Seasonal Linear Predictivity in National Football Championships

Giuseppe Jurman

Abstract
Predicting the results of sport matches and competitions is a growing research field, benefiting from the increasing amount of available data and novel data analytics techniques. Excellent forecasts can be achieved by advanced statistical and machine learning methods applied to detailed historical data, especially in very popular sports such as football (soccer). Here, we show that despite the large number of confounding factors, the results of a football team in longer competitions (e.g., a national league) follow a basically linear trend that is also useful for predictive purposes. In support of this claim, we present a set of experiments of linear regression compared to alternative approaches on a database collecting the yearly results of 746 teams playing in 22 divisions spanning up to five different levels from 11 countries, in 25 football seasons, for a total of 181,160 matches grouped in 9386 seasonal time series.

Keywords: football; analytics; linear models; championships

Introduction
Predicting sport results in the last few years has ceased being just an art for initiated specialists to enter the realm of data analytics, thus providing a further support to the claim of considering as science many aspects of several sports. In particular, interest in forecasting the results of sporting competitions has grown in the last few years, essentially because of two key factors: the arising need for more reliable predictive models by betting agencies and the increasing number of available sources collecting data at different levels of detail. However, the predictability of results is still a debated issue, mainly because of the random effects affecting the outcome of a match, with football (soccer) as a major example. Clearly, structural effects have an even larger impact: as noted by Parasich, the average precision (53%) reached by bookmakers in forecasting the outcome of a match (home win, draw, away win) seems not so positive anymore if you consider that the home team wins 46% of matches, and thus the null constant model always predicting a home win achieves 46% precision.

Many algorithms from statistics and machine learning have recently been used to overcome such randomness bias in order to achieve good predictive performance, which have been applied to data catching diverse aspects of the game, with different historical spans, or at various levels of detail, even publicly available online as infotainment resources. Generalized linear or polynomial models and logistic or probit regressions have been used in the literature since the mid-2000s, using as variables points or goals or even adding economic parameters. More recent are statistical approaches based on Bayesian or Poissonian predictors, or even Weibull counts, where individual player’s performance is also included as a model covariate. Further statistical approaches involving for instance Markov chain Monte Carol, hierarchical models, or moving averages have been published, indicating that a shared agreement on a grounded modeling is still far from being acknowledged. More recently, machine learning models have become a major trend in the field, and all the best-known algorithms have
appeared on the football forecasting arena (e.g., nearest neighbours, Gaussian and Poisson processes, Random Forest, Support Vector Machines, MultiLayer Perceptron just to name a few\textsuperscript{63–64}) to even deep learning approaches.\textsuperscript{65–67} Finally, complex networks strategies based on team structure\textsuperscript{68–71} can provide complementary insights on the game not covered by more classical methodologies, aiming for instance at ranking teams\textsuperscript{72} or at assessing players’ rating and success.\textsuperscript{73,74} In general, when powerful learning methods and/or a substantial wealth of training data are used—and social network data are playing an increasingly crucial role\textsuperscript{32,33,75,76}—the predictive accuracy that can be reached is excellent, and the occurring randomness is effectively dealt with.

In this article, we want to demonstrate that despite the existing randomness and other confounding factors, there are situations where sporting results are driven by very simple (e.g., linear) trends, and these trends can be captured by basic techniques and a limited amount of training data. In detail, the philosophy driving the research is the ambition of filling a gap in the literature of the predictive models in soccer. In fact, very little can be found about simple baselines derived by analyzing a restricted number of fundamental features. In particular, there is no reference method evaluating the dynamics of earned points throughout a season by only using the time series of points itself: this is the niche where this article positions itself, also showing that the dynamics is fundamentally linear, and establishing at the same time a base reference for more complex models.

As in Rue and Salvesen,\textsuperscript{51} we focus on a longer competition such as a national league, and we show the outcome of forecasting the last part of a season by using only the results of the initial portion of the campaign (in Heuer and Rubner,\textsuperscript{26} the authors restrict their attention to the last 17 matches of the season). Here, we restrict our analysis to national football (soccer) championships and the simplest possible (predictive) statistical technique (i.e., linear regression as in Goddard\textsuperscript{25} and Rocha et al.\textsuperscript{30}), also compared to polynomial, autoregressive integrated moving average (ARIMA),\textsuperscript{77} and exponential smoothing state space models.\textsuperscript{78} We also compare to models from the Pythagorean Expectation family,\textsuperscript{79} a class of algorithms that have gained interest in the last few years in various team sports, originating from the Pythagorean Theorem of Baseball by James\textsuperscript{80} and later improved. Goals for and goals against are used as point predictors, with parameters fitted using a Weibull distribution. Note that linear regression has already been used to forecast future league points, using as predictors some economic indicators such as turnover, profit/loss before tax, net debt, interest owed on any debt, and the club’s wage bill.\textsuperscript{24} In particular, we want to assess to what extent such a simple approach used only on the current season results, without any historical data, can be effectively used to predict the behaviour of a team in the final portion of a tournament in terms of both the total number of earned points and the final ranking in the championship table.

