Introduction

Panel data analysis deals with a regression procedure where individual subjects as well as information about different times is taken into account [1]. The update of estimators with time can be related to Bayesian approaches [2,3] as explicitly discussed, e.g., in [4]. For Gaussian statistics there exists a direct connection between Bayesian inference and a regression analysis; see, e.g., [5]. Actually, Bayesian inference to soccer has recently been discussed in Ref. [6].

Of key interest is the knowledge about the quality of the estimator. Here we simplify the general result by using the assumption that the underlying property of the subject does not change between the final measurement and the prognosis time interval. This does not necessarily hold for the time of earlier measurements. However, due to the random noise, by which the most recent measurement may be disturbed, it may still be favorable to take into account older pieces of information. Having an explicit expression of the estimator quality it is possible to judge the relevance of the available information for the prediction process in a detailed manner. Furthermore, we can define the limit of optimum prediction and judge, how far a specific prediction process one has to identify appropriate observables which reflect the strength of the individual teams as close as possible. A criterion to distinguish different observables is presented. Surprisingly, chances for goals turn out to be much better suited than the goals themselves to characterize the strength of a team. Routes towards further improvement of the prediction are indicated. Finally, two specific applications are discussed.

Optimizing the Prediction Process: From Statistical Concepts to the Case Study of Soccer

Andreas Heuer1,2*, Oliver Rubner1,2

1 Institute of Physical Chemistry, WWU Muenster, Muenster, Germany, 2 Center of Nonlinear Science CeNoS, WWU Muenster, Muenster, Germany

Abstract

We present a systematic approach for prediction purposes based on panel data, involving information about different interacting subjects and different times (here: two). The corresponding bivariate regression problem can be solved analytically for the final statistical estimation error. Furthermore, this expression is simplified for the special case that the subjects do not change their properties between the last measurement and the prediction period. This statistical framework is applied to the prediction of soccer matches, based on information from the previous and the present season. It is determined how well the outcome of soccer matches can be predicted theoretically. This optimum limit is compared with the actual quality of the prediction, taking the German premier league as an example. As a key step for the actual prediction process one has to identify appropriate observables which reflect the strength of the individual teams as close as possible. A criterion to distinguish different observables is presented. Surprisingly, chances for goals turn out to be much better suited than the goals themselves to characterize the strength of a team. Routes towards further improvement of the prediction are indicated. Finally, two specific applications are discussed.

Citation: Heuer A, Rubner O (2014) Optimizing the Prediction Process: From Statistical Concepts to the Case Study of Soccer. PLoS ONE 9(9): e104647. doi:10.1371/journal.pone.0104647

Editor: Dominik Wodarz, University of California Irvine, United States of America

Received December 3, 2013; Accepted July 16, 2014; Published September 8, 2014

Copyright: © 2014 Heuer, Rubner. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publication Fund of University of Muenster. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: andheuer@uni-muenster.de
this general scheme is applied to the prediction of soccer matches, using the German premier league (Bundesliga) as an example. It can be shown that all assumptions, used in the previous Section, are fulfilled to a very good approximation. Furthermore, it is shown that chances for goals possess a very high information content about the individual team strengths and are, thus, chosen for the respective covariates. Subsequently, the theoretical results are compared with the explicit bivariate regression analysis. The specific setting is chosen such that one wants to predict the outcome of the second half of a season, based on knowledge of a variable number of matches from the first half of the same season as well as all matches of the previous season. In particular we discuss the dependence of the prediction quality on the number of matches, taken into account. Furthermore, it is shown, how the present concepts can be applied to the prediction of single matches. We end with a discussion.

The Statistical Background of Prediction

Variables

We consider two successive time intervals, in which we measure the independent variables \(X\) and \(Y\). For the later application to soccer this might be the accumulated goal difference during the previous season and during the present season, measured individually for each team. Here we consider differences in order to capture both the offensive and defensive strength. Specifically, we perform this analysis after half of the present season is over. Naturally, this can be easily generalized to other situations. The aim is to predict the goal difference \(Z\), i.e. the dependent variable, of each team during the second half of the season. This setup is sketched in Fig.1. The prediction quality can be explicitly expressed and compared with the theoretical optimum.

Regression

First, we briefly review some key relations of regression analysis. We start with the linear relation \(Z^{} = b^{} Y^{}\) for the independent variable \(Y\) and the dependent variable \(Z\). Note that we assume all variables fulfill the condition that their first moment is zero. Generalisation is, of course, straightforward. The regression problem requires the minimisation of \(\langle (Z - Z^{})^2 \rangle\) with respect to \(b\) where \(Z^{} = b^{} Y^{}\) is the predictor of \(Z\). Substituting the resulting value of \(b_{\operatorname{opt}} = \frac{\operatorname{corr}(Y,Z) \sqrt{\operatorname{Var}(Z)/\operatorname{Var}(Y)}}\) yields for the optimum quadratic variation, denoted \(\chi^2\) (1),

\[
\chi^2 (Y) = \operatorname{Var}(Z) (1 - |\operatorname{corr}(Y,Z)|^2)
\]

where \(\operatorname{Var}(Z)\) denotes the variance of the distribution of \(Z\) and

\[
\operatorname{corr}(Y,Z) = \frac{\langle Y Z \rangle^{}}{\sqrt{\operatorname{Var}(Y) \operatorname{Var}(Z)}}
\]

is the Pearson correlation coefficient between the variables \(Y\) and \(Z\). Eq.1 has a simple intuitive interpretation: The higher the correlation between the variables \(Y\) and \(Z\), the better the predictability of \(Z\) in terms of \(Y\).

