

Premier League via Dirichlet Regression

Igor Ferreira do Nascimento^{a,b}, Pedro Henrique Melo Albuquerque^{a,*}, Yaohao Peng^a

^a *University of Brasília, Campus Darcy Ribeiro, Brasília, Distrito Federal, 70910-900, Brazil.*

^b *Piauí Institute of Technology, R. Álvaro Mendes, 94, Teresina, Piauí, 64000-040, Brazil.*

Abstract

Keywords: Premier League; Dirichlet Regression

*Corresponding author

Email addresses: igor.nascimento@ifpi.edu.br, pedroa@unb.br, peng.yaohao@gmail.com

URL: www.lamfo.unb.br

1. Introdução

O fascínio por eventos incertos instiga o ser humano desde os remotos tempos dos antepassados. Os historiadores XXXX mostram que desde a época o ser humano demonstrava interesse por saber ao menos algo sobre o que é incerto. Ao longo dos anos, a teoria dos jogos proporcionou o desenvolvimento de modelos probabilísticos para determinar o nível de determinadas ocorrências.

O objetivo do trabalho é apresentar modelos probabilísticos para representar os resultados da Premier League por meio de covariáveis relacionado às habilidades dos jogadores do time. Além disso, é apresentada uma estimativa para o resultado final da tabela de classificação.

2. Sports Forecast

Sport is a human activity normally related to the playful environment but we can not forget the betting. The work like Kain and Logan (2014) analyse the dissonance between the outcome of sports and the sports betting market explained by fact the set betting lines is straight related with the profit goal instead the outcome probability. On the other hand, the paper Štrumbelj and Šikonja (2010) analysed some European soccer leagues and show that exist evidence that bookmakers learn over time about soccer prediction results. Anyway, this is interesting to either bettors or bookmakers researchers in this area.

As redes de apostas como, XXX, YYY, GGG são encontrados ao redor do mundo e fazem girar cifras ainda maiores de recursos no ato de acreditar em um determinado resultado. Esportes como o baseball e o futebol americano possuem grande suporte dos cientistas de dados para apresentar pequenas melhorias nos resultados individuais dos atletas, mas em conjunto representa grande impacto para o time como todo. De encontro com tais anseios está o desenvolvimento de tecnologias de captação, armazenamento, tratamento e análise de dados relacionados às atividades esportivas.

O futebol é, se não o mais, um dos esportes com as movimentações financeiras de destaque na atualidade. Atualmente, existem elenco de jogadores em times que passam a casa dos 10 dígitos, sobre tudo na Europa Ocidental. Os times Espanhóis e Britânicos lideram os valores financeiros do elenco. No entanto, essa aparente hegemonia é, cada vez mais, desafiada com o surgimento de novos mercados da bola, tais como o mercado norte americano e asiático.

Dito isso, pode-se destacar que os resultados esportivos são de interesse da sociedade, seja ela do ponto de vista financeiro ou a mera busca por entretenimento.

30 3. Método

3.1. Regressão Dirichlet

A distribuição Dirichlet permite modelar um vetor composicional de dimensão $k \leq 2$ $Y = (y_1, y_2, \dots, y_K)$, com parâmetros de escala $\alpha_1, \alpha_2, \dots, \alpha_K \geq 0$, sendo $\sum_{j=1}^K y_j = 1$ um padrão $K - 1$ simplex. Dessa forma, as probabilidades presentes na distribuição multinomial tem como
 35 priori conjugada tal distribuição, com densidade probabilidade:

The Dirichlet distribution is able to model a composite vector of dimension $k \leq 2$ $Y = (y_1, y_2, \dots, y_K)$, with scalings parameters $\alpha_1, \alpha_2, \dots, \alpha_K \geq 0$, being $\sum_{j=1}^K y_j = 1$ a standard $K - 1$ simplex. Thus the probabilities present in the multinomial distribution have as a priori conjugated such distribution, with probability density:

$$P(Y|\alpha) = \frac{\prod_{j=1}^K \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^K \alpha_j)} \prod_{j=1}^K y_j^{\alpha_j-1} \quad (1)$$

40 The work of Hijazi and Jernigan (2009) that observed a good ajust upon longitudinal covariates changes using Dirichlet Regression. In field of sports the work of Null (2009) show a competitive accuracy modeling player abilities in Major League Baseball.