Analysis

Data description

Data were extracted from the Football-Data repository,\textsuperscript{81} and they include the results of all matches for 513 European national championships over the 25-year range 1993/94–2017/18. In detail, data for 22 divisions of 11 countries at five different levels are studied, giving a total of 9386 series for 746 unique teams. Championships grouped by league and country are shown in Figure 1.

For our purposes, all 9386 time series are described by the independent variable rounds and by the dependent variable points, keeping track of the accumulated points gained by a team during the rounds of a season-long campaign, as shown in Figure 2.

Methods

All models were computed in the R environment\textsuperscript{82} using the packages \texttt{stats} for linear/quadratic/cubic and \texttt{stats} for the ARIMA and the exponential smoothing state space (ETS) models. Confidence intervals were computed via the Student’s bootstrap procedure\textsuperscript{83,84} using the version described in Davison and Hinkley\textsuperscript{85} and implemented in the \texttt{boot.ci} function of the \texttt{boot} R package.

In detail, let $T$ be a team participating in a league whose season consists of $n$ rounds, and let $T_i$ be the number of points earned by $T$ after the $i$th round, so that $T_n$ is the total number of points at the end of season. Let $t_i$ be an integer between 1 and $n-1$, and let $L_T^{t_i}$ be a model trained on $(1, T_1), \ldots, (n-t_i, T_{n-t_i})$. Define $T_n = [L_T^{t_i}(n)]$ as the estimated number of total points earned by $T$ as the largest integer smaller than the extrapolation of $L_T^{t_i}$ computed on the point $n$. In Figure 3, an example is shown of the linear modeling of Schalke 04 season in the Bundesliga 2013/14, where the final number of earned points is predicted for $t_i = 10$.

Finally, quantitative comparison between tournament standings (predicted and actual) is computed by mean of
Spearman’s rank correlation coefficient $\rho^{86}$ and by the total absolute displacement $d^{87}$ defined as the normalized sum of the differences between rankings (see Appendix A for mathematical details and examples on the metric $d$).

**Results**

In what follows, we will estimate the total number of points earned by a team by mean of a linear model trained on the first $n - t_i$ matches of the seasons, for several values of $t_i$, for $n$ the total number of matches in the season. Furthermore, we will derive, for each championship, the estimated final league table to be compared to the actual standing.

**Simulations**

As a synthetic benchmark, $10^4$ series are randomly generated with $n = 38$, $T_i = T_{i-1} + \xi$ with $T_0 = 0$, and $\xi$
FIG. 2. Time series of the points earned by Juventus FC (black), AS Roma (red), and Internazionale FC (blue) during the 2014/15 Serie A campaign. The $x$-axis shows the 38 matchdays, and the $y$-axis shows the accumulated points. Color images are available online.

FIG. 3. Points earned by Schalke 04 in the Bundesliga 2013/14 season (T, black square) and their approximation (circles) through a linear model (gray line) trained on the first 24 rounds (blue filled circles) and extrapolated on the last 10 rounds (P, white and red circles), highlighted in the yellow box. In the bottom-right yellow table, the comparison between the real points (T) and the predicted points (P) on the last 10 rounds. Color images are available online.
equal to 0, 1, or 3 with a probability of 1/3. Then $T_n = \lfloor L^n(n) \rfloor$ is computed together with CI for $L$ the linear, the quadratic, the cubic, the ARIMA, or the ETS model, for $t = 1, \ldots, 19$. Both the ARIMA and the ETS models are automatically optimized by the default R call, and the results are reported in Table 1. The ARIMA model is consistently the best, but the linear model represents a solid and simpler alternative not needing any optimization, while the quadratic and the cubic models have poor performances, especially when the training set becomes smaller: when estimating the final number of points using the first 28 rounds (out of 38), the linear, ARIMA, and ETS models show an average error of 4 points, while the quadratic and the cubic models have an error of 7 and 14 points, respectively. Thus, the linear trend is a good approximation of the null model with random results. The critical feature that supports the fair performance of the linear model in the longitudinal data of teams’ yearly campings is the fact that the successive increments $\xi$ are non-negative and small. In fact, allowing $\xi$ to take larger values quickly worsens the fit of the linear model, as shown from the three examples reported in Table 2.

The above experiment was replicated with different sets of values for $n$, namely $f_0, 1, 2, 3, 4, 5, 6, 7$, and in all three cases the average error of the linear model was larger than the one resulting from the true setting $f_0, 1, 3$.

