Journal of Data Science **15**(2017), 293-312

INVESTIGATING THE UNDERLYING CAUSAL NETWORK ON EUROPEAN FOOTBALL TEAMS

Pedro H.R. Cerqueira[1], Luiz R. Nakamura[2], Rodrigo R. Pescim[3], Roseli A. Leandro[1]

[1]*Departamento de Ciências Exatas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo*

[2]*Departamento de Informática e Estatística, Universidade Federal de Santa Catarina*

[3]*Departamento de Estatística, Universidade Estadual de Londrina*

Abstract: Football, or soccer, is considered one of the most important col- lective sports in the world. Managers, specialists and fans are always trying to find out the important keys to have a good team. The evaluation of the team quality may present many variables and subjective concepts, and for this reason, it is not simple to answer the following question: How to define quality? Another point that should be considered is the importance of aspects such as offensive and defensive: Which one is more important to measure quality of a football team? For this task, we propose the use of a causal model with latent variables as a tool to measure the subjectivity of the team quality and how it can be affected by other aspects. Information from the four most important football leagues in the world (England, Germany, Italy and Spain) during three seasons (2011-2012; 2012-2013; 2013-2014) was collected. We defined the latent variables in the model and evaluated the relationships among them. The results show that the offensive aspect exert more influence on team quality than defensive aspect, which reflects directly on the players market strategies.

Key words: Collective sports, latent causal models, match analysis, soccer, structural equation model.

1. Introduction

Football or soccer, name assigned in USA, is a collective sport played by two teams with eleven players each. Football is considered one of the most popular

and important sports in the world, being played in every nation without exception (Reilly and Williams, 2003). One example of its popularity is the number of countries affiliated to the Fédération Internationale de Football Association (FIFA), which is higher than the ones affiliated to the United Nations (UN) and International Olympic Committee (IOC) (Louzada et al., 2014).

The simplest way to explain the game objectives is to focus in the results. Both teams play against each other in a field with their players composed mostly of a goalkeeper, defenders, midfielders and strikers. They try for 90 minutes, split in two periods, to score goals and avoid then. It is considered a goal when the ball crosses the goal line between the goalposts (Reilly and Williams, 2003; Louzada et al., 2014). Unlike many other sports, instead of only two possible results (win and lose), football allows a third one, the draw. The team that scores the most goals is considered to be the winner, and when both teams score the same amount of goals it is considered a draw. The score is assigned to each team at the end of the event match: three points are given for the winner team, zero points for the loser team and if a draw occurs one point is assigned for each team (Reilly and Williams, 2003; Louzada et al., 2014).

Football also carries an economic issue that can be seen in many different aspects. Every year the amount of spent money with players transfers potentially increases, mainly in European most important leagues, such as Barclays Premier League (England), Bundesliga (Germany), Serie A (Italy), BBVA league (Spain), among others. Some cities use the football games as a tourist attraction, specially every four years when the world cup happens, the biggest event of football wherein currently 32 nations, initially divided into eight groups of four teams each, is held in a different country, mobilizing not only those participating countries but many others around the world (Lee and Taylor, 2005).

Although the FIFA World Cup is the most famous and important football event, it has a huge "lucky" effect on it, since each team does not play against all other teams, i.e., an arbitrary team could be benefited or harmed depending on which teams are in the same group as yours in the first phase of the championship and hence, it does the search for quality to be more complex and in some situa- tions it is almost impossible to find a pattern and/or a consensus about quality. In order to avoid the "lucky" effect, we performed our study using the four most important football leagues of the world (English, German, Italian and Spanish) for three different seasons (2011–2012, 2012–2013 and2013–2014).

Different scientific studies focused in specific objectives related to football have been widely proposed in the last decades. For instance, in medical sciences some studies are related to fitness, e.g. performing better strategies to improve the strength, stamina and to avoid injuries of the players, or trying to assess whether a football player is able to return without any risk (for further details,

see Delvaux et al., 2014; Meckel et al., 2014; Stubbe et al., 2015). Also, some studies have been aimed in predictions about possible results in a specific match or a championship as in Goddard (2005), Suzuki et al. (2010) and Louzada et al. (2014), or analyzing external factors that may directly influence the match outcome as proposed by Nevill et al. (1996), Taylor et al. (2008) and Staufenbiel et al. (2015) or even to study the football betting market as in Dixon and Coles (1997), Dixon and Pope (2004) and Goddard and Asimakopoulos (2004). Recently, the study of game-related statistics has been receiving considerable attention from researchers, football industry and specialists as a powerful mechanism to improve a team, since it can measure its quality and highlight its most important players (see, e.g. Poulter, 2009; Castellano et al., 2012; Moura et al., 2014).

However, the evaluation of the team quality may present many variables and subjective concepts and for this reason it is not simple to answer the following question: How to define quality of a football team? Moreover, another point that should be considered is the importance of offensive and defensive aspects. Which one can be considered more important to measure the quality of a team? A suitable answer for these questions can be derived using the concept of causal models under latent variables, that allow us to measure those subjective concepts of the team quality and how it could be affected.