We object model the scale parameters by covariates X_r , and use this covariates to forecast the final classification result. Thus the model is:

$$g[\alpha_j(X)] = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + \dots + \beta_{jR}X_R + \varepsilon_j \quad (2)$$

45 Being $g(\cdot) = \log(\cdot)$ has a transformation possible. The sample Y with size N the likelihood is:

$$L(\beta|X, Y) = \prod_{i=1}^N \left[\frac{\prod_{j=1}^K \Gamma(\alpha_j(X))}{\Gamma(\sum_{j=1}^K \alpha_c(X))} \prod_{j=1}^K y_{ij}^{\alpha_j(X)-1} \right]$$

The coefficients β_{jr} in each α_j of Equation 2 could be measured solvind the minimum problem as.

$$\underset{\beta}{\text{maximize}} \quad \log(L(\beta|X, Y)) - \lambda [\phi(||\beta||) + (1 - \phi)(||\beta^2||)] \quad (3)$$

Being

$$\log(L(\beta|X, Y)) = \sum_{i=1}^N \log \left[\Gamma \left(\sum_{j=1}^K \alpha_j(X_{ri}) \right) \right] - \sum_{j=1}^K \log [\Gamma(\alpha_j(X_{ri}))] + \sum_{j=1}^K [1 - \alpha_j(X_{ri})] \log(Y_{ji})$$

Thus, we using a maximization procedure to find the β coefficients in equation 2.

50 Nosso modelo, $K = 3$, sendo α_{Home} o parâmetro associado a probabilidade do time mandante vencer, $\alpha_{Visitor}$ a probabilidade do time visitante vencer e α_{Tie} a probabilidade para o empate dos times.

As variáveis regressoras são apresentadas a seguir.

3.2. Base de dados

55 Foram consideradas como regressoras as habilidades encontradas no site do jogo virtual FIFA. São elas **Aceleração, Altura, Cabeceio, Carrinho, Chute de longe, Cobr. falta, Combativ., Contr. de bola, Cruzamento, Div. em pé, Dribles, Duração do contrato, Elast., GL, Finalização, Fôlego, Força, Força de Chute, Idade, Lançamento, Manejo, Marcação, Passe curto, Perna boa, Perna ruim, Peso, Pique, Classificação, Posicion. GL, Reação**
60 e **Reflexos.**

Foram coletadas informações durante o período de 2008 e 2016.

Similarmente ao *Premier League*, os dados dos disponibilizados no site da FIFA acompanham os times de cada temporada. Os dados são atualizados mais de uma vez durante semana de jogos e foram considerados os valores da versão mais atualizada de cada temporada.

65 Seja X_r^{home} o valor da variável regressora r para o time mandante do confronto e, analogamente, $X_r^{visitor}$ o valor para a mesma variável do time visitante. A equação do modelo 2 é modelada pela diferença entre a habilidade na variável r entre o time mandante e o visitante. $X_r = X_r^{home} - X_r^{visitor}$. Dessa forma, valores muito positivos para a variável X_r significam superioridade do time mandante e os negativos significam que as habilidades do time visitante são superiores aos anfitriões.

70 O gráfico do painel (a) da imagem 1 mostra que, em média, os jogos em que a medida X_r é maior do que zero, houve vitória do time mandante. Do forma similar, nos casos em que X_r foi menor do que zero o time visitante venceu. Os casos tem empate, há equilíbrio entre os times e a medida X_r é próxima, em média, de zero. Em anexo estão as medidas para as demais variáveis.

75 O gráfico do painel (b) da imagem 1 mostra a evolução da variável aceleração ao longo do período de análise para os principais times da *Premier League*. É possível perceber que todos times

