Football analytics: a Higher-Order PLS-SEM approach to evaluate players' performance

Analisi statistica nel calcio: un approccio Higher-Order PLS-SEM per valutare la performance dei calciatori

Mattia Cefis and Maurizio Carpita

Abstract Nowdays, data science is applied in several area of our life, and also many applications in sports fields are increasing. In this context, we are focusing on football (e.g. soccer); thanks to this work we have the aim to give a new approach in the evaluation of football players' performance given from the EA Sports experts and available on Kaggle in the KES dataset. For this purpose, we adopt a Higher-Order PLS-SEM approach to the *sofifa* KPIs (e.g. Key Performance Indicators) in order to compute a composite indicator and compare it with the well-known overall index from EA Sports. The final goal is to suggest a new performance index for helping coaches and scouting staff of professional teams to take strategic decisions, in order to evaluate impartially players' performance.

Abstract *Oggi la data science è applicata in diversi contesti della nostra vita e anche in ambito sportivo le sue applicazioni sono in crescita. Nel nostro contesto ci siamo focalizzati sul calcio e con questo lavoro proponiamo un approccio innovativo all'analisi della performance dei calciatori partendo da quella già offerta dagli esperti di EA Sports e disponibile sulla piattaforma Kaggle grazie al KES dataset. A tale scopo adottiamo un approccio Higher-Ordered PLS-SEM agli indici di performance di sofifa per calcolare un nuovo indice composito, confrontandolo con quello di EA Sports. L'obiettivo finale è quello di proporre un nuovo indice di performance per aiutare allenatori e l'area scouting di una società calcistica a prendere decisioni strategiche e a valutare oggettivamente i calciatori.*

Key words: football performance indicators, PLS-SEM, composite indicators.

Mattia Cefis
University of Brescia, Department of Economics and Management, e-mail: mattia.cefis@unibs.it

Maurizio Carpita
University of Brescia, Department of Economics and Management
e-mail: maurizio.carpita@unibs.it

1 Introduction

The latest developments in sports research, especially in football, are driven from a sort of a new "data-culture" approach. Players' performance evaluation is becoming a strategic key for football coaches and for the management of a football team. We know that players' performance on the soccer field has been extensively measured and described by soccer experts: in literature, very important are the detailed classification by the experts from Electronic Arts (EA)[1]. In their opinion, players' performance can be thought as a multidimensional construct made up of 6 performance composite indicators (e.g. *defending*), each of which consists of several, more specific skills (e.g. *marking*, *standing tackle* and *sliding tackle* as elements for the *defending* dimension), which combined form an *overall* index that sums up the performance; here the main problem is that experts' opinions are not statistically supported [2, 3].

In this paper, our goal is to propose the use of the Higher Order PLS-SEM approach, starting from the data a relevant Data Science platform (e.g. Kaggle) in order to build a new composite index and to compare it with the well-known *overall* index from EA Sports experts, in order to give a significant statistics support to the experts' opinion.

2 Literature overview and data employed

In order to give an overview about literature, we can say that there are two main approaches in football analytics: an explorative method oriented on analysis and classification of the KPIs (e.g., Key Performance Indicators) with the aim to evaluate players' performance [2, 3] and another one oriented in the prediction of football match results [4]. Furthermore, in order to evaluate the single player's performance there exist different methods: for example Pappalardo [8] adopted a SVM observing match outcome, Schultze and Wellbrock [10] created a rating performance index thanks to a plus-minus metric, Carpita [1] adopted an unsupervised method to classify different area of performance. We will focalize our attention on this last issue (e.g. evaluation of single player's performance), in fact our goal is to explore players' performance variables (e.g. KPIs), in order to evaluate some different strategic skills of each one; it can be useful for understanding any key choice of coaches, as well as to guide player transfer decisions, transfer fees and contract negotiations or to improve future predictive modelling.

In the European framework, the Kaggle European Soccer (KES) database is the biggest open one devoted to the soccer leagues of European countries: it contains data about 10000 players and 21000 matches of the championship leagues of 10 countries and 7 seasons from 2009/2010 to 2015/2016. It is composed mainly by two big tables:

[1] Link to the website: https://www.easports.com/

- The Match table contains the date, the positions (X and Y coordinates) on the pitch for the 22 players of the two teams and the final result of each match.
- The Player Attributes table contains other 29 variables (e.g. KPIs), with periodic player's performance on a 0–100 scale with respect to different abilities.

For our work we are interest just in the Player table and in particular we will take into account just midfielder's players from Italian Serie A 2015/2016, with stats relying the beginning of the season, in order to have a toy dataset of 106 players and 29 KPIs for each one. As said in the introduction, for what concerns attributes' description, experts of Electronic Arts (EA/*sofifa*) Sports are considered the main authority: players' performance is defined as a multidimensional entity made up of 6 latent traits (e.g. *attacking*, *skill*, *movement*, *power*, *mentality*, *defending*), but they are not statistically supported [2, 3]. Our goal is to apply to these KPIs a Higher-Order PLS-SEM model, in order to create a new synthetic composite indicator and compare it with the *overall* index of EA Sports experts.

