IMA Journal of Management Mathematics (2009) **20**, 121−131
doi:10.1093/imaman/dpn027
Advance Access publication on October 30, 2008

A random point process model for the score in sport matches

PETR VOLF†
Institute of Information Theory and Automation, Pod vodarenkou vezi 8, 18208 Prague, Czech Republic

[Received on 20 August 2007; accepted on 28 September 2008]

A sequence of goals scored during sport match is modelled as a realization of two dependent random point processes. It is assumed that the scoring intensity of each team has several components depending on time or on factors describing the teams and other conditions of the match. This dependence is modelled with the aid of a semi-parametric multiplicative regression model of intensity. A method of model evaluation is presented and demonstrated on a real data set. Prediction obtained from the model via the Monte Carlo simulation is compared with real results.

Keywords: sport statistics; random point process; scoring intensity; Cox's regression model; football.

1. Introduction

Poisson processes and their generalizations are used frequently for the probabilistic modelling of occurrence of certain events in time. Main characteristics of such a process is its intensity. Influence of other factors, covariates, is as a rule expressed via a regression model. In the present contribution, a random point process model is applied to the modelling of score development during a match (like for instance the ice hockey or football). A main advantage of dynamic point process models is their ability to estimate the team performance during the match period. It can be a valid source of information for the team management. Dynamic model also offers a possibility of 'on-line' prediction, i.e. the prediction of future match development from the information on the match 'history' and its actual state.

The paper is organized as follows. First, static Poisson models used commonly for the modelling of matches results are recalled. Then, a random point process model of scoring is introduced. It will generalize to some extent the model of Dixon & Robinson (1998). In our formulation, the intensity of scoring of each team consists of the product of two parts: a time-dependent non-parametric baseline intensity and a regression part dependent on factors like the rival team defence strength and the match state. The procedure of model evaluation is described, then it is applied to the analysis of real data from the football World Championship 2006. The model is also used for random generation of artificial results. Finally, a generalization with time-dependent defence parameters is considered.

2. Models of score in sport statistics

A basic model presented in Maher (1982) assumes that the numbers of goals scored by home and away teams in any particular game are independent Poisson variables: in a match where a home team i plays against an away team j, let X_{ij} and Y_{ij} be the numbers of goals scored by the home and away sides,

†Email: volf@utia.cas.cz

respectively. Then,

$$\begin{aligned} X_{ij} &\sim \text{Poisson}(\lambda_1 = \alpha_i \beta_j \gamma), \\ Y_{ij} &\sim \text{Poisson}(\lambda_2 = \alpha_j \beta_i), \end{aligned} \tag{1}$$

where X_{ij} and Y_{ij} are independent, $\alpha_i, \beta_i > 0$, α_i measure the attack strength of team i and β_i measure its defence ability (the smaller β_i, the better defence). Finally, $\gamma > 0$ is a parameter characterizing the effect of home field.

Such a model has at least two weak aspects. First, it is known from experience that the frequencies of certain results do not correspond to model (1), for instance, in football the draws 0:0, 1:1 and also results 1:0, 0:1 are more frequent than predicted by the model. Hence, an improvement uses the 'inflated' version of Poisson model which is the mixture of (1) with certain fixed additional probability of those more frequent results. This fact is discussed for instance in Dixon & Coles (1997). Further improvement leads to the bivariate Poisson model with dependent components. For the score-modelling purposes, such a model has been utilized for instance in Karlis & Ntzoufras (2003).

Though the models based on Poisson distribution yield the match score, they are used mostly just for the prediction of the winner or draw result. As underlined also in Dixon & Robinson (1998), the prediction of exact score is not reliable sufficiently, there are always several different results with comparably high probabilities. Another set of models, namely the trinomial classification models, have been designed for direct evaluation of probabilities of the victory (home, away) or the draw. They can utilize for instance the logistic regression models or other classification techniques.

The models described above do not contain any component describing temporal aspects of intensities, their possible development and variation. First, the performance of teams (i.e. their parameters) can change during each season. It can be incorporated by updating the parameters, giving more weight to recent than to older results, as in Dixon & Coles (1997). Further, Rue & Salvesen (2000) introduced a dynamic autoregressive model for parameters innovation during the season and considered the Bayes approach to solution (with the Markov chain Monte Carlo computations). In the same context, Crowder *et al.* (2002) used the state-space modelling technique, namely a normal approximation leading to a variant of Kalman filter method.