Table 1. Prediction error of the linear (L), quadratic (Q), cubic (C), ARIMA (A), and ETS (E) models on $10^4$ simulated time series on $n = 38$ rounds

<table>
<thead>
<tr>
<th>n - $t_s$</th>
<th>$L_l$</th>
<th>$L_m$</th>
<th>$L_u$</th>
<th>$Q_l$</th>
<th>$Q_m$</th>
<th>$Q_u$</th>
<th>$C_l$</th>
<th>$C_m$</th>
<th>$C_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>6.76</td>
<td>6.85</td>
<td>6.95</td>
<td>18.94</td>
<td>19.21</td>
<td>19.48</td>
<td>74.54</td>
<td>75.50</td>
<td>76.60</td>
</tr>
<tr>
<td>20</td>
<td>6.38</td>
<td>6.47</td>
<td>6.57</td>
<td>17.02</td>
<td>17.30</td>
<td>17.54</td>
<td>60.78</td>
<td>61.71</td>
<td>62.53</td>
</tr>
<tr>
<td>21</td>
<td>5.99</td>
<td>6.08</td>
<td>6.18</td>
<td>15.04</td>
<td>15.28</td>
<td>15.50</td>
<td>50.08</td>
<td>50.81</td>
<td>51.52</td>
</tr>
<tr>
<td>22</td>
<td>5.63</td>
<td>5.92</td>
<td>6.01</td>
<td>13.36</td>
<td>13.54</td>
<td>13.73</td>
<td>41.95</td>
<td>42.58</td>
<td>43.22</td>
</tr>
<tr>
<td>23</td>
<td>5.27</td>
<td>5.55</td>
<td>5.43</td>
<td>12.04</td>
<td>12.28</td>
<td>12.50</td>
<td>35.01</td>
<td>35.54</td>
<td>36.05</td>
</tr>
<tr>
<td>24</td>
<td>4.99</td>
<td>5.06</td>
<td>5.14</td>
<td>10.75</td>
<td>10.91</td>
<td>11.07</td>
<td>28.89</td>
<td>29.31</td>
<td>29.74</td>
</tr>
<tr>
<td>26</td>
<td>4.54</td>
<td>4.60</td>
<td>4.67</td>
<td>8.60</td>
<td>8.73</td>
<td>8.86</td>
<td>20.17</td>
<td>20.45</td>
<td>20.78</td>
</tr>
<tr>
<td>27</td>
<td>4.29</td>
<td>4.35</td>
<td>4.42</td>
<td>7.63</td>
<td>7.73</td>
<td>7.84</td>
<td>18.00</td>
<td>17.06</td>
<td>17.30</td>
</tr>
<tr>
<td>28</td>
<td>4.09</td>
<td>4.15</td>
<td>4.21</td>
<td>6.92</td>
<td>7.02</td>
<td>7.11</td>
<td>15.78</td>
<td>15.98</td>
<td>16.19</td>
</tr>
<tr>
<td>30</td>
<td>3.67</td>
<td>3.73</td>
<td>3.79</td>
<td>5.54</td>
<td>5.62</td>
<td>5.70</td>
<td>13.97</td>
<td>14.02</td>
<td>14.07</td>
</tr>
<tr>
<td>31</td>
<td>3.39</td>
<td>3.44</td>
<td>3.49</td>
<td>4.90</td>
<td>4.98</td>
<td>5.06</td>
<td>12.61</td>
<td>12.71</td>
<td>12.81</td>
</tr>
<tr>
<td>32</td>
<td>3.18</td>
<td>3.23</td>
<td>3.28</td>
<td>4.35</td>
<td>4.41</td>
<td>4.47</td>
<td>11.31</td>
<td>11.41</td>
<td>11.51</td>
</tr>
<tr>
<td>33</td>
<td>3.00</td>
<td>3.04</td>
<td>3.09</td>
<td>3.80</td>
<td>3.86</td>
<td>3.92</td>
<td>9.96</td>
<td>10.02</td>
<td>10.08</td>
</tr>
<tr>
<td>34</td>
<td>2.76</td>
<td>2.80</td>
<td>2.84</td>
<td>3.29</td>
<td>3.34</td>
<td>3.39</td>
<td>8.74</td>
<td>8.84</td>
<td>8.94</td>
</tr>
<tr>
<td>35</td>
<td>2.56</td>
<td>2.60</td>
<td>2.64</td>
<td>2.87</td>
<td>2.91</td>
<td>2.95</td>
<td>7.71</td>
<td>7.81</td>
<td>7.91</td>
</tr>
<tr>
<td>36</td>
<td>2.33</td>
<td>2.37</td>
<td>2.40</td>
<td>2.48</td>
<td>2.52</td>
<td>2.56</td>
<td>6.80</td>
<td>6.90</td>
<td>7.00</td>
</tr>
<tr>
<td>37</td>
<td>2.10</td>
<td>2.14</td>
<td>2.18</td>
<td>2.48</td>
<td>2.52</td>
<td>2.56</td>
<td>5.94</td>
<td>6.04</td>
<td>6.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n - $t_s$</th>
<th>$E_l$</th>
<th>$E_m$</th>
<th>$E_u$</th>
<th>$A_l$</th>
<th>$A_m$</th>
<th>$A_u$</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>8.34</td>
<td>8.48</td>
<td>8.62</td>
<td>6.80</td>
<td>6.91</td>
<td>7.03</td>
</tr>
<tr>
<td>20</td>
<td>7.91</td>
<td>8.03</td>
<td>8.16</td>
<td>6.20</td>
<td>6.30</td>
<td>6.40</td>
</tr>
<tr>
<td>21</td>
<td>7.23</td>
<td>7.34</td>
<td>7.47</td>
<td>5.83</td>
<td>5.92</td>
<td>6.01</td>
</tr>
<tr>
<td>22</td>
<td>6.79</td>
<td>6.90</td>
<td>7.01</td>
<td>5.53</td>
<td>5.62</td>
<td>5.70</td>
</tr>
<tr>
<td>23</td>
<td>6.43</td>
<td>6.54</td>
<td>6.64</td>
<td>5.26</td>
<td>5.35</td>
<td>5.43</td>
</tr>
<tr>
<td>24</td>
<td>5.98</td>
<td>6.08</td>
<td>6.18</td>
<td>5.00</td>
<td>5.07</td>
<td>5.15</td>
</tr>
<tr>
<td>25</td>
<td>5.48</td>
<td>5.57</td>
<td>5.66</td>
<td>4.64</td>
<td>4.71</td>
<td>4.78</td>
</tr>
<tr>
<td>26</td>
<td>5.11</td>
<td>5.20</td>
<td>5.29</td>
<td>4.31</td>
<td>4.37</td>
<td>4.45</td>
</tr>
<tr>
<td>27</td>
<td>4.77</td>
<td>4.84</td>
<td>4.92</td>
<td>4.08</td>
<td>4.14</td>
<td>4.21</td>
</tr>
<tr>
<td>28</td>
<td>4.41</td>
<td>4.48</td>
<td>4.55</td>
<td>3.81</td>
<td>3.86</td>
<td>3.92</td>
</tr>
<tr>
<td>29</td>
<td>4.00</td>
<td>4.06</td>
<td>4.13</td>
<td>3.55</td>
<td>3.60</td>
<td>3.66</td>
</tr>
<tr>
<td>30</td>
<td>3.71</td>
<td>3.76</td>
<td>3.82</td>
<td>3.28</td>
<td>3.33</td>
<td>3.38</td>
</tr>
<tr>
<td>31</td>
<td>3.33</td>
<td>3.38</td>
<td>3.43</td>
<td>3.02</td>
<td>3.06</td>
<td>3.11</td>
</tr>
<tr>
<td>32</td>
<td>3.00</td>
<td>3.04</td>
<td>3.09</td>
<td>2.76</td>
<td>2.80</td>
<td>2.84</td>
</tr>
<tr>
<td>33</td>
<td>2.60</td>
<td>2.64</td>
<td>2.69</td>
<td>2.43</td>
<td>2.46</td>
<td>2.50</td>
</tr>
<tr>
<td>34</td>
<td>2.29</td>
<td>2.32</td>
<td>2.35</td>
<td>2.15</td>
<td>2.18</td>
<td>2.21</td>
</tr>
<tr>
<td>35</td>
<td>1.91</td>
<td>1.94</td>
<td>1.97</td>
<td>1.82</td>
<td>1.85</td>
<td>1.88</td>
</tr>
<tr>
<td>36</td>
<td>1.54</td>
<td>1.56</td>
<td>1.59</td>
<td>1.49</td>
<td>1.51</td>
<td>1.53</td>
</tr>
<tr>
<td>37</td>
<td>1.10</td>
<td>1.11</td>
<td>1.13</td>
<td>1.08</td>
<td>1.09</td>
<td>1.11</td>
</tr>
</tbody>
</table>