2.1 The proposed Higher-Order PLS-SEM approach

PLS-SEM [11], also called PLS Path-Model, is a very interesting tool that offers us a valid alternative to the well-known covariance-based model [6]. Its goal is to measure causality relation between concepts (e.g. latent variables, the 6 *sofifa* latent traits in our case), starting from some manifest variables (e.g. MVs, in our case the *sofifa* KPIs), thanks to an explorative approach: the explained variance of the endogenous latent variables (e.g. LVs, variables that we see as a sort of outcome, the performance in our case) is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression [7]. Another essential point of PLS-SEM is that does not require any preliminary assumptions for the data, so it's called a soft-modelling technique. PLS-SEM estimates simultaneously two model:

- Measurement (or outer) model ⇒ links MVs (e.g. KPIs in our case) to their LVs (e.g. the 6 *sofifa* dimensions). Each block of MVs $\mathbf{X}_g$, $g = 1,...,G$ (with $G = 6$) must contain at least one MV and this relation can be treated in two ways: reflective (where the MVs are the effects of their own LV) and formative (where the MVs are the causes of their own LV). In our work we will assume a formative structure for the outer model where each LV ξ_g is considered to be formed by its KPIs following a multiple regression:

$$\xi_g = \mathbf{X}_g \mathbf{w}_g + \delta_g \tag{1}$$

and

$$E[\delta_g | \mathbf{X}_g] = \mathbf{0} \tag{2}$$

where $\mathbf{w}_g$ is the vector of the outer regression weights and δ_g is the vector of error terms. So, the vector of the outer weights for the g-th LV is estimated by least squares:

$$\mathbf{w}_g = (\mathbf{X}_g^T \mathbf{X}_g)^{-1} \mathbf{X}_g^T \xi_g \tag{3}$$

- Structural (or inner) model ⇒ thanks to this model LVs are divided into two groups: exogenous and endogenous. The first one does not have any predecessor in the path diagram, the rest are endogenous. For the j-th endogenous variable in the model, the linear equation of its own structural model is:

$$\xi_j = \beta_0 + \sum_{r=1}^{R} \beta_{rj} \xi_r \ + \zeta_j \tag{4}$$

where R is the number of exogenous LVs that affect the endogenous one and β_{rj} is so called path coefficient, a sort of linkage between the r-th exogenous LV and the j-th endogenous LV and ζ_j is the error term.

Moreover, for our work we will assume a PLS-SEM with Higher-Order Constructs, also known as Hierarchical Models [9]. In this framework we can include LVs that represent an "higher-order" of abstraction. In fact, for our purpose, we will assume players' performance as extra-latent construct of higher (second) order. Since this LV is virtual, and so without any apparent MVs, literature suggested us an interesting technique in order to modelling this framework: a two-step or patch approach [9]. In the first step of this approach, we can compute thanks to PCA (e.g. Principal Component Analysis) the scores of the lower-order LVs (e.g. the first principal component -I PC- of each one), while in the second one we can apply the classical PLS-SEM using the computed scores as MVs for the endogenous (e.g. the performance) LV. In our work, we will build two different frameworks, following the experts' suggestion[2] , in order to replicate the EA Sports *overall*:

- In the first framework, with the classical *sofifa* LVs classification (6 groups of LVs), we assume a conceptual structure behind the performance [9] with the presence of 3 endogenous LVs: *attacking*, *defending*, and the player's performance (e.g. PLS Path in Fig. 1). Note that for the performance (the only II order construct), we used the I PCs of *movement*, *defending* and *attacking* as MVs.
- In the second framework we take in consideration the EA FIFA cards ability classification (a little bit different classification of the same 29 MVs into others 6 LVs); here we assume just one endogenous Higher-Order LV (e.g. performance) influenced directly from the others 5 exogenous (Fig. 1).

For the work we used the R package *plspm* [9] and bootstrap for the validation of the models. In the next section we will share our results and a brief discussion.