The amount of researchs using causal models have increased during the past decades and it became an important tool to verify causal relationship between systems that contain observed variables, specially in human sciences, where they are usually trying to study causal effects concerning to subjective aspects such as intelligence, aspirations or political interventions (further details can be seen in Haavelmo, 1943; Duncan et al., 1968; Bollen, 1995; Lee and Zhu, 2000; Bollen, 2002; Ferron and Hess, 2007; Greene, 2011).

For the use of causal models two aspects ought to be considered: Graph analysis (GA) and structural equation model (SEM). The GA involves searching for causal structures that qualitatively represent how variables are causally connected, while the SEM with a well-known causal structure allows to infer the magnitude of causal relationships. Also, SEM can be considered as multiple-trait regression models in which some response variables may be represented as covari- ates in the right-side of the equations for the other response variables (Lee and Zhu, 2000; Lee and Tang, 2006; Rosa et al., 2011).

In the literature, causal models have been widely studied under two approaches. The first approach uses the latent variables and then relates the causal structure among latent variables. This approach is interesting for situations in which subjectivity aspects or unmeasured variables are used and their relationships are able to infer the causality. The second approach uses the structure

without latent variables, that should be considered when the variables are measured and the relationship between them are used to infer causality (Duncan et al., 1968; Lee and Zhu, 2000; Bollen, 2002; Lee and Tang, 2006; Rosa et al., 2011).

The GA has some particular notations that should be mentioned and are necessary for a better comprehension of this model: i) variables inside a circle are called latent variables; ii) variables inside a rectangle are the observed variables; iii) arrows represent a causal effect; and iv) double arrows represent correlation. Moreover, we classify the explanatory variables as exogenous and the response variables as endogenous, and also this notation can be extended to latent variables.

The remainder of the paper is outlined as follows. In Section 2, we introduce a brief description of the data set. We discuss some statistical inference for the causal models via structural equation model such as maximum likelihood ratio method and some model selection criteria, in Section 3. The results given in Section 4 reveal the usefulness of the selected causal model under latent variable for analyzing real data. Concluding remarks are addressed in Section 5.

2. Data set

We choose to use championships as league in order to minimize the "lucky" effects such as a bad day, bad draw, or any external intervention, that could happen in championships as cups. The data used in this paper comes from the four most important football leagues affiliated to Union of European Football Association (UEFA) in Europe (Barclays Premier League from England, Bundesliga from Germany, Serie A from Italy and BBVA league from Spain) related to the past three seasons (2011–2012, 2012–2013 and 2013–2014).

All of these leagues present the same structure, where all teams play against each other twice, i.e., home and away game. Despite all similarity among them, Bundesliga is composed by 18 teams in a total of 34 games whereas the other leagues under study represent 20 teams in a total of 38 games. Another difference among those leagues is the way that the teams who will be playing the UEFA Leagues (Champions and Europe) and the number of relegations are chosen. To avoid any problems with the different amount of games for each championship, we used all information per game.

The information evaluated in this study consists in 32 different variables: win (total, home and away), draw (total, home and away), lose (total, home and away), points rate (total, home and away), goals favor, goals against, goals difference, shots, shots on goal, clean sheet, offsides, fouls, yellow and red cards, fouled (received fouls), tackles, interception, possession, dribble, shot conceded, pass accuracy, position, classification to UEFA league and relegation. This data

set is available for consulting at http://www.whoscored.com. Table 1 presents a descriptive summary for some variables of data set divided by leagues using the three above mentioned seasons.

Table 1: Descriptive statistics for some observed variables

Barclays Premier League (England)					
Variable	SD	Min	Median	Mean	Max
Win (%)	16.80	10.526	31.579	37.76	73.68
Home points rate (%)	18.07	21.100	49.100	53.42	96.50
Away points rate	15.31	14.000	35.100	38.41	73.70
Goals favor	0.449	0.737	1.237	1.395	2.684
Clean sheet (%)	10.91	2.632	27.632	27.85	52.63
Dribbles	2.080	3.000	7.350	7.297	12.60
Shots	2.514	9.900	13.400	13.87	19.40
Shots on goal	1.044	2.500	4.300	4.580	6.842
Tackles	1.438	15.800	19.150	19.05	22.30
Interception	2.623	9.800	15.150	15.08	20.50
Fouls	0.983	8.400	10.800	11.00	12.80
Pass accuracy (%)	4.820	69.500	79.100	79.38	86.00
Yellow cards	0.238	1.053	1.553	1.559	2.053
Red cards	0.047	0.000	0.079	0.075	0.237
Received fouls	0.952	8.700	10.300	10.43	13.10
Offsides	0.427	1.300	2.200	2.198	3.400
Bundesliga					
Variable	SD	Min	Median	Mean	Max
Win (%)	17.06	11.765	35.294	37.96	85.29
Home points rate (%)	17.74	7.800	52.900	53.12	90.20
Away points rate	16.51	13.700	36.300	38.85	92.20
Goals favor	0.479	0.706	1.382	1.492	2.882
Clean sheet (%)	10.92	8.824	23.529	24.78	61.76
Dribbles	2.621	7.900	13.350	13.70	20.20
Shots	2.114	8.900	12.800	13.14	18.70
Shots on goal	1.062	3.100	4.600	4.820	7.600
Tackles	2.372	16.800	22.100	22.25	26.50
Interception	3.523	9.600	18.800	18.41	27.60
Fouls	1.797	10.700	16.100	15.79	19.90
Pass accuracy (%)	3.611	69.500	77.750	77.44	88.30
Yellow cards	0.310	0.971	1.838	1.781	2.235
Red cards	0.052	0.029	0.088	0.098	0.235
Received fouls	1.284	11.500	15.100	15.09	17.40
Offsides	0.591	1.600	2.800	2.730	4.100