For each model, the mean error (m) and the lower (l) and upper (u) Student’s bootstrap confidence intervals (CIs) are reported. ARIMA, autoregressive integrated moving average; ETS, exponential smoothing state space.
Team performance prediction

For the 9386 seasonal time series $T$, we estimate $\tilde{T}_n$ for $t_s = 1, \ldots, 20$, with linear, quadratic, cubic, ARIMA, and ETS models. As a comparative baseline, we also interpolate $\tilde{T}_n$ as $\tilde{T}_n = \frac{T_n}{s}$. In Table 3, we list the mean and the confidence intervals for $|\tilde{T}_n - T_n|$ for the five models.

Results for the six models are all mutually statistically different, with $p$-values $< 10^{-16}$. While quadratic and cubic models have very poor predictivity (with the cubic model performing even worse than the interpolation baseline), linear, ETS, and ARIMA models produce similar and better results, supporting the claim of an overall linear trend in points evolution. These results imply that overall, the seasonal trend is not linear, since the nonlinear ARIMA model fits better (and Pearson’s correlation between the ARIMA and linear models is 0.76), but the simpler linear model represents a valid approximation. In Figure 4, we show two cases (both in Dutch Eredivisie) where the nonlinearity instead is particularly evident, namely AZ Alkmaar in 2004/05 and Nijmegen in 2007/08. In both cases, the trend for the last part of the season is very different from the initial part.

In what follows, as a representative case, we set $t_s = 10$, that is, for each series, we use all the rounds except for the last 10 as the training set, and we predict the final figure of earned points. Overall, when $t_s = 10$, the average prediction error of the final number of points is 4.43 for the linear model, 3.93 for the ARIMA model, and 4.62 for the ETS model, but the boxplots for the three models almost overlap, as shown in Figure 5. In this case, the average prediction error for the interpolation baseline model is 14.24, with a confidence interval of 14.07–14.40. Consistency of the results with those obtained in the general situation indicates that choosing $t_s = 10$ as the representative case is a meaningful choice (the full linear prediction results for all teams, leagues, and seasons can be found in the Supplementary Data).

