UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DOTTORATO DI RICERCA IN ANALYTICS FOR ECONOMICS AND MANAGEMENT

SECS-S/01 STATISTICA

CICLO

XXXV

STATISTICAL METHODS AND TOOLS FOR FOOTBALL ANALYTICS

NOME DEL DOTTORANDO/A
Dr. Mattia Cefis

NOME DEL RELATORE
Prof. Maurizio Carpita

NOME DEL SECONDO RELATORE
Prof. Eugenio Brentari

Declaration of Authorship

I, Dr. Mattia CEFIS, declare that this thesis titled, "Statistical Methods and Tools for Football Analytics" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

"Possiamo vivere nel mondo una vita meravigliosa se sappiamo lavorare e amare, lavorare per coloro che amiamo e amare ciò per cui lavoriamo."

Lev Tolstoj

UNIVERSITY OF BRESCIA

Abstract

Ph.D. course in Analytics for Economics and Management
Department of Economics and Management

Doctor of Philosophy

Statistical Methods and Tools for Football Analytics

by Dr. Mattia CEFIS

Machine learning and digitization tools are exponentially increasing in these last years and their applications are reflected in different areas of our life: in particular, this thesis aims to focus on football (i.e. soccer for Americans), the most practised sport in the world. Due to needing of professional teams, analytics tools in football are becoming a crucial point, in order to help technical staff, scouting and clubs management in policy evaluation and to optimize strategic decisions; for this reason, different statistical applications have been developed, one for each chapter, corresponding to published or submitted scientific articles. In the first part are presented the main activities I attended during my PhD, then the first chapter is dedicated to literature review, by an original bibliometric analysis relying football analytics development in the decade 2010-2020. The following chapter is designated for in-depth the Partial Least Squares Structural Equation Modeling (PLS-SEM) framework, in order to study and create some original composite indicators for players performance using data provided by Electronic Arts (EA) experts and available on the Kaggle data science platform; in particular, a Third-Order PLS-PM approach was adopted on the *sofifa* Key Performance Indices, in order to compute a composite indicator differentiated by role. In the next chapter the PLS-SEM model has been refined and validated, applying both Confirmatory Tetrad Analysis (CTA) and Confirmatory Composite Analysis (CCA), using EA *sofifa* data relying the most recent football season (2021/2022); the final results underline how some sub-areas of performance have different significance weights depending on the player's role; as concurrent and predictive analysis, the new Player Indicator (PI) *overall* was compared with a benchmark (the EA *overall*) and with some performance quality proxies, such as players' market value and wage, showing interesting and consistent relations. At this point, these original composite indicators have been introduced as regressors in the last chapter for improving in terms of prediction performance the expected goal (xG) model; it is one emerging tool in the field of football analytics, that aims to predict goal and measure the quality of each shot, by applying a supervised machine learning approach (logit model) on different scenarios for sample balanced techniques. In particular, some performance composite indicators obtained by the PLS-SEM and some original tracking variables are significant for the classification model, contributing to increase the goal prediction probability, compared with a benchmark.

Italian abstract - *Gli strumenti di digitalizzazione e di machine learning hanno avuto una crescita esponenziale nel corso degli ultimi anni e tutto ciò ha riguardato di riflesso i più svariati settori della nostra vita: in particolar modo, questa tesi ha l'obiettivo di focalizzarsi sulla sport analytics, in particolare sul calcio, lo sport più praticato al mondo. A causa della crescente necessità dei club professionistici, gli strumenti analitici nel calcio stanno diventando uno snodo cruciale per aiutare gli staff tecnici, le aree scouting e i management nell'ottimizzare e nel prendere decisioni; per questa ragione, in questa tesi sono state sviluppate diverse applicazioni statistiche, una per ogni capitolo, ognuna corrispondente ad un articolo scientifico pubblicato o in revisione da parte di una rivista scientifica. Nell'introduzione della tesi sono elencate le principali attività svolte durante il periodo di dottorato, seguite dal primo capitolo dedicato alla revisione della letteratura, effettuato in forma analitica grazie ad un originale analisi bibliometrica sugli ultimi 10 anni di produzione scientifica.*

Il secondo capitolo è dedicato ad un approfondimento metodologico sul Partial Least Squares Structural Equation Modeling (PLS-SEM), metodologia statistica utilizzata per la creazione di indicatori compositi volti ad analizzare la performance dei giocatori, tramite l'utilizzo di dati forniti dagli esperti di Electronic Arts (EA) e disponibili sulla piattaforma di data science Kaggle; nella seconda parte del capitolo è presente l'applicazione sviluppata, in particolare un modello gerarchico del terzo ordine utilizzando i Key Performance Indices di sofifa per calcolare un indicatore composito differenziato per ogni ruolo.

Nel terzo capitolo il modello sviluppato nel capitolo precedente è stato rifinito e validato per ogni ruolo, applicando una Confirmatory Tetrad Analysis (CTA) e una Confirmatory Composite Analysis (CCA), utilizzando i dati relativi ai più recenti campionati (stagione 2021/2022); i risultati ottenuti sottolineano come le diverse aree e sottoaree di performance hanno diversi pesi e valori a seconda del ruolo del giocatore. Infine, con lo scopo di valutare la validità predittiva del modello, il nuovo indicatore composito (PI) overall *è stato confrontato con un benchmark (EA* overall*) e con delle variabili proxy come il valore di mercato e l'ingaggio dei giocatori, ottenendo dei risultati interessanti e significativi.*

A questo punto, nell'ultimo capitolo gli indicatori compositi sviluppati in precedenza sono stati introdotti come regressori nel modello di expected goal (xG), con lo scopo di migliorarne l'accuratezza predittiva. Il modello xG è infatti uno dei modelli emergenti nel mondo della football analytics e ha lo scopo di prevedere i goal e misurarne la qualità. Per fare questo è stato applicato un modello logistico classico ed un modello logistico aggiustato su diversi scenari per campioni bilanciati. Nella fattispece, alcuni indicatori compositi e altri nuovi regressori (variabili di tracking) sono risultati significativi per il modello di classificazione, contribuendo a migliorare l'accuratezza nella predizione dei goal, confrontandolo con un benchmark.

Acknowledgements

Il tempo vola, e senza nemmeno accorgermene sono arrivato alla fine di questo bellissimo percorso di dottorato, che mi ha permesso di crescere, sotto tutti gli aspetti, di conoscere ed imparare nuove metodologie e tecniche statistiche, ma soprattutto di conoscere tante bellissime persone, che mi hanno dato tanto: ci tengo a ringraziare tantissimo per la pazienza ed il fondamentale supporto durante tutta quest'esperienza il mio supervisor, Prof. Maurizio Carpita ed il mio tutor, Prof. Eugenio Brentari. Inoltre voglio ringraziare tutto il fantastico team degli statistici dell'Università di Brescia, dal Dr. Manlio Migliorati alle prof.sse Marica Manisera, Paola Zuccolotto e la Dr.ssa Silvia Golia per i loro preziosi consigli e la loro grande disponibilità. Ci tengo a ringraziare tutti i professori e colleghi di dottorato conosciuti in questi 3 anni. Inoltre ringrazio il Dr. Josè Maria Oliva Lozano per avere accettato la mia proposta di visiting e per la collaborazione che è nata grazie al periodo trascorso presso l'Università di Almerìa (ES). Ringrazio il BDSport Group (https://bodai.unibs.it/bdsports) dell'Università di Brescia per il prezioso supporto grazie al quale è stato possibile finanziare la mia trasferta presso l'Università di Vienna e per avermi permesso di conoscere tante figure legate al mondo dello sport e della statistica, tra cui Math&Sport, azienda che ringrazio e con cui ho avuto la fortuna di collaborare durante quest'esperienza di dottorato.

Ringrazio tutta la mia famiglia ed in particolare i miei nonni, che mi sono sempre stati vicino e mi hanno supportato in ogni momento, non facendomi mai mancare niente, dandomi consigli preziosi, soprattutto nei momenti difficili. Ringrazio di cuore gli amici e tutte le persone che mi sono sempre state accanto e che mi hanno sempre sostenuto, sono stati fondamentali. Voglio dedicare questa tesi a tutte le persone che mi vogliono bene, per me sono importantissime, perchè ho imparato che la serenità e data da ciò che ci circonda, e per esperienza posso dire che è la cosa più importante nella vita.

Contents

List of Figures

List of Tables

Introduction

Nowadays, football analytics is a growing theme for researchers and teams both, encouraging different areas of study: from players' market estimation to performance evaluation, from injures prevention to data-tracking analysis, all thanks a data-driven approach. Like we will see in the following chapters, data and statistics could be very precious for helping policy makers to take some strategic decisions. In this thesis it has shown an initial setup thanks a literature review in Chapter 1, to give an idea about the state of the art about the last decade, then the aim is to build some composite performance indicators thanks an innovative statistical approach: the Partial Least Squares Structural Equation Modeling (PLS-SEM, respectively in Chapter 2 and Chapter 3), in order to evaluate and split players' performance in different sub-areas, for helping technical and scouting staff of a soccer team. These composite indicators will be used for improving in terms of prediction accuracy the well-known Understat (`www.understat.com`) expected-goal model, integrating it with some players' tracking data too (Chapter 4). Finally, the conclusion of the thesis is provided.

Thanks to this introduction the goal is to give a clearer idea about the logical path of this thesis, also underlining my growing way, with my contributions and related conferences or courses I attended during my PhD experience.

My contributions and attended conferences

We give below an overview of the contributions of this PhD thesis in terms of published or accepted peer-reviewed papers, in international conferences or scientific journals. The list is in chronological order, and for each paper there is a brief description.

- Cefis, M. [31], "Football Analytics: performance analysis differentiate by role", Book of abstracts of DSSR conference, 2020.

  This was a speech contribution at the DSSR (Data Science Social Research) conference of December 2020, in the session named "Sport analytics on the pitch". In this work I focused on data visualization about strategic key performance indicators for football players, in particular underling the different statistical distribution among roles' key performance indices.

- Football Data Analyst course (`www.wylab.net`): I attended this course from march to may 2021, organised from Wylab. This course had a data-driven approach to football, with some interesting topics useful for the research path: data visualization, the expected goal model, the playrank algorithm [102], key performance indices about teams and players. I made as output of this course a project work with others classmates, thanks to players' event data and using some machine learning techniques by Python, with title: "The analytical scouting: data and insight to detect the target player. The case-study of Bologna F.C."

- Cefis, M. and Carpita, M. [30]: "Football analytics: a Higher-Order PLS-SEM approach to evaluate players' performance", Book of short papers SIS 2021.

  This was a contribution talk at the 50^{th} SIS (Italian Statistical Society) conference on June 2021, in the session entitled "Statistical Methods and Models for the analysis of sports data". In this work we focused to build a model to evaluate players' performance in an innovative way, thanks a PLS-SEM model . I contributed as first author in the discussion of the idea with my co-author, in implementing on *R* the work and in writing about it.

- Cefis, M., "Using Higher-Order PLS-PM model for performance analysis in football", Link to the presentation: `youtu.be/sZdryIGEDEs`.

  This was a contribution talk at the ISI (63^{rd} World Statistics Congress) conference of July 2021, in the session named "Advances in Sports Statistics". As only author, I focused to improve and develop a PLS-PM model of third-order after reviewing experts and football scientist advice on the theoretical model, with the aim to split and to create a composite indicator to evaluate players' performance.

- Cefis, M. and Carpita, M. [32]: "PLS-SEM with CCA for football goalkeeper's performance indicators", Book of short papers IES 2022.

  This was a contribution talk at the IES (Innovation and Society 5.0) conference on January 2022 (Capua -CE- Italy), in the session named "SEM with PLS: Theory and Applications". In this work we focused on building a model to evaluate goalkeepers' performance in an innovative way, by a Second Order formative-formative PLS-SEM model and to provide a confirmatory composite analysis (CCA) on that one. I have contributed in the discussion of the idea as first author, in implementing on R the work and in writing about it.

- Visiting period at University of Almerìa (Spain, April-May 2022): during this experience I collaborated with a researcher and sport scientist (Dr. Josè Maria Oliva Lozano) with a project for LaLiga (the main Spanish football league), relating the soccer performance analysis. I analysed datasets concerning players athletics features, with the objective to create a customized physical performance indicator, by a PLS-SEM approach.

- Cefis, M. [25], "Football analytics: a bibliometric study about the last decade contributions", Electronic Journal of Applied Statistical Analysis. Vol 15, No 1 (2022).

  As only author, I submitted this full paper to the Electronic Journal of Applied Statistical Analysis (EJASA, `siba-ese.unisalento.it`), with the aim to emphasize the growing of statistical and machine learning tools in football analytics about last decade, thanks an original bibliometric analysis by the well known SCOPUS database. The paper was published in May 2022. It is the practical application proposed in Chapter 1.

- Cefis, M. and Carpita, M.: "The Higher-Order PLS-SEM Confirmatory Approach for Composite Indicators applied on Football", AUEB online workshop.

  This was a contribution talk at the 6^{th} AUEB SAW 2022 online workshop on May 2022, relying sport analytics. In this event I presented the complete work focused on the creation of a football players' composite performance indicator, based on CTA and CCA as confirmatory approach.

- Cefis, M. and Brentari, E., "Formative vs Reflective constructs: a CTA-PLS approach on a goalkeepers' performance model", SIS 2022 conference.

  This was a contribution talk in an invited session related sports analytics at the 51^{st} Statistical Italian Society (SIS) 2022 conference on June 2022 in Caserta (Italy). By this conference I have presented, as main author, the complete work focused on the creation of a football goalkeepers' composite performance indicator, based on CTA and on CCA as confirmatory approach.

- Cefis, M. [26], "Observed heterogeneity in players' football performance analysis using PLS-PM", Journal of Applied Statistics (2022).

  As only author, I submitted this full paper to the Journal of Applied Statistics, with the goal to propose an exploratory approach using the PLS-SEM in the creation a composite performance indicator in for football analytics. In particular, focus was pointed on the heterogeneity observed among leagues and roles. The paper was published in July 2022. It is the practical application proposed in Chapter 2.

- Cefis, M. and Carpita, M. [33], with title "The Higher-Order PLS-SEM Confirmatory Approach for Composite Indicators of Football Performance Quality".

  As first author, we submitted this full article at the Computational Statistics Journal, as extension of the previous article published in the Journal of Applied Statistics. Here the main purpose was to in depth the early model introducing a Confirmatory Tetrad Analysis (CTA) in order to evaluate statistically the nature of each first order latent construct and to validate each model-role by a Confirmatory Composite Analysis (CCA). This is the application that will be presented in Chapter 3.

- Cefis, M. and Carpita, M., "An innovative xG Model for football analytics", IACSS & ISPAS 2022 conference.

  This was a contribution talk at the 13^{th} World Congress of Performance Analysis of Sport 2022 & 13^{th} International Symposium on Computer Science in Sport on September 2022 in Wien (Austria), in the Performance Analysis session. By this conference I have presented, as main author, the complete work focused on the creation of an innovative expected goal model based on some composite performance indicators and some tracking and event data.

- Cefis, M. and Carpita, M., "Higher-Order PLS-SEM for football analytics", ECDA 2022 online conference.

  This was a contribution talk in an online invited session entitled "Advances in PLS Structural Equation Modelling" at the European Conference on Data Analysis, on September 2022. By this conference I have presented, as main author, the complete work done in the paper submitted at the Computational Statistics Journal.

- Cefis, M. and Carpita, M. [29], entitled "A new xG Model for football analytics".

  As first author, we submitted in October 2022 this full article to the Journal of the Operational Research Society, and it's now under review. Here the main purpose was to introduced in the well known expected goal (xG) model as new regressors the composite indicators developed in Chapter 3 and some tracking variables, in order to improve the goal detection capability of the model. This application will be presented in Chapter 4.

Chapter 1

State of the art

To make this thesis more self-contained, this chapter presents an overview of the literature about football analytics, thanks to an innovative approach: the Bibliometrix *R* package [2]. This interesting tool let us to optimise the stages of data-analysis and data-visualization both, about literature, managing data directly from the famous bibliographic database SCOPUS[1]. After a brief introduction (Sec. 1.1), we will present the data management phase (Sec. 1.2), followed from a results analysis and a discussion about the more significant papers, respectively in Sec. 1.3 and Sec. 1.4. Finally, a conclusion about this chapter is given in Sec. 1.5.

More specific concepts, are described in the corresponding chapters throughout the thesis.

1.1 Introduction

Nowadays, we can consider football clubs as real firms, while until some years ago we were in the so-called patronage era. Because of this, now the aim of football clubs is to optimize their own financial statements, in order to avoid any penalty from the maximum football authority (for example, UEFA[2] for Europe clubs). As summary, the main earnings for a football club [19] derive from:

- Pay TV
- Stadium tickets and official merchandising
- Sponsors
- Players' transfer market

With regard to this, it is logical to affirm that a successful team, with a good game-system and which often plays international competitions (for example, UEFA Champions League) causes a virtuous circle ([19], see also Fig. 1.1) : new sponsors, fans, UEFA bonus and increasing in the players' value.

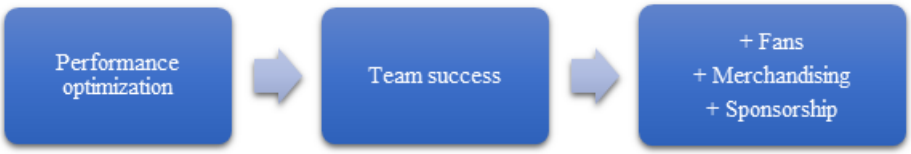


FIGURE 1.1: Virtuous circle in football world

[1] www.scopus.com
[2] www.uefa.com

So, we can say that all this virtuous circle is strongly influenced from sports results; for this reason, for a football club is extremely important to optimize them. In this last decade, for many sports and also for football is developing a digital revolution, where the crucial theme is: how to optimize players and team performance, in order to reach positive results on the pitch. Many teams and researchers are trying to answer this question.

1.1.1 Previous works and guideline

Until now, bibliometric reviews on sports have been focused on different topics but not directly on football analytics: for example there is a bibliometric analysis on sports science [129], some others focused on technology of the sport [5], until the more recently focalized on the role of social media in sports [93]. So, following the introduction made in the previous paragraph and in order to provide a guide for football analysts and data scientists, our goal is to propose an original overview of the literature about football analytics, thanks to an innovative approach: the Bibliometrix *R* package [2]. This interesting tool let us to automate the stages of data-analysis and data-visualization both, about literature, managing data directly from the famous bibliographic database SCOPUS. As previous step we tried to take in consideration also another famous database (i.e. Web of Science), but merging different databases is one of the most challenging topics of bibliometrics literature: in fact SCOPUS and Web of Science have very different records and many metadata, such as authors' names, affiliations, and references, that are stored with not compatible formats. Furthermore, we noticed that our query (Sec. 1.2) applied on SCOPUS produced 215 documents as output, while from Web of Science only 73, of which 67 already found in the SCOPUS database; for these reasons we decided to adopted just documents from SCOPUS.

1.2 Data extraction and preparation

As disclosed at the beginning of this chapter, data were extracted from SCOPUS. The goal was to collect all documents about football or soccer analytics (1.1), searching these words in the title, abstract and keywords of each article: before the year 2010 we observed that scientific production produced maximum one article per year, and as consequence the decision to focus just on the last ten years (decade 2010-2020, with more significant production); furthermore, we kept in consideration only documents in English language. This query ran on July, the 26^{th} of 2021.

$$(Football\ or\ Soccer)\ and\ analytics \tag{1.1}$$

$$\begin{aligned} &not \\ &(Rugby\ or\ Cricket\ or\ Hockey \\ &or\ American\ Football \\ &or\ Australian\ Football) \end{aligned} \tag{1.2}$$

After a first review, we noticed that in the output there were included some bias articles (for example, about American or Australian football, or other sports): for this reason, in order to automate the extraction, we attached below (1.1) an extra part (1.2). By (1.2) we could exclude noisy documents from our research, then the dataset was converted (thanks a special function provided by the *Bibliometrix R* package) into a data-frame, with cases corresponding to articles and variables to field tags.

In the final dataset we obtained a total of 215 documents over the last decade. It is not a high number, but we must take in mind that soccer is one of the last sports where analytics achieved: in practise, as we will see in Sec. 1.3, football analytics revolution is on the cutting edge just from the last years. Before beginning the bibliometric analysis, we adjusted by hand some typos in the keywords and in the authors' names from the dataset, in order to avoid redundancy and misunderstanding in the results.

1.3 Results and analysis

In this paragraph we will analyze results of our bibliometric analysis focusing on different aspects: in Sec. 1.3.1 we will see an overview about the scientific production in this last decade, while in Sec. 1.3.2 we will show some statistics about the authors, in Sec. 1.3.3 we will focus on the keywords, in Sec. 1.3.4 we will propose an in-depth analysis about the most productive countries and universities; eventually, some in-depth graphs are shown in Sec. 1.3.5.

1.3.1 Overview results

As said before, this analysis was performed by the *Bibliometrix* package of *R*; for first, here are shown some general results, in order to understand how the bibliometric dataset is composed and the documents production trend over the last decade. In Tab. 1.1 we can see a preliminary classification of the documents: it's clear the prevalence of conference papers and articles.

TABLE 1.1: Documents classification

Document types	Nb. of docs
Article	78
Book chapter	2
Conference paper	113
Conference review	11
Data paper	1
Editorial	3
Review	7
Total	215

In order to give an overview of the most relevant sources, considering all documents listed in Tab. 1.1, the plot in Fig. 1.2 is presented: the most relevant sources (i.e. with more than 15 documents) are the Journal of Lecture Notes in Computer Sciences, the CEUR workshop proceedings and the International Journal of Sports Science and Coaching.

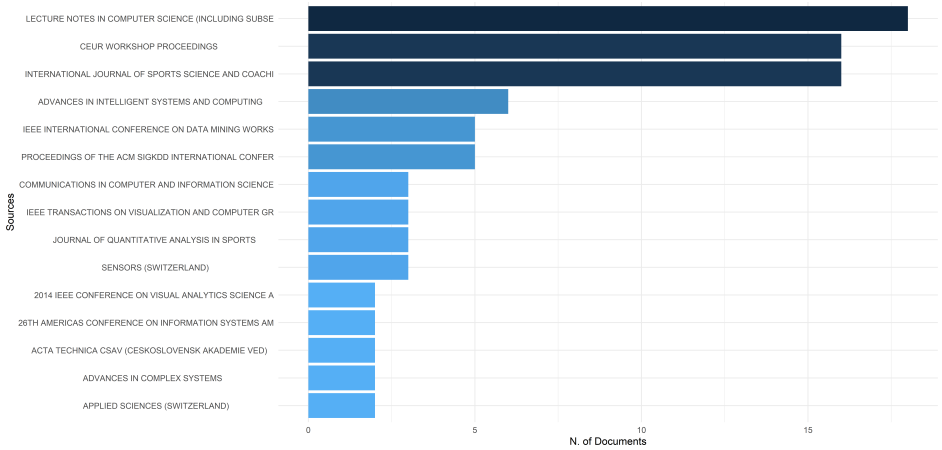


FIGURE 1.2: The most relevant sources in football or soccer analytics

In Fig. 1.3 instead we can see the time-series of documents production over the last decade: this evolution shows us a significant growing in the football analytics production, with a peak in 2019 and a stabilization in 2020.

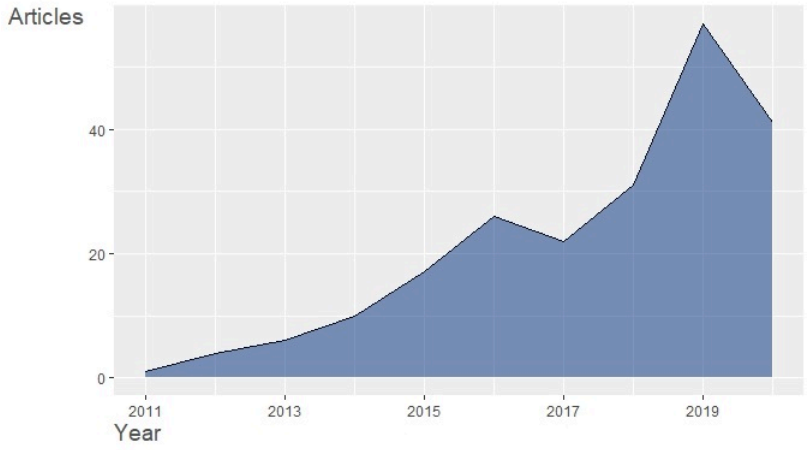


FIGURE 1.3: Annual scientific production

All this let us to underline how soccer analytics is an emerging and attractive topic in the research world.

1.3.2 Authors analysis

In this paragraph we underline the most active authors; in Tab. 1.2 is suggested their ranking by the well-known dominance factor: it is a ratio indicating the fraction of multi-authored articles in which a scholar appears as the first author. Consequently, an index near to 1 indicates very high dominance (i.e. in this table are considered authors with dominance factor greater than 0.10).

TABLE 1.2: Authors' ranking by dominance factor

Ranking	Name	Dominance factor
1	Stein M.	0.73
2	Bransen L.	0.50
3	Pappalardo L.	0.50
4	Fernandez J.	0.25
5	Lucey P.	0.20
6	Stensland H.	0.20
7	Cintia P.	0.17
8	Davis J.	0.14
9	Halvorsen P.	0.14
10	Janetzko H.	0.13
11	Van H. J.	0.11

Now, here below (Fig. 1.4) we propose an interesting plot that take in consideration not only the volume of the authors' production, but also the number of citations per year over the last decade: for this reason there is not perfect correspondence between Tab. 1.2 and Fig. 1.4 authors. In addition, take in consideration that the diameter of circles is proportional to the number of published articles, while their darkness is proportional to the total number of citations received per year.

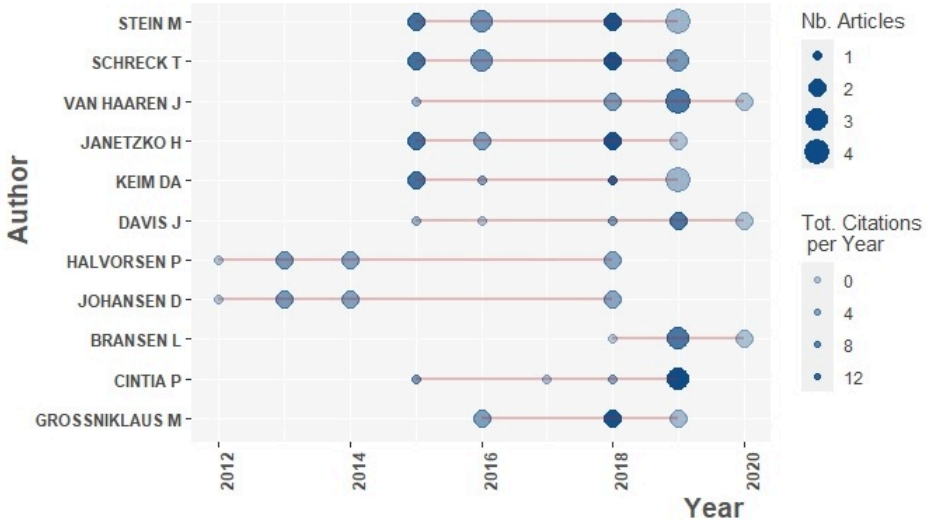


FIGURE 1.4: Top-Authors' production over the years

From Fig. 1.4 we can emphasize that activity of the most productive and cited authors is concentrated in the last five years, except for Halvorsen and Johansen. For example Cintia (University of Pisa -Italy) had an increasing of his production and obtained more than 10 citations in 2019, whereas Van Haaren (University of Leuven -Belgium-) contributed with more than 5 articles and received 15 citations in the last three years; remarkable also the contribution offered by Schreck (University of Graz -Austria-) and Stein (University of Konstanz -Germany-, with also the highest dominance factor, see Tab. 1.2) between 2015 and 2019 (more than 10 articles and 20 citations received for each one).

1.3.3 Keywords analysis

In this paragraph the aim is to investigate about research topics, show what are the most relevant keywords used from authors and their connection thanks to different plots [45]. As preliminary analysis, in Fig. 1.5 are shown the most used keywords from the authors, thanks a word cloud plot (i.e. the words size is proportional to their frequency). It is interesting to notice how, excepting the keywords used in the initial query (i.e. football, soccer and analytics, that we expected in this result), there are also ”sports” (the most used one) and typical analytics tools such as ”data mining”, ”learning systems”, ”visualization”, ”artificial intelligence” and ”machine learning”.

FIGURE 1.5: The most used keywords from authors

For the next, *Bibliometrix* allows using the *conceptualStructure* function to perform multiple correspondence analysis to draw a conceptual structure of the field and K-means clustering to identify clusters of documents that express common concepts, all summarised by a network plot (Fig. 1.6); this graphic let us to explain co-occurrence, where keywords and rectangles size are proportional to the production, while thickness of ties to the strength of co-occurrence. Different colours represent clusters, created from a K-means clustering procedure, to identify groups of documents that express common concepts [2]. In particular, co-word analysis aims to map the conceptual structure of a framework using the word co-occurrences (i.e. in this case the keywords) in a bibliographic collection.

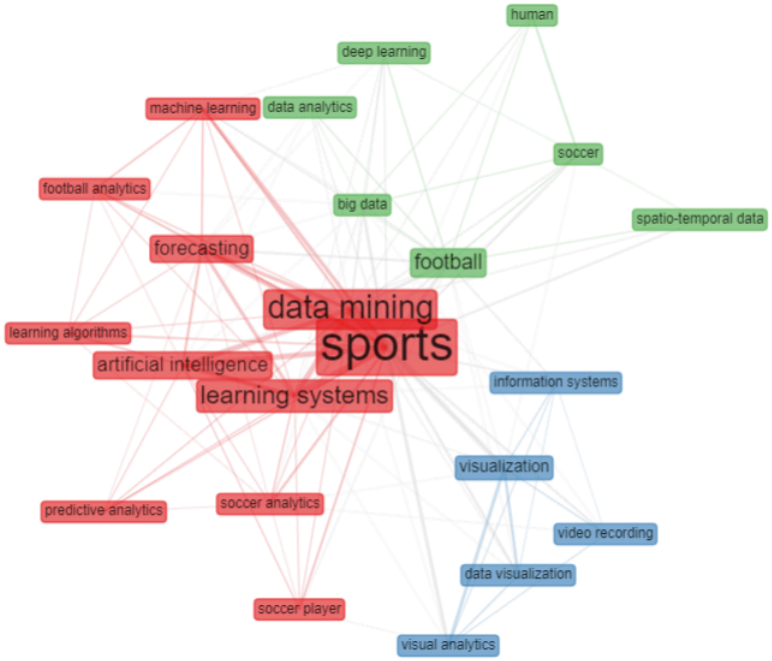


FIGURE 1.6: Keywords co-occurrence network

From Fig. 1.6 we can highlight how the red cluster is the most representative (i.e. 11 keywords), with focus on technical tools (i.e., machine learning, data mining, forecasting, artificial intelligence, player analysis and prediction) while the blue one is focalized on visualization tools and the green one on more general topics such as big data and deep learning.

Now, in order to represent co-occurrences network in a simpler view (i.e. a 2-dimensions plot), we can see the thematic map (Fig. 1.7; for this plot we must take in consideration that the words used in the initial query (i.e. football and soccer) have been excluded, in order to have a clearer interpretation. As comment, this graphic lets us to understand:

- In the top-right quadrant (high density and centrality) we can see the motor themes.
- In the bottom-right quadrant (high density and low centrality) there are the basic themes.
- In the top-left quadrant (low density and high centrality) we find niche themes.
- In the bottom-left quadrant (low density and low centrality) there are emerging or discovering themes.

Take in mind that circle size is proportional to the cluster word (i.e. in this case keywords) occurrences.

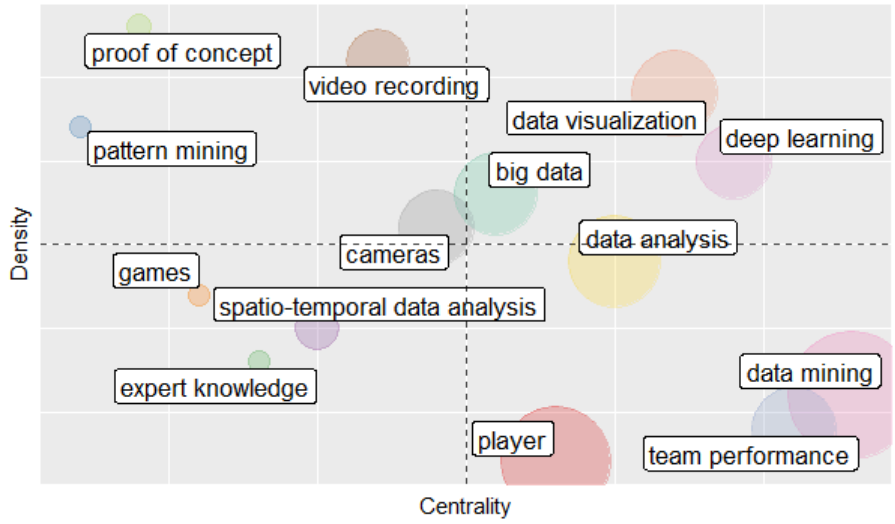


FIGURE 1.7: The thematic plot

From Fig. 1.7 we see how technical tools are the motor, they are often applied for basic themes (i.e., player, team performance and data mining), while niche themes are mainly video recording and cameras; finally, considerable emerging themes are spatio-temporal (also called as data tracking analysis), that is strictly related with video recording and cameras themes. Also expert knowledge is a crucial emerging theme, since it could be very useful in comparison with analytic results.

As insight, we analyse in Fig. 1.8 the top-five keywords evolution over the last decade. It is interesting since we can see the topics trend applied into football analytics research.

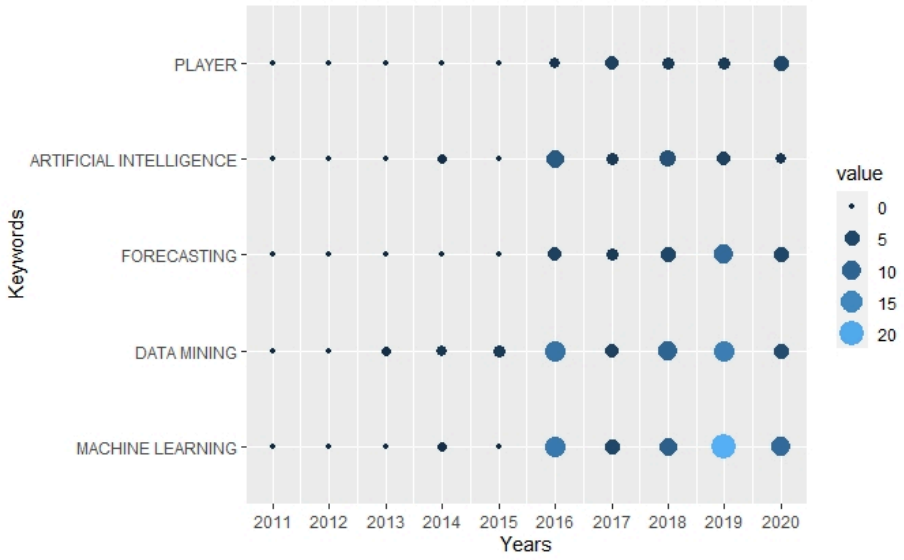


FIGURE 1.8: Keywords evolution over the ten years

It's correct to highlight that in Fig. 1.8 circle and brightness are proportional to the number of contributes. We emphasize the increasing of employment for these keywords, moreover until the year 2016, then a little decreasing and a new increasing

in the last two years: the most employed keywords are machine learning and data mining.

