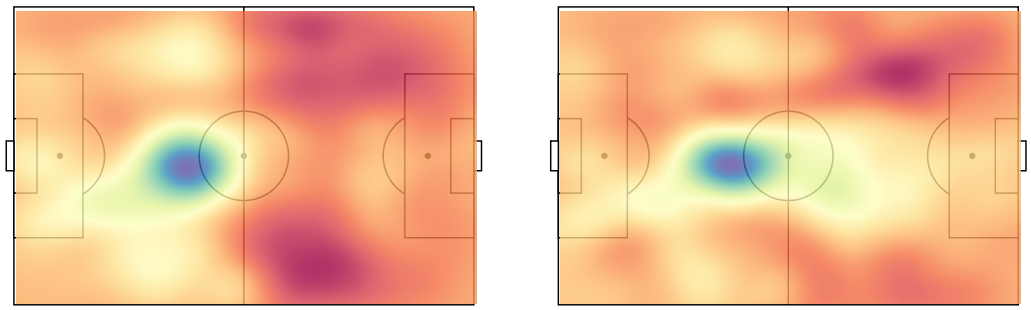
(a) $\widehat{\lambda}_{353|352} - \widehat{\lambda}_{352}$ (b) $\widehat{\lambda}_{353} - \widehat{\lambda}_{352}$

Figura D.27: Variação predita e observada, $T + 1 = 353$.



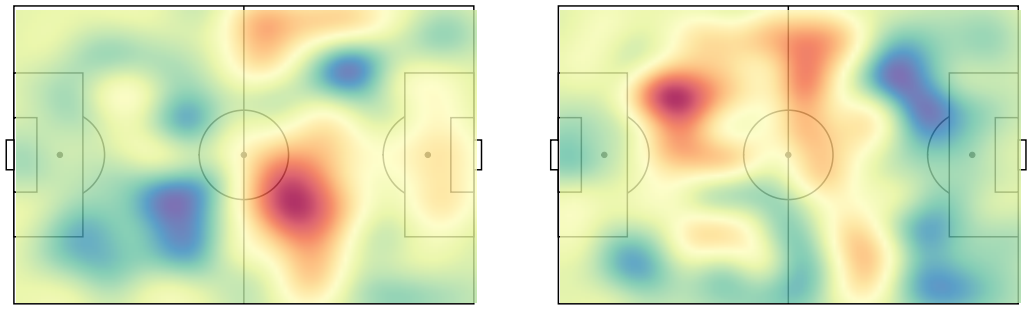
(a) $\widehat{\lambda}_{354|353} - \widehat{\lambda}_{353}$ (b) $\widehat{\lambda}_{354} - \widehat{\lambda}_{353}$

Figura D.28: Variação predita e observada, $T + 1 = 354$.



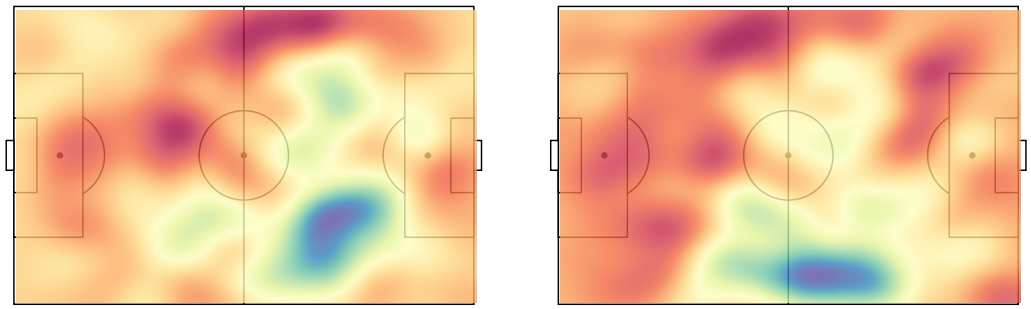
(a) $\widehat{\lambda}_{355|354} - \widehat{\lambda}_{354}$ (b) $\widehat{\lambda}_{355} - \widehat{\lambda}_{354}$

Figura D.29: Variação predita e observada, $T + 1 = 355$.