For the present work we are mainly dealing with the bivariate regression \(X = a^{} X + b^{} Y\). Via normal equations, one can obtain general expressions for the regression coefficients \(a\) and \(b\). Interestingly, the prediction quality of the bivariate prediction can be analogously expressed to Eq.1 and reads

\[
\chi^2 (X,Y) = \chi^2 (Y) (1 - |\operatorname{corr}(X,Y,Z,Y)|^2)
\]

where the partial correlation coefficient

\[
\operatorname{corr}(X,Y,Z,Y) = \frac{\operatorname{corr}(X, Z, Y) \operatorname{corr}(Y, Z)}{\sqrt{1 - \operatorname{corr}(X, Y)^2} \sqrt{1 - \operatorname{corr}(Y, Z)^2}}
\]

is used and \(\chi^2 (Y)\) is defined as in Eq.1. The second factor on the right-hand side of Eq.3 explicitly contains the additional information of the variable \(X\) as compared to \(Y\). One can easily show that in agreement with expectation Eq.3 is completely symmetric in \(X\) and \(Y\).

Here we present a straightforward derivation of Eq.3. Let \(dX\) denote the solution of the regression problem \(X = dX\). Accordingly, \(dX\) is the solution of the regression problem \(X = dY\). In the first step one defines the new variables \(Z = X - dY\) and \(Z = X - dX\). For these new variables the correlation with \(Y\) is explicitly taken out. A straightforward calculation shows that the Pearson correlation coefficient \(\operatorname{corr}(X,Z)\) is exactly given by the partial correlation coefficient \(\operatorname{corr}(X,Y,Z,Y)\).

Now we consider the regression problem of interest \(Z = a^{} X + b^{} Y\). In a first step it is formally rewritten as

\[
Z = dY + (X - dY Y) + (b - dX Z + adY X) Y
\]

Using the above notation and introducing the new regression parameter \(\tilde{b}\) we abbreviate this relation via

\[
\tilde{Z} = a^{} X + \tilde{b}^{} Y
\]

By construction the observable \(Y\) is uncorrelated to \(\tilde{X}\) and \(\tilde{Z}\). Therefore the independent variable \(Y\) does not play any role for the prediction of \(\tilde{Z}\) so that effectively one just has a single-variable regression problem. Therefore one can immediately write

\[
\chi^2 (X,Y) = \operatorname{Var} (\tilde{Z}) (1 - |\operatorname{corr}(\tilde{X},\tilde{Z})|^2)
\]

The first factor is identical to \(\chi^2 (Y)\) whereas the Pearson correlation coefficient in the second factor is identical to \(\operatorname{corr}(X-Y,Z-Y)\). This concludes the derivation of Eq.3.

Prediction for individual subjects/teams

As introduced, the variables \(X,Y,Z\) denote the output of a team or, more generally, of some subject during three successive time intervals. For the first time interval, the outcome of team \(i\) is
denoted $x_i$. Conceptually, this value has contributions from the true underlying team strength $s_{X,i}$ as well as from random non-predictable effects $e_{X,i}$, i.e.

$$x_i = s_{X,i} + e_{X,i}.$$  \hfill (8)

Thus, only in the absence of random effects the team strength $s_{X,i}$ could be directly identified with the outcome $x_i$. In what follows we use the terminology of soccer but this approach can be directly applied to other cases where the observable is the sum of the properties of the respective subject and some random effects.

Following the previous discussion we only consider observables $X$ for which the first moment disappears after averaging over all teams. Naturally, the same holds for the team strength observable $S_Y$. Squaring Eq.8 and averaging over all teams yields

$$\text{Var}(X) = \text{Var}(S_X) + \text{Var}(e_X).$$  \hfill (9)

Analogous relations hold for $\text{Var}(Y)$ and $\text{Var}(Z)$.

For the evaluation of the prediction quality Eq.3 one needs to calculate individual correlations such as $\text{corr}(Y,Z)$. A straightforward calculation yields

$$\text{corr}(Y,Z) = \frac{\text{corr}(S_Y,S_Z)}{\sqrt{1 + \text{Var}(e_Y)/\text{Var}(S_Y)} \sqrt{1 + \text{Var}(e_Z)/\text{Var}(S_Z)}}.$$  \hfill (10)

Again, analogous expressions hold for $\text{corr}(X,Z)$ and $\text{corr}(X,Y)$.

Eq.10 allows one to identify two distinct reasons why the correlation of $Y$ and $Z$ is smaller than unity. First, the team strength may change with the two time intervals, i.e. $\text{corr}(S_Y,S_Z)<1$. Second, the random effects, which influence the observables $Y$ and $Z$, may play an important role ($\text{Var}(e_Y),\text{Var}(e_Z)>0$).