1.3.4 Countries analysis

Now, in this section attention is relied on countries analysis, in order to discover what are the most productive ones and the network of universities collaboration. Notice that in Fig. 1.9 Multiple Country Publications (MCP) indicates, for each country, the number of documents in which there is at least one co-author from a different country and so it measures the international collaboration intensity of a country; instead, Single Country Publications (SCP) index measures the number of documents in which author and co-authors are from the same country. We can see how Germany, USA and Italy are the most productive countries, with an interesting difference: while for Germany and USA a part of their production derive from collaboration with authors from other countries, Italy do not contribute with anyone. Austria, Belgium and China shows a higher rate of collaboration with other countries (MCP) than their own SCP.

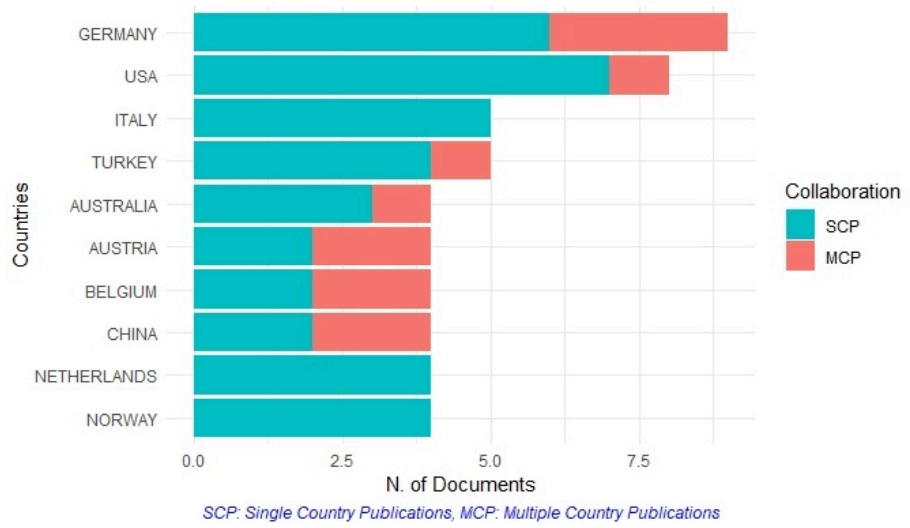


FIGURE 1.9: The most productive countries

In order to have a clearer idea than before about countries collaboration and their rate of production we can see a summary plot in Fig. 1.10, the country collaboration map. Thanks this graphic, the darkness of each country is proportional to each own production (i.e. grey states have no production), while lines thickness among countries is proportional to their collaboration rate. This plot emphasize the relevant relationship between USA and Australia (i.e. the strongest one), and some others intercontinental relations respectively among centre Europe and Brazil, Spain and Japan. Since football analytics is an emerging theme, there are many countries with zero or very poor relations, for example Canada, Argentina, Italy, north Europe, middle East and India. It could be proficient to encourage a more global cooperation for the next years.

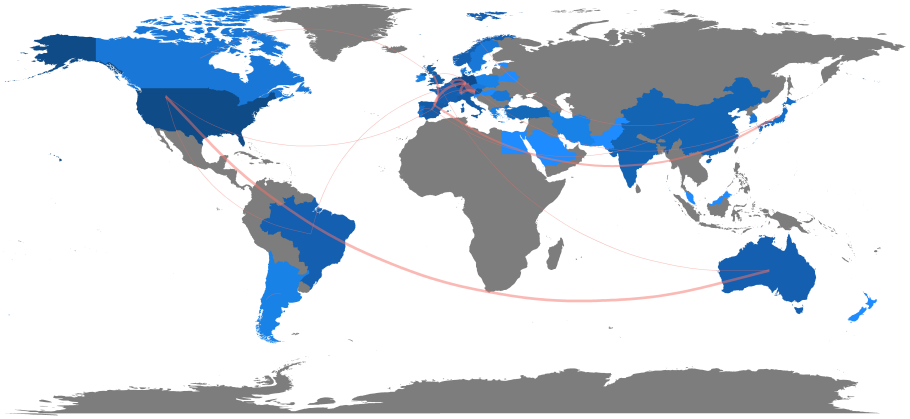


FIGURE 1.10: Country collaboration map

In the following we present the research groups collaboration network: we must keep in mind the guidelines explained in the Sec. 1.3.3. Furthermore, we show just networks with at least two research groups involved in.

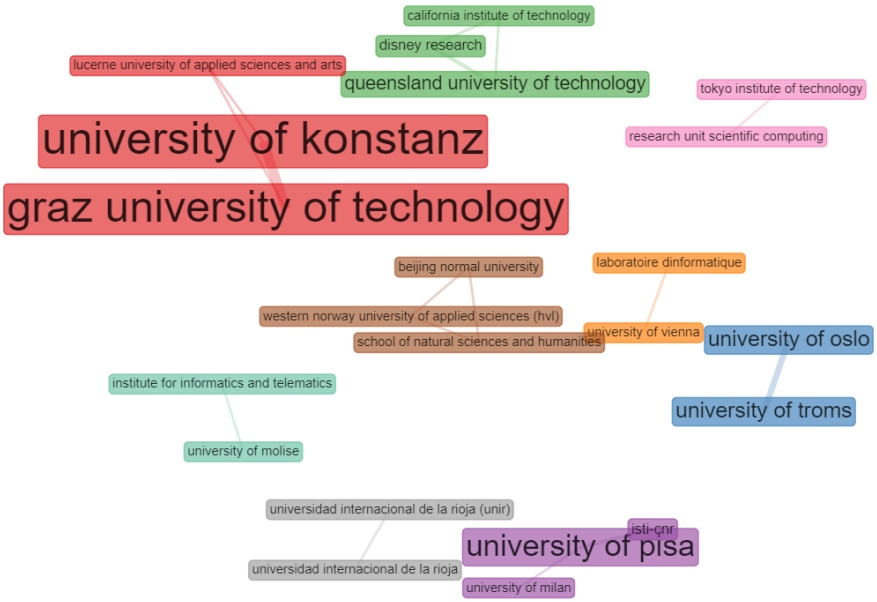


FIGURE 1.11: Research collaboration network

We can see some clear clusters, where the red, green, brown and purple are the most representative ones (i.e. clusters with more than two research groups linked):

- Red cluster is a European group: it is composed from Dutch, Austrian and Switzerland universities.
- The green is an intercontinental cluster (i.e. research groups from USA and Australia).
- Brown cluster is another intercontinental group among China and Norway.

- Purple cluster is an example of single country group: here we find only Italian research institutes.

Each one of the remaining little clusters is composed mainly from strictly continental research groups or from single group.

1.3.5 In-depth analysis

In this final paragraph, we propose an in-depth analysis thanks two interesting graphs: in Fig. 1.12 we can see a three-field plot [45, 2], where there are linked authors, keywords and sources, taking in consideration only the articles; in Fig. 1.13 instead we can see thematic evolution between first and last five years of the decade, with focus just on conferences.

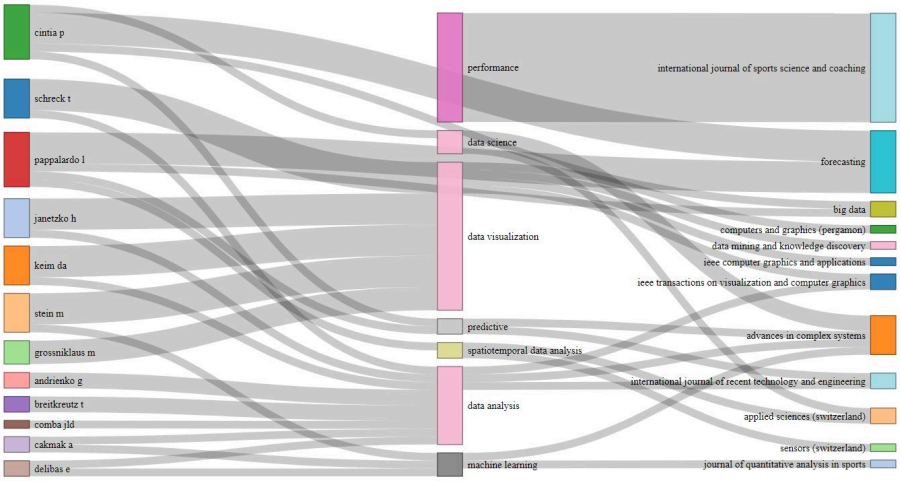


FIGURE 1.12: Three-fields articles plot: connection between authors, keywords and sources

In Fig. 1.12 we can see top authors-keywords-sources linkage: note that the height of rectangles is proportional to the number of documents produced. Notice that the best source for football analytics article is the International Journal of Sport Science and Coaching while data visualization and performance are the most relevant keywords.

In the last plot (Fig. 1.13) we can see keywords thematic evolution, where height of rectangles is proportional to the number of documents produced.

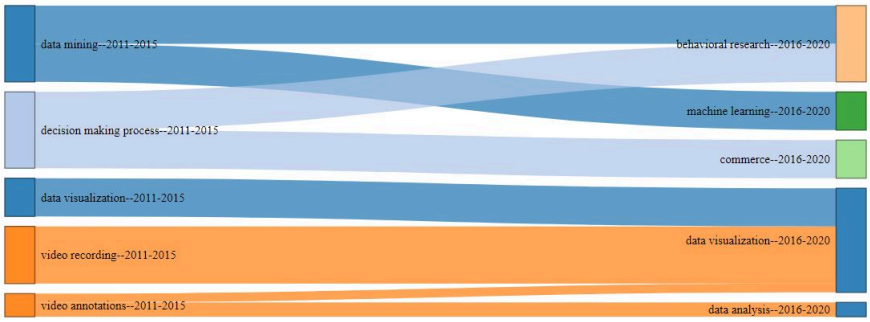


FIGURE 1.13: Thematic conference trend over the last ten years

We can explain the graphic, for example, in this way: conferences focused on data mining in the first five years, in the last half of the decade (2016-2020) moved respectively to behavioural research and machine learning topics. In addition to this, it is interesting to underline the recently growing of data visualization topic conferences (higher rectangle in the last five years than before).

1.4 Discussion

As viewed until now, we can say that football analytics is an increasing topic in sports research; in particular, the most interconnection keywords used by authors bright us to sum up three crucial theme for football analytics:

- *Technical keywords and tools*, where the recurring ones are data mining, machine learning and artificial intelligence, tools nowadays applied in many branch of our life.
- *Data visualization*, because presentation of results is fundamental in every sector, and also football does not do exception; for instance, it is crucial to show results in the simplest and incisive way to the coaches and technical staff.
- *Performance* is the core of each analysis, in fact nowadays it is crucial for a football team to be able optimizing it.

Since Soccer analytics is an emerging topic, there aren't too much collaboration between research groups yet. For example Italy has productions just in its own country; the main groups are located within continent, except some sporadic case (for example USA and Australia, China and Norway).

Here below are presented, as insights, three relevant articles, produced from authors with the highest dominance factor (see Tab. 1.2):

- Stein et al. [124]: their work, result from a collaboration of different research groups (from Germany, Portugal, Austria and Switzerland), has its own focus on movement and visual analysis on soccer. They suggest a tool that covers the automatic detection of region-based faulty movement behaviour, as well as the automatic suggestion of possible improved alternative movements. They compare their work with experts knowledge, with an interesting result: an agree index of 83%. So, we can say that their approach could effectively supports analysts and coaches investigating matches. This contribution was published on the Journal of Sports Sciences.
- Pappalardo et al. [102]: this is an example of single-country work, in fact it is produced just from Italian research groups. The aim of this work is to create a sort of synthetic index in order to evaluate objectively football players performance, thanks some event-match data and machine learning/big data techniques. The final goal is to support teams in scouting, and so to evaluate players impartially. This article was published on the ACM Transactions on Intelligent Systems and Technology.
- Bransen and Van Haaren [16]: these Dutchmen authors, thanks artificial intelligence and learning systems, propose a novel approach to measure players' on-the-ball contributions from passes during games. Their method measures the expected impact of each pass on the scoreline. This document was published on the 5th Workshop on Machine Learning and Data Mining for Sports Analytics.

1.5 Chapter conclusion

As summary, we have seen the important growing of football analytics topics over the last ten years, although it is a niche theme. We have shown how the main goal for researchers and football teams is to support policy-evaluation, thanks the more recent techniques of machine learning and artificial intelligence. Furthermore, another relevant topic is data visualization, in order to show statistic results in the simplest and efficient way to the people without an analytics background. It has been illustrated how this topic has not involved an intercontinental collaboration yet, except some sporadic cases.

We have written this chapter with the goal to guide readers in this new world of football analytics, to show the relevance it could have for teams and to emphasize the role of analytical tools. The final goal is to guide researchers and practitioners in this new frontier of football research, highlighting the importance of this data-driven revolution.

The direction is traced, we have seen what already exist, but since this is a "young" theme, there are also many emerging topics to improve and investigate. Eventually, it could be interesting to encourage an exchange between researchers and teams' experts, in order to create a bridge with the club needs: experts and statisticians collaboration could be the future for football. In the following chapters, some original applications for football analytics will be presented.

Chapter 2

The PLS-SEM approach

The latest developments in sports research, especially in football, are more and more oriented on a data-driven approach [25]. Players' performance evaluation is becoming a strategic key for football coaches and for the management of a football team; for this reason, in this chapter we will propose a very interesting tool, already known in the psychometric world, but at the same time innovative for the sport research field: the PLS-SEM (e.g. Partial least squares Structural Equation Modeling, called also PLS-PM, as Path Modeling), for evaluating and measuring players' latent performance by some composite indicators. In particular, this chapter will be structured in the following way: the first part (Sec. 2.1) is focalized on the methodological framework, whereas in Sec. 2.2 is presented an exploratory application. Eventually, a conclusion of the chapter is given in Sec. 2.3.

2.1 Methodology

In order to present an overview of this section, we can start speaking about the big family of indicators which belongs PLS-SEM: the composites. Composite indicators had an exponential growing in the last 20 years, used to evaluates and supervise issues in a wide range of topics, for example: economy, society, industry, health. As written in the Handbook of Composites ([48]), a composite is formed when individual indicators are compiled into a single index on the basis of an underlying model, in fact it is used to measure multidimensional concepts; so, its purpose is to summarize a complex phenomenon and monitoring it over the time, in order to help policy makers to take strategic decisions. Due to its easier interpretation than a battery of many separate indicators, a composite also tends to facilitate communication with citizens and media, promoting accountability. On the other hand, may be also a risk to simplify too much a topic and invite simplistic policy conclusions. Another crucial theme is the simple indicators selection and their weights computation: it could be the subject of political dispute. Furthermore, composite indicators may lead to inappropriate policies if dimensions of performance that are difficult to measure are ignored.

So, the debate is open, and like all methodological frameworks, also composites have their pros and cons, as introduced above; we think that collaboration between statisticians and experts, in addition to a clear explanation of the methodology, could take towards a general agreement. For this reason, we think that model-based composite indicators, and in particular PLS-SEM, could be a significant tool in this sense. This method had a relevant evolution in the last 30 years, with many applications in social sciences and psychometric [84, 88]. Recently, some bibliometric reviews have been proposed [43]; in particular, it is important to highlight how the application of PLS-SEM in the field of sport analytics is new and original.

In the following, we will present an introduction of the topic (Sec. 2.1.1), then we will focus on the PLS-SEM theory (Sec. 2.1.2), followed from some in-depth topics, like how to assess a PLS-SEM (Sec. 2.1.3), the higher-order approach (Sec. 2.1.4), how to manage moderating effects and the heterogeneity between observations (Sec. 2.1.5 and Sec. 2.1.7). Finally, a focus among pros and cons of PLS-SEM is given in Sec. 2.1.8.

2.1.1 Model-Based Composite Indicators

In recent years, model-based composite indicators had a significant growing in different research fields, in particular PLS-SEM has becoming crucial for social science. The forerunners of the PLS-SEM were two iterative procedures created from Herman Wold those used least squares estimation to develop solutions for single and multicomponent models and for canonical correlation [134]. Starting from this procedure, there has been different implementations during the years: Herman Wold still developed the Non-linear Iterative Partial Least Squares (NIPALS, [136]), followed from a generalized version of the PLS algorithm focused on the inclusion of latent variables in path models [135, 137, 132, 92]. Two relevant and well-known procedures evolved from Wold's works are the Principal Components Regression (PCR) and the Partial Least Squares Regression (PLS-R). As summary, the first one is focused on reducing the dimensionality of the independent features without taking into account the relationship between them and the dependent variables. Instead, PLS-R was originally designed to reduce the problem of multicollinearity in regression models: its goal is to optimize variance extracted from the independent variables (i.e. dimension reduction) and simultaneously maximize the variance explained in the dependent variables. This last technique was developed from the Wold's son, Herman, in the field of analytical chemistry.

Now, we are arrived to the more interesting result developed starting from the Wold's generalized PLS algorithm: PLS-SEM (Partial Least Squares Structural Equation Modeling, [66]), also known as PLS-PM (Path Modeling). This technique determines the parameters of a set of equations in a path model by combining principal component to assess the measurement models with path analysis to estimate relationships between latent variables [75]. Wold [137], proposed his "soft-model basic design" underlying PLS-SEM as an alternative to Joreskog [86] Covariance-Based SEM (CB-SEM). CB-SEM is suitable as confirming theory approach, and it is characterized from more restrictive assumptions in terms of data distribution and sample size, while PLS-SEM is appropriate as exploratory and theory development tool both, but we will examine it more accurately in the following paragraph.

Nevertheless both approaches were developed about the same time, CB-SEM became more famous thanks the LISREL software since the late 1970s, while the first user-friendly commercial software for PLS-SEM was SmartPLS ([108], with a graphical user interface). So, from 2005 PLS-SEM applications and related software grew exponentially, and the more relevant tool, in addition to SmartPLS, is XLSTAT-PLSPM[1], an optional tool integrated in MS Excel, with implemented inside the REBUS segmentation approach [58] too, for treating unobserved heterogeneity. For free application, in the last 15 years were developed some interesting package on *R* software, dedicated to PLS-SEM approach, for example: *plspm* [113], *semPLS* [101], *seminr* [104], till the more recent *csem* package [99]. In the following sections, we will use mainly the PLS-SEM nomenclature, since the term path modelling is preferred

[1] `www.xlstat.com/en/products/xlstat-plspm`

over structural equation modelling in the PLS community [113] even though both terms are often used interchangeably.

2.1.2 Theory under PLS-SEM

In general, structured equation modelling (SEM, [10, 87]) aims to measure the causality relation between concepts (i.e. latent variables, not directly observable) starting from some observed indicators (i.e. manifest variables). The main two approaches for estimating parameters in a SEM model are:

- **Covariance-based method**: based on the covariance matrix between manifest variables.
- **Component-based method**: based on the research of particular latent components.

In particular, PLS-SEM algorithm is a component-based method, and before analysing it in detail we want to make its concept clearer. It is important the use of PLS-SEM for creating composite indicators, that is our main objective [43]. As viewed above, this technique aims to measure causality relation between concepts, that we will call latent variables (LVs), starting from some manifest (MVs), by an exploratory approach: the explained variance of the endogenous latent variables (e.g. the outcomes) is maximized by estimating partial model relationships in an iterative sequence of ordinary least squares regression [101]. Another essential point is that PLS-SEM does not require any preliminary assumptions for the data distribution, so it's called a *soft-modelling* technique.

PLS-SEM is characterized from three frameworks: the structural (inner) model, the measurement (outer) model and the weighting scheme (i.e. its own distinctive component). Graphically, it is represented in a user friendly way, thanks the so-called path diagram (Fig. 2.1): contrary to the CBSEM approach, in the PLS context each MV (in each rectangle) is just connected to one LV (in the circle). Moreover all arrows connecting a LV with its block of MVs must point in the same direction and the connections between LVs and MVs are referred as measurement or outer model. In particular, a LV that point towards another one is named exogenous, while the outcome LV is called as endogenous. For what concerning the outer model (Fig. 2.1), when all arrows pointing outwards, it is called a Mode A model (reflective measurements); a model with all arrows pointing inwards is called a Mode B model (formative measurements).

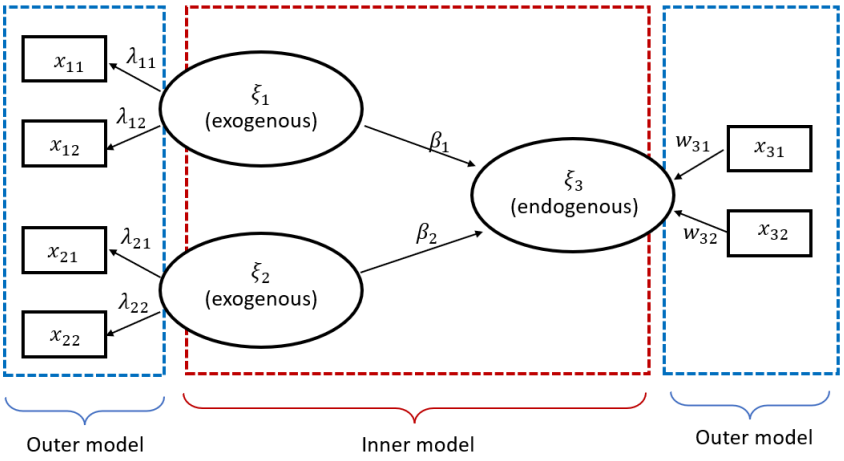


FIGURE 2.1: PLS-SEM: an example of path diagram

The **measurement (outer) model** relates observed variables (MVs) to their latent variables (LVs, also named factors). Such as mentioned before, in the PLS framework one MV can only be linked to one LV, and all MVs related to one LV form a block; so, each LV has its own block of indicators (at least one, like in Fig. 2.1). Before proceeding we assume that all MVs in our data matrix $\mathbf{X}$ are normalized and that each block $\mathbf{X}_g$ of MVs is positively correlated for all LVs $\boldsymbol{\xi}_g$, $g = 1, ..., G$. Now, we are ready to understand the two type of outer model:

- **Reflective measurement (Mode A)**: as we can see in Fig. 2.2 a typical example of reflective model is the case of intelligence test, where each block of MVs (in this case each question of the test) reflects its LV; note that reflective indicators are interchangeable, in fact if we remove an item we do not alter the underlying concept (in this case the intelligence). It assumes also uni-dimensionality for each block of MVs (just one latent concept is reflected on different indicators).

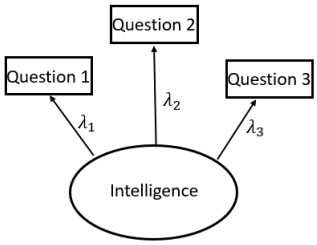


FIGURE 2.2: Example of reflective path model

In this framework, each block of MVs can be written as a multivariate regression:

$$\mathbf{X}_g = \boldsymbol{\xi}_g \boldsymbol{\lambda}_g^T + \mathbf{F}_g, \qquad E[\mathbf{F}_g | \boldsymbol{\xi}_g] = \mathbf{0} \tag{2.1}$$

Where $\boldsymbol{F}_g$ indicates the error terms and $\boldsymbol{\lambda}_g$ is the matrix of loadings, that can be estimated thanks OLS. Keep in mind that, since we normalized data at the beginning, we hold the constraint to have unit variance. Take in consideration that, since each MV is a simple linear regression between its correspondent LV and the estimated loading, we haven't any multicollinearity problem between the indicators of each block. The main thing to verify in this case is the

uni-dimensionality of each block, in order to validate our model, but we will deepen this topic in Sec. 2.1.3.

- **Formative measurement (Mode B)**: formative indicators are considered as causing (i.e. forming) a latent variable (i.e. an emerging construct). In this case, there may be theoretical or conceptual reasons to consider a block as formative: this implies a strong consensus among experts about how the latent variable is formed [113]. For instance, in Fig. 2.3 we can see the well-being as LV, caused from its own MVs (i.e. health, income and employment). Compared with the reflective model, in this framework if we omit one MV we lose a part of the concept [47]; that is, formative indicators are not supposed to be correlated and for this reason, they cannot be evaluated in the same way of reflective measures.

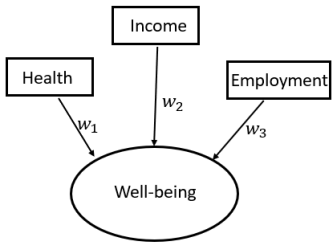


FIGURE 2.3: Example of formative path model

From the theoretical point of view, each LV is considered to be formed by its MVs following a multiple regression:

$$\boldsymbol{\xi}_g = \mathbf{X}_g \mathbf{w}_g + \delta_g, \qquad E[\delta_g | \mathbf{X}_g] = \mathbf{0} \tag{2.2}$$

And here below the weighting block is estimated by least squares:

$$\begin{aligned} \hat{\mathbf{w}_g} &= (\mathbf{X}_g^T \mathbf{X}_g)^{-1} \mathbf{X}_g^T \boldsymbol{\xi}_g \\ &= VAR(\mathbf{X}_g)^{-1} COV((\mathbf{X}_g, \boldsymbol{\xi}_g) \\ &= COR(\mathbf{X}_g)^{-1} COR((\mathbf{X}_g, \boldsymbol{\xi}_g) \end{aligned} \tag{2.3}$$

Keep in mind that $\mathbf{X}_g$ is a matrix when the LV $\boldsymbol{\xi}_g$ is measured by a block of more than one MV: noting, in that case $VAR(\mathbf{X}_g)$ refers to covariance matrix. It is useful to compare the outer weights of each MV in order to determine which indicators contribute most effectively to the construct. Attention must be paid in order to avoid misinterpreting relative small absolute values of weights as poor contributions. If we are considering the elimination of some indicators, this should be done based on multicollinearity (see better in Chapter 3).

The **structural (inner) model** describes causality relation between LVs. For this purpose, it divides LVs in two groups: exogenous and endogenous; in particular, the first one does not have any predecessor in the path diagram, the rest are endogenous. For example, if we suppose to have just one endogenous LV (and the rest as exogenous), the linear equation of its own structural model is:

$$\xi_{end} = \sum_{g=1}^{G} \beta_{g,end} \xi_g \ + \zeta_{end} \tag{2.4}$$

Where the error terms ζ are assumed to be centred, β_g links the *g-th* exogenous LV to the endogenous one and it can be estimated in different ways [137, 92]:

- Centroid scheme:
$$\beta_g = sign[Cor(\xi_g, \xi_{end})] \tag{2.5}$$
It shows some problems with very low correlation (i.e. $\simeq 0$).

- Factorial scheme:
$$\beta_g = Cor(\xi_g, \xi_{end}) \tag{2.6}$$

- Structural scheme: β_g is an OLS coefficient of ξ_g that impact on the endogenous ξ_{end}.

Before focusing on the model assessment, we summarize the PLS-SEM algorithm, since it follows an iterative procedure that alternate outer and inner model estimation until convergence is achieved, as follow:

1. Arbitrarily choice of the weights.
2. Measurement model: outer estimates for each LV.
3. Structural model: inner estimates for each endogenous LV.
4. Uploading the outer weights, following the relative scheme (A or B) and return to the point 2 until convergence is achieved (the relative difference between outer weights respect to the previous iteration is less than a small fixed threshold).

When convergence is achieved, the algorithm calculates each LV such as linear combination of its own MVs by weights obtained from the iterations. Finally, PLS-SEM procedure computes path coefficients of the inner model by following the scheme specifed (2.4).

2.1.3 Assessment and validation

When we work with composites and in particular with a PLS-SEM model, we are not sure from the beginning that is the perfect approach for our framework: so, it is useful a method able to assess and validate our model. In order to reach this goal, due to exponentially growing of PLS-SEM, many researchers tried to build some guidelines [80]. Here we propose an in-depth analysis about three important assessment index in the PLS-SEM literature:

- **Communality index**: it measures the goodness of outer model; in particular, it specifies how many variability of one MV is explained by its own LV. For example, given a block of MVs g, it is computed as follow:
$$COM_g = \frac{1}{p_g} \sum_{p=1}^{p_g} Cor^2(x_{pg}, \xi_g) \tag{2.7}$$
Where, in (2.7) p_g indicates the total number of MVs in the block g. As consequences, for each MV x_{pg} in the model, communality is the squared of the correlation coefficient between it and the correspondent LV. Derived from (2.7),

reliability of LVs is measured by the amount of variance that the LV captures from its own indicators, as average communality:

$$\overline{COM} = \sum_{g=1}^{G} p_g COM_g \ / P \tag{2.8}$$

Where G is the total number of blocks and P is the total number of MVs.

- **Redundancy index**: it measures quality of the structural model for each endogenous LV, keeping in consideration also the measurement model. Called J the total number of endogenous LV, it states the powerful of exogenous LVs to predict the outcomes. We can express it as the product between the $j-th$ endogenous block and R^2 index of its structural relation:

$$RED_j = COM_j \times R^2(\xi_j, \xi_m^{'}) \tag{2.9}$$

Where $\xi_m^{'}$ indicates all LVs related to the outcome ξ_j. In order to measure the global quality of the structural model, it is useful to compute the average redundancy for all the blocks of endogenous LVs:

$$\overline{RED} = \frac{1}{J} \sum_{j=1}^{J} RED_j \tag{2.10}$$

- **GoF index**: the last, but not the least, we find the goodness of fit (GoF); it is an intermediate solution, because it considers both the inner and outer model in its computation:

$$GoF = \sqrt{\overline{COM} \times \overline{R^2}} \tag{2.11}$$

With $\overline{R^2}$ computed as follow:

$$\overline{R^2} = \frac{1}{J} \sum_{j=1}^{J} R^2(\xi_j, \xi_m^{'}) \tag{2.12}$$

This is the most synthetic index in PLS-SEM world, but we must pay attention and interpret it with prudence, as suggested from the scientific community, since it is just a geometric mean between the inner and outer model performance and it does not say us if the model is replicable in other context.

With the aim to validate the index (and so, also the model) described above, all the software introduced at the beginning of Sec. 2.1 propose a bootstrap validation, which let us to obtain a confidence interval for each parameter, testing their significance.

Finally, in this paragraph we showed the most important assessment index in the PLS-SEM framework, then we have seen how to evaluate it, in particular how to assess and verify the reliability of the outer and the inner model. We tried to give some guidelines, by recent growing literature in the PLS-SEM world.

2.1.4 Higher order PLS-SEM

Now, we talk about an important extension applicable in the PLS-SEM framework: the Higher-Order Construct Models, also known as Hierarchical Models. As the

name says, they contain LVs of "higher-order" (HOCs) and the conceptual idea behind them is that they are supposed to be at a higher level of abstraction. We have already seen that LVs represent abstract of theoretical concepts, but sometimes we need extra LVs representing other constructs [113]. In Fig 2.4 we can see the well-known example of higher-order construct (HOC), that comes from psychometric literature: the General Intelligence Ability (second-order), that is supposed to be reflected by three lower order constructs (first order, LOCs) ability (verbal, numerical and spatial). The application of hierarchical models is often limited to a second-order structure, but in literature there are applications with third, fourth or higher order latent variables.

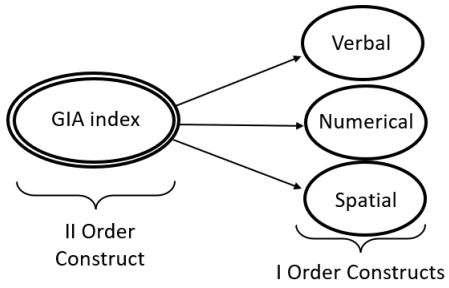


FIGURE 2.4: Example of higher-order construct: the GIA index

Noting that there are two types of hierarchical models in literature: when the higher-order LV is reflected by some lower order LVs we refer as *molecular* model (Fig. 2.4), while if the higher-order LV is formed by some lower-order constructs we refer as a *molar* model. There is also a more recently classification that, in addition, keeps in consideration the relation among the first-order latent variables and their manifest variables [3]. Following this suggestion, we can classify four types of higher order constructs (Fig. 2.5):

- The Type I is the *Reflective-Reflective Measurement Model*: it is one of the most frequently applied framework, it is used when both the HOC and the LOCs are reflective constructs.
- The Type II is the *Reflective-Formative Measurement Model*: here the LOCs are selectively measured constructs that do not share a common cause but rather form a general concept that fully mediates the impact on subsequent endogenous variables [38]. In the last ten years, this type of model has become the most widely used in empirical applications [3, 39].
- The Type III is the *Formative-Reflective Measurement Model*: in this case, the HOC is a common concept of several specific formative LOCs. There aren't many examples in the empirical literature, but an interesting application is the firm performance, seen as a reflective HOC measured by several different LOCs indices [3].
- The Type IV is the *Formative-Formative Measurement Model*: this framework is appropriate when both the HOC and LOCs are formative constructs. This last application has only a lack: there is a large need for guidelines on the use of this framework in PLS-SEM [3].

Whereas for the reflective relationship there are some available guidelines [41, 49], for what concern formative constructs the situation is different: despite some empirical studies indicate the predominance of formative hierarchical models, a clear

guideline on their use is lacking in the literature [91]. It is important to underline

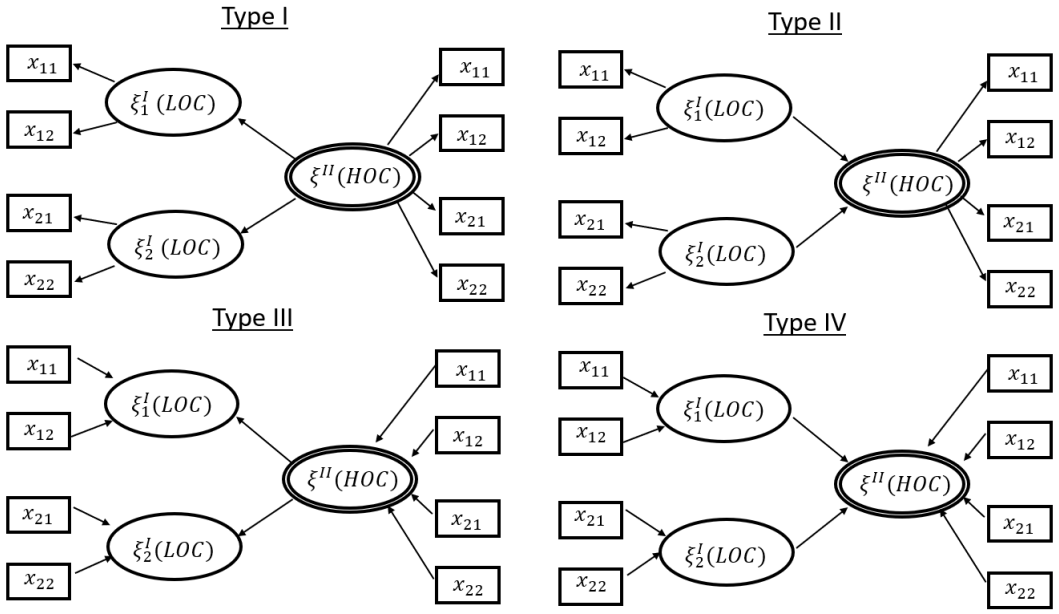


FIGURE 2.5: The four types of higher order constructs

also the fact that a higher-order LV, since it is an "artificial" latent variable, it does not have by default any MVs, and we remind that LV with no indicators has no place in the PLS-SEM models. In order to fix this problem, there exist two main approaches used in literature for the HOC estimation:

- **Repeated indicators approach**: this is the simplest technique in the framework of hierarchical models [131]. It consists of taking all MVs of the lower order LVs and using them as MVs for the higher-order construct (Fig. 2.6), then apply to this framework the PLS-SEM algorithm. The main weakness of this approach is that all indicators must be treated just in reflective way [113], because of multicollinearity problems.