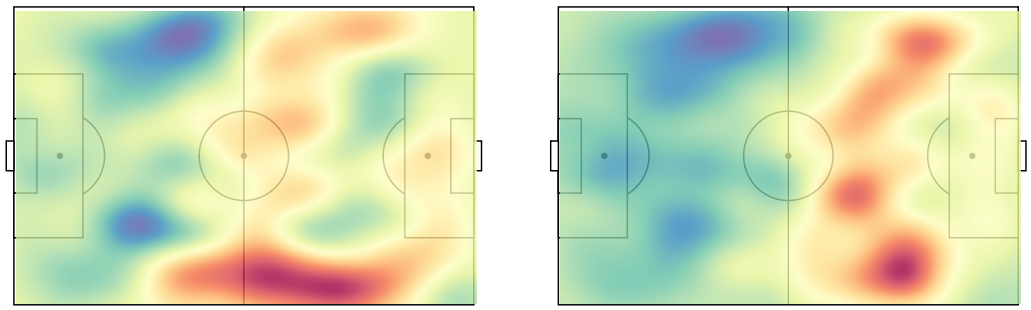
(a) $\widehat{\lambda}_{356|355} - \widehat{\lambda}_{355}$ (b) $\widehat{\lambda}_{356} - \widehat{\lambda}_{355}$

Figura D.30: Variação predita e observada, $T + 1 = 356$.



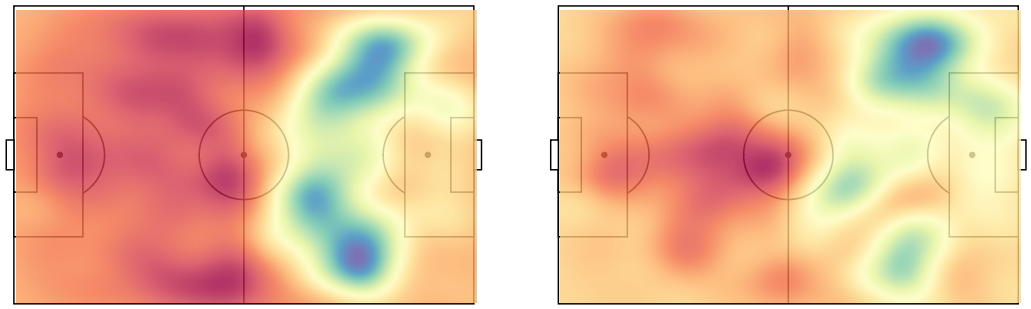
(a) $\widehat{\lambda}_{359|358} - \widehat{\lambda}_{358}$ (b) $\widehat{\lambda}_{359} - \widehat{\lambda}_{358}$

Figura D.31: Variação predita e observada, $T + 1 = 359$.



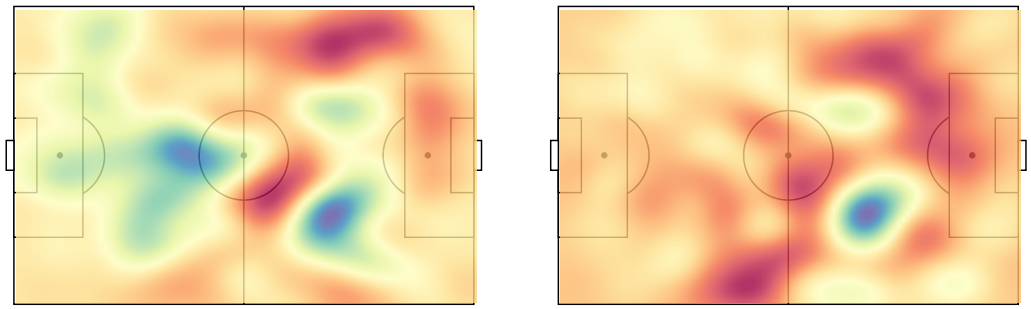
(a) $\widehat{\lambda}_{360|359} - \widehat{\lambda}_{359}$ (b) $\widehat{\lambda}_{360} - \widehat{\lambda}_{359}$

Figura D.32: Variação predita e observada, $T + 1 = 360$.



