From Training to Match Performance: An Exploratory and Predictive Analysis on F.C. Barcelona GPS Data

by

Javier Fernández

Submitted to the Facultat d'Informàtica de Barcelona
in partial fulfillment of the requirements for the degree of

Master of Science in Artificial Intelligence

at the

Universitat Politécnica de Catalunya

October 2016

Author .....
Facultat d'Informàtica de Barcelona
October 15th, 2016

Certified by .....
Marta Arias and Ricard Gavalà
Thesis Supervisors

From Training to Match Performance: An Exploratory and Predictive Analysis on F.C. Barcelona GPS Data

by

Javier Fernández

Submitted to the Facultat d'Informàtica de Barcelona
on October 15th, 2016, in partial fulfillment of the
requirements for the degree of
Master of Science in Artificial Intelligence

Abstract

FIFA has recently allowed the use of electronic performance and tracking systems (EPTS) in professional football competition, providing teams with novel and more accurate data, regarding physical player performance. The analysis of this kind of information will provide teams with competitive advantages, by gaining a deeper understanding of the relation between training and match load, and individual player's fitness characteristics. In order to make sense of this physical data, which is inherently complex, machine learning algorithms that exploit both non-linear and linear relations among variables could be of great aid on building predictive and explanatory models. This study provides a methodology based on machine learning and statistical methods to relate the physical performance players during training sessions, and their performance in the following matches. The analysis is carried out over F.C. Barcelona B, season 2015-2016 data. The study is structured in four main phases. The first phase is based on data collection and processing in order to generate datasets suited to the application of artificial intelligence algorithms. A second exploratory phase provides a in-deep analysis of the characteristics of the data that allows to validate its quality and physical coaches main believes. Then, two phases consisting in unsupervised and supervised analysis are carried out. The first one approaches relations between training adaptability through variations and further match performance, through the use of cluster analysis, in time-based data. The second one is based on predicting future match physical variables through the application of linear and non-linear learning algorithms. The study has found remarkable relations between training variations and match performance, as well as able to predict 11 of 17 physical variables, along proposing a practical metric for regression analysis. Multiple believes from football world have been validated, and a new schema for structuring these variables have been proposed.

Acknowledgments

I would like to thank Daniel Medina whose dedication on finding paths to merge data analytics into the professional sports world made this work possible. Special thanks to Antonio Gómez who patiently transmitted an invaluable perspective from the sports and physical conditioning world, and dedicated great time in helping to reach the most practical results possible. Special mention to physical coaches Eduard Pons and Rafel Pol, who contributed with remarkable insights along the way, which drove great part of this project. Thanks also to F.C. Barcelona for providing a creative and collaborative culture which makes it a delightful place to work. Also, to Raúl Peláez and Marc Subira, who kindly supported this work with insights and ideas. I am very thankful to Marta Arias and Ricard Gavaldá who believed in this project and where a crucial part for making it even possible to start. Their great advice and creative freedom were an important fuel for reaching value. Thanks to the Universitat Politécnica de Catalunya (UPC) for supporting the project and keeping a high standard which makes the world to invest in their name. And more remarkably, thanks to my wife Gaby whose support and energy is the best ally anyone can have.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This sections presents the main motivations that drive the design and development of this study, focused on understanding relations between training and match physical performance in professional football. The thesis outline is further presented as a guide of how to read and approach this work, and finally the most relevant background studies are commented.

1.1 Motivation and Purpose

Professional football has attracted the attention of the data science community in the last decade due to the increasing availability of quantitative data. Currently, multiple types of information can be gathered directly or indirectly from both official matches and training sessions, providing physical performance data, tactical characteristics of the game, medical data and even genetic information of players. The numerous types of metrics and in-game detailed events has contributed to the improvement of critical tasks such as team tactics evaluation, opponent analysis, player scouting and training design [23, 27]. Given this scenario, the application of in-depth analysis of this data is believed to provide a significant competitive advantage in the following years [2]. However, the use of advanced data mining techniques on this data, in order to detect complex patterns that might bring a deeper understanding of the game, is still in its initial steps of development.

One of the cornerstones for reaching competitive performance in sports is the process and evolution of training, both in its physical and tactical components, among others. Players and coaches devote at least half of the weekly training time on physical conditioning, in order to ensure the team is reaching adequate levels of fitness. It is noticeable, however, that few of the current studies are oriented to the analysis of physical information of the players. This is due mainly to the difficulty of having access to this data through training and competition, which is considered highly valued by football clubs [40]. Typically, such information is gathered through the use of electronic performance and tracking systems (EPTS) which include GPS and microsensor technology such as accelerometers, gyroscopes and magnetometers. Collecting this information was not allowed during official football competition until the recent authorization of the Football Association Board (IFAB), for the 2015-2016 season [30]. These devices have been increasingly adapted and accepted in sports such as Rugby, Australian football, Cricket and Hockey [9]. Despite some concerns over the reliability of GPS measurement of accelerations, especially at low sample rates, it has been an important parameter for analysing the activity profile in team sports [38].

At F.C. Barcelona these tools have been used for monitoring load and many other physical variables at training sessions in the last four years, and this season, for first time, at official competition. These EPTS are aiding the evaluation of the applied training methodology, the *structured training* [34], a system that sets the baselines for the planning and adaptation of the training activities along the season. Within 3 weeks periodization frames [31], physical coaches design strategies to induce player adaptation taking into account training activities and the competition, considering the latter the most relevant stimulus to optimize the athlete's capabilities. The information that is provided by EPTS devices becomes then highly important to analyse the physical demands of the sessions and the performance of both individual players and the team as a whole. However, this also presents to coaches a wide set of new variables, most of which were not previously quantified, that need to be understood

and incorporated within the weekly design and analysis process. Also, the availability of matches data provides the opportunity to relate physical performance during competition and training, guiding a more fine-grained design of player adaptation, and adding information for better understanding of each player's fitness profile.

Beyond the availability of new data, it becomes essential that efforts to analyse and make sense of this data can be translated into practice. As proposed by Aaron J. Coutts, the laborious and slow-paced research effort based on robust and detailed analysis, must be able to produce findings and results that can be applied by fast-working practitioners [21], which commonly act (and need to act) quickly, intuitively and emotionally. Latest EPTS devices provide over a hundred variables that aim to quantify the different physical efforts and responses of players. However, this amount of information makes infeasible for physical coaches to perform a one-to-one variable analysis in a frequent basis and be able to reach conclusions quickly. This opens the door for statistical analysis for exploring the relations among variables, understanding which are more informative, and providing mechanisms for simplifying the fast-paced periodical analysis.

The key role of physical conditioning for performance, and the availability of training and matches data, provides an interesting research question on whether there exist significant relations between physical performance of players during training and the measured performance in subsequent matches. And more specifically, if the patterns and characteristics that may arise from this relations, can be simplified and expressed in terms that allow to acknowledge them and apply them in practice. For this matter, this study focuses on the analysis of these relations between training and matches physical information, from a data mining-driven approach. Both exploratory and predictive analysis are performed in order to answer different questions related to the main research objective, using F.C. Barcelona B team training and match physical performance data from season 2015-2016.

From an exploratory point of view, data is first transformed, plotted, and analysed in

order to understand its main characteristics and validate some of the most relevant observations of physical coaches regarding the quantitative and qualitative aspects of the data. Then, time series related analysis and unsupervised learning methods are applied in order to answer a related question: does the variation of physical values in a weekly basis have a relation with the forthcoming match performance?. This idea derives from the physiological characteristics of physical performance, which present oscillating patterns and are key to understanding the process of adaptation of the player to the training stimulus. This is highly related to the idea of *deterministic chaos* present in biological systems [1], due to which players are expected to evidence different adaptational behaviors along the season trainings. Finding relations between these variations and match performance, might provide new knowledge for understanding how to quantify and evaluate each player's optimal range of performance, commonly known as its optimal fitness profile.
A predictive analysis is also carried out on this data, with the purpose of understanding up to which extent is possible to estimate or predict future physical performance in matches given training performance. Also, for variables which its predictive power is considered acceptable, this part of studies aims to understand which other variables have the biggest influence for explaining the former. The main idea is to be able to aid physical coaches to understand the current physical state of players from historical data, and understanding which variables are most significant for explaining others. Machine learning algorithms that exploit either linear or non-linear relations among variables are applied, within regression analysis, while feature selection methods are also applied to refine the prediction power and improve understandability. Also, a specific metric of regression quality is proposed, which is considered to be better fit to the practical requirements of the sports world.

In order to contextualize and ease the analysis of this type data, three main categories are proposed that group variables together regarding the origin of their measurements and their nature. These groups: locomotor, mechanical, and metabolic, provide a higher-level categorization that is expected to orient future analysis on

EPTS physical performance data. This is the first study, up to our knowledge, to relate training and match physical values directly registered from player using EPTS devices during training and matches for a whole season, in professional football. Some of the different findings presented in this study are to be applied at F.C. Barcelona during season 2016-2017 and the upcoming ones, for aiding the player assessment process on training.

1.2 Derived Publications

From the results of this study two papers have been derived and accepted for presentation and publication, as detailed below. Both papers are attached at the end of this report.

- Fernandez J., Medina D., Gomez A., Arias M and Gavaldà R. From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data. Presented at European Conference of Machine Learning (ECML-PKDD), Sports Analytics Workshop. September 2016 [32].

- Fernandez J., Medina D., Gomez A., Arias M and Gavaldà R. From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data. To be presented at International Conference on Data Mining (ICDM), Data Mining for the Analysis of Performance and Sucess Workshop. December 2016 [33].

1.3 Thesis Outline

From a methodological point of view this study is divided in four main phases, which correspond from the initial collection and processing of data, to the application of exploratory and predictive analysis on the processed data. The different sections of this report are, in turn, structured to reflect the information that corresponds to each of these phases. In Chapter 1 the motivation and purpose of the study is discussed and

initially sketched, while also presenting the most relevant background studies related to sports analytics. Then, Chapter 2 outlines the theoretical background information that is required to understand both the data mining related techniques applied and some key concepts from sports science and F.C. Barcelona training structure. Chapter 3 presents a complete methodological description of the four different phases in which the study has been split. In each subsection, all the transformations, algorithms and methods applied are explained in detail, in order to gain a technical overview of the proposed approach for each phase. Afterwards, experiments are presented in Chapter 4, describing specifically the characteristics of the datasets used in each experiment, and the statistical methods, tables and plots to be used in the results section to support the experiments. Then, Chapter 5 presents the obtained results where most relevant observations and findings are highlighted. Finally, Chapter 6 presents detailed conclusions from each of the phases of this study, as well as overall remarking observations. The initial hypothesis or research questions are addressed in contrast with the final conclusions. Suggestions for future work are also presented at the end of Chapter 6.

1.4 Related Work

Last decade has presented a considerably increase in sports analytics-related research. This is also the case in football given the recent rise of different types of information sources, which include video-tracking data, match events tagging and physical performance measurement. From this information, multiple type of analysis have been developed, including tactical analysis, performance prediction, injury prediction and physical performance analysis, among others.

Electronic performance and tracking systems (EPTS) which are composed by global position systems (GPS), accelerometer, gyroscope and a magnetometer, and are carried directly by players, have been commonly used by many professional teams during the last 5 years for training sessions. From this data the physical demands of

football related training activities has started to be quantified and analysed in professional football [45]. Through use of additional sensors such as heart rate meter and EPTS data, the relations between training load and physical responses of players has been studied [13]. This has lead to more refined understanding of physical demands of training thorough the comparison of different types of physical variables such as distance, speeds at different rates and metabolic energy expenditure, taking into account long-term training design [35]. The quality of physical variables coming from EPTS data, in particular those related with GPS measurements, has been analysed through multiple devices and software, finding out that data collected with different models might vary greatly [8]. A deeper study suggests that the search for standards in this measurements is critical to properly orient research, and particularly to translate the findings effectively into practice [29]. The recent approval from FIFA of using this devices during official competition has provided the possibility of gathering data from the same devices and using the same procedure, so ensuring a more proper analysis of training and match data.

Another recurrent task on modern sports analytics is the application of predictive model for estimating future behaviour or performance. Using both training and matches data, and based in heart rate variability, two recent studies have provided predictive models for estimating physical match performance in the Australian Football sport [20, 36]. Both linear and non-linear models are applied, and different feature selection methods are tested in order to model inter-variable relations. However, results are provided in terms of the r-squared metric, which is harder to interpret and translate into practice. Predictive models have also been approached for National Basketball League (NBA) injury analysis [49] through the application of the Random Forest algorithm for predicting injuries based on time-windowed historical data. Other studies have also focused on statistical analysis of player injury based on physical player and team performance [4], however is recognized as a complex and highly variable subject, that requires to deal with highly dynamic and complex system [19] which might still lack from appropriate or sufficient data.

Beyond the physical performance area, tactical analysis covers a great part of current research, due to the increasing availability of event and tracking data. Most of these studies focus on highly detailed pattern detection, based on temporal data an common events such as passes [6], player movements and coverage of space [14], and even complex interactions to define unique playing styles [26]. A great part of this research is based on several definitions of performance and success, which is in continuous evolution and debate. Performance of teams have been analysed by quantifying players individual performance through statistical models and match results [24], using a wide set of match general statistics for correlation and distribution analysis [15], and relating known statistical distribution to passing patterns and team success [16]. Also, tactical patterns have been approached through complex networks analysis to understand team interactions for different types of events, and provide better visualization capabilities to the findings [7, 17, 53].

Chapter 2

Background

This section presents the most relevant algorithms, methods and concepts related to the methodological approach presented in this study, and its implementation.

2.1 Unsupervised Learning Algorithm and Methods

2.1.1 Dynamic Time Warping

Dynamic time warping (DTW) allows to measure the similarity between two temporal series, while being less sensitive to signal transformations as shifting, uniform amplitude scaling or uniform time scaling [46]. Its main purpose is to find an optimal alignment between two time-dependent sequences which are warped in a nonlinear fashion to match each other. [43]. Since its able to find appropriate warping path between time-series of different lengths, it has become an effective and rather simple approach for dealing with time-dependent data. It has been applied successfully in a variety of problems such as speech recognition, signature recognition, financial stock classification, and many other fields within data mining. The application of DTW produces a distance or dissimilarity matrix for each pair of series in the dataset. In order to calculate specific distances between data-points (within each series) any appropriate distance measure can be applied, becoming a flexible tool that can adjust to similarity measures specifically designed for data. The most commonly used dis-

tance is Euclidean, when leaning to favor vector magnitudes, while Angular distance is used when angular orientation of data vectors becomes more relevant. Providing the possibility to find a dissimilarity matrix between series, requiring it has become a powerful tool for classification and clustering of time-based data.

2.1.2 Cluster Analysis

Cluster analysis deals with the problem of finding natural groups among examples in data, which represent underlying, and hopefully sufficiently significant, differences in data [39]. The idea of obtaining representative concepts within data provides a wide scope of alternatives, and frequently requires to add to the analysis insightful observations based on expert knowledge. The approach to find these clusters, is carried out in a unsupervised way, in the sense that no expected classification or categorization of data is known beforehand. A critical aspect of the clustering procedure is being able to evaluate the validity or significance of the obtained groups. For this matter two types of indices are typical used: internal and external indices [28]. The first, deal with the evaluation of the clustering results based purely on current data information, without taking into account an expert defined expected categorization. The second type of indices use external data which refers to a priori expected grouping, in order to validate the obtained results.

Two popular and simple clustering algorithms are K-means and K-mediods. Both are centered in the idea of grouping data in accordance to the closest of K mean values calculated, referred as centroids. For K-means centroids are randomly initialized, and recomputed iteratively based on the mean distance of their closest points. In the case if K-mediods, the centroid always corresponds to a data-point within the data, which allows the algorithm to be computed directly on dissimilarity matrices. K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between cluster points and the centroid. Both have showed a similar performance in practice [39]. Many other methods and clustering techniques exist that have shown more specialized treatment of data and less or none dependence of

similarity distances, however only these two methods are covered since their applied in this study.

2.2 Supervised Learning Algorithms and Methods

2.2.1 Random Forest and Variable Importance

Random Forest is an ensemble learning algorithm that derives from the successful idea of applying the bootstrap aggregating (bagging) technique on decision trees. The general procedure consists of repeatedly select random samples and fit decision trees on them, and then using a majority vote when predicting on unseen data. Since trees can tend to fit noise in data, and led to over-fitting, the use of many trees on different training samples, and then deciding by averaging or voting, allows to decrease the variance of the model without increasing the bias [48]. Also, in order to avoid strong predictors to be selected too frequently in trees and increasing the correlation among them, at each candidate split a random subset of the features is selected. For building a Random Forest the number of trees and the amount of features at each split must be decided a priori. The choice of this parameters will depend highly on data, but can be fairly approximated through cross-validation. Random forest produce an out-of-bag error (OOB) consisting in the mean prediction error of each training sample, using only the trees that do not contain that sample. This error is commonly accepted as good approximation of the generalization error of the model.

Random Forest provide also a way of ranking the predictor features, allowing to assess and understand the characteristics of the model, more deeply. For this variable ranking two types of heuristic are used, the Gini node impurity and the mean decrease of accuracy. The Gini criterion measures the average of decrease in node impurity in the forest when a variable forms the split [3]. The other heuristic is centered in measuring the difference in accuracy between the original model and the one obtained when the values of a variable are randomly permuted. If the variable is

significantly influential on predicting the dependent variable, then the accuracy of the model should drop considerably. Features are ranked based on the score obtained in this permutation process. Recent studies have shown that variable importance ranking through Random Forests can be biased in presence of highly correlated variables [12].

2.2.2 Support Vector Machines and Kernel Trick

Support Vector Machines (SVM) are a very popular method for classification and regression tasks. Its main approach is to look for an optimal separating hyperplane between two classes while maximizing the margin between classes. The points lying on the boundaries are called support vectors, and the middle of the margin corresponds to the optimal separating hyperplane [47]. Beside the maximization of the margin, a regularization parameter (lambda) is set to define the degree of importance that is given to points lying in the wrong side of the hyperplane (miss-classifications). For non-linearly separable problems, this regularization parameter is set to lower values in order allow a higher degree of miss-classifications but also obtaining a less strict separation space. In case of dealing with non-linear problems, SVM can perform efficiently through the application of a method known as the kernel trick. A kernel is a specific function that maps original features into a new feature space. This approach is centered in the idea that data that is not separable linearly in its current space might be linearly separable in a high-dimensional space. Using the kernel trick, data does not need to be explicitly transformed, but using a function that can be expresses as an inner-product in a different space. The described task can be formulated as a quadratic optimization problem, thus allowing efficient problem solving.