Table 1: Descriptive statistics for some observed variables (continued)

Serie A (Italy)					
Variable	SD	Min	Media	Mean	Max
Win (%)	15.447	10.526	34.211	36.974	86.842
Home points rate (%)	15.924	22.800	57.000	55.258	100.000
Away points rate (%)	15.613	8.800	33.300	36.052	78.900
Goals favor	0.366	0.632	1.250	1.320	2.105
Clean sheet (%)	10.299	7.895	30.263	29.868	57.895
Dribbles	1.777	5.300	8.550	8.832	13.100
Shots	2.098	9.200	13.100	13.322	19.100
Shots on goal	0.922	2.800	4.300	4.427	6.900
Tackles	1.730	17.600	22.350	22.100	25.300
Interception	2.345	12.600	16.150	16.530	21.900
Fouls	1.575	12.200	15.450	15.120	18.000
Pass accuracy (%)	3.151	73.300	79.900	80.473	86.400
Yellow cards	0.305	1.605	2.303	2.316	2.895
Red cards	0.074	0.026	0.132	0.149	0.342
Received fouls	1.326	11.900	14.200	14.330	17.700
Offsides	0.618	1.400	2.400	2.505	4.300
BBVA League (Spain)					
Variable	SD	Min	Media	Mean	Max
Win (%)	16.216	10.526	34.211	38.421	84.211
Home points rate (%)	16.193	28.100	52.600	56.497	96.500
Away points rate (%)	15.732	12.300	31.600	35.790	87.700
Goals favor	0.576	0.737	1.224	1.397	3.184
Clean sheet (%)	8.298	10.526	26.316	28.728	52.632
Dribbles	2.217	3.800	6.850	7.265	14.100
Shots	2.187	9.400	12.750	12.923	19.600
Shots on goal	1.150	3.100	4.300	4.583	7.900
Tackles	1.929	17.500	21.450	21.770	27.300
Interception	7.454	13.200	17.600	21.027	36.800
Fouls	1.826	10.600	14.350	14.293	18.000
Pass accuracy (%)	4.942	67.100	77.400	76.577	89.500
Yellow cards	0.485	1.474	2.605	2.667	3.763
Red cards	0.074	0.026	0.171	0.164	0.342
Received fouls	1.220	10.300	13.450	13.413	16.800
Offsides	0.536	1.700	2.400	2.600	4.200

Some variables were omitted because they summarize and express the same information. The greater average performance in home games is observed in the

BBVA League, while on the away games is on the Bundesliga. If we consider all leagues, the performance in home games is almost 60% percent greater than the performance of away games. In general, the best attack in a season scores almost four times more than the worst attack. For any team belonging to the English league, we can see that at least 2% of the games were finished without being scored, while for teams from other leagues this minimum percentage vary from 7.895 up to 10.526. Bundesliga presents the team who avoided being scored the most (61.765%). In average, around 25% of the shots made were in the goal direction in all leagues. All teams along the season obtained at least more than 60% of passing accuracy, received at least almost one yellow card per game and in average received a red card each 10 games.

We split the data set into five possible groups that present similar characteristics in the game field, e.g the amount of fouls and cards, since a player could receive a card according to the amount of fouls in the game or their intensity, or the shots or offsides, since several shots on goal come from through ball.

3. Causal Inference

In this section, we are interested in creating some variables which could represent different subjective aspects, such as offensive, defensive, quality, etc. After that, we are able to use the causal models under the latent variables framework to perform the inferential procedures. In order to achieve the best possible model, we propose different relationships between variables with different latent structures and also allow the covariance relationships between all observed and latent variables.

