Japanese Journal of Statistics and Data Science (2022) 5:885–898
https://doi.org/10.1007/s42081-022-00173-z

ORIGINAL PAPER

Data Science: Present and Future

Analysis and visualization of team performances of football games

Tsuneshi Obata[1] · Shizue Izumi[2]

Received: 28 January 2022 / Revised: 21 May 2022 / Accepted: 30 June 2022 /
Published online: 12 August 2022

Abstract

It is well known that the number of goals in a football game follows a Poisson distribution very well. Therefore, the intensity parameters of Poisson distributions, followed by the number of goals gained/lost by a team, can be regarded as the indices of the offensive/defensive performance of the team. Teams belonging to a professional football league have many games throughout the season. The performances of such teams should be considered not to be constant throughout the season. A Poisson regression model with varying coefficients is proposed to analyze and visualize the time-varying performance indices of football teams. It is also applicable to estimating the performances up to the middle of a season, and estimated performances can be used to predict the future game's outcome probabilities. These methods are demonstrated with the data of the Japanese professional football league.

Keywords Football data · Intensity parameter · Poisson regression · Prediction · Varying coefficients

1 Introduction

It is well known that the number of goals by a team in a football game follows a Poisson distribution very well (e.g., Chu, 2003). Based on it, many statistical models which assume that scores gained are generated from a Poisson distribution with a particular intensity parameter have been discussed (Izumi & Konaka, 2016; Saraiva et al., 2016;

✉ Shizue Izumi
shizue-izumi@biwako.shiga-u.ac.jp

Tsuneshi Obata
t-obata@oita-u.ac.jp

[1] Oita University, Oita, Japan

[2] Shiga University, Hikone, Japan

Koopman & Lit, 2015). The intensity parameter of the Poisson distribution can be considered as the index of team performance.

Izumi and Konaka (2016) discussed the two-stage championship held in the 2015 and 2016 seasons in Japan Professional Football League (J.League) Division 1. They proposed some probability models that determine the number of goals. Saraiva et al. (2016) proposed a model that the performances of teams and the effect of home advantage determine the number of goals and applied it to the professional football leagues in England and Brazil. However, they both assumed that teams' performances never vary through a season. Koopman and Lit (2015) proposed the model that the intensity parameter of the next game is determined stochastically from that of the previous game, like a Markov model. They treated the time discretely.

The varying coefficient models proposed by Hastie and Tibshirani (1993) are applied to the fields of medical science and social science. For example, Izumi et al. (2015, 2017) proposed the method to visualize the transition of effects of binary covariates for count data and demonstrated it with the Peace Declaration of Hiroshima and Nagasaki. The varying coefficient models can be applied to estimate the effect depending on time or spatial position as seen in the literature, including them.

Izumi and Obata (2018) discussed introducing the varying coefficient model into the analysis of football data. This paper considers the continuous time-varying transition of the team performance by introducing the varying coefficient model, based on their discussion. It can be used smoothing for counting data by taking the observed time, and the observed number of goals gained/lost as x-axis and y-axis, respectively, to extract continuous performance. We propose the visualization of the estimated transition of team performance from semi-parametric regression with linear spline basis functions.

We can also use our model for the prediction of game results. In the middle of the season, our model can estimate the performance transition function up to that time. Based on this estimated performance function, both teams' offensive/defensive performances in the next game can be predicted. The predicted performances enable us to evaluate the probabilities of the number of goals and the game outcome. Obata and Izumi (2018) considered using the estimated performances to predict the outcome of the game outcome. This paper proposes a method to predict the outcome probabilities of football games.

This paper consists as follows. Section 2 reviews the varying coefficient model. In Sect. 3, we propose the model to estimate the time transition of team performances through a season. An example of an application to the data of J.League Division 2 in the 2018 season is shown. In Sect. 4, our model is applied to estimate the performances up to the middle of the season. We propose the method for predicting the probabilities of the outcomes of the succeeding game using these estimated performances. The overall appropriateness of the results of the prediction is also discussed. Section 5 concludes this paper.