The subsequent discussion is based on the mathematical identity

$$\frac{\text{corr}(X,Y)}{\text{corr}(X,Z)\text{corr}(Y,Z)} = \frac{\text{corr}(S_X,S_Y)}{\sqrt{1 + \text{Var}(e_Y)/\text{Var}(S_Y)} \sqrt{1 + \text{Var}(e_Z)/\text{Var}(S_Z)}} \left(1 + \frac{\text{Var}(e_Z)}{\text{Var}(S_Z)}\right).$$  \hfill (11)

As a first step of simplification we want to estimate the team strength $S_Z$ rather than $Z$ itself. Then the prediction quality is denoted by $\hat{\chi}^2(X,Y)$. All relations remain identical except $\text{Var}(e_Z)=0$ in the evaluation of quantities, occurring in Eq.3. Naturally, one has the simple relation

$$\hat{\chi}^2(X,Y) = \hat{\chi}^2(X,Y) + \text{Var}(e_Z).$$  \hfill (12)

As the second step we consider the special case that the team strengths are the same in the second and third time interval, belonging to $Y$ and $Z$, respectively. Actually, it has been already shown in Ref. [21] that apart from short-time fluctuations the team strength remains constant during the course of a season. As a consequence one has $S_Y \equiv S_Z$, i.e. nearly the same team strength in the first and the second half of a season. Mathematically, we assume a strict equality. The corresponding empirical result will be discussed further below. A mathematical consequence is (see below for specific data) $\text{corr}(S_X,S_Y) = \text{corr}(S_X,S_Z)$. Under this assumption, Eq.11 can be rewritten as

$$\text{corr}(X,Y) = \text{corr}(X,S_Z) \text{corr}(Y,S_Z).$$  \hfill (13)

Inserting this relation into Eq.3 for the prediction quality of $S_Z$ the general expression simplifies significantly and one obtains

$$\hat{\chi}^2(X,Y) = \frac{\text{Var}(S_Z)}{(1 - \text{corr}(Y,S_Z))^2}(1 - \text{corr}(X,S_Z))^2 \frac{1}{1 - \text{corr}(X,Y)^2}. \hfill (14)$$

This is the key relation to be used when estimating the quality of the prediction. Apart from the assumption of constant properties during the final two time intervals, this relation is generally valid.

**Application to the Case of Soccer Prediction: Concepts**

**General**

Our general goal is the prediction of the future results of soccer matches. Specific data are taken for the German premier league (Bundesliga), employing information about all matches between the seasons 1995/96 and 2010/11. During a season a team has 34 matches.

Our goal is the prediction of the aggregated results $z_i$ of each team $i$ of the second half of a season, based on knowledge about $N_Y$ match results $y_i$ from the first half of the season as well as the $N_Z$ results $x_i$ from the previous season. As the dependent variable $z_i$ we choose the goal difference but a similar analysis could be also performed for points; see again Fig.1. Of course, due to the generality of our approach also different prediction problems can be handled. For the explicit calculations of the goal differences we correct for the home advantages [5] so that the statistical properties are independent of the home advantage.

**Disentangling random and systematic effects**

For our analysis it is essential to decompose the variables $X,Y$ and $Z$ into its systematic parts $(S_X,Y,Z)$ and its random contributions $(e_{X,Y,Z})$; see Eq.8. As mentioned above, $z_i$ will be identified as the goal difference of team $i$ after $N_Z$ matches, normalised by $N_Z$. In case of matches under identical conditions the random effects are averaged out as reflected by the standard scaling relation $\text{Var}(e_Z) \propto 1/N_Z$ where the proportionality constant is denoted $V_Z$. Thus, we have

$$\text{Var}(e_Z) = \frac{V_Z}{N_Z}.$$  \hfill (15)

By studying the dependence of $\text{Var}(Z)$ on $N_Z$ the systematic and random contributions to $\text{Var}(Z)$, as expressed in Eq.9, can be identified. Of course, analogous relations hold for $X$ and $Y$.

Strictly speaking, the scaling with the inverse number of the matches breaks down for $N_Z$ close to unity because then different strengths of the opponents no longer average out. In practice it turns out that for $N_Z > 4$ the difference of the $N_Z$ opponents has sufficiently averaged out. This dependence on the number of considered matches has been explicitly analysed in Ref. [5,22]. For the present set of data we obtain $\text{Var}(S_Z)=0.21$ and $V_Z=2.95$. Actually, $V_Z$ is very close to the total number of goals per match (2.85). This expectation is compatible with the assumption of independent Poisson processes.
Choice of observables

The goal is to predict the goal difference \( Z \) or, alternatively, the
team strength \( SZ \). A natural choice for the independent variables \( X \)
and \( Y \) are the goal differences in the respective time intervals.
In what follows, goal differences are denoted as \( \Delta G \). However, as
will be shown below, this choice is far from optimum. Generally
speaking, one aims for observables which contain as much
information as possible about the team strength.