We also add here the comparison with algorithms from a well-known family of algorithms, variants of the original Pythagorean Expectation (PE) method, originally developed for baseball in James,80 recently adapted for football by Hamilton79 and then further optimized by several authors.88–91 The basic PE model has the form

$$\text{earned points} = \frac{GF^\lambda + GA^\lambda}{\lambda \cdot \text{#rounds}}$$

and thus is different from the other model, since it is not a model for the dynamics of the time series.
gained points. In the original article,\(^7\) \(c_1 = c_2 = c_3\) were estimated on the available data by a Weibull distribution and \(\lambda = 1\), while the following variants have different values for the three \(c\) parameters and \(\lambda\) is \(> 1\) to account for ties and 3-point wins. In what follows, we tried several different sets of parameters, and we report the results for the best performing values \(c_1 = 1.5\), \(c_2 = 1.08\), \(c_3 = 1.13\), and \(\lambda = 2.31\). Although all these models are known to achieve reasonably good diagnostic results across different leagues and seasons in modeling the team’s points, they do not perform similarly well in prediction: for \(t_s = 10\), the average prediction error of the final number of points is 6.93.

Consider the (linear) predictivity \((|T_n - \tilde{T}_n|)\) for \(S\) of 258 teams that are present for 220/25 seasons in the available data. Figure 6 shows the histogram of the average differences between prediction and actual values. The set of values \(|T_n - \tilde{T}_n|\) for \(S\) is Gaussian-like, with a range of 3.00–6.33 (min and max corresponding to Lorient and Dundee Utd., respectively) and a mean and median of approximately 4.4. Smaller values indicate more linear behavior of a team throughout all the considered seasons, while larger values mark the presence of one or more seasons where the sequence of results had a nonlinear trend. On the same task, again the PE algorithm has poorer performances, with an average error 7.39. In Table 4, the values \(|T_n - \tilde{T}_n|\) are listed for the top10 Union of European Football Associations ranking teams (current standing at June 2018). Among a number of teams such as Bayern Munich, Juventus, and Manchester City whose linear trend is quite consistent through all the considered seasons
Barcelona’s case emerges. Barcelona’s high value $|T_n - T_n| = 5.76$ is due to a number of seasons (1993/94, 2002/03, 2005/06, 2003/04, 2007/08, 2008/09, and 2010/11) where the seasonal trend was markedly nonlinear, mostly because the last matches followed a very different pattern from the initial part of the season. As an example, consider the situation in the 2003/04 campaign shown in Figure 7. The seasonal pattern is nonlinear, but it is piecewise linear, with the first and second halves of the season following two distinct linear approximations whose corresponding slopes are 1.29 and 2.52, respectively, thus in ratio almost 1:2. A similar situation happened to Juventus in the 2015/16 season, when they collected 12 points in the first 10 rounds, and 79 in the following 28 rounds.

Furthermore, differences between various countries and leagues are small for every value of $t_s$. For example, for $t_s = 10$, the value of $|T_n - T_n|$ ranges between 4.19 for Portugal and 4.63 for The Netherlands, while for leagues, the minimum 4.19 is reached by the Portuguese Primeira Liga and the maximum 4.63 by the Dutch Eredivisie.

Finally, differences between teams ending in different zones of the final standing are also small. For $t_s = 10$, the values (with confidence intervals) of $|T_n - T_n|$ for all teams finishing first to fifth is 4.32 (4.19–4.46), for all teams filling the bottom five positions is 4.21 (4.07–4.35), while for the teams in the five positions at the middle of the table the corresponding values are slightly larger 4.54 (4.39–4.67) indicating a less precise linear predictivity for these teams. This reflects the fact that a team whose dynamic throughout the year is far from linear will hardly reach high or low positions in the standing, which instead include teams performing consistently good (or bad) during the campaign.

We also applied the five models to another quantity that is often used as a predictor: goal difference. Here, the nonlinearity is far more evident, but the ARIMA model also performs poorly. In the case $t_s = 10$, the

---

**Table 4.** $|T_n - T_n|$ for the top 10 Union of European Football Associations ranking teams at November 2015 for $t_s = 10$

| Rank | Team                  | $|T_n - T_n|$ | Rank | Team                  | $|T_n - T_n|$ |
|------|-----------------------|--------------|------|-----------------------|--------------|
| 1    | Real Madrid CF        | 4.48         | 6    | Sevilla FC           | 4.24         |
| 2    | Club Atlético de Madrid | 5.20        | 7    | Paris Saint-Germain | 4.72         |
| 3    | FC Bayern München     | 3.72         | 8    | Manchester City      | 3.68         |
| 4    | FC Barcelona          | 5.76         | 9    | Arsenal FC           | 4.28         |
| 5    | Juventus              | 3.72         | 10   | Borussia Dortmund    | 4.84         |

---

**FIG. 6.** Histogram of $|T_n - T_n|$ for the set $S$ of 258 teams having more presences ($\geq 20/25$). Color images are available online.