3.1 Structural equation model

Here, we use the structural equation model to estimate the effects above. We can note that SEM consists in two distinct parts. The first part is due to the development of a set of equations related to the causal relations between latent variables (further details in Bollen, 1995; Lee and Zhu, 2000; Bollen, 2002; Lee and Tang, 2006). The model can be expressed as

$$\eta = B\eta + \Gamma\xi + \zeta$$

where η represents the vector of latent endogenous variables, B is the matrix of loading coefficients that gives effects of η_j on η_i with diagonal equal to zero, Γ represents the matrix of loading coefficients that gives the effects of ξ_j on η_i, ξ is the vector of latent exogenous variables which follows a multivariate normal distribution with mean 0 and covariance matrix given by Φ, and ζ is the vector of errors for the latent variable η, which has multivariate normal distribution

with mean 0 and covariance matrix given by Ψ. Here, we assume that ξ and ζ are not correlated.

We can note that η has a multivariate normal distribution with mean equals to 0 and covariance matrix given by $(I - B)^{-1} [\Gamma\Phi\Gamma^t + \Psi] (I - B)^{-1^t}$. The second part of SEM is used to verify how the observed variables are related to latent variables. The model for the observed variables structure can be written as

$$Y = \Lambda y\, \eta + \varepsilon \qquad and \qquad X = \Lambda x\, \xi + \delta,$$

where X is the matrix of observed variables related to latent exogenous variables with dimension ($m \times kx$), Y is the matrix of observed variables related to latent endogenous variables with dimension ($n \times ky$), ε with dimension ($n \times ky$), and δ, with dimension ($m \times kx$), represent the matrix of errors in equations with covariance matrix given by $\Theta\varepsilon$, with dimension ($p \times p$), and $\Theta\delta$, with dimension ($q \times q$), respectively.

The joint probability density function for the observed variables X and Y follows a multivariate normal distribution $N(0, \Sigma)$ and the covariance matrix, Σ, is given by

$$\boldsymbol{\Sigma}(\Theta) = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

where

$\Sigma_{xx} = \Lambda_x \Phi \Lambda'_x + \Theta_\delta$,

$\Sigma_{yy} = \Lambda_y[(I - B)^{-1}][\Gamma\Phi\Gamma' + \Psi][(I - B)^{-1}]'\Lambda'_y + \Theta_\varepsilon$,

$\mathrm{Cov}(\eta, \xi) = (I - B)^{-1}\Gamma\Phi$,

And

$\Sigma_{xy} = \Sigma'_{yx} = \Lambda_y (I - B)^{-1}\Gamma\Phi\Lambda'_x$,

As described by Crisci (2012), we are interested in solving $S = \Sigma(\Theta)$, where S represents the covariance matrix based on the empirical data and the difference between S and $\Sigma(\Theta)$ is named as discrepancy function. Using the method based on maximum likelihood ratio, we have the total log-likelihood function for Θ, represented by

$$F_{ML} = \log|\Sigma(\Theta)| + tr\big(S\Sigma'(\Theta)\big) - \log|S| - (p + q)$$

where p and q are the parameters related to each covariance matrix. The max- imum likelihood estimate (MLE) $\hat{\Theta}$ of Θ is the solution of the score vector for Θ

3.2 Model selection

In this section, we shall apply different measures as tools to verify (among all models considered) which should usually be taken as the best model for describing the given data set.

In the SEM's context, a model is considered suitable if the covariance structure implied by the model is similar to the covariance structure of the sample data, as indicated by an acceptable value of goodness-of-fit index (GFI) (Cheung and Rensvold, 2002). In the literature, the most popular GFI used in SEM is the χ^2 statistic. However, a problem arises because χ^2 statistic has a sample size dependence. For instance, the χ^2 statistic provides a highly sensitive statistical test for large sample sizes, but not a practical one.

To overcome this problem, many authors have been proposed GFIs as alternative to χ^2 statistic in last decades. Some of them are the Comparative Fit Index (CFI) (Bentler, 1990), Tucker-Lewis Index (TLI) (Tucker and Lewis, 1973), Normed Fit Index (NFI) (Bentler and Bonett, 1980) and root mean squared error of approximation (RMSEA) (Steiger, 1989). In this paper, we performed the methods suggested by Bollen (1995) and Kline (2011), i.e the CFI, TLI and RMSEA.

Comparative Fit Index (CFI)

The CFI is an incremental fit index that measures the relative improvement in the fit of the proposed model over that of a baseline model, typically the independence model. Its formula can be expressed as

$$\mathrm{CFI} = 1 - \frac{max\left(\hat{C}_m - df_m, 0\right)}{max\left(\hat{C}_b - df_b, 0\right)}$$

where $\hat{C}m$ and $\hat{C}b$ are the sample minimum discrepancy for the proposed and baseline models, respectively and dfm and dfb are the degrees of freedom for the proposed and baseline models.

Tucker-Lewis Index (TLI)

The TLI is an incremental fit index which was developed against the disadvantage of Normed Fit Index regarding being affected by sample size. TLI is

calculated as givebelow

$$\text{TLI} = \frac{(X_\sigma^2/df_b)(X_m^2/df_m)}{(X^2/df_b) - 1}$$

where χ_m^2 and dfm are the chi-square and degrees of freedom for the proposed model while χ^2 and df_b are the chi-square and degrees of freedom for the baseline model. The bigger TLI value indicated better fit for the model. The most advantage of this fit index is the fact that TLI is not affected significantly from sample size.