2 Varying coefficient model

Let us consider response count data $y(t_1), y(t_2), \ldots, y(t_{n-1}), y(t_n)$ observed, respectively, at n observation times $t_1 < t_2 < \cdots < t_{n-1} < t_n$. We assume that response data

$y(t_1), \ldots, y(t_n)$ can be considered as observed values of time-varying random variable $Y(t)$ following a Poisson distribution $\mathrm{Po}(\lambda(t))$, where intensity parameter $\lambda(t)$ varies through the time t. If responses depend on time-varying p covariates $a_1(t), \ldots, a_p(t)$, Poisson regression model can be written as

$$\begin{aligned} E(Y(t)|a_1(t), \ldots, a_p(t)) &= \lambda(t|a_1(t), \ldots, a_p(t)) \\ &= \exp\left[\sum_{j=1}^{p} \beta_j(t) a_j(t)\right], \end{aligned} \tag{1}$$

where $\beta_j(t)$ are the time-varying effects and called as varying coefficients. The varying coefficient model was proposed originally by Hastie and Tibshirani (1993).

If the number of observation times n is sufficiently large, linear spline functions with $r(< n-2)$ knots $\kappa_1, \ldots, \kappa_r$ can be considered as varying coefficients $\beta_j(t)$. Linear spline function is continuous polyline, and its basis are represented as

$$x(t) = (1, t, (t-\kappa_1)_+, \ldots, (t-\kappa_r)_+)^\top, \tag{2}$$

where

$$(t-\kappa)_+ = \begin{cases} (t-\kappa), & (t-\kappa > 0), \\ 0, & (\text{otherwise}). \end{cases}$$

Positions of knots are set so that each interval includes a sufficient number of data. Izumi et al. (2015) propose arranging knots so that each interval contains the same number of observations using quantile of data. Varying coefficients $\beta_j(t)$ are represented as

$$\beta_j(t) = x(t)^\top b_j, \quad j = 1, \ldots, p, \tag{3}$$

where b_j is the vector of regression coefficients.

3 The transition of team performance

3.1 Model for estimation

If the number of goals gained $Y_O^{(k)}$ by team k is assumed to follow a Poisson distribution, there exists an intensity parameter $\lambda_O^{(k)}$ such that

$$Y_O^{(k)} \sim \mathrm{Po}(\lambda_O^{(k)}).$$

These parameters may differ from team to team. So, the magnitude of $\lambda_O^{(k)}$ can be interpreted as the index of the offensive performance of team k. In the same manner, there exists $\lambda_D^{(k)}$ such that

$$Y_D^{(k)} \sim \mathrm{Po}(\lambda_D^{(k)}),$$

where $Y_D^{(k)}$ denotes the number of goals lost by team k. The magnitude of $\lambda_D^{(k)}$ can be interpreted as the index of the defensive performance of team k. To be more precise, $-\lambda_D^{(k)}$, instead of $\lambda_D^{(k)}$, should be called the defensive index because small $\lambda_D^{(k)}$ indicates high defensive performance. However, we call $\lambda_D^{(k)}$ as the index of defensive performance. We omit the superscript of team k unless misleading hereafter.

In most professional football leagues, teams play many games in a long season. It must be natural that the offensive/defensive performances of a team vary throughout a season. Therefore, it can be supposed that there exist time-varying intensity parameters $\lambda_O(t)$ and $\lambda_D(t)$ such that

$$Y_u(t) \sim \mathrm{Po}\left(\lambda_u(t)\right), \quad u = O, D,$$

where $Y_O(t)$ and $Y_D(t)$ denote the numbers of goals gained and lost by a team at the game held at time t, respectively. In this paper, we regard these time-varying offensive index $\lambda_O(t)$, defensive index $\lambda_D(t)$, or $-\lambda_D(t)$, and total index $\lambda_T(t) = \lambda_O(t) - \lambda_D(t)$ as the transitions of team performances.

To estimate these indices from actually observed numbers of goals gained and lost by a team, Poisson regression models with varying coefficients are suitable. We consider the simplest models in which we set $p = 1$ and $a_1(t) = 1$ in Eqs. (1) and (3) as the simplest model, i.e., we fit the observed number of goals gained and lost to the following model:

$$\begin{aligned} E(Y_u(t)) &= \lambda_u(t) = \exp\left[\beta_u(t)\right] \\ &= \exp\left[x(t)^\top b_u\right], \quad u = O, D. \end{aligned} \tag{4}$$

Let the number of games that each team has in a season be n, then the actual number of goals gained and lost $y_u(t_1), y_u(t_2), \ldots, y_u(t_n)$ in games held at $t_1 < t_2 < \cdots < t_n$ are observed. By fitting these observed data to the model (4), we can obtain the estimates $\hat{b}_u$ of the coefficients b_u and the estimate of the team performance indices

$$\hat{\lambda}_u(t) = \exp\left[\hat{\beta}_u(t)\right] = \exp\left[x(t)^\top \hat{b}_u\right]$$

$(u = O, D)$.