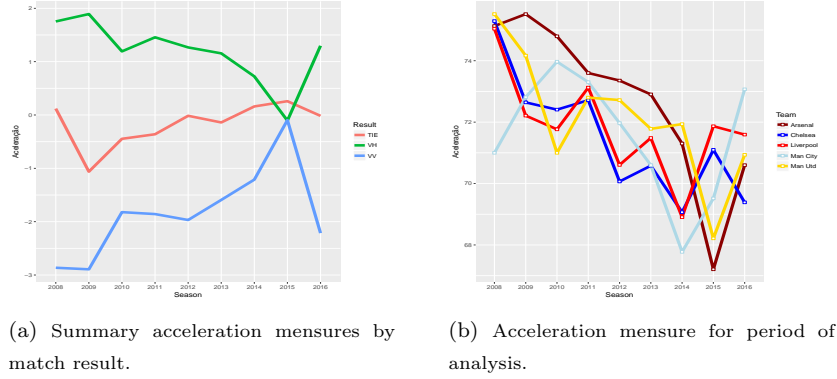


Figura 1: (a) Summary acceleration measures by match result. Tie (TIE), Victory of Home team (VH) and Victory of visitor team (VV). (b) Skill acceleration measure for period of analysis.

tiveram entre 2008 e 2014 uma diminuição do valor médio da variável aceleração. O processo de renovação de elenco do Chelsea, Liverpool e Manchester City iniciou-se em 2015. Já para Arsenal e Manchester United a renovação ocorreu no ano seguinte em 2015.

O início de renovação do Chelsea foi campeã pode ajudar a explicar o sucesso em 2015 e 2017. Uma queda dessas e outras informações das equipes inglesas podem ajudar a compreender a zebra de 2016, quando a equipe de pouca expressão Leicester foi campeã.

Pretendemos explicar e prever o resultado dos jogos da Premier League, baseado nas habilidades informadas pelo site FIFA dos jogadores de cada time.

A relação positiva é observada quando o aumento da variável regressora X_r , que representa o nível de superioridade do time mandante, contribui para a vitória do time mandante. Nesse raciocínio, é esperado nos jogos em que houve vitória do time mandante, a média dessa variável é maior do que os valores dos jogos empate, que por sua vez é maior do que os que ocorreram derrota.

$$\bar{X}_r^{VH} \geq \bar{X}_r^{Tie} \geq \bar{X}_r^{VV}$$

De maneira análoga, a relação negativa é observada quando o diminuição de X_r contribui para a derrota do time mandante. Nesse caso, é esperado que nos jogos em que houve vitória do time mandante, a média dessa variável é menor do que os valores dos jogos empate, que por sua vez é menor do que os que ocorreram derrota.

$$\bar{X}_r^{VH} \leq \bar{X}_r^{Tie} \leq \bar{X}_r^{VV}$$

Além desse resultado geral, é esperado que ao longo de **cada temporadas analisadas** esse comportamento se repita. O número de vezes em que esse resultado é observado pode sugerir o

95 nível dessa relação.

A tabela 1 apresenta essas análises para cada variável e dá indícios para o processo de seleção das variáveis relevantes para o modelo. É possível identificar, que as variáveis “chute de longe” e “controle de bola” um comportamento adequado com a relação positiva. Também é possível notar que as variáveis “idade” e “classificação na tabela” têm relação negativa com a vitória do mandante.

100 A base de dados foi dividida em:

- **base de treino:** jogos das temporadas de 2008 e 2014. A base servirá para estimar os parâmetros do modelo da equação 2.
- **base de tuning:** jogos das temporadas de 2015. A base servirá para estimar os parâmetros λ e ϕ do processo de **regularização** na equação 3.
- 105 • **base de teste:** jogos das temporadas de 2016. A base servirá para avaliar *out-of-sample* o modelo.

Os parâmetros de ϕ e λ foram avaliados por meio do *grid-search*, sendo $\phi = (0, 0.1, 0.2, 0.3, \dots, 0.9, 1)$ e $\lambda = (0, 0.5, 1, 1.5, 2, 2.5, 3)$.

4. Resultados

110 4.1. Modelo 1

O modelo 1 utiliza todas as variáveis disponíveis e utiliza medida de dissimilaridade para empate, vitória ou derrota do time visitante. A tabela 2 apresenta o resultado para o modelo e é possível perceber que o empate não é identificado corretamente pelo modelo.

A tabela 3 apresenta a simulação para o resultado da temporada de 2017/2018 baseada nos
115 parâmetros estimados.