Root Mean Squared Error of Approximation (RMSEA)

In recent years, the RMSEA has become regarded as one of the most informative fit indexes due to its sensitivity to the number of estimated parameters in the model. (Diamantopoulos and Siguaw, 2000) In other words, the RMSEA favours parsimony in that it will choose the model with the lesser number of parameters (Hooper et al., 2008).

The RMSEA is computed based on sample size and the non-centrality parameter and degrees of freedom for the proposed model given by

$$\text{RMSEA} = \sqrt{\frac{\hat{F}_\theta}{df_m}}$$

where $\hat{F}_\theta = \max\left(\frac{\hat{c} - df_m}{n}, 0\right)$ and dfm is the degrees of freedom for the proposed model.

For the first two measures (CFI and TLI) values close to one indicate the better models and for the RMSEA values smaller than 0.05 are considered better acceptable models. All the computation were performed using lavaan, simsem, semPlot and sem Tools packages available in the statistical software R (R Core Team, 2015).

4. Results and discussion

For the data set described in Section 2, we create five latent variables based on the observed variables in order to explain several subjective aspects that specialists usually bring forward during discussion regarding football. These aspects are defensive, offensive, discipline, creation and quality. Subsequently, we introduce the causal relationship among all latent variables and consider a structure for selecting the best model based on the three measures mentioned in Section 3.2. In order to estimate the selected model, we consider the maximum likelihood ratio

method, discussed in Section 3.1. Table 2 lists the estimates of the parameters and the relationships between observed and latent variables.

Table 2: Relationship between latent variables and their exogenous variable

Latent	Exogenous variable	Estimate
Offensive	Goals Favor	1.000
	Shots	0.776 (0.055)
	Shots on Goal	0.905 (0.037)
	Offsides	0.208 (0.070)
	Wins	1.091 (0.035)
Creation	Passes	1.000
	Possession	1.098 (0.059)
	Interception	-0.288 (0.089)
	Dribbles	0.530 (0.097)
Defense	Goals Against	1.000
	Clean Sheet	-0.874 (0.042)
	Shots conceded	0.841 (0.062)
Discipline	Fouls	1.000
	Yellow Cards	1.131 (0.282)
	Red Cards	0.932 (0.279)
Quality	Points rate	1.000
	classification	-0.847 (0.035)
	Goals Difference	0.972 (0.017)
	Home Points Rate	0.935 (0.023)
	Away Points Rate	0.925 (0.025)
	Position	-0.933 (0.024)

4.1 Latent variable for the football data

In this section, we create five latent variables (offensive, defensive, creation, discipline and quality) based on the observed variables from the football data set. We also give some comments about the relations (positive or negative) between latent and observed variables.

Offensive

It is suggested that the latent variable offensive is composed by goals favor, shots, shots on goal, offsides and wins, and all these variables are positively related to offensive. Also, it is possible to verify that wins and goals favor are the

variables that present more contributions to offensive aspect. On the other hand, it is possible to observe that offsides is the variable that contributed less (around 21% of the effect of goals). These relations make sense since for any victory at least a score is needed. Further, we can observe a high relation between shots on goal and goals, as well as the offsides and goals, since a lot of creation in football come from through ball.

Creation

Creation is positively related to percentage of passes completed, possession and dribbles while it is negatively related to interceptions. Passes and ball possession are the variables that can better explain the creation variable. In this case, interception is considered negative since these results come from the fact that the ball is in possession of the other team.

Defensive

The latent variable defensive is defined by goals against, shots conceded and clean sheet. The first two variables are positively related to defensive while the clean sheet has a negative relation. In absolute values, we observed that clean sheet and shots conceded are equivalent. These relations are well expressed, under the game point of view, since a team that spends more goals without being scored is expected to receive less goals during the whole season.

Discipline

Discipline was positively related to fouls, yellow and red cards. We can observe that the difference between the smaller and greater value is around 22%. These relationships can explain what actually happens on a football field, since the players can receive the cards for several reasons such as the amount of fouls or the intensity of fouls.

Quality

Quality is positively related to points rate, goals difference, home and away points while it is negatively related to classification to European leagues and posi- tion. These relations can be explained by the reason that for a good classification it is expected a higher punctuation at home and away games. Goal difference has the same explanation because more victories implies more goals in favor than against.

4.2 Causal relationship between latent variables

After the development of the offensive, defensive, creation, discipline and quality variables, we proposed the causal relationship among them and several scenarios were provided to achieve a structure which could be represented for the best model. The values of CFI, TLI and RMSEA measures for the best model are 0.982, 0.974 and 0.069, respectively. All relationships between latent and observed exogenous variables are presented in Table 2 and the causal relationship among all of latent variables is displayed in Figure 1.