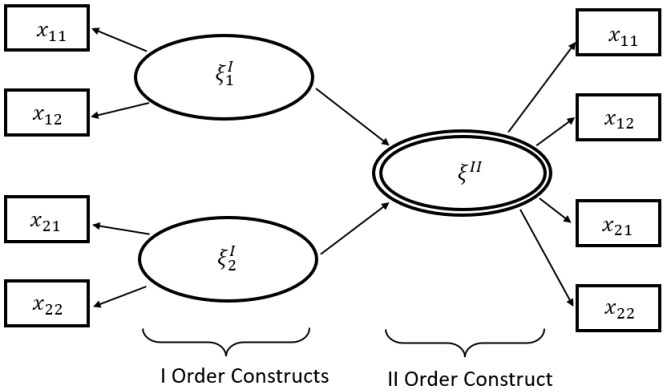


FIGURE 2.6: The repeated indicators approach

- **Two-step approach**: as its name suggest, this technique is divided in two steps, where in the first we have to apply PCA or FA (i.e. factor analysis) on each

block of LOC, then we must save their scores (e.g. the first principal component for each one, [113]); in the second phase of the procedure we will apply PLS-SEM using the computed scores as MVs of the higher order LV (Fig. 2.7).

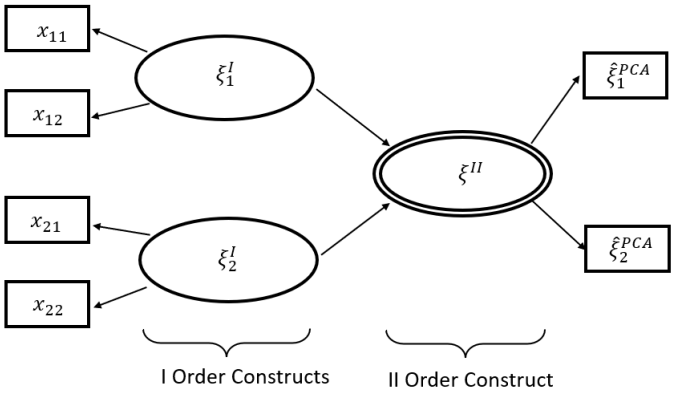


FIGURE 2.7: The two-step approach

In particular, $\hat{\xi}_1^{PCA}$ and $\hat{\xi}_2^{PCA}$ are the scores (i.e. first principal component) from the lower order LVs (respectively, the LVs number one and two). Moreover, also this approach has its own weak point: the first is that just one component is chosen for each block of lower LVs; the second is that this component has a strong representative power but a weak predictive power [41, 49].

In order to overstep some limits of the previous methods, researchers are studying different solutions. Here below we propose two interesting techniques [41, 49], tested at the moment only for the case of second-order constructs:

- **Mixed two-step approach**: this method is an extension of both the two techniques described above, and it has the goal to solve the issue related to the predictive power of the component for each first-order construct. In order to do this, it provides for the following steps (Fig. 2.8):

  1. We have to apply a repeated indicators approach.
  2. We must save the scores of first order LVs, use them as MVs for the second order LV and re-apply PLS-SEM in this new framework.

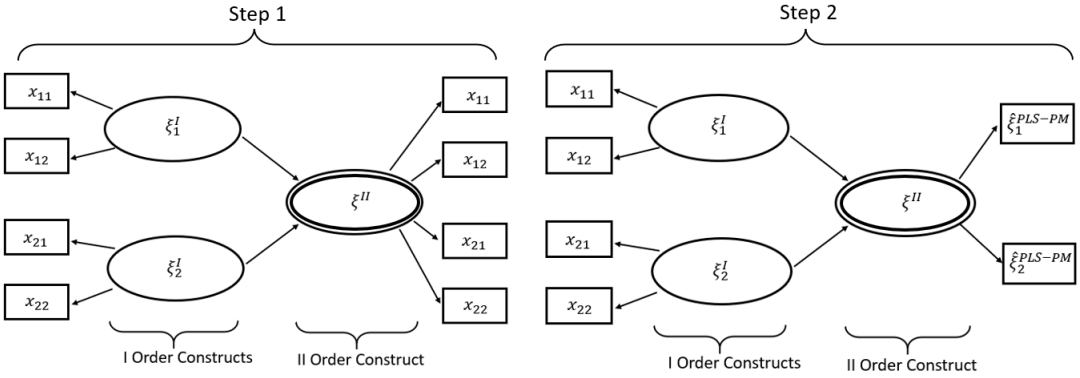


FIGURE 2.8: The mixed two-step approach

- **PLS-CR approach**: this other method aims to solve the problem regarding the choice of components number for each block of MVs. In fact, PLS Component Regression gives the possibility of choosing the number of components to be extracted manually or according to a criterion. Moreover, it provides components that are at the same time representative of their blocks and predictive of the second-order construct. In particular, first-order LVs are considered as blocks of predictors and the second-order construct as a block of response variables. It follows these steps (Fig. 2.9):

  1. We must apply PLS regression for each block of first-order LVs.
  2. Then, once h components for each block have been obtained, these will be the MVs of the higher-order LV.
  3. Finally, we must apply PLS-SEM to the framework obtained at the previous step.

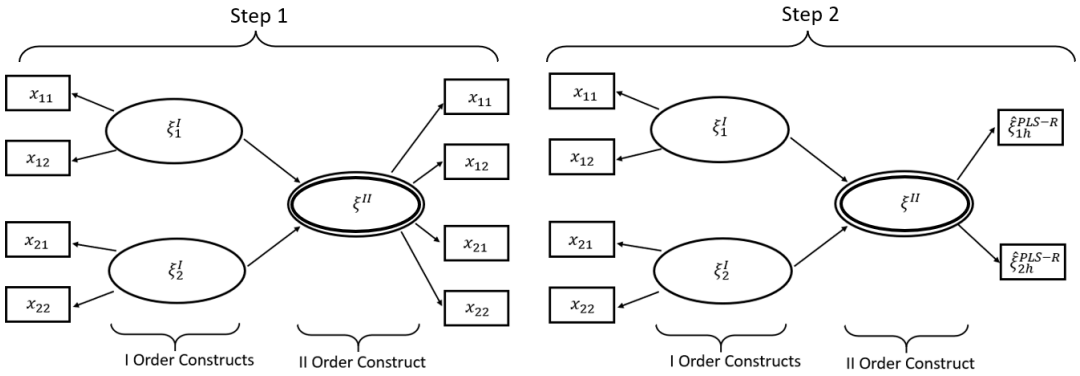


FIGURE 2.9: The main steps of PLS-CR approach

Researchers made some simulation studies, in order to compare these different approaches; for example, the last two methods seen before slightly outperform, in terms of prediction accuracy, the two-step approach. Furthermore, they are preferable in terms of the BIAS and MSE of the estimates (i.e. smaller errors than two-step method). Finally, working with a small sample (saying, less than 100 units), the mixed two-step approach is recommended (i.e. better quality of the model), while with a large sample size the two algorithms have similar results, with performance slightly better for PLS-CR. Again, [49] tested all four types of HOCs (Fig. 2.5) thanks a simulation study, confirming how the mixed two step and the PLS Regression approaches are always the best choices, in terms of bias and MSE (i.e. mean square error) of the estimates. At the moment for third or higher-order PLS-SEM the best tested method is the two-step approach yet, but researchers are moving also to improve these frameworks.

2.1.5 Moderating effects in PLS-SEM

At this point, we do a brief panoramic about moderating or interaction effects, that are another question to take into account when we work with PLS-SEM. In particular, they represent the influence that a third variable has on the relationship between an independent (i.e. in our case the exogenous LV) and a dependent (i.e. in our case the endogenous LV) variable (Fig. 2.10). Moreover, the moderator variable, that for

simplicity we call MoV, can be qualitative (e.g. gender, ethnicity) or quantitative (e.g. age, income).

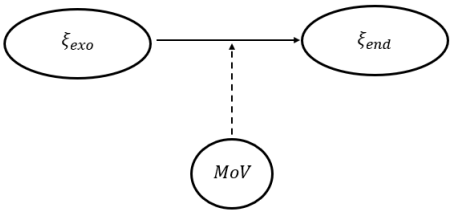


FIGURE 2.10: Example of interaction effect between two LVs

In order to study moderating effects, there are two main options:

- **Group comparisons**: it is useful when MoV is qualitative, in order to evaluate the observed heterogeneity between units (see better in Sec. 2.1.7); for example, some software and tools cited in Sec. 2.1.1 have implemented a function that compare, thanks a bootstrapping resampling or a permutation test, two or more groups and test if difference among them is significant or not [113].
- **Moderator constructs**: it is better to apply this one when the MoV is treated as LV. As consequence, under this approach, moderator variables are just considered in the structural model. It has preferred to use this approach when MoV is quantitative, but we can also if it is qualitative. There are three main ways to study the effects of latent MoV:
    1. **Product indicator approach**: as its name says, thanks to this method we have to add another LV, that we can call as $Interaction$, whose MVs are all possible products between all MVs of the exogenous LV and all MVs of the MoV. Then, thanks to a bootstrap validation we can see if the interaction terms are significant or not.
    2. **Two-stage PM approach**: it consists on two stage, where in the first we must apply the classic PLS-SEM to the whole model; in the second step we have to take the scores obtained in the first stage for creating the interaction term and perform a second PLS-SEM including the scores as indicator of the constructs [113].
    3. **Two-stage regression approach**: in its first stage we have to apply the PLS-SEM classic, while in the second we must perform a regression analysis with the scores of the first stage [113].

2.1.6 Mediation effect in PLS-SEM

Mediation occurs when a third mediator variable intervenes between two other related constructs. More precisely, a change in the exogenous construct causes a change in the mediator variable, which, in turn, results in a change in the endogenous construct in the PLS path model. Analyzing the strength of the mediator variable's relationships with the other constructs allows substantiating the mechanisms that underlie the cause-effect relationship between an exogenous construct and an endogenous construct. In the simplest form, the analysis considers only one mediator variable, but the path model can include a multitude of mediator variables simultaneously [75].

2.1.7 Heterogeneity in PLS-SEM

Until now, we have implicitly assumed that data adopted in our PLS-SEM analysis came from a homogeneous population, but in reality, this assumption is often unrealistic. In fact, people (with their peculiarity), corporations (with their structure), or environments (with their dynamism) are frequently different, and probably if we analyse all data without take into account this heterogeneity is likely to produce misleading results and incorrect conclusions [116]. For example, as mentioned in Sec. 2.1.5, if we consider a dataset of consumers, and we take into account as moderator variable the people's gender, from a technical prospective we must split our dataset into two consumer groups and this imply to estimate two different models. Very relevant, if we fail to recognize the heterogeneity between groups and we analyse the whole dataset, the path coefficients estimates offer an unrealistic picture of the model relationships [75]. As consequence, it is very important to identify, assess, and, if there is, treat heterogeneity in our data. When differences between two or more groups of units are referred to observed features, such as gender, age, or nationality, this is called **observed heterogeneity**, and it is used to treat it by splitting the dataset and creating some different PLS-SEM models, one for each group (e.g. one for males and another for females) and to evaluate their differences (Sec. 2.1.5). Opposite, if differences between two or more groups of data do not emerge a priori from observable characteristics but they appear in the structural path coefficients, we talk about **unobserved heterogeneity**. In order to take into account unobserved heterogeneity, researchers firstly tried to apply the classical clustering technique, such as the well-known K-means algorithm [94] on the indicator data, or LV scores derived from a proceeding analysis of the entire dataset [115]. The typical approach by apply a clustering method followed in sequence from a regression or other analysis on the obtained segmented data has called as tandem analysis; but for the case of PLS-SEM there is a problem: clustering techniques ignore path model relationships specified at the beginning of our analysis, that is essential for a PLS-SEM model. Therefore, some others researchers has shown how traditional clustering performs very poorly in identifying group differences in PLS-SEM [116]. Seeing the limitations of sequential approaches, methodological research has suggested some specific methods to identify and treat unobserved heterogeneity, commonly referred to as latent class techniques. In this context, the most relevant algorithms (i.e. able to account for sources of heterogeneity in the structural model) recognised from the PLS-SEM community, each one with its own pros and cons, are:

- **FIMIX-PLS**: the finite-mixture segmentation approach [65], is the pioneer technique for identifying unobserved heterogeneity in PLS-SEM models. Its peculiarity is that it assumes each endogenous LV distributed as a finite mixture of conditional multivariate normal densities; it uses then these densities to estimate probabilities of segment memberships for each observation (proportional assignment) to optimize the likelihood function. At the beginning, this algorithm provides a random split of the observations and in each iteration there is a proportional assignment of all units to all segments based on the conditional multivariate normal densities, in order to optimize the likelihood function. FIMIX-PLS stops when there is a very small improvement in the log-likelihood (i.e. under a fixed value) or if the maximum number of iterations has reached.

- **REBUS-PLS**: this algorithm [58, 130, 113] is a sort of "upgrade" respect the previous techniques, in fact it does not require any preliminary assumptions; its

only limit (not to underestimate), it is feasible just with reflective blocks. It is a so called distance-based clustering approach, based on communality residuals of all LVs and structural residuals of all endogenous ones.As initial step, it provides a hierarchical classification based on communality and structural residuals of the overall model, while in each iteration it assigns all observations to the closest segment. REBUS-PLS routine ends when stability of the classes' composition holds, in alternative when the maximum number of iterations is reached.

- **PLS-POS**: it is the only method [4, 75] that take into account for sources of heterogeneity in formative constructs. On the other hand, it does not hold for reflective measures. It has not the pre-clustering phase: in fact it provides randon split of units and assignment to the closest segment according to the distance measure. So, also PLS-POS is a distance-based clustering approach, based on structural residuals of all endogenous LVs with an extension that also accounts for heterogeneity in formative measures. For each iteration, this procedure assigns only one observation to the closes segment and assures improvement of an objective criterion (R^2 of all endogenous LVs) before accepting the change. PLS-POS stops when there is an infinitesimal improvement in the objective criterion (i.e. or maximum number of iterations).

In Tab. 2.1 we have tried to summarise all pros and cons for each one of the previous methods, in order to give a clearer comparison among them.

TABLE 2.1: A summary comparison between clustering algorithms

Feature	FIMIX-PLS	REBUS-PLS	PLS-POS
Distrib. assumptions	Yes	No	No
Pre-clustering	No	Yes	No
Distance measure	No	Yes	Yes
OK for reflective measures?	No	Yes	No
Ok for formative measures?	No	No	Yes
Ok for structural model?	Yes	Yes	Yes
Observations assignment	All	All	Only one
Stop criterion	Small impr./max it.	Stability/max it.	Small impr./max it.

2.1.8 PLS-SEM limits

Until now we have seen all peculiarities and features of PLS-SEM, its non-parametric nature, but when we apply this technique we must pay attention also to some critical point, emphasize in different researches. For example [56] proofed the inconsistency of PLS-SEM in the reflective approach by showing adverse results in the hypothesis test. The solution proposed is a framework called PLS consistent (i.e. just for reflective constructs), that performs particularly well when the initial data are not normally distributed. [111] focus on the lack of methodological justification for PLS-SEM, mainly on the use of PLS weights, that for them have no firm basis in statistical theory, since it is an heuristic-method. They showed that PLS-SEM is regression with scale scores and thus has very limited capabilities to handle the wide array of problems for which applied researchers use structured equation modelling (SEM). So, [110] proposed an alternative way to estimate a PLS-SEM model, working with data covariance matrix, designed to be computationally efficient. Some other extensions

of the base PLS-SEM have been developed, in order to overcome the PLS-SEM limits: one of these is the GSCA (i.e. Generalized Structured Component Analysis, [83]), that replaces factors by exact linear combinations of observed variables. It employs a well-defined least squares criterion to estimate model parameters. As a result, GSCA avoids the principal limitation of partial least squares (i.e. the lack of a global optimization procedure) while fully retaining all the advantages (e.g. less restricted distributional assumptions).

Like all the statistician techniques, also PLS-SEM have pros and cons; the main things we must take into account before to be inclined towards a PLS-SEM analysis, are essentially three: we have to know very well the data to be analysed, we must have clear in mind the objective of our analysis and it is recommended to collaborate with experts in order to develop a consistency theory under our path model. With this purpose, in the following part of this chapter an exploratory application [26] for football analytics will be presented.

2.2 The application

In recent years, football, the most watched sport in the world, has been moving towards a data-driven revolution [25]; in particular, football analytics was born with the aim to predict the results of a match, and there have been many papers on this subject [23]. Furthermore, starting a few years ago, this field of research has been moving towards the evaluation of players' performance, which is becoming a strategic key for football coaches in the management of the team. For this purpose, different approaches have been developed: for example, Pappalardo [102] adopted a support vector machine (SVM) observing match outcomes to evaluate players' performance, Schultze and Wellbrock [120] created a rating index employing a plus-minus metric, and Carpita [21] adopted an unsupervised method to classify different areas of performance. Attention will be focused on this last issue: in fact our goal is to explore players' key performance indices (KPIs), in order to evaluate and weight different strategic skills; this can be useful for understanding any of a coach's key choices, as well as to guide decisions to transfer a player, contract negotiations, and to improve future predictive modelling. We must keep in mind that players' performance has been tried to measure by using data coming from a survey organised by Electronic Arts (EA) Sports experts that combine the subjective evaluations of over 9000 scouts, coaches and season ticket holders -who watch as many live matches as possible- into ratings for over 18000 players (EA FIFA[2] ratings, [95]); they constantly maintain this database with systematic and periodical data collection. These data are also integrated in one of the most famous football videogame FIFA by EA Sports[3] experts. In particular, the EA Sports experts consider 6 performance dimensions (latent traits) defined by 6 composite indicators, each one with specific KPIs which can be combined into the well-known EA *overall* indicator: using these data, McHale et al. [98] already developed a player performance rating system for the English Premier League, but without taking into account the player's position or role; Matano et al. [95] combined FIFA ratings and an adjusted plus-minus approach into a single metric, whereas Kirschstein and Liebscher [89] used these data to predict and assess the market value of a soccer player. More recently, Biecek and Burzykowski [7] tried to use these data for the 2018/2019 season to evaluate and combine players' performance and market value.

[2] `www.FifaUltimateTeam.it`

[3] `www.easports.com`

Along with other widely available data (i.e., players' wage and monetary value), performance attributes guide strategies for forming competitive sports teams: rather than relying exclusively on subjective and error-prone intuition, scouts, technical directors and coaches turn to plausible, available and up-to-date data to select players for their teams or to determine the team line-up [6]. A number of studies confirm how experts' evaluations proved useful not only for skills classification, but also for the prediction of players' monetary value as well as merchandising potential [44, 89].

But at this point the main problem is that experts' opinions are not statistically supported [20, 22] and furthermore it is not clear how they keep in consideration players' heterogeneity (their roles on the field for example): Carpita et al. [21] found some significant differences in performance KPIs depending on the player's role, studying their densities and presenting a preliminary model, but without taking into consideration the viewpoint of the sports scientist. In fact, It has been verified that there is no agreement among experts in defining a unique model of performance. For this, sport scientists have proposed different approaches, for example, from a technical-tactical point of view [60] or keeping in consideration the ball-in-play time of a match [105], to studying the physical peak demand for each player. In order to improve the preliminary model defined by Carpita et al. [21] from a statistical point of view and to take into consideration different opinions from both *sofifa* experts and sport scientists, a partial least-squares structured equation model (PLS-SEM, [133]) that uses a third-order approach ([113]) has been developed, improving the early bird PLS-SEM second-order model presented by Cefis and Carpita [30], with the final goal to explore and measure players' performance in its totality (i.e. an overall indicator) and its subgroups (e.g. the 6 areas of performance defined by EA experts). Infact, as pointed by experts, a player needs some complementary abilities on the field [82] and it is useful also taking in consideration heterogeneity among them (i.e. role and league). The goal is to make easier for policy makers, managers and sport scientists the players evaluation, by different sub-areas of performance, that is more and more crucial for a football team. In addition, from a sport scientist point of view, sport evaluation is a fundamental moment in the training process of athletes and teams and is an indispensable support for the coach [57].

After this introduction, this application is organised as follows: data and roles classification are presented (Sec. 2.2.1), then an in-depth analysis regard the framework developed in Sec. 2.2.2 and finally some results are shown in Sec. 2.2.3.

In summary, this is an exploratory work, that aims to replicate the EA *overall* indicator by an innovative Third-Order PLS-SEM measurement model, taking into consideration both statistical evidence and experts opinion, validate it, create a full model (i.e. with all players) and compare this one by splitting data among roles and leagues, in order to take into consideration the observed heterogeneity.

2.2.1 Data and role classification

For this application data provides from EA experts and available on the famous Kaggle data science platform by Leone[4] has been used; in particular, the focus will be on all players' stats from the top 5 European Leagues (e.g., Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). This dataset contains another 28 variables (e.g. KPIs), with periodic player's performance on a

[4] `www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset`, named "Fifa 20 complete player dataset"

0–100 scale with respect to different abilities, classified by *sofifa* experts into 6 latent traits of performance: *attacking, skill, movement, power, mentality, defending* (for a detailed classification, see Tab. 2.2); caused of the purposes it has been used data relying on the beginning of the season 2018/2019, so our dataset was composed from the stats of about 2662 players (note that goalkeepers have been excluded from the analysis, due their singularity role).

TABLE 2.2: Classification of KPIs.

sofifa LV	KPIs/MVs	MVs label
Attacking	Crossing	*att1*
	Finishing	*att2*
	Heading accuracy	*att3*
	Short passing	*att4*
	Volleys	*att5*
Skill	Dribbling	*ski1*
	Curve	*ski2*
	FK accuracy	*ski3*
	Long passing	*ski4*
	Ball control	*ski5*
Movement	Acceleration	*mov1*
	Sprint speed	*mov2*
	Agility	*mov3*
	Reaction	*mov4*
	Balance	*mov5*
Power	Shot power	*pow1*
	Jumping	*pow2*
	Stamina	*pow3*
	Strength	*pow4*
	Long shot	*pow5*
Mentality	Aggression	*men1*
	Interception	*men2*
	Positioning	*men3*
	Vision	*men4*
	Penalties	*men5*
Defending	Marking	*def1*
	Standing tackle	*def2*
	Sliding tackle	*def3*

In addition, to classify roles, the main advice of football experts has been followed [82] in order to get the specific role of each player (and not the classical three roles, such as defender, midfielder or forward): we can see better this classification in Fig. 2.11. Recall that goalkeepers have been excluded from the analysis, due to their singular role.

FIGURE 2.11: Players' roles classification by experts on the pitch

2.2.2 The framework developed

By following the suggestions provided by experts [105, 60], the PLS-SEM framework was developed to manage simultaneously these two models:

- The measurement (outer) model, that links MVs (KPIs) to their corresponding LVs. Each block of MVs $\mathbf{X}_g, g = 1, \ldots, G = 6$ (Tab. 2.2) must contain at least one MV and this relation has been treated in a formative way (MVs are the causes of their own LV, [36]). In particular, it has been assumed each LV ξ_g as formed by its KPIs following a multiple regression (2.13), where $\mathbf{w}_g$ is the vector of outer regression weights and δ_g of error terms, with their conditional expected value assumed to be zero (2.14). Finally, the vector of outer weights for the g-th LV is estimated by OLS (2.15, Mode B).

$$\xi_g = \mathbf{X}_g \mathbf{w}_g + \delta_g \tag{2.13}$$

$$E[\delta_g | \mathbf{X}_g] = \mathbf{0} \tag{2.14}$$

$$\mathbf{w}_g = (\mathbf{X}_g{}^T \mathbf{X}_g)^{-1} \mathbf{X}_g{}^T \xi_g \tag{2.15}$$

  In this case (i.e. with formative constructs) PLS-SEM computes for the outer model also the loadings (i.e. λ), which represent correlations between MVs and their own correspondent LV estimated [113].

- The structural (inner) model, that divides LVs in two groups: exogenous and endogenous. The first one does not have any predecessor in the path diagram, the rest are endogenous (Fig. 2.12). For the j-th endogenous LV in the model, the linear equation of its own structural model is defined in (2.16), by standardized data; in particular, R represents the number of exogenous LVs that affect the endogenous one and β_{rj} is so called path coefficient, a linkage between the r-th exogenous LV and the j-th endogenous one, where ζ_j is the error term.

$$\xi_j = \sum_{r=1}^{R} \beta_{rj} \xi_r + \zeta_j \tag{2.16}$$

Where the error terms ζ_j are assumed to be centred, β_{rj} is so called path-coefficient, it links the r-th exogenous LV with the endogenous one and it can be estimated by following one of these more recently approaches [92]:

- Factorial scheme: β_{rj} is the correlation coefficient between the endogenous LV ξ_j and the exogenous one ξ_r.
- Structural scheme: β_{rj} is an OLS coefficient of ξ_r that impact on the endogenous ξ_j.

For the purpose of this work, it was used the factorial scheme. So, you must keep in mind, in order to avoid misinterpretation, that β coefficients which you will find respectively in Fig. 2.13 and in Fig. 2.18 are the correlations between each exogenous and endogenous LVs and so they are not regression coefficients.

A Higher-Order construct model (HOC, also called hierarchical model) is adopted, so LVs that represent superior levels of abstraction can be included. We have already seen that LVs represent abstract of theoretical concepts, but sometimes we need extra LVs representing other constructs [113]. In particular, a third-order model is used. In fact, for the purpose of this project, players' performance was built as an extra-latent construct of higher (third) order, formed from two extra LVs (second order constructs), namely, *Off_phase* (the phase where a player is in attack, i.e. in the opposite midfield, with or without ball possession) and *_phase* (the phase where a player is in defense, i.e. in his own midfield, with or without ball possession) [105, 60]. It has been assumed that the initial 6 *sofifa* LVs (first order constructs) contribute to the second-order LVs in the following way: all first order LVs except *defending* shape the *Off_phase*, while all LVs except *attacking* contribute to *Def_phase* (Fig. 2.12).

Since our framework have a third order construct, in order to avoid some collinearity problems, a two-step or patch approach is used [113]: in the first step of this method, a principal component analysis (PCA) is used to obtain scores of the lower-order LVs (the first principal component - I PC - of each one), and in the second step standard PLS-SEM uses these PCs as MVs for the higher order LVs. Consequence of this approach, we assumed a reflective relation between the HOCs and their own MVs. Of course, this approach has its own weak points: the first is that just one component is chosen for each block of lower LVs; the second is that this component has a strong representative power but a weak predictive power [24]. In order to exceed that limits, in the last years have been developed some new techniques in the estimates of HOCs, like the Mixed Two Step approach or the Partial Least Squares Component-Regression approach, but at the moment they are tested only for the case of second order constructs [41].

In particular, the MVs adopted for *Off_phase* are the I PCs of *attacking, skill, movement, power* and *mentality* (labelled as *offcomp1* to *offcomp5* in the graphs of Sec. 2.2.3), while for *Def_phase* the I PCs of *movement, power, mentality, defending* and *skill* (*defcomp1* to *defcomp5* in the plots of Sec. 2.2.3). Finally, the I PCs of *Off_phase* and *Def_phase* are used as the MVs for Performance.

For the measurement (outer) model estimation of each HOC, as already introduced, a reflective (Mode A) relation between the HOCs and their own MVs was assumed, by construction (i.e. PCA): each block of MVs reflects its own LV; note that reflective indicators are interchangeable, in fact if we remove an item we do not alter the underlying concept. It assumes also uni-dimensionality for each block of MVs

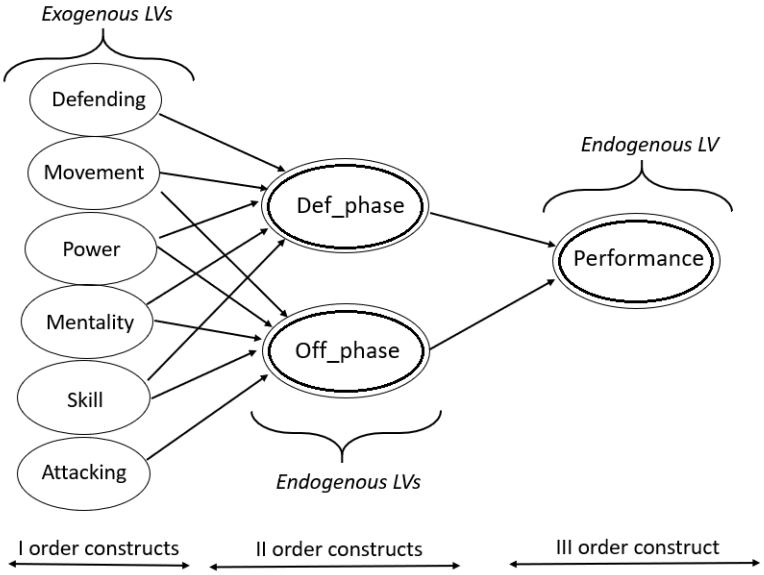


FIGURE 2.12: Path diagram: the third-order inner model for players' performance

(just one latent concept is reflected on different indicators).

$$\mathbf{PC}_g^I = \boldsymbol{\xi}_g^{HOC}\boldsymbol{\lambda}_g^T + \mathbf{F}_g, \qquad E[\mathbf{F}_g|\boldsymbol{\xi}_g^{HOC}] = \mathbf{0} \tag{2.17}$$

Where $\boldsymbol{F}_g$ indicates the error terms and $\boldsymbol{\lambda}_g$ is the matrix of loading coefficients, that can be estimated by OLS:

$$\hat{\boldsymbol{\lambda}}_g^T = ((\boldsymbol{\xi}_g^{HOC})^T\boldsymbol{\xi}_g^{HOC})^{-1}(\boldsymbol{\xi}_g^{HOC})^T\mathbf{PC}_g^I \tag{2.18}$$

Where PC_g^I represents the block g (i.e. as a matrix) of the first principal components scores (i.e. the MVs) by that specific HOC g; for example, considering the *Off_phase* as $\boldsymbol{\xi}_g^{HOC}$, its block g of MVs (PC_g^I) is composed by the first principal components scores of *attacking, skill, movement, power* and *mentality*. Keep in mind that, since we normalized data at the beginning, we hold the constraint to have unit variance, so the equality in (2.18) is valid. Since each MV is a simple linear regression between its correspondent LV and the estimated loading coefficient, we haven't any multicollinearity problem between the indicators of each block [133].

In order to evaluate the performance of the model, a GoF (Goodness of Fit, [133]) index is computed; this is the most synthetic performance index in the PLS-SEM context, but we must pay attention and interpret it with prudence, as generally suggested [79], since it is just a geometric mean between the inner and outer model performance and it does not tell us if the model is replicable in other context.

Until now the model has been described, implicitly assuming that the data came from a homogeneous population, but in reality, this assumption is often unrealistic. In fact, players (with their particularities) are frequently different, and analysing all the data without taking into account this heterogeneity is likely to produce misleading results and incorrect conclusions [116]. So, the innovation of this study is considering observed heterogeneity among players' roles and leagues; it is usually treated by splitting the dataset and creating different PLS-SEMs, one for each group and evaluating their differences. In order to do this, two tests have been performed:

- "A test for multigroup comparison using partial least squares path modeling" [90] proposed by Klesel, Schuberth, Henseler and Niehaves; in this approach the model-implied variance-covariance matrix (for both MVs and LVs) is compared across groups. It measures the distance between the model-implied variance-covariance matrices by the well-known squared Euclidean distance. If more than two groups are compared, it uses the average distance over all groups.
- An intuitive multi-group test has been performed to evaluate group differences in the estimates: this approach is based on the confidence intervals (CIs) constructed around the bootstrap estimates (i.e. it works by comparing two groups at a time). If the parameter of one group is covered by the CI of the other one and/or vice versa, it can be concluded that there is no group difference [114].

Finally, as an in-depth analysis to test its predictive power, a comparison with a benchmark model is provided; GSCA (Generalized Structured Component Analysis, [83]) is used for this purpose, that replaces factors by exact linear combinations of observed variables. For the work the *R* packages *csem* [99] and *seminr* [122] have been used; it was applied a bootstrap validation (i.e. 1000 resampling) for the model in order to assess the path significance. In the next section, the results are shown.

2.2.3 Results

In this section some results are shown, organised in this way: first of all there will be a focus on the estimates for the full model and its performance. Then some output considering the heterogeneity observed among leagues and roles. Finally, a paragraph concerning an in-depth analysis of the model for the midfielder is provided, followed by a global comparison of the models created.

The full model

As a starting point, a full PLS-SEM (i.e. using all 2662 players) is created; in Fig. 2.13 the parameter estimates and their statistical significance are shown, with circles representing LVs and rectangles MVs. The directions of the arrows for the outer model of LOCs (on the left) from MVs to LVs represent the formative framework whereas for the HOCs the arrows point from LVs to MVs (e.g. the I PCs) and represent the reflective model; the thickness of the arrows is proportional to the strength of their effect. Above each arrow of the outer model are located the loadings (i.e. λ) between each MV and the corresponding LV. For the inner model, β above each arrow represents the correlation (i.e. the path coefficient) among each couple of exogenous and endogenous LVs, following the factorial scheme (2.16). Asterisks next to each value represent its statistical significance (after 1000 bootstrap resampling); dotted arrows mean negative values for the corresponding parameters.