How to capture the information content of a given observable?
For this discussion we restrict ourselves to the prediction problem
\( Y \rightarrow Z \) to be solved via a simple univariate regression as
summarised above (see Eq.1). For this analysis we use
\( N_{Y} = N_{Z} = 17 \), i.e. all matches from the first and second half of
the season, respectively. The quality of the prediction is captured
by \( \text{corr}(Y,Z) \). The larger the value \( \text{corr}(Y,Z) \), the better the
prediction and thus the higher the information content of \( Y \) about
the team strength. From the empirical data we obtain
\( \text{corr}(Y = \Delta CY, Z = \Delta CZ) = 0.56 \).

Can one increase \( \text{corr}(Y,Z) \) significantly beyond the value of
0.56 by using other observables? The scoring of goals is the final
step in a series of match events. One may thus hope that there
exist other match characteristics which are even more informative
about the team strength. A possible candidate is the number of
chances for goals. They are provided by a professional sports
journal (www.kicker.de) for all seasons, considered in this work.
We denote the chances for goals as \( C_{Y} \) and the goals as \( G_{Z} \). The
sign indicates whether it refers to the considered team (+) or the
opponent of that team (−).

Next we define the scoring efficiencies \( p_{\pm} \) via the relation

\[
G_{Z} = C_{Y} \cdot p_{\pm},
\]

Here, \( p_{\pm} ( = G_{+}/C_{+} ) \) denotes the probability that the team is able
to convert a chance for a goal into a real goal and \( 1 - p_{\pm} \) that the
team manages not to concede a goal after a chance for a goal of
the opponent. Averaging over all teams and seasons one obtains
\( \langle p_{\pm} \rangle = 0.24 \). Thus, every forth chance for a goal ends up in a
goal.

In Fig.2 the actual scoring efficiencies \( p_{+} \) after a season are
shown together with the respective values of \( \Delta C \). Very clearly, the
goal efficiencies are widely distributed between approx. 15% and
35%. On average, better teams with a larger value of \( \Delta C \) have a
slightly better efficiency to score goals and more likely avoid to
concede goals (correlation coefficients \( \pm 0.26 \)). Despite this small
 correlation, the large scatter of \( p_{\pm} \) cannot be explained in terms of
\( \Delta C \).

This large unexplained variance seems to imply that the scoring
efficiencies strongly vary from team to team in an a priori
unknown way. As a consequence the chances for goals would
hardly contain additional information about the expected number
of goals, which a team is going to score in the future. In particular,
the estimation of the team strength, which is defined on the basis
of goals, would hardly be improved by taking into account the
chances for goals.

With the definition \( \Delta C = C_{Y} - C_{Z} \) this statement is equivalent
to the presence of a weak correlation between \( \Delta Cy \) and \( \Delta GZ \).
However, this preliminary conclusion is wrong. Rather the
 correlation coefficient turns out to be \( \text{corr}(Y = \Delta CY, Z = \Delta CZ) = 0.65 \)
which is much larger than the value of \( \text{corr}(Y = \Delta GZ, Z = \Delta GZ) = 0.56 \).
Stated differently, the chances for goals are by far more informative
for the prediction of the team strength than the goals themselves!

Why chances for goals are so informative

This observation could be rationalized under the hypothesis that
the scoring efficiencies are very similar for all teams. Qualitatively,
one can argue in this limit that random effects are stronger for
goals than for chances for goals, since the number of goals is
typically smaller than the number of chances for goals. To quantify
this aspect, we consider a simple example of a fictive coin-tossing
tournament where the head appears with probability \( p \) which in
this simple example is given by \( 1/2 \). A team is allowed to toss the
coin \( M \) times per round. In the first round this results in \( g_{1} \) times
tossing the head. Thus, in the first round one has observed the
number of tosses \( M \) as well as the number of heads \( g_{1} \). In the
relation to soccer \( M \) would correspond to the number of chances
for goals and \( g_{1} \) to the number of goals in that match. In order to
keep the argument simple we assume that \( M \) is a constant whereas
in a real soccer match \( M \) can vary. How to predict the expected
number of heads \( g_{2} \) in the next round? Here we consider two
different approaches. (1) The prediction is based on the
achievement of the first round, i.e. on the value of \( g_{1} \). Then the
best prediction is \( g_{2} = g_{1} \). The variance of the statistical error of
the prediction can be simply written as \( \sum_{g_{1} = 0}^{M} p(g_{1}) (p(g_{2} = g_{1} - g_{2})^{2} \)
where \( p(g) \) is the binomial distribution. A straightforward
calculation yields for this variance a value of \( 2Mp(1 - p) \). (2)
The prediction is based on the knowledge of tossing attempts \( M \). If
furthermore the value of \( p \) is known the optimum prediction is, of
course, \( pM \). The variance of the statistical error is given by the
binomial distribution, i.e. by \( Mp(1 - p) \). Stated differently,
knowing the number of attempts to reach a specific goal (here
tossing a head) is more informative than the actual number of
successful outcomes as long as the probability \( p \) is well known.
Note that in this limit the common value of the scoring efficiency is
very well determined because it results from averaging over all
teams.