4.2. Modelo 2

O modelo 2 adiciona medida de similaridade para modelar o empate. Isso ocorre devido o valor de X_r representar uma medida de **dissimilaridade**, isto é, quanto maior o valor X_r mais diferente são as habilidades do time mandante e visitante. Caso as habilidades dos dois times sejam iguais o

	Média			Indicador positivo				Indicador negativo			
Skill	\bar{X}_r^{VH}	\bar{X}_r^{TIE}	\bar{X}_r^{VV}	NVH	NVV	NTIE	NP	NVH	NVV	NTIE	NN
Aceleração	1.204	-0.167	-1.797	88.9	88.9	88.9	88.9	11.1	0.0	0.0	11.1
Altura	0.024	-0.022	-0.019	33.3	44.4	22.2	33.3	11.1	44.4	11.1	11.1
Cabeceio	0.431	-0.195	-0.523	66.7	66.7	55.6	66.7	0.0	22.2	0.0	0.0
Carrinho	0.574	-0.119	-0.822	77.8	77.8	66.7	77.8	0.0	11.1	0.0	0.0
Ch. de longe	1.294	-0.257	-1.863	100.0	88.9	88.9	100.0	0.0	0.0	0.0	0.0
Cobr. falta	1.316	-0.170	-1.975	88.9	88.9	77.8	88.9	0.0	0.0	0.0	0.0
Combativ.	0.457	-0.334	-0.442	77.8	55.6	44.4	77.8	11.1	11.1	11.1	11.1
Contr. bola	1.702	-0.480	-2.323	100.0	88.9	88.9	100.0	0.0	0.0	0.0	0.0
Cruzamento	1.672	-0.525	-2.235	100.0	88.9	88.9	100.0	0.0	0.0	0.0	0.0
Div. em pé	0.649	-0.142	-0.922	88.9	88.9	77.8	88.9	0.0	0.0	0.0	0.0
Dribles	1.621	-0.297	-2.355	100.0	88.9	88.9	100.0	0.0	0.0	0.0	0.0
Duração Do Contrato	0.395	-0.146	-0.508	100.0	88.9	88.9	100.0	0.0	0.0	0.0	0.0
Elast. GL	0.196	-0.023	-0.297	66.7	55.6	33.3	66.7	11.1	11.1	11.1	11.1
Finalização	0.752	-0.054	-1.167	77.8	77.8	77.8	77.8	22.2	11.1	11.1	22.2
Fôlego	0.435	-0.098	-0.616	55.6	66.7	55.6	55.6	33.3	22.2	22.2	33.3
Força	0.255	0.024	-0.434	44.4	88.9	44.4	44.4	11.1	11.1	11.1	11.1
Força chute	1.241	-0.416	-1.636	88.9	88.9	88.9	88.9	0.0	11.1	0.0	0.0
Idade	-0.277	0.057	0.396	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0
Lançamento	1.816	-0.267	-2.697	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0
Manejo	-0.215	0.171	0.195	44.4	44.4	44.4	44.4	55.6	55.6	55.6	55.6
Marcação	0.606	-0.295	-0.718	88.9	77.8	66.7	88.9	0.0	0.0	0.0	0.0
Passe curto	1.777	-0.374	-2.540	100.0	88.9	88.9	100.0	0.0	0.0	0.0	0.0
Perna boa	0.021	0.008	-0.041	44.4	44.4	33.3	44.4	44.4	22.2	22.2	44.4
Perna ruim	0.215	0.037	-0.381	66.7	100.0	66.7	66.7	0.0	0.0	0.0	0.0
Peso	0.024	-0.022	-0.019	33.3	44.4	22.2	33.3	11.1	44.4	11.1	11.1
Pique	1.124	-0.157	-1.677	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0
Classificação	-2.241	1.144	3.290	0.0	0.0	0.0	0.0	100.0	100.0	100.0	100.0
Posicion. GL	-0.223	0.154	0.224	33.3	33.3	33.3	33.3	66.7	55.6	55.6	66.7
Reação	1.634	-0.463	-2.228	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0
Reflexos	-0.235	0.222	0.182	33.3	33.3	33.3	33.3	66.7	55.6	55.6	66.7

Tabela 1: Tabela indicativa do tipo de relação entre a habilidade dos times e o resultado do jogo. Na seção positiva, a coluna VH apresenta a proporção de temporadas em que o valor médio da habilidade para os casos em que o time mandante venceu foi superior às médias dos casos de empate e derrota, indicando uma relação **positiva** entre a variável X_r (superioridade do time mandante) e vitória do mandante. Na seção negativa, essa mesma coluna apresenta a proporção de temporadas em que o valor médio da habilidade foi **inferior** às médias de empate e derrota. De maneira análoga, nesse seção essa coluna indica uma relação **negativa** entre a variável X_r . Na seção positiva, a coluna Tie apresenta a proporção de temporadas em que o valor médio da habilidade foi superior às médias de empate e derrota, indicando uma relação **positiva** entre a variável X_r . A coluna TIE apresenta a proporção de temporadas em que o valor médio da habilidade foi superior às médias de empate e derrota.