Fig. 2.13 shows that *Off_phase* has a stronger correlation on the performance than *Def_phase* (0.57 vs 0.43, both significant); it is interesting noting also how all LVs are significant (p-values $< 1\%$, ***). Concerning the inner model related to the *Off_phase*, the *attacking* abilities have the strongest path coefficient ($\beta_{attacking} = 0.3$), followed by generic *skill* ($\beta_{skill} = 0.26$); we can note also that *mentality*, *power* and *movement* have a similar value on the *Off_phase*. Regarding *Def_phase*, the strongest one is that of the generic *skill* abilities ($\beta_{skill} = 0.36$), followed by mentality ($\beta_{mentality} = 0.28$); it is curious how *defending* has the lowest path coefficient on *Def_phase*. About MVs,

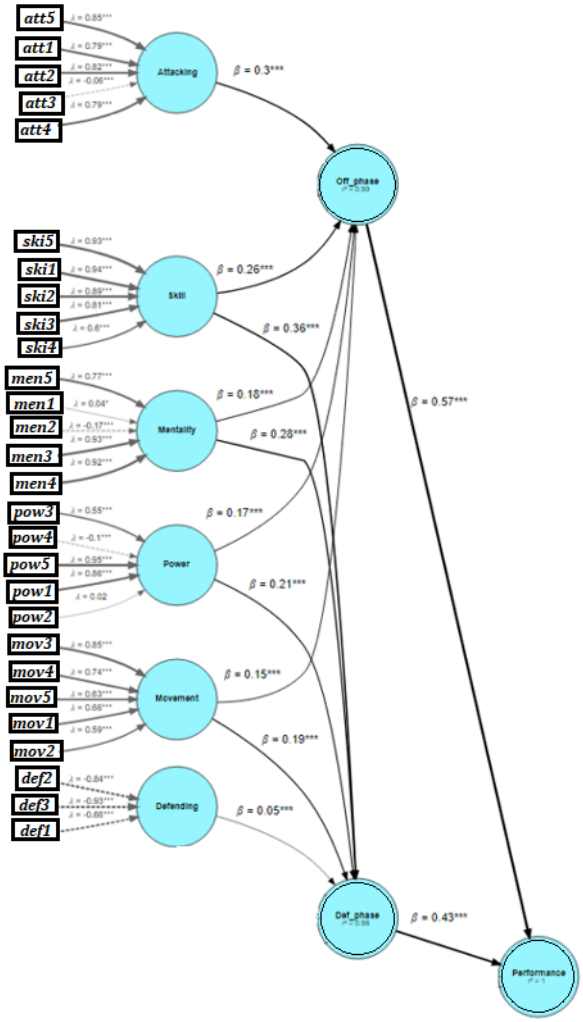


FIGURE 2.13: Output of the full model considering all 2662 players

some of them have negative loadings with the related LV, making interpretation difficult for some of them (e.g. *defending*); other ones have very low loadings (near to zero) for example "att3" related to *attacking*, "men1" for *mentality* or "pow4" for *power*.

About the performance of this model, compared to the benchmark *overall* indicator from EA Sports, this new indicator has a medium correlation (0.65, Fig. 2.14). It has also a good GoF index (0.76) after the bootstrap validation. The scatterplot in Fig. 2.14 compares EA *overall* performance versus the performance indicator of the PLS-SEM (i.e. standardized values), suggesting a pattern that depends on the role: this figure shows clear differences in the PLS-SEM indicator depending on the role, while this does not seem to be true for the EA *overall*.

In summary, the indicator obtained with this model has a medium correlation with the benchmark indicator and a good performance in terms of GoF, but some of the LVs and MVs are difficult to interpret; some of them have also a weak path (i.e. near to zero). It has been shown how there could be a pattern depending on the role, which it will be analysed in-depth in what follows.

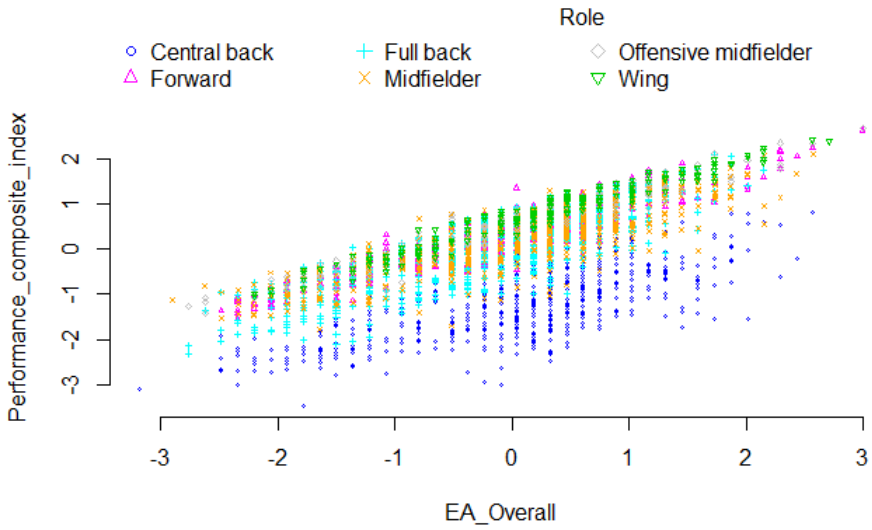
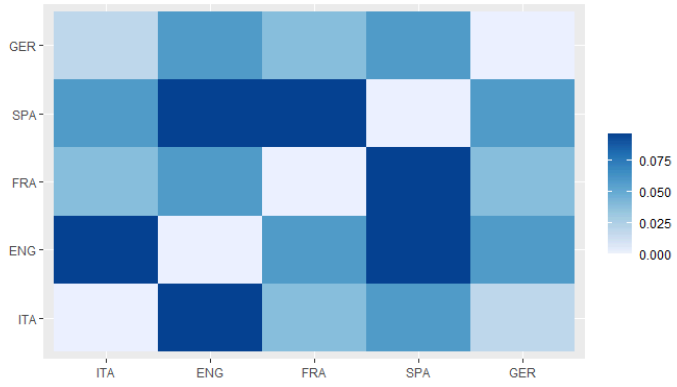


FIGURE 2.14: PLS-SEM *overall* performance indicator vs EA *overall* performance indicator by roles

Heterogeneity observed among leagues

In this paragraph, heterogeneity among the top 5 EU leagues is examined in order to see if it works as confounding factor for the model built in Sec. 2.2.3. Preliminary results of the test of Klesel et al. suggest rejecting (p-value $< 5\%$) its null hypothesis (i.e. equality across leagues) for what concern MVs, while to accept it for constructs (LVs, p-value $> 5\%$): it seems there are not significant differences between the variance-covariance matrices of the different leagues for the LVs, while there are for the MVs. Then, the output of the second test is presented using an heat-map (Fig. 2.15) with the non-overlapping rate from the bootstrap CI: this rate is the proportion between the number of non-overlapping CIs and the total number of CIs (i.e. it includes path coefficients of the inner model and weights of the outer model both, so 52 estimates in total).

FIGURE 2.15: Heatmap of non-overlapping rate from 95% bootstrap CIs by league



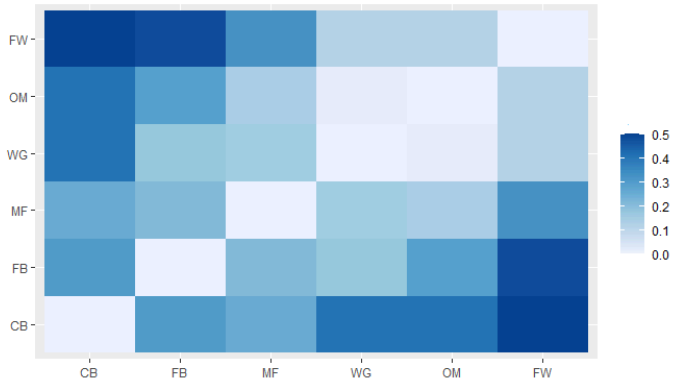
Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

In Fig. 2.15 we can see the leagues with more differences in the PLS-SEM estimates have a darker colour (i.e. dark blue), in particular ENG vs SPA, FRA vs SPA and ITA vs ENG. In every case, this does not seem to be relevant, because the maximum rate for that is 8%: this means for example that the comparison of the English Premier League and Spain's LaLiga differs just for 4 estimates out of 52. So, considering these leagues separately does not seem useful, and so it does not contribute to confound the general model.

Heterogeneity observed among roles

The other aspect about the heterogeneity observed in the dataset is due to the different roles (Sec. 2.2.1). This is crucial, since the aim is to investigate whether it is a possible confounder. The same method is used as in the previous subsection: the first multi-group test suggests rejecting (p-value $< 5\%$) its null hypothesis (i.e. equality across roles) for both MVs and LVs variance-covariance matrices. Then, an heat-map of the non-overlapping rate of bootstrap CIs among different roles defined by experts [82] is presented (Fig. 2.16). As in the previous interpretation, a darker colour (i.e. dark blue) indicates a high rate of non-overlapping (i.e. high diversity) between roles. Colours from light blue to white mean there is significant similarity between the different roles.

FIGURE 2.16: Heatmap of non-overlapping rate from 95% bootstrap CIs by role



Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

Unlike the case with the leagues, Fig. 6 shows how for roles the situation is different: the colour scale reaches a rate of 50% (dark blue) in the top left (also bottom right, since the matrix is symmetric). For instance, looking at the role of CB and the differences between its estimates and those of the other roles (the first column of the heat-map), there is a low rate of non-overlapping for CB versus FB and MF (less than 30%), while there is a medium high rate versus offensive roles (i.e. FW, OM and WG). But, if we look at the last column, concerning FW, they have a high rate of differences versus CB and FB, a medium rate with MF, while this is minimal versus OM and WG. The only role that seems to have more equilibrated estimates is, as could be expected, the MF (no dark blue rates in its own column). Summarizing, we can see very significant differences between the estimates for the roles, more so than when considering leagues. In order to have a clearer idea, a summary plot is presented in Fig. 7 with path coefficients (i.e. inner model) and their 95% bootstrap CIs.

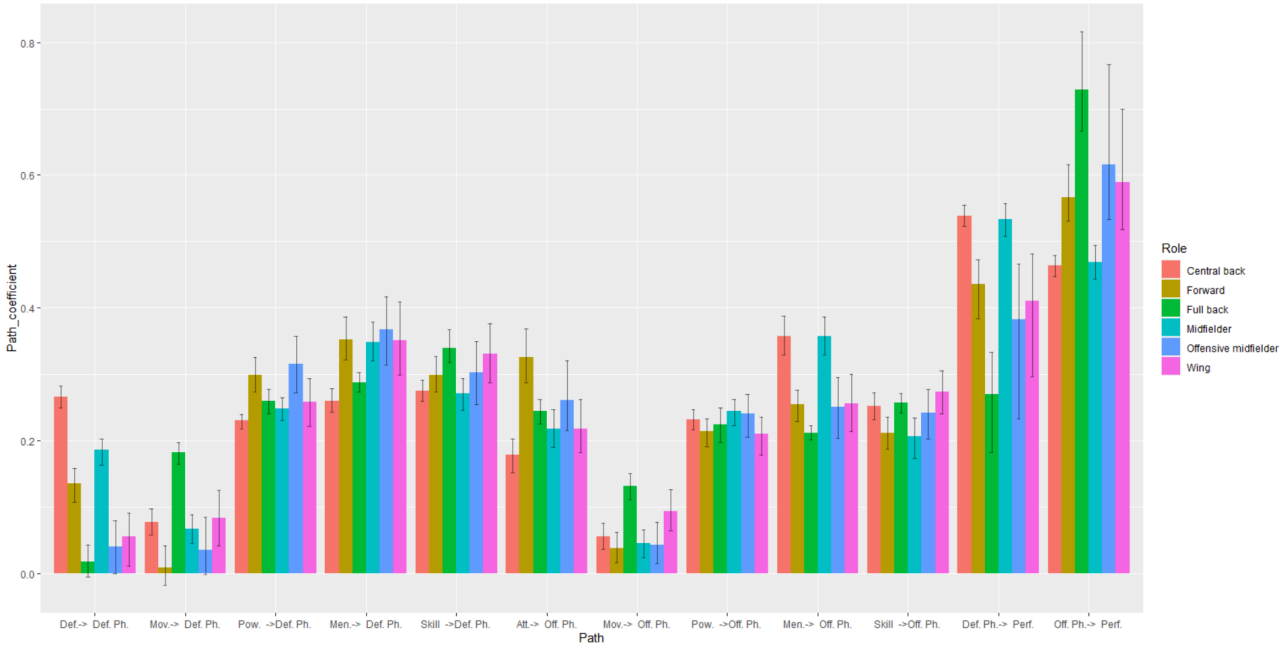


FIGURE 2.17: Estimated path coefficients by role and 95% CI after 1000 bootstrapping

It is interesting to note how *Off_phase* and *Def_phase* generally have the highest path coefficient (in absolute value) on the performance composite indicator. Basically, *movement* LV has the lowest path coefficients versus both *Off_phase* and *Def_phase*, in some cases also not statistically significant (e.g. movement versus *Def_phase* for FW). Also *defending* towards *Def_phase* has very low correlations: in this case only CB and MF have a real impact (greater than 0.15); for FB instead this coefficient is not significant. In every case, for each group of path coefficients, there are clear differences between the roles. Below, there are some specific comments for each path:

- *skill* $\rightarrow$ *Off_phase*: this is the most equilibrated path among roles, in fact it ranges from 0.2 of midfielders to 0.28 of wings.
- *skill* $\rightarrow$ *Def_phase*: we can see here how the differences are much greater than the previous one. *skill* has the highest correlation with *Def_phase* for FB (almost 0.4) while the lowest for MF.
- *power* $\rightarrow$ *Off_phase*: this has the highest value for CB, MF and OM.
- *power* $\rightarrow$ *Def_phase*: *power* is crucial for the *Def_phase* performance of FW and OM.
- *Off_phase* $\rightarrow$ performance: here we can see how the composite *Off_phase* is very high (greater than 0.45) for all roles, with a peak greater than 0.7 for FB (this is the only strange estimate, which could be investigated in a future research project, since FB is not properly an offensive role). Other roles with a value greater than 0.5 are WG, OM and FW (the typical offensive roles).
- *Def_phase* $\rightarrow$ performance: here we can see how the composite *Def_phase* is quite high (greater than 0.35) for all roles, with a peak greater than 0.5 for CB and MF (i.e. the only two where it has a significantly greater impact than the *Off_phase*).

- *movement* → *Off_phase*: athletic abilities, as said before, do not have a very high impact for any roles in the *Off_phase* except for the FB (which seems right, since they typically run along the wing of the field).
- *movement* → *Def_phase*: here, estimates are slightly higher than the previous case, and also in this situation FB has the highest correlation.
- *mentality* → *Off_phase*: this set of abilities is important in the *Off_phase* for all roles, with a peak for MF and CB (path estimates greater than 0.3).
- *mentality* → *Def_phase*: for the *Def_phase*, *mentality* abilities are important for almost all roles (path greater than 0.3), except for FB and CB, with a slightly lower path (still greater than 0.2).
- *defending* → *Def_phase*: strictly *defending* abilities are important only for MF and CB, as expected.
- *attacking* → *Off_phase*: also in this case, strictly *attacking* abilities have the highest impact for FW on the *Off_phase* (greater than 0.3).

Since the more equilibrated values (the role with less differences in path coefficients from the other ones) belong to the MF model, an in-depth analysis and a more accurate validation of this one is presented in the following subsection.

In-depth analysis: the midfielders' model

The midfielder's PLS-SEM takes into account just 621 players from the full dataset. After the bootstrap validation, its output is presented in Fig. 2.18 (note that it has the same logic for its interpretation as Fig. 2.13).

Unlike the initial full model, it is immediately clear that all path coefficients (i.e. correlations, for the inner model) and loadings (for the outer model) are significant and concordant. The midfielder has a higher weight for the *Def_phase* (0.53) than the *Off_phase* (0.47) on the global performance. For both *Def_phase* and *Off_phase*, *mentality* is the LV with highest impact, while *movement* is the lowest.

Compared to the benchmark *overall* indicator from EA Sports, this new specific indicator improves its correlation a lot (0.93 versus the 0.65 of the full model). It also has a good GoF index (0.76) after the bootstrap validation. In order to evaluate the predictive power of this model, a 5-fold cross-validation (i.e. the default value) was performed and in Fig. 2.19 we can see its output for each PC (i.e. principal component, that works as MV) of the composite performance indicator. All predictions follow a strong linear relationship with the actual values, except for *defcomp1* and *offcomp3* (i.e. the first PCs of *movement*): they have a poor correlation between the actual and predicted values, confirmed from their high prediction error indices (Tab. 2.3); furthermore, recall how in Fig. 2.18 *movement*, despite its significance, has a very poor impact on *Off_phase* and *Def_phase* (i.e. path coefficients less than 0.1). Despite our goal was to create a measuring model and not a predictive one, for completeness the power of prediction was compared with a benchmark model, such as GSCA [83], and in Tab. 2.3 we can see the MAE and RMSE (i.e. respectively mean absolute error and root mean squared error, defined from the difference between predicted and actual values, [99]) for each MV of the endogenous (such as *Off_phase*, *Def_phase* and performance) in these two frameworks.

In Tab. 2.3, in bold are the PCs whose MAE (RMSE) index is lower in the target (PLS-SEM) than the benchmark: following the literature [71, 121], since there is an

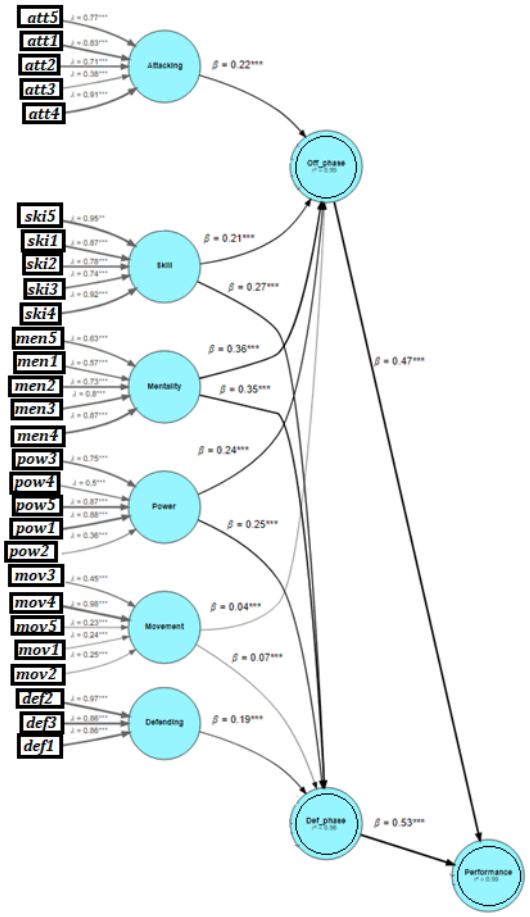


FIGURE 2.18: Output of the midfielder PLS-SEM

equal proportion of PCs (MVs) whose MAE (RMSE) is lower in the target model (6 on 12), it can being reasonable to affirm a medium predictive power for PLS-SEM in the MF player's role.

Comparison of the models

As the final result of our analysis, a comparison among the different models per role was done, by considering the GoF index and the correlation with the *overall* indicator from EA Sports: the results are shown in Tab. 2.4.

All PLS-SEMs by player's role have a GoF greater than or equal to that of the full model, the highest is that of WG (0.79). Hence, every model has a good value of the GoF. Concerning instead the correlation with the well-known *overall* indicator from EA Sports, all roles have a significant higher value (from 0.86 for CB to 0.97 for WG and OM) than considering the full model (0.65): this is crucial, since it supports the confounding function of players' role. In this case, heterogeneity observed among roles let us to improve in a relevant way models' performance result.

2.3 Chapter conclusion

Summarising, in the first part of the chapter the PLS-SEM features have been described. Then, an in-depth analysis about PLS-SEM inner and outer model has been provided, followed by some measurement and structural model assessments. In

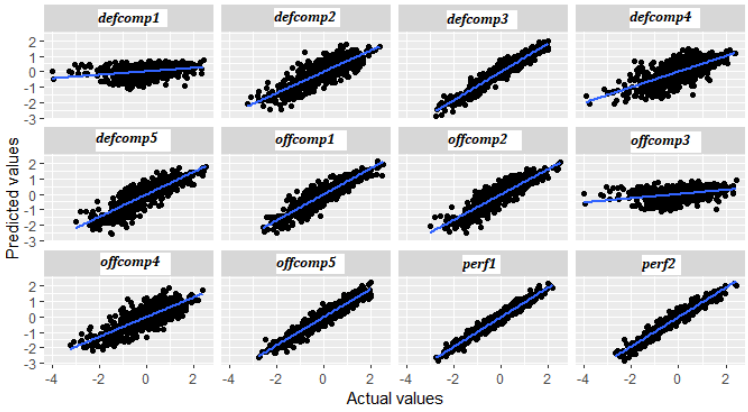


FIGURE 2.19: Midfielder PLS-SEM: actual vs predicted (standardized) values for PCs of performance

TABLE 2.3: PLS-SEM vs GSCA for MF player's role

MV	MAE (PLS-SEM)	MAE (GSCA)	RMSE (PLS-SEM)	RMSE (GSCA)
offcomp1	0.3074	0.2553	0.3899	0.3277
offcomp2	0.3352	0.2950	0.4182	0.3686
offcomp3	0.7065	0.6681	0.9388	0.8828
offcomp4	0.4789	0.4935	0.6000	0.6177
offcomp5	0.2277	0.3199	0.2863	0.4073
defcomp1	0.7153	0.6864	0.9509	0.9098
defcomp2	0.4401	0.3944	0.5542	0.4976
defcomp3	0.2153	0.3314	0.2722	0.4259
defcomp4	0.5432	0.5578	0.6856	0.7045
defcomp5	0.4097	0.4276	0.5133	0.5311
perf1	0.1607	0.3130	0.2030	0.3969
perf2	0.1729	0.1022	0.2229	0.1285

following, the so called "extension" phases of a PLS-SEM analysis have been presented, for example the different approaches in a higher-order model, in order to obtain a higher level of abstraction; in addition, evaluating the importance to take into account moderating effects and heterogeneity in the data, in order to avoid misinterpreting in the final results. Starting from that preface, in the second part of the chapter an exploratory application has been developed, with the objective to measure football players performance.

The preliminary point of this application were the EA Sports experts, which are the ultimate authority on soccer performance measurement: they constantly maintain a database of realistic players' performance attributes resulting from careful and systematic data collection. According to the experts, performance variables make up several broader, theoretical dimensions. The initial part of this research was dedicated to a Partial Least Squares Structural Equation Modeling (PLS-SEM) considering all the players: it had a bit of incoherence (e.g. difficulty of interpretation, negative loadings and weights) in some latent variables (LV) and a medium correlation with the Electronic Arts (EA) *overall*; then, considering the observed heterogeneity as a possible confounding cause, the leagues and different roles have been taken into

TABLE 2.4: PLS-SEM comparison by player's role

Role	n	GoF	Corr. with EA *overall*
CB	524	0.77	0.86
FB	732	0.76	0.90
MF	621	0.76	0.93
OM	176	0.78	0.97
WG	201	0.79	0.97
FW	408	0.78	0.95
Full	2662	0.76	0.65

Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

account: the model was replicated for each different league and it revealed only very low differences (e.g. maximum non-overlapping rate of 8% between the English Premier League and Spanish LaLiga) in the estimates for both inner and outer models; furthermore, the test of Klesel et al. accepted the hypothesis of equality among leagues for the variance-covariance matrices of the LVs: hence, the league does not seem to be a confounding factor. Considering players' roles, the output changes: here, differences in the path estimates are evident across all roles (non-overlapping rate maximum at 50%), especially between opposite positions (e.g. central back vs forward or full back vs forward), which is also confirmed from the test of Klesel et al. (p-value $< 5\%$); furthermore, considering these models by roles, it's right to underline both their good goodness of fit (GoF) and a considerable increase in the correlation with the benchmark index (i.e. *overall*). Despite was no the main goal of this application, also an in-depth analysis of the midfielder (MF) model was presented, evaluating its predictive power compared with the benchmark (GSCA): it showed that the PLS-SEM in this framework had a medium predictive power.

Considering the path coefficients by role (Fig. 2.17) and their bootstrap validation, the results are quite sensible also from a logical point of view; only for FB (a typically more defensive role) it was noticed a too high value for *Off_phase* on performance (more than 0.7), despite a very low estimate for *Def_phase* (0.3).Maybe this could be due to the influence of some offensives full backs, but in any case for future research it should be interesting to carry out a specific confirming composite analysis (CCA, [42, 119, 72]), in order to improve and further validate the model: for example, by tacking in consideration possible collinearity problems, or making a predictive validity, for example evaluating better and improve the predictive power of *defcomp1* and *offcomp3* for the midfielder PLS-SEM, or considering the procedure of the Weight PLS (WPLS) algorithm [37] when using the PLS technique for assessment. Other interesting issues for future projects could be to take into consideration unobserved heterogeneity among players, maybe using a PLS-POS algorithm [4], or using this indicator as starting point to improve other existing football analytics models (e.g. the expected goal model).

Overall, reminding that the core of this application was to create a players' performance measuring model and not a predictive one, the presented research tried to create an innovative model for players' performance and at the same time build two other interesting indicators, *Off_phase* and *Def_phase*, employing a Third-Order PLS-SEM approach, with the objective of helping football policy makers in an impartial

evaluation of their players, specific for each role. Another advantage derived from this PLS-SEM approach is the possibility to split performance into its sub-areas, in order to study in-depth the players' condition. In the next chapter you could see the forward steps of this application.

Chapter 3

PLS-SEM insight: CTA and CCA

In this chapter we will make an in-depth analysis about two important issue that emerged in the Chapter 2: the choice about reflective-formative constructs and a more solid method to confirm and validate the model. It has been thought to dedicate a chapter due the relevance of these topics for PLS-SEM and because we have dedicated one application them; in fact the chapter is organised as follow: we will talk about the debate among reflective-formative constructs and the solution proposed (i.e. the CTA analysis) in Sec. 3.1, whereas in Sec. 3.2 we will talk about the confirmatory composite analysis (CCA). Finally, the extension of our application proposed in the previous chapter is presented in Sec. 3.3, followed by a conclusion of the chapter in Sec. 3.4.

3.1 Reflective vs Formative constructs: the CTA analysis

As seen in the previous chapter, PLS-SEM for its own measurement model allows two types of constructs, respectively reflective and formative: the first one implies that the LV exists independently from the measures used (i.e. causality from construct to items), whereas the second is determined as a combination of its own indicators (i.e. causality from items to construct) [15, 14]. Although PLS-SEM represents a good method for estimating complex cause–effect-relationship models for many scopes, as already introduced in this thesis there is a lack in some part of its theory: in particular, there are just few works relying the assessment of using formative measurement models. While for reflective constructs exist several tests to assess their reliability, for what concern formative constructs researchers are just basing on theory and experts opinion, since the classical procedures to assess reflective constructs (e.g. confirming factor analysis and internal consistency evaluation) are not suitable for the formatives [55]. These problems have caused by a measurement model misspecification: as consequence, this can lead a bias in the inner model estimation and lead to incorrect assessments of relationships in PLS-SEM [64].

In order to overstep that limits, some researchers stressed this crucial point: in particular some of them have applied the confirmatory tetrad analysis (CTA) by Bollen [10] for drawing conclusions about the appropriateness of using formative measurement models as compared to reflective ones [64], that we will face in Sec. 3.1.1.

3.1.1 In depth-analysis: the CTA-PLS

The CTA-PLS module is a tool developed for the smartPLS software [108] that facilitates the evaluation of cause–effect relationships for latent variables and their specification of indicators in measurement models. Before proceeding we must specify what we mean with "tetrad": a tetrad τ is the difference between the product of two

pairs of covariances. For instance, the six covariances of a block of four MVs permit the formation of three tetrads:

$$\begin{aligned} \tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24} \\ \tau_{1342} &= \sigma_{13}\sigma_{42} - \sigma_{14}\sigma_{32} \\ \tau_{1423} &= \sigma_{14}\sigma_{23} - \sigma_{12}\sigma_{43} \end{aligned} \tag{3.1}$$

Take in consideration that, while the construction of tetrads like in (3.1) requires four indicator variables at a time, CTA is also applicable to measurement models of more or less than four indicators [13]. At the beginning, the CTA was introduced in CBSEM applications ([12], CTA-SEM): these authors propose the concept of vanishing tetrads using a covariance (correlation) data matrix to complement standard procedures of model evaluation, and provide methods for selecting model-implied non-redundant vanishing tetrads and significance testing. In particular, the concept of vanishing tetrad is important also for the successive CTA-PLS, so we have to make clear this point: a vanishing tetrad equals zero and all model-implied non-redundant tetrads vanish in reflective measurement models. In this sense, researchers developed the following hypothesis test:

$$\begin{aligned} H_0 &: \tau = 0 \\ H_1 &: \tau \neq 0 \end{aligned} \tag{3.2}$$

Following the hypothesis test in (3.2), if the test supports H_0 (i.e. non-significant test, $pvalue > 0.05$), it involves vanishing tetrads, implying a reflective measurement model; otherwise, if the test is significant, it suggests us a formative measurement model.

CTA-PLS is born following the confirmatory approach of testing model-implied vanishing tetrads and the application of CTA [12] to help distinguish between formative and reflective measurement models in PLS path modeling. Although CTA-PLS uses a similar evaluation process, the approach differs from CTA-SEM for PLS methodological assumptions in both the single tetrad testing approach and the simultaneous tetrad testing procedure [64], infact:

- CTA-PLS builds on the statistical test for every single model-implied vanishing tetrad. To overcome the limitations regarding distributional assumptions, it includes a bootstrapping routine [11].
- Results for the single non-redundant vanishing tetrad significance tests provide a basis for choosing whether a reflective approach does not conform to the empirical data. A rejection of the reflective mode provides support for a formative indicator specification.

We must take in mind that neither CTA-SEM nor CTA-PLS are applicable for correlations (covariances) close to zero in the measurement model, so in the original correlation (covariance) matrix of MVs [12].

Since in CTA-PLS all tests are made for all tetrads in each block of MVs, a multiple testing problem is involved: in order to deal with this issue, a Bonferroni adjustment of the significance levels is used. It assures that the error rate does not exceed the level α for all the n desired tests; the non-parametric nature of the Bonferroni approach meets the PLS assumptions. In particular, the resultant test statistic, asymptotically approaches a χ^2 distribution with the degrees of freedom equal to the number of tested tetrads. So, Bonferroni approach lets to compute simultaneous confidence intervals for multiple tetrad tests: when a confidence interval for

a difference does not include zero, the H_0 of (3.2) is rejected. Testing with confidence intervals has the advantage that they give more information by indicating the direction and something about the magnitude of the difference; in addition, if the hypothesis is not rejected, the power of the procedure can be measured by the width of the interval [64].

Although some differences we observed, the practical application of CTA-PLS is similar to the one of CTA-SEM [12]. Now, in order to have a clearer idea than before, the CTA-PLS routine is presented:

1. First, it forms and computes all vanishing tetrads for the measurement model of a given latent variable. In general:

$$C_{n,4} = \frac{n!}{(n-4)! * 4!} \tag{3.3}$$

   By (3.3) are computed the number of sets of four variables, each resulting in three vanishing tetrads for measurement models with n MVs for each block (3.1); so, the total number of tetrads for each one is:

$$\#\tau = 3 * C_{n,4} \tag{3.4}$$

   Despite CTA-SEM, it is interesting to highlight that a LV with less than four MVs require an inclusion of indicators from another one to form a set of four MVs to perform CTA-PLS.

2. In the second step, CTA-PLS identifies model-implied vanishing tetrads; however, we must just take in mind that the expectations for formative indicator specifications is different: none of the tetrads shall vanish in formative measurement models.

3. In order to improve the successful steps, at this point CTA-PLS eliminates redundant model-implied vanishing tetrads: it happens when the same pair of covariances appears in two model-implied vanishing tetrads.

4. At this point, it performs a statistical significance test for each vanishing tetrad, determining whether the value is significantly different from zero. CTA-PLS follows, as introduced before, some Bollen's suggestions, like using a bootstrap routine [11]. Infact, by generating a high number of bootstrap subsamples (e.g. 5000) and computing their relevant tetrads allows to obtain the bootstrap estimated standard error (SE) for every tetrad, and the t-value of the Student's t distribution (e.g. $t_{obs} = \tau / SE(\tau)$). The structure of the hypothesis test was introduced in (3.2). Bootstrapping provides a non-parametric alternative but we must account for the problem that the generation of the data follows the alternative hypothesis (H_1). An examination of the statistical correspondence between tests of significance and confidence intervals when the null hypothesis concerns a particular parameter value allows addressing this problem: specifically, bootstrapping confidence intervals is an appropriate approach for this purpose [64]. In addition, the bias correction of the bootstrap confidence interval provides an appropriate means to test the model-implied non-redundant vanishing tetrads in CTA-PLS, as specified in the following:

$$t_{obs} - b_B \pm {v_B}^{1/2} z_{1-\alpha/2} \tag{3.5}$$

where b_B is the bootstrap estimates of the bias and v_B is the bootstrap estimates of the variance. Only if the corresponding (i.e. $1 - \alpha$, two-tailed) confidence interval includes (does not include) the parameter (i.e. the zero), an acceptance (rejection) of H_0 is established.

5. In the last step, CTA-PLS evaluates the results for all model-implied non redundant vanishing tetrads per measurement model by accounting for multiple testing issues. It assess the conformity of a reflective indicator with the empirical data. A reflective measurement model does not meet the empirical data if at least one of the model-implied vanishing tetrads is significantly different from zero. CTA-PLS employs a procedure for testing n hypotheses $Hp_1, Hp_2 \ldots, Hp_n$ with test statistics $T_1, T_2 \ldots, T_n$ for each LV (i.e. measurement model). Thus, in order to take into accounting for multiple testing issues, the probability of rejecting the null hypothesis requires adjustment [64]. The correction used by CTA-PLS for this problem is the Bonferroni adjustment [11], which consists of rejecting H_i, for any $i = 1, ..., n$, if the associated statistic test T_i is significant at the $\alpha' = \alpha/n$ adjusted level of the test, where n is the number of hypotheses to be tested. For instance, if we have 5 vanishing tetrads to test, and $\alpha = 0.05$, then for each individual test we should use in (3.5) a critical value α' equal to $0.05/5 = 0.01$. The procedure finishes with a sensitivity analysis, a reliable foundation for evaluating the analytical results [12].

As summary, we can say that CTA-PLS is an important tool for testing theoretical concepts and for an empirical evaluation of the mode (i.e. reflective or formative) of the measurement models; when we apply it, there are three practical steps to follow:

- A theoretical a-priori specification for each outer model (i.e. for each LV), generally defined by experts.
- The integrative CTA-PLS evaluation of measurement models, in order to make consistent CTA-PLS with PLS-SEM assumptions.
- The practical application by a software (for example, smartPLS) that includes the CTA-PLS module.

In any case, it should be better also a posterior re-examination of the constructs, in order to assess possible misspecifications of the measurement models. Infact, researchers recommand a-priori theoretical specification and posterior re-examination along with empirical data, that are essential to better understand the structure of the outer models [64]. In particular, from a practical point of view, after setting all the measurement models in a reflective way, then a CTA-PLS can be applied in order to understand what LVs are confirmed as reflective constructs, and what are disconfirmed, moving over a formative one [126]; then, after that it can be applied the PLS-SEM using the type of constructs provided by the CTA-PLS.

Concluding, the CTA-PLS allows to evaluate the direction mode for each measurement model with respect to empirical data but switching the mode of outer models (e.g., from reflective to formative) without further consideration, however, it does not represent the final result of this analysis, unless additional supporting theoretical or conceptual reasoning provides clarification.