2.3 Dimensionality Reduction

2.3.1 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a dimensionality reduction technique that is particularly well suited for low-dimensionality visualization of high-dimensionality data. This algorithm fits within the area of manifold learning, since it relies on the assumption that data lies in low-dimensional manifolds, which allows to perform a reduction that keeps, up to certain degree, relevant structures of data that can then be visualized in 2 or 3 dimensions [42]. The algorithm has been shown to perform considerably better than other low-dimensional visualization techniques such as Sammon mapping, Isomap and Locally Linear embedding, in multiple types of problems [42]. The dimensionality reduction carried out by t-SNE exploits non-linear relations among features, which allows to find relevant structures when these kind of relations exists, unlike linear methods such as principal component analysis (PCA) or linear discriminant analysis (LDA). The algorithm consists of two stages. The first stage constructs a probability distribution over pairs of high-dimensional objects, in order to reach a high probability of picking similar data-points and a low probability of picking dissimilar data-points. The second stage defines a probability distribution on the low-dimensional map, and minimizes the Kullback-Leibler divergence between the two distributions, using a gradient descent approach. It must be noted that t-SNE follows a stochastic process, so the low-dimensional mapping varies from different executions. Also, the method lacks of the capacity of obtaining a low-dimensional mapping function to reduce unseen data based on previous examples, so is not well fit for data transformation applied to predictive analysis. The low-dimensional mapping keeps considerably well the inner cluster distances, but not that accurately the external distances. This means that points that appear grouped together in a low-dimensional space tend to represent accurately the high-dimensional shape, however, distances among the different groups (those where points are naturally clustered together) are not so well preserved [42].

2.3.2 Feature Selection: Recursive Feature Elimination

Feature selection consists on reducing the dimensionality of data by picking or selecting those features that represent more accurately the underlying concepts represented in data. In the case of supervised learning, feature selection might be more directly addressed to retain features that are more strongly associated with the target variable. In general, the main advantages of these methods are avoiding overfitting while improving model performance, building faster and cost-effective models, and most importantly, allowing to build more interpretable models by preserving the semantic of original variables [52]. These advantages come at the price of adding additional complexity to the model-building procedure and the possible loss of information that may get unnoticed by the method used. Literature refers to three main types of feature selection methods: filter methods, which exploit intrinsic properties of the data; wrapper methods which embed the model hypothesis search with the feature subset search; and embedded methods where the search of features is mixed with the model building procedure [52]. Recursive feature elimination (RFE) is a type of wrapper method which consists on building a predictive model the initial set of features, then based on the result variables are ranked according to their influence to the final prediction, and a set of the worst ranked varible (usually just the last one) is deleted, and the process is repeated. The optimal set of features is selected according to the results of every subset of variables. The algorithm used to fit the data must provide the capacity of ranking feature importance, such as partial least squares, random forest and linear models, among many other variations reachable through bagged trees, boosted trees and multivariate adaptive regression splines (MARS). The model assessment of the RFE process should be carried out through cross-validation in order to obtain a more stable prediction error, and avoid prediction bias by features or model overfitting.

2.3.3 Principal Components Analysis (PCA)

PCA is widely used method based on transforming the original feature space into orthogonal components, which are linearly uncorrelated. These are called the principal components, and are constructed in such a way that the first component comprises the largest possible variance in data, and each next component has also the highest variance possible while complying with the restriction of being orthogonal to the preceding components. The method aims to extract the most important information from data, allowing to compress the size of the feature space, and simplifying the description of the dataset in order to ease the analysis of data structure and underlying concepts [37]. PCA bases its calculation on projections based on the correlation matrix among features, which limits the method to exploit linear relations between variables. Usually, for data mining applications, a subset of components is chosen by setting a minimum expected threshold of accumulated variance. For visualization purposes, if the first two or three principal components do not cover a sufficient amount of data variance, the visual assessment might fail to grasp detailed underlying characteristics of data, but still preserve the most general concepts.

2.4 Statistical Methods

2.4.1 ANOVA and PostHoc Tests

The analysis of variance (ANOVA) responds to a series of statistical models typically used to analyse the differences among group means. It is particularly useful to asses if there exists significant differences with three or more groups, and it has been commonly applied in practical problems. ANOVA uses the F-test (F-Statistic test) to determine whether the variability between group means is larger than the variability of the observations within the groups. Means will not be equal (at least not all) if that ratios is sufficiently large. Depending on whether one or two independent variables are used it is referred to as one-way-ANOVA or two-way-ANOVA. When differences in means are obtained with a significant level of confidence, a PostHoc analysis can

be carried out in order to determine which pairs of groups are actually different. The main idea of this analysis is to look for subgroups of smaller parts of the samples in which remarkable patterns might be undetected by a wider approach. There exist multiple methods for applying this type of tests, such as Tukey's, which is applied in this study, which consists of a Student's t-test between every pair of groups but avoiding family-wise errors. These type of errors refer to false discoveries obtained when applying multiple hypotheses tests [51].

2.4.2 Effect Size and Standardized Difference of Means

When performing and statistical test on a null hypothesis, such as the equality of mean between two groups, is typical to rely on the probability of the hypothesis being correct (p-value) under a determined confidence interval. This process has the issue that it does not provide an idea of how strong are the differences found between groups in case of rejection of the null hypothesis. Thus, the seemly significant effect obtained might be a result of chance, when the differences between means are too small or trivial. In order to deal with this the concept of effect size rises, by providing a measure of the size of differences. For the case of comparing two means, Jacob Cohen provided the Cohen's d effect size measure, which consists in the difference of the means of the two groups divided by the average of their standard deviations [18]. This referred study also provided a set of ranges to standardize the categorization of these differences. Being d the calculated value, $d < 0.2$ is considered a trivial effect, $d < 0.2$, $0.2 < d < 0.5$ a small effect, $0.5 < d < 0.8$ a medium effect and $d >= 0.8$ a large effect. A more recent study redefined Cohen's d limits for practical applications, specially those related with sports [50]. Following this schema, $d < 0.2$ is considered a trivial effect, $d < 0.2$, $0.2 < d < 0.6$ a small effect, $0.6 < d < 1.2$ a moderate effect and $1.2 < d < 2.0$ a large effect, and $d >= 2.0$ a large effect.

2.5 Sports Science

2.5.1 Structured Training Methodology

Physical conditioning and training design has been approached with several different types of structures. Such is the case of models such as block periodization, tactical periodization and structured training. This last model, proposed by Francisco Seirul.lo, conforms the main methodological approach for training design at F.C. Barcelona. The model is strongly based in the theories of complexity, by taking the player as the center of training process, and aiming to see it as multi-component and complex being. The training stimuli is variated from week to week in order to seek the adaptation of players to different types of demands, thus improving its different components such as physical, cognitive, tactical and social. This involves the idea of providing a schema in which the player is promoted to adapt to the training demands and evolve in each of its structures, beyond the strictly physical conditions [31]. Within the *structured training* the *structured microcycle* is proposed as a recurrent unit that evolves during the season in order to provide the best development of the player, corresponding to a 3-weeks periodization structure. A *player optimization* is sought through the application of training situations that causes imbalance in one of the subjects structures in order to promote its adaptation, so forcing a continuous auto-organization process [31]. This implies that physical demands for players during training are structured within consequent cycles but are not strictly defined, so the measured physical player values can provide uncertainty and richness in its analysis.

Within the methodology a set of different types of training sessions are defined. Each training day is labelled in strict relation with the following match day, as defined within F.C. Barcelona's training structure. Match day is labelled as MD, the following two days MD+1 and MD+2, and the previous days MD-1 up to MD-4. Each day-type follows specific design rules for training drills. Sessions MD-4 and MD-3 are oriented to strength and resistance, respectively, and also are the more demanding, presenting the higher differences in absolute values and distribution among players.

The structured training methodology considers matches within the training process, as one of the most important stimuli for player physical conditioning.

2.5.2 EPTS Technology. StatsSports Viper Pod

Electronic performance and tracking systems (EPTS) are devices compound with microsensor technology such as global positioning systems (GPS), accelerometeres, gyroscopes and magnetometers. Despite some concerns over the reliability of GPS measurements of accelerations, especially at low sampling rates, these devices have become a relevant tool for analysing the activity profile in team sports, including football. The recent approval by FIFA of its use during official competition, and recent efforts for achieving standardization of measurements have increased the interest for this devices in professional sports. Such is the case of professional sections at *FC Barcelona* where these tools are used for monitoring training load, running speed, traveled distances and many other physical variables. Due to its portable nature, teams can be monitored in both home and away matches, during pre-season or during competitions at a relatively affordable cost [22].

StatsSports technology provides a EPTS solution named Viper Pod, which is used by many of world leading sport clubs, in different kind of disciplines, including Basketball, Football and American Football. The device is conformed by four processors, state of the art GPS module, 3-D accelerometer, 3-D gyroscope, 3-D digital compass , Long range radio ad a Heart rate receiver. These components log data at a rate of up to 100Hz and stream data at over 50Hz. The Viper system both samples and processes GPS data at 10Hz using the newest generation of GPS chipset. The data is available at 10Hz over live streaming and at 10Hz for subsequent download and post session analysis. From this devices over a hundred physical performance variables on different intensity zones can be obtained during training and match sessions.

Chapter 3

Methodology

This section covers the main methodological aspects of the four different proposed phases of this study. For each phase, the main research questions are presented as well as the reasoning supporting the different methods and algorithms selected. Each process is comprehensively described, including technical details and limitations.

3.1 From Data to Dataset: Data Collection And Processing (Phase 1)

F.C. Barcelona has collected both training and matches physical performance measurements, for season 2015-2016, using the *StatsSports GPS Viper Pod* devices, which are carried by individual players. The resulting tracking information is manually segmented by physical coaches, which cut parts of the session where the player was not involved in specifics drills. During this process, a software integrated with the device allows to obtain the overall and segmented results of the session distributed over a hundred variables. Physical coaches have selected 18 variables considered the most relevant performance information of the players, since they constitute measurements at the highest intensities which tend to differ more from player to player, than lower intensity measurements. This section describes the original source data regarding physical performance of player and the different pre-processing procedures carried

out in order to contextualize the information.

3.1.1 Characteristics of Source Data

The original data is gathered in each training session and match, where the EPTS device registers, for each player, 18 physical variables. Depending on its nature and definition, variables can be a result of pure Global Positioning System (GPS) or accelerometer data, as well as a mathematical calculation from any of these two measurements and also gyroscope and magnetometer data, which is provided by the same device. Physical variables are obtained directly from the *StatsSports GPS Viper Pod* software, as summaries (averages or sums) per session, after coaches manually cut the parts of the session where players were not participating in a specific training drill or match activity, such as planned recesses. For most physical variables the registered information is split in six intensity zones, so multiplying the number of available numerical variables. Based on expert knowledge and field observation, physical coaches have chosen to use solely the information regarding summarized data from a set of the highest intensity zones (typically from zone four to zone six), for their daily analysis of physical performance. The reason for this choice is that these high intensity ranges are expected to provide the most rich information about physical performance of players since its related to their response on the most demanding efforts, while lower zones tend to present similar and mostly noisy values among players. Based on this, the referred 18 specific physical performance variables were selected for this study in order to adjust to physical coaches criteria, while removing the physical information obtained in different intensity zones.

The selected physical variables are shown in Table 3.1 where their specific characteristics are described. Beside physical measurements, additional meta variables are provided that allow to contextualize the data. These are: the total duration time of the session (TIME), a player unique id (PID), player's last name (LAST), player's position (POS), the session date (SDAT), and a session id (SID). The categorization of these variables in three groups is further explained in Section 3.1.3.

After the different processing steps are performed, explained below, three different datasets are generated. A training information dataset, a solely match performance dataset, and a final dataset with both the training and matches datasets merged together. This is necessary since some of the procedures require training and match information to be separated.

3.1.2 Data Processing and Variable Extraction

Each training session and match physical performance information is provided in independent Comma Separated Value (CSV) files, which go through a series of processing steps. Initially this data files are processed, transformed to a tabular format, and then finally merged. Each row of the original files correspond to the registered physical information of a specific player in a specific session (training session or match), which in turn is uniquely identified. Once the data is merged it goes through multiple processing steps in order to contextualize the information and finally provide a structured dataset which can drive the analyses presented in this study and any other to be performed in the future.

Day Type Association

At F.C. Barcelona, each training day is labelled in strict relation with the following match day. Match day is labelled as MD, the following two days MD+1 and MD+2, and the previous days MD-1 up to MD-4. This sets up a cyclical labelling schema for referring to days in the week, according to a training structure point of view. Each day-type follows specific design rules for training drills. Sessions MD-4 and MD-3 are oriented to strength and resistance, respectively, and also are the more demanding, presenting the higher differences in absolute values and distribution among players. Sessions MD+1 are typically recovery days (with higher load for players that did not play the last match) and have a low load level, as well as sessions MD-1. MD+2

Table 3.1: Description of selected physical variables splitted in three groups: locomotor, metabolic and mechanical.

Locomotor Variables	
Name and Acronym	**Description**
Travelled Distance (DIS) [41]	Total distance travelled during session drills or matches
Sprints (SPR) [41]	Number of times over $5.5m/s$ during > 1
High Speed Running (HSR) [41]	Travelled meters when speed $> 5.8m/s$
Max Speed (MAX) [41]	Maximum speed reached by the player
Ratio HI/LI (RHL)	The ratio of travelled distances at high intensity ($> 5.8m/s$) and low intensity ($< 5.8m/s$)

Metabolic Variables	
Name and Acronym	**Description**
Average Metabolic Power (AMP) [41]	Energy expended by the player per second per kg, measured in W/Kg
High Metabolic Load Distance (HML) [41]	Distance travelled by a player when the metabolic power is $> 25.5W/Kg$
High Metabolic Efforts (HEF) [44]	The number of separate movements/efforts undertaken in producing HML distance
Load Percentage (PER)	Proportion of AMP with respect to an average 9.5 AMP in matches
Equivalent Metabolic Distance (EMD) [41]	Distance in metres that an athlete would need to cover at a constant speed to expend the total amount of energy.
Speed Intensity (SPI) [41]	Total exertion of a player in a session based on time spent at each speed values.

Mechanical Variables	
Name and Acronym	**Description**
Fatigue Index (FAI) [41]	Accumulated DSL from the total session volume, in terms of speed. (DSL/SPI)
Dynamic Stress Load (DSL) [41]	Total of the weighted impacts, based on accelerometer values over 2g
Lower Speed Loading (LSL) [41]	Load associated with the low speed activity alone
Total Loading (TLO) [41]	The total of the forces on the player over the entire session based on accelerometer data alone
Accelerations (ACC) [41]	Number of increases of speed during at least 0.5 s ($> 3m/s^2$)
Decelerations (DEC)[41]	Number of decreases of speed during at least 0.5 s ($< 3m/s^2$)
Step Balance (STE) [41]	Ratio of left step impact to the sum of the left step impact and right step impact

typically correspond to mid-load intensity sessions. The matchday (MD) is considered among the club's training schema the most demanding in physical terms and a critical part for the physical conditioning of players.

Original data does not present the day-type association in most sessions, however it is considered an important variable to contextualize the information and ease data-related studies. Day-type was added to data following backward labelling recursive process. This consisted in organizing data in descent time from the session date, identifying the following match and labelling the previous training days with the corresponding label, up to the previous match is reached. The complete day-label assignation was then manually validated by physical coaches.

Time Normalization

Each training session and match have different duration time. This can even be different for each player, depending on the actual time spent. In order to provide a fairly comparison of values, each variable is normalized by dividing by the total time of duration of the session or match. Variables that already represent averages or maximums are kept as originally measured, such as AMP, FI, PER, STE and MAX. This type of normalization is commonly performed in most studies related to the analysis of physical information during training. [25]

Load Percentage Variable

In order to design training loads, coaches within the club take as reference the Average Metabolic Power (AMP) of the most demanding session day (matches) to create a relative evaluation of the training load. Given this, training load is designed to reach a specific percentage of the total expected match load, in average. In order to reflect this value of designed load, which is not tracked or annotated electronically by coaches, a new variable is created: Load Percentage (PER). This variable is created as the ratio between the AMP registered by a player and the average AMP registered by all players (excluding goalkeepers) during matches. The variable is added both

for training sessions and match sessions. For the latter, this value will provide a proportion of how much or less demanding was the match in terms of metabolic load, compared with the average match for these players. Since the ratio is the same for all the players, the exact obtained valued for PER is not relevant, but the relations among them are preserved.

Adding Historical Training and Match Variables

Since physical variables are providing insight on player's adaptation to training and further physical performance, and given that data is obtained in a timed manner, the aggregation of these variables along time might provide useful information. This is also supported by the idea, derived from the sciences of complexity, that the physical adaptation of a player tends to be reflected by physiological variations, which are related in time (so they are not random or solely related with the stimulus) [1]. Based on this, for each of the physical variables two additional variables are added to dataset, representing the average value of that variable shown by a player in the last 3-week matches and training sessions, respectively. We refer to this last two set of variables as historical matches and historical training information. Additionally, summarized information is added to matches data such as the average training minutes, average fatigue and total (training plus match) load in the previous three weeks. For all this summarization, only the MD-3 training information is used, since is considered the one demanding the highest effort among all training sessions and avoids the redundancy that might arrive from adding summarized data from additional training days.

Player Information Normalization

Player's last name and position present multiple inconsistencies along the season. Firstly, the same position in the field might appear with different names. For this matter all player positions are normalized to one of the following six positions in the field: center-back, full-back, center midfielder, attacking midfielder, winger, striker.

Additionally, a new variable is created for providing a more general description of positions: defender (center-back, and full-back), midfielder (center midfielder and attacking midfielder) and attacker (winger, striker). For specific sessions and matches of some players the positional information was missing, so a manual process of correction was carried out by verifying the position with coaches and adding the missing corresponding value. Also, Goalkeepers are deleted from the database since they face considerably different physical challenges and training schema than field players, so loosing the similarity that exists among training sessions days. A second issue is that players' names commonly appear with different text, where letters are missing or the first name or surrogate name is used instead. This was also normalized through an automated process based on predefined rules.

3.1.3 Structuring Physical Variables

The high amount of physical variables and the inherent differences of the units and ranges difficult the analysis and further communication of observation and findings. Also, some of them are expected to present a high linear correlation, since they are originated from similar or related calculations and measured with the same device. In order to provide a more compact structure for these variables, tree main groups are proposed regarding the origin of measurement and their nature: metabolic, mechanical and locomotor. Metabolic-related variables are associated with energy expenditure and exertion, and mechanical variables relate with intensity changes and impacts, following the classification criteria used for similar variables in a recent study [25]. The first two groups contain variables which are calculated in most cases with a combination of GPS and accelerometer with higher influence of GPS in the first one and higher influence of the accelerometer in the second one. The third group, locomotor, refers to calculations associated to simple direct measurements of travelled distance and speed, that are obtained solely through GPS. The relation between the different variables conforming these groups is better detailed in experiments from Section 4.2 where the correlation between each of the predictor variables in MD-3 is presented. It is expected to be observed that metabolic and locomotor variables tend to present

higher correlation. Also moderate correlation between locomotor and metabolic variables is expected to exist since most of the metabolic variables are created through calculations that take into account locomotor variables.

3.1.4 Selection of Day-Type MD-3

For Phases 3 and 4 of current study, the training dataset is reduced to comprise only the information of day-type MD-3. As explained before and more thoroughly explained in Phase 2 results, MD-3 provides the highest demanding effort of the whole training cycle for players, as well as providing the most similar stimulus to matches. Performing the different analyses presented in this study while using all the training day-types, would force to find mechanism to synthesize the physical information from each day in a compact and unified form. This process is expected to be highly complex, and it has not been approached in recent literature on the subject. Also, physical coaches tend to analyse this data by day-type at a time, instead of doing a, possibly inaccurate, summary of the whole training cycle (1-week). Moreover, this goes beyond the scope of this study.