Another type of dynamic models, studied, e.g. in Dixon & Robinson (1998), describes the variation of the scoring intensity or at least of certain parameters during the match. It leads then to the concept of random point process and to the regression models of time-varying intensity commonly used in statistical survival analysis. This is the type of model we shall consider in the rest of our paper.

There exist also other attempts to take into account the match development. For instance, Croxson & Reade (2007) have developed a model based on bivariate Poisson distribution which adapts its parameters to actual match state. On this basis, the probabilities of final results are updated during the match. Again, though the bivariate Poisson model yields the whole probability distribution of final score, the authors analyse just probabilities of home, away win or draw.

Dixon and Robinson consider a piecewise-constant intensity which changes after the 'birth' of a new goal, so that it depends on actual state of the match. Intensity is parametrized for each team separately. This part is then multiplied by corresponding attack and defence parameters. In their most complex model VI, they add an intensity component depending directly on time, as a linear (or other parametric) trend function. Formally, the resulting intensity model can be written as

$$\lambda(t) = a \cdot b(t) + \xi \cdot t,$$

where a is a basic part and $b(t)$ is a piecewise-constant intensity of birth process. The last term is the common trend of changes of intensity during the match ($\xi \geqslant 0$).

The model we shall present here consists also of two parts, a regression part and a baseline intensity. However, in our setting, the baseline intensity is a non-parametric function reflecting the team's scoring ability. The defence strength of the rival and actual match state then influence the regression part of intensity. At least in the first stage of the analysis, it has a sense to visualize the variation of the team attack strength as a function of time. Then, in a further step, it is naturally possible to use a proper parametric function suggested by the shape of obtained non-parametric estimate. Thus, our model differs from the model of Dixon & Robinson (1998) and is also more general due to its semi-parametric form.

3. Random point processes in time

Let us now recall briefly random point process models for time sequences of events. In a homogeneous Poisson process, the events occur with a constant intensity. In a non-homogeneous variant, the intensity is a non-negative, bounded, measurable function $\lambda(t)$, $t \geqslant 0$. More generally, the intensity can depend on some explanatory variables, covariates, such a dependence is modelled via regression models.

The values of covariates can again be given by an observed random process depending on time. Let us denote it by $Z(t)$. Then, the regression model for intensity assumes that the random point process behaviour is governed by a (bounded and smooth, say) hazard function $h(t, z)$ from $[0, T] \times \mathscr{X}$ to $[0, \infty)$, where $\mathscr{X}$ is the domain of values of $Z(t)$. The intensity of point process is then

$$\lambda(t) = h(t, Z(t)),$$

so that it is actually a random process too. In order to make this setting tractable from the point of view of mathematical theory, it is assumed that the intensities at t (and the process $Z(t)$ too) are predictable and depend just on the history of system before t (are adapted to a proper set of σ-algebras, a filtration). Corresponding theory as well as the methodology of statistical analysis is collected in many papers and monographs devoted to statistical survival analysis, e.g. in Andersen *et al.* (1993).

3.1 *Examples of regression models*

The idea to separate a common hazard rate from the influence of covariates led to the multiplicative model, called also the proportional hazard model,

$$h(t, z) = h_0(t) \cdot \exp(b(z)).$$

Function $h_0(t)$ is the baseline hazard function and $b(z)$ is the regression (response) function. If the response function is parametrized, we obtain semi-parametric Cox's model. Its most popular form assumes that $h(t, z) = h_0(t) \exp(\beta z)$.

Alternatives are for instance the Aalen's additive regression model or the accelerated time model used frequently in reliability analysis. In the present paper, the Cox's model is utilized, hence we shall employ 'maximum partial likelihood' estimators of parameters β and the Breslow–Crowley estimator of the increments of cumulated baseline hazard function $H_0(t) = \int_0^t h_0(s)\mathrm{d}s$. Function $h_0(t)$ is then obtained by kernel smoothing of these increments.