**FIG. 7.** Points earned by FC Barcelona (black dots) in La Liga 2003/04 and the corresponding linear models for the first (red line) and second (blue line) halves of the season. Color images are available online.
Confidence interval of the absolute prediction error is 5.86–6.06 for the linear model and 5.33–5.50 for the ARIMA model.

Championship outcome prediction

Let us now consider predicting the final outcome not of a single team but rather of an entire championship. As a performance measure, we use the normalized total absolute displacement \(d\) and Spearman’s rank correlation \(\rho\) outlined in the Methods.

As a first result, in Figure 8 we plot, for each \(1 \leq t_s \leq 20\), the distribution of the normalized total absolute displacements \(d\) for the 511 championships included in the considered data set. The 95% Student’s bootstrap confidence intervals \([l, u]\) are not reported in Figure 8 because they are too narrow. For each \(t_s\), we have \([l, u] \subset \left[ \frac{d}{1.038}, 1.037d \right]\). As a function of \(t_s\), the median of \(d\) is very close to \(\bar{d}\) (the ratio between the mean and the median of \(d\) ranges between 0.988 and 1.057), and it has an almost linear trend significantly smaller than the null model value \(\approx \frac{1}{2}\) even for large values of \(t_s\). For example, for \(t_s = 10\), we have \(d = 0.1874\) and \(\bar{d} = 0.879\), which, for a tournament with 20 teams, means that on average the linear model can guess the final ranking of each team with an error of 1.874 positions. In 25 cases (with \(t_s \leq 7\), the actual final ranking was perfectly predicted by the linear model. Note that for this task, the baseline model using the standing at \(t_s = 10\) to predict the final ranking trivially has a very good performance, achieving \(d = 0.196\) and \(\bar{d} = 0.814\). The fact that predicting the final ranking is an easier task than predicting the final number of points is not unexpected, since ranking variations in the last rounds of a championship are limited because differences in points can become quite large, even for teams that are close in the standing. This observation also explains the very good performance of the baseline model, whose accuracy is far better than the interpolating baseline for the point prediction task.

Moreover, even the best PE algorithm performs worse than the linear model, achieving an average displacement of \(\bar{d} = 0.264\) and an average Spearman rank correlation of \(\bar{\rho} = 0.741\).

No significant difference in the table prediction performance is also detected when comparing the top leagues (Premier League, Serie A, Ligue 1, La Liga, Bundesliga, Eredivisie, and Primeira Liga) with all the other considered leagues: \(\bar{d}\) for the former championships is
0.184 (0.175–0.192), while for the latter it is 0.189 (0.180–0.198; respectively). 

A crucial task championship outcome prediction is to forecast the final top and bottom of the table, that is, the teams qualifying for European tournaments (Champions League and Europa League) and the teams facing relegation. Define the true positive rate (TPR) as the fraction of championships (out of 425) where all the teams finishing in top \( k \) (or bottom \( k \)) positions were correctly predicted by a linear model. In Table 5, the TPR is shown for increasing \( t_s = 1, \ldots, 20 \), for the first/last \( k = 3 \) and \( k = 6 \) positions. Overall, the performance of the linear model is quite good for a wide range of values of \( t_s \). For \( t_s < 10 \), the TPR is >0.9 for all cases. Moreover, predictions for \( k = 3 \) is slightly noisier than \( k = 6 \), while in both cases predicting the bottom of the table is slightly harder than guessing the top teams. This is due to the fact that when the amount of points is small, as happens in the relegation zone, a single fluctuation (i.e., an unexpected win) can perturb the whole bottom part of the standing with a far larger impact than at the top.

Example: EPL 12/13

We conclude with a particularly favorable example (English Premier League 2012/13 relegation zone) where the linear model predictivity is better than the more complex combinations of algorithm and human knowledge, which translate into the odds offered by betting services. In Table 6, the corresponding relegation odds are reported for six betting agencies: (B1) Betting Expert, B2 (bwin), B3 Bet365, B4 Ladbrokes, B5 SportBookReview, and (B6) William Hill, together with the average odds. Although the betting odds were suggesting Norwich and Southampton, for instance, as likely candidates (with 2.50 and 2.16 average odds), quite unexpectedly (average odds 4.95) Queen’s Park Rangers suffered relegation instead.

In this case, the linear model performs effectively,
consistently predicting QPR, Reading, and Wigan as the relegated teams, for each $t = 1, \ldots, 20$.

Conclusions
A high level of linearity may be unexpected when dealing with football results, where a large number of confounding factors influence the outcome of both a single match and an entire tournament. Here, we show that when considering long tournaments such as national championships, linear trends are quite widespread, and linear models can also work as effective predictors. Although more refined predictors such as ARIMA or ETS have a better fit, the linear model indeed represents a consistent compromise between performance and simplicity. In particular, we tested the linear forecast of the total number of earned points by a team during a season, and the final team ranking in the table, where the model is trained only on the initial portion of the season. In both cases, we demonstrate that even such a minimalist approach and without using historical data can achieve good predictive performances.

Acknowledgments
We thank Ernesto Arbitrio and Luca Coviello for the support with the data web scraping and preprocessing.