3.2 Understanding The Data: Initial Exploration (Phase 2)

Physical coaches at F.C. Barcelona have developed a series of beliefs and knowledge regarding the characteristics of EPTS performance information, due to their continuous monitoring of physical variables and expertise in the field. Most of analyses carried out in the past on these variables consisted in by visualization of mean values for each variable, determination of thresholds based on team mean, and comparison between players. In order to perform more complex analyses regarding the relations among these variables, it becomes critical to deeply understand the inherent characteristics of the data, validating most of the main beliefs stated by physical coaches, and even reaching new knowledge, through an initial exploration. To this effect we

propose, in collaboration with physical coaches, a series of questions and hypothesis to be addressed by exploratory analysis. In each case we outline the analytical approach for answering these questions, while a more detailed procedure, followed in this study, is described in Section 4.2.

How is data distributed per variable and session type?

One of the main objectives of the structured training methodology at F.C. Barcelona is managing the distribution of training loads along training weeks, and the season as a whole. Matchday (MD) is believed to demand a considerably more intense effort which should be evidenced in most of variables. Also when closer to MD, training sessions are expected to show lower intensity effort, which should be magnified at MD-3 and MD-4. It is expected that the average metabolic power (AMP) presents a range of values during MD which corresponds to the known standard for professional football, since its the main variable used for designing training load. In order to assess this, data for each variable and player is plotted through boxplots which compare variables and session types, through mean, standard deviation, and quartiles. In the case visual assessment is not sufficiently clear, an ANOVA test is performed to validate there are significant difference in the mean of the different groups, and a POST-HOC test is performed to asses for each pair of groups if the difference is present.

Is there a clear difference between session types when considering relations among variables?

Beyond the specific differences that might arise from comparing variables one-on-one in different session types, there is the believe that there is a clear overall difference between MD and the training sessions, in terms of per-minute normalized intensity and effort. Also MD-3 and MD-4 are expected to be the closest to MD in these same terms. In order to address this is desired to perform an exploratory analysis which takes into account relations among variables. A visualization approach is followed through plotting the first two components from principal component analysis (PCA),

which in case of containing a considerably high amount of variance might provide useful insights on the relations between session types. Also, t-distributed stochastic neighbor embedding (t-SNE) is applied since its ability for mapping high-dimensional data into lower dimensional spaces. For both cases, only in the presence of highly evident differences, interesting answers for this main question might arise.

Are there differences or similarities between positions in the field?

From general football knowledge and the opinion of expert physical coaches, the different positions in football are expected to provide differences in physical efforts. These differences are expected to be independent from the specific player characteristics up to some degree. A similar procedure using boxplot analysis is proposed.

How much are the variables linearly correlated? Is there high correlation between variables of the same group?

Given that physical variables are measured using the same device on specific activities, and that some calculations come from combinations from these, it is expected to find relevant linear correlations. When performing multivariate analysis, high correlations between variables might become an important factor to consider, since it might affect the accuracy and interpretability of results. For assessing this, the Pearson correlation between each pair of variables is calculated and plotted.

Do the physical values during matches vary considerably depending on different ranges of historical values?

Historical aggregated information might result valuable for structuring and analysing further physical responses in matches. Along the continuous process of training design coaches attempt to balance and handle the training and match load and minutes. Also, a variable such as Fatigue Index (FAI), which is not currently taken into account by coaches, might also be relevant during this design process. To address question we propose to structure different historical load, minutes and fatigue, in different groups or ranges, and visualize if there is a considerably variation of the different

physical variables for each group. Historical information consists in the average (and sum) value for team load, train and match minutes, and fatigue in the last 3-weeks (according to micro-cycle length). Since each of this historical values corresponds to 1-dimensional data, groups are found through cluster analysis, so reaching the most natural separation of this data. The number of groups in each case is defined by majority decision on a set of different internal measures for cluster evaluation. The followed process is more accurately detailed in Section 4.2.

Does training physical variables present a oscillatory tendency?

From the study of physiological characteristics of physical conditioning coaches hold that when evidencing different adaptational behaviours during training, players tend to present an oscillatory tendency on physical values. These oscillations are also attributed to the structured training schema in which this variations are pursued. In order to understand if the current data validates this belief, physical variables are individually plotted along the season weeks. Only MD-3 session type measurements are used for this purpose, based on its similitude to MD and in order to simplify the analysis.

3.3 Weekly Variations And Match Performance: A Time-Based Exploratory Analysis (Phase 3)

As detailed in Section 2.5.1, F.C. Barcelona training methodology, the structured training, organizes training design within a 3-weeks periodization structure (the structured microcycle), involving continuous variation of physical values. Based on the idea of *deterministic chaos* present in biological systems [1], players are expected to evidence different adaptational behaviors along the season trainings. Also, training sessions by day-type are intended to be as different as possible in terms of specific exercises, although keeping similar load levels, so the measured physical player values can provide uncertainty and richness in its analysis. Given this, is plausible to

think that periodical variation of physical values could provide valuable information regarding the adaptability and fitness of the player.

In this phase of the study we propose a methodology to answer the question of whether there exist significant relations between the periodical variation of physical training variables, and the subsequent performance of players during matches. For this, we analyse specific patterns of variations in the data in time, as an approach of measuring the level of adaptation of players to the training stimulus. Physical coaches have instructed that a higher level of oscillation of physical values in time is commonly associated to better adaptation to training, being also true the opposite case with lower variations, although the specific ranges for players have not yet been defined. Machine learning algorithms are used in order to exploit the contribution of the high amount of measured variables as a whole, all of which are expected to contribute explaining the player's dynamic up to some extent. Following a fully unsupervised approach, two methods are proposed for finding patterns in temporal data, and then cluster analysis is performed to find natural groupings from these patterns, on training data. This temporal data corresponds to consecutive MD-3 training sessions of three weeks for a specific player, which is then associated with the player's match performance of the next week. In other words, clusters found in training windows, are used to label a player's physical performance during the next match associated with the corresponding time window.

For this phase, both training dataset and matches dataset are used. The training dataset is cleaned up to contain only physical information from day-type MD-3, for the reasons detailed in Section 3.1.4. Also, physical coaches have selected 15 variables from the overall 18, that are considered to be more representative regarding inter-week variation. These are: DIS, SPR, HSR, MAX (locomotor variables), FAI, DSL, LSL, TLO, ACC, DEC, STE (mechanical variables), and AMP, HML, HEF, PER (metabolic variables). Excluded variables are RHL, SPI and EMD. A complete overview of the whole methodology can be observed in Figure 3.1. Here, a similar

pipeline is followed up to creating two time-based datasets (with and without aggregation of time window values). For the first one, dynamic time warping is applied (DTW) for obtaining a distance matrix of 3-week window frames of physical values (15-dimensional datapoints). The second dataset is already a direct summary of the 3-week window frames. In both distance matrix and summarized dataset, cluster analysis is applied to obtain a natural grouping of this physical data, and further comparing with matches.

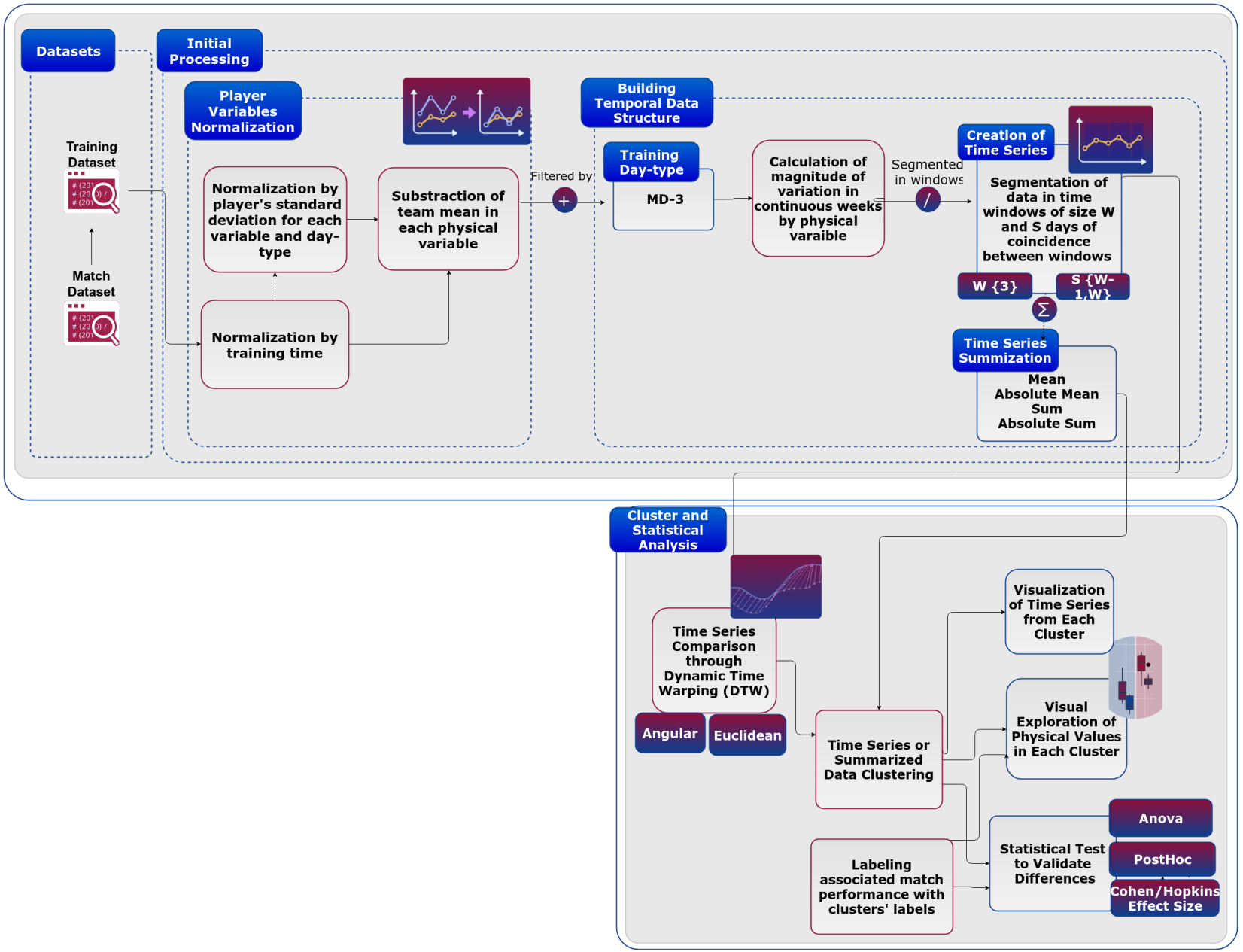


Figure 3.1: Diagram presenting the main procedures involved in the time-based exploratory analysis of data.

3.3.1 Calculating Physical Variations

The absolute values of physical variables can differ considerably from one player to another, due to their inherent differences in physical characteristics and position in the field, beyond the actual fitness state of players. With the objective of minimizing these differences and aiming for a more fair comparison of player values regarding its fitness state, the absolute values of each variable are normalized. The applied normalization consists of expressing each value as the number of player's standard deviation of that variable (at day-type MD-3) along the season. For doing this, the standard deviation of each player and each variable is calculated, and the absolute value measured in each training day (MD-3) is divided by this standard deviation value. In a more general sense, this transforms absolute values into a measure of how much is a player deviating from his own mean in a training session. This is believed to present a more reliable view on player's deviation from a more typical state, and hopefully provide a more reliable axis of comparison between players.

After doing this, we would like to obtain the degree of variability from week to week on each physical value. The idea is to measure the difference between registered values from two consecutive weeks, as presented in Figure 3.2.

Each value V_i represents the absolute difference between a value registered at sessions S_{i+1} and S_i. From this, two datasets were built. The first dataset consists of time-series of W window size, which is used for the dynamic time warping procedure, explained belowe. A sliding window approach is followed by using a fix-sized (W) window of consecutive weeks. The time-series dataset is conformed by groups of W rows containing the 15 selected physical variables, corresponding to a player in a specific period of the season. The selected window size during experiments is 3 in order to match the methodology of the club. Windows are moved SW steps each time, so to control the degree of coincidence of values between windows. The value of SW was selected following Equation (3.1) to avoid an excessive overlap between windows and to avoid a too strict separation that would reduce significantly the

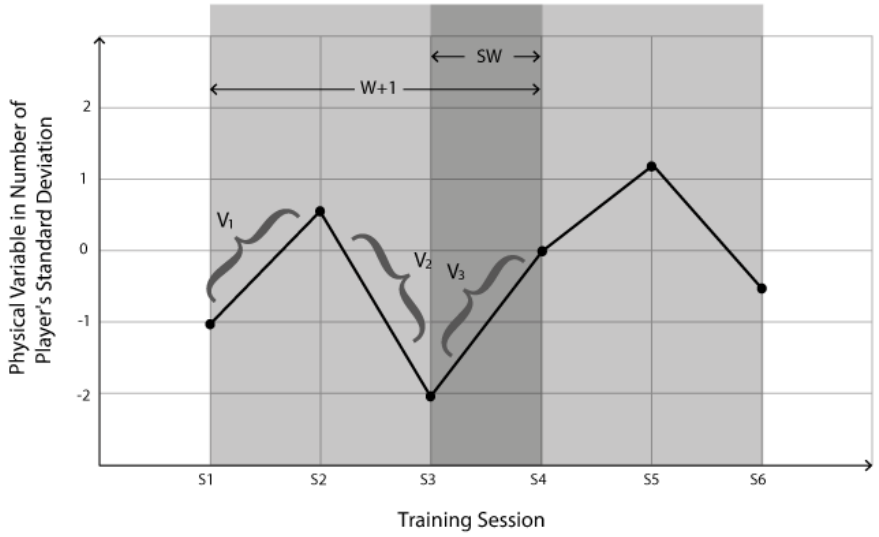


Figure 3.2: Representation of a series of measured values of a particular variable during weekly training sessions (x axis). V_i values refer to the difference of values registered at sessions S_{i+1} and S_i. W is the size of the sliding window, used to build time-series and summarized datasets. SW refers to the amount of weeks to slide each time.

amount of data. The second dataset built summarizes each group of W rows in each variable, by calculating the average of absolute differences. This represents a more compact representation of the variation information of the player, where each window is transformed into a single averaged 15-dimensional datapoint. Equation (3.2) describes the performed calculations, where P_{jvd} corresponds to the absolute average of window differences of a variable v of a player j, measured in the window frame d, substracted by the mean of P_{ivd} for every other player i. P corresponds to the set of all possible players.

$$SW = W - (W/3) \tag{3.1}$$

$$P_{jvd} = \frac{\sum_{i=2}^{W+1} \|S_i - S_{i-1}\|}{W} - \frac{\sum_{i \neq j}^{|P|} P_{ivd}}{|P|} \tag{3.2}$$

3.3.2 Matching Time Patterns through Dynamic Time Warping

Dynamic time warping (DTW) was applied over the time-series dataset in order to calculate similarity between windowed variations along the season on different players. The idea is to find variation patterns that are more similar to each other, independently from the specific player or position. A distance or dissimilarity matrix is found for each pair of series in the dataset. Two distance measures were initially taken into consideration: euclidean and angular distances. Euclidean distance favors vector magnitudes over angular orientation, while angular distances does the opposite. Based on expert opinion the nature of the oscillation patterns is expected to be informative when matching series of similar magnitude than matching specific changes in angles, specially regarding the small size of window frame (3-weeks). In other words, the similarity of the path of oscillation is considered noisy and less relevant, than the absolute variation manifested by the player, in order to approximate the physiological response and the adaptation capabilities. Based on this, Euclidean distances were selected. Once the dissimilarity matrix is found, the *k-mediods* algorithm is applied for finding a natural clustering of the time series.

3.3.3 Clustering Degrees of Variations

Both for the time-series dataset and the summarized dataset, cluster analysis is applied to find natural groupings regarding the variation of the measured physical variables. It is critical to observe that the clustering procedure is applied to multidimensional data that involves the 15 measured physical variables at a time, aiming to incorporate the relation between each of the variables. For the time-series dataset the *k-mediods* algorithm is applied, since its capability of being applied to distance matrices and the flexibility of controlling the number of clusters, which is preferred as low as possible for easier applicability. For the summarized dataset, *k-means* is used instead. Having a similar performance than *k-mediods* and similar characteristics, becomes a more fairly comparable approach between results. The selection of num-

ber of clusters is performed by calculating multiple internal indexes (and not external indices since there is no labelled data) and choosing the majority vote. The specific indices used during this study are explained in the experiments section (Section 4.3).

Once the training sessions information is clustered, each of the window-frames is associated with next upcoming match (the match after the series). This generates a new dataset containing the absolute values of each physical variable registered by the associated player in that following match. Theses matches are labelled with the previous training series cluster number in order to associate characteristics of the clusters in training to characteristics in the match.

The characteristics of the different clustered groups can be assessed through statistical tools such as the standardized difference of means. This measure, which represents effect size, is recently being applied in sports for comparison of performance data between players, and its particularly helpful for discriminating between groups with practical implications [50]. For the clusters found in training data (for both datasets) the effect size for each variable is calculated, in order to evaluate how big is the difference between both groups regarding each variable. This allows to understand which are the main characteristics of the found groups. Then the physical information of the next match after each short series is found, and labelled accordingly the associated training clusters. This provides a new clustering for matches of the same size of the training clustering. For this new clusters the standardize difference of means are also calculated for each variable and further analysed, in order to understand if these associated matches groups present remarkable characteristics and differences.

3.4 Estimating Future Match Performance: A Predictive Analysis (Phase 4)

Within the weekly analysis performed by coaches, physical variables are overview from player to player, and specific measures such as the team mean and the player mean are used to assess individual performance. The main objective of this analysis is to estimate the current fitness state of the player, which might lead to decisions such as recommending lining-up or not the player in the next match. For the latter, coaches typically make use of the currently measured data and the overall visually perceived performance of the player during the microcycle. However, the great amount of variables available, and the complexity of understanding the relations among them, makes unfeasible for physical coaches to perform a one-to-one variable analysis in a frequent basis and be able to reach conclusions quickly. Under the hypothesis that match physical performance could be predicted from past training physical performance data, it becomes clear that a model capable to estimate future physical performance would become an useful tool for assessing players' fitness state, and simplifying the fast-paced periodical analysis.

For this phase of study we approach the question of whether is possible to predict future physical performance variables based on past performance. If possible, a derived question arises regarding which variables are most relevant to explain each other one. For this matter a regression analysis is proposed that involve the application of machine learning algorithms that exploit either linear or non-linear relations among variables. Given the relatively high amount of predictors or features, two different feature selection strategies are evaluated with the aim of reducing the noise caused of highly correlated variables which occur with high frequency, facilitating variable analysis and increasing prediction accuracy. Since the results are intended to be applied in practice, a specific Normalized Root Mean Squared Error (NRMSE) is proposed for model evaluation. The expanded dataset that includes aggregated historical variables is used, in order to evaluate their influence in explaining future outcomes. Finally,

variable importance from Random Forest will be calculated in order to approximate the relative importance of predictors and targets.