4. Process of score development

In the match of team i playing against team j, let $\lambda_{0i}(t)$ be the attack intensity part of team i and α_j the defence parameter of j. Further, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ are the regression parameters expressing the influence of different factors, as the actual state of score, the power play in ice hockey, one missing player (red card) in football, the advantage of home field, etc. The state of these factors will be described by an m-dimensional vector of corresponding indicators $\mathbf{z}_{ij}(t)$, some of them may be switched on or off during the match. Then,

$$\lambda_{ij}(t) = \lambda_{0i}(t)\mathrm{e}^{\alpha_j}\,\mathrm{e}^{\boldsymbol{\beta}\mathbf{z}_{ij}(t)}, \quad \lambda_{ji}(t) = \lambda_{0j}(t)\mathrm{e}^{\alpha_i}\,\mathrm{e}^{\boldsymbol{\beta}\mathbf{z}_{ji}(t)}, \tag{2}$$

are the intensities of scoring of team i against team j (the first expression) and vice versa, at time t.

It is seen that expressions (2) have a form of Cox's model. In the follow-up, we shall take $a_i = \mathrm{e}^{\alpha_i}$ and treat it as a heterogeneity component specific for the defence of team i (compare so-called frailty models in survival analysis). Such an approach is convenient from the computational point of view. One step of global iteration, which can sometimes converge rather slowly, is then divided into two alternating simpler steps. Note also that we actually deal with two interacting processes, their interaction is modelled with the aid of shared components and covariates (here just indicators).

4.1 *Computation procedure*

In this part, we shall recall (and adapt to our case) the method of evaluation of Cox's model components. It is described elsewhere, for instance, also in Andersen *et al.* (1993). The data—record of each match—should contain the times of goals and other important time moments of covariate changes, so that at each time we know actual covariate values. Then, we can construct the likelihood function. Its logarithm is

$$L = \sum_i \sum_j \left\{ \int_0^T (\log \lambda_{0i}(t) + \log a_j + \boldsymbol{\beta}\mathbf{z}_{ij}(t))\mathrm{d}N_{ij}(t) - \int_0^T \lambda_{0i}(t) a_j \exp(\boldsymbol{\beta}\mathbf{z}_{ij}(t))\mathrm{d}t \right\}.$$

The indices are over all matches, $i \neq j$, so that each match is recorded twice, from the point of view of both teams. Indices are repeated if there are more matches of the same teams. T is the match period (here 90 min) and $N_{ij}(t)$ is the counting process of goals of team i against team j: $\mathrm{d}N_{ij}(t) = 1$ at moments of goals, $\mathrm{d}N_{ij}(t) \equiv 0$ otherwise. Hence, the first integral is just a finite sum.

Let us denote $\mathrm{d}\Lambda_{0i}(t) = \lambda_{0i}(t)\mathrm{d}t$ the increments of cumulative baseline intensities, and formally compute, by putting $\partial L/\partial \mathrm{d}\Lambda_{0i}(t) = 0$, the Breslow–Crowley estimate of these increments (for fixed i and t):

$$\mathrm{d}\hat{\Lambda}_{0i}(t) = \frac{\sum_{j\neq i} \mathrm{d}N_{ij}(t)}{\sum_{k\neq i} a_k \exp(\boldsymbol{\beta}\mathbf{z}_{ik}(t))}. \tag{3}$$

Again, $\hat{\Lambda}_{0i}(t)$ are stepwise functions with steps $\mathrm{d}\hat{\Lambda}_{0i}(t) > 0$ at the times of goals scored by team i. Further, by putting $\partial L/\partial a_j = 0$, we obtain the estimate of defence parameter a_j:

$$\hat{a}_j = \frac{\sum_{i\neq j} N_{ij}(T)}{\sum_{k\neq j} \int_0^T \mathrm{d}\Lambda_{0k}(t) \exp(\boldsymbol{\beta}\mathbf{z}_{kj}(t))}. \tag{4}$$

The estimate compares the number of goals obtained by team j in all matches (in the numerator) with the sum of cumulated intensities of obtaining the goal (again in all matches) if the defence parameter is set to one. The integral in the denominator is again in fact a finite sum.