By drawing the graphs of these estimates $\hat{\lambda}_O(t)$, $-\hat{\lambda}_D(t)$, and $\hat{\lambda}_T(t)$, the transitions of team performances can be visualized. These three curves run along with the observed numbers of goals gained, the observed numbers of goals lost, and the observed goal differentials, respectively.

Table 1 The number of goals gained and lost by Oita Trinita in the 2018 season

Date (month/day)	Sec.	Days	Gained	Lost	Opponent
2/25	1	0	4	2	Tochigi
3/4	2	7	2	2	Yamagata
3/11	3	14	0	1	Okayama
3/17	4	20	0	0	Tokyo V.
(omitted)					
11/10	41	258	2	1	Kanazawa
11/17	42	265	1	1	Yamagata

3.2 Example

Let us see an example. We apply the previously mentioned model to the data of the 2018 season of J.League Division 2 (J2). J2 consists of 22 teams. All teams had two round-robin games; therefore, each played 42 games from Feb. 25 (0th day) to Nov. 17 (265th day). We use days from the season's opening day as the unit of time. Table 1 summarizes the data of the games by Oita Trinita in the 2018 season. The meanings of each column are as follows:

Date: the date when the game was held,
Sec.: the section number,
Days: the number of days since the opening day of the season,
Gained: the number of goals gained,
Lost: the number of goals lost,
Opponent: the name of the team against.

We apply our model mentioned in the previous subsection to such data. The number of games n is 42. All games are numbered as sections ("Sec." column in Table 1). Ordinarily, sections are numbered in the order of the date games are scheduled. However, games may be held against the order of the sections due to inclement weather or others. The "Days" column corresponds to a series of observation times, $t_1 = 0, t_2 = 7, \ldots, t_{42} = 265$. The observation times are ordered according to the actual time passage, even if the games are held in irregular order. The "Gained" column corresponds to the observed number of goals gained. That is $y_O(t_1) = 4, y_O(t_2) = 2, \ldots, y_O(t_{42}) = 1$. The "Lost" column corresponds to the observed number of goals lost. That is $y_D(t_1) = 2, y_D(t_2) = 2, \ldots, y_D(t_{42}) = 1$.

As for Eq. (2), we use three knots such as $\kappa_q = \frac{q}{4} \cdot 265$ $(q = 1, 2, 3)$ that each interval contains around ten games.

Figure 1 shows the estimated transition of performances of Oita Trinita. The top figure illustrates the estimated offensive index function $\hat{\lambda}_O(t)$ (solid line) with three knots (vertical bars). Solid dots express the observed numbers of goals gained. In the same way, the central illustrates the estimated defensive index function $\hat{\lambda}_D(t)$ and the observed numbers of goals lost. These are plotted by multiplying a minus one so that