3.2 The CCA analysis

Here is presented a guideline to the more recent confirming composite analysis (CCA, [119, 72]). This procedure is a systematic methodological process for confirming measurement models in PLS-SEM; respect to the well known exploratory factor analysis (EFA) and the confirmatory factor analysis (CFA), CCA includes a series of steps to confirm both reflective and formative measurement models (Tab. 3.1). CCA is the most suitable and complete tool to confirm and validate a PLS-SEM model despite some years ago the most used approach by researchers was the following: first, a EFA analysis was recommanded, in order to examine the underlying structure of multi-item scales, following by a CFA [72].

TABLE 3.1: A brief comparison between EFA, CCA and CFA

Approach	**EFA**	**CCA**	**CFA**
Total Variance	X	X	-
Common Variance	X	-	X
Exploratory	X	X	-
Confirmatory	-	X	X

Early applications of EFA were based on the common factor model, which variables are assumed to be a function respectively of common, specific and error variance [17, 67]: common variance is the indicator's variance that is shared with all other indicators in the analysis, specific variance is the indicator's variance that is only associated with the indicator, while error variance is the indicator's variance that is due to unreliability, bias, or randomness. EFA assumes that any indicator may be associated with any factor. However, EFA only needs to be applied when there is no established theory describing the underlying constructs for a set of MVs [72].

For what concern CFA, it was introduced as assessing measurement quality for CBSEM models, as confirmatory factor analysis; CFA is both a qualitative and statistical process suitable to evaluate some relevant measures, such as construct reliability, discriminant validity or goodness of fit. In summary, by applying CFA on CBSEM, a researcher is testing the hypothesis that a proposed theoretical relationship exists between the observed variables (MVs) and their underlying latent constructs. The final objective is to confirm the measurement properties of a set of MVs for measuring a specified latent construct. In the early year, also for PLS-SEM was applied the CFA analysis to measurement model confirmation, but, since there are slightly differences between CBSEM and PLS-SEM (i.e. as already underlined), some researchers proposed a new technique, more suitable for PLS-SEM: as consequence, it is born the concept of CCA as a process for confirming measurement models in PLS-SEM [80].

So, from a practical point of view, the two main possible approaches are the principal axis analysis, which extracts factors using only common (shared) variance (EFA and CFA, Tab. 3.1), whereas the other one is the principal component analysis (PCA), which extracts factors using total variance (EFA and CCA, Tab. 3.1).

The three methods have similarities but many differences. For instance, the statistical objective of EFA is data reduction through exploration of response patterns, while for CCA and CFA is confirmation of measurement theory. Again, EFA often ends with the identification of factors, while CCA and CFA begin with proposing

theoretical constructs to be confirmed, and almost always moves on to structural modeling after the composite confirmation [72].

In order to validate the measurement model, CCA drives us with some steps, depending if we are working with reflective or formative indicators. By [72], we try to guide you by the fundamental steps to evaluate and validate the outer model, in these two cases:

- **Reflective measurement** $\Rightarrow$ as seen before, in this case indicators are seen as a manifestation of the empirical surrogates (proxy variables) for the latent variable. In order to assess them, CCA suggest us these following steps:

  1. **MVs preliminary check**: we must assess the indicator loadings (i.e. λ, the root squared of the communalities) and their significance; more precisely, loadings must be equal or greater than 0.7. Take in consideration that loadings in the range of $0.4 < \lambda < 0.7$ are acceptable if composite reliability and the AVE index meet their thresholds [126]. In order to test λ significance, since does not exist a closed form solution for the confidence intervals, we suggest a bootstrapping procedure to obtain t-statistics ([68]). Indicator loadings with confidence intervals excluding zero are statistically significant.
  2. **Indicator reliability**: it can be measured by the square of each indicator loading with its own LV; this index provides a measure of the amount of variance shared between the MV and its associated construct [9].
  3. **Composite reliability**: at this point, it is very important to check the unidimensionality (i.e. the coherence) of each block of MVs. We can do it in different ways:

     PCA: making a principal component analysis (PCA) on each block and checking if we have just one eigenvalue greater than 1.

     Cronbach's alpha: the rule of thumb of this criteria is that it must be equal or greater than 0.7.

     Composite reliability: also this other index must be equal or greater than 0.7. Because indicators are not equally reliable, composite reliability, which is weighted, it is more accurate then Cronbach's alpha statistic, that is unweighted ([69]).

     We have to make attention if a reliability index is 0.95 or higher, in this case the individual items are measuring the same concept, and are therefore redundant. If a block is not uni-dimensional we could proceed in different ways: we could remove the correspondent MVs, otherwise we could split the multidimensional block in different unidimensional underblocks. Eventually, as extreme solution, we could adopt instead a formative approach.
  4. **Convergent validity**: it can be measured by the average variance extracted index (AVE), obtained by averaging the indicator reliabilities of a construct. It quantifies the average variance shared between the LV and its MVs. AVE should be 0.5 or higher.
  5. **Discriminant validity**: this step has the goal to ensure that a reflective construct has the strongest relationships with its own indicators. It can be assessed using cross-loadings, Fornell-Lacker criterion [61] or the heterotrait-monotrait ratio of correlations (HTMT, [77, 70]). The first procedure provides that the outer loading of an item should be greater on its respective

latent variable than its cross-loadings on other latent variables. According to Forner–Lacker criterion, the square root of AVE, of each of the latent variable, should be greater than its correlation with other latent variables. Finally, the HTMT ratio is the average of the heterotrait-heteromethod correlations (i.e, the correlations of indicators across constructs measuring different phenomena) over the average of the monotrait-heteromethod correlations (i.e. the correlations of indicators within the same construct). In particular, if $HTMT < 0.9$ [127, 42] or its own bootstrap confidence interval does not contain 1, then discriminant validity has been established between two reflective constructs [77].

6. **Nomological validity**: it is an additional method to assess construct validity; in practise, it works correlating the LV score of each construct with one or more other concepts (LVs) in the nomological network. This framework is a sort of representation of the interrelationships between concepts, for example in psychological test [50].
7. **Predictive validity**: it evaluates how well a construct score predicts on some outcomes (i.e. criterion measures) and it is measured by correlation. In particular, predictive validity involves using the construct score to predict the score of a criterion variable that is collected at a later point in time. Nevertheless there is not a unique tested method of predictive validity, construct invariance in PLS-SEM measurement models can be tested by applying the MICOM procedure [78, 123]. For instance, construct invariance is most often applied with cross-cultural studies and it can be assessed with MICOM.

- **Formative measurement** $\Rightarrow$ in this case, LVs are formed (i.e. caused) from a set of MVs. Formative indicators cannot be evaluated at the same manner of the reflective ones and so they must be interpreted in a different way [38]. It will be presented in the following what CCA recommands:

  1. **Convergent validity**: as written by [70], it is the extent to which the formative construct is positively correlated with a reflective measure(s) of the same construct using different MVs. The relationship between multi-item formative LV and the reflective measure of the same construct is usually examined using correlation. Convergent validity can be measured on the size of path coefficient between two constructs; in particular, [70] recommend a minimum path coefficient of 0.70, and in general, the larger its size, the stronger is its indication of convergent validity. In alternative to this criterion, we could check if a revision of the theoretical formative LV is possible by removing, revising, adding one or more MVs. All this process of convergent validity is also referred to as redundancy analysis from [38].
  2. **Indicator multicollinearity**: we remind that for formative LVs, high correlation between indicators creates problems of multicollinearity. In order to see if this problem is present we can check the VIF (i.e. variance inflation factor) index: if its value is equal or lower than 5 (i.e. use 3 as threshold for a more conservative approach, [72]) multicollinearity is not a problem. When multicollinearity holds, we can evaluate whether one or more MVs can be removed. But we must pay attention in this case, in fact a formative indicator should never be eliminated based solely on statistical

criteria. Another option to solve this problem is to develop higher-order constructs that are supported by measurement theory [3].

3. **Size and significance of indicator weights**: if the previous steps indicate the formative measurement model meets recommended guidelines, now we must examine the size and significance of the indicator weights. The amount of contribution (relevance) of the MVs is based on the size of the outer model weights; larger weights indicates a higher contribution. Since PLS-SEM is a non-parametric statistical tool, significance of the outer weights is determined using bootstrapping. In general, the level of statistical significance is $\alpha = 0.05$. If an outer weight relying a MV is non-significant (i.e. $pvalue > 0.05$) then it must be checked its outer loading: if it is statistically significant then the MV can be retained, otherwise it can be removed from the outer model [126].
4. **Contribution of indicators**: now, in order to evaluate this point, we can take in consideration size and significance of the loadings. In particular, one loading is considered relevant in forming the LV when it is greater or equal to 0.5 and statistically significant. If it does not hold, we can remove or retain the LV based on a theoretical assessment of its relevance obtained from experts' knowledge.
5. **Assess predictive validity**: this step assesses the extent to which a construct score predicts scores on some outcomes measure. Note that predictive validity involves using the LV score to predict the score of an outcome LV that is collected at a later point in time, like for the reflective constructs.

Structural model assess $\Rightarrow$ at this point, after evaluated our PLS-SEM measurement model by the CCA approach, we are ready to assess the structural model, by following these relevant steps [72]:

1. **Evaluate structural model collinearity**: like for the formative constructs, also for the inner model we must evaluate the presence of possible multicollinearity problems. Also in this case VIF values can be examined and if they are under the threshold of 5 then multicollinearity is not a problem (i.e. with 3 more conservative threshold, [126]).

2. **Size and Significance of Path Coefficients**: after verified the first step, now we must check the size and significance of the path coefficients; they are standardized values that may range between $+1$ and -1: the closer they are to 0 , the weaker they are in predicting endogenous LVs. Opposite, the nearer they are to $|1|$ the stronger they are in prediction. In order to validate our path coefficients, there is a bootstrap validation (i.e. they are significance if their confidence interval does not contain zero).

3. **Model goodness (in-sample prediction)**: since our inner model is a substantially a multiple regression model, the most often used metric for in-sample prediction for all endogenous LVs is R^2. We must pay attention with this metric, because in multiple regression R^2 is proportional with the number of independent variables, so we risk to overfit the performance of our model. For this reason, some researchers also examine the adjusted R^2, useful when we have too many nonsignificant predictor constructs in the structural model [74]. Adjusted R^2 improves the classical R^2 value downward based on the sample size and the number of predictive LVs.

4. **Effect size (in-sample prediction)**: this metric measures the in-sample predictive ability of each independent LV. For example, effect size (i.e. f^2) for a given predictor is computed as the difference between two R^2 index: the first is the R^2 computed in a new model without that predictor, while the second R^2 is computed for the full model. We can interpret its value as follow [46]:

   - $0.02 \leq f^2 \leq 0.15$: small effects size.
   - $0.15 \leq f^2 \leq 0.35$: medium effects size.
   - $f^2 \geq 0.35$: large effects size.

5. **Predictive relevance (primarily in-sample prediction)**: in this step we can compute the third assessment metric, the Q^2 value, also referred to as blindfolding [62, 125]. We can interpret its value in this manner:

   - $Q^2 \ll 0$: not meaningful value (i.e. lack of predictive relevance).
   - $Q^2 \gg 0$: meaningful value.

   And in particular:

   - $0.25 \leq Q^2 \leq 0.5$: medium predictive relevance.
   - $Q^2 \geq 0.5$: large predictive relevance.

6. **PLS predict (out-of-sample prediction)**: this last step is maybe the most important of our work, since it let us to validate our inner model. In fact all metrics seen until now are useful in evaluating the predictive power of a model based on in-sample [118]. The main weakness of in-sample prediction is that it uses the same sample as training and test dataset, with very high risk of overfitting (i.e. our model fits too well for the current dataset, but if we try to test it with other data we obtain very poor prediction). The out-of-sample prediction for PLS-SEM was recently proposed [122]: it consists on estimating our model on a training sample, then using results of that model to predict other data in a separate test sample. Then, this PLSpredict function [121] has been implemented on the well-known smartPLS software [108] and also on some open source tools such as *R*. PLSpredict package lets us to choose some options for validating the model, in particular we can select different proportions for training and test set if we want to divide our whole dataset in two part, or we can apply a k-fold cross validation, selecting the number k of folds. In general, the recommended minimum size for the holdout sample would be $n = 30$ ([72]) and in case $n < 30$ then we should interpret results cautiously. For the assessment of our model, in out-of-sample prediction we can use some well-known statistics:

   - **MAE (mean absolute error)**: it evaluates the average magnitude of the errors in a set of predictions without considering their direction. It is preferred to refer this index if the prediction errors distribution is unbalanced [121].
   - **RMSE (root mean squared error)**: it measures the square root of the average of the squared differences between predictions and the actual values.

   Furthermore, in order to have more robust results, it could be useful to compare RMSE and MAE to a naive values obtained by a benchmark (for example, a linear regression model, i.e. LM). A LM predicts each of the endogenous LV

from all indicators of the exogenous ones. But we must take in consideration that LM does not include the PLS-SEM model structure [53]. Comparing MAE and RMSE statistics between PLS-SEM and our benchmark, we could interpret results as follow [71, 121]:

- When the RMSE or MAE index is higher for all PLS-SEM outcomes compared to the benchmark ⇒ **lacks predictive power**.
- When the RMSE or MAE index is higher for the majority of our PLS-SEM outcomes compared to the benchmark ⇒ **low predictive power**.
- When the RMSE or MAE index is higher for an equal or minority proportion of our PLS-SEM outcomes compared to the benchmark ⇒ **medium predictive power**.
- When the RMSE or MAE index is higher for none of our PLS-SEM outcomes compared to the benchmark ⇒ **high predictive power**.

Summarizing, PLS-SEM approach can perform confirmatory and assessment analysis of reflective and formative composite structures via CCA, which broadens the applicability of both PLS-SEM and CCA. In addition, CCA guidelines can be applied for the higher-order constructs [42, 117]. Despite EFA and CTA, CCA introduced the evaluation and confirmation of formative constructs, and it is the best choice when prediction is the statistical objective of the research (i.e. variance extracted from exogenous LVs in CCA is focused on the prediction of endogenous ones) [73]. Some applications on empirical data have been recently developed in order to apply a CCA on different PLS-SEM frameworks, also considering hierarchical constructs, showing the reliability of this procedure [42].

3.3 The application

This application aims to use a computationally intensive nonparametric approach with an original application in the sport field, creating a new composite indicator by role (the PI *overall*), starting from a Third-Order PLS-SEM measurement model presented in the previous chapter, taking into consideration both statistical evidence and expert opinion. As recall, PLS-SEM is based on bootstrap sample replications: infact bootstrap confidence intervals provide additional information on the stability of coefficient estimates and in the following we will use the Bias-Corrected and accelerated (BCa) method [75] for constructing them. By this study, in order to take into consideration the formative or reflective nature of each first order latent variable, a Confirmatory Tetrad Analysis (CTA) will be adopted, whereas a formative structure was assumed for higher order constructs; finally, in order to improve the quality of the model a more recent Confirmatory Composite Analysis (CCA) will be applied, including a predictive validity evaluation with the benchmark indicator EA *overall* and some performance quality proxies, such as the player's market value and wage. Data and model description are presented in Sec. 3.3.1, whereas final results are given in Sec. 3.3.2, 3.3.3 and 3.3.4. An in-depth analysis relying goalkeepers is proposed in Sec. 3.3.5.

3.3.1 Data and application design

The dataset used in this study is maintained by Electronic Arts (EA), a video game company (FIFA series) and available online[1]. Software developers are constantly

[1] `www.sofifa.com`

maintaining a database with performance measures of soccer players and the process of performance measuring can be described as a professionally conducted survey of an interested audience. The dataset is a free available online platform (Kaggle[2]), with data concerning the beginning (September) of the season 2021/2022; the focus has been on all players' stats from the top 5 European Leagues (Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). This Kaggle dataset contains 29 player's abilities (KPIs), with player's performance on a 0-100 scale with respect to different abilities, classified by EA Sports experts into 6 latent traits: *attacking, skill, movement, power, mentality* and *defending* (see Tab. 3.2 for more details). The dataset was composed by data relying 2650 players; since on Kaggle there are only data about the beginning of the season, we took some others data (i.e. relying December 2021) directly from the *sofifa* website, in order to apply the predictive validity analysis (Sec. 3.3.3). In addition, to classify roles, the main suggestions of football experts have been followed, in order to get the specific role of each player and customize their performance: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW) [26, 82]; note that goalkeepers have been excluded from the analysis, due to their singularity and because they have specific performance indices [20, 25].

[2] `www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset`

TABLE 3.2: Statistics of the EA Sports KPIs with experts' classification for the top 5 European leagues in the 2021/2022 season

sofifa LV	Index (MVs)	MVs label	Mean	Std	Skew	Q1	Q2	Q3
	Crossing	*att1*	60.17	14.60	-0.52	50	63	71
	Finishing	*att2*	55.39	17.72	-0.32	41	59	70
Attacking	Heading accuracy	*att3*	61.90	12.35	-0.26	53	63	71
	Short passing	*att4*	69.95	9.06	-0.70	65	71	76
	Volleys	*att5*	52.14	16.09	-0.01	39	52	65
	Dribbling	*ski1*	68.04	11.81	-0.88	62	70	76
	Curve	*ski2*	58.31	15.53	-0.27	47	60	70
Skill	FK accuracy	*ski3*	50.30	15.95	0.27	38	48	63
	Long passing	*ski4*	63.09	12.09	-0.59	56	65	72
	Ball control	*ski5*	70.78	9.27	-0.65	65	72	77
	Acceleration	*mov1*	69.58	11.71	-0.49	63	70	78
	Sprint speed	*mov2*	69.72	11.61	-0.49	63	70	78
Movement	Agility	*mov3*	68.43	12.23	-0.52	61	69	77
	Reaction	*mov4*	69.20	9.35	-0.39	63	70	76
	Balance	*mov5*	67.76	12.38	-0.50	60	69	76
	Shot power	*pow1*	65.61	13.11	-0.62	57	68	76
	Jumping	*pow2*	67.29	12.11	-0.46	60	68	76
Power	Stamina	*pow3*	69.28	11.60	-0.43	62	70	77
	Strength	*pow4*	67.66	12.52	-0.54	60	69	77
	Long shot	*pow5*	56.99	16.76	-0.48	45	60	70
	Aggression	*men1*	64.73	14.15	-0.59	56	67	75
	Interception	*men2*	56.33	20.50	-0.58	39	63	73
Mentality	Positioning	*men3*	60.88	15.92	-0.68	52	64	73
	Vision	*men4*	62.24	13.42	-0.54	54	64	72
	Penalties	*men5*	55.28	13.43	0.14	45	55	65
	Composure	*men6*	67.56	10.33	-0.42	61	69	75
	Marking	*def1*	56.52	19.06	-0.55	41	62	72
Defending	Standing tackle	*def2*	58.18	20.37	-0.67	40	66	74
	Sliding tackle	*def3*	55.17	20.71	-0.58	36	63	72

In the PLS-SEM framework, the 29 indices (abilities) are the MVs, whereas the initial 6 *sofifa* traits are the LVs (i.e. the first order constructs). As already introduced before, a hierarchical model is used, but despite the early application presented in Chapter 2, a mixed two-step (hybrid) approach [41] for estimating the HOCs was preferred (i.e. it showed some stable results in terms of the BIAS and MSE of the estimates, [49]): as recall, a player's performance was built as an extra-latent construct of higher (third) order, formed from two extra LVs (second order constructs), namely, *Off_phase* (the phase at which a player is in attack, with or without ball possession) and *Def_phase* (the moment at which a player is in defense, with or without ball possession) [60]. In particular, the estimation phase for this higher order framework has followed these steps:

1. The second order latent variables (*Off_phase* and *Def_phase*) have been estimated by a mixed two step approach [41] in formula (3.6). For each one as regressors have been used the output scores relying the first order constructs, coming from a previous implementation of the PLS-SEM algorithm using a repeated indicators approach (indicated with $\hat{\boldsymbol{\xi}}_q^{\mathbf{I}}$ in the following formula, as a row vector):

$$\boldsymbol{\xi}_q^{\mathbf{II}} = \hat{\boldsymbol{\xi}}_q^{\mathbf{I}} \mathbf{w}_q^{\mathbf{II}} + \delta_q^{\mathbf{II}} \tag{3.6}$$

It has been inner structure used in the previous chapter: all first order LVs except *defending* cause the *Off_phase*, whereas all LVs except for *attacking* causing the *Def_phase* (Fig. 3.1).

At this point, the scores of the second HOCs have been used for the only third order construct ($\hat{\boldsymbol{\xi}}_q^{\mathbf{II}}$ as a row vector):

$$\boldsymbol{\xi}_q^{\mathbf{III}} = \hat{\boldsymbol{\xi}}_q^{\mathbf{II}} \mathbf{w}_q^{\mathbf{III}} + \delta_q^{\mathbf{III}} \tag{3.7}$$

Here the block q for the macro composite performance (PI *overall*) is formed by two MVs, *Off_phase* and *Def_phase*.

In particular, for both (3.6) and (3.7) $\mathbf{w}_q$ is the column vector of outer regression weights and δ_q of error terms, with their conditional expected value assumed to be zero.

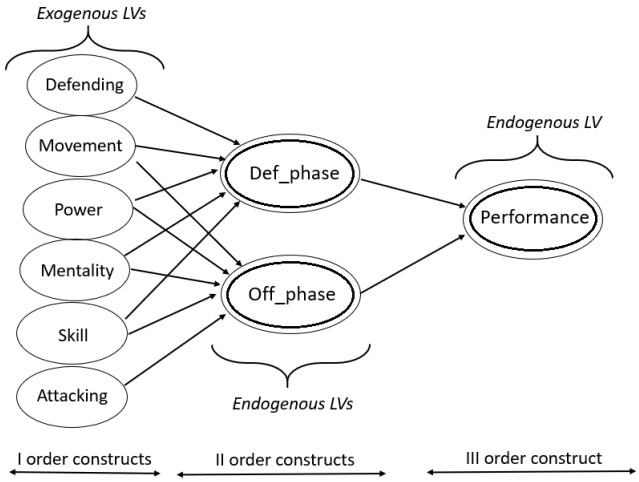


FIGURE 3.1: PLS-SEM: the path diagram of the third-order inner model

At this point, the analysis followed these steps: the preliminary analysis focused on the CTA, in order to established the nature of each construct; then the validation step by the CCA was applied, removing problematic MVs. Finally, following previous guidelines, a PLS-SEM for each player's role was implemented comparing their inner and outer estimates and their performance, by different measures of fit: the Goodness of Fit (GoF) index [133] and the Standardized Root Mean Square Residual (SRMR, [72]).

For this research the *R* software packages csem [99] and seminr [122] have been used; furthermore, the smartPLS [106] has been adopted for the CTA. We carried out a bootstrap validation (5000 replications) by using a BCa approach [75] for building

the confidence intervals (CIs) in order to assess weights and path coefficients. In the next sections, the results are shown.

3.3.2 The CTA output among each model-role

Following the roles' classification [82], 6 different PLS-SEMs were estimated, one for each player's role. In this phase the goal is firstly to investigate the nature of the constructs for each model-role: so, a CTA (with 5000 bootstrap samples for constructing the Bonferroni adjusted CIs) was adopted to detect the nature of each first order LV block. As interpretation, we adopted the method proposed by Hair et al. ([75], Chapter 3): if at least one of the nonredundant tetrads is significantly different from zero, we rejected the reflective measurement model and we assumed a formative specification. Following the previous criteria, in our case all the LVs for each model-role can be assumed as formative; for example, in Tab. 3.3 we show the CTA output for the central back (CB) model in which there is at least one tetrad (for each LV) that does not contain zero (i.e. does not vanish): 2 for the *attacking* block, 3 for the *skill*, 4 for the *movement*, 2 for the *power*, 5 for the *mentality* and 2 out of 2 for the *defending*; keep in mind that for the *defending* LV we cant't apply the tetrad, since this block has just 3 MVs [12]: since all the other LVs are formative, we assumed the same structure for *defending*.

TABLE 3.3: CTA output for the Central Back (CB) model (95% Bonferroni bias corrected bootstrap -two tailed- CIs with 5000 replications)

sofifa LV	Tetrad (τ)	τ (Sample Mean)	CI Low adj.	CI Up adj.
	att1,att2,att3,att4	2402.45	1626.82	3250.34
	att1,att2,att4,att3	1831.78	1012.14	2706.83
Attacking	*att1,att2,att3,att5*	-506.20	-1107.73	90.97
	att1,att3,att5,att2	21.19	-607.13	631.96
	att1,att3,att4,att5	-281.25	-601.38	39.40
	ski1,ski2,ski3,ski4	781.20	126.96	1443.19
	ski1,ski2,ski4,ski3	-3694.45	-5339.51	-2163.29
Skill	*ski1,ski2,ski3,ski5*	357.68	-73.05	784.47
	ski1,ski3,ski5,ski2	-3872.07	-5256.56	-2571.75
	ski1,ski3,ski4,ski5	343.73	-399.71	1110.47
	mov1,mov2,mov3,mov4	1181.01	192.33	2171.17
	mov1,mov2,mov4,mov3	2166.10	1228.13	3115.84
Movement	*mov1,mov2,mov3,mov5*	3895.69	2640.66	5255.62
	mov1,mov3,mov5,mov2	-662.85	-1100.09	-227.10
	mov1,mov3,mov4,mov5	-334.97	-1012.18	329.24
	pow1,pow2,pow3,pow4	-19.81	-372.25	332.13
	pow1,pow2,pow4,pow3	-228.24	-607.10	148.76
Power	*pow1,pow2,pow3,pow5*	101.38	-171.02	391.38
	pow1,pow3,pow5,pow2	-2150.94	-3284.02	-1081.29
	pow1,pow3,pow4,pow5	-311.94	-634.96	-1.43
	men1,men2,men3,men4	3319.12	2165.93	4527.78
	men1,men2,men4,men3	3596.57	2478.02	4767.85
	men1,men2,men3,men5	1900.04	906.19	2934.41
	men1,men3,men5,men2	94.95	-171.85	367.51
Mentality	*men1,men2,men3,men6*	404.02	-216.83	1037.80
	men1,men2,men4,men5	1974.28	1083.08	2912.84
	men1,men2,men5,men6	306.78	-211.56	846.19
	men1,men3,men4,men6	332.29	-239.13	905.82
	men1,men3,men6,men5	-2105.95	-3298.68	-956.21

3.3.3 The CCA output

After the results of the CTA, we estimated the PLS-SEM for each player's role, by using a formative approach for each measurement model: at this point we are ready to evaluate each LV, following the CCA guidelines for formative constructs. We can summarise the MVs removed for each model by role, by following the CCA suggestions (Tab. 3.4). The CB is the model with the lowest number of MVs removed (just 4), whereas the wing is the highest one (14).

TABLE 3.4: The MVs removed for each model by role by CCA

Model by role	Non-sign. outer weights	Non-sign. loadings	Multicollinearity
CB	*def2*	*mov1*	-
	mov5	*mov2*	-
FB	*def2*	*att3*	*ski5*
	mov1	*men1*	*def3*
	mov5	*men2*	-
	-	*pow2*	-
	-	*pow4*	-
MF	*mov5*	*att3*	*def2*
	pow4	*mov1*	*ski5*
	ski2	*mov2*	-
	def3	*mov5*	-
	-	*pow2*	-
OM	*att3*	*mov2*	*men4*
	men1	*pow2*	*ski1*
	mov1	-	*ski5*
	pow2	-	-
	ski3	-	-
WG	*att1*	-	*men3*
	att3	-	men4
	def2	-	*men6*
	def3	-	*ski1*
	mov1	-	*ski5*
	mov2	-	-
	mov3	-	-
	mov5	-	-
	ski3	-	-
FW	*def3*	*mov1*	*men6*
	men1	*mov2*	-
	mov2	*mov5*	-
	ski1	-	-
	ski3	-	-

Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

After removing problematic MVs, each model by role has been re-estimated: in order to evaluate each model assessment, by Tab. 3.5, the sample size, the GoF index and the SRMR of PLS-SEM by role are provided (introduced in Sec. 3.3.1). If we look at all the stats, each model by role is better than considering the full model: all GoFs are greater than 0.8 except for the full model, all SRMRs meet the threshold (i.e. lower than 0.10) except for the full model. So, also from a global point of view, taking into consideration the CTA and the CCA analysis for each model-role it improves the models' quality and their interpretation.

TABLE 3.5: PLS-SEM performance by player's role assessment

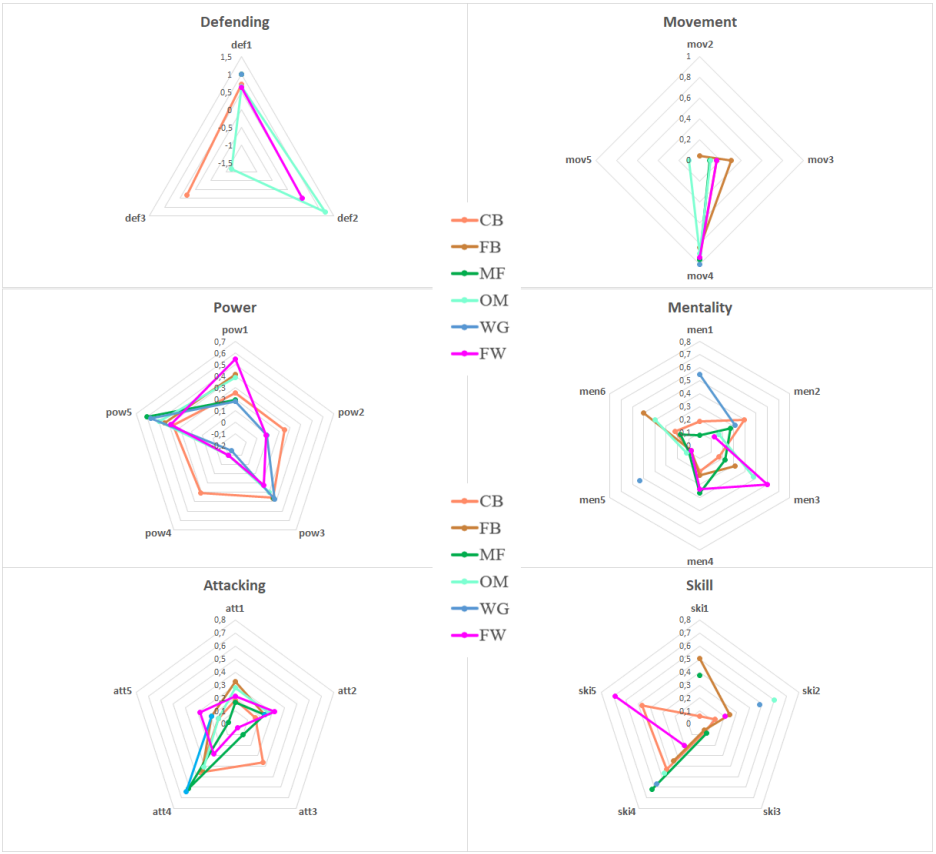
Model by role	n	GoF	SRMR
CB	535	0.814	0.094
FB	765	0.856	0.093
MF	626	0.874	0.082
OM	218	0.854	0.085
WG	117	0.886	0.070
FW	389	0.854	0.072
Full Dataset	2650	0.768	0.158

Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

Now, it is proposed a summary of the outer weights by the radar plot (Fig. 3.2), that sums up the 6 first order LVs scores by role. Since the previous CCA suggested removing some MVs, in each radar plot we have some non-continuous segments or some isolated point (e.g. the *skill* LV); in any case, we can do some interesting considerations for each LV:

- *Defending*: as we expected, CB model gets the highest weight for *def3*, whereas for *def1* there are similar scores; *def2* is present only for OM and FW as might be expected, since it refers to a *standing tackle*, an ability mainly used by offensive players.
- *Movement*: here we have similar scores for what concern *mov4*, whereas *mov3* has the highest impact for the FB, underlying the importance of the *reaction* for this important role. *mov5* is significant only for OM, with a poor impact on the movement performance.
- *Power*: here it is interesting to underline how CB has the highest scores for 3 items out of 5 and *pow1* (*shot power*) is the most important variable for FW, as we could expect.
- *Mentality*: mental skills are becoming more and more relevant in modern football for all roles; from our results it emerges that some mental abilities are significant only for one role (e.g. *men1* and *men6* for WG), others have similar scores (*men4* and *men2*) among roles. Again, another interesting evidence emerges concerning *men3* (*positioning*), that has the most relevant impact for the offensive roles (FW and OM) and *men2* (*interception*) for the CB: these considerations agree with football experts opinions ([82]).
- *Skill*: this latent trait supports the classical ball-possession ability, typical for offensive and play-maker roles; for example we have at the top OM and WG relying *ski2* (*curve*) and MF regarding *ski4*. Due to modern football (and their offensive duties), it is interesting to emphasize how FB is at the top score for what concern *ski1* (the *dribbling* impact). *ski5* is the top variable for the FW model.
- *Attacking*: for this group there are similar scores among roles for *att1* and *att2*; the outer weight of *att3* (*heading accuracy*) confirms that it has the highest impact for the CB model. *att4*, which refers to *short passing*, has the top impact for MF and WG, whereas *att5* (*volleys*) for FW.