Finally, parameters $\boldsymbol{\beta}$ are estimated separately from the so-called partial likelihood. It can be obtained also when estimators (3) are inserted back to L, so that in this case the partial likelihood equals to the profile likelihood. Namely, the logarithm of partial likelihood is now

$$L_p = \int_0^T \sum_i \sum_j \log \frac{\exp(\boldsymbol{\beta} \boldsymbol{z}_{ij}(t))}{\sum_k \sum_l a_l \exp(\boldsymbol{\beta} \boldsymbol{z}_{kl}(t))} \mathrm{d}N_{ij}(t), \tag{5}$$

again with $i \neq j$ and $k \neq l$. The estimate of parameters $\boldsymbol{\beta}$, i.e. the solution of a set of equations $\partial L_p / \partial \beta_r = 0$ ($\boldsymbol{\beta}$ can be multidimensional), is found as a rule by the Newton–Raphson algorithm. Hence, iterated estimation of all model components proceeds in the following way:

1. Set all initial $\hat{a}_j = 1$.
2. Compute $\hat{\boldsymbol{\beta}}$ from (5) by several iterations of the Newton–Raphson algorithm.
3. Compute increments $\mathrm{d}\hat{\Lambda}_{0i}(t)$ from (3).
4. Compute new $\hat{a}_j$ from (4). Normalize them in order to ensure their uniqueness.

For instance, we kept the mean of $\hat{a}_j$s equal to 1.

Steps 2–4 are repeated till convergence. In the following example, the convergence was rather fast, 10–20 cycles sufficed. In Step 2, just five iterations of the Newton–Raphson algorithm were enough. Final shapes of scoring intensities are obtained by kernel smoothing,

$$\hat{\lambda}_{0i}(t) = \frac{1}{d} \int_0^\infty W\left(\frac{t-s}{d}\right) \mathrm{d}\hat{\Lambda}_{0i}(s).$$

A frequent choice of the kernel function is the standard Gauss density, $W(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, d is a conveniently selected smoothing parameter.

5. Example

The objective of this example is to show the method applicability and the form of obtained outputs. We apply model (2) to a rather small data from the football World Championship in Germany 2006.

We were interested in the analysis of the performance of eight teams participating in quarterfinals. We took into account all their matches (together 40), including their results in groups and eight-finals. That is why we introduced also ‘imaginary’ teams: number nine represented all four unsuccessful eight-finals rivals and numbers 10, 11 and 12 all other teams met in first three rounds in the groups. From one point of view, such an aggregation might bias the estimation; on the other side, the data contained too small information on those remaining 24 participants that we preferred to collect these teams to such a small set of ‘average’ teams.

Let us first summarize certain basic statistics. For instance, Fig. 1(a) shows the distribution of times of all goals in a histogram form. It has to be taken into account that the conclusions based on rather small data are not reliable, still the graph shows certain non-uniformity of distribution, with rather large number of goals achieved in first 10 min, and (what is more common in other studies) a higher frequency in final periods of matches (including ‘extra-time’ goals assigned to the last minute). Prolonged times (2×15 min) of several matches were not taken into account. Together, 86 goals were scored in regular period of 90 min. Red card was given 16 times, weakened teams obtained nine goals after such a punishment.

Figure 1(b) shows the distribution of times to the first goal (i.e. the duration of state 0:0) and Fig. 1(c) shows the time between the first and second goals, shape of its distribution is more similar to (expected) exponential form.