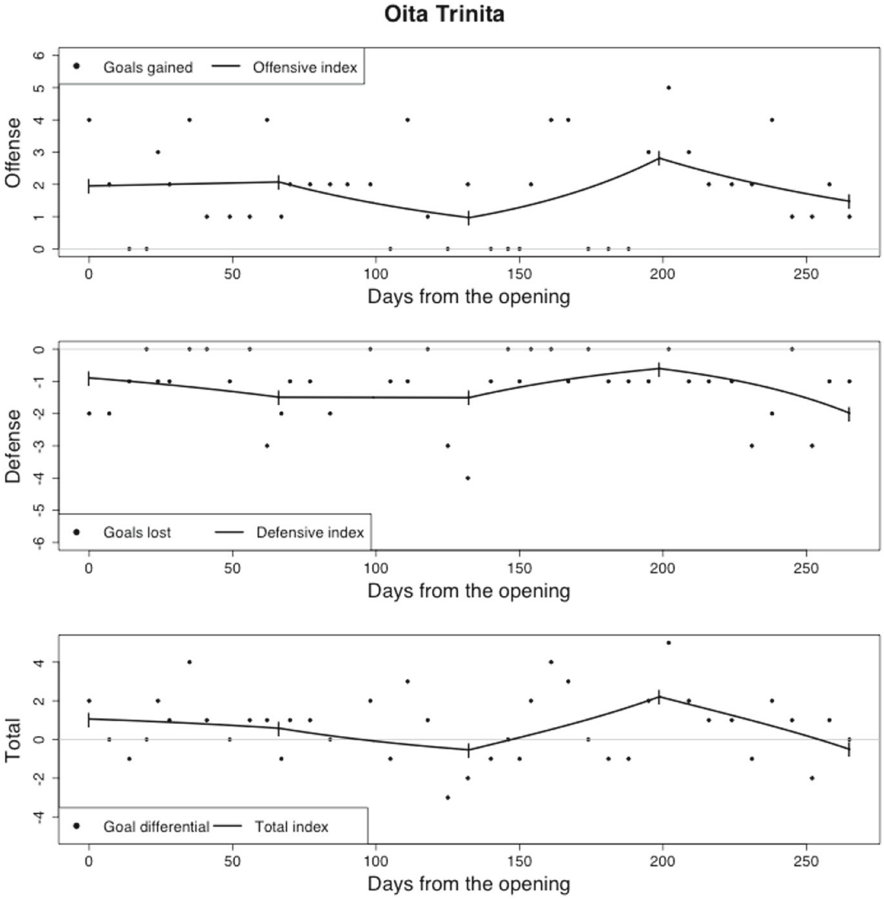

Fig. 1 Estimated transition of performances of Oita Trinita in the 2018 season

the higher value indicates a better performance. The bottom illustrates the estimated total index function $\hat{\lambda}_T(t)$ and the observed goal differentials.

4 Prediction of outcome probabilities

4.1 Prediction of outcome probabilities

In the previous section, we obtained the transitions of team performances. In this section, we propose the method to obtain the outcome probabilities of the succeeding game using a proposed model.

Let us consider the situation to predict the game's outcome—which team will win?—in which competing teams k and l at the time t^*, in advance.

Suppose that team k have finished $i^{(k)}$ games held at $t_1 < \cdots < t_{i^{(k)}}$ so far. By fitting the Poisson regression model (4) to the observed number of goals gained/lost

$y_u^{(k)}(t_1), \ldots, y_u^{(k)}(t_{i^{(k)}})$, the transitions of performances $\lambda_u^{(k)}(t)$ of the team up to the time $t_{i^{(k)}}$ can be estimated ($u = O, D$). Similarly for team l, the transitions of performances $\lambda_u^{(l)}(t)$ can be estimated ($u = O, D$). As we will see later, the allocation of the knots should be considered carefully.

The number of goals gained by team k in the next game can be supposed to follow a Poisson distribution with certain intensity $\lambda^{(k)}$. If trends of performances of team k continue, this parameter $\lambda^{(k)}$ will depend on the intensity of goals gained by team k at time t^*, $\lambda_O^{(k)}(t^*)$. However, if the defensive performance of the opponent, team l, is high, team k may not fully show their offensive performance. Conversely, if team l has low defensive performance, team k may be able to score more than expected. So that, the intensity of goals lost by team l at time t^*, $\lambda_D^{(l)}(t^*)$, also has influence to $\lambda^{(k)}$. Consequently, it is natural to suppose that the value of $\lambda^{(k)}$ can be represented as some kind of average of $\lambda_O^{(k)}(t^*)$ and $\lambda_D^{(l)}(t^*)$. Note that the smaller $\lambda_D^{(l)}(t^*)$ means the higher defensive performance of team l.

So, we assume that the number of goals gained by team k—this equals to the number of goals lost by team l—at the next game $Y^{(k)}$ follows the Poisson distribution whose intensity parameter equals a geometric mean of $\lambda_O^{(k)}(t^*)$ and $\lambda_D^{(l)}(t^*)$ as follows:

$$Y^{(k)} \sim \mathrm{Po}(\lambda^{(k)}),$$

where

$$\lambda^{(k)} = \sqrt{\lambda_O^{(k)}(t^*) \cdot \lambda_D^{(l)}(t^*)} = \exp\left[\frac{\beta_O^{(k)}(t^*) + \beta_D^{(l)}(t^*)}{2}\right].$$

The reason why we use a geometric mean is the structure of $\lambda_u(\cdot) = \exp[\beta_u(\cdot)]$. In the same way, the number of goals gained by team l at the next game $Y^{(l)}$ is assumed to be

$$Y^{(l)} \sim \mathrm{Po}(\lambda^{(l)}),$$

where

$$\lambda^{(l)} = \sqrt{\lambda_O^{(l)}(t^*) \cdot \lambda_D^{(k)}(t^*)} = \exp\left[\frac{\beta_O^{(l)}(t^*) + \beta_D^{(k)}(t^*)}{2}\right].$$