This hypothesis seems to contradict the results Fig.2, as
presented above. However, a priori the large fluctuations of \( p_{\pm} \)
in Fig.2 do not necessarily contradict the presence of a rather
uniform value of \( p_{\pm} \) for all teams. Rather, this apparent
disagreement can be easily resolved by discussing in more detail
the possible reasons for the strong fluctuations of \( p_{\pm} \) when
comparing different teams. In general, these fluctuations are a
superposition of two effects: (i) true differences between teams and
(ii) statistical fluctuations, reflecting the random effects in the 34
soccer matches of the season. In analogy to the previous discussion
both effects can be disentangled by studying the dependence of the
variance of \( p_{\pm} \) on the number of matches \( N \), which has been used
for the averaging. The results of this analysis is shown in Fig. 3.
One can see that the extrapolation to large \( N \), i.e. the systematic

![Figure 2. The efficiency factors \( p_{\pm} \) as a function of the
differences of the chances for goals \( \Delta C \).
doi:10.1371/journal.pone.0104647.g002](https://www.plosone.org/figure/2.0104647.g002)
team-specific variance of $p_{\pm}$, yields a value (0.0002) which is much smaller than the variance for $N=34$ (0.0012), i.e. after averaging over a whole season. Thus, the large fluctuations in Fig.2 are mainly of statistical nature and the efficiency to score a goal from a proportionality factor, basically identical to the definition of the team strength as defined via the goals (corresponding to $S_Z$). Both results are very promising with respect to the ability to predict soccer matches. In particular, the key approximation, entering Eq.14, is indeed very well fulfilled.

We mention in passing [22] that a closer analysis reveals that the team strength fluctuates with a small amplitude of approx. $A = 0.17$ and with a decorrelation time of approx. 7 matches. Since we average over a larger number of matches and, furthermore, restrict ourselves to the prediction of the total second half, these temporal fluctuations are to a large extent averaged out and do not show up in the present statistical analysis.

In case that the team strength $S_Y$ is perfectly known, i.e. $Y = S_Y$, Eq.10 yields (using $\epsilon_Y = 0$) $corr(S_Y, Z) = 1/\sqrt{1 + Var(S_Z)/[17Var(S_Z)]} = 0.74$. One may compare this limit of optimum prediction with the case where $Y$ was calculated based on the the chances for goals (correlation of 0.65) or based on the goals (correlation of 0.56). This clearly reveals that using the chances for goals instead of the goals yields a significant step towards the theoretical optimum.

The final unknown in our prediction scheme are values of $corr(S_X, S_Y)$ and $corr(S_Y, S_Z)$ which can be determined in analogy to $corr(S_Y, S_Z)$. Explicit calculation yields $corr(S_X, S_Z) = 0.88$ and $corr(S_Y, S_Y) = 0.86$. Both values are identical within statistical errors $(corr(S_X, S_Z) - corr(S_X, S_Y) = 0.02 \pm 0.02)$. This is compatible with the observation that the team strength does not vary within a season but vary within the summer break. For future purposes we use the average value of $corr(S_X, S_Y, Z) = 0.87$ for the characterization of the correlation of the team strength between two seasons.

**Application to the Case of Soccer Prediction: Results**

**Prediction of team strength**

To check our analytical results we perform an explicit multivariate regression analysis to estimate $Z$ based on knowledge of $X$ and $Y$ by using standard algorithms. To capture the dependence on the information content of the first half of the present season we also vary the number of considered matches $N_Y$. To improve the statistical quality of the data for $N_Y < 17$ we always average over different random selections of $N_Y$ matches from the first half of the season. For the determination of $D_X$ we choose all matches, i.e. $N_X = 34$ (thus taking the whole season). To check the relevance of the information from the previous season we alternatively set $X = 0$, i.e. ignore the information from the previous season.

One technical aspect needs to be mentioned. In a given season two or three teams have just been promoted. Thus, no data about the previous season are available. Therefore, we set the value of $x_i$ for the differences of the chances for goals for the promoted team to a constant value $x_{\text{prom}}$. This value is determined by the condition that the resulting average value $x_i$ (averaged over all teams of the present season) is zero.

The numerical results are shown in Fig.4. We start with the case $X = 0$. One can see that (trivially) for $N_Y = 0$ the standard deviation in the estimation of the team strength is identical to the standard deviation of the $S_Y$-distribution because no team-specific information has been used. The longer the season, the more information is available to distinguish between stronger and weaker teams. Using the information of the complete first half of the season ($N_Y = 17$) the statistical uncertainty decreases to 0.22.

**Table 1. The different systematic and random contributions of the observables, relevant for this work.**

| $X = D_X$ | $V_X$ | $Y = D_Y$ | $V_Y$ | $Z = D_Z$ | $V_Z$ |
|----------|-------|----------|-------|----------|-------|
| 2.32     | 14.1  | 2.66     | 14.2  | 0.21     | 2.95  |

doi:10.1371/journal.pone.0104647.t001

---

**Figure 3. The variance of the distribution of scoring efficiencies in dependence of the number of match days.**

doi:10.1371/journal.pone.0104647.g003
We have repeated the same calculation by identifying $Y$ with the goal differences $\Delta G_Y$. The prediction quality is significantly worse and one obtains an uncertainty of 0.30 rather than 0.22 after $N_Y = 17$ matches.

When additionally incorporating the information from $X$, the statistical uncertainty is already quite small at the beginning of the season (0.3). Of course, when increasing $N_Y$ it further decreases. Even after 17 matches the additional gain of using $X$ is significant (0.19 vs. 0.22). Thus, despite the slight decorrelation of the team strength during the summer break it is advantageous to take into account the information from the previous season even after half of the present season has been played.