FIGURE 3.2: The PLS-SEM outer weights summary by role

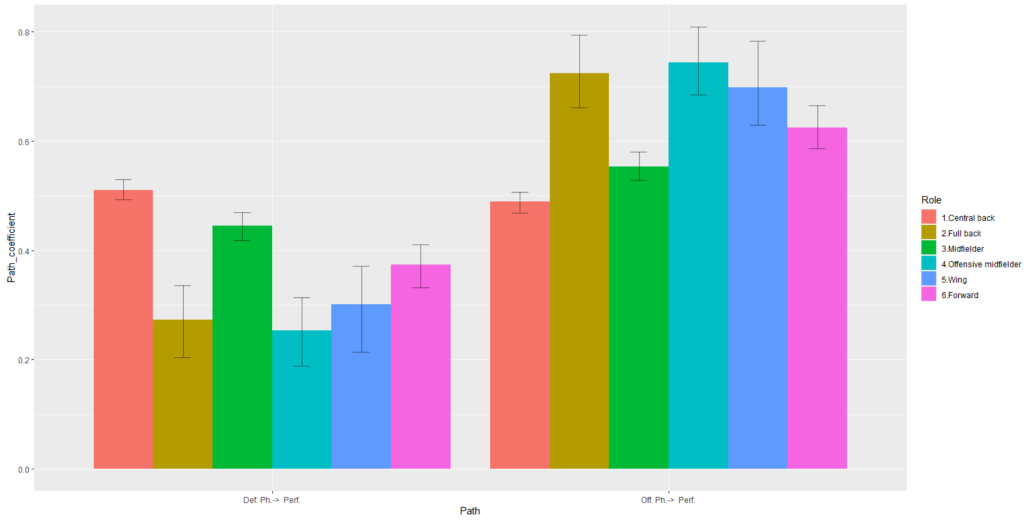


Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

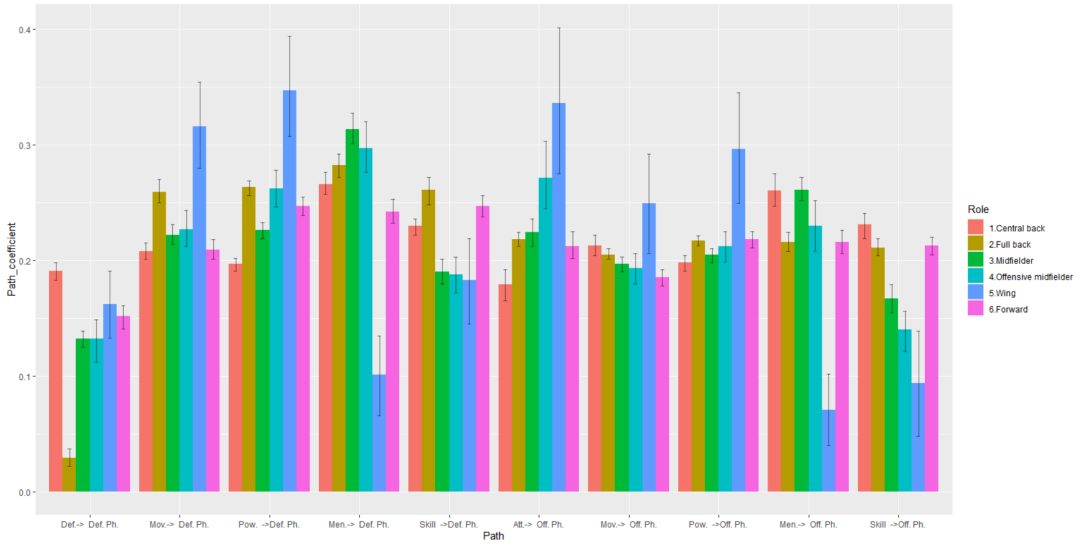
In addition, Fig. 3.3 shows two summary plots with path coefficients and their 95% bootstrap BCa CIs after 5000 replications, concerning the inner model by role. It is interesting, by Fig. 3.3a, to note how for all roles except for CB the *Off_phase* construct has an higher path coefficient on the performance composite indicator (III order construct) than the *Def_phase*; focusing on the *Def_phase* indicator, the lowest (but significant, i.e. they do not contain zero) path estimates concern the typical offensive roles and the FB. The opposite scenario there is if we focus just on the *Off_phase* indicator: the highest path estimates that impact on the PI *overall* are related to the offensive roles and the FB. Moving our attention towards the links between I and II order constructs (Fig. 3.3b) we can see differences between each role's inner estimates, in some case more evident. Below, there are some specific comments for each path, reading the plot from left to right:

- *defending* → *Def_phase*: the *defending* LV shows the highest path coefficient for CB (as expected) and similar values for the other roles, except FB (the lowest value, but statistically significant).
- *movement* → *Def_phase*: here, estimates are globally slightly higher than the previous case, with WG and FB (typical dynamic roles) those have the highest correlations (i.e. greater than 0.25).
- *power* → *Def_phase*: *power* is crucial for the *Def_phase* of WG (nearly 0.4).

- *mentality* → *Def_phase*: for the *Def_phase*, mental abilities are important for almost all roles (path greater than 0.25), except for WG, with a path of 0.1, but significant.
- *skill* → *Def_phase*: in this case there are similar scores between MF, OM and WG. *skill* has the highest correlations (greater than 0.2) with the *Def_phase* for what concern CB, FB and FW.
- *attacking* → *Off_phase*: also in this case, strictly *attacking* abilities have the highest impact (greater than 0.25) for offensive roles (OM and WG) except for FW, that has a similar score of FB and MF. As expected, in this case CB has the lowest one.
- *movement* → *Off_phase*: athletic abilities have slight differences among roles, except for WG, the highest one.
- *power* → *Off_phase*: also in this case the highest value of the path coefficient is for the WG model (almost 0.3); others roles have less pronounced heterogeneity.
- *mentality* → *Off_phase*: this LV seems to be important in the *Off_phase* for all roles (all path greater than 0.2) except for WG (less than 0.1), with a peak for MF and CB (path estimates greater than 0.25).
- *skill* → *Off_phase*: contrary to what we would expect, skill (or abilities in the ball-possession phase) is crucial for the *Off_phase* of CB, FB and FW (correlation greater than 0.2). A lower impact for MF, OM and WG.



(A) Inner model II vs III order constructs



(B) Inner model I vs II order constructs

FIGURE 3.3: Estimated path coefficients by role and 95% BCa -two tailed- bootstrap CIs (5000 replications)

In conclusion, this analysis marks again the importance of estimating the model by role: there have emerged more or less pronounced differences between the roles for both the outer and the inner model.

3.3.4 Players ranking and predictive validity analysis

Tab. 3.6 shows the top 20 players in the Europe ranking (September 2021) considering our PI overall and observing also their Off and Def phase indicators; in addition, the benchmark (EA overall) and some proxies are presented (all variables are standardized). Top players have very high scores in both the offensive and defensive indicators (there are many central backs and forwards) and the majority of them play in the English Premier League (9 over 20).

TABLE 3.6: The top players' ranking based on PI *overall*

Name	League	Role	Mkt. Value	Wage	EA *overall*	*Def_Phase*	*Off_Phase*	PI *overall*
Sergio Ramos	FR	CB	24	12	2.18	2.94	3.16	2.97
D. Alaba	SP	CB	36	20	1.62	2.67	2.95	2.71
V. van Dijk	EN	CB	86	23	2.32	2.52	2.35	2.50
Cristiano Ronaldo	EN	FW	45	27	2.60	2.24	2.53	2.47
R. Lewandowski	GE	FW	120	27	2.74	2.26	2.45	2.41
H. Kane	EN	FW	130	24	2.46	2.35	2.40	2.40
A. Griezmann	SP	FW	53	22	1.76	2.42	2.27	2.31
L. Suárez	SP	FW	45	14	2.18	2.25	2.30	2.29
Marquinhos	FR	CB	91	14	2.04	2.29	2.22	2.28
H. Son	EN	FB	104	22	2.32	2.20	2.29	2.25
K. Mbappé	FR	FW	194	23	2.60	2.12	2.24	2.22
M. Hummels	GE	CB	44	10	1.90	2.18	2.27	2.20
K. De Bruyne	EN	MF	126	35	2.60	2.00	2.52	2.20
M. Acuña	SP	FB	37	5	1.62	2.24	2.09	2.17
Bruno Fernandes	EN	OM	108	25	2.18	2.19	2.08	2.14
Azpilicueta	EN	CB	25	13	1.48	2.08	2.13	2.09
L. Messi	FR	WG	78	32	2.88	1.96	2.28	2.07
Thiago Silva	EN	CB	10	11	1.76	2.06	2.06	2.07
M. Rashford	EN	FB	78	15	1.76	2.03	2.06	2.05
L. Bonucci	IT	CB	18	11	1.76	2.01	2.02	2.01

Legend: English Premier League (EN), French Ligue 1 (FR), German 1. Bundesliga (GE), Italian Serie A (IT), Spanish Primera Division (SP)

In the last part of this results section we focus on the predictive validity of our new performance indicator, that is the last step of the CCA (Sec. 3.2): correlation of our new PI *overall* obtained by PLS-SEM with an existing benchmark (EA *overall*) and some other performance proxies: the players' market value and the players' wage in a later point of time; for this purpose, we used data from the *sofifa* website for December 2021 (original data refers to September 2021). In Tab. 3.7 there is a summary of this work, considering both concurrent (i.e. the correlation with data relying the same time) and predictive validity; the first consideration regards the comparison (for all the sub-tables) among model by role and the full dataset: in all the three cases considering the full dataset (i.e. without taking into consideration heterogeneity by role) leads to a lower correlation for both concurrent and predictive validity than considering separate roles. Again, for all the cases we can see a very slightly decreasing or equal correlation moving from September 2021 to December 2021: it means that our framework has also a good predictive validity. The comparison between our indicator and the benchmark (Tab. 3.7a) reveals very high correlation terms (near or higher than 0.90, just considering the model by role). Tab. 3.7b and Tab. 3.7c show similar scores, generally slightly higher when we compare our PI *overall* with the players' wage: anyway, for both cases we can see medium and medium-high correlations, considering model by role (always greater than 0.60).

TABLE 3.7: The predictive validity: analysis with some proxies

(A) PI *overall* vs EA *overall*

Model by role	Sept. 2021	Dec. 2021
CB	0.914	0.905
FB	0.910	0.897
MF	0.952	0.943
OM	0.977	0.967
WG	0.977	0.966
FW	0.962	0.953
Full Data	0.670	0.661

(B) PI *overall* vs Mkt. value

Model by role	Sept. 2021	Dec. 2021
CB	0.615	0.600
FB	0.606	0.606
MF	0.650	0.648
OM	0.679	0.665
WG	0.758	0.753
FW	0.616	0.609
Full Data	0.512	0.505

(C) PI *overall* vs Wage

Model by role	Sept. 2021	Dec. 2021
CB	0.699	0.690
FB	0.657	0.656
MF	0.674	0.671
OM	0.684	0.672
WG	0.762	0.760
FW	0.701	0.701
Full Data	0.530	0.526

Legend: central backs (CB), full backs (FB), midfielders (MF), offensive midfielders (OM), wings (WG) and forwards (FW)

Take in mind that final results (i.e. the path diagrams) for each model-role are shown by Fig. A1-A2-A3 in the Appendix A.

3.3.5 In-depth analysis: the goalkeepers model

Thanks a similar approach used before for movement players, now we focus attention on a singular role, the goalkeepers. Also in this case it has been applied both the CTA (for first order constructs) and the CCA analysis. In particular, a second order PLS-SEM model has been developed, in order to build a refined composite indicator dedicated to goalkeepers and comparing it with the well-known EA *overall* and others proxies (i.e. goalkeepers' market value and wage).

For this application it has been still used data provided from *sofifa* experts and available on the Kaggle[3] data science platform; in particular, we will focus on all goalkeepers' stats from the top 5 European Leagues (e.g., Italian Serie A, German Bundesliga, English Premier League, Spanish LaLiga and French Ligue1). Despite movement players, this dataset contains 31 variables, with abilities, classified by *sofifa* experts into 6 latent traits: *attacking, skill, movement, power, mentality* and *goalkeeper features*; note that, after a preliminary check, we did not take into account the *defending* block for this model, since its skills are strictly related with movement players. Note that the classification provided by *sofifa* experts is available online[4]. We took into account data relying the beginning of the season 2021/2022, so the dataset was composed by stats about 331 goalkeepers.

Relying the model, it has been assumed the goalkeepers' macro-composite performance as extra-latent construct of second order, influenced directly from the others 6 lower order constructs (LOCs). Since the HOC is without any apparent MVs, we adopted a mixed two-step approach [39, 49] for modelling this framework, using a bootstrap validation (i.e. 5000 resampling) for the model in order to assess the path significance.

[3] `www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset`

[4] https://sofifa.com/player/192985/kevin-de-bruyne/220030/

Preliminary CTA-PLS output suggested us the following classification for the LOCs:

- Reflective constructs (i.e. all vanishing tetrads in each block): *attacking*, *mentality* and *power*.
- Formative constructs (i.e. at least one tetrad does not vanish in each block): *gk_features*, *movement* and *skill*.

At this point we run the model following the CTA-PLS advice and then we assessed each LV removing problematic MVs as the CCA analysis [126]:

- Reflective constructs: we removed some MVs with reliability problems (i.e. loadings < 0.7), in particular crossing, heading accuracy and short passing that refers to the *attacking* LV, aggression, vision and penalties relying *mentality*, and jumping, strength and long shot for *power*.
- Formative constructs: here we removed MVs with collinearity problems (i.e. VIF> 5) or outer weights non-significant; agility relying the *movement* construct, whereas diving, positioning and speed for the *gk_features* block.

The final model is showed in Fig. 3.4: in the light blue circle there are formative constructs, whereas in the light blue rectangles there are reflective constructs; finally, in the white circle there is the HOC. We can see how *GK_Features* (as we expected) have the strongest impact on the macro-composite indicator (i.e. beta coefficient significant and equal to 0.28 for the inner model). It's interesting to note how for each LV the strongest MV (i.e. with highest weight or loading) is a typical variable strictly related with the goalkeepers ability [82], for example: long passing for *skill*, reaction for *movement*, shot power for *power*, positioning for *mentality*, short passing for *attacking*. Other comforting results derived from the GoF index, that is 0.792 (i.e. the geometric mean between the inner and the outer model performances) and from the SRMR (standardized root mean square residual, the difference between the observed correlations and the model-implied correlation matrix), equals to 0.096 (i.e. under the threshold of 0.10) [126].

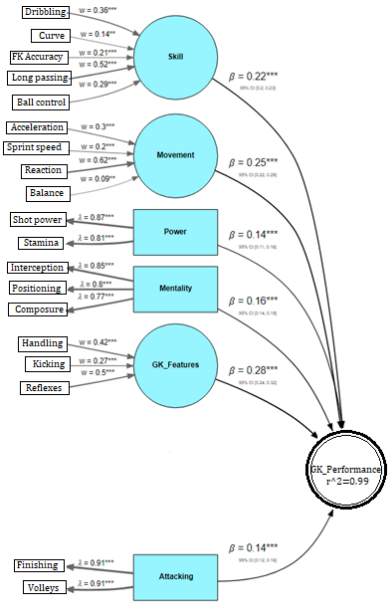


FIGURE 3.4: Goalkeepers' path diagram and estimates significant (95% BCa -two tailed- bootstrap CIs with 5000 replications)

TABLE 3.8: The goalkeeper predictive validity: correlation GK performance indicator vs some proxies

Proxies	Sept. 2021	Dec. 2021
EA *overall*	0.858	0.851
Wage	0.605	0.600
Market Value	0.585	0.571

In order to check the concurrent validity, we compared our scores with some criteria measures (Tab. 3.8), such as the EA *overall*, wage and players' market value, with interesting results: all medium-high correlations and significant, the highest between our indicator and the EA *overall*.

3.4 Chapter conclusion

Summarising, in the first part of the chapter we described two important techniques: the Confirmatory Tetrad Analysis, for evaluating the nature of each first order latent variable and the Confirmatory Composite Analysis, for confirming and refining each construct. Starting from that preface, we developed in the second part of the chapter an exploratory application, that is the natural continuation of the early framework developed in the previous chapter. With the goal to create some robust composite indicators for measuring the football player's performance quality, the computational approach by a Partial Least-Squares Structural Equation Modeling (PLS-SEM) by a hierarchical approach was adopted, creating an original application in the sport field. Following experts classification, there were developed 6 different models corresponding to the 6 roles, applying the CTA and the CCA. The CTA framework aimed to confirm by a statistical approach the nature of each first order construct: for all constructs (first order) of each role CTA defined a formative approach, whereas we assumed a formative structure for higher order ones; then, in order to refine each model by role, a CCA was followed: some problematic manifest variables (MVs) were removed (with non-significant outer weights or loadings, or due to multicollinearity problems), the highest number for the wing (WG) model (14), the lowest for the central back (CB), just 4. Considering the outer weights by role there are significant differences, and we also showed agreement with football experts' opinions. A final comparison among each model by role and the full model was provided by several assessment stats: all the indices agreed with the right choice to split players by role. In order to further validate our final composite indicator, we also assessed it for predictive validity, computing correlations of our Player Indicator (PI) *overall* with the benchmark index EA *overall*, players' market values and wages. Results are good, with higher correlations for the EA *overall* (greater than 0.90 for all the models by role) and medium-high correlation for the other two proxies (more than 0.60). Overall, recalling that the objective of this research was to create a player performance measurement model, the presented research tried to replicate an innovative model for players' performance quality and at the same time build two other interesting indicators, *Off_phase* and *Def_phase* employing a Third-Order PLS-SEM approach and validate them by a CTA and a CCA; the objective was to help football team management (coach, technical staff and scouting), in an impartial evaluation of their players, specific for each role. Another advantage derived from this new approach is the possibility to split performance into its sub-areas, in

order to study in-depth the player's abilities. From a methodological point of view, for next works it could be interesting to in-depth both the CTA for higher order constructs and to compare others HOCs estimation approaches [40]. In a similar way, as in-depth analysis, we developed a PLS-SEM second order model specific for goalkeepers, showing interesting results. Other interesting paths for future research could be, as pointed by experts [82], taking into consideration others heterogeneity observed factors (for example different seasons), using the Measurement Invariance of the COmposite Models (MICOM) procedure [75], or focused on unobserved heterogeneity among players, using the Prediction-Oriented Segmentation in PLS (PLS-POS) method [75]. In addition, it will be interesting also considering a different weights scheme based on the role of the players [37]. In particular, in the next chapter these indicators will be used for improving other existing models of football (e.g. the expected goal model [63]).

Chapter 4

An original application: the Expected Goal Model

In this chapter we will integrate as regressors the composite indicators created in the previous chapters in a well-known model: the expected goal (xG) model. The xG model is more and more used in the football world for evaluating each shot accuracy and measuring the offensive production of a team during a match. As innovation, by this chapter we want to introduce some original variables (the composite indicators) for improving and customizing the xG model. The chapter is organised as follow: in Sec. 4.1 an introduction related the expected goal world is provided, then in Sec. 4.2 the data preparation phase is explained, whereas Sec. 4.3 is devoted to illustrate the methods; practical application with results and the chapter conclusion are given respectively in Sec. 4.4 and Sec. 4.5.

4.1 Expected goal review

In this chapter we want to refine and improve, in terms of prediction accuracy, the well-known expected goal (xG) model, that overpass the most basic and frequently used metric in football to summarize the team performance: the shot, that can be a misleading metric, since it does not consider the quality of the goal-scoring opportunity from which it arises [63, 59, 1]. The main idea of the xG model is to assign a quality metric on each shot: to do so, it assigns a value between zero and one to each shot which represents the probability for a shot resulting in a goal, using a machine learning probabilistic classifier [109]. During the last years, the xG model is becoming increasingly popular and it is more and more used in the football world as proxy for measuring players' finalization performance and teams' offensive production during a match [59]. For this reason, some studies and websites have treated this topic: for example Rathke [103] and Umami et al. [128] examined shots taking in consideration only distance and angle to goal, whereas Fairchild et al. [59] made a spatial analysis of shots of the Mayor League Soccer, by a logistic regression. Another recent work [112] tried to quantify the effectiveness of defensive playing styles in the Chinese Football Super League by using xG.

The main lack is that currently xG models are based just on event data and do not take in consideration players' features. Here we want to present an original xG model by adding some composite indicators relying the players' performance and obtained by a Partial Least Squares Structural Equation Modeling (PLS-SEM, [34]), in order to take in consideration shooters and goalkeepers features. In addition, a little step forward has been done in a more recent study by Anzer and Bauer, those refined the existing xG model by adding at the classical event data some synchronized positional data, by using an extreme gradient boosting algorithm [1].

In summary, the objective is to merge and synchronize data from different sources (e.g. Understat[1] for event data, Math&Sport for tracking data and Sofifa[2] for the players' performance indicators) for refining and improving the xG in terms of model sensitivity and performance [109]. From the methodological point of view, a logistic regression (LR) model will be applied on different samples scenarios, using some machine learning sample-balanced techniques (SBT, [100, 35], since the target -the GOAL- is a rare event [103]): Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-Sampling Examples (ROSE). The benchmark to compare results will be the xG model provided by Understat (they trained neural network prediction algorithms with a large dataset -more than 100000 shots- over 10 parameters for each one) and the software for the analysis will be R (version 4.1.3, r-project.org).

4.2 Data description and preparation

Our tracking data provider, due some policy rules, gave us data relying a random sample of 53 official matches (i.e. drawing along all the season) of the Italian Serie A 2019/2020. Based on these matches, a merge among data coming from different sources was done, in particular:

- Event data: the source of these data was the well known free website Understat, that provides event data for the top five European leagues. This kind of data let us to have and compute, for example, information about the angle of each shot, if it is an head shot or not and about the shooter; additionally, the time (first or second half, seconds) of the shot and its location on the pitch (by x and y coordinates) are provided. In order to scrape these data it has been used the *R* package *worldfootballR*.

- Tracking data: for what concern these players' positional data, Math&Sport provided us the data; they are collected with a semi-automatic procedure, that captures the location of every player on the pitch (by x and y coordinates) with a temporal frequency of 16 Hz and the corresponding match time for each tracking frame is specified; in addition, for each frame the player with ball possession is tagged by a code. By these data we could know the shooter, the goalkeeper position and how many players disturb the shooter at the moment of the shot (i.e. the number of opponents between the shooter and the goal, except the goalkeeper).

- Sofifa data: these data, as recall, are provided by Electronic Arts (EA, easports.com) experts and they combine the subjective evaluations of over 9000 scouts, coaches and season-ticket holders into ratings for over 18000 players; they are free available on the Kaggle data science platform (kaggle.com) and they have been used to create some composite indicators to sum up and measure players' performance starting from the 29 Key Performance Indices (KPIs) for movement players and 31 KPIs for goalkeepers (Fig. 4.1), by using a PLS-SEM [76, 27] approach: in addition, these models have been validated by Confirmatory Tetrad Analysis and Confirmatory Composite Analysis [34] (see Chapter 3). Since these performance data are continuously update (i.e. every month, [7]), we computed and saved for our xG model the last composite indicator available before the date of the match for each player.

[1] `www.understat.com`

[2] `www.sofifa.com`

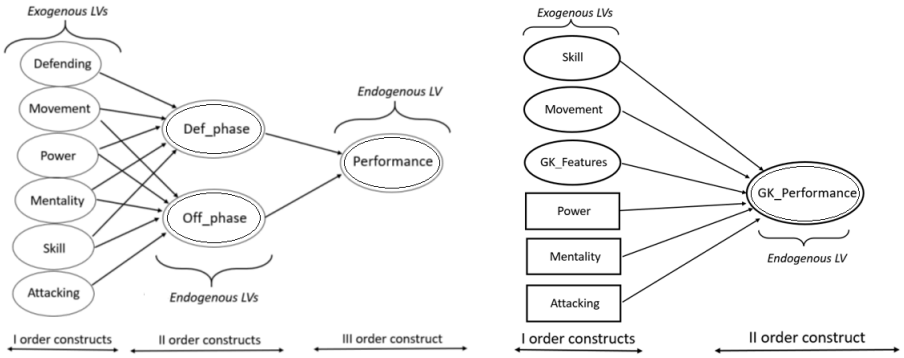


FIGURE 4.1: Movement players [34] and Goalkeepers [28] inner models used to construct the composite indicators with the PLS-SEM approach

4.2.1 The tracking features

For what concern the coordinates on the pitch, we modified them in order to have a standardized measure, representing respectively the percentage of length and width of the pitch (Fig. 4.2).

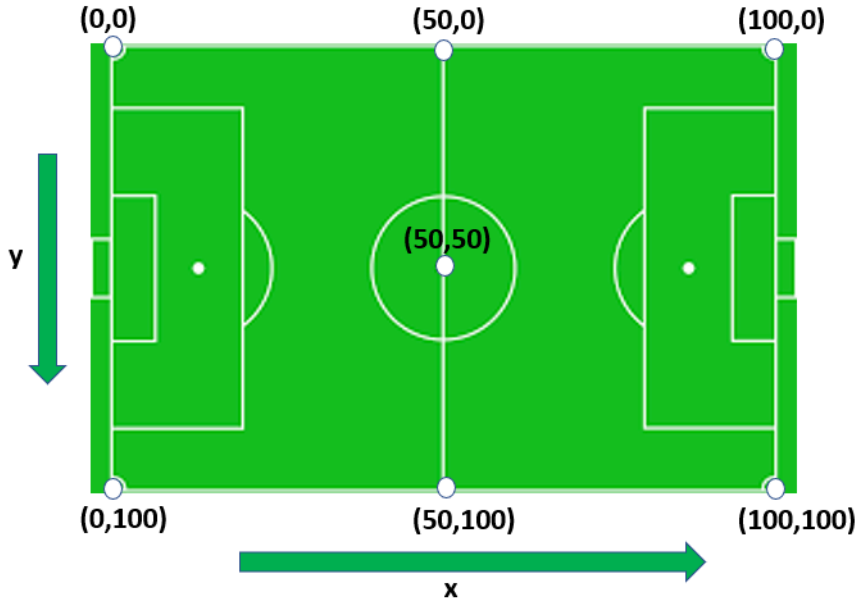


FIGURE 4.2: Standardized x and y coordinates on the pitch

Then, in order to compute the number of players around the shooter, we proceeded in this way: we counted how many players (excluding goalkeeper) are in front of the shooter (x coordinate greater than the x of the shooter) and around him (i.e. y coordinate around the 10% respect the y of the shooter). In Fig. 4.3 we can see an example: in this case the shooter (blue point) has around him two opponents (red points).

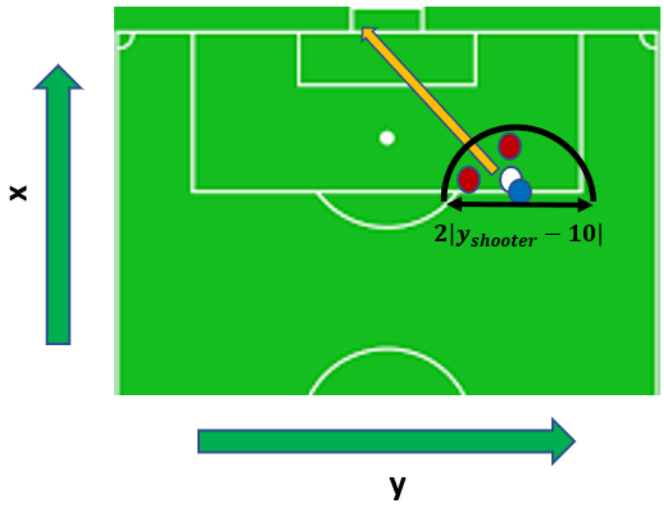


FIGURE 4.3: Example of one shooter (blue point) with two opponents players (red points)

4.2.2 Data merging phase

In order to create the final dataset for the xG model, first of all we did some Extraction Transformation Loading (ETL) steps, merging the three different sources data by using the *R* package *sqldf*:

1. Data cleaning: from the tracking dataset we removed by a filter all the row-data tagged as stopped game (for example when the ball is out of the pitch or when a player is injured on the pitch) whereas for the event dataset we kept just the shot events.

2. Players' code legend table: a support table with respectively players' names and IDs was created, in order to simplify the successful join among the three sources data.

3. "Adjusted" Left Join event vs tracking: a naive procedure for each match was developed in order to merge tracking and event data; after a preliminary check, we noted that, because of different sources, the time frame of each actions was not perfectly the same: so, we decided to use a combined primary key (match and player code joined together) that matched the nearest time frame among the two sources. In the other words, by this procedure, we performed an "adjusted" left join between event and tracking data for each football game: it is called "adjusted" because we downloaded from the tracking dataset the time frame closest that matched with our combined primary key.

4. Sofifa merging: the last ETL phase consisted of joined to the shot dataset the players' composite indicators created by the PLS-SEM, by using the players' code; in particular, for each shot we added the indicators relying the shooter and the opponent goalkeeper.

Due to their limited number, we excluded 9 free kicks, 2 penalties, the variable "headed" (just 13 headed shots) and 3 outliers (Fig. 4.4) from the final dataset; so, at the end of the ETL procedure, the dataset was composed by a sample of 660 shots and 23 features for each-one, relying 53 matches of the season 2019/2020 (Italian

Serie A): in Tab. 4.1 are described the features used and their sources. Take in consideration that there are 3 binary features: GOAL (i.e. the outcome, 1=goal and 0=no goal), previous dribbling before the shot (1=yes, 0=no) and favourite foot, i.e. if the player shots on goal with his favourite foot or not: 1=yes, 0=no; in addition, we converted the variable that counts the number of opponents player around the shooter into two dummies: the first one (i.e. D1.OpponentsPlayer) is equal to 1 when the opponents number is greater than 0, whereas the second one (D2.OpponentsPlayer) is active when that number is greater than 1; the others 18 features are continuous. In addition, the shots distribution concerning our final dataset is showed in Fig. 4.4.

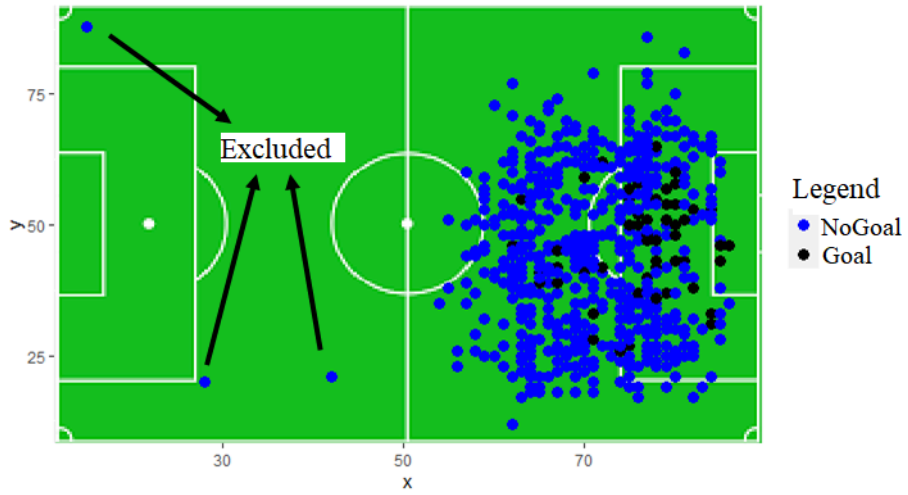


FIGURE 4.4: Distribution on the pitch of the 660 shots for the 53 matches of the Italian Serie A (Season 2019/2020)

TABLE 4.1: Statistics of the variables for the sample of 660 shots of 53 matches of the Italian Serie A (Season 2019/2020)

Variable description	Source Dataset	Mean	Std.	Q1	Q2	Q3
GOAL (Yes/No)	Understat	0.10	0.30	-	-	-
x (shooter coordinate, %)	Understat	83.22	7.64	77.00	83.00	89.00
y (shooter coordinate, %)	Understat	49.98	15.06	37.00	50.00	63.00
Favourite foot (Yes/No)	Understat	0.77	0.42	-	-	-
Previous dribbling (Yes/No)	Understat	0.32	0.47	-	-	-
Angle of shot (degree)	Understat	37.39	20.81	20.22	39.42	52.80
Previous ball distance (%)	Tracking	14.43	8.41	8.60	14.33	17.09
Possession duration (sec.)	Tracking	6.49	4.81	5.68	6.49	6.49
D1.OpponentsPlayer	Tracking	0.13	0.33	-	-	-
D2.OpponentsPlayer	Tracking	0.05	0.21	-	-	-
GK x coordinate (%)	Tracking	96.74	5.25	96.55	97.67	98.41
GK y coordinate (%)	Tracking	50.08	5.06	47.37	50.04	53.07
Defending	Sofifa (PLS-SEM)	0.06	0.99	-0.62	0.14	0.61
Mentality	Sofifa (PLS-SEM)	0.38	1.00	-0.28	0.41	1.17
Movement	Sofifa (PLS-SEM)	0.41	1.03	-0.27	0.32	1.00
Power	Sofifa (PLS-SEM)	0.49	1.06	-0.16	0.53	1.10
Skill	Sofifa (PLS-SEM)	0.47	0.91	-0.15	0.49	1.12
GK Attacking	Sofifa (PLS-SEM)	-0.03	1.01	-0.78	0.2	0.77
GK Features	Sofifa (PLS-SEM)	0.68	0.78	0.32	0.82	1.12
GK Mentality	Sofifa (PLS-SEM)	0.26	0.98	0.08	0.43	0.82
GK Movement	Sofifa (PLS-SEM)	0.55	0.73	0.12	0.53	1.02
GK Power	Sofifa (PLS-SEM)	0.45	0.92	0.10	0.45	1.19
GK Skill	Sofifa (PLS-SEM)	0.06	0.79	-0.51	-0.19	0.52

4.3 Logistic regression and sample balanced techniques

In this section we will discuss about the methodological framework used for the analysis: since we deal with a classification problem i.e. the Goal (Y = 1) or NoGoal (Y = 0) based on a set of regressors (18 continuous and 4 binaries), we applied a logistic model with parameters estimated by maximum likelihood [85]. We preferred this statistical model than others because of its easy implementation/interpretation concerning the regressors effects and because the real focus of this study is to introduce new predictors in the expected goal (xG) model, in order to improve the goal probability estimation. In the context of the xG, this model lets to estimate the conditional probability of goal for a given shot i by its set of features values $\mathbf{x}$, as row vector, and estimate parameters $\hat{\beta}$ in (4.1). Note that the regression coefficients are estimated by maximum likelihood [85], with the corresponding maximum of the log-likelihood in (4.2).

$$xG = P(Goal|\mathbf{X}) = \frac{e^{\mathbf{X}\hat{\beta}}}{1 + e^{\mathbf{X}\hat{\beta}}} \tag{4.1}$$

$$logL(\hat{\beta}) = \sum_{i=1}^{n} y_i log(xG_i) + \sum_{i=1}^{n} (1 - y_i) log(1 - xG_i) \tag{4.2}$$

Then, the typical good of fit index used in logistic regression is the McFadden Pseudo-R^2 [97], that ranges between 0 and 1:

$$Pseudo - R^2 = 1 - \frac{logL(\hat{\beta})}{logL(\hat{\beta_0})} \tag{4.3}$$

Where $logL(\hat{\beta})$ is the max log likelihood value for the fitted model (including all the predictors) and $logL(\hat{\beta_0})$ is the max log likelihood value for the null model which includes only the intercept as predictor; regards its interpretation, already from a range between $[0.20 - 0.40]$ it represents excellent fit, considering that $(2 \cdot Pseudo - R^2)$ has roughly the same interpretation of the standard R^2 for linear regression [97].

$$OR = \frac{xG}{1 - xG} = \frac{P(Goal|\mathbf{X})}{P(NoGoal|\mathbf{X})} \tag{4.4}$$

Finally, in the logistic regression framework, the odds ratio estimate (OR), in order to establish if one regressor can be considered as risk ($OR > 1$), neutral ($OR = 1$) or protective ($OR < 1$) factor for the event, is the suitable metric (4.4).

4.3.1 The metrics used to evaluate the model

In the field of classification problems, the more appropriate performance measures may be derived from the confusion matrix (Tab. 4.2), which compares the predicted labels to the actual ones.

TABLE 4.2: The Confusion Matrix

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

In order to provide comprehensive assessments of imbalanced learning problems, the most frequently adopted performance measures are based on different propensity towards false negatives (FN) and false positives (FP) [81]:

- Accuracy: it is the ratio of correct predictions over the total number of instances evaluated, as $(TP + TN)/(TP + TN + FP + FN)$.
- Sensitivity (Recall): it measures the proportion of true positive (TP) that are correctly classified $TP/(TP + FN)$.
- Specificity: it measures the proportion of true negative (TN) that are correctly classified $TN/(TN + FP)$.
- Precision: it computes the fraction of examples classified as positive that are truly positive $TP/(TP + FP)$.
- F1: it represents the harmonic mean between precision and recall, as $(2 \cdot Precision \cdot Recall)/(Precision + Recall)$.