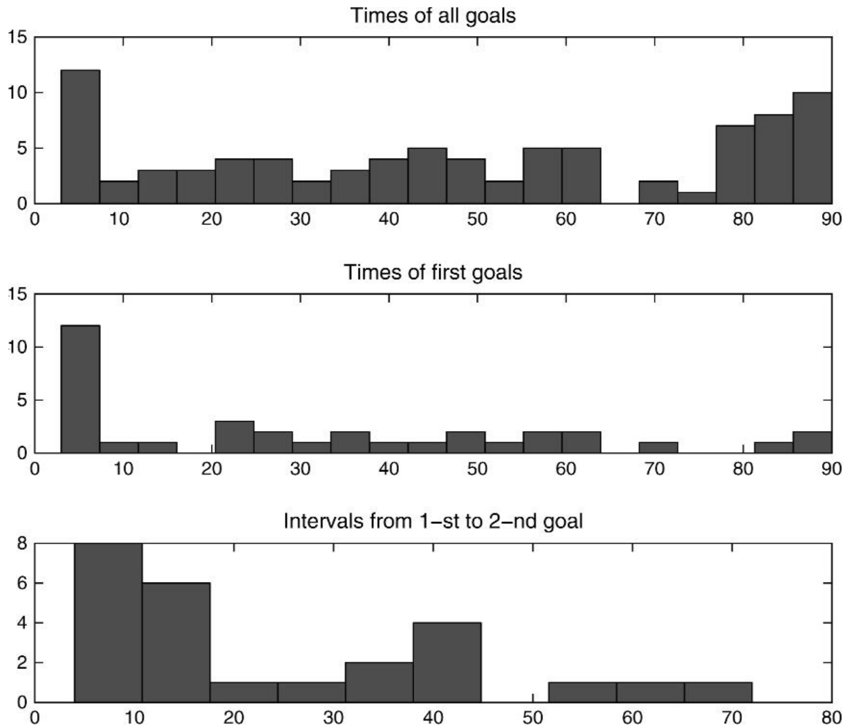

FIG. 1. Histograms of distributions of scoring times: (a) all goals, (b) first goals and (c) times from first to second goal.

After such an initial analysis, model (2) has been utilized for the analysis of matches of leading eight teams and of four others representing the rest of teams. We selected the following factors—covariates (from the point of view of team i, $i, j = 1, \ldots, 12$):

$z_{ij1}(t) = 1$, when the rival had one player missing (red card), $z_{ij1}(t) = 0$, otherwise,
$z_{ij2}(t) = 1$, if actual score was positive, $z_{ij2}(t) = 0$ otherwise,
$z_{ij3}(t) = 1$, if actual score was negative, $z_{ij3}(t) = 0$ otherwise.

Thus, the 'pure' model without covariates corresponded to the intensity in balanced score state, without red card. Consequently, the following functions and parameters should be estimated (we preferred the 'frailty' form of defence parameters a_j):

$$\lambda_{0j}(t), \quad a_j, \quad j = 1, \ldots, 12, \quad \beta_r, \quad r = 1, 2, 3.$$

As the number of analysed data was small compared to the number of unknown components, we could not afford to select more different covariates. Still the confidence of some parameters was rather vague. Table 1 contains estimated parameters and their 95% confidence intervals (approximate, because based on asymptotic normality). Parameters a_j for $j = 1, \ldots, 8$ correspond to the following teams (first four are in the final order): Italy, France, Germany, Portugal, Argentina, Ukraine, England, Brazil and parameters $\beta_1, \ldots, \beta_3$ then to indicators $z_1(t), \ldots, z_3(t)$, respectively. Figure 2 displays the scoring intensities $\lambda_{0j}(t)$ (kernel smoothed from the Breslow–Crowley estimate) for first four teams. Certain

TABLE 1 *Estimated parameters and approximate* 95% *confidence intervals*

β_1	0.5045 (−0.3398, 1.3488)		
β_2	−0.4395 (−1.0031, 0.1240)		
β_3	0.8243 (0.0713, 1.5773)		
a_1	0.2630 (0, 0.6276)	a_7	0.5069 (0, 1.2094)
a_2	0.4120 (0, 0.8783)	a_8	0.5481 (0, 1.3077)
a_3	0.5686 (0.0114, 1.1258)	a_9	1.1695 (0.5078, 1.8312)
a_4	0.6899 (0.0852, 1.2946)	a_{10}	1.2148 (0.4969, 1.9327)
a_5	0.4950 (0, 1.0552)	a_{11}	1.8848 (1.0373, 2.7323)
a_6	2.7090 (0.7022, 4.7158)	a_{12}	1.5383 (0.7846, 2.2921)