Furthermore, we compare in Fig. 4 the actual uncertainty of the prediction of $Z$ with the theoretical expectation as expressed by Eq.12 and Eq.14. One finds a very close agreement with the actual data. This serves as a consistency check of our whole procedure and just reflects the fact that the assumptions, underlying the derivation of Eq.14, are fulfilled very well.

Finally, we explicitly apply this formalism to the prediction of a specific season of the Bundesliga. We aim to predict the goal difference of the 2nd half based on previous information. The regression problem reads $\Delta G_Z = a(N_Y) \Delta C_X + b(N_Y) \Delta C_Y(N_Y)$ where the weighting factors depend on the number of matches, included from the first half of the present season. They are listed in Tab.2 for different values of $N_Y$. Naturally, for $N_Y = 0$ the estimation is only based on $\Delta C_X$. Here the regression coefficient can be also calculated analytically using the values, mentioned in this work. Specifically, one gets $a(N_Y = 0) = 17\text{corr}(\Delta C_X, \Delta G_Z)/\text{Var}(\Delta C_X) = 17\text{corr}(\Delta C_X, S_Z)/\text{Var}(S_X)\text{Var}(S_Z)/\text{Var}(\Delta C_X) = 17 - 0.88\sqrt{2.32\cdot 0.21/(2.32 + 14.1/34)} \approx 3.8$ which is very close to the numerically determined value of 3.71. As expected, more information during the present season, i.e. larger $N_Y$, leads to a stronger weighting of $\Delta C_Y$. After $N_Y = 12$ matches the information contents of the previous season is basically equal to that of the first matches of the present season.

Based on these regression parameters we explicitly predict the goal difference of the second half for the two cases $N_Y = 0$ and $N_Y = 17$. We present data for the season 2007/08. Both predictions for $\Delta G_Z$ are listed in Tab.3 together with the actual values of $\Delta G_Y$ and $\Delta G_Z$ during that season.

One can see that for most cases the prediction before the season and in the middle of the season agree quite well, i.e. no dramatic reevaluations of the team strength as compared to the previous year was necessary. Notable exceptions are Munchen (estimation of +10 before the season and +21 after half of the season) and Leverkusen (increase from +2 to +9). Obviously, this reevaluation reflects that the fact that both teams played much better during the first half of that season (goal differences of +23 and +16 for Munchen and Leverkusen, respectively) than expected beforehand.

The final column also contains information about the logarithm of the market value (taken from www.transfermarkt.de) as an independent variable for a trivariate regression problem. The scaling of the team strength with the logarithm of the market value has been explicitly shown in previous work [22]. The resulting modifications in the estimation of $\Delta G_Z$ are small but significant. When averaging over all years between 2001/02 and 2010/11, for which the market value is available, it turns out that the prediction quality improves by 0.02 for $N_Y = 17$. Thus, relative to 0.19 a further significant improvement can be achieved.

### Prediction of single matches

Please note that the estimation of $\Delta G_Z$ is the basis for many other types of prediction. Since $\Delta G_Z$ is nothing else than the team strength, this value can be directly taken to estimate individual matches. For example, on the 18th match day of the season 2007/08 Cottbus was playing vs. Leverkusen. As shown in Refs. [21,22] the expected goal difference during a match of team $i$ and $j$ in the Bundesliga is given by the difference of the team strength of both teams plus some team-independent contribution, reflecting the home advantage. Nonlinear effects can be neglected. For this specific match the expected outcome was (using the final column in Tab.3): $(-12)/17 - (+8)/17 + 0.3 = -0.9$, using the home advantage of approx. 0.3 during that season. Thus, the best estimation for the resulting goal difference of that match, based on the available information used in this work, is -0.9. Actually, the final result was 2:3.

Here is a brief summary of the different prediction steps, following the general procedure in [21] and in agreement with previous work [e.g.10].

1. Calculation of the team strength via a linear regression approach. As main parameters enter $\Delta C_X$, $\Delta C_Y$, and the logarithm of the market value of the team at the beginning of the season. Naturally, for a match on the M-th match day one uses $N_Y = M - 1$. Minor further improvements can be reached by introducing an index for promoted teams and by taking into account short-time fluctuations of the team strength by using the results of the last seven matches as an individual parameter [22]. In total, this ends up in a five-dimensional regression analysis. The regression parameter have been obtained from comparison of all seasons between 1995/96 and 2010/11, excluding the season which predictions are performed.

[Table 2. The two regression parameters as a function of $N_Y$.]

| $N_Y$ | $a(N_Y)$ | $b(N_Y)$ |
|------|---------|---------|
| 0    | 3.71    | 0       |
| 4    | 3.20    | 0.81    |
| 8    | 2.60    | 1.70    |
| 12   | 2.33    | 2.33    |
| 17   | 1.86    | 2.77    |

doi:10.1371/journal.pone.0104647.t002
2. Calculation of the sum of goals by a corresponding regression analysis, taking into account the goals, scored in the present season so far, and the goals of the previous season [21]. However, for the calculation of the outcome of individual matches this step is by far less important than the estimation of the team strength.