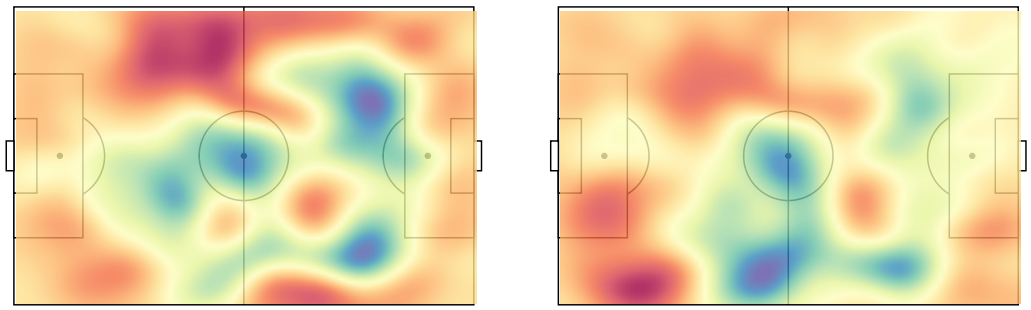
(a) $\widehat{\lambda}_{361|360} - \widehat{\lambda}_{360}$ (b) $\widehat{\lambda}_{361} - \widehat{\lambda}_{360}$

Figura D.33: Variação predita e observada, $T + 1 = 361$.



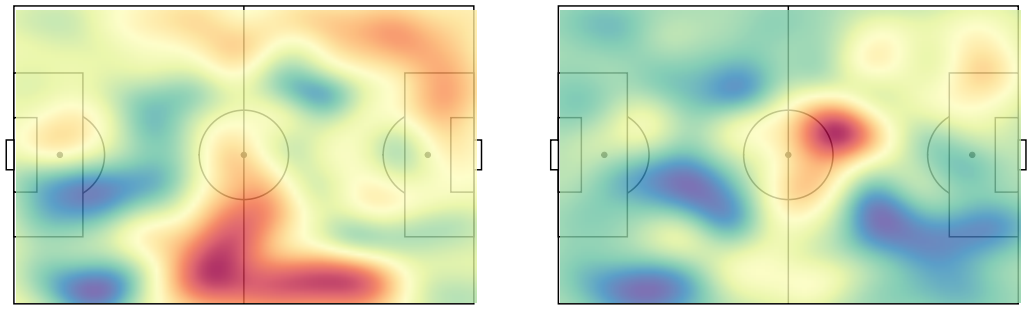
(a) $\widehat{\lambda}_{362|361} - \widehat{\lambda}_{361}$ (b) $\widehat{\lambda}_{362} - \widehat{\lambda}_{361}$

Figura D.34: Variação predita e observada, $T + 1 = 362$.



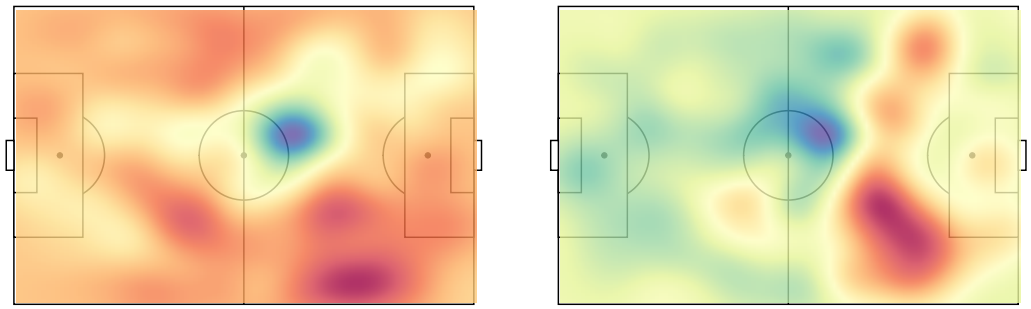
(a) $\widehat{\lambda}_{363|362} - \widehat{\lambda}_{362}$ (b) $\widehat{\lambda}_{363} - \widehat{\lambda}_{362}$

Figura D.35: Variação predita e observada, $T + 1 = 363$.