Author Disclosure Statement
No competing financial interests exist.

Supplementary Material
Supplementary Data

References
23. t1egen11. The best predictor for future performance is expected goals. Available online at http://t1egen11.net/2015/01/05/the-best-predictor-for-future-performance-is-expected-goals/ (last accessed March 12, 2019).
Appendix A: The Displacement Metric

Let $T = \{z_1, \ldots, z_n\}$ be the teams involved in a given tournament. Consider now the standing $S$ after a certain matchday of the tournament, that is, the ranked list $S = [z_{x_1}, \ldots, z_{x_n}]$ for $\{x_1, \ldots, x_n\} = \{1, \ldots, n\}$. Let $\text{rk}_S$ be the ranking map, that is, the function associating to each team $z_i$ its position inside the standing $S$, and define $\tau_S = (\text{rk}_S(z_1), \text{rk}_S(z_2), \ldots, \text{rk}_S(z_n))$. Then, $\tau_S$ is a permutation of the first $n$ natural numbers, that is, a member of the symmetric group $S_n$. Thus, to each one of all possible $n!$ standings, $S$ is biunivocally associated with a permutation $\tau_S$. Given two standings $R, S$ on $T$, we define the distance $D$ between $R$ and $S$ as the total absolute displacement between $\tau_R$ and $\tau_S$:

$$D(R, S) = \sum_{i=1}^{n} |\text{rk}_R(z_i) - \text{rk}_S(z_i)| = \sum_{i=1}^{n} |\tau_R(i) - \tau_S(i)| .$$

In order to compare meaningfully distances computed in tournaments with different numbers of competing teams, $D$ is normalized by its maximum value, as in Mitchell[27]:

$$\max_{\tau_R, \tau_S \in S_n} D(\tau_R, \tau_S) = \max_{\tau_R, \tau_S \in S_n} \sum_{i=1}^{n} |i - \tau_R(i)| = \frac{n^2}{2} ,$$

where Id is the identical permutation. We can thus define the normalized distance $d$ as follows:

$$d(R, S) = \frac{D(R, S)}{\max_{\tau_R, \tau_S \in S_n} D(\tau_R, \tau_S)} = \frac{\sum_{i=1}^{n} |\tau_R(i) - \tau_S(i)|}{\frac{n^2}{2}} .$$

Furthermore, computing the expected value of $d$ over the whole permutation group $S_n$ allows the comparison of a given value of the normalized distance with

### Abbreviations Used

- ARIMA = autoregressive integrated moving average
- ETS = exponential smoothing state space
- PE = Pythagorean Expectation
- TPR = true positive rate

### Table A1. Set $T$ of teams playing in Italian Serie A 2014/15

<table>
<thead>
<tr>
<th>Index</th>
<th>Team name</th>
<th>Index</th>
<th>Team name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z1</td>
<td>Atalanta</td>
<td>Z11</td>
<td>Lazio</td>
</tr>
<tr>
<td>Z2</td>
<td>Cagliari</td>
<td>Z12</td>
<td>Milan</td>
</tr>
<tr>
<td>Z3</td>
<td>Cesena</td>
<td>Z13</td>
<td>Napoli</td>
</tr>
<tr>
<td>Z4</td>
<td>Chievo</td>
<td>Z14</td>
<td>Palermo</td>
</tr>
<tr>
<td>Z5</td>
<td>Empoli</td>
<td>Z15</td>
<td>Parma</td>
</tr>
<tr>
<td>Z6</td>
<td>Fiorentina</td>
<td>Z16</td>
<td>Roma</td>
</tr>
<tr>
<td>Z7</td>
<td>Genoa</td>
<td>Z17</td>
<td>Sampdoria</td>
</tr>
<tr>
<td>Z8</td>
<td>Hellas</td>
<td>Z18</td>
<td>Sassuolo</td>
</tr>
<tr>
<td>Z9</td>
<td>Inter</td>
<td>Z19</td>
<td>Torino</td>
</tr>
<tr>
<td>Z10</td>
<td>Juventus</td>
<td>Z20</td>
<td>Udinese</td>
</tr>
</tbody>
</table>