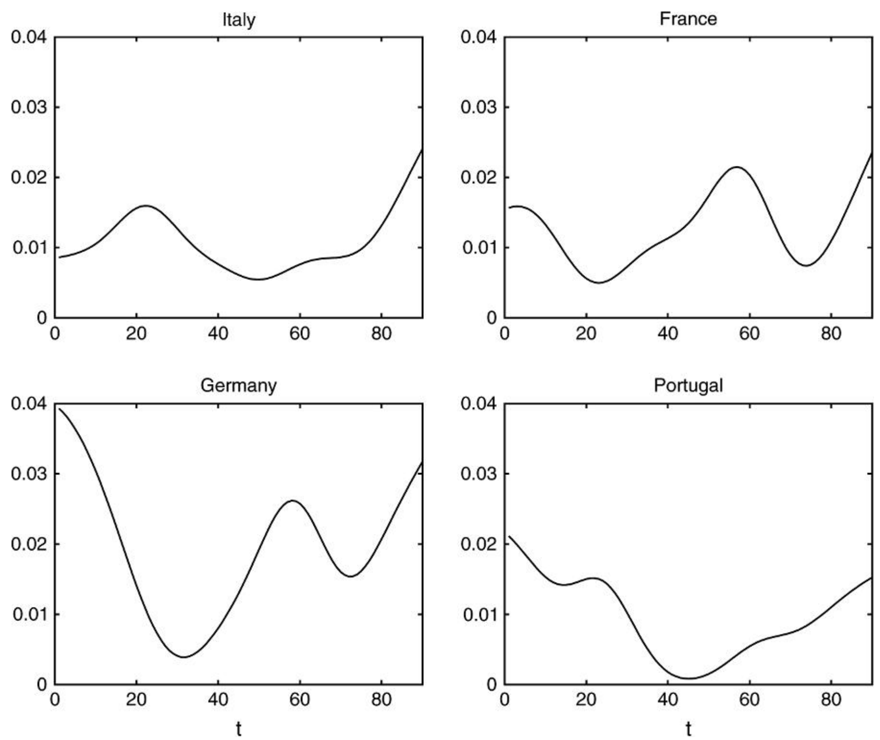

FIG. 2. Estimated attack intensities of first four teams.

differences of teams attack performance can be traced from their shapes and also from the area below the curves.

The influence of a covariate is taken as significant when the confidence interval of corresponding parameter β_r does not contain zero. It is here the case of β_3 only, suggesting that the effort of loosing team increases. With regards to defence parameters a_j, they are taken as significant when their confidence interval does not contain one, hence here it concerns just the teams number 1, 2 and 11. Again, confidence intervals are rather wide, though they vary and still yield some information on the defence strength of teams.

5.1 *Randomly generated results*

An important advantage of studied model is connected with the possibility to simulate a 'virtual' match course, using estimated intensities and parameters. A part of covariates depends on actual match state and is derived throughout the match development, others (e.g. the red-card punishment) are actually also random point processes running in parallel with the process of scoring (and can also depend on it). Its intensity has to be estimated, and then such a covariate process has to be simulated simultaneously with the process of the match. Thus, we deal with a scheme of several processes influencing each other.

We have employed this method to repeat 'artificially' the matches of the World Championship 2006, starting from quarterfinals and assuming the same pairs of teams. For simplicity, as there were 16 red cards in 40 matches, the red-card occurrence was modelled by a Poisson process with constant intensity $\lambda^* = 16/(90 \times 2 \times 40) \approx 0.002$. Thousand simulations of each match were performed. The final order of teams was the same as in reality, except that simulations preferred England to Portugal (their real match ended without goals and the Portugal team passed on after penalties). Except the final score of each match, the prediction procedure generated also the times of all goals.

By another set of simulations, we checked the influence of increased number of data to confidence of estimates. Figure 3 shows the cumulated baseline intensity $\Lambda_0(t) = \int_0^t \lambda_0(s)\mathrm{d}s$ estimated for the English team, first from small real data (actually just five matches) and then from simulated data (together 22 matches with different championship participants). Naturally, as simulated data were generated with the help of the model evaluated from real cases, the shapes are rather similar. Simultaneously, it is seen how the width of confidence band decreases with increased amount of data (the plots contain approximated