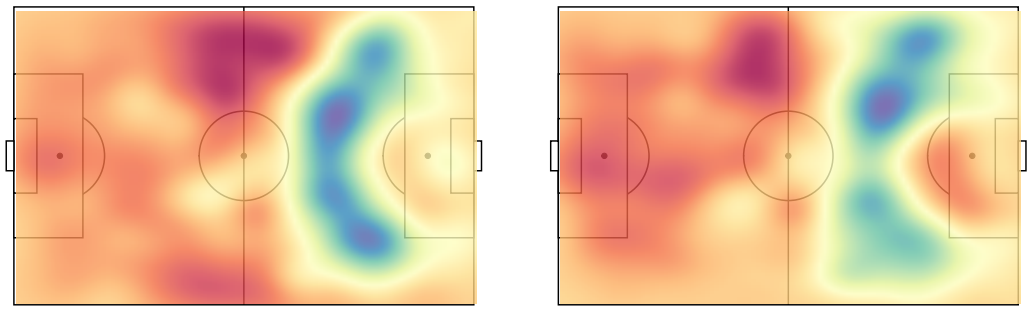
(a) $\widehat{\lambda}_{364|333} - \widehat{\lambda}_{363}$ (b) $\widehat{\lambda}_{364} - \widehat{\lambda}_{363}$

Figura D.36: Variação predita e observada, $T + 1 = 364$.



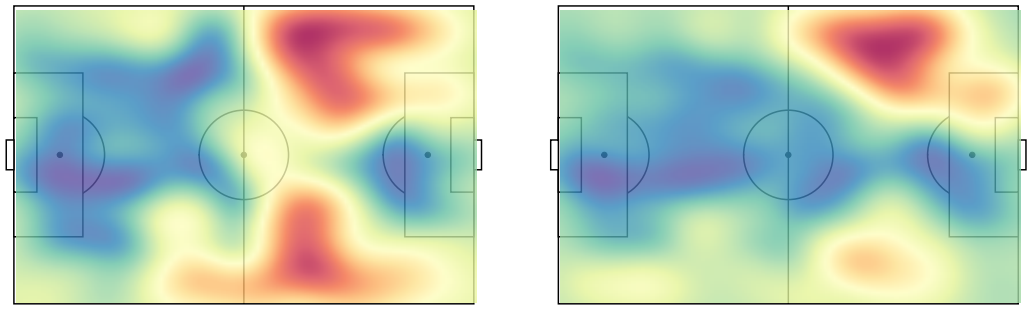
(a) $\widehat{\lambda}_{365|364} - \widehat{\lambda}_{364}$ (b) $\widehat{\lambda}_{365} - \widehat{\lambda}_{364}$

Figura D.37: Variação predita e observada, $T + 1 = 365$.



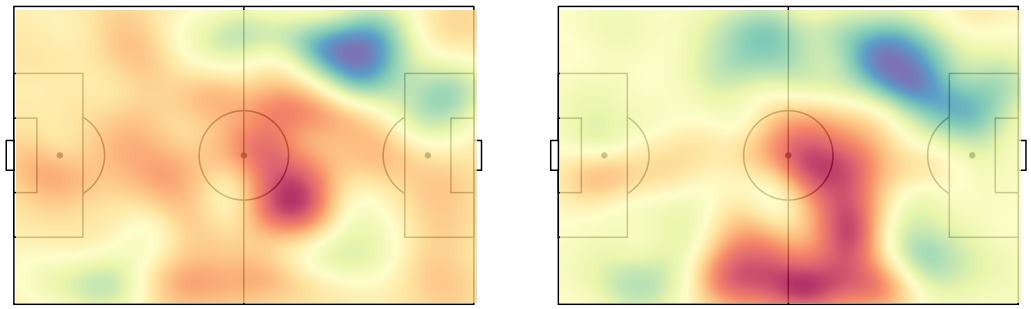
(a) $\widehat{\lambda}_{366|365} - \widehat{\lambda}_{365}$ (b) $\widehat{\lambda}_{366} - \widehat{\lambda}_{365}$

Figura D.38: Variação predita e observada, $T + 1 = 366$.