Seasons	Global	Home	Tie	Visitor
2008-2014	53.52	83.86	3.44	49.13
2015	46.32	78.98	3.74	41.38
2016	57.63	83.42	1.19	56.88

Tabela 2: Medidas do modelo. A coluna Global é a acurácia geral para o modelo. A coluna Home, Tie e Visitor apresentam os valores da acurácia para os casos de vitória, empate e derrota do time mandante, respectivamente.

Team	PLC	CCL	CEL	RET
Man Utd	16.50	49.02	14.96	1.77
Man City	16.08	48.46	14.94	1.58
Tottenham	12.90	43.01	15.79	1.98
Chelsea	10.99	37.00	14.25	3.62
Arsenal	9.32	35.04	14.77	3.83
Southampton	8.70	36.16	14.13	3.08
Newcastle	4.51	22.01	12.36	7.78
Crystal Palace	3.85	19.93	11.86	9.30
Stoke City	3.31	17.91	11.30	9.36
Everton	2.79	16.15	10.55	11.90
Leicester	1.85	10.63	8.78	16.67
Liverpool	1.77	12.71	8.93	14.09
Watford	1.69	9.36	7.82	16.87
West Brom	1.52	10.38	8.18	17.85
Bournemouth	1.52	8.61	7.53	19.35
Swansea	1.14	8.78	7.97	19.50
Burnley	0.60	4.68	5.03	29.40
West Ham	0.46	4.25	4.81	31.67
Brighton	0.40	4.31	4.06	32.00
Huddersfield	0.10	1.62	1.98	48.40

Tabela 3: Detailed table of competition where the values are probability of the simulation. Premier League Champion (PLC), Classify to UEFA Champions League (CCL), Classify to UEFA Europa League (CEL) Relegated team (RE).

120 valor X_r zero, inviabilizando a regressão do parâmetro α_{Tie} . Dessa forma, propomos uma medida de **similaridade** para regressão do parâmetro α_{Tie} relativo ao empate. Faremos:

$$X_r^{Tie} = \gamma \exp \{-\sigma X_r^2\} \quad (4)$$

Nessa estratégia, os parâmetros λ , ϕ , σ e γ são determinados utilizando a procedimento *tuning*. O modelo para α_{Tie} captará o nível de similaridade dos times. A tabela 4 apresenta o resultado do modelo, com o percentual de acerto para cada resultado incorporando a transformação 4 para a
125 medida de empate.

Seasons	Data	Global	Home	Tie	Visitor
2008-2014	Training	53.11	83.05	1.80	50.43
2015	Tuning	47.11	83.44	3.74	37.93
2016	Test	58.16	84.49	3.57	55.05

Tabela 4: Medidas do modelo que incluem o ajuste para similaridade do empate. A coluna Global é a acurácia geral para o modelo. A coluna Home, Tie e Visitor apresentam os valores da acurácia para os casos de vitória, empate e derrota do time mandante, respectivamente.

A tabela 5 apresenta a simulação para a classificação do campeonato com os parâmetros estimados.

Team	PLC	CCL	CEL	RET
Liverpool	39.00	86.00	7.00	0.00
Man City	31.00	78.00	12.00	0.00
Chelsea	13.00	62.00	17.00	0.00
Leicester	6.00	34.00	24.00	0.00
Tottenham	5.00	32.00	24.00	2.00
Man Utd	4.00	26.00	19.00	0.00
Stoke City	1.00	26.00	26.00	0.00
Everton	1.00	14.00	8.00	3.00
West Ham	0.00	9.00	13.00	10.00
West Brom	0.00	11.00	17.00	4.00
Watford	0.00	0.00	3.00	21.00
Swansea	0.00	1.00	5.00	17.00
Southampton	0.00	6.00	6.00	16.00
Newcastle	0.00	3.00	3.00	9.00
Huddersfield	0.00	0.00	0.00	64.00
Crystal Palace	0.00	1.00	5.00	21.00
Burnley	0.00	0.00	0.00	84.00
Brighton	0.00	2.00	1.00	18.00
Bournemouth	0.00	0.00	1.00	27.00
Arsenal	0.00	9.00	9.00	4.00

Tabela 5: Detailed table of competition where the values are probability of the simulation with similarity for tie. Premier League Champion (PLC), Classify to UEFA Champions League (CCL), Classify to UEFA Europa League (CEL) Relegated team (RE).