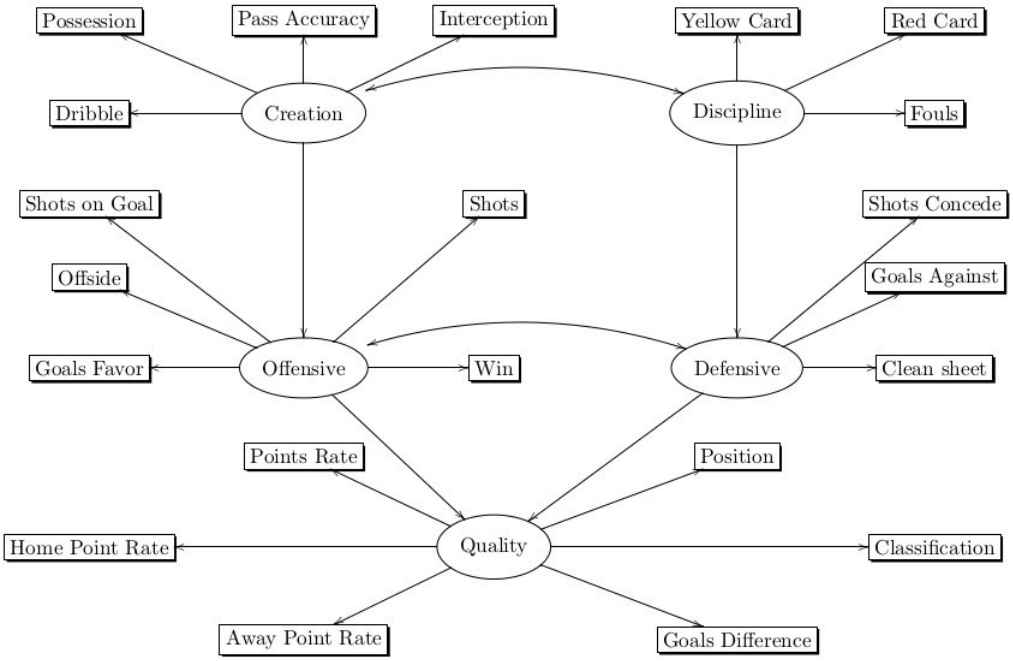

Figure 1: Path Diagram representing the relationship between observed variables and latent variables. The correlation among observed variables are omitted to show a cleaner path diagram.

We can observe that the offensive and defensive characteristics are correlated to each other as well as discipline and creation aspects, without any causal meaning and for this reason these relations are expressed by two-headed curved arrows. On the other hand, we observe that creation and discipline present direct cause effects on offensive and defensive aspects, respectively, which are represented by two single headed straight arrows. Also, it is possible to visualize that the offensive and defensive characteristics affect directly the football team quality.

Table 3 shows that the creation variable is considered as cause of the offensive variable which exerts an effect equals to 1.088. In the same way, we can observe

that the discipline variable causes an effect on the defensive aspect, considering the effect equals to 2.554. Moreover, both variables (defensive and offensive) affect the team quality with coefficients equal to -0.194 and 0.920, respectively. According to Table 3, we can also infer that the discipline variable has an indirect effect (discipline effect × defensive effect) in quality equal to -0.4954 while the creation variable has an indirect effect (creation effect × offensive effect) in quality equal to 1.00096.

Table 3: Estimated effects (standard errors in parentheses) of the exogenous and endogenous latent variables

Endogenous latent	Exogenous latent variable	Estimate
Offensive	Creation	1.088 (0.128)
Defensive	Discipline	2.554 (0.698)
Quality	Defensive	-0.194
	Offensive	0.920 (0.042)

Based on Table 3, we can assume that interventions can be used in order to improve offensive characteristics, which implies a gain in the team quality almost five times more than interventions realized on defensive aspects. We also note that the indirect effect provided by creation is almost the same as the direct effect provided by offensive in relation to team quality.

The negative effect between defensive and quality variables is explained by the fact that defensive variable is related to goals against and then, the more goals conceded by a team, the worse is its quality. The positive effect between discipline and defensive variables can be explained by the fact that fouls generate more chances for the team shot on goal and thus more scores may be done.

4.3 Observed and latent variables correlated

In this Section, we also make some comments about the correlations (positive or negative) of the observed and latent variables. Table 4 shows the correlation among some observed variables which were considered statistical significant for the fitted model and the covariance between latent variables.