One of the most frequently used tools for evaluating the accuracy of a classifier is the Receiver Operating Characteristics (ROC) curve. As the classification threshold

varies, the predicted label is assigned to the examples and the confusion matrix represented. The true positive rate (sensitivity of the classifier) is then plotted versus the false positive rate (1-specificity of the classifier) for each considered value of the classification threshold. The classifier performs as better the steeper the ROC curve becomes, that is, the larger the area underlying the curve (AUC) is (1 as maximum possible value).

4.3.2 The Imbalanced Training Sample Problem

By observing Tab. 4.1 stats, it's clear how our target feature is an imbalanced variable, as the proportion of goal is "only" the 10% considering all the shots in our dataset (i.e. the goal is a rare event). In these circumstances is sometimes recommended the use of models like the Gompit (complementary log-log) [18], but in this case the main goal is to improve the goal detection capability of the xG model; supporting the last sentence, it has been largely reported that this class imbalance heavily compromises the training process, because the model tends to focus on the prevalent class and to ignore the rare event (the scarcity of data leads to poor model's accuracy and the model struggles to correct classify the rare event) [100, 54].

In order to overcome this problem, the issue of class imbalance can be addressed in two ways: one is to assign distinct costs to training examples and the other is to balance by some resampling techniques the original dataset [35]; in particular, we will treat with the second group of techniques. Remedies following the "resampling" approach include various approaches: the most common are random oversampling with replacement the rare class and random undersampling (without replacement) the prevalent class. Oversampling, in its simplest form, duplicates examples of the minority class, while undersampling removes some data from the most frequent class [100]. Generally, before applying these approaches the dataset is used randomly split in two parts: 2/3 of the data are used as training set and the remaining 1/3 as test set; oversampling and/or undersampling work directly on the training set, then the classification model that coming in output is performed in the test set. Both undersampling and oversampling have some weak points [96]: the first one may discard potentially useful data, thus reducing the sample size, while the second may increase the likelihood of overfitting, since it is bound to produce ties in the sample, especially as the sampling rate increases. In addition, the augmented sample increases the computational effort of the learning process.

In order to overcome the limitations of classic under and oversampling techniques, some other approaches have been developed: in particular, increasing attention has been recently paid to the novel strategy of generating new artificial examples that are "similar" to the observations belonging to the minority class. In the next sections, we will discuss about two of these approaches: the Synthetic Minority Oversampling Technique (SMOTE, [35]) and the Random OverSampling Examples (ROSE, [100]). From a practical point of view, for both algorithms, the synthetic training set can be used to estimate the classification model, whereas the original data remain free of being used as test set (alternatively, cross-validation or smoothed bootstrap methods, [100]). In addition, when we balance a dataset thanks some techniques like ROSE or SMOTE, we must compensate for the effects of our modifications to the training data [8, 52]. For the Bayes' Theorem, posterior probabilities are proportional to the prior ones, which can be estimated as the relative frequency in each category. Therefore, the estimated posterior probabilities (expected goal) obtained using artificially balanced data set can be corrected (calibrated, xG^*) using

the following formula (4.5):

$$xG^* = \frac{\frac{0.1}{0.5}xG}{\frac{0.1}{0.5}xG + \frac{(1-0.1)}{(1-0.5)}(1 - xG)} \tag{4.5}$$

as in our case 10% and 50% are the real and artificial (balanced) sample sizes of the rare class respectively. Additionally, (4.5) will be used for calibrating the test set probabilities (see better in Sec. 4.4).

4.3.3 SMOTE

This algorithm was proposed by Chawla et al. [35] and it is an approach in which the minority class is oversampled by creating "synthetic" examples rather than by oversampling with replacement. The rare class is oversampled taking in consideration each rare class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbours. Depending upon the amount of oversampling required, some neighbours are randomly chosen: for instance, if we set $k = 5$ nearest neighbours and if the amount of oversampling needed is 100%, only one neighbor is chosen and one sample is generated in that direction. In Fig. 4.5 we show how the algorithm works considering as example the statistical unit u_i: it chooses the 5 nearest neighbours and then it randomly selects one of them (u_{ik}^{Knn}); the procedure that creates the artificial unit u_j^{SMOTE} can be summarized by (4.6).

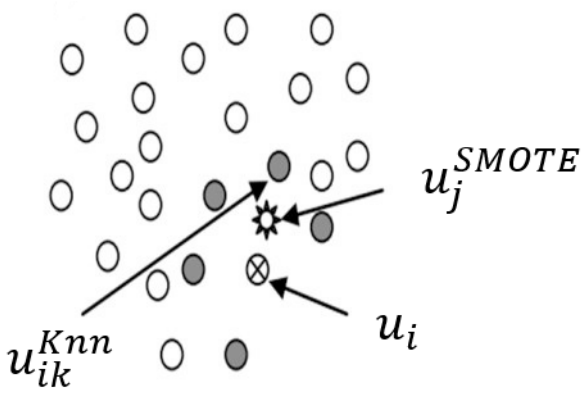


FIGURE 4.5: Minority observations cloud: example of SMOTE procedure

$$u_j^{SMOTE} = u_i + (u_{ik}^{Knn} - u_i) \cdot \delta_i \tag{4.6}$$

In order to focus on (4.6), synthetic samples are generated in the following way:

1. Compute the difference between the features vector under consideration and its nearest neighbour;
2. Multiply the difference of the step 1 by a random number (δ) between 0 and 1;
3. Add the quantity coming out from the step 2 to the features vector under consideration.

The procedure described causes the selection of a random point along the line segment (u_j^{SMOTE}). So, this approach effectively forces the decision region of the minority class to become more general. In particular, since our dataset contains 4 nominal covariates (i.e. binary) we will use the SMOTE for Nominal-Continuous (SMOTE-NC) features [35]: here the continuous variables of the new synthetic minority class sample are created using the classical SMOTE approach, whereas for the nominal features are given the values occurring in the majority of the k-nearest neighbours. The main drawback of SMOTE is that the majority of the new observations will be located near the most dense zones of the minority class, as consequence there are just few units near the marginal points.

4.3.4 ROSE

The Random OverSampling Examples (ROSE) algorithm is based on the generation of new artificial data from classes, according to a smoothed bootstrap approach, proposed by Menardi and Torelli [100]. Considering the initial training sample with sample size n, the ROSE procedure for generating one artificial example follows these steps:

- Select randomly one of the two classes assuming equiprobability;
- Called $n_j < n$ the sample size of the class j selected randomly select a statistical unit i with its features vector $\mathbf{x}_i$ with probability $p_i = \frac{1}{n_j}$;
- Sample $\mathbf{x}$ from $K_{\mathbf{H}_j}(\cdot, \mathbf{x}_j)$, with $K_{\mathbf{H}_j}$ a probability distribution centred at $\mathbf{x}_i$ and $\mathbf{H}_j$ a matrix of scale parameters.

In practise, the algorithm draws from the training set an observation belonging to one of the two classes and generate a new example in its neighbourhood (the width of the neighbourhood is determined by $\mathbf{H}_j$). Note that $K_{\mathbf{H}_j}$ is usually chosen in the set of unimodal and symmetric distributions. In particular, once a label class y_j has been selected, this technique generates new examples starting from that one by generating data from the kernel density estimate of $f(\mathbf{x}|y_j)$ (4.7).

$$\hat{f}(\mathbf{x}|y_j) = \sum_{i=1}^{n_j} p_i \cdot Pr(\mathbf{x}|\mathbf{x}_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} \cdot K_{\mathbf{H}_j}(\mathbf{x} - \mathbf{x}_i) \tag{4.7}$$

The repeated implementation of the three steps explained above let to create a new artificial training set (let's call $\mathbf{T}_s^*$) with size s where approximately the same number of examples belong to the two classes. Note that the size s can be set to the original training set size n or chosen in any other way. ROSE combines techniques of oversampling and undersampling by generating an augmented sample of data thus helping the classifier in estimating a more accurate classification rule, because the same attention will be addressed to both the classes (e.g. strengthening the process of learning as well as estimating the distribution of the chosen measure of the model accuracy).

ROSE allows to overcome the limits of both apparent error (i.e. overfitting problems, a too high accuracy in the classifier's performance) and holdout method (i.e. non-advisable in unbalanced learning because the scarcity of rare class prevents their use in both training and test set). It was proofed that, by generating ROSE examples, the logit model shows better performances (in terms of classification accuracy and precision) than the classical decision trees [100].

4.4 Results and discussion

From a practical point of view, we developed a routine in R by using a stratified 3-Fold cross validation for evaluating the model fit and computing the performance measures, with 5000 replications (Fig. 4.6); it was performed for each framework (ROSE, with *R* package *ROSE* and SMOTE, with *R* package *RSBID*) and it works for each one of the 5000 replication in this way:

1. A stratified (i.e. keeping constant as in the original dataset the proportion of Goal and NoGoal in both the training and the test set) 3-fold cross validation has been performed, balancing each time the training set before applying the logistic regression, using all the regressors introduced in Tab. 4.1 (quantitative regressors have been standardized): at this point, for each fold the model fit was measured by computing the $Pseudo - R^2$; then, the corresponding quantitative regressors on the test set were standardized too, before computing xG(4.1) and the correction xG^*(4.5). As a conclusion of this step, the out-of-sample performance metrics (Sec. 4.3.1) have been computed, using the xG^* probabilities in (4.5).

2. For each 3-fold cross validation procedure, the average for each performance metric and for the $Pseudo - R^2$ have been computed;

3. Repeat the steps 1 and 2 for 5000 times, each time with different (random) split of the dataset for what concern the stratified 3 fold cross validation.

In addition, a similar routine (without balancing the training set) has been performed for the imbalanced dataset, whereas for the benchmark (i.e. Understat) we directly had the xG probability. Note that xG(4.1) has been used for the Imbalance out-of-sample estimates.

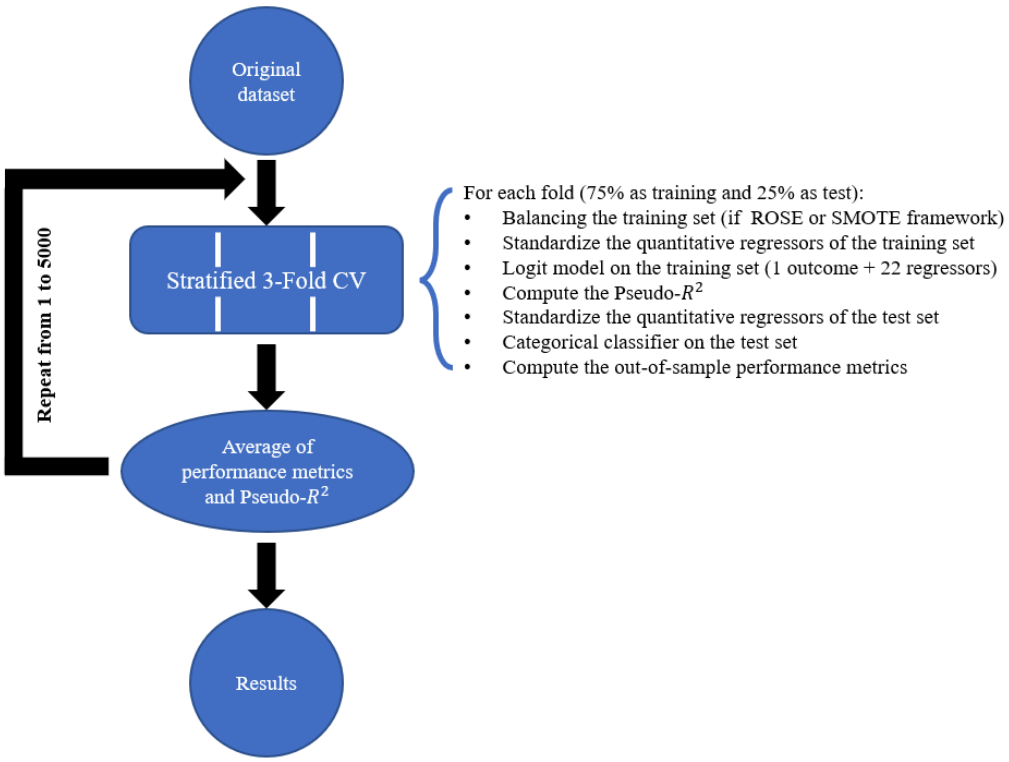
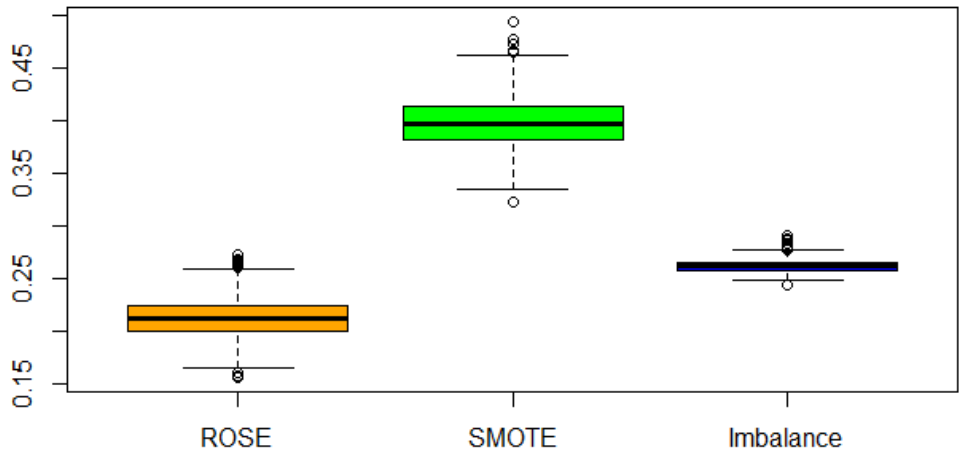


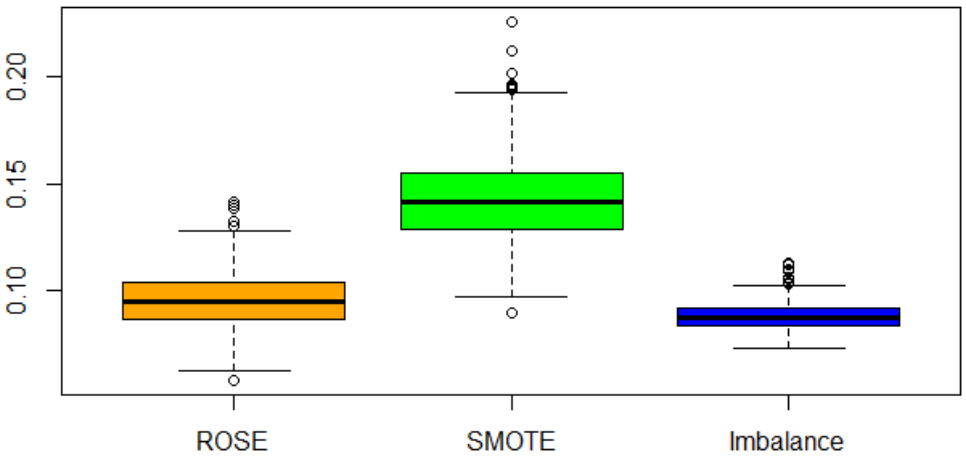
FIGURE 4.6: The estimation process

4.4.1 The logistic regression model output

Here we show preliminary results concerning the logistic regression model: first of all, none regressor showed collinearity problems (i.e. all Variance Inflation Factors -VIFs- lower than 5, [107]); additionally, for evaluating the model fitting on each training set, the distribution of the Pseudo-R^2 obtained with 5000 replications of the stratified 3-fold CV has been computed and showed by a boxplot (Fig. 4.7a): the SMOTE approach shows better performance (McFadden between 0.32 and 0.50), whereas for ROSE and imbalance the Pseudo-R^2 has respectively a range between 0.15 and 0.29 and 0.25 and 0.30. In addition, the difference between the Pseudo-R^2 distribution of our model and a nested model with just 3 regressors, as x, y and angle of shot [103, 128] for each framework is showed in Fig. 4.7b: in all the cases our complete model shows better performance in terms of model fitting (the boxplot difference distributions do not contain zero).



(A) The boxplot of the Pseudo-R^2 obtained with 5000 replications among frameworks relying the complete model



(B) The boxplot of the difference between the Pseudo-R^2 (complete vs nested) among frameworks (5000 replications)

FIGURE 4.7: The Pseudo-R^2 McFadden distributions

Now, focusing on regressor estimates and their odds ratios, we show in Tab. 4.3 a

summary after 5000 iterations. To assess the statistical significance of the regressors the following notation is used [51]:

- *: if, for the $i-th$ regressor, the interval $\hat{\beta}_i \pm 1 * Std_{\hat{\beta}_i}$ does not contain 0.
- **: if, for the $i-th$ regressor, the interval $\hat{\beta}_i \pm 2 * Std_{\hat{\beta}_i}$ does not contain 0.
- ***: if, for the $i-th$ regressor, the interval $\hat{\beta}_i \pm 3 * Std_{\hat{\beta}_i}$ does not contain 0.

We can see how not all the regressors in the model have significant estimated coefficients: for all frameworks are significant (at least with one *) the estimate of the coefficients of x, angle of shot, previous ball distance for what concern the Understat features block, whereas we find D2.OpponentsPlayer and the goalkeeper position (at least two **) for the tracking block. Then, also some performance features like movement, GK Mentality and GK Skill are relevant for both frameworks. The approach with the highest number of significant features is the imbalance (17 with three asterisks). Again, it's interesting to in depth this analysis by observing the odds ratios related with the relevant regressors: as recall, an OR greater (lower) than 1 means that feature is preparatory (protective) for the goal. For instance, the x regressor has an $OR > 1$ for each framework, it means that the position to shot is preparatory to goal (i.e. high x position -near the box- means high probability to score a goal); similar situation ($OR > 1$) for GK y coordindate and Movement, two original composite performance indicators. An example concerning the opposite situation is the angle of shot ($OR < 1$), it means that is protective to goal (i.e. high angle of shot means low probability to score a goal); comparable circumstance for our original regressors D2.OpponentsPlayer, GK y coordinate and GK mentality.

TABLE 4.3: The Logistic Regression coefficients and odds ratios estimates after 5000 iterations for the 53 matches of the Italian Serie A (Season 2019/2020)

Regressor	Coeff. ROSE	Coeff. SMOTE	Coeff. Imbl.	Odds ROSE	Odds SMOTE	Odds Imbl.
x	0.94***	2.09***	1.74***	2.57	8.10	5.70
y	0.02	-0.05	-0.09***	1.02	0.95	0.91
Favourite foot	-0.16	0.10	0.17***	0.85	1.10	1.19
Previous dribbling	0.21*	0.33*	0.46***	1.23	1.39	1.59
Angle of shot	-0.53***	-1.40***	-1.09***	0.59	0.25	0.34
Previous ball distance	-0.09*	-0.22*	-0.06**	0.92	0.80	0.94
Possession duration	-0.18	-0.25*	-0.22***	0.84	0.78	0.80
D1.OpponentsPlayer	-0.74	-1.02*	-0.71	0.48	0.36	0.49
D2. OpponentsPlayer	-5.69***	-6.13***	-5.68***	0.00	0.00	0.00
GK x coordinate	-0.33***	-0.59**	-0.33***	0.72	0.56	0.72
GK y coordinate	0.27***	0.48***	0.46***	1.32	1.62	1.59
Defending	0.14*	0.13	0.16***	1.15	1.14	1.17
Mentality	0.03	-0.23	-0.40***	1.03	0.79	0.67
Movement	0.12*	0.40*	0.27***	1.13	1.49	1.31
Power	-0.01	-0.07	0.14***	0.99	0.93	1.16
Skill	0.08	0.23*	0.21***	1.09	1.26	1.23
GK_Attacking	0.02	0.05	0.03*	1.02	1.05	1.03
GK_features	0.03	0.32*	0.22***	1.03	1.38	1.24
GK_Mentality	-0.15*	-0.44**	-0.33***	0.86	0.64	0.72
GK_Movement	0.10*	0.15	0.07**	1.10	1.17	1.08
GK_Power	-0.17*	-0.20	-0.17**	0.84	0.82	0.84
GK_Skill	0.10*	0.27**	0.20***	1.11	1.31	1.22

Finally, in order to illustrate the expected goal probability directly on the pitch, an heatmap (using the coefficients of Tab. 4.3 related the imbalance framework) is provided in Fig. 4.8: the field zones have been divided in deciles and the colour shade is proportional to the estimated probability to score a goal (xG).

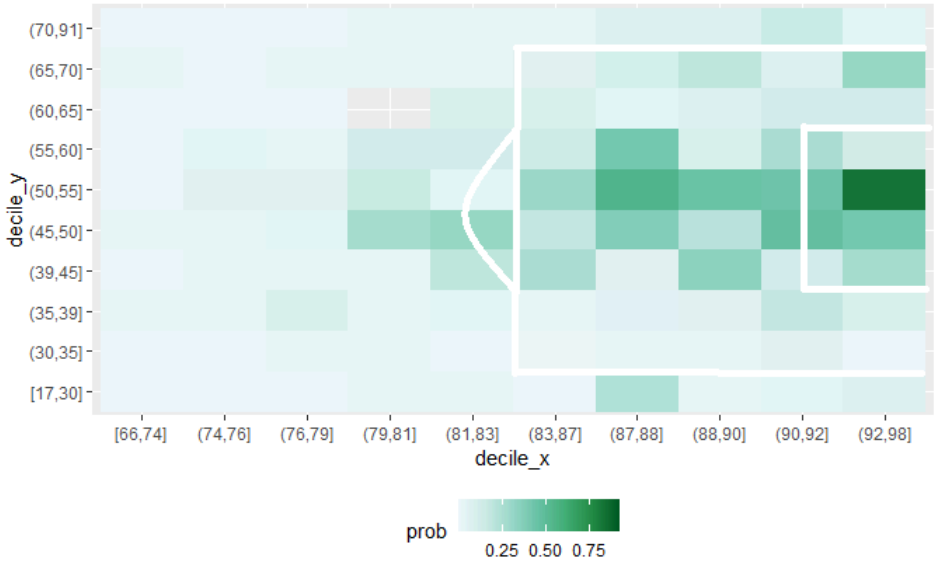


FIGURE 4.8: The xG heatmap for the 53 matches of the Italian Serie A - Season 2019/2020

4.4.2 Classification performance of the Logit Model

As explained in Sec. 4.3.1, for binary classification problems usually the probabilistic classifier (xG) is transformed in the categorical one (Goal, NoGoal) using the threshold 0.5. For instance, in our situation, an expected goal (shot quality) greater than this threshold will be classified as a goal. In Tab. 4.4 are proposed all the classification performance metrics (Sec. 4.3.1) and their average scores (5000 replications), comparing them with the benchmark (that is a punctual value, directly provided from Understat). Take in consideration that asterisks in Tab. 4.4 (Tab. 4.5) must be interpreted in a similar way of coefficients and odds ratios related the previous paragraph (Sec. 4.4.1), but in this case the confidence interval must not be include the corresponding benchmark (Understat) metric to be significant.

TABLE 4.4: The performance classification metrics averaged after 5000 replications for the sample of 660 shots of 53 matches of the Italian Serie A -Season 2019/2020- compared with the benchmark (classification threshold=0.5)

Metric	ROSE	SMOTE	Imbl.	Understat
Accuracy	0.89*	0.86***	0.90	0.91
Sensitivity	0.14*	**0.36*** **	0.15	0.16
Specificity	**0.98***	0.91***	**0.98***	0.96
Precision	**0.35***	**0.32****	**0.51*** **	0.22
F1	0.16	**0.33*** **	0.23	0.19
AUC	0.72	0.73	**0.74***	0.72

In Tab. 4.4 the metrics those outperform the benchmark are emphasized in bold. Using the classical classification threshold (0.5), both ROSE and Imbalance significantly outperform in terms of specificity and precision the benchmark (Understat) whereas SMOTE seems able to detect better the goals (sensitivity equals to 0.36 vs 0.16 of the benchmark) and to improve Understat in terms of F1 and precision. The

AUC metric is similar for all the frameworks, except for the Imbalance, that significantly outperforms Understat (0.74 vs 0.72).

Despite 0.5 is the classical classification threshold used for the categorical classifier,we noted that the Goal event is a rare event (Tab. 4.1, 10% of Goal): for this reason, we decided to simulate some scenarios with different classification thresholds (from 0 to 0.5, using a pace of 0.001), for each framework. For each simulation have been saved the average sensitivity and specificity rate after 5000 replications, except for Understat, for which we directly had the xG probability and so the related probabilistic classifier value: we can see by Fig. 4.9 that the two balanced approaches have a similar equilibrium threshold point (0.135 for ROSE and 0.155 for SMOTE), whereas it is at about the half for the two imbalanced frameworks (0.075 for imbalance and 0.065 for the benchmark); then, in correspondence of each equilibrium threshold, all the 4 frameworks have a similar rate (around 0.65 for both specificity and sensitivity). In addition, the frameworks curves seem to have a similar slope, except for SMOTE, that is a little less pronounced: when threshold is 0.5, it has a higher sensitivity rate than the other approaches.

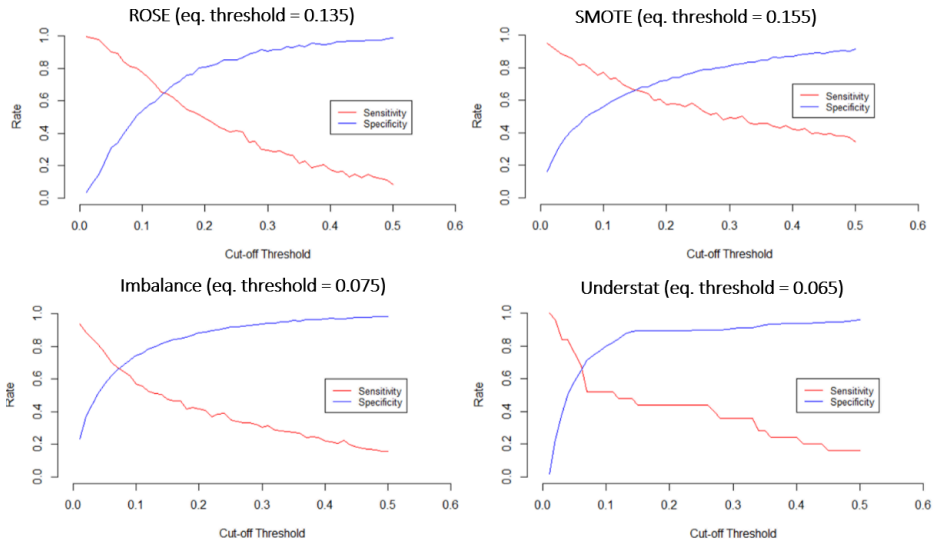


FIGURE 4.9: The classification thresholds performance scenarios for each framework

At this point, we still performed the 5000 replications using for each framework the corresponding equilibrium point as threshold: we chose this point since it is the only one in which the proportion of real Goal predicted (sensitivity) and real NoGoal predicted (specificity) are the same, minimizing both the errors, as trade-off point. Results are showed in Tab. 4.5: for the AUC metric we have the same situation of Tab. 4.4; in particular, since we chose an equilibrium point, we have lost some percentage of accuracy and specificity than before, but the benchmark (Understat) has been outperformed for what concern the percentage of goal correctly predicted (precision) and for the F1 metric.

TABLE 4.5: The performance classification metrics averaged after 5000 replications for the sample of 660 shots of 53 matches of the Italian Serie A -Season 2019/2020- compared with the benchmark (classification threshold=equilibrium point of each framework)

Metric	ROSE	SMOTE	Imbl.	Understat
Accuracy	0.66	0.66	0.67	0.66
Sensitivity	0.66	0.66	0.66	0.65
Specificity	0.66	0.66	0.66	0.65
Precision	**0.17*****	**0.17*****	**0.18*****	0.11
F1	**0.28*****	**0.28*****	**0.28*****	0.19
AUC	0.72	0.73	**0.74***	0.72

4.4.3 In depth-analysis: some real cases

Now, in order to emphasize the importance of the new regressors (players'composite indicators and their positions on the pitch) introduced in the model, we propose some real goal, comparing their expected goal for each framework and introducing some variation, in order to better understand how the xG changes.

In the first real case (Fig. 4.10) we propose a goal scored from a high distance during the match Bologna vs Inter (November 2019), in a situation with a high number of opponents in front of the shooter: in particular, here we have a good player as shooter (Soriano, Bologna) and a top player as goalkeeper (Handanovic, Inter). Then we propose the expected goal for each framework (xG^* for ROSE and SMOTE) and others two scenarios (Tab. 4.7): the first one putting a top player as shooter (Ronaldo), whereas the second one leaving the same top player as shooter (Ronaldo, Juventus) and moving a normal goalkeeper (Skorupski, Bologna). We can see how in the real scenario our balanced frameworks increase the goal prediction accuracy (higher xG than the benchmark); the xG for the imbalance approach is very similar (2.0% vs 2.1%). It's interesting to note how introducing firstly a top player as shooter (Scenario 1), then a normal goalkeeper (Scenario 2) the expected goal increase in both three frameworks, emphasizing the importance of introducing players' performance indices in the model, as innovation of this work.

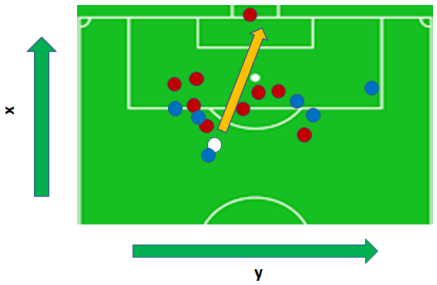
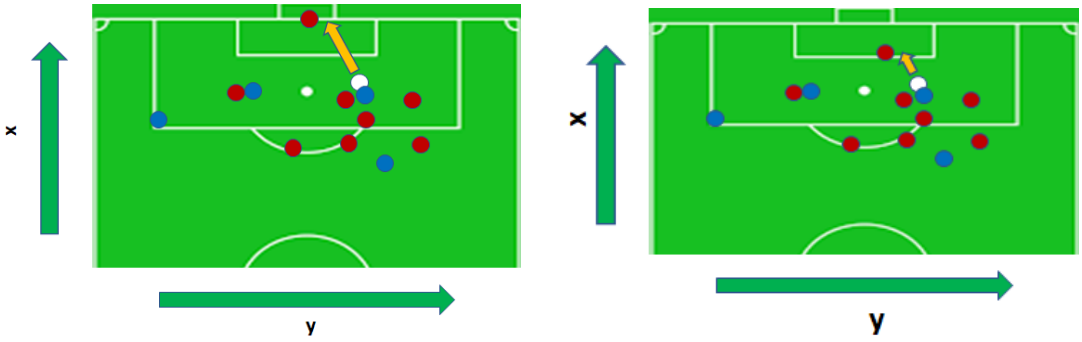


FIGURE 4.10: Real case 1: goal from the distance

TABLE 4.6: The real case 1: expected goal for each framework and different scenarios

Situation	Shooter	GK	xG^* ROSE	xG^* SMOTE	xG Imbl.	xG Understat
Real case 1	Soriano (Bol)	Handanovic (Int)	4.1%	2.3%	2.0%	2.1%
Scenario 1	Ronaldo (Juv)	Handanovic (Int)	5.4%	3.3%	3.1%	2.1%
Scenario 2	Ronaldo (Juv)	Skorupski (Bol)	7.1%	4.4%	4.5%	2.1%

In the second real case (Fig. 4.11a) we propose a goal scored from a low distance, in a very favourable situation (Atalanta vs Brescia, July 2020). Here the alternative scenario considers the goalkeeper position: whereas in the real case he is aligned at his own goal post, in the alternative scenario he is slightly out of goal (Fig. 4.11b). We can see also in the real scenario 2 that our balanced frameworks increase the goal prediction accuracy (higher xG than the benchmark), with SMOTE and imbalance those have a similar xG: as expected (more favourable chance), we have in a higher xG than the real case 1. It's interesting to underline that moving the goalkeeper position outside the box (Scenario 1) increase a lot the xG of the shot.



(A) Real case 2: goal near the box (B) Scenario 1: moving the GK position

FIGURE 4.11: The second real case and its alternative scenario on the pitch

TABLE 4.7: The real case 2: expected goal for each framework and its alternative scenario

Situation	Shooter	GK	xG^* ROSE	xG^* SMOTE	xG Imbl.	xG Understat
Real case 2	Pasalic (Ata)	Andrenacci (Bre)	14.6%	22.1%	21.7%	11.0%
Scenario 1	Pasalic (Ata)	Andrenacci (Bre)	30.0%	58.5%	52.2%	11.0%

4.5 Chapter conclusion

In this chapter is proposed an improvement of the current expected goal (xG) Model, one of the emerging tool in the field of football analytics. The main idea under this model is to assign a quality metric (probability) to goal for each shot. The main idea under this model is to assign a quality metric (probability) to goal for each shot. The main lack of the current xG frameworks is that they take in consideration only the classical event data: in order to overcome this weakness and to customize the model, we integrated some players performance composite indicators obtained from a PLS-SEM approach and some tracking data, such as the goalkeepers position and the number of opponents players. To do this one, we used and merged data relying 53

matches coming from the Italian Serie A, with a final dataset composed by 660 shots and 23 features (1 outcome and 22 regressors). In addition, from a methodological point of view, another important issue that has not taken in consideration until now is the nature of the goal: it is infact a rare event (i.e. 1 goal over 10 shots), as consequence we have an imbalance sample. It has been largely reported that this class imbalance heavily compromises the process of learning, because the model tends to focus on the prevalent class and to ignore the rare event: in order to overcome this limit, we applied ROSE and SMOTE, two balanced algorithms. As baseline framework we adopted the imbalance approach and as benchmark the Understat xG.

We developed a routine applying 5000 times the 3 Fold Cross-Validation, with randomly split every time, using a logistic regression model as learning method; then, we summarized and saved each model performance. As preliminary analysis, we compared the McFadden Pseudo-R^2 of our complete model with a nested one (i.e. the most used in previous literature), proofing a better fitting: the Pseudo-R^2 McFadden difference distribution always showed better performance for the complete model (i.e. it does not contain the zero). The beta coefficients distribution said us that there are different significant regressors for every framework: in particular, for all frameworks are significant x, angle of shot, previous ball distance (Understat features block), whereas we found D2.OpponentsPlayer and the goalkeeper position for what concern the tracking block. Then, also some performance features like movement, GK Mentality and GK Skill are relevant for both frameworks. The approach with the highest number of significant features is the imbalance. Again, by observing the odds ratio, we found that the x regressor has an $OR > 1$ for each framework, proofing that the position to shot is preparatory to goal (i.e. high x position -near the box- means high probability to score a goal); similar situation ($OR > 1$) for GK y coordindate and Movement, two original composite performance indicators. Example of regressor with $OR < 1$ were the angle of shot, emphasizing its protective role to score a goal; comparable circumstance for our original regressors D2.OpponentsPlayer, GK y coordinate and GK mentality.