Apesar de pequena, o tratamento para a regressão do coeficiente do empate contribui para melhorar o poder de previsibilidade do modelo.

O modelo 3 restringe as variáveis baseado nas análises preliminares do modelo 1 e 2. O primeiro critério seleciona as variáveis para modelar a vitória e a derrota do time mandante. A partir da tabela 1 foram selecionados aqueles que possuem a variável NP e NN com valor superior ou igual a 80%, isso é, variáveis que possuem relação positiva ou negativa em mais de 80% das temporadas. A figura A.2 destaca as variáveis selecionadas. Com esse critério são selecionadas as variáveis Aceleração, Ch. de longe, Cobr. falta, Contr. bola, Cruzamento, Div. em pé, Dribles, Duração Do Contrato, Força chute, Idade, Lançamento, Marcação, Passe curto, Pique, Posição e Reação.

Com esse critério são selecionadas as variáveis Aceleração, Cobr. falta, Contr. bola, Cruzamento, Dribles, Lançamento, Passe curto, Pique e Reação. Todas as variáveis selecionadas para o modelo do empate estão no modelo de vitória ou derrota do time mandante.

Seasons	Data	Global	Home	Tie	Visitor
2008-2014	Trainning	46.14	98.11	3.60	0.43
2015	Tuning	42.89	96.18	9.35	1.72
2016	Test	39.47	47.59	15.48	44.04

Tabela 6: Medidas do modelo que incluem o ajuste para similaridade do empate. A coluna Global é a acurácia geral para o modelo. A coluna Home, Tie e Visitor apresentam os valores da acurácia para os casos de vitória, empate e derrota do time mandante, respectivamente.

5. Considerações finais

BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA
BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA

Team	PLC	CCL	CEL	RET
Man City	18.00	55.00	10.00	6.00
Man Utd	17.00	51.00	15.00	0.00
Arsenal	15.00	35.00	15.00	3.00
Tottenham	14.00	43.00	18.00	0.00
Liverpool	7.00	31.00	9.00	3.00
Newcastle	6.00	21.00	19.00	13.00
Chelsea	6.00	32.00	12.00	6.00
Southampton	5.00	27.00	11.00	8.00
Everton	4.00	17.00	10.00	7.00
Crystal Palace	2.00	15.00	10.00	13.00
Bournemouth	2.00	10.00	5.00	24.00
West Brom	1.00	5.00	4.00	34.00
Swansea	1.00	8.00	10.00	18.00
Stoke City	1.00	12.00	12.00	12.00
Leicester	1.00	9.00	9.00	8.00
West Ham	0.00	14.00	14.00	18.00
Watford	0.00	4.00	6.00	27.00
Huddersfield	0.00	1.00	4.00	36.00
Burnley	0.00	5.00	3.00	33.00
Brighton	0.00	5.00	4.00	31.00

Tabela 7: Detailed table of competition where the values are probability of the simulation with variable selection and similarity for tie. Premier League Champion (PLC), Classify to UEFA Champions League (CCL), Classify to UEFA Europa League (CEL) Relegated team (RE).

References

- 145 Hijazi, R.H., Jernigan, R.W., 2009. Modeling Compositional Data Using Dirichlet Regression Models. Journal of Applied Probability and Statistics .
- Kain, K.J., Logan, T.D., 2014. Are Sports Betting Markets Prediction Markets? Journal of Sports Economics doi:10.1177/1527002512437744.
- 150 Null, B., 2009. Modeling Baseball Player Ability with a Nested Dirichlet Distribution. Journal of Quantitative Analysis in Sports doi:10.2202/1559-0410.1175.
- Štrumbelj, E., Šikonja, M.R., 2010. Online bookmakers' odds as forecasts: The case of European soccer leagues. International Journal of Forecasting doi:10.1016/j.ijforecast.2009.10.005.