Table 4: Correlation among observed variables

Variable 1	Variable 2	Estimate	Variable 1	Variable 2	Estimate
Goals Favor	Goals Difference	0.068	Shots	Pass	0.186
	Shots on Goal	0.084		Interception	-0.103
	Clean Sheet	-0.161		Shots on Goal	0.263
	Goals Against	0.129		Fouls	-0.094
Shots on Goal	Clean Sheet	-0.062	Shots Conceded	Shots	-0.16
	Pass	0.087		Fouls	-0.111
	Posse	0.123		Goals Difference	-0.042
	Goals Against	0.061		Yellow Card	-0.118
	Goals Difference	0.026		Shots on Goal	-0.102
Offside	Interception	0.225	Dribble	Shots on Goal	0.081
	Fouls	0.16		Yellow Card	-0.201
	Pass	-0.085		Fouls	0.341
Win	Fouls	-0.013	Goals Against	Goals Difference	-0.049
	Shots Conceded	-0.016		Clean Sheet	-0.132
Pass	Interception	-0.189	Yellow Card	Red Card	0.429
	Fouls	-0.15	Home Point Rate	Away Point Rate	-0.135
Possession	Shots Conceded	-0.117	Fouls	Yellow Card	0.416
	Fouls	-0.119		Red Card	0.333
	Classiftcation	0.109		Possession	-0.026
	Shots	0.198		Interception	0.104
	Pass	0.221			

Offensive, defensive, creation and discipline

We can observe that offensive and defensive are negatively related to each other as well as the creation and discipline, both results are perfectly explained, given that the team which presents more creation is more disciplined because it has more possession of the ball during the match and consequently its number of fouls and cards (yellow and red) will be smaller. In the same sense, we can explain the negative correlation between offensive and defensive aspects, considering that, during a match, more shots and time on offense represent less opportunities to the opponent.

Fouls, yellow and red cards

For the observed variables, we note that yellow cards are positively correlated to red cards and fouls, and also fouls are positively related to red cards by the reason that, in general, many red cards are given after the yellow cards.

Offsides, interception, goals against, clean sheet

The offsides are positively related to interception, in the game context this relation is explained by the reason of some interceptions are directly linked to the offensive aspect and on most of the time some players are not paying properly

attention and do not follow the game speed. Goals against present a negative covariance in relation to clean sheet. This fact leads us to observe that games with lots of goals are not the standard during the championship.

Possession, pass accuracy, shots conceded

Possession is positively related to shot and pass accuracy, and negatively related to shots conceded and fouls. These results are widely discussed by specialists because more possession leads to less opportunities for the opponent and then it concedes less shots and prevents the defense to do many fouls. On the other hand, more possession means that a team usually presents a better pass accuracy and thus, it creates more chances to shots.

5. Concluding remarks

In this paper, we proposed the use of causal models under latent variables for the task of measuring the football teams quality. We noted that this approach allowed us to measure those subjective concepts of the teams quality and how it could be affected by others characteristics. In order to avoid the "lucky" effect, we performed our study using the four most important football leagues of the world (English, German, Italian and Spanish) for the last three seasons. We also discussed some statistical inference for the causal models through the structural equation model using the maximum likelihood ratio method and selected the model by CFI, TLI and RMSEA measures. The results revealed that the team quality is explained by offensive aspect around five times more than the defensive characteristic and also the creation variable exerted an important effect on team quality. Furthermore, the results expresses the strategies related to the players market well, where the most valuable players (higher salaries and sponsorship values), generally presents offensive skills which appears more developed, such as, midfielders, forwards and strikers. The importance of the players with offensive skills is noted in the best player of the year awards, where in 24 editions, only once a player which plays on the first half of the field received the prize. Moreover, we have evidenced that the stand for the usage of causal models as an efficient tool to explain and quantify is useful in terms of the relationships which are always treated as opinions for many specialists.

Acknowledgments

The authors are grateful to the editor and an anonymous referee for helpful comments and suggestions. We gratefully acknowledge grants from CNPq and CAPES (Brazil).

References

[1] Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin* **107**, 238–246.

[2] Bentler, P.M. and Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* **88**, 588–606.

[3] Bollen, K.A. (1995). *Structural Equations with Latent Variables*, New York: John Wiley Sons Inc.

[4] Bollen, K.A. (2002). Latent variables in psychology and the social sciences.

[5] Annual Review of Psychology **53**, 605–634.

[6] Castellano, J., Casamichana, D. and Lago, C. (2012). The use of match statis- tics that discriminate between successful and unsuccessful soccer teams. *Journal of Human Kinetics* **31**, 139–147.

[7] Cheung, G.W. and Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Mul- tidisciplinary Journal* **9**, 233–255.

[8] Crisci, A. (2012). Estimation methods for the structural equation models: maximum likelihood, partial least squares e generalized maximum entropy. *Jour- nal of Applied Quantitative Methods* **7**, 3–17.

[9] Delvaux, F., Rochcongar, P., Bruyére, O., Bourlet, G., Daniel, C., Diverse, P., Reginster, J.Y. and Croisier, J.L. (2014). Return-to-play criteria after hamstring injury: actual medicine practice in professional soccer teams. *Journal of Sports Sciences and Medicine* **13**, 721–723.

[10] Diamantopoulos, A. and Siguaw, J.A. (2000). *Introducing LISREL*. London: SAGE Publications Ltd.

[11] Dixon, M.J. and Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C – Applied Statistics* **46**, 265–280.