[2] For details see the website: https://www.fifauteam.com/fifa-19-attributes-guide/

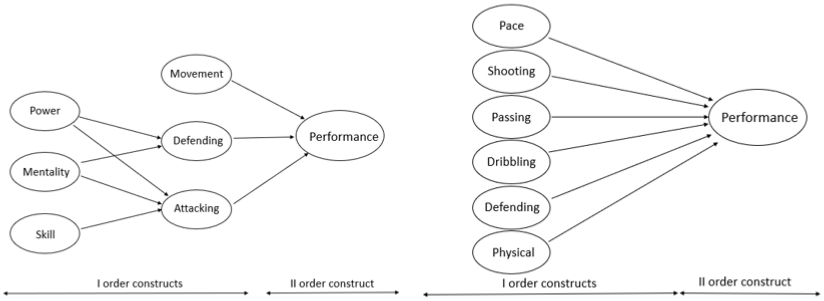

Fig. 1 PLS Path *sofifa* model vs FIFA cards model

3 Results and discussion

Preliminary results are showed in Fig. 2, where we can see an example of loadings (in a formative way) for the *defending* LV in both *sofifa* and FIFA cards model. We can see immediately the differences in these two classifications (3 KPIs for the first and 5 for the second): loadings are not exactly the same but are very high in both the cases.

In Table 1 instead we can see a comparison regards some assessments index between our two models: the unidimensionality holds in both frameworks, while the goodness of fit index (e.g. GoF) is good (e.g. > 0.7, [9]) and reveals that the second framework is a bit better than the first one. Then we computed the *rho* index (e.g. the correlation) between our Higher-Ordered PLS-SEM performance index and the true *overall* index computed from EA experts. It shows us a very high concordance (*rho* > 0.9) between our index and the EA index, in both frameworks.

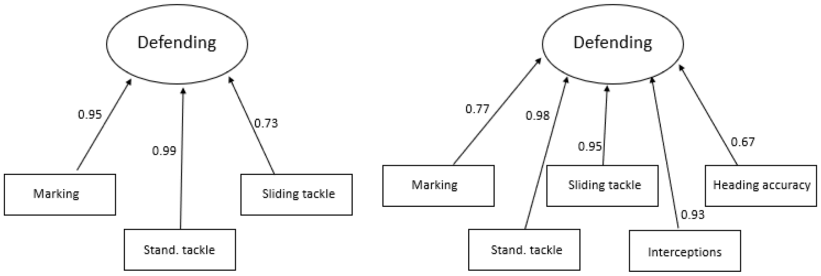

Fig. 2 Loadings comparison for *defending* LV between *sofifa* vs FIFA cards model

In Table 2 we can see the output of the bootstrap validation (with 200 samples) and how both models have a significative R^2 index for their own endogenous LVs. Interesting to note how in the second model all MVs and LVs are significative for their respectively outer and inner model.

Table 1 The two models goodness index comparison

Model	Unidim. LVs	GoF	Corr. with the EA *overall* index
sofifa	OK	0.71	0.94
FIFA cards	OK	0.82	0.93

Table 2 The two models validation comparison

Model	Non-sign. MVs	Non-sign. LVs	CI 95% for R^2 of the endogenous LVs
sofifa	1	1	$A.:[0.89;0.95]$ $D.:[0.67;0.83]$ $P.:[0.95;0.98]$
FIFA cards	0	0	$P.:[0.98;0.99]$

In summary, we have seen how both models are good and so we reapplied them across data of others European leagues and players' roles, discovering some little differences between the path coeffiecients: because of this, for future research it could be interesting, as in-depth analysis, to focus on the problem of observed and unobserved heterogeneity for players' performance (e.g. roles, leagues, teams...), maybe thanks the REBUS-PLS algorithm [5].

References

1. Carpita, M., Ciavolino, E., Pasca, P.: Players' role-based performance composite indicators of soccer teams: A statistical perspective. Social Indicators Research (2020): 1-16.
2. Carpita, M., Golia, S.: Discovering associations between players' performance indicators and matches' results in the European Soccer Leagues. Journal of Applied Statistics (2020): 1-16.
3. Carpita, M., Ciavolino, E., Pasca., P.: Exploring and modelling team performances of the Kaggle European Soccer database. Statistical Modelling 19.1 (2019): 74-101.
4. Carpita, M., et al.: Discovering the drivers of football match outcomes with data mining. Quality Technology & Quantitative Management 12.4 (2015): 561-577.
5. Esposito Vinzi, V., et al.: REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. Applied Stochastic Models in Business and Industry 24.5 (2008): 439-458.
6. Jöreskog, K.G.: Structural analysis of covariance and correlation matrices. Psychometrika 43.4 (1978): 443-477.
7. Monecke, A., Leisch, F.: semPLS: structural equation modeling using partial least squares. Journal of Statistical Software, 48 (3),(2012) 1-32.
8. Pappalardo, L., et al.: PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. ACM Transactions on Intelligent Systems and Technology (TIST) 10.5 (2019): 1-27.
9. Sanchez, G.:. PLS path modeling with R. Berkeley: Trowchez Editions 383 (2013)
10. Schultze, S.R., Wellbrock, C.M.: A weighted plus/minus metric for individual soccer player performance. Journal of Sports Analytics 4.2 (2018): 121-131.
11. Wold, H.: Encyclopedia of statistical sciences. Partial least squares. Wiley, New York, (1985): 581-591.