If the $Y^{(k)}$ and $Y^{(l)}$ can be considered independent, the joint distribution of $\left(Y^{(k)}, Y^{(l)}\right)$ is expressed as

$$\begin{aligned}
&\Pr\left\{Y^{(k)} = i, Y^{(l)} = j\right\} \\
&\quad = \Pr\left\{Y^{(k)} = i\right\} \cdot \Pr\left\{Y^{(l)} = j\right\}, \quad i, j = 0, 1, 2, \ldots .
\end{aligned}$$

Hence, probabilities of the outcome of the next game competing teams k and l can be predicted as

$$\begin{aligned}\Pr\{\text{team } k \text{ wins at the next game}\} &= \sum_{i>j} \Pr\left\{Y^{(k)} = i, Y^{(l)} = j\right\},\\ \Pr\{\text{ends in a draw at the next game}\} &= \sum_{i=j} \Pr\left\{Y^{(k)} = i, Y^{(l)} = j\right\},\\ \Pr\{\text{team } l \text{ wins at the next game}\} &= \sum_{i<j} \Pr\left\{Y^{(k)} = i, Y^{(l)} = j\right\}.\end{aligned}$$

4.2 Example

We demonstrate our prediction method using the 2018 season J2 data again. Remind that each team in the league plays $n = 42$ games.

Estimation of the performances in the middle of the season does not work well if each interval does not contain sufficient data. So, we change the number and the allocation of knots of linear spline according to the number of usable data. Let i^* be the number of games up to the current time. Similar to Sect. 3.2, the basis of time is set as follows with the maximum number of knots is three:

$$x(t) = \begin{cases} (1, t)^\top, & (5 \le i^* \le n/4),\\ (1, t, (t-\kappa_{11})_+)^\top, & (n/4 < i^* \le n/2),\\ (1, t, (t-\kappa_{21})_+, (t-\kappa_{22})_+)^\top, & (n/2 < i^* \le 3n/4),\\ (1, t, (t-\kappa_{31})_+, (t-\kappa_{32})_+, (t-\kappa_{33})_+)^\top, & (3n/4 < i^* \le n),\end{cases}$$

and the allocation is set as follows:

$$\begin{cases} \kappa_{11} = t_{i^*}/2, & (n/4 < i^* \le n/2),\\ \kappa_{21} = t_{i^*}/3, \quad \kappa_{22} = 2t_{i^*}/3, & (n/2 < i^* \le 3n/4),\\ \kappa_{31} = t_{i^*}/4, \quad \kappa_{32} = t_{i^*}/2, \quad \kappa_{33} = 3t_{i^*}/4, & (3n/4 < i^* \le n),\end{cases}$$

where t_{i^*} denotes the time when i^*th game held. Therefore, each interval contains more than five pieces of data.

An example of performance estimation of Oita Trinita up to the 19th game is shown in Fig. 2. The performances are estimated using linear spline with one knot ($i^* = 19$ and $t_{i^*} = 111$).

For example, we show the prediction process of the outcome of the game of Oita Trinita vs. Avispa Fukuoka held at time $t^* = 118$. It was the 20th game for both. First, the performance index functions of both teams up to the 19th game, $\hat{\lambda}_O^{(\text{Oita})}(t)$, $\hat{\lambda}_D^{(\text{Oita})}(t)$, $\hat{\lambda}_O^{(\text{Fukuoka})}(t)$ and $\hat{\lambda}_D^{(\text{Fukuoka})}(t)$ are estimated. The values of these functions at $t^* = 118$ are