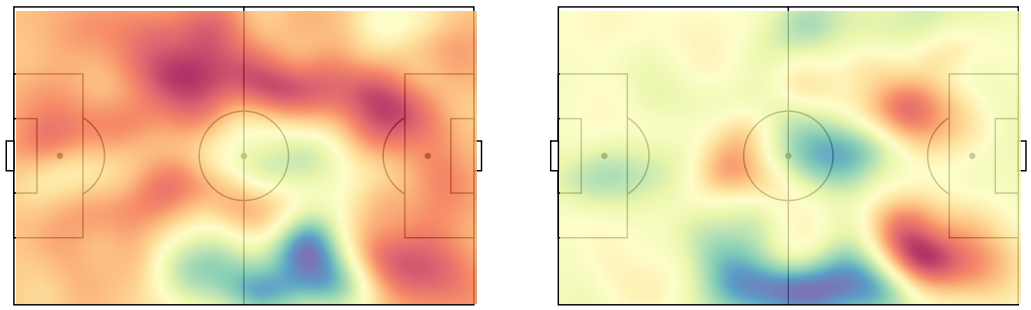
(a) $\widehat{\lambda}_{367|366} - \widehat{\lambda}_{366}$ (b) $\widehat{\lambda}_{367} - \widehat{\lambda}_{366}$

Figura D.39: Variação predita e observada, $T + 1 = 367$.



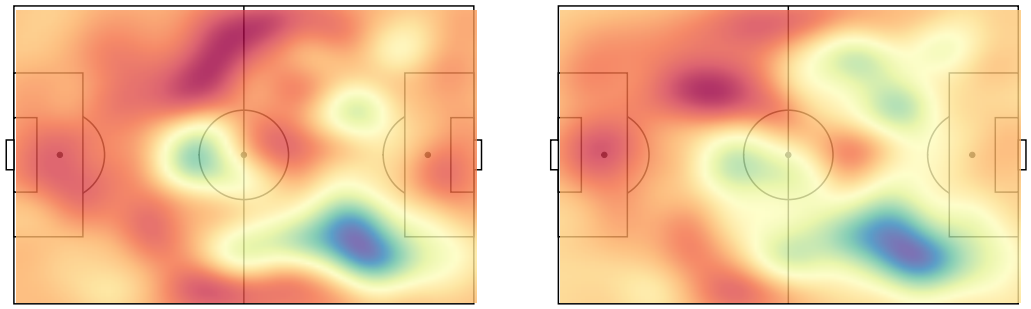
(a) $\widehat{\lambda}_{368|367} - \widehat{\lambda}_{367}$ (b) $\widehat{\lambda}_{368} - \widehat{\lambda}_{367}$

Figura D.40: Variação predita e observada, $T + 1 = 368$.



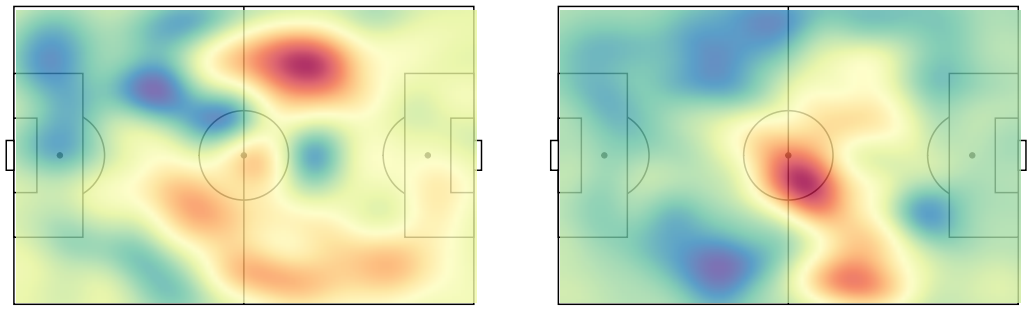
(a) $\widehat{\lambda}_{369|368} - \widehat{\lambda}_{368}$ (b) $\widehat{\lambda}_{369} - \widehat{\lambda}_{368}$

Figura D.41: Variação predita e observada, $T + 1 = 369$.