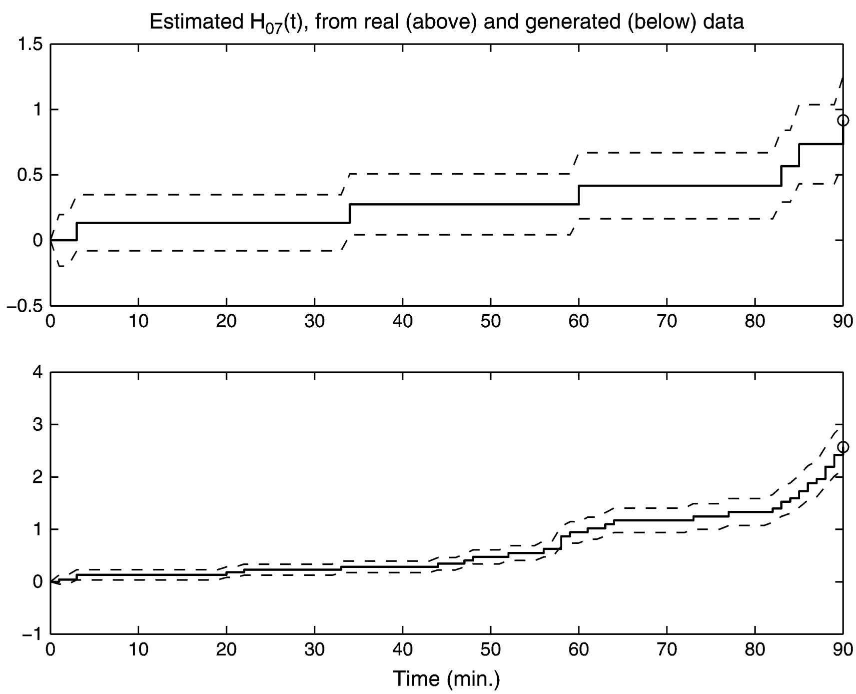

FIG. 3. Estimates of the cumulated attack rate $\Lambda_{07}(t)$ for team 7, with approximate 95% confidence bands, above from real data and below from artificial data.

95% confidence bands). As regards the defence parameters, for instance, the estimate of the English team parameter from the same 22 simulated matches was $a_7 = 0.4678$ (± 0.2051) (compare it with the value in Table 1).

In the next part of example, we shall show how the dynamic model can be used for the prediction of match future development conditionally on its actual state. However, let us first present one possible model generalization.

6. Time-dependent defence parameters

A variant of Cox's model with time-dependent parameters has been proposed by several authors in the nineties, an overview can again be found in Andersen *et al.* (1993). The estimate of variability of regression parameter is in a standard setting based on a piecewise-constant (i.e. histogram-like) approximation. The proof of consistency of such an estimation uses the idea that the width of histogram intervals is adapted properly to the number of data.

In the example presented here, we again preferred a piecewise-constant model of transformed parameters $a_j(t) = \exp(\alpha_j(t))$ because it led to a more comfortable (and stable) computations. First, the time interval $(0, T)$ was divided into several intervals and piecewise-constant estimates of $a_j(t)$ were obtained. It was repeated several times for different divisions. Finally, for each j, estimated $a_j(t)$ were averaged and then smoothed by a Gauss kernel function. Resulting curves for four selected teams are presented in Fig. 4.

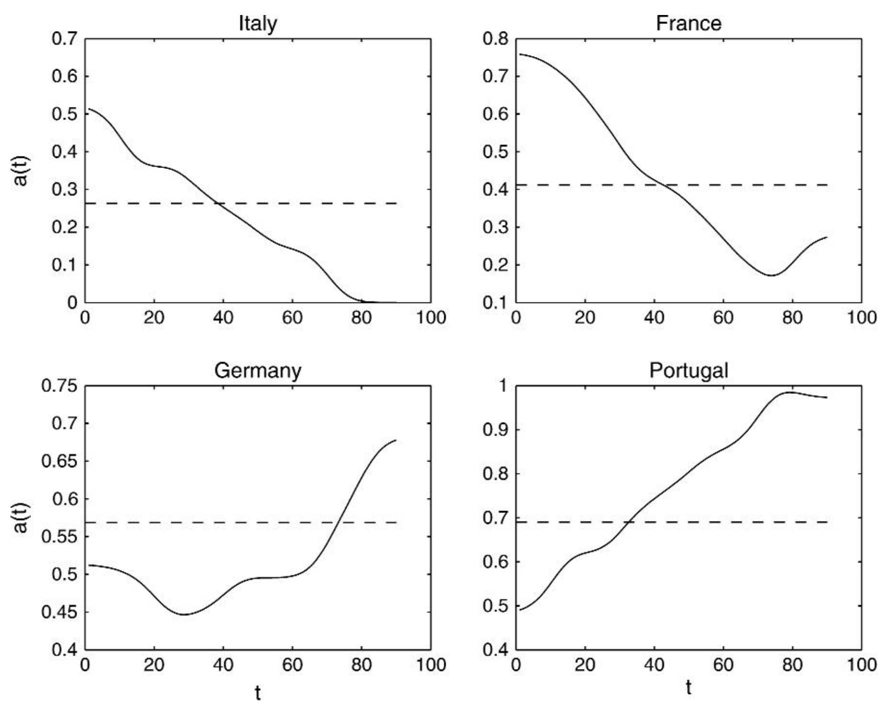

FIG. 4. Defence parameters estimated as time varying, for teams 1–4. Constant lines correspond to values of constant parameters from Table 1.