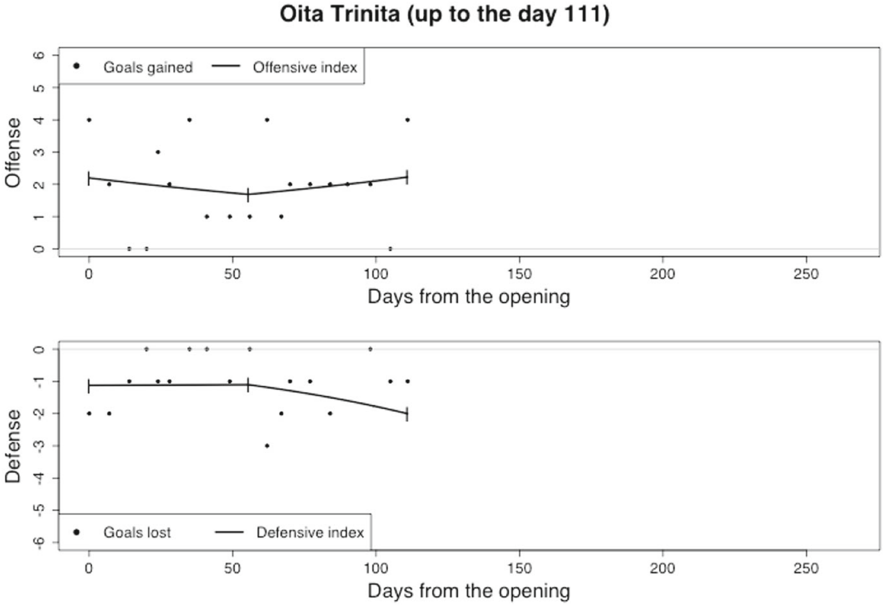

Fig. 2 Estimated transition of performances of Oita Trinita after 19 games finished

$$\hat{\lambda}_O^{(\text{Oita})}(118) = 2.304, \quad \hat{\lambda}_D^{(\text{Oita})}(118) = 2.160,$$
$$\hat{\lambda}_O^{(\text{Fukuoka})}(118) = 0.552, \quad \hat{\lambda}_D^{(\text{Fukuoka})}(118) = 0.516.$$

The numbers of goals gained by both teams, $Y^{(\text{Oita})}$ and $Y^{(\text{Fukuoka})}$, are assumed to follow the following distributions, respectively:

$$Y^{(\text{Oita})} \sim \text{Po}\left(\sqrt{2.304 \times 0.516}\right) = \text{Po}(1.094),$$
$$Y^{(\text{Fukuoka})} \sim \text{Po}\left(\sqrt{0.552 \times 2.160}\right) = \text{Po}(1.092).$$

The joint distribution of $\left(Y^{(\text{Oita})}, Y^{(\text{Fukuoka})}\right)$ is shown in Table 2. Joint probabilities of the number of goals beyond five (i.e., $i, j = 6, 7, \dots$) are omitted from the table because the values are less than 0.001. Finally, the probabilities of outcomes can be predicted as

$$\begin{aligned}&\Pr\{\text{Oita wins at the 20th game}\}\\ &\quad = \sum_{i>j} \Pr\left\{Y^{(\text{Oita})} = i, Y^{(\text{Fukuoka})} = j\right\} = 0.354,\\ &\Pr\{\text{ends in a draw at the 20th game}\}\\ &\quad = \sum_{i=j} \Pr\left\{Y^{(\text{Oita})} = i, Y^{(\text{Fukuoka})} = j\right\} = 0.292,\end{aligned}$$

Table 2 The joint distribution of the number of goals gained by Oita and Fukuoka

		Fukuoka					
		0	1	2	3	4	5
Oita	0	0.112	0.123	0.067	0.024	0.007	0.001
	1	0.123	0.134	0.073	0.027	0.007	0.002
	2	0.067	0.073	0.040	0.015	0.004	0.001
	3	0.025	0.027	0.015	0.005	0.001	0.000
	4	0.007	0.007	0.004	0.001	0.000	0.000
	5	0.001	0.002	0.001	0.000	0.000	0.000

$$\begin{aligned}&\Pr\{\text{Fukuoka wins at the 20th game}\}\\&\quad=\sum_{i<j}\Pr\left\{Y^{(\text{Oita})}=i,\,Y^{(\text{Fukuoka})}=j\right\}=0.353.\end{aligned}$$

Thus our method using the data up to the 19th game predicts that Oita Trinita has a slightly better chance of winning against Avispa Fukuoka at the 20th game. The actual result of this game was that Oita Trinita achieved a 1–0 victory over Avispa Fukuoka. This case suggests that our method may be valid. Thus, we decided to examine other cases.

We applied such a process to all games after the sixth section in the 2018 season of J2. Table 3 shows a part of the predicted probabilities of outcomes using the data up to the just previous game. The meanings of each column are as follows:

Sec.: the section number,
Home team: the home (H) team,
Away team: the home (A) team,
Prob. H wins: the predicted probability that the home team wins,
Prob. draw: the predicted probability that the game ends in a draw,
Prob. A wins: the predicted probability that the away team wins,
Actual outcome: the actual outcome of the game.

In some cases, like Chiba vs. Kyoto in Section 6, Okayama vs. Omiya in Section 42, and Kumamoto vs. Ehime in Section 42, our prediction was appropriate, and in others was not.

4.3 Overall appropriateness of our prediction

To see the overall appropriateness of our prediction method, suppose the situation of betting on the outcome of the next game using our predicted probabilities. There are three alternatives: “home win”, “draw” or “away win”, and one can bet on one of these outcomes.