3. Estimation of the team-independent home advantage in the corresponding season in analogy to the previous step [22].

4. Calculation of the expectation value of goals of both teams from steps 1–3.

5. Estimating possible final scores by assuming independent Poisson processes.

6. Correcting for the effect that draws are more likely than expected on the expense of matches with goal differences $+1$ [23].

Note that in earlier work goals rather than chances for goals were employed. We would like to stress again that the critical part of this endeavor is the determination of the team strength as described in this work.

To characterize the quality of the present approach we have compared the predictions of single matches with odds from Oddset, using data between the seasons 2002/03 and 2006/07, where the odds were available to us. Specifically, we used the scaled inverse odds as an estimate of the respective probabilities for a win of the home team, a draw, or a win for the away team. An objective measure is the parameter

$$K = -\langle \ln (\text{probability for win, draw, loss}) \rangle$$

(17)

where the probability for the actual outcome is taken as the argument of the logarithm. One can show that the value of $K$ is a

### Table 3.

The predictions of the goal difference of the second half of the Bundesliga-season 2007/08 for each team, based on the differences of chances for goals $\Delta C_Y$ of the previous season (3rd column) or, additionally, on the differences of chances for goals $\Delta C_Y$ of the first 17 matches of the present season (4th column).

|                  | $17\Delta G_Y$ | $17\Delta G_Z$ | $17\Delta G_Z,0(N_Y = 0)$ | $17\Delta G_Z,17(N_Y = 17)$ | plus market value |
|------------------|----------------|----------------|---------------------------|-----------------------------|------------------|
| B. München      | 23             | 24             | 10                        | 21                          | 23               |
| Bremen          | 18             | 12             | 11                        | 15                          | 14               |
| Hamburg         | 11             | 10             | 3                         | 9                           | 10               |
| Leverkusen      | 16             | 1              | 2                         | 9                           | 8                |
| Schalke         | 9              | 14             | 8                         | 12                          | 11               |
| Karlsruhe       | $-2$           | $-13$          | $-8$                      | $-6$                        | $-7$             |
| Hannover        | $-1$           | $-1$           | 3                         | $-1$                        | $-2$             |
| Stuttgart       | $-1$           | 1              | 9                         | 5                           | 6                |
| Frankfurt       | $-4$           | $-3$           | 2                         | $-3$                        | $-4$             |
| Dortmund        | $-4$           | $-8$           | 0                         | 0                           | 2                |
| Wolfsburg       | 0              | 12             | $-4$                      | $-5$                        | $-2$             |
| Hertha          | $-5$           | 0              | $-5$                      | $-8$                        | $-5$             |
| Bochum          | $-2$           | $-4$           | $-1$                      | $-4$                        | $-7$             |
| Bielefeld       | $-19$          | $-6$           | $-6$                      | $-11$                       | $-10$            |
| Rostock         | $-10$          | $-12$          | $-8$                      | $-11$                       | $-13$            |
| Nürnberg        | $-7$           | $-9$           | 1                         | 1                           | 1                |
| Cottbus         | $-10$          | $-11$          | $-8$                      | $-10$                       | $-12$            |
| Duisburg        | $-12$          | $-7$           | $-8$                      | $-13$                       | $-12$            |

The estimation in the final column also involves information about the market value. The actual goal differences of the first half of that season and the second half are included in the first two columns, respectively.

doi:10.1371/journal.pone.0104647.t003

### Table 4.

The K-value for the regression model during the seasons 2002/03 and 2006/07 as well as for the Oddset-odds.

|                  | first 10 matches of season | all 34 matches |
|------------------|-----------------------------|----------------|
| Only home advantage | 1.073                      | 1.057          |
| + matches of present season | 1.054                      | 1.013          |
| + matches of previous season | 1.027                      | 1.004          |
| + market value | 1.019                      | 1.000          |
| Oddset         | 1.025                      | 1.012          |
| Difference     | 0.006 ± 0.009              | 0.012 ± 0.004  |

The impact of adding additional information to the model is listed.

doi:10.1371/journal.pone.0104647.t004
minimum if the predicted probabilities for a win, a draw, and a loss are identical to the true probabilities. Analogous measures can be already found in literature, e.g. [10,24]. One can see in Tab.4 the additional consideration of new information indeed gives rise to a lower value of $K$. Furthermore, restricting the choice of matches to those taking place during the first 10 match days, the prediction becomes worse (larger $K$). In particular, the additional impact of the market value is larger, if restricting oneself to the first 10 matches of the season. When averaging over all matches in these seasons [22], it turns out that the $K$-value of the present approach is smaller than the $K$-value for the Oddset-odds by 0.012 ± 0.004. Thus, the comparison yields a highly significant improvement of the present model as compared to the Oddset-odds. The size of this improvement is non-negligible if compared to the variations of $K$ when adding different pieces of information; see Tab.4.

Discussion

The main goal of this work is to provide a theoretical framework which allows one to determine the quality of the prediction. Conceptually, it is related to the Bayesian approach because it takes into account the impact of additional information as well as the impact of decorrelations on the estimation of future events. As a formal framework we have used a multivariate regression approach.