ApêndiceA. Selected variables

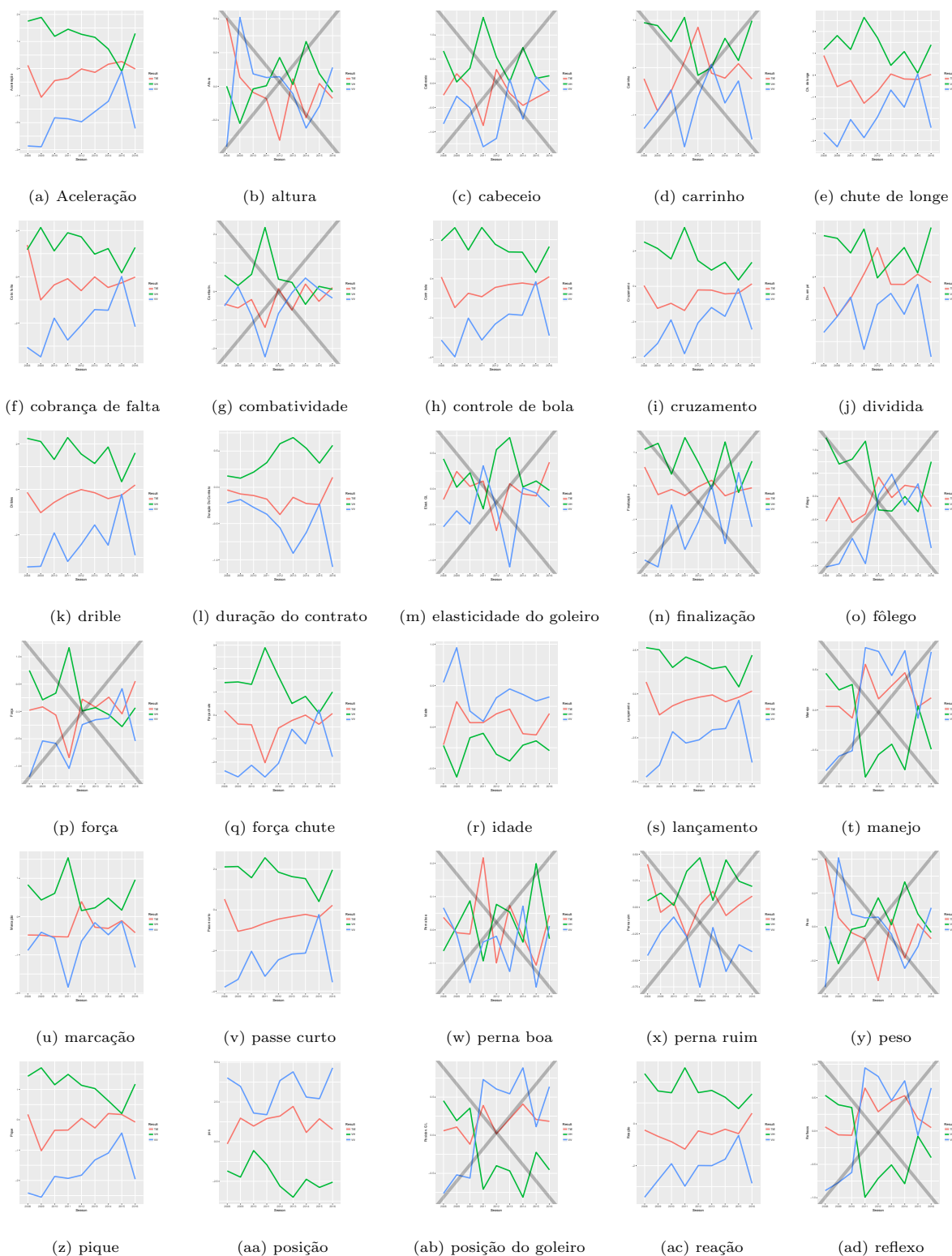


Figura A.2: Medidas resumo ao longo das temporadas para cada variável disponível. O marcador X é inserido para os gráficos das variáveis não utilizadas para o modelo, por apresentar inconsistências nas relações positivas e negativas.