We introduce the following ten betting strategies:

D1: Bet on the outcome whose predicted probability by our method is the largest.
D2: Bet on “home win” always.
D3: Bet on “draw” always.

Table 3 Predicted probabilities of outcomes in the 2018 season of J2

Sec.	Home team	Away team	Prob. H wins	Prob. draw	Prob. A wins	Actual outcome
6	Mito	Yokohama FC	0.318	0.144	0.538	Draw
6	Tochigi	Kumamoto	0.209	0.265	0.526	Home won
6	Chiba	Kyoto	0.510	0.199	0.291	**Home won**
6	Tokyo V.	Yamagata	0.316	0.272	0.412	Draw
6	Niigata	Tokushima	0.126	0.225	0.648	Home won
(omitted)						
42	Okayama	Omiya	0.236	0.328	0.435	**Away won**
42	Sanuki	Kyoto	0.478	0.316	0.206	Away won
42	Kumamoto	Ehime	0.360	0.323	0.317	**Home won**

Underline indicates the largest among three predicted probabilities
Boldface indicates that our prediction is consistent with the actual outcome

Table 4 Proportions of hitting the right answers by D1–D6

D1 (our prob.)	D2 (always home)	D3 (always draw)
0.430	0.383	0.248
D4 (always away)	D5 (recent 3)	D6 (recent 5)
0.369	0.396	0.428

D4: Bet on “away win” always.
D5: Bet on the outcome whose probability is the largest based on the recent three games.
D6: Bet on the outcome whose probability is the largest based on the recent five games.
S1: Try a random sampling of three outcomes with predicted probabilities by our method and bet on the result.
S2: Try a random sampling of three outcomes with equal probabilities and bet on the result.
S3: Try a random sampling of three outcomes with probabilities based on the recent three games and bet on the result.
S4: Try a random sampling of three outcomes with probabilities based on the recent five games and bet on the result.

“Probabilities based on the recent games” used in D5, D6, S3, and S4 are determined by the following procedure:

1. Calculate the arithmetic means of the numbers of goals gained by both teams, $\bar{\lambda}^{\mathrm{H}}_{O}$ and $\bar{\lambda}^{\mathrm{A}}_{O}$, of the recent three/five games. Similarly, calculate the arithmetic means of the numbers of goals lost by both teams, $\bar{\lambda}^{\mathrm{H}}_{D}$ and $\bar{\lambda}^{\mathrm{A}}_{D}$, of the recent three/five games.
2. Calculate the geometric mean of the number of goals gained by the home team and the number of goals lost by the away team, $\bar{\lambda}^{\mathrm{H}} = \sqrt{\bar{\lambda}^{\mathrm{H}}_{O} \cdot \bar{\lambda}^{\mathrm{A}}_{D}}$. Similarly, calculate the geometric mean of the number of goals gained by the away team and the number of goals lost by the home team, $\bar{\lambda}^{\mathrm{A}} = \sqrt{\bar{\lambda}^{\mathrm{A}}_{O} \cdot \bar{\lambda}^{\mathrm{H}}_{D}}$.
3. Obtain the probabilities of three outcomes from independent joint distribution of $\mathrm{Po}(\bar{\lambda}^{\mathrm{H}})$ and $\mathrm{Po}(\bar{\lambda}^{\mathrm{A}})$, similarly as mentioned in the previous subsection.

Six (D1–D6) are deterministic, and four (S1–S4) are stochastic. The strategies D1 and S1 use our predicted probabilities.

We applied these strategies to all games after the sixth section in the 2018 J2 season previously mentioned.

For deterministic strategies, Table 4 shows the proportions of hitting the correct answers. Strategy D1, which uses our prediction, seems better than others, though not significantly.

For stochastic strategies, we had 10,000 trials each. Figure 3 shows the distributions of the proportions of hitting the correct answers by four strategies. The medians are 0.3686, 0.3342, 0.3612 and 0.3612 for S1, S2, S3 and S4, respectively. Strategy S1, which uses our prediction, seems slightly better than the others.