For what concern the model performance metrics, we initially simulated a scenario with different classification thresholds, for each framework, then we in depth the analysis by using the classical threshold of 0.5 and the equilibrium point for each one. In the first case both ROSE and Imbalance significantly outperform in terms of specificity and precision the benchmark whereas SMOTE improves Understat in terms of sensitivity, F1 and precision. The AUC index is similar for all the frameworks, except for the Imbalance, that significantly outperforms Understat (0.74 vs 0.72). In the second case our frameworks outperformed the benchmark for what concern precision and F1 metrics; the imbalance improved the AUC, too. Finally, we proofed by some real cases that by modifying the players' performance indices and positions on the pitch, the xG changes, increasing the goal prediction probability, that's an innovation for this model.

As summary, the original approach presented in this chapter seems to suggest that some performance indicators (developed during this thesis) and the tracking variables are helpful to detect the goals, improving the benchmark model. As future works, it should be interesting to in-depth this topic by using more data, relying a major number of matches, more football seasons and leagues, and maybe comparing other classification model (for example, Gompit [18]) performances.

Conclusion

As summary, in the thesis a literature review was proposed in Chapter 1, revealing a strong growing related the football analytics production over the last decade; then, in Chapter 2 the focus was pointed versus the creation of some composite performance indicators for football players by a Partial Least Squares Structural Equation Modeling (PLS-SEM) approach: after a preliminary overview of the method in the first part of the chapter, a preliminary Third-Order model was developed as application in the second part. For this purpose, the *sofifa* data have been used, provided from EA Sports experts, which are the ultimate authority on soccer performance measurement, infact they constantly maintain a database of realistic players' performance attributes resulting from careful and systematic data collection; according to the experts, performance variables make up several broader, theoretical dimensions. In this phase, the heterogeneity observed among leagues and players'roles have been evaluated, revealing the second as confounding factor: for this reason different early models have been developed, one for each role. Then, in Chapter 3 each model-role was refined by applying a Confirmatory Tetrad Analysis (CTA), for evaluating statistically the nature of each first order latent construct, and a Confirmatory Composite Analysis (CCA), for removing problematic manifest variables. In a similar way, as in-depth analysis, a PLS-SEM second order model specific for goalkeepers, was developed. A final comparison among each model by role and the full model was provided by several assessment stats: all the indices agreed with the right choice to split players by role. In order to further validate our final composite indicator, the predictive validity was evaluated, computing correlations between the Player Indicator (PI) *overall* with a benchmark (EA *overall*), players' market values and wages.

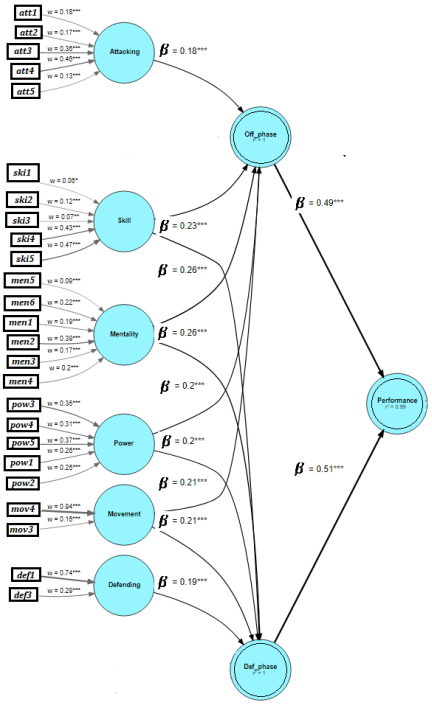
By these chapters the main objective was to create an innovative indicator for players' performance customized for each role, able to capture the different latent traits of the performance; for sure, the first goal was to help football team management (coach, technical staff and scouting), in an impartial evaluation of players. Another advantage derived from this new approach is the possibility to split performance into its sub-areas. Then, the other goal was to integrate those original composite indicators with others tracking variable (Math&Sport) and introduce them as regressors in the expected goal (xG) model (Chapter 4). The classical xG model, based just on event data was tried to be improved by introducing our original variables, applying a logit model with three different frameworks: the imbalance approach and two balanced methods, SMOTE and ROSE. These approaches, compared with a benchmark (the Understat xG), showed important performance, in particular the imbalance approach outperformed the benchmark in terms of specificity, precision and AUC index. In summary, the original approach presented in the last chapter suggested that some performance indicators (created in the previous chapters) and the tracking variables are helpful to detect the goals, improving the benchmark model.

As future steps, for what concern the PLS-SEM approach, it could be interesting

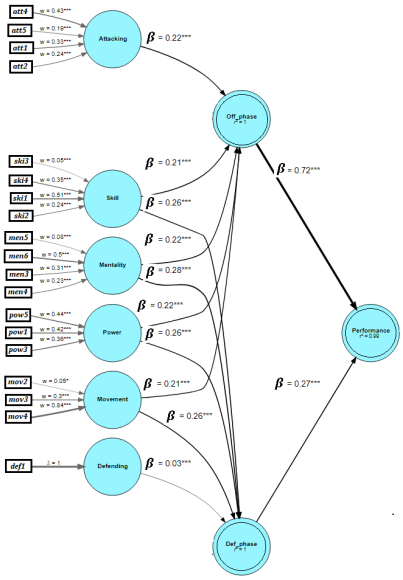
to in-depth both the CTA for higher order constructs and comparing others estimation approaches [40]. Other interesting paths for future research could be, as pointed by experts [82], taking into consideration others heterogeneity observed factors (for example different seasons) maybe using the Measurement Invariance of the COmposite Models (MICOM, [75]). Instead, concerning the xG model, first of all it should be interesting to in-depth this topic by augmenting the shots sample size, in order to investigate if this approach confirm the preliminary results; in addition, focusing on specific tactical situations, such as studying the xG in function of the distance between the goalkeeper and the shooter, or introducing the composite indicators scores for all the players. By the xG model approach based on composites the aim was to refine and improve the existing one, in order to help policy makers and stakeholders for evaluating the teams production during a match, based on the shots quality.

Appendix A

The PLS-SEM Third-Order by role



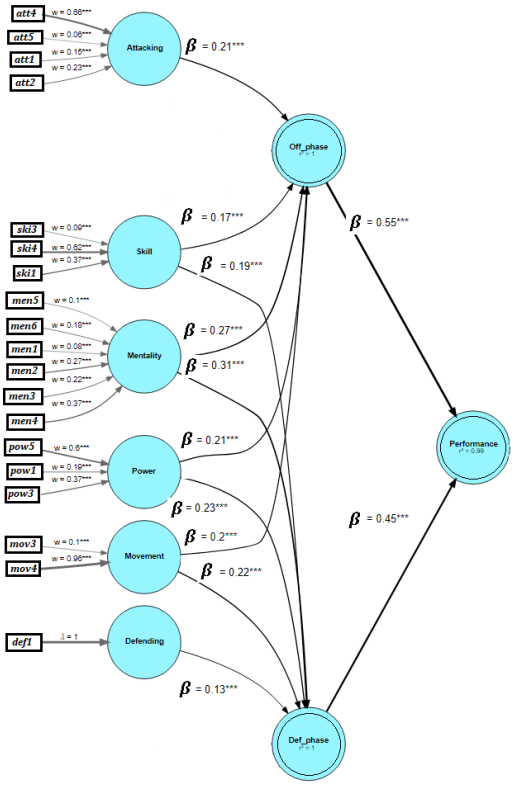
(A) The CB path diagram



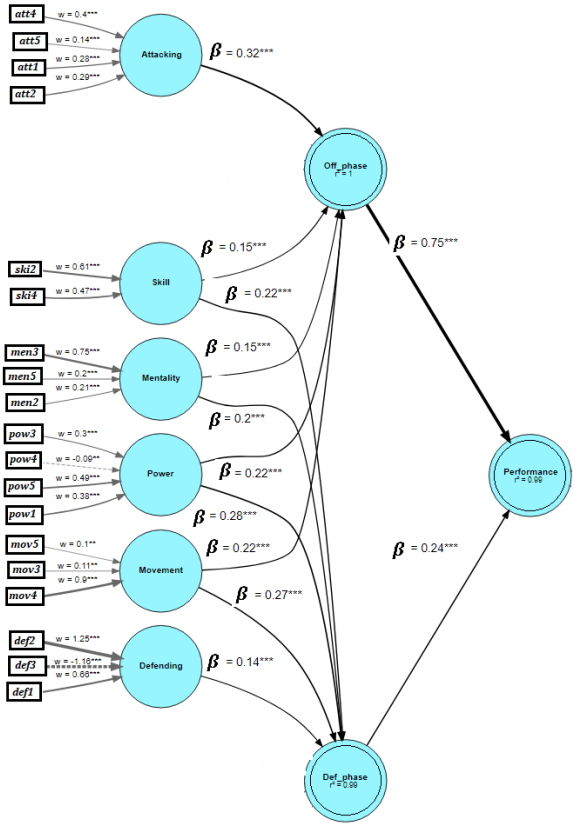
(B) The FB path diagram

FIGURE A1: Path diagram by defensive roles and estimates significant (95% BCa -two tailed- bootstrap CIs with 5000 replications)

Legend: central backs (CB), full backs (FB)



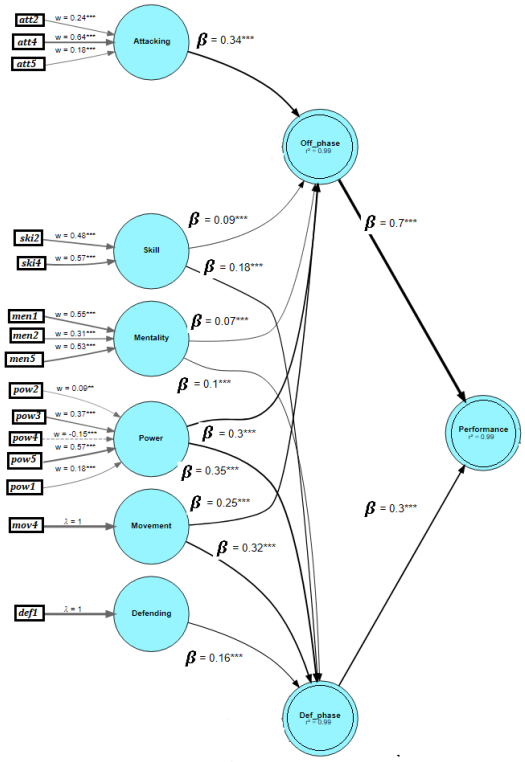
(A) The MF path diagram



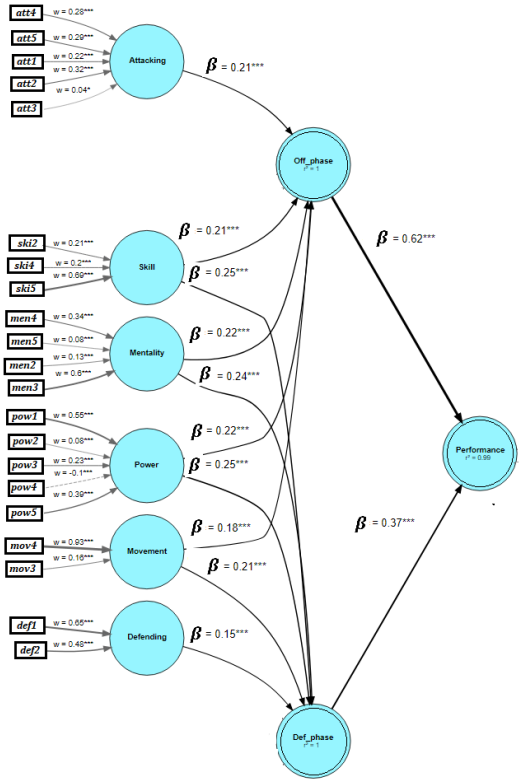
(B) The OM path diagram

FIGURE A2: Path diagram by midfielder roles and estimates significant (95% BCa -two tailed- bootstrap CIs with 5000 replications)

Legend: midfielders (MF), offensive midfielders (OM)



(A) The WG path diagram



(B) The FW path diagram

FIGURE A3: Path diagram by offensive roles and estimates significant (95% BCa -two tailed- bootstrap CIs with 5000 replications)

Legend: wings (WG), forwards (FW)

Bibliography

[1] Gabriel Anzer and Pascal Bauer. "A goal scoring probability model for shots based on synchronized positional and event data in football (soccer)". In: *Frontiers in Sports and Active Living* 3 (2021), p. 53.

[2] Massimo Aria and Corrado Cuccurullo. "bibliometrix: An R-tool for comprehensive science mapping analysis". In: *Journal of Informetrics* 11.4 (2017), pp. 959–975. URL: https://doi.org/10.1016/j.joi.2017.08.007.

[3] Jan-Michael Becker, Kristina Klein, and Martin Wetzels. "Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models". In: *Long range planning* 45.5-6 (2012), pp. 359–394.

[4] Jan-Michael Becker et al. "Discovering unobserved heterogeneity in structural equation models to avert validity threats". In: *MIS quarterly* (2013), pp. 665–694.

[5] Patrizia Belfiore, Antonio Ascione, and Davide Di Palma. "Technology and sport for health promotion: A bibliometric analysis". In: *Journal of Human Sport and Exercise* 10.4 (2019), pp. 932–942.

[6] Iraia Bidaurrazaga-Letona et al. "Identifying talented young soccer players: conditional, anthropometrical and physiological characteristics as predictors of performance.[Identificación de jóvenes talentos en fútbol: características condicionales, antropométricas y fisiológicas como predictores del rendimiento]." In: *RICYDE. Revista Internacional de Ciencias del Deporte. doi: 10.5232/ricyde* 11.39 (2014), pp. 79–95.

[7] Przemyslaw Biecek and Tomasz Burzykowski. *Explanatory model analysis: Explore, explain and examine predictive models*. Chapman and Hall/CRC, 2021.

[8] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[9] William Black and Barry J Babin. "Multivariate data analysis: Its approach, evolution, and impact". In: *The great facilitator*. Springer, 2019, pp. 121–130.

[10] Kenneth A Bollen. "Measurement models: The relation between latent and observed variables". In: *Structural equations with latent variables* (1989), pp. 179–225.

[11] Kenneth A Bollen. "Outlier screening and a distribution-free test for vanishing tetrads". In: *Sociological Methods & Research* 19.1 (1990), pp. 80–92.

[12] Kenneth A Bollen and Kwok-fai Ting. "A tetrad test for causal indicators." In: *Psychological methods* 5.1 (2000), p. 3.

[13] Kenneth A Bollen and Kwok-fai Ting. "Confirmatory tetrad analysis". In: *Sociological methodology* (1993), pp. 147–175.

[14] Denny Borsboom, Gideon J Mellenbergh, and Jaap Van Heerden. "The concept of validity." In: *Psychological review* 111.4 (2004), p. 1061.

[15] Denny Borsboom, Gideon J Mellenbergh, and Jaap Van Heerden. "The theoretical status of latent variables." In: *Psychological review* 110.2 (2003), p. 203.

[16] Lotte Bransen and Jan Van Haaren. "Measuring football players' on-the-ball contributions from passes during games". In: *International Workshop on Machine Learning and Data Mining for Sports Analytics*. Springer. 2018, pp. 3–15.

[17] Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

[18] A Colin Cameron and Pravin K Trivedi. "Microeconometrics with STATA". In: *College Station, TX: StataCorp LP* (2009).

[19] Luciano Canova and Carlo Canepa. "La scienza dei goal: numeri e statistica applicati allo sport più bello del mondo". In: *La scienza dei goal* (2016), pp. 1–174.

[20] Maurizio Carpita, Enrico Ciavolino, and Paola Pasca. "Exploring and modelling team performances of the Kaggle European Soccer database". In: *Statistical Modelling* 19.1 (2019), pp. 74–101.

[21] Maurizio Carpita, Enrico Ciavolino, and Paola Pasca. "Players' role-based performance composite indicators of soccer teams: A statistical perspective". In: *Social Indicators Research* (2020), pp. 1–16.

[22] Maurizio Carpita and Silvia Golia. "Discovering associations between players' performance indicators and matches' results in the European Soccer Leagues". In: *Journal of Applied Statistics* (2020), pp. 1–16.

[23] Maurizio Carpita et al. "Discovering the drivers of football match outcomes with data mining". In: *Quality Technology & Quantitative Management* 12.4 (2015), pp. 561–577.

[24] Rosanna Cataldo et al. "Developments in Higher-Order PLS-PM for the building of a system of Composite Indicators". In: *Quality & Quantity* 51.2 (2017), pp. 657–674.

[25] Mattia Cefis. "Football analytics: a bibliometric study about the last decade contributions". In: *Electronic Journal of Applied Statistical Analysis* 15.1 (2022), pp. 232–248.

[26] Mattia Cefis. "Observed heterogeneity in players' football performance analysis using PLS-PM". In: *Journal of Applied Statistics* 0.0 (2022), pp. 1–20. DOI: 10.1080/02664763.2022.2101044.

[27] Mattia Cefis. "Observed heterogeneity in players' football performance analysis using PLS-PM". In: *Journal of Applied Statistics* (2022), pp. 1–20.

[28] Mattia Cefis and Eugenio Brentari. "Formative vs Reflective constructs: a CTA-PLS approach on a goalkeepers' performance model". In: *Book of the Short Papers, 51st Scientific Meeting of the Italian Statistical Society* (2022), pp. 323–328.

[29] Mattia Cefis and Maurizio Carpita. "A New xG Model for Football Analytics". In: *Journal of the Operational Research Society* (2022).

[30] Mattia Cefis and Maurizio Carpita. "Football analytics: a Higher-Order PLS-SEM approach to evaluate players' performance". In: *Book of Short Papers SIS 2021* (2021), pp. 508–513.

[31] Mattia Cefis and Maurizio Carpita. "Football Analytics: performance analysis differentiate by role". In: *Third international conference on Data Science & Social Research -Book of Abstracts*. 2020, p. 22.

[32] Mattia Cefis and Maurizio Carpita. "PLS-SEM with CCA for football goalkeeper's performance indicators". In: *Book of Short Papers IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment* (2022), pp. 288–293.

[33] Mattia Cefis and Maurizio Carpita. "The Higher-Order PLS-SEM Confi rmatory Approach for Composite Indicators of Football Performance Quality". In: *Computational Statistics* (2022), pp. 1–24. DOI: 10.1007/s00180-022-01295-4. URL: https://doi.org/10.1007/s00180-022-01295-4.

[34] Mattia Cefis and Maurizio Carpita. "The Higher-Order PLS-SEM Confi rmatory Approach for Composite Indicators of Football Performance Quality". In: *Submitted at Journal of Computational Statistics* (2022).

[35] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[36] Jun-Hwa Cheah et al. "A comparison of five reflective–formative estimation approaches: reconsideration and recommendations for tourism research". In: *Quality & Quantity* 53.3 (2019), pp. 1421–1458.

[37] Jun-Hwa Cheah et al. "Sampling weight adjustments in partial least squares structural equation modeling: guidelines and illustrations". In: *Total Quality Management & Business Excellence* 32.13-14 (2021), pp. 1594–1613.

[38] Wynne W Chin et al. "The partial least squares approach to structural equation modeling". In: *Modern methods for business research* 295.2 (1998), pp. 295–336.

[39] Enrico Ciavolino. "General distress as second order latent variable estimated through PLS-PM approach". In: *Electronic Journal of Applied Statistical Analysis* 5.3 (2012), pp. 458–464.

[40] Enrico Ciavolino and Mariangela Nitti. "Simulation study for PLS path modelling with high-order construct: A job satisfaction model evidence". In: *Advanced dynamic modeling of economic and social systems*. Springer, 2013, pp. 185–207.

[41] Enrico Ciavolino and Mariangela Nitti. "Using the hybrid two-step estimation approach for the identification of second-order latent variable models". In: *Journal of Applied Statistics* 40.3 (2013), pp. 508–526.

[42] Enrico Ciavolino et al. "A confirmatory composite analysis for the Italian validation of the interactions anxiousness scale: a higher-order version". In: *Behaviormetrika* 49.1 (2022), pp. 23–46.

[43] Enrico Ciavolino et al. "A tale of PLS Structural Equation Modelling: Episode I—A Bibliometrix Citation Analysis". In: *Social Indicators Research* (2022), pp. 1–26.

[44] Dennis Coates and Petr Parshakov. "The wisdom of crowds and transfer market values". In: *European Journal of Operational Research* 301.2 (2022), pp. 523–534.

[45] Manuel J Cobo et al. "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field". In: *Journal of informetrics* 5.1 (2011), pp. 146–166.

[46] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[47] Tim Coltman et al. "Formative versus reflective measurement models: Two applications of formative measurement". In: *Journal of Business Research* 61.12 (2008), pp. 1250–1262.

[48] Joint Research Centre-European Commission et al. *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing, 2008.

[49] Corrado Crocetta et al. "Higher-order PLS-PM approach for different types of constructs". In: *Social Indicators Research* 154.2 (2021), pp. 725–754.

[50] Lee J Cronbach and Paul E Meehl. "Construct validity in psychological tests." In: *Psychological bulletin* 52.4 (1955), p. 281.

[51] Geoff Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2013.

[52] Andrea Dal Pozzolo et al. "Calibrating probability with undersampling for unbalanced classification". In: *2015 IEEE symposium series on computational intelligence*. IEEE. 2015, pp. 159–166.

[53] Nicholas P Danks and Soumya Ray. "Predictions from partial least squares models". In: *Applying partial least squares in tourism and hospitality research*. Emerald Publishing Limited, 2018.

[54] Selçuk Demir and Emrehan Kutluğ Şahin. "Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes". In: *Avrupa Bilim ve Teknoloji Dergisi* 34 (2022), pp. 142–147.

[55] Adamantios Diamantopoulos and Heidi M Winklhofer. "Index construction with formative indicators: An alternative to scale development". In: *Journal of marketing research* 38.2 (2001), pp. 269–277.

[56] Theo K Dijkstra and Jörg Henseler. "Consistent partial least squares path modeling". In: *MIS quarterly* 39.2 (2015), pp. 297–316.

[57] Tiziana D'Isanto et al. "Assessment of sport performance: Theoretical aspects and practical indications". In: *Sport Mont* 17.1 (2019), pp. 79–82.

[58] Vincenzo Esposito Vinzi et al. "REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling". In: *Applied Stochastic Models in Business and Industry* 24.5 (2008), pp. 439–458.

[59] Alexander Fairchild, Konstantinos Pelechrinis, and Marios Kokkodis. "Spatial analysis of shots in MLS: a model for expected goals and fractal dimensionality". In: *Journal of Sports Analytics* 4.3 (2018), pp. 165–174.

[60] Cristoforo Filetti et al. "A study of relationships among technical, tactical, physical parameters and final outcomes in elite soccer matches as analyzed by a semiautomatic video tracking system". In: *Perceptual and Motor Skills* 124.3 (2017), pp. 601–620.

[61] Claes Fornell and David F Larcker. "Evaluating structural equation models with unobservable variables and measurement error". In: *Journal of marketing research* 18.1 (1981), pp. 39–50.

[62] Seymour Geisser. "A predictive approach to the random effect model". In: *Biometrika* 61.1 (1974), pp. 101–107.

[63] Sam Green. "Assessing the Performance of Premier Leauge Goalscorers". In: *OptaPro Blog* (2012). URL: `http://www.optasportspro.com`.

[64] Siegfried P Gudergan et al. "Confirmatory tetrad analysis in PLS path modeling". In: *Journal of business research* 61.12 (2008), pp. 1238–1249.

[65] Carsten Hahn et al. "Capturing customer heterogeneity using a finite mixture PLS approach". In: *Schmalenbach Business Review* 54.3 (2002), pp. 243–269.

[66] Joe F Hair, Christian M Ringle, and Marko Sarstedt. "PLS-SEM: Indeed a silver bullet". In: *Journal of Marketing theory and Practice* 19.2 (2011), pp. 139–152.

[67] Joseph F Hair. "Multivariate data analysis". In: (2009).

[68] Joseph F Hair, Christian M Ringle, and Marko Sarstedt. "Partial least squares: the better approach to structural equation modeling?" In: *Long Range Planning* 45.5-6 (2012), pp. 312–319.

[69] Joseph F Hair, Marko Sarstedt, and Christian M Ringle. "Rethinking some of the rethinking of partial least squares". In: *European Journal of Marketing* (2019).

[70] Joseph F Hair et al. "Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modeling methods". In: *Journal of the Academy of Marketing Science* 45.5 (2017), pp. 616–632.

[71] Joseph F Hair et al. "When to use and how to report the results of PLS-SEM". In: *European business review* (2019).

[72] Joe F Hair Jr, Matt C Howard, and Christian Nitzl. "Assessing measurement model quality in PLS-SEM using confirmatory composite analysis". In: *Journal of Business Research* 109 (2020), pp. 101–110.

[73] Joe F Hair Jr et al. "PLS-SEM or CB-SEM: updated guidelines on which method to use". In: *International Journal of Multivariate Data Analysis* 1.2 (2017), pp. 107–123.

[74] Joseph F Hair Jr et al. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications, 2016.

[75] Joseph F Hair Jr et al. *Advanced issues in partial least squares structural equation modeling*. saGe publications, 2017.

[76] Joseph F Hair Jr et al. *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook*. 2021.

[77] Jörg Henseler, Christian M Ringle, and Marko Sarstedt. "A new criterion for assessing discriminant validity in variance-based structural equation modeling". In: *Journal of the academy of marketing science* 43.1 (2015), pp. 115–135.

[78] Jörg Henseler, Christian M Ringle, and Marko Sarstedt. "Testing measurement invariance of composites using partial least squares". In: *International marketing review* (2016).

[79] Jörg Henseler and Marko Sarstedt. "Goodness-of-fit indices for partial least squares path modeling". In: *Computational statistics* 28.2 (2013), pp. 565–580.

[80] Jörg Henseler et al. "Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013)". In: *Organizational research methods* 17.2 (2014), pp. 182–209.

[81] Mohammad Hossin and Md Nasir Sulaiman. "A review on evaluation metrics for data classification evaluations". In: *International journal of data mining & knowledge management process* 5.2 (2015).

[82] Michael David Hughes et al. "Moneyball and soccer-an analysis of the key performance indicators of elite male soccer players by position". In: (2012).

[83] Heungsun Hwang and Yoshio Takane. "Generalized structured component analysis". In: *Psychometrika* 69.1 (2004), pp. 81–99.

[84] Heungsun Hwang et al. "A concept analysis of methodological research on composite-based structural equation modeling: bridging PLSPM and GSCA". In: *Behaviormetrika* 47.1 (2020), pp. 219–241.

[85] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[86] Karl G Jöreskog. "Analysis of covariance structures". In: *Multivariate analysis–III*. Elsevier, 1973, pp. 263–285.

[87] David Kaplan. *Structural equation modeling: Foundations and extensions*. Vol. 10. Sage Publications, 2008.

[88] Gohar F Khan et al. "Methodological research on partial least squares structural equation modeling (PLS-SEM): an analysis based on social network approaches". In: *Internet Research* (2019).

[89] Thomas Kirschstein and Steffen Liebscher. "Assessing the market values of soccer players- A robust analysis of data from German 1. and 2. Bundesliga". In: *Journal of Applied Statistics* 46.7 (2019), pp. 1336–1349.

[90] Michael Klesel et al. "A test for multigroup comparison using partial least squares path modeling". In: *Internet research* 29 (2019), pp. 464–477.

[91] Adrian Leguina. *A primer on partial least squares structural equation modeling (PLS-SEM)*. 2015.

[92] Jan-Bernd Lohmöller. "Predictive vs. structural modeling: Pls vs. ml". In: *Latent variable path modeling with partial least squares*. Springer, 1989, pp. 199–226.

[93] Samuel López-Carril et al. "The rise of social media in sport: a bibliometric analysis". In: *International Journal of Innovation and Technology Management* 17.06 (2020), p. 2050041.

[94] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.

[95] Francesca Matano et al. "Augmenting adjusted plus-minus in soccer with FIFA ratings". In: *arXiv preprint arXiv:1810.08032* (2018).

[96] Kate McCarthy, Bibi Zabar, and Gary Weiss. "Does cost-sensitive learning beat sampling for classifying rare classes?" In: *Proceedings of the 1st international workshop on Utility-based data mining*. 2005, pp. 69–77.

[97] Daniel McFadden. "Quantitative methods for analysing travel behaviour of individuals: some recent developments". In: *Behavioural travel modelling*. Routledge, 2021, pp. 279–318.

[98] Ian G McHale, Philip A Scarf, and David E Folker. "On the development of a soccer player performance rating system for the English Premier League". In: *Interfaces* 42.4 (2012), pp. 339–351.

[99] Mehmet Mehmetoglu and Sergio Venturini. *Structural equation modelling with partial least squares using Stata and R*. CRC Press, 2021.

[100] Giovanna Menardi and Nicola Torelli. "Training and assessing classification rules with imbalanced data". In: *Data mining and knowledge discovery* 28.1 (2014), pp. 92–122.

[101] Armin Monecke and Friedrich Leisch. "semPLS: structural equation modeling using partial least squares". In: (2012).

[102] Luca Pappalardo et al. "PlayeRank: data-driven performance evaluation and player ranking in soccer via a machine learning approach". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.5 (2019), pp. 1–27.

[103] Alex Rathke. "An examination of expected goals and shot efficiency in soccer". In: *Journal of Human Sport and Exercise* 12.2 (2017), pp. 514–529.

[104] Soumya Ray and Nicholas Danks. "January 18, 2018". In: *dim (mobi)* 1.250 (2018), p. 24.

[105] Andrea Riboli et al. "Effect of formation, ball in play and ball possession on peak demands in elite soccer". In: *Biology of Sport* 38.2 (2021), p. 195.

[106] Christian Ringle, Dirceu Da Silva, and Diógenes Bido. "Structural equation modeling with the SmartPLS". In: *Brazilian Journal Of Marketing* 13.2 (2015).

[107] Christian M Ringle, Sven Wende, Jan-Michael Becker, et al. "SmartPLS 3". In: *Boenningstedt: SmartPLS GmbH* 584 (2015).

[108] Christian M Ringle, Sven Wende, and Alexander Will. *SmartPLS 2.0 (beta)*. 2005.

[109] Pieter Robberechts and Jesse Davis. "How data availability affects the ability to learn good xG models". In: *International Workshop on Machine Learning and Data Mining for Sports Analytics*. Springer. 2020, pp. 17–27.

[110] Mikko Rönkkö. *Introduction to matrixpls*. 2016.

[111] Mikko Rönkkö et al. "Partial least squares path modeling: Time for some serious second thoughts". In: *Journal of Operations Management* 47 (2016), pp. 9–27.

[112] Lingfeng Ruan et al. "Quantifying the effectiveness of defensive playing styles in the Chinese Football Super League". In: *Frontiers in Psychology* (2022), p. 2461.

[113] Gaston Sanchez et al. "Package 'plspm'". In: *Citeseer: State College, PA, USA* (2013).

[114] Marko Sarstedt, Jörg Henseler, and Christian M Ringle. "Multigroup analysis in partial least squares (PLS) path modeling: Alternative methods and empirical results". In: *Measurement and research methods in international marketing*. Vol. 22. Emerald Group Publishing, 2011, pp. 195–218.

[115] Marko Sarstedt, Erik Mooi, et al. "A concise guide to market research". In: *The Process, Data, and* 12 (2014).

[116] Marko Sarstedt and Christian M Ringle. "Treating unobserved heterogeneity in PLS path modeling: a comparison of FIMIX-PLS with different data analysis strategies". In: *Journal of Applied Statistics* 37.8 (2010), pp. 1299–1318.

[117] Marko Sarstedt et al. "How to specify, estimate, and validate higher-order constructs in PLS-SEM". In: *Australasian Marketing Journal (AMJ)* 27.3 (2019), pp. 197–211.

[118] Marko Sarstedt et al. "On the emancipation of PLS-SEM: A commentary on Rigdon (2012)". In: *Long range planning* 47.3 (2014), pp. 154–160.

[119] Florian Schuberth, Jörg Henseler, and Theo K Dijkstra. "Confirmatory composite analysis". In: *Frontiers in psychology* 9 (2018), p. 2541.

[120] Steven R Schultze and Christian-Mathias Wellbrock. "A weighted plus/minus metric for individual soccer player performance". In: *Journal of Sports Analytics* 4.2 (2018), pp. 121–131.

[121] Galit Shmueli et al. "Predictive model assessment in PLS-SEM: guidelines for using PLSpredict". In: *European Journal of Marketing* (2019).

[122] Galit Shmueli et al. "The elephant in the room: Predictive performance of PLS models". In: *Journal of Business Research* 69.10 (2016), pp. 4552–4564.

[123] Rudolf R Sinkovics et al. "Testing measurement invariance of composites using partial least squares". In: *International marketing review* (2016).

[124] Manuel Stein et al. "Where to go: Computational and visual what-if analyses in soccer". In: *Journal of sports sciences* 37.24 (2019), pp. 2774–2782.

[125] Mervyn Stone. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 111–133.

[126] Saundra M Tabet et al. "An analysis of the world health organization disability assessment schedule 2.0 measurement model using partial least squares–structural equation modeling". In: *Assessment* 27.8 (2020), pp. 1731–1747.

[127] Thompson SH Teo, Shirish C Srivastava, and LI Jiang. "Trust and electronic government success: An empirical study". In: *Journal of management information systems* 25.3 (2008), pp. 99–132.

[128] Izzatul Umami, Deden Hardan Gautama, and Heliza Rahmania Hatta. "implementing the Expected Goal (xG) model to predict scores in soccer matches". In: *International Journal of Informatics and Information Systems* 4.1 (2021), pp. 38–54.

[129] G. Vigneshwaran and R. Kalidasan. "STUDY OF PUBLICATIONS OUTPUT ON SPORTS SCIENCE–A BIBLIOMETRIC ANALYSIS". In: *Ganesar College of Arts and Science* (2018), pp. 256–260.

[130] Vincenzo Esposito Vinzi, Laura Trinchera, and Silvano Amato. "PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement". In: *Handbook of partial least squares*. Springer, 2010, pp. 47–82.

[131] Bradley Wilson. "Using PLS to investigate interaction effects between higher order branding constructs". In: *Handbook of partial least squares*. Springer, 2010, pp. 621–652.

[132] Herman Wold. "A. 1985. Partial least squares". In: *Kotz S, Johnson N L. Encyclopedia of Statistical Sciences. New York: Wiley* (1985), pp. 581–591.

[133] Herman Wold. "Encyclopedia of statistical sciences". In: *Partial least squares. Wiley, New York* (1985), pp. 581–591.

[134] Herman Wold. "Estimation of principal components and related models by iterative least squares". In: *Multivariate analysis* (1966), pp. 391–420.

[135] Herman Wold. "Model construction and evaluation when theoretical knowledge is scarce: Theory and application of partial least squares". In: *Evaluation of econometric models*. Elsevier, 1980, pp. 47–74.

[136] Herman Wold. "Nonlinear iterative partial least squares (NIPALS) modelling: some current developments". In: *Multivariate analysis–III*. Elsevier, 1973, pp. 383–407.

[137] Herman Wold. "Soft modeling: the basic design and some extensions". In: *Systems under indirect observation* 2 (1982), p. 343.