TABLE 2 *Proportions of predicted final results conditioned by the state of match at time T_1*

T_1 (min)	Score at T_1	Italy–France proportion of results	T_1 (min)	Score at T_1	Italy–Portugal proportion of results
0	0:0	3:5:2	0	0:0	5, 5:4:0, 5
7	0:1	2:5:3	10	0:1	4, 5:4, 5:1
30	0:1	1:4:5	30	0:1	3, 5:5:1, 5
45	0:1	1:3:6	45	0:1	2, 5:5, 5:2
60	0:1	0, 5:3:6, 5	60	0:1	2:5:3
7	1:0	6:3:1	80	0:1	1:4:5

6.1 *Prediction conditioned by actual match state*

Let us imagine that we have observed a match until a time T_1 and we wish to model its future development. Hence, we start to generate two new scoring processes (for both teams) with intensities

$$\lambda_{ij}(t) = \lambda_{0i}(t)a_j(t)\mathrm{e}^{\boldsymbol{\beta}\mathbf{z}_{ij}(t)}, \quad \lambda_{ji}(t) = \lambda_{0j}(t)a_i(t)\mathrm{e}^{\boldsymbol{\beta}\mathbf{z}_{ji}(t)},$$

with time running from T_1 to T and with actual states of covariates (they reflect also the state of score). We considered different times T_1 and different states at this time. For each case, 1000 possible match continuations were generated. The left side of Table 2 displays results of several such examples for the match Italy–France. In real match, France scored first at 7th min and Italy levelled score at 19th min. The table shows the rounded proportions of victories, draws and losses from generated 1000 matches (from the point of view of Italy). We see that while the non-conditioned prediction preferred Italy slightly (with proportion 3:5:2), the fact that France scored first changed the probabilities and, naturally, the chance of Italy to level (or even to win) decreased with increasing time T_1. The right part of table shows similar results for the match between Italy and Portugal (they did not meet during the championship). Results reflect quite different shapes of attack and defence functions of both teams. In this case, the model preferred Italy even when the half-time state was 0:1 for Portugal. Note also rather large (more or less realistic) proportion of draws in both examples.

7. Conclusion

The objective of the present study was to demonstrate advantages of random point process model for the score development in a sport match. The results show that even a small data study can reveal interesting patterns of teams characteristics and can be used for the analysis of team performance and then to appropriate measures. Moreover, the analysis is quite interesting also from the statistical methodology point of view. Naturally, a reliable analysis and prediction in the framework of a model with many parameters and several non-parametric components need much larger learning data.

Use of large data set would enable also the testing of model fit, either by statistical procedures developed for this purpose in the survival analysis field or by comparing predictions with real results. This is what remains to be done. Nevertheless, as the model generalizes standard Poisson models, we believe that its predictive ability is at least comparable with others. And, as underlined above, the main benefit consists, first, in the time-dependent form of model components and, further, in the possibility of match course prediction, even conditionally on actual match state.

Funding

Grant Agency of Academy of Sciences, Czech Republic (IAA101120604).

References

ANDERSEN, P. K., BORGAN, O., GILL, R. D. & KEIDING, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer.

CROWDER, M., DIXON, M. J., LEDFORD, A. & ROBINSON, M. E. (2002) Dynamic modelling and prediction of English football league matches. *Statistician*, **51**, 157–168.

CROXSON, K. & READE, J. (2007) Goal arrival in soccer: a high frequency study of efficiency in betting pricing. *Proceedings of the First International Conference on Mathematics in Sport* (D. F. Percy, P. A. Scarf & C. L. Robinson eds). Manchester: Institute of Mathematics and Applications, pp. 50–54.

DIXON, M. J. & COLES, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Appl. Stat.*, **46**, 265–280.

DIXON, M. J. & ROBINSON, M. E. (1998) A birth process model for association football matches. *Statistician*, **47**, 523–538.

KARLIS, D. & NTZOUFRAS, I. (2003) Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc. Ser. D Stat.*, **52**, 381–394.

MAHER, M. J. (1982) Modelling association football scores. *Stat. Neerl.*, **36**, 109–118.

RUE, H. & SALVESEN, O. (2000) Prediction and retrospective analysis of soccer matches in a league. *Statistician*, **49**, 399–418.