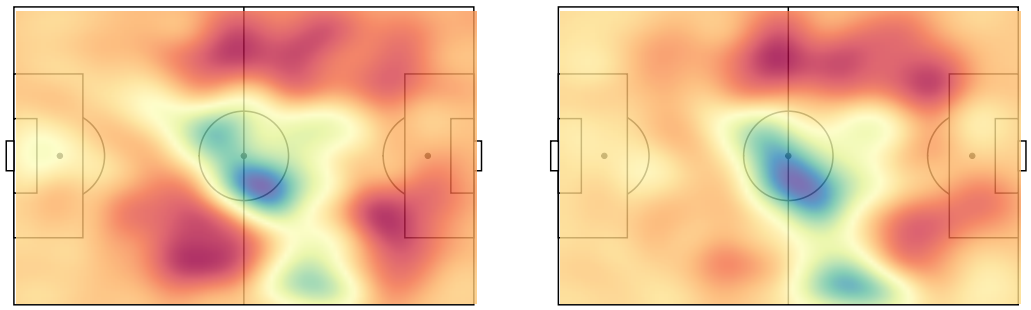
(a) $\widehat{\lambda}_{370|369} - \widehat{\lambda}_{369}$ (b) $\widehat{\lambda}_{370} - \widehat{\lambda}_{369}$

Figura D.42: Variação predita e observada, $T + 1 = 370$.



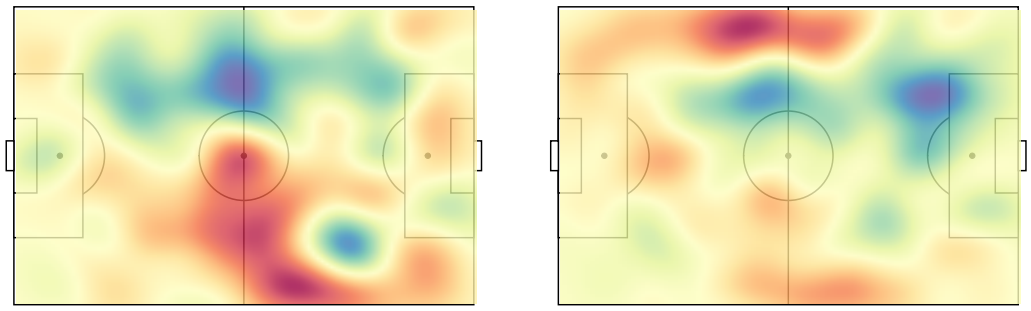
(a) $\widehat{\lambda}_{371|370} - \widehat{\lambda}_{370}$ (b) $\widehat{\lambda}_{371} - \widehat{\lambda}_{370}$

Figura D.43: Variação predita e observada, $T + 1 = 371$.



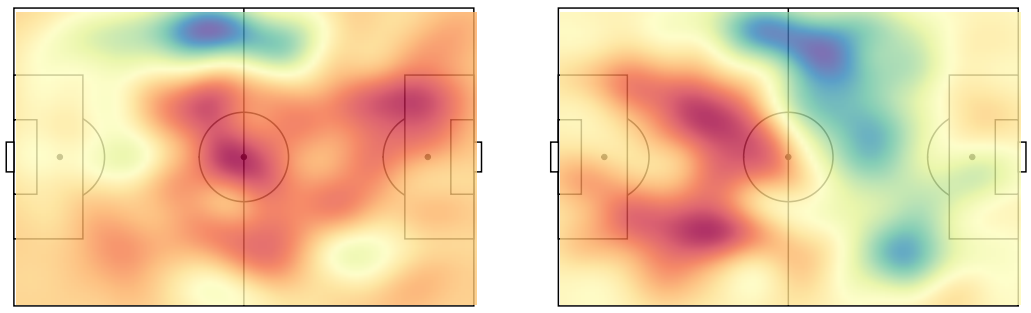
(a) $\widehat{\lambda}_{372|371} - \widehat{\lambda}_{371}$ (b) $\widehat{\lambda}_{372} - \widehat{\lambda}_{371}$

Figura D.44: Variação predita e observada, $T + 1 = 372$.



(a) $\widehat{\lambda}_{373|372} - \widehat{\lambda}_{372}$ (b) $\widehat{\lambda}_{373} - \widehat{\lambda}_{372}$

Figura D.45: Variação predita e observada, $T + 1 = 373$.



(a) $\widehat{\lambda}_{374|373} - \widehat{\lambda}_{373}$ (b) $\widehat{\lambda}_{374} - \widehat{\lambda}_{373}$

Figura D.46: Variação predita e observada, $T + 1 = 374$.