[12] Dixon, M.J. and Pope, P.F. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting* **20**, 697–711.

[13] Duncan, O.D., Haller, A.O. and Portes, A. (1968). Peer influences on aspira- tions: A reinterpretation. *American Journal of Sociology* **74**, 119–137.

[14] Ferron, J.M. and Hess, M.R. (2007). Estimation in SEM: A concrete example.
[15] Journal of Educational and Behavioral Statistics **32**, 110–120.

[16] Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting* **21**, 331–340.

[17] Goddard, J. and Asimakopoulos, I. (2004). Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting* **23**, 51–66.
[18] Greene, W.H. (2011). *Econometric analysis*, 7.ed., New York: Macmillan.

[19] Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12.

[20] Hooper, D., Coughlan, J. and Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *The Electronic Journal of Business Research Methods* **6**, 53–60.

[21] Kline, R.B. (2011). Principles and practice of structural equation modeling, New York: Guilford Press.

[22] Lee, C.K. and Taylor, T. (2005). Critical reflections on the economic impact assessment of a mega-event: the case of 2002 FIFA World Cup. *Tourism Management* **26**, 595–603.

[23] Lee, S.Y. and Tang, N.S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* **71**, 541– 564.

[24] Lee, S.Y. and Zhu, H.T. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology* **53**, 209-232.

[25] Louzada, F., Suzuki, A.K. and Salasar, L.E.B. (2014). Predicting match outcomes in the English Premier League: which will be the final rank?. *Journal of Data Science* **12**, 235–254.

[26] Meckel, Y., Harel, U., Michaely, Y. and Eliakim, A. (2014). Effects of a very short-term preseason training procedure on the fitness of soccer players. *The Journal of Sports Medicine and Physical Fitness* **54**, 432–440.

[27] Moura, F.A., Martins, L.E.B. and Cunha, S.A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences* **32**, 1881–1887.

[28] Nevill, A.M., Newell, S.M. and Gale, S. (1996). Factors associated with home advantage in English and Scottish soccer matches. *Journal of Sports Sci- ences* **14**, 181–186.

[29] Poulter, D.R. (2009). Home advantage and player nationality in international club football. *Journal of Sports Sciences* **27**, 797–805.

[30] R Core Team. (2015). *R: A language and enviroment for statistical comput- ing*, R Foundation for Statistical Computing, Vienna, Austria. Software available at http://www.R-project.org/.

[31] Reilly, T. and Williams, A.M. (2003). Introduction to science and soccer. In *Science and Soccer*, T. Reilly and A.M. Williams, eds., 2nd ed., London and New York: Routledge, pp. 1–6.

[32] Rosa, G.J.M., Valente, B.D., de los Campos, G., Wu, X.L., Gianola, D. and Silva, M.A. (2011). Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution* **43**, 1046–1057.

[33] Staufenbiel, K., Lobinger, B. and Strauss, B. (2015). Home advantage in soccer

[34] – A matter of expectations, goal setting and tactical decisions of coaches?.

[35] Journal of Sports Sciences **33**, 1932–1941.

[36] Steiger, J.H. (1989). EzPATH: Causal modeling, a Supplementary Module for SYSTAT and SYGRAPH, PC/MS-DOS. Illinois: Evanston.

[37] Stubbe, J.H., van Beijsterveldt, A.M.C., van der Knaap, S., Stege, J., Verhagen, E.A., van Mechelen, W. and Backx, F.J.G. (2015). Injuries in professional male soccer players in the Netherlands: a prospective cohort study. *Journal of Athletic Training* **50**, 211–216.

[38] Suzuki, A.K., Salasar, L.E.B., Leite, J.G. and Louzada-Neto, F. (2010). A Bayesian approach for predicting match outcomes: the 2006 (association) football world cup. *Journal of the Operational Research Society* **61**, 1530– 1539.

[39] Taylor, J.B., Mellalieu, S.D., James, N. and Shearer, D.A. (2008). The influence of match location, quality of opposition, and match status on technical performance in professional association football. *Journal of Sports Sciences* **26**, 885–895.

[40] Tucker, L.R. and Lewis, C. (1973). A reliability coefficient for maximum likeli- hood factor analysis. *Psychometrika* **38**, 1–10.

Received June 25, 2014; accepted September 28, 2014.

Pedro H. R. Cerqueira
Departamento de Ciências Exatas
Escola Superior de Agricultura “Luiz de Queiroz”
Universidade de São Paulo
13418-900, Piracicaba, São Paulo, Brazil

Luiz R. Nakamura
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina
88040-900, Florianópolis, Santa Catarina, Brazil

Rodrigo R. Pescim
Departamento de Estat´ıstica
Universidade Estadual de Londrina
86057-970, Londrina, Paraná, Brazil

Roseli A. Leandro
Departamento de Ciências Exatas
Escola Superior de Agricultura “Luiz de Queiroz”
Universidade de São Paulo
13418-900, Piracicaba, São Paulo, Brazil