The prediction of soccer results is a particularly nice case study of this approach due to the availability of well-defined data and due to the popular interest in this matter. Beyond the application of the analytical results it turned out to be essential to search for observables (here: chances for goals) with a high information content.

One interesting question arises: is the residual statistical error of $S_Z$ for $N_Y=17$ small or large? This question may be discussed from two different perspectives. First, one may want to predict the outcome of the second half of the league. Then the uncertainty is given by $17\sqrt{\hat{\gamma}(X,Y)} = 17\sqrt{\hat{\gamma}(X,Y) + V_Z}/17$. These values are plotted for different prediction scenarios in Fig.5. One can see how the additional information decreases the uncertainty of the prediction. Most importantly, the no man's land below an uncertainty of $\sqrt{17V_Z}=7.1$ cannot be reached by any type of prediction. The art of approaching this perfect prediction thus resorts to decrease the present value of 7.8 to a value closer to 7.1. Second, one may be interested in the prediction of a single match. This case is somewhat different. Since the team fluctuations are very difficult to predict, the fluctuation amplitude $A=0.17$ (see above) serves as a scale for estimating the highest possible quality of match prediction. If the uncertainty is much smaller than $A$ any further improvement would be irrelevant due to the non-predictable fluctuations of the team strength. However, in the present case the statistical error after $N_Y=17$ is close to $A$ so that a further reduction of $\sqrt{\hat{\gamma}(X,Y)}$ would still be relevant for prediction purposes of individual matches.

Repeating this analysis for the prediction of the points in the second half of the season the statistical uncertainty of the estimation corresponds to approx. 6 points (standard deviation). This corresponds to lose rather than to win two matches or vice versa.

References

1. Wooldridge JM (2002) Econometric analysis of cross section and panel data. Cambridge: MIT Press.
2. Goldstein M, Wooff D (2005) Statistical Models: Theory and Practice. Cambridge University Press.
3. Freedman D (2007) Bayes Linear Statistics, Theory and Methods. Wiley.
4. Laird N, Ware J (1982) Random-effects models for longitudinal data. Biometrics 38: 963–974.
5. Heuer A, Rubner O (2009) Fitness, chance and myths: an objective view on soccer results. Er Phys J B 67: 445–458.
6. Constantinou A, Fenton NE, Neil M (2012) pi-football: A Bayesian network model for forecasting Association Football match outcomes. Knowl-Based Syst 36: 322–339.
7. Lee A (1997) Modelling scores in the premier league: is manchester united really the best? Chance 10: 15–19.
8. Dixon M, Coles S (1997) Modelling association football scores and inefficiencies in a football betting market. Appl Statist 46: 263–280.
9. Dixon M, Robinson M (1998) A birth process model for association football matches (pages 523–538), The Statistician 47: 525–538.
10. Rue H, Salvesen O (2000) Prediction and retrospective analysis of soccer team standings and streaks. Eur Phys J B 67: 473–481.
11. Maher M (1982) Modelling association football scores. Statistica Neerlandica 36: 109–118.
12. Bittner E, Nussbaumer A, Janke W, Weigel M (2007) Self-affirmation model for football goal distributions. Europhys Lett 76: 58002.
13. Bitter E, Nussbaumer A, Janke W, Weigel M (2009) Football fever: goal distributions and non-gaussian statistics. Eur Phys J B 67: 459–471.
14. Koning R (2000) Balance in competition in dutch soccer. The Statistician 49: 419–431.
15. Goddard J (2005) Regression models for forecasting goals and match results in association football. International Journal of Forecasting 21: 331–340.
16. Hvattum LM, Arntzen H (2010) Using elo ratings for match result prediction in association football. International Journal of Forecasting 26: 460–470.
17. Andersson P, Edman J, Ekman M (2003) Predicting the world cup 2002 in soccer: Performance and confidence of experts and non-experts. International Journal of Forecasting 21: 565–576.
18. Song CU, Boulier BL, Stekler HO (2007) The comparative accuracy of judgmental and model forecasts of american football games. International Journal of Forecasting 23: 405–413.
19. Forrest D, Simmons R (2000) Forecasting sport: the behaviour and performance of football tipsters. International Journal of Forecasting 16: 317–331.
20. Heuer A, Mueller C, Rubner O (2010) Soccer: is scoring goals a predictable poissonian process? Europhys Lett 89: 38007.
21. Heuer A (2012) Der perfekte Tipp: Statistik des Fussballspiels. Wiley–VCH.
22. Heuer A, Rubner O (2012) How does the past of a soccer match influence its future? PLoS ONE 7/11: e47678.
23. Goddard J, Asimakopoulos I (2004) Football results and the efficiency of fixed-odds betting. J Forecast 32: 51–66.
24. Judge T, Hurst C (2008) How the rich (and happy) get richer (and happier): Relationship of core self-evaluations to trajectories in attaining work success. Journal of Applied Psychology 93: 849–863.
25. Schnitzlein D (2014) How important is the family? evidence from sibling correlations in permanent earnings in the usa, germany, and denmark. J Popul Econ 27: 69–89.
26. Ambrey C, Fleming C (2014) The causal effect of income on life satisfaction and the implications for valuing non-market goods. Economics Letters 123: 131–134.