### Table A2. Actual (A) and predicted (P) table of Serie A 2014/15 after matchday 20

| Pos. | A  | A P | T | Team  | $\tau_A$ | $\tau_P$ | $|\tau_A - \tau_P|$ |
|------|----|-----|---|-------|----------|----------|-----------------|
| 1    | Juvent | Juvent | Z1 | Atalanta | 15       | 14       | 1               |
| 2    | Roma   | Roma  | Z2 | Cagliari | 17       | 18       | 1               |
| 3    | Napoli | Lazio | Z3 | Cesena   | 19       | 19       | 0               |
| 4    | Lazio  | Napoli| Z4 | Chievo   | 18       | 16       | 2               |
| 5    | Sampdoria | Genoa | Z5 | Empoli   | 16       | 13       | 3               |
| 6    | Fiorentina | Milan | Z6 | Fiorentina| 6        | 8        | 2               |
| 7    | Genoa  | Sampdoria | Z7 | Genoa    | 7        | 5        | 2               |
| 8    | Palermo Fiorentina | Z8 | Hellas | 14       | 12       | 2               |
| 9    | Udinese | Inter | Z9 | Inter    | 11       | 9        | 2               |
| 10   | Milan  | Udinese | Z10| Juvent | 1        | 1        | 0               |
| 11   | Inter  | Torino | Z11| Lazio    | 4        | 3        | 1               |
| 12   | Sassuolo | Hellas | Z12| Milan    | 10       | 6        | 4               |
| 13   | Torino | Empoli | Z13| Napoli   | 3        | 4        | 1               |
| 14   | Hellas | Atalanta | Z14| Palermo  | 8        | 15       | 7               |
| 15   | Atalanta | Palermo | Z15| Parma    | 20       | 20       | 0               |
| 16   | Empoli | Chievo | Z16| Roma     | 2        | 2        | 0               |
| 17   | Cagliari | Sassuolo | Z17| Sampdoria| 5        | 7        | 2               |
| 18   | Chievo | Cagliari | Z18| Sassuolo | 12        | 17       | 5               |
| 19   | Cesena | Cesena | Z19| Torino   | 13       | 11       | 2               |
| 20   | Parma  | Parma  | Z20| Udinese  | 9        | 10       | 1               |

$D(A, P) = \sum_{i=1}^{20} |\tau_A(i) - \tau_P(i)| = 38$.

In the last column, the absolute displacement $|\tau_A - \tau_P|$ is reported between A and P for the corresponding team $z_i$, and its total is indicated in the last row. The corresponding permutations $\tau_A$ and $\tau_P$ are computed with respect to the set of teams $T$. 
the null hypothesis of the distance with a random standing:

\[
E_S(d) = \frac{1}{|S_n|} \sum_{\tau \in S_n} d(\text{id}, \tau)
\]

\[
= \frac{1}{n!} \left[ \frac{n!}{2} \right] \sum_{i=1}^{n} \sum_{i=1}^{n} [i - \tau(i)]
\]

\[
= \frac{1}{n!} \left[ \frac{n!}{2} \right] \sum_{i=1}^{n} \sum_{i=1}^{n} [i - \tau(i)]
\]

\[
= \frac{1}{n!} \left[ \frac{n!}{2} \right] 2 \sum_{i=1}^{n} \sum_{i=1}^{n} (n-1)\frac{(i-1)(i-n)}{2}
\]

\[
= \frac{1}{n!} \left[ \frac{n!}{2} \right] \frac{(n-1)(n-2)(n+1)}{3}
\]

\[
= \frac{2}{3} - \frac{2}{3n^2 \text{ mod } 2},
\]

which is \(\frac{2}{3}\) for odd \(n\)’s and \(\frac{2}{3} - \varepsilon_n\) for even \(n\)’s, with \(\varepsilon_n\) positive, decreasing to 0 and smaller than 0.06 for \(n \geq 10\). Thus, regardless of the number of playing teams, the distance \(d\) between two standings in the same championship is a number ranging between 0 (for identical rankings) and 1 (for maximally different standings), with \(d \approx \frac{2}{3}\) for randomly chosen standings. Hereafter, we show an example of the use and the interpretation of the distance \(d\).

**Example**

Suppose we want to assess the error of a predictive algorithm \(P\) in forecasting the standing of a tournament after a given matchday, using metric \(d\) as the evaluation measure. In particular, we test \(P\) in two situations: (1) round 20 of Italian Serie A 2014/15 and (2) the final round (18) of the South American qualifiers for the 2010 FIFA World Cup.

(1) Italian Serie A 2014/15 involved 20 teams, composing the set \(T\) as shown in Table A1. The initial assignment of the \(z_i\) labels with the team is arbitrary, and any other choice would work instead.

After round 20, the table, labeled as A, read as reported in Table A2. Suppose now that algorithm \(P\) predicts the championship table as in Table A2, labeled as P. The first step in evaluating the difference between standings A and P is the derivation of the corresponding permutations \(\tau_A\) and \(\tau_P\), and then the computation of the sum of all displacements \(\tau_A - \tau_P\). As shown in the last row of Table A2, this reads as:

\[
D(A, P) = \sum_{i=1}^{20} |\tau_A(i) - \tau_P(i)| = 38,
\]

thus the final normalization provides the value of the distance \(d\):

\[
d(A, P) = D(A, P) \cdot \frac{1}{\left[ \frac{n!}{2} \right]} = 38 \cdot \frac{1}{200} = 0.19,
\]

which is a small number, indicating a good similarity between standings A and P, quite distant from the random value 0.6.

(2) In the second case study, we compare the actual A and the predicted P final standings of the South American qualifiers for the 2010 FIFA World Cup, whose competing teams are listed in Table A3. Following the same approach of case (1), we build the analogous Table A4. Here, the absolute total displacement \(D(A, P) = 14\), apparently much smaller than in case (1), but the normalized distance \(d(A, P)\) results \(14/(10^2/2) = 0.28\), showing instead a worse performance of the predictive algorithm \(P\) in case (2) compared to case (1).