3.4.1 Re-structuring the Dataset

For this phase of the study both training and matches datasets are used. Training information is filtered to contain exclusively MD-3 training day information, as in previous cases. Also, the calculated historical variables are also included, from both training and matches. It must be clarified that historical information of a given training or match only accounts for the previous 3-weeks (or less in case of first three weeks of the season) to that event. Using information from future events will incur in a clear bias in the data. The selection of only MD-3 training information allows to avoid the issue of having historical variables repeated among rows with the same target variable, which would also tend to greatly bias the trained model and provide erroneous results. The variable STE is excluded from this part of this study since is considered to be highly dependent from specific events during the match. Thus, matches physical data involves the remaining 17 variables.

This dataset is still not ready for performing a regression analysis. For this matter we need to add a target variable that is associated with each training session. Following this idea each of the physical variables in matches is used as a target variable for prediction, thus implying the generation of 17 different datasets, which contain the same predictor variables but different targets. Following the selection of MD-3 for training data and since the use of match variables as target for prediction, each dataset is reduced to contain strictly the training sessions and aggregated information of players that played the next match. Since for the resulting datasets the target variable in each case corresponds to one of the 17 match variables to predict, the original prediction task is transformed into 17 independent prediction tasks. After adding the historical training and matches variables, and the additional context variables the number of predictors raises up to 66. Data is further standardized in order to equalize the importance of each variable to avoid unbalancing due to specific different

ranges and magnitude of values among variables.

3.4.2 Feature Selection

Considering the high number of predictor variables and given the high correlation among some of these variables, feature selection seems like highly desirable. The main advantages and drawbacks of these methods have been detailed in Section 2.3.2. For this study we have considered two feature selection approaches: pairwise-correlation selection (COR) and recursive feature elimination (RFE). The first approach, which can be roughly considered a filter method, consists on finding the pairwise Pearson correlation among the predictor variables and removing variables that are above a certain threshold. The second approach was applied by using Random Forest variable importance ranking, which have shown high performance in multiple types of problems, especially those where variables do not vary greatly in their scale of measurements [48]. The COR procedure becomes relevant given the high correlation among some of the predictor variables, as shown in Figure 1.9, which is known to impact negatively on final regression (or classification) error in most machine learning tasks. The COR procedure is always applied before the RFE, since high correlation of predictor variables has been shown to bias the selection of features by wrapper methods, and particularly in the case of random forest [12]. Also, RFE is performed using cross-validation, where average feature ranking is used in order to obtain an unbiased estimator of importance.

3.4.3 Regression Analysis

A regression analysis procedure is carried out that seeks to evaluate how predictable these variables are, with the given data. For each target variable multiple combinations of pre-processing steps are applied to also multiple different algorithms. In order to evaluate a wider range of possible relations among variables (features and targets included), both linear and non-linear learning algorithms are applied. Since

the contribution of feature selection methods or a specific algorithm for building the best model can not be clearly anticipated, is recommendable to try different combinations of the pre-processing steps and final statistical methods. The specific details of the performed experiments are presented in Section 4.4.

The objective of this analysis is to obtain the best possible model in terms of minimizing prediction error. In order to approximate as much as possible the generalization error, nested cross-validation is used. It is critical to observe that recent studies have shown that when parameter selection is involved within a cross-validation procedure for model building, the average fold error will be biased to the model selection procedure, and thus the obtained error will be lower than the actual generalization capabilities of the model, leading to erroneous results [10]. We deal with this problem using nested cross-validation, where the outer cross-validation estimates the generalization error of a model, while the inner cross-validation optimizes its parameters. As a consequence, different outer fold models will possibly use different parameters. The variance of the errors among the outer folds will also provide an idea of how good or valid the parameter selection procedure is for each algorithm. All the pre-processing steps such as standardization of data and further feature-selection are applied to each of inner folds of the nested-cross-validation process. Not doing so, would lead to data leakage and thus to an optimistically biased error estimation [11].

For evaluating the performance of regression as well as for the wrapper-methods on feature selection, the mean square error (MSE) is used and minimized; see Equation 3.3. From this error we derive and additional error metric: normalized root mean square error (NRMSE), described in Equation 3.4. NRMSE is used as the ratio of root mean square error and the standard deviation of the target variable. This expresses the magnitude of the obtained error in terms of number of standard deviations of the target variables. Depending on the variable, an expert practitioner can assess if the provided error is acceptable or not for her analysis objectives, by defining an sufficient threshold in terms of standard deviations. Also, through this metric,

prediction error becomes directly comparable between variables (independently from very specific units).

$$MSE = \frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n} \tag{3.3}$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}}}{\sigma(y)} \tag{3.4}$$

3.4.4 Variable Importance Analysis

For each of the independent variables, predictor variable importance is calculated in order to provide a much clear and practical interpretation of their effect. The variable importance ranking from Random Forest is used, as well as a feature selection procedure, further described in Section 4.4. This approach is based on calculating the mean increase error (MIE), as an analogous to most typical mean decrease accuracy, which is obtained when predictor variables are randomly permuted. Variables are ranked based on the impact they have in final prediction error when removed. The parameters of the best performing model for Random Forest during the regression stage are selected and a new model is built analogously. Variable importance in each of the folds is averaged to produce a final variable importance ranking that is expected to provide the most reliable representation of the influence of the predictor variables. The choice of Random Forest derives from its capacity to provide variable importance ranking on prediction, and is sustained by the obtained prediction results for this method (to be presented later). The selection of one specific approach simplifies the overall explanation of the importance of variables, since the objective is to grasp the general influence of the different variables among the three defined groups.

Recent studies have shown that variable importance ranking through Random Forests can be biased in presence of highly correlated variables [12]. In order to deal with this, a procedure of removing the highly correlated variables is performed (leaving just one of two, based on pairwise correlation) prior to model fitting and

variable importance calculation. Alternative methods have been proposed in order to approach this problem in a more elegant way [12], although at the price of higher computational costs. These more expensive methods are left out for future work. Results from variable importance are then visualized in order to reach conclusions about influence of predictors for each target variable.

Chapter 4

Experiments

For each phase a series of specific experiments were performed following the methodology proposed in the previous section. For the first phase, regarding the processing of the original data and generating datasets for further use, the final structure of the obtained dataset will be completely explained in this section. This will allow to understand and follow properly the design of experiments of the following phases and the further results. For each of the remaining phases the specific algorithms applied are defined, as well as indicating the characteristics of the datasets to be produced during each phase.

4.1 Main Dataset Structure

The original data consisted in 187 CSV files, where 153 corresponded to individual training sessions, and 34 to matches along Season 2015-2016 of Barcelona B team. Each file contains 117 variables with both physical information and session and player identification. As detailed in Section 3.1, these files are read, interpreted and structured to provide two main datasets corresponding to training sessions and matches sessions information, following the series of transformations and addition of variables described in the mentioned section. After the selection of the 18, previously explained, specific physical variables, and the further addition of historical variables, the resulting datasets are conformed by 66 features. Both datasets present the struc-

Table 4.1: Description of the features present in the processed training and matches datasets

Feature	Description
Physical Variable	Each of the 18 selected physical variables. Each feature entry corresponds to the measurement performed on that variable in that session.
Historical Training Physical Variable	For each of the 18 selected physical variables, a new variable is added with the average value in the previous 3 training session of type MD-3 [1]. Additional historical variables of Training Minutes, and Training Fatigue Index are added.
Historical Matches Physical Variable	For each of the 18 selected physical variables, a new variable is added with the average value in the previous 3 matches sessions (MD). Additional variables for Training Minutes, Match Minutes, and Match Fatigue are added.
Player Id	An unique numerical identification of the player.
Player Position	A numerical identification of the player position
Player Last Name	The last name of the player
Total Time	Total amount of time spent in the training session or match
Day Label	A unique numerical identification of the session type
Session Id	A numerical identification of the specific session or match.
Next Match Session Id	A numerical identification of the id of the next upcoming Match. Just present in the training dataset.

ture detailed in Table 4.1. The training information dataset contains 2478 entries, while the matches dataset contains 473 entries, covering physical information data of 42 different players. A third dataset is created from the merge of these two, which allows to compare training and matches directly. These datasets are used as the main input throughout the next three phases of the study.

4.2 Exploration Plots (Phase 2)

The exploratory analysis of the data is focused on answering the main questions proposed in Section 4.2. For this purpose, data is plotted and filtered in several ways in order to discover patterns that validate physical coaches main believes, and also to possibly reveal new relations that might be used as baseline knowledge for the next to phases of the study. Here, we detail the different experiments carried out. First, a boxplot is presented where physical variables are compared one by one, and filtered by session type. The boxplot shows how data is distributed among four equal parts, while the mean and standard deviation of the data is indicated. All the variables which do not refer to maximum values are normalized by the time spent in the session, to provide a fair comparison. Also, since the differences between session types are expected to be considerably high, specially matchdays, data is normalized

within a [0..1] scale, so visual comparison is proportional. Afterwards, linear relations among variables are exploited through PCA and t-SNE dimensionality reduction, in order to visual asses differences and similarities between groups. First, the two principal components of the PCA reduction are plotted and labelled according to session type. Then, t-SNE is applied to reduce the high dimensional data down to two dimensions, and plotted and labelled accordingly. The same two types of analysis, through boxplots and linear dimensionality reduction techniques, are applied to data but taking into account the player positions. In this case, only MD session types is used for filtering data. For t-SNE a perplexity of 6 is used, while reaching 2000 iterations and a theta optimization value of 0.1.

After visualizing data distribution and first relations between variables, the correlation between variables is studied. In order to asses this, a correlation plot is presented where for each pair of variables the Pearson correlation is calculated. Variables are grouped according to their associated variable type in order to detail characteristics that might arise between locomotor, mechanical and metabolic groups. The next part of the exploration focuses on understanding the contribution of historical information to reveal specific performance patterns on the selected match variables. The historical information of player load, training minutes and fatigue is split in several groups, and for each group a boxplot of a set of selected physical variables is presented. The grouping of each historical variable is performed through cluster analysis. The number of groups or clusters in each case, are decided by applying the K-Means algorithm, using K values in the [2..6] range. Five internal indices metrics are used to decide the optimal number of clusters in each case. These indices are: C-index, C-H index, DB index, Silhouette index and the Ratkowsky-Lance index [5]. Since each historical variable is 1-dimensional, the clustering procedure allows to divide the variable in K natural groups. For example, if K value of 2 is selected for the historical value of training minutes, then the first group will refer to the lower set of trained minutes and the second to the higher set of trained minutes. Clustering analysis provides a more fair division than splitting directly based on an arbitrary threshold. For each historical variable divided in K groups the registered match physical value

of a set of variables is plotted. These variables are DIS and RHL (Locomotor), AMP and PER (Metabolic) and FAI, ACC and DSL (Mechanical). The variables in each group are representatives of the higher correlation subgroups that will be observed in the linear correlation diagram.

For each match variable plotted by historical variable groups, ANOVA and POST-HOC (Tukey) tests are performed to ensure that there exists significant differences in means between the K groups, thus possibly showing interesting patterns. For space reasons, only the plots showing remarkable patterns are presented. ANOVA and POST-HOC results are presented in detailed.

A last part of this exploratory analysis is focused in visualizing the expected oscillatory pattern of physical variables during training. For a set of physical variables, a line plot and interpolation curve of the mean variable value and each position at each training week is presented.

4.3 Unsupervised Performance Analysis (Phase 3)

Following the procedures described in the methodology, the training set is first transformed to reflect the week to week variations of physical variables. From this, the two time-based datasets mentioned in Section 3.3 are created. For both, a window size of 3 and a sliding window size of 2 are used [2]. The first dataset consists of a 3-weeks time-series of physical variables, while the second is created by summarizing each physical variable in each short time series, thus producing a summarized dataset. The absolute average of variations is used for the summarization. In case there is a missing training session within the 3-week window, this series is removed to avoid missing data in the time series. Both datasets consist of 112 samples (either time-series or data-points). For each series, the next match session coming after its

[2]During experiments multiple sizes of windows (3, 6, 9) and sliding window steps (1, 2, 3) were tested, but this configuration have been selected according to physical coaches criteria and to simplify the presentation of results

last training session is obtained. The list of all the associated matches is used to filter the matches datasets, producing a new matches dataset with 82 samples [3].

On the first dataset, dynamic time warping (DTW) is applied using Euclidean distances, and producing a distance matrix between each series. On this distance matrix the k-mediods algorithm is applied for clustering the data. On the second dataset k-means is applied instead. The selection of the number of clusters, k, is performed by calculating five internal indices and selecting the number of clusters picked by the majority. These indices are: C-index, C-H index, DB index, Silhouette index and the Ratkowsky-Lance index [5]. For both procedures, matches are labelled with the cluster label associated with the previous training window. A comparison between the same variables in each clusters is performed by calculating the standardized difference of means. This calculates the effect size of the difference of means, in case a t-test statistical significant difference is found between means. The limits for the effect sizes followed the ones suggested by Hopkins [50] which are recommended in sports related data and for practical applications (trivial effect: < 0.2, small effect: $0.2 - 0.6$,, moderate effect: $0.6 - 1.2$, large effect $1.2 - 2.0$ and, very large: $>= 2.0$), with a confidence interval of 90%.

The results are presented in Section 5.2 where the standardized difference of means in the different groups is presented, along the mean and standard deviation values of each variable, observing the most relevant relations between training variations and match performance.

4.4 Supervised Performance Prediction (Phase 4)

This fourth phase is divided in two main parts, the prediction of match physical performance variables, and the explanation of the most influential variables for each

[3]It must be observed that associated matches have a lower sample size, since not all the players who fitted into the match previous 3-week window actually played the next match

predicted variable. The predictive model building involves a the application of pre-processing filters, and a grid search for hyperparameters in each of the selected algorithms, through a nested cross-validation procedure. The pre-processing procedures consists on data standardization and dimensionality reduction through principal component analysis (PCA) or feature selection. Random Forest (RF) and Radial Basis Function Kernel Support Vector Machines (KSVM) were selected as the set of algorithms that exploit non-linear relations among variables. On the other hand, Linear Support Vector Machines (LSVM) and Linear Regression (LREG) were used as methods that are based on exploiting linear spaces. Regarding feature selection, for the parwise-correlation filter a threshold of 0.8 Pearson linear correlation was chosen, while for the RFE procedure the set of features achieving the lowest mean square error (MSE) where selected. For each algorithm the set of pre-processing combinations were the following. First at each fold, standardization is applied by transforming each data column to have mean 0 and unit variance. Then, the COR filter was either applied or not. For the cases where the filter was applied, the following combinations were also applied: COR+RFE and COR+PCA. PCA was not applied to KSVM since the kernel function is already transforming the feature space. This approach provides 4 different combinations for each algorithm, except only 3 in the case of KSVM. For each algorithm a parameter selection phase is carried out by testing different parameter combinations. For Random Forest both number of trees ($[50, 100, 250, 500, 750]$) and the number of variables sampled as candidate at each split are tried ([ncol/3,ncol/4] where *ncol* refers to the total number of predictor variables). For KSVM the tested parameters are the gamma parameter of the Gaussian kernel ($[0.0001, 0.001, 0.01, 0.1, 1, 10, 100]$) and the cost of missclassifications ($[0.0001, 0.001, 0.01, 0.1, 1, 10, 100]$). The same cost of missclassifications list is used for LSVM.

The amount of data available is considered insufficient for building a separate Test set beside the Training and Validation sets build during cross-validation. This is why the whole dataset is used during the nested-cross validation procedure (split in sub-

sequent training and validation sets) which, as explained in Section 4.4, is expected to provide a performance error close to the true generalization power of the model, on similar data. For the outer and inner cross-validations 5 and 2 folds are used respectively. For each combination, the mean normalized root mean squared error (NRMSE) of the outer folds and their standard deviation are reported. The degree of variance among the folds is assessed to analyse the stability of the model selection procedure. The mean NRMSE is used to determine the best performing combinations in each case.

For assessing the variable importance on each of the target variables, Random Forest was used, by applying the COR filter within an analogous nested cross-validation procedure where the average best ranking features among folds were selected. Random Forest variable importance metrics have been extensively used in literature. The mean increase error (MIE) obtained by the variable importance ranking is expressed in terms of NRMSE. So, the impact of variables is measured in terms of how many standard deviations of the target variable would be added to the prediction error if the variable was missing. In order to visualize the importance of variables a chord diagram is used where the proportional influence of each of the predictor variables is observed. This is further explained in the results section.

Chapter 5

Results

This section presents the results obtained in each of the main phases of the study. First, a comprehensive set of plots are presented in order to visually assess the characteristics of data and answer general questions and believes regarding physical behaviour of players. Then, the unsupervised approach results are detailed, presenting the found relations between weekly variation of variables and match performance. Finally, the main results on the supervised approach are presented.

5.1 Initial Exploration Results

The exploratory analysis of data is a critical step for understanding its main characteristics and to devise some initial underlying patterns. Also, the study attempts to validate some of the physical coaches believes on this data and physical conditioning in general, as a way of assessing the quality of the information. Figures 1.1, 1.2 and 1.3 present a boxplot distribution for each of the selected physical values, distributed among the proposed physical groups: locomotor, metabolic and mechanical. At first glance, the inherent structure of training design can be observed, where matchday (MD) presents consistently higher values and higher variability than the training sessions. MD-1, and MD-2 arise clearly as the day where the physical demands are the lowest along the week, which coincides with their association to recovery days. Although MD-3 and MD-4 are focused on different types of exercises, they show to be

the most similar to MD, as it is also expected. MD-4 tends to present slightly higher values than session MD-3, and similar ranges of variability, in terms of data standard deviation and the range of distribution of the data points. MD-3 standard deviations tends to be higher for most cases, in comparison with MD-4, which is consistent with the idea that this type of sessions present typically the most diverse variation of activities. It is noticeable that MD+1 and MD+2 days present a considerably higher variability in terms of metabolic and mechanical efforts, while being lower in terms of locomotor measurements. This is expected since in these sessions players that were not involved in the previous match (MD) are exposed to stronger intensity efforts than players who did participate in the match. Although variables are normalized to ease comparison, is clear to observe than MD presents considerably higher and less stable demands, specially for metabolic and locomotor efforts. Despite the intention of coaches of providing training sessions the most similar possible to matches during the week (in particular in sessions MD-3 and MD-4), it becomes evident that matches, which in opposition to training sessions can not be controlled, are considerably different from the rest of the training sessions. Load percentage (PER) which is variable used to measure and control demanded effort in sessions reflects clearly the expected differences in demands from session to session. It becomes noticeable that variables that do not depend directly on session specific conditions or demands, such as Fatigue Index (FAI) and Step Balance (STE) are very similar in terms of data distribution. Both variables are aimed to measure player's evidenced fatigue in different ways, and arise as possibly more informative variables in terms of individual players. However, STE presents a very low level of variability, and is expected to be more difficult to fully assess, than FAI.

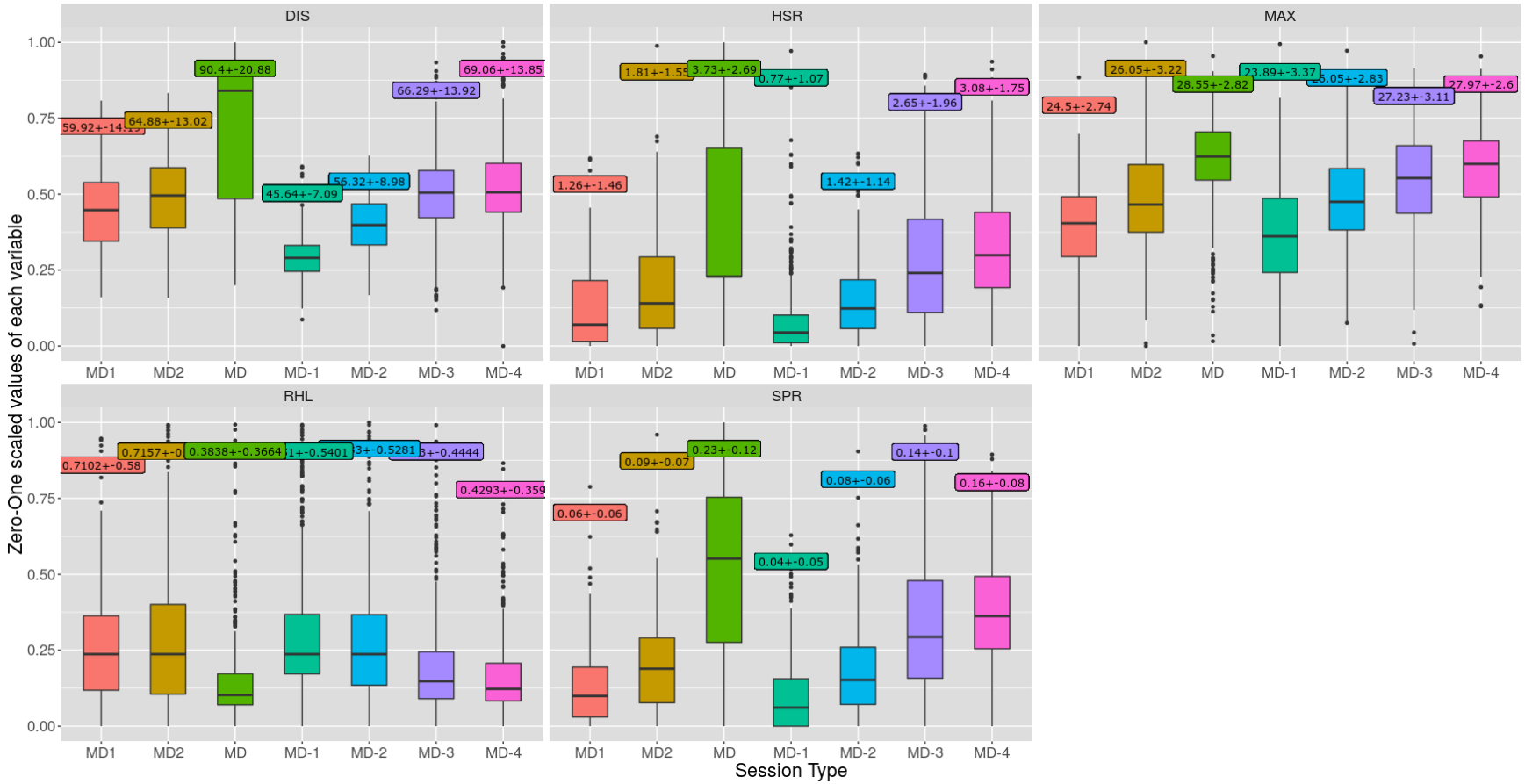


Figure 1.1: Boxplot distribution of the Locomotor physical variables. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

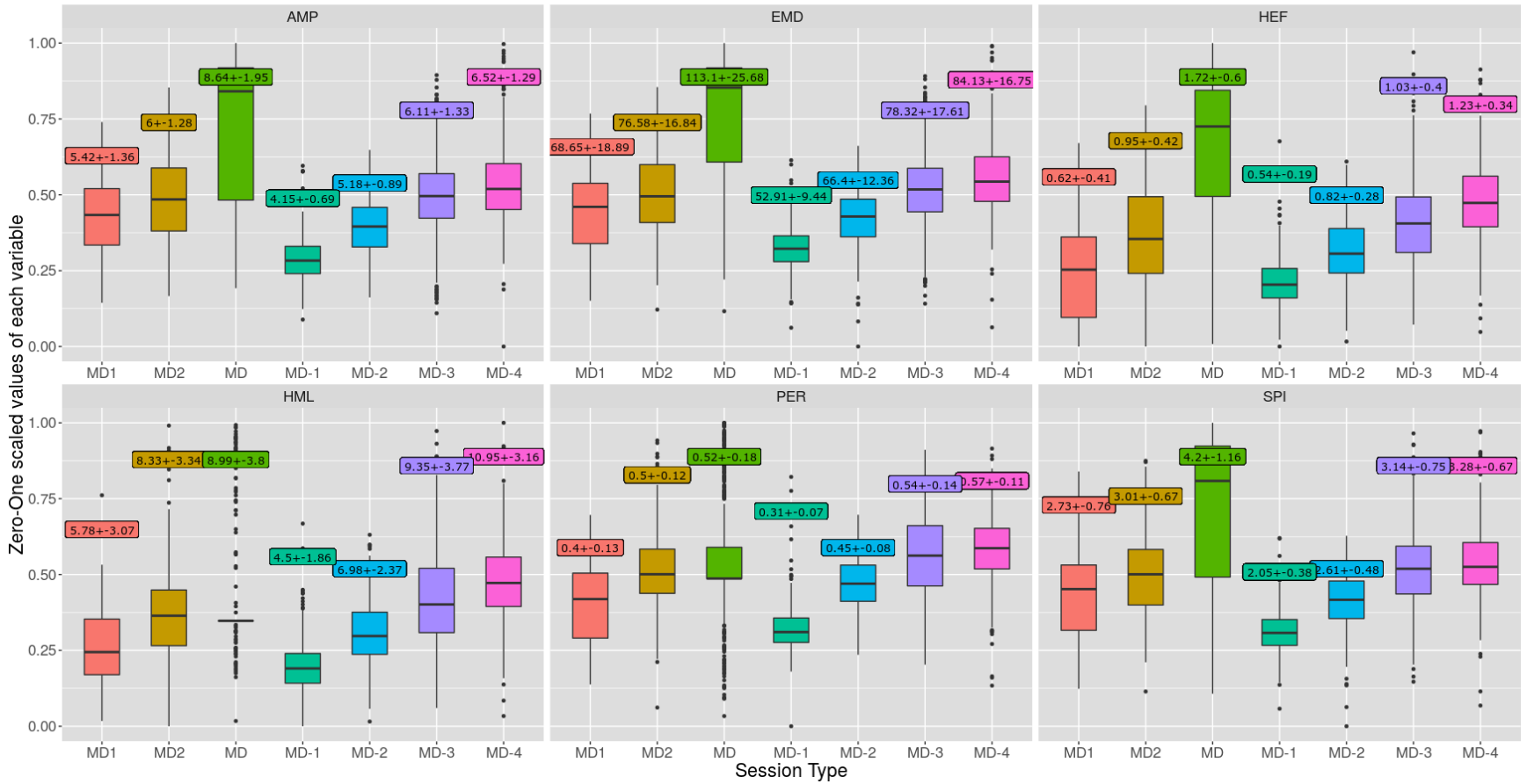


Figure 1.2: Boxplot distribution of the Metabolic physical variables. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

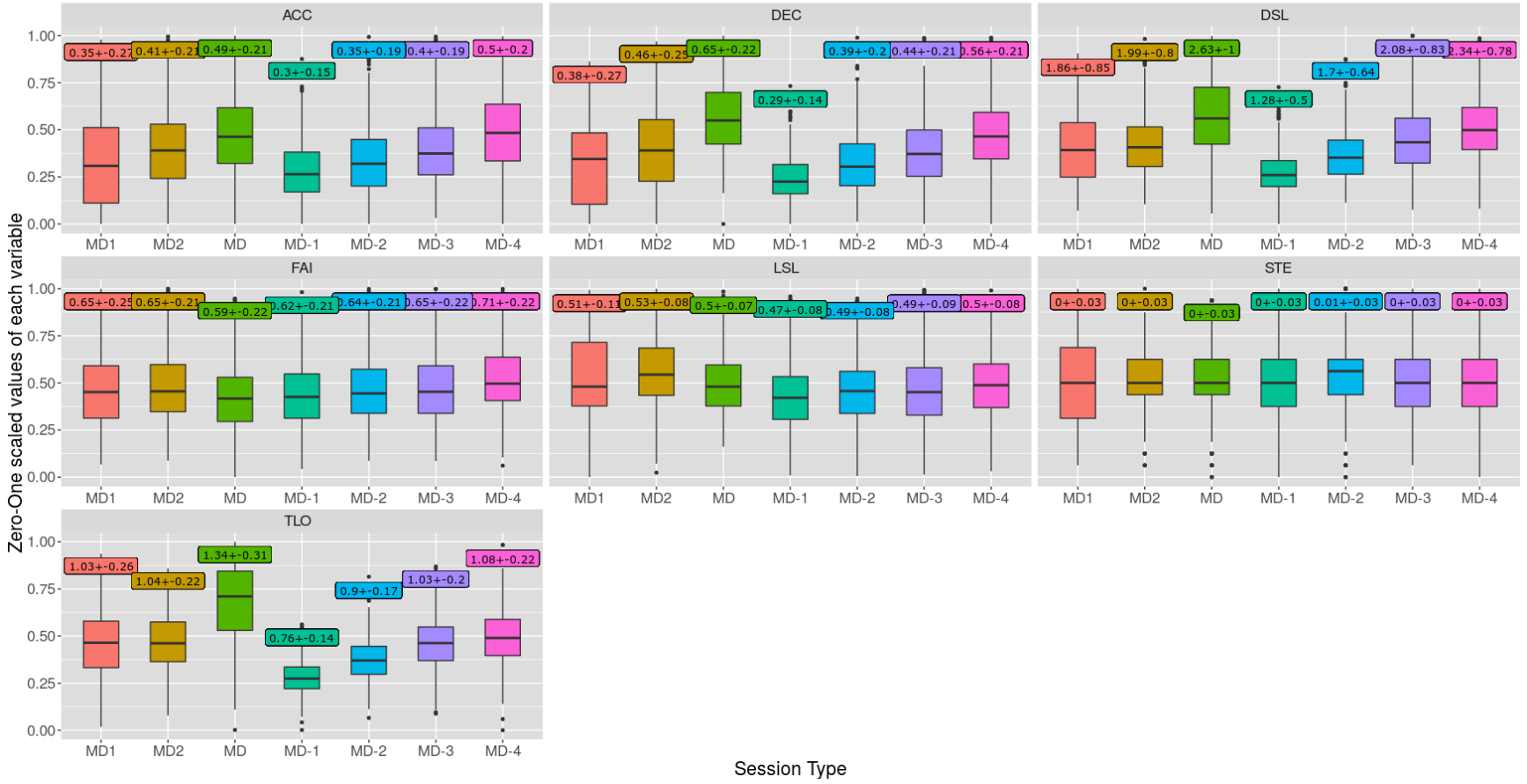


Figure 1.3: Boxplot distribution of the Mechanical physical variables. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

Beyond one on one comparison of variables, is interesting to analyse what can the relation among variables show with respect to differences between sessions. Figure 1.4, presents the first two principal components of data obtained by the application of principal component analysis (PCA), accounting for 66% of variance in data. Also, Figure 1.5 presents a similar representation, but using instead a two-dimensional reduction of the data obtained through t-distributed stochastic neighbour embedding method (t-SNE). From the PCA visualization it becomes clearly contrasted the differences between matches and training sessions. Although MD-3 and MD-4 sessions appear to be closer to MD data points, the differences between training sessions can not be clearly assessed. This is attained, most probably, to the rather low level of variance that is represented by the two first components. A level of variance higher than 95% is reached at the 9th principal component, among the 18 variables. The t-SNE plot presents a more clear visualization. MD is clearly separated from the rest, while MD-3 and MD-4 appear as the most similar sessions to matchday. MD-3 shows a wider dispersion of data, while for MD-4 data-points appear closer to each other. MD-2 also shows to be closer to these previous three, reflecting the higher intensity efforts demanded to players that do not participate in MD. MD-1 and MD-2 appear as practically non-separable, and far from MD demands. Although both dimensionality reduction techniques exploit linear relations on variables, the separation between different types of sessions becomes sufficiently clear to validate the behaviour expected by physical coaches.

Up to this point we have shown relations among session types using player information individually without any other kind of discrimination. In order to understand differences that might arise from player's specific positions, we analyse the distribution of data presented in Figures 1.6, 1.7 and 1.8. It can be observed that central midfielders (MC) tend to present higher values and variation for metabolic effort and locomotor measurements. This characteristic is shared by positions that incur in longer translations along the field, such as lower backs (LB) and wingers (AW/WN). For attacking positions, such MC, AW/WN and Strikers (ST) the varia-

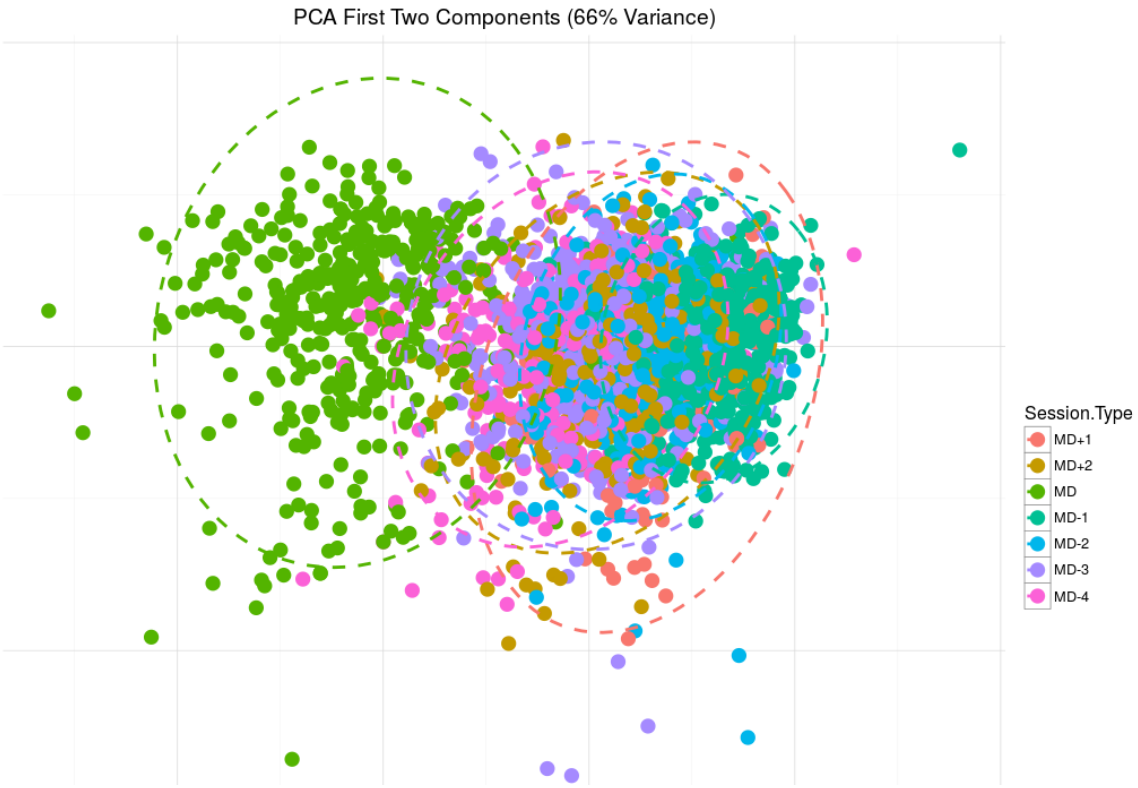


Figure 1.4: First two principal components of a PCA dimensionality reduction on data, comprehending 66% of variance

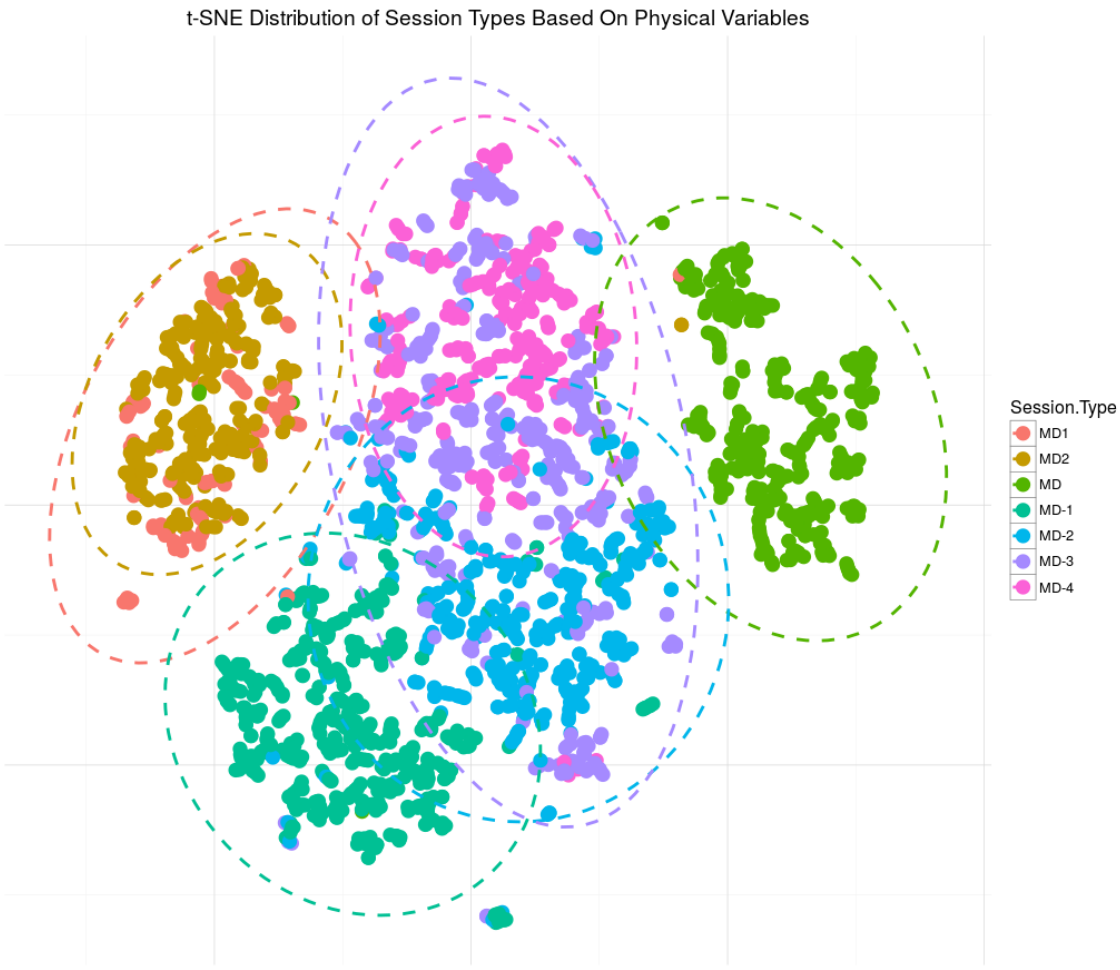


Figure 1.5: 2-dimensional plot of physical variables, highlighting session types, produced by t-SNE dimensionality reduction

tion of metabolic variables is higher than defending positions. Central defenders (CB) show a tendency for lower values and considerably lower variability than the rest of positions in all the three groups of variables, which is related to their more static role within the game. PER values show to be similar among positions, similarly to the sessions comparison performed previously. However the variability of this value is considerably high for all the positions, thus strengthening the idea of this variable being more dependent of match specific conditions and player's own fitness state. Although their exists a variation in absolute intensity effort for the different positions, there is no evidence of clear separation of demands and relation among variables, from one position to another. It becomes feasible to believe that data from players can be compared directly without strictly requiring separation between positions. This can be assumable despite the idea that separating player position might provide a more refined insight on player's response to training, which might, however, be impractical in the case of not having sufficient data for each position.

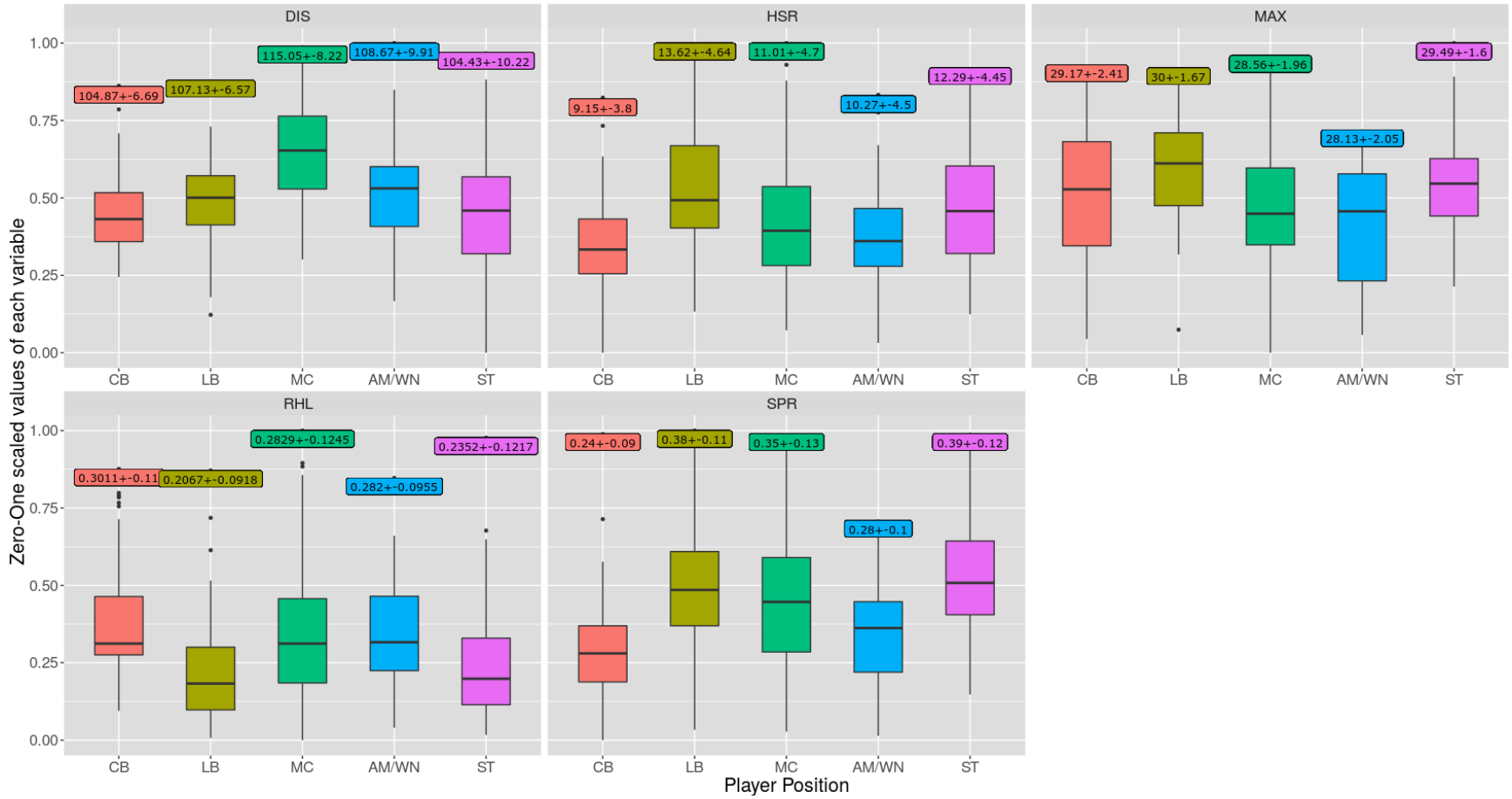


Figure 1.6: Boxplot distribution of the Locomotor physical variables on matchday, distributed by player position. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

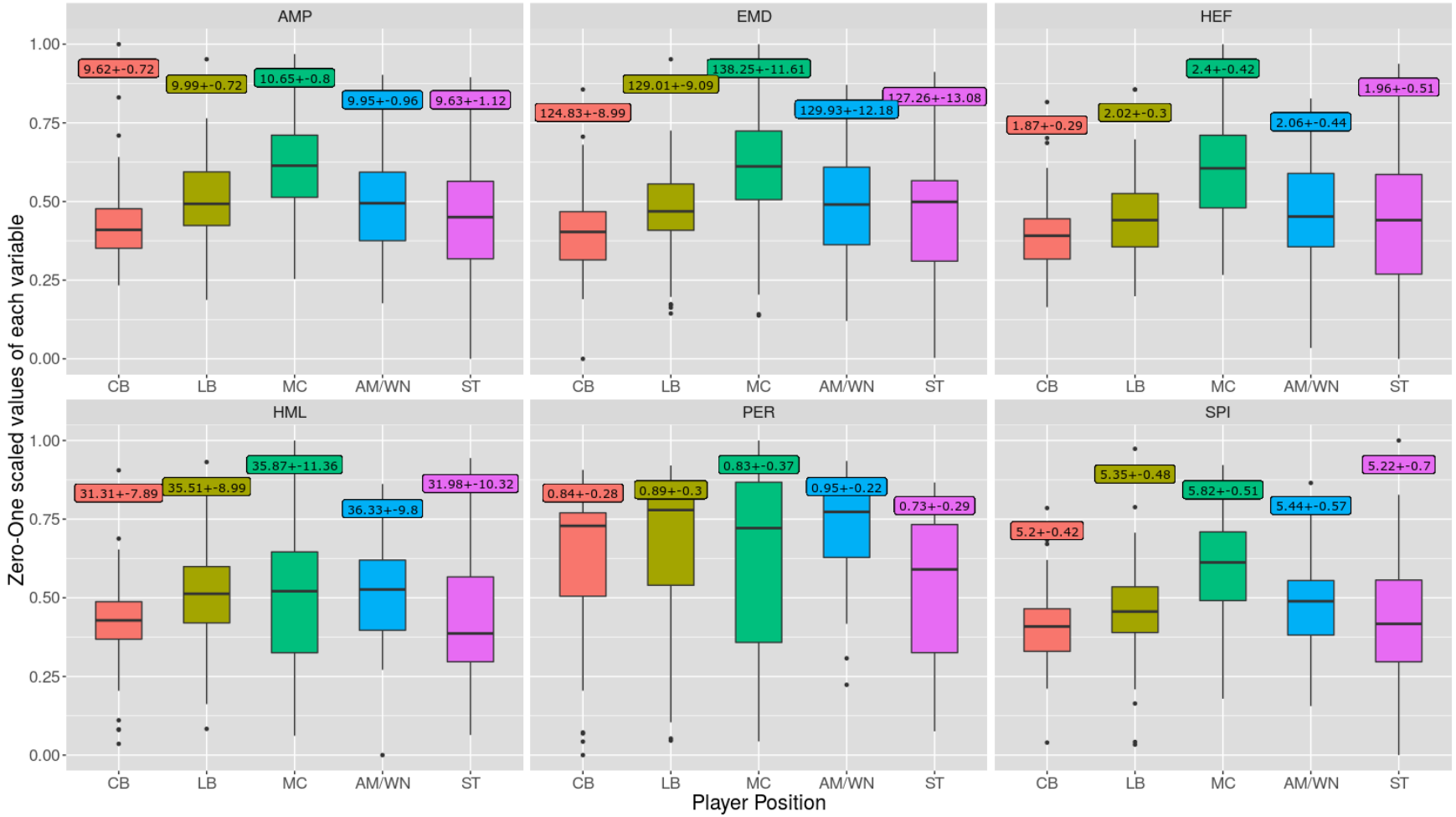


Figure 1.7: Boxplot distribution of the Metabolic physical variables on matchday, distributed by player position. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

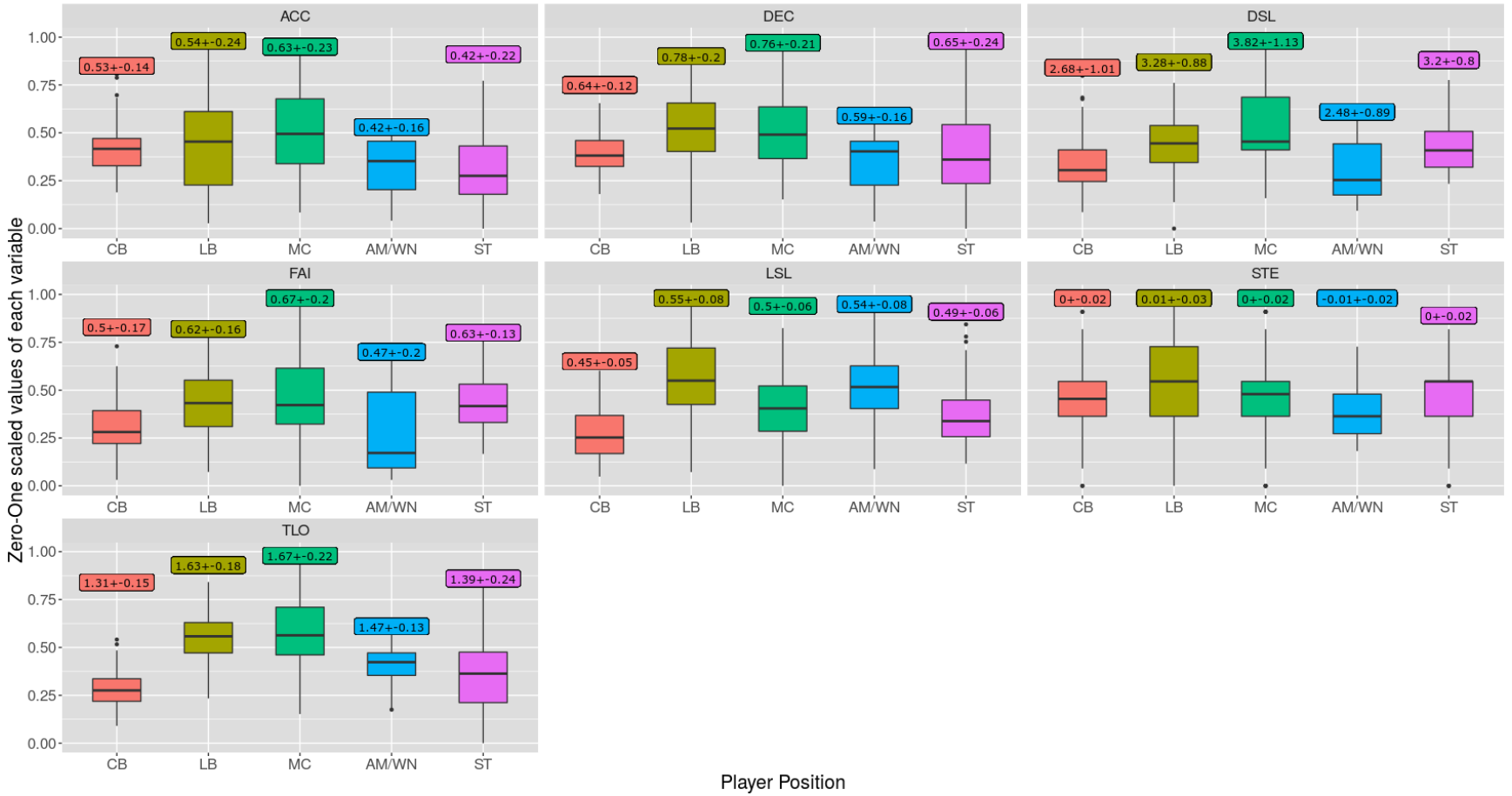


Figure 1.8: Boxplot distribution of the Mechanical physical variables on matchday, distributed by player position. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

Previous plots show a tendency of certain groups of variables to be distributed in similar ways, along sessions and positions. Also, the proposed segmentation of physical variables in groups is expected to evidence correlation between variables that provide from similar measurements. In order to assess this, Figure 1.9 presents the Pearson linear correlation of the selected variables, among the different groups. It can be observed that metabolic variables tend to present higher linear correlation among them. This is particular evident for Average Metabolic Power (AMP), which summarizes many of the information within other metabolic-related variables. Also, a tendency of medium to high correlation is presented in locomotor variables. For mechanical variables, correlation is lower, where the Fatigue Index (FAI) is very unlikely correlated with the rest of the variables, with exception of DSL which is specifically used for its calculation. A clear relation is also shown between locomotor variables Distance (DIS) and High Speed Running (HSR), and metabolic variables, since they are related with continued effort during larger distances. The evidence of correlations between variables is evident, and should be taken into account when performing multi-variate analysis on this data.

Another interesting element to analyse is the relation between historical information and match-related physical performance values. For different sizes of aggregation windows, three historical variables where created regarding average Fatigue, accumulated training minutes, and accumulated load. Table 5.1 presents for 7 different variables, if statistical differences where found when building matches physical variables against several natural clusters for each historical variables. The groups that are presenting significant differences between means are described, in order to guide the analysis of the corresponding plots. For simplicity just results with 6-week windows are shown, but 3-week and 9-week aggregation windows tend to show similar patterns. For each combination which presented significant differences boxplots of the distribution of the data where generated. Figures 1.10, 1.11 and 1.12, present the obtained results for fatigue, training minutes and load historical variables, respectively. In the case of historical fatigue, it can be seen that DSL and FAI (Fatigue) values

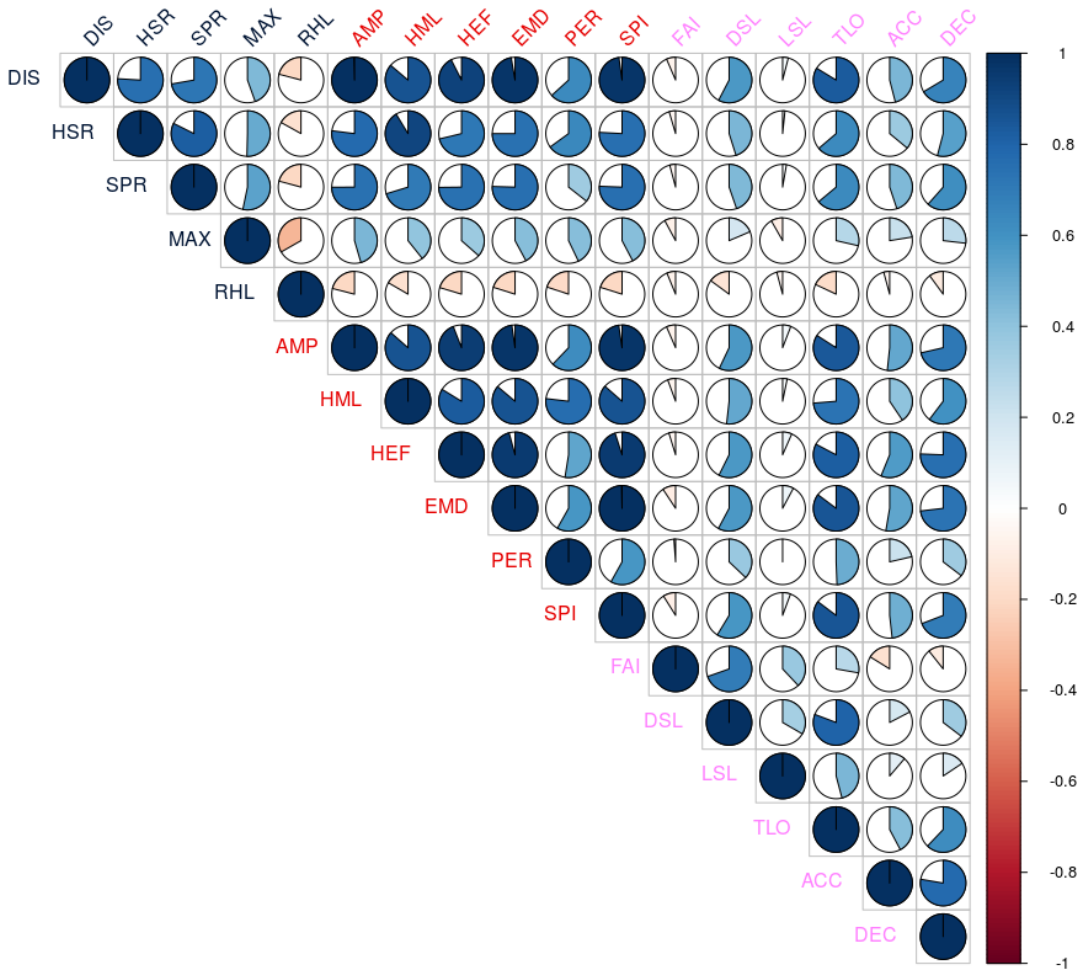
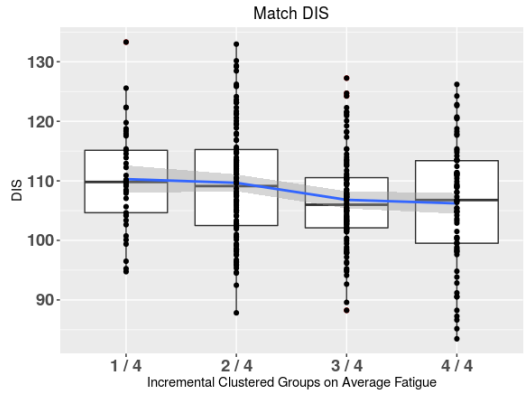


Figure 1.9: Pairwise Pearson correlation of the target variables from both training and matches data. Variables are organized following the three structured groups from top to bottom: locomotor (blue or dark grey), metabolic (red or medium dark grey), mechanical (pink or light grey). A filled circle refers to full correlation, where blue and red colors refer to positive or negative correlation respectively

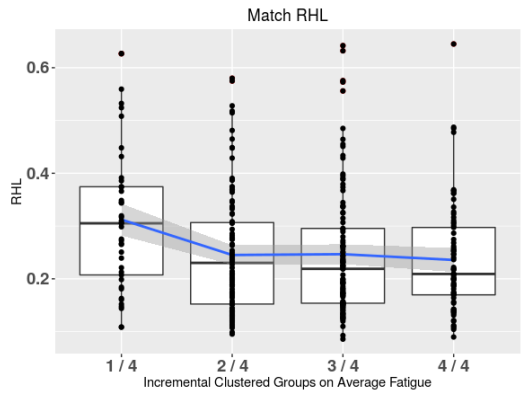
are linearly related to the increment of fatigue. So, at higher levels of fatigue during 6-weeks, players are showing higher levels of fatigue during matches. It can be seen that when the historical fatigue becomes greater the absolute values of ACC, AMP, RHL and DIS become lower. For AMP, DIS and ACC, each representing a different group of variables, there seems to exist a breaking point at the second lower level of historical fatigue. At this point, the higher values are registered, while the lower are shown at the lowest level of fatigue. This lowest level might also be associated to players that have trained the less minutes, so in possible lower fitness state. This observations might point onto the direction of the existence of an optimal fitness level in which higher physical efforts can be reached, and a less optimal state in which players, either by being tired or less trained cannot reach higher levels. It should be clarified, however, that higher physical registers are not a direct synonymous of better physical state.

Table 5.1: Comparison between clustered groups of historical variables and 7 selected match performance variables. Results of ANOVA test between groups is presented, along with the specific inter-group differences found through Tukey post-hoc test

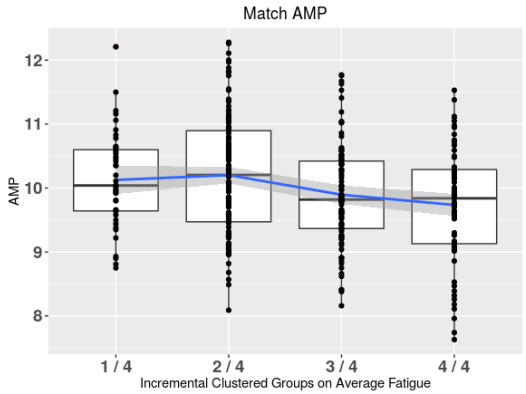
Variable	Historical Var. Filter	ANOVA $p < 0.05$	Post-hoc group differences
DIS	6-Week Avg. FAI	YES	2/4-1/4, 3/4-1/4
	6-Week Training Minutes	NO	
	6-Week PER	YES	3/3-2/3
RHL	6-Week Avg. FAI	YES	3/4-1/4
	6-Week Training Minutes	YES	3/3-2/3
	6-Week PER	YES	2/3-1/3
AMP	6-Week Avg. FAI	YES	2/4-1/4, 3/4-1/4
	6-Week Training Minutes	NO	-
	6-Week PER	YES	3/3-2/3
PER	6-Week Avg. FAI	NO	-
	6-Week Training Minutes	NO	-
	6-Week PER	YES	2/3-1/3
FAI	6-Week Avg. FAI	YES	2/4-1/4, 3/4-1/4, 3/4-2/4, 4/4-2/4, 4/4-3/4
	6-Week Training Minutes	NO	-
	6-Week PER	NO	-
ACC	6-Week Avg. FAI	YES	2/4-1/4
	6-Week Training Minutes	YES	2/3-1/3
	6-Week PER	YES	3/3-2/3
DSL	6-Week Avg. FAI	YES	2/4-1/4, 3/4-1/4, 3/4-2/4, 4/4-2/4, 4/4-3/4
	6-Week Training Minutes	NO	–
	6-Week PER	NO	-



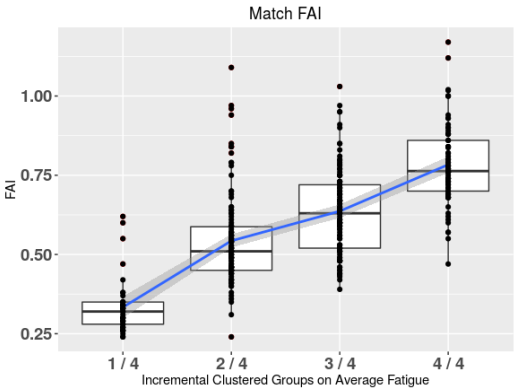
(a) DIS against 6-Week Fatigue



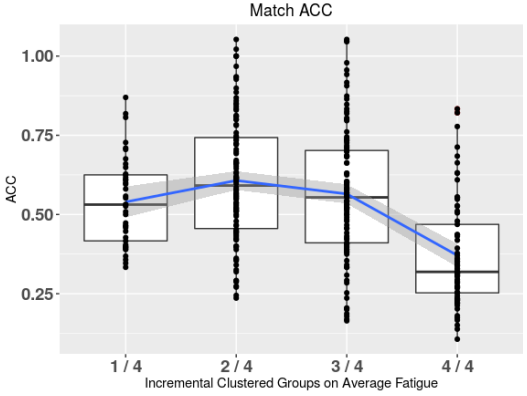
(b) RHL against 6-Week Fatigue



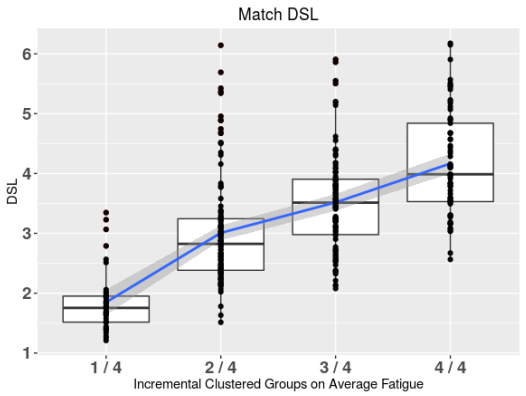
(c) AMP against 6-Week Fatigue



(d) FAI against 6-Week Fatigue

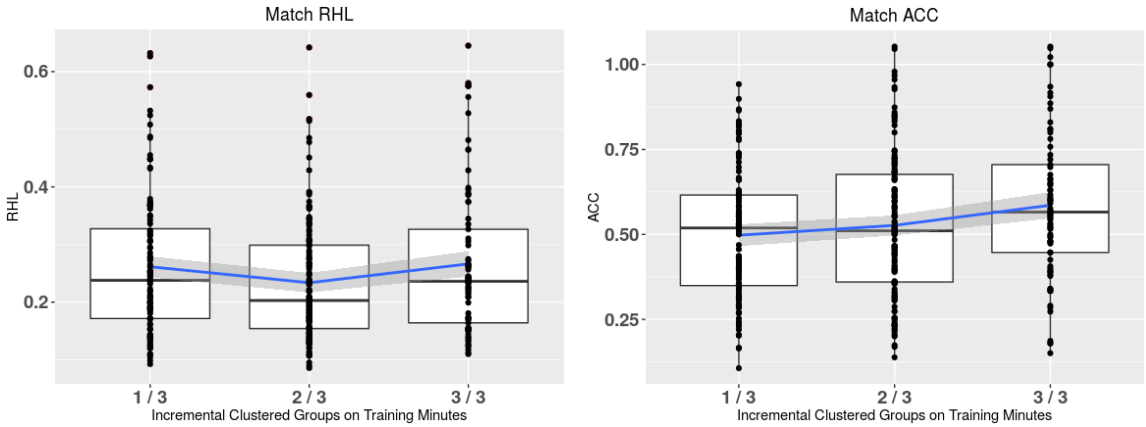


(e) ACC against 6-Week Fatigue



(f) DSL against 6-Week Fatigue

Figure 1.10: Comparison of 6 Variables Split By 4 Different Incremental Groups of 6-Week Average Fatigue

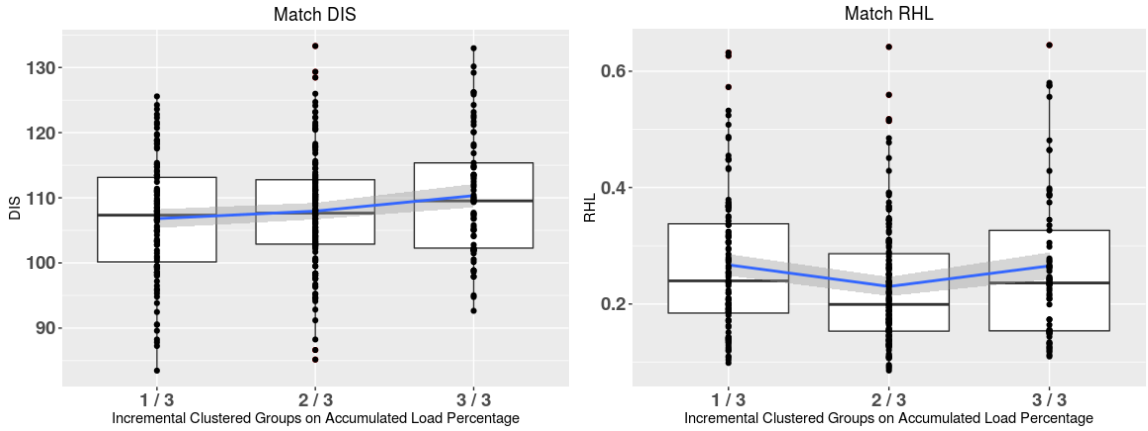


(a) RHL against 6-Week Training Minutes (b) ACC against 6-Week Training Minutes

Figure 1.11: Comparison of 2 Variables Split By 3 Different Incremental Groups of 6-Week Training Minutes

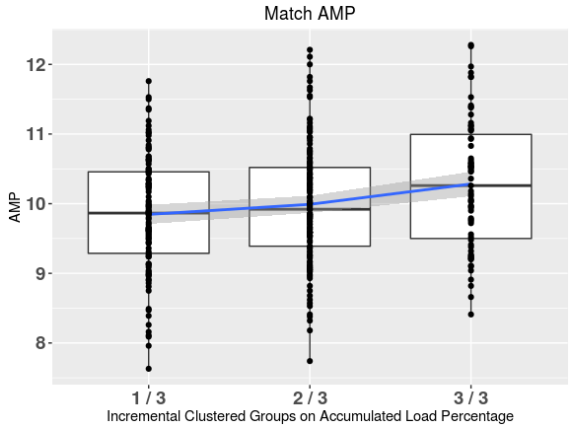
For accumulated load, is interesting to observe that when load increases in this period of time, the three group representative variables DIS, AMP and ACC, tend to show higher values. This might respond to the idea of the influence of match minutes in the overall physical conditioning of players, which is addressed below. In the case of training minutes, no remarkable pattern can be found, beyond the idea that when accumulating higher training time the number of accelerations in the match slightly increment.

Figure 1.13 contrast the influence of historical variables, physical variables, and players that played less and more match minutes. This is interesting since match minutes are seen as a critical part of player conditioning, but also as a demanding activity for players, so important differences arises. It can be observed that for players that have played the lower number of match minutes, but that have participated in at least one match in this time window, values tend to be higher and also wider disperse for ACC, DIS and PER shown variables. The number of ACC its notably increased with higher accumulated load, but is higher for player that played less. This could be due to higher presence of fatigue for players who present higher loads and have played most matches. For players showing higher levels of fatigue is notable that the registered match ACC tends to be lower, and also finding a possible optimal state

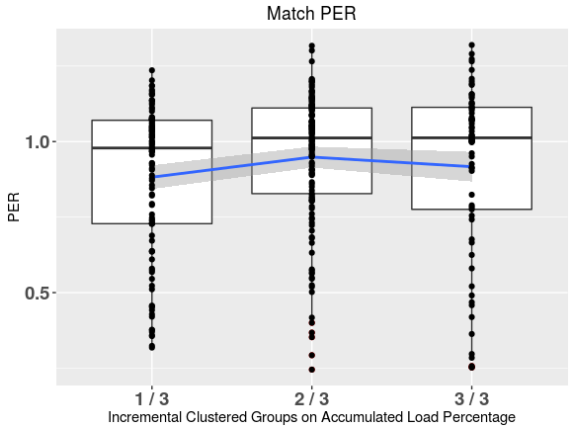


(a) DIS against 6-Week Load

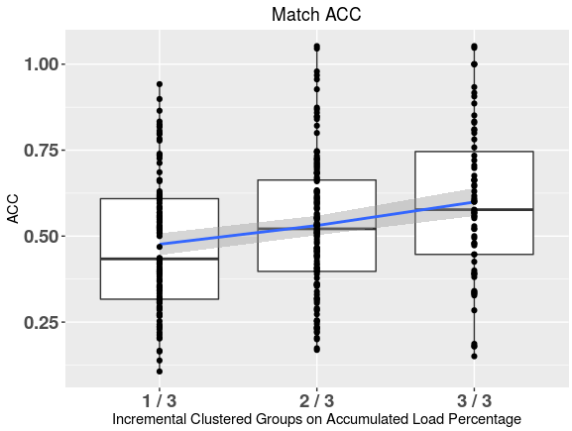
(b) RHL against 6-Week Load



(c) AMP against 6-Week Load



(d) PER against 6-Week Load



(e) ACC against 6-Week Load

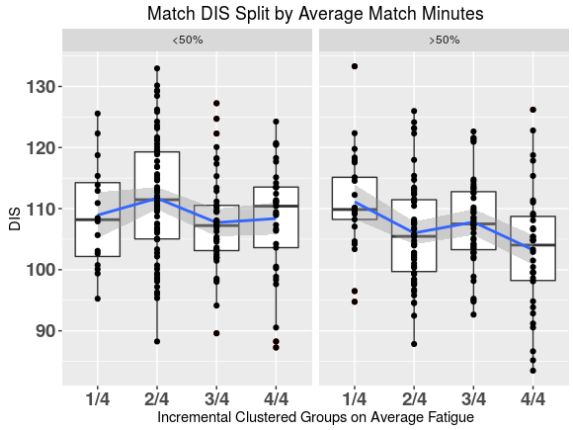
Figure 1.12: Comparison of 5 Variables Split By 3 Different Incremental Groups of 6-Week Accumulated Load Percentage

at second fatigue level, such as previously described. Is interesting to observe that players presenting lower accumulated load presents higher levels of load in matches if they are in the group who played the most. This evidences the influence of matches for physical conditioning, and its impact in players fitness.

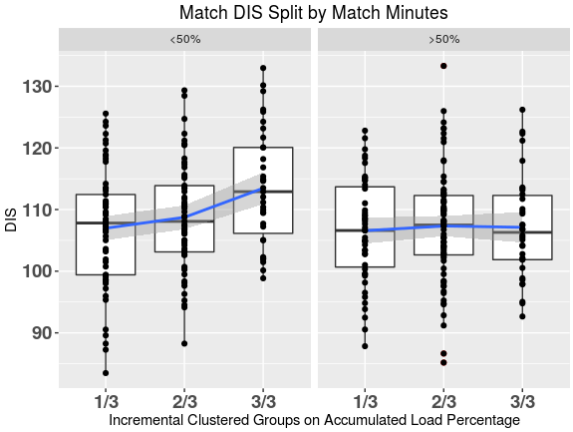
The last part of this phase addresses the expected oscillatory tendency along a season time which might respond to the physiological nature of players and training design. Figure 1.14 presents the mean values for the team and different positions for 7 physical variables representing the three proposed categories. For most variables the oscillatory tendency is clear, with higher and lower pikes along the season, and without showing chaotic behaviour. It is critical to observe that in comparison with team mean, the positions do not present much differences between them, and tend to follow a similar path as the team mean. This adds value to the believe that, although players have differences in their individual state of fitness, coaches aim to guide players through a designed schema. However, individual differences among players is expected to be found. Consistently with previous results, FAI and DSL do not present the same tendency as directly registered values. This adds up to the idea that these variables are reflecting physical information that is less possible to control and more related with the players own fitness state.

5.2 Unsupervised Analysis Results

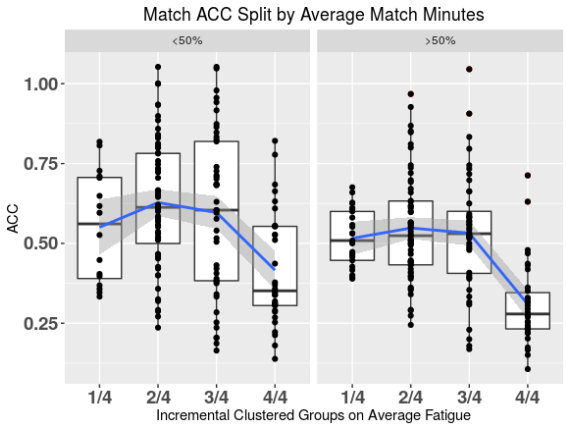
Results for both the time-series and the summarized datasets are presented together, since they follow an identical approach in its evaluation. After performing the transformations described in Section 3.3 on the training dataset, the sample size of the obtained training sessions dataset, related to weekly variations, is 112. On the other hand, the resulting matches dataset, produced by association to training session windows consists of a sample size of 82. It must be observed that associated matches have a lower sample size, since not all the players who fitted into the match previous 3-week window actually played the next match. From this weekly variation dataset



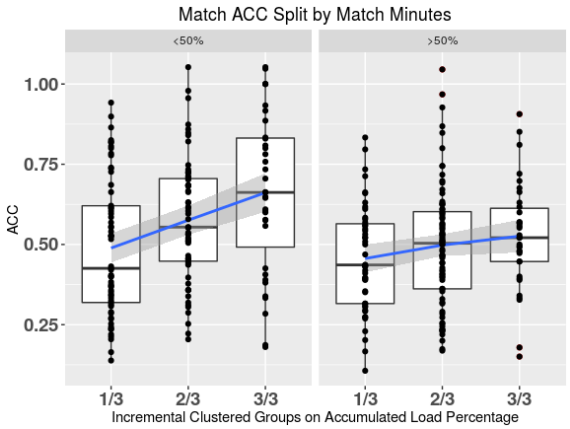
(a) DIS against 6-Week Fatigue



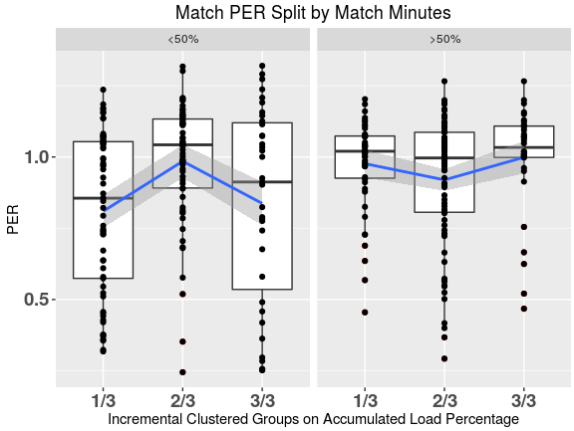
(b) DIS against 6-Week Accumulated Load



(c) ACC against 6-Week Fatigue



(d) ACC against 6-Week Accumulated Load



(e) PER against 6-Week Accumulated Load

Figure 1.13: Comparison of 6 Variables Split By 4 Different Incremental Groups of 6-Week Average Fatigue

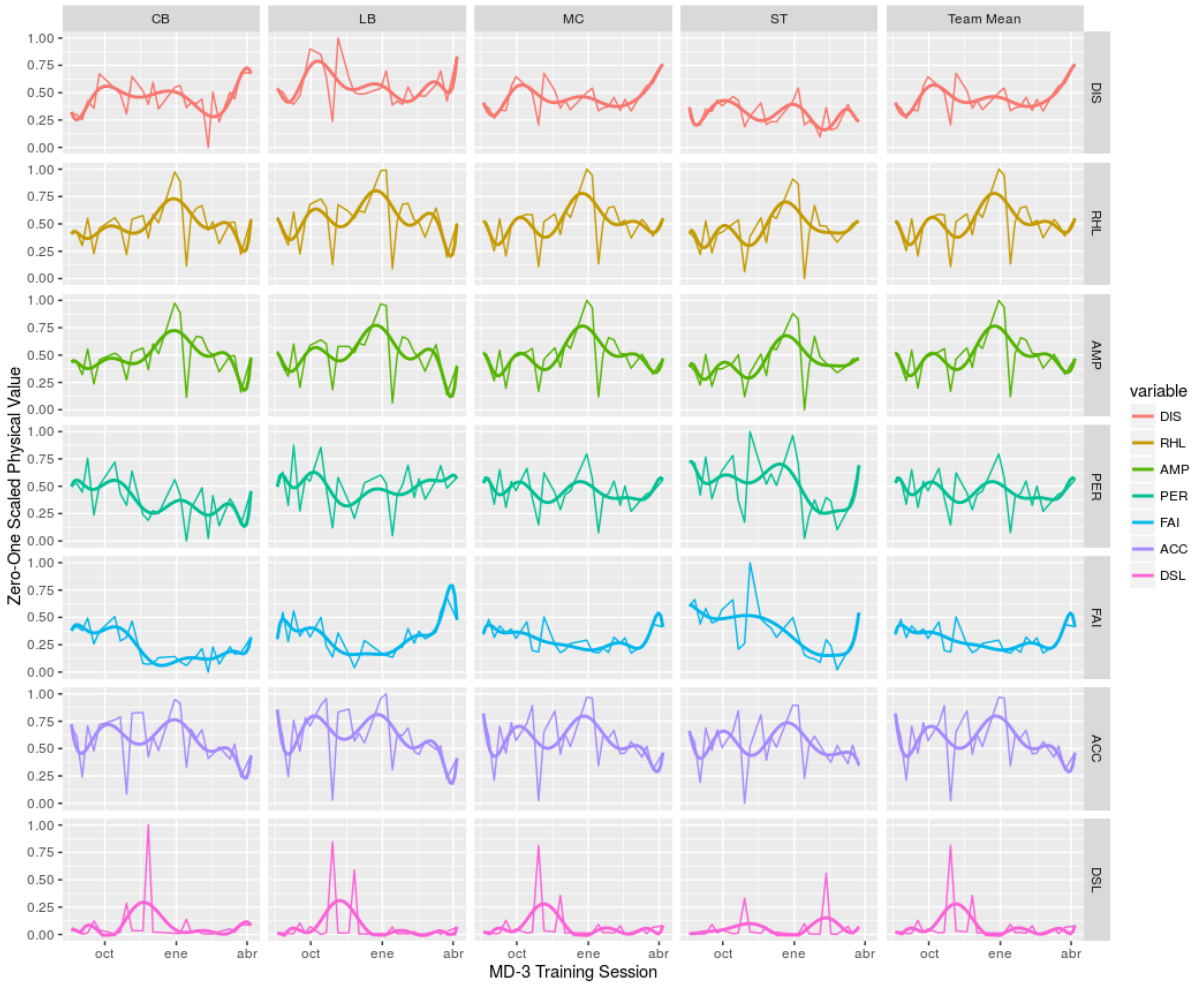


Figure 1.14: Mean values registered by the team and segmented by every position in the field, for 7 physical variables from the three variable groups. Values are scaled in a [0..1] range.

the time-series and summarized datasets versions where obtained. It is critical to recall, as explained previously, that both derived datasets contain the same sample size. Using the mentioned five different indices, the optimal number of clusters was selected in both cases. For the time-series dataset, the optimal K value for the K-Mediods algorithm was 2; obtained by running the algorithm over the distance matrix obtained through Dynamic Time Warping. For the summarized dataset, the optimal K value according to the K-means algorithm was also 2. In both cases, this number of clusters that reduces the most each internal distance, was selected by four of the five different indices. Naturally, the exact examples belonging to the found clusters are different. For each example, the corresponding match performance example is labelled with the associated cluster number. For each of the variables conforming the two groups (in each dataset) the standardized difference of means was calculated for describing the effect size, using the limits detailed in Section 4.3.

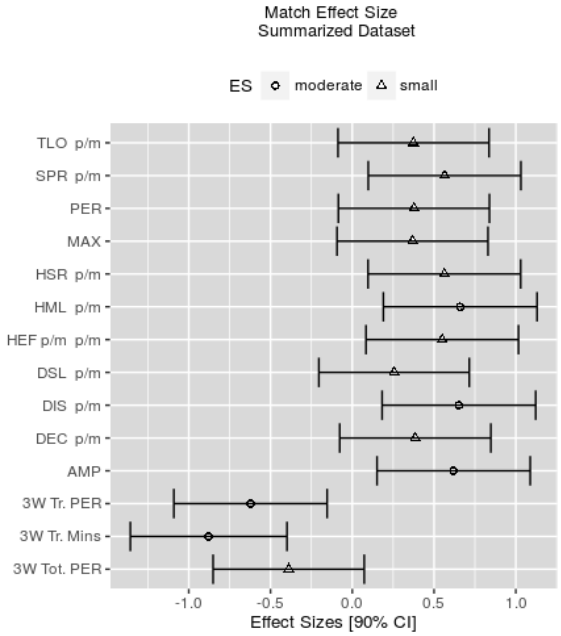
Detailed results from the training data clustering and the associated matches groups are presented in table 5.3. It can be clearly observed that for the summarized dataset almost every variable in training registered a moderate to large effect size in the first cluster in comparison with the second. Moreover, this effect is leaned onto the same side for most variables. We are observing the detection of two groups: one where the average magnitude of variations of each variable is higher (*high variation group*), and one where the variations are lower (*low variation group*). It is critical to observe that separation among groups (for each variable) is not absolute, and there exists ranges of values which overlap. This has to do with multivariate nature of the clustering procedure, and coincides with the original expectation of this study. It can also be observed that for the time-series dataset few variables where able to stand out just with a small size effect. Even with the selection of euclidean distance to favor magnitudes, the cluster analysis over the DTW matrix was not able to find a clear separation between groups. The procedure over the summarized dataset, instead, did find a considerably separation between training groups so the analysis over associated matches is easier to interpret and translate to practical applications.

Figure 2.15 presents the effect sizes for the associated matches in both datasets. It can be observed for both cases that variables registering high intensity efforts, energy consumption and distance travelled appear with higher magnitude in the *high variation group* consistently, while the total load percentage and training minutes in the previous three weeks are considerably low in this same group. HMP, AMP and DIS present moderate effect size in the summarized dataset while in the time-series dataset just HML presents a moderate effect size, toward the same tendency. The rest of variables registering intensity efforts show just a small effect size. Three-weekly PER and training minutes show also a moderate effect in differences, towards lower values. Sample size for associated matches allows to conclude with certainty about moderate size effects. Small effects should be taken into account, but must be further validated with the future increase of availability of data. Although it was not an objective of the planned experiments, it is important to remark that higher variation and lower variation tendencies where not associated with specific players or positions, but rather distributed among different individuals (with repetition) along the season.

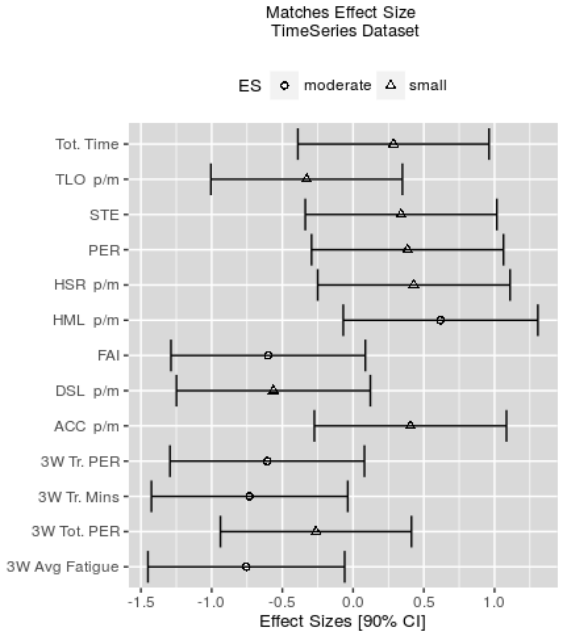
This results are showing an interesting tendency: players who present higher variation in 3-week windows during training sessions, are also presenting higher registers for 11 over 15 physical variables. And also, these players who present higher variation are associated with lower values of training minutes and accumulated load, in comparison with the team (among players who have participated in matches in that period). The association with matches is performed blindly, in that sense that no match information is used for this calculations. Thus, is considered remarkable to find actual differences in the mean of both matches groups, specially considering the previously stated differences between MD-3 training sessions and matchdays. This provides a first insight on the relations between training adaptability and match performance, that should be validated with further availability of new data.

Table 5.2: Mean and standard deviation for each physical variable in each of the clustered groups. For both training data and the associated matches, values obtained in both summarized and timeseries datasets are presented. The standardized difference of means *SDM* is presented for each case. Training results refer to the absolute average of variation while matches results refer to the actual measured physical values.

	Training (mean ± SD)						Matches (mean ± SD)					
	Summarized			Timeseries			Summarized			Timeseries		
Variable	Cluster 1	Cluster 2	SDM	Cluster 1	Cluster 2	SDM	Cluster 1	Cluster 2	SDM	Cluster 1	Cluster 2	SDM
DSL p/m	0.89 ± 0.29	0.58 ± 0.28	*Moderate*	0.41 ± 0.198	0.41 ± 0.18	*Trivial*	3.60 ± 1.30	3.26 ± 1.37	*Small*	3.37 ± 1.17	4.29 ± 1.84	*Small*
ACC p/m	1.011 ± 0.45	0.67 ± 0.31	*Moderate*	0.45 ± 0.26	0.49 ± 0.23	*Trivial*	0.56 ± 0.22	0.52 ± 0.24	*Trivial*	0.65 ± 0.23	0.561 ± 0.21	*Small*
DEC p/m	0.94 ± 0.38	0.51 ± 0.23	*Large*	0.43 ± 0.25	0.38 ± 0.19	*Trivial*	0.80 ± 0.27	0.69 ± 0.26	*Small*	0.82 ± 0.29	0.785 ± 0.307	*Trivial*
SPR p/m	0.74 ± 0.35	0.59 ± 0.28	*Small*	0.36 ± 0.17	0.40 ± 0.20	*Trivial*	0.37 ± 0.11	0.311 ± 0.09	*Small*	0.37 ± 0.12	0.38 ± 0.20	*Trivial*
HSR p/m	0.98 ± 0.46	0.53 ± 0.26	*Large*	0.39 ± 0.17	0.43 ± 0.28	*Trivial*	13.30 ± 4.73	10.78 ± 4.31	*Small*	12.58 ± 5.27	10.686 ± 3.837	*Small*
AMP	0.62 ± 0.36	0.29 ± 0.14	*Large*	0.28 ± 0.10	0.25 ± 0.13	*Small*	10.35 ± 1.08	9.65 ± 1.17	*Moderate*	10.27 ± 1.09	10.26 ± 1.50	*Trivial*
HML	0.58 ± 0.24	0.40 ± 0.20	*Moderate*	0.33 ± 0.14	0.25 ± 0.13	*Small*	39.12 ± 10.05	32.05 ± 10.01	*Moderate*	36.09 ± 9.11	30.892 ± 7.95	*Moderate*
HEF p/m	0.74 ± 0.31	0.40 ± 0.22	*Large*	0.34 ± 0.17	0.29 ± 0.19	*Small*	2.23 ± 0.52	1.94 ± 0.5	*Small*	2.31 ± 0.5	2.20 ± 0.64	*Trivial*
FAI	0.93 ± 0.39	0.71 ± 0.34	*Moderate*	0.47 ± 0.24	0.48 ± 0.21	*Trivial*	0.62 ± 0.19	0.64 ± 0.25	*Trivial*	0.602 ± 0.18	0.75 ± 0.29	*Moderate*
DIS p/m	0.58 ± 0.36	0.27 ± 0.15	*Large*	0.25 ± 0.10	0.21 ± 0.20	*Small*	111.5 ± 10.98	104 ± 11.3	*Moderate*	109.5 ± 10.72	110.71 ± 15.41	*Trivial*
TLO p/m	0.83 ± 0.25	0.39 ± 0.21	*Large*	0.35 ± 0.16	0.32 ± 01.9	*Small*	1.59 ± 0.23	1.48 ± 0.32	*Small*	1.56 ± 0.22	1.67 ± 0.38	*Small*
MAX	0.80 ± 0.39	0.63 ± 0.32	*Trivial*	0.40 ± 0.198	0.42 ± 0.22	*Trivial*	29.6 ± 2.11	28.79 ± 2.32	*Small*	29.612 ± 2.57	29.28 ± 1.61	*Trivial*
STE	1.05 ± 0.57	0.97 ± 0.53	*Trivial*	0.61 ± 0.35	0.596 ± 0.317	*Trivial*	0.006 ± 0.03	0.008 ± 0.026	*Trivial*	0.012 ± 0.022	0.002 ± 0.029	*Small*
PER	0.57 ± 0.35	0.27 ± 0.15	*Large*	0.23 ± 0.12	0.22 ± 0.19	*Trivial*	0.96 ± 0.25	0.85 ± 0.29	*Small*	0.86 ± 0.31	0.73 ± 0.34	*Small*
3W Training PER	–	–	–	–	–	–	5.26 ± 1.24	8.29 ± 1.68	*Moderate*	7.04 ± 1.09	7.45 ± 1.79	*Moderate*
3W Training Minutes	–	–	–	–	–	–	796 ± 171	964 ± 201	*Moderate*	725 ± 186.50	873.30 ± 202.87	*Moderate*
3W Total PER	–	–	–	–	–	–	7.71 ± 1.08	8.29 ± 1.68	*Small*	7.04 ± 1.09	7.45 ± 1.79	*Small*
3W Average FAI	–	–	–	–	–	–	0.65 ± 0.14	0.67 ± 0.18	*Trivial*	0.629 ± 0.14	0.77 ± 0.22	*Moderate*



(a) Effect sizes in matches' clusters for the summarized dataset



(b) Effect sizes in matches' clusters for the time-series dataset

Figure 2.15: Effect size differences in group mean values in standardize units for matches groups found through the summarized dataset (a) and the time-series dataset (b). Trivial effect sizes are not shown.

5.3 Supervised Analysis Results

In this section we present the results from the regression analysis for predicting match performance variables, and the subsequent analysis of variable importance.

5.3.1 Variable Prediction

The results from a applying the different mentioned algorithms, feature selection, and dimensionality reduction methods are presented in Table 5.3 and Table 5.4, using the NRMSE metric described earlier. Values under 0.75 NRMSE are considered good results in the sense that they can be translated into practice. This threshold was arbitrarily selected together with physical coaches. The desired threshold was achieved in 11 out of 17 target variables, mostly distributed among metabolic and mechanical groups. From the locomotor variables group it can be seen that only DIS was able to be successfully predicted, but results were below threshold for the other 4 variables (HSR, SPR, MAX, RHL). This situation might respond to a high association of these variables with specific match dynamics beyond the current fitness state of the player such as the opposition team's tactical game, the score or any other variable beyond the strictly physical performance.

The algorithms exploiting non-linear relations among the variables such as Random Forest and RBF-Kernel SVM showed significantly better results that the linear approaches, and achieved a successful threshold in most of the combinations. Also, the feature selection method based on removing highly correlated variables (COR) showed to be a critical resource in this set of combinations, helping to achieve the best result in each of the successfully predicted variables. Recursive feature elimination (RFE) allowed to improve slightly most of the results, however its high computational cost provides doubt regarding its usefulness in this context. Principal Component Analysis (PCA) did not provide a considerable improvement with the exception of few isolated cases. It is noticeable that, for most of the models performing under 0.75 NRMSE, the variation of prediction among folds of the outer loop from the nested

Table 5.3: Mean prediction error and standard deviation in NRMSE units among folds, for non-linear algorithms. Dark Gray cells indicate the best NRMSE, and Light Gray cells the models achieving under 0.75 NRMSE

.

Variable	Random Forest			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.74 ± 0.07	0.64 ± 0.05	0.80 ± 0.10	0.66 ± 0.06
HSR (LC)	0.97 ± 0.02	0.99 ± 0.05	0.99 ± 0.06	1.03 ± 0.06
SPR (LC)	0.94 ± 0.04	0.87 ± 0.04	0.89 ± 0.05	0.88 ± 0.03
MAX (LC)	1.12 ± 0.22	0.86 ± 0.07	1.09 ± 0.12	0.93 ± 0.05
RHL (LC)	1.33 ± 0.25	1.15 ± 0.05	1.39 ± 0.22	1.23 ± 0.05
AMP (MB)	0.71 ± 0.02	0.62 ± 0.05	0.72 ± 0.03	0.60 ± 0.03
HML (MB)	1.02 ± 0.07	1.01 ± 0.05	1.04 ± 0.00	1.03 ± 0.06
HEF (MB)	0.77 ± 0.02	0.69 ± 0.02	0.76 ± 0.07	0.70 ± 0.03
EMD (MB)	0.79 ± 0.03	0.70 ± 0.05	0.79 ± 0.02	0.72 ± 0.04
PER (MB)	0.92 ± 0.06	0.80 ± 0.06	0.95 ± 0.04	0.79 ± 0.04
SPI (MB)	0.76 ± 0.03	0.67 ± 0.03	0.80 ± 0.04	0.70 ± 0.03
FAI (MC)	0.72 ± 0.03	0.71 ± 0.01	0.85 ± 0.06	0.72 ± 0.01
DSL (MC)	0.68 ± 0.02	0.80 ± 0.05	0.93 ± 0.06	0.77 ± 0.05
LSL (MC)	0.98 ± 0.11	0.96 ± 0.05	1.03 ± 0.08	0.99 ± 0.12
TLO (MC)	0.69 ± 0.03	0.77 ± 0.04	0.87 ± 0.02	0.72 ± 0.04
ACC (MC)	0.64 ± 0.04	0.65 ± 0.04	0.80 ± 0.04	0.63 ± 0.04
DEC (MC)	0.70 ± 0.01	0.64 ± 0.02	0.79 ± 0.05	0.64 ± 0.03

Variable	RBF-K SVM		
	PLAIN	COR	COR+RFE
DIS (LC)	0.66 ± 0.06	0.67 ± 0.08	0.67 ± 0.04
HSR (