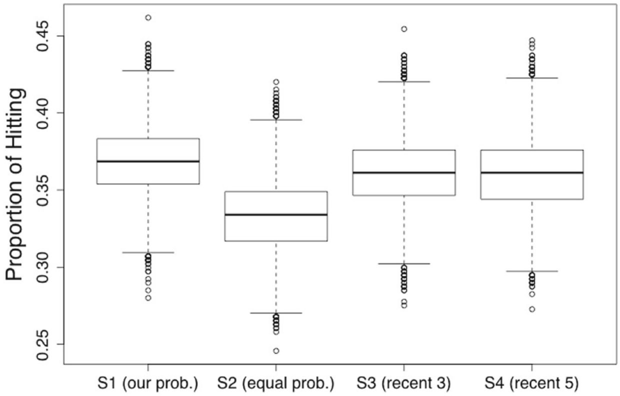

Fig. 3 Distributions of the proportion of hitting the right answers by S1–S4

5 Conclusions

If the number of goals in a football game follows a Poisson distribution, we can regard the intensity parameter of the distribution as an indicator of the team's performance. Besides, the intensity can be supposed to vary over time.

In this paper, at first, a method for estimating the time-varying performance of football teams was proposed. We introduced the varying coefficient model into sports data analysis and demonstrated that it could be used to estimate the time transition of team performances. For the estimation of semi-parametric varying coefficient, the mixed effect model can be used as in Izumi et al. (2015, 2017). However, we used the simple generalized linear model here because it is necessary to carefully interpret the meaning of estimated coefficients for the mixed effect model.

Our performance estimation is also available for the data up to the middle of the season. We proposed a method to predict the probabilities of the next game's outcome based on estimated performance up to the previous game. It was applied to the one-year data of the Japanese professional football league, and we examined the appropriateness of our method.

We used the simplest varying coefficient model, which does not contain other covariates than time. The effect of home advantage in a football game is discussed in several works of literature (e.g., Goumas, 2014; Koopman & Lit, 2015; Saraiva et al., 2016). Other factors, e.g., weather or the number of audiences, which may influence the football team's performance, can also be considered. The addition of these environment covariates is worth considering.

The combination of players and team formation are also essential factors in the performances of football teams (Hirotsu & Ueda, 2015). Considering these factors may improve the results in the phases of performance estimation or outcome prediction.

In the prediction phase, we assume independence between the number of goals of both teams in a game. However, the appropriateness of this assumption should be discussed more carefully.

Finally, careful investigation of errors and confidence intervals of estimation and prediction remains for future research.

Acknowledgements The authors would like to thank the reviewers and the editors for helpful comments that improved the paper. This study was carried out under the ISM Cooperative Research Programs (2019-ISMCRP-4401 and 2020-ISMCRP-4601). This work was supported partly by JSPS KAKENHI Grant Numbers JP17K00047 and JP19H04355.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Chu, S. (2003). Using soccer goals to motivate the Poisson process. *INFORMS Transactions on Education, 3*(2), 64–70. https://doi.org/10.1287/ited.3.2.64

Goumas, C. (2014). Home advantage in Australian soccer. *Journal of Science and Medicine in Sport, 17*, 119–123. https://doi.org/10.1016/j.jsams.2013.02.014

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B, 55*, 757–796.

Hirotsu, N., & Ueda, T. (2015). Measuring efficiency of a set of players of a soccer team and differentiating players' performances by their reference frequency. In *Proceedings of the fifth international conference on mathematics in sport*, 66–71.

Izumi, S., & Obata, T. (2018). Visualization of time transition of team performance in soccer league (in Japanese). *Communications of the Operations Research Society of Japan, 63*(10), 628–634.

Izumi, S., Satoh, K., & Kawano, N. (2015). Statistical classification and visualization based on varying coefficients model for longitudinal text data (in Japanese). *Bulletin of the Computational Statistics of Japan, 28*(1), 81–92.

Izumi, S., Tonda, T., Kawano, N., & Satoh, K. (2017). Estimating and visualizing the time-varying effects of a binary covariate on longitudinal big text data. *The International Journal of Networked and Distributed Computing, 5*(4), 243–253. https://doi.org/10.2991/ijndc.2017.5.4.6

Izumi, T., & Konaka, E. (2016). Statistical analysis of two-stage and postseason format of J1 football league (in Japanese). *Transactions of the Operations Research Society of Japan, 59*, 21–37.

Koopman, S., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A, 178*, 167–186.

Obata, T., & Izumi, S. (2018). Predicting the outcome probability of the soccer match based on the estimation of time-varying team performance. In *Proceedings of the Asia Pacific industrial engineering and management systems conference*, 1–5.

Saraiva, E. R., Suzuki, A. K., Filho, C. A. O., & Louzada, F. (2016). Predicting football scores via Poisson regression model: applications to the National Football League. *Communications for Statistical Applications and Methods, 23*(4), 297–319. https://doi.org/10.5351/CSAM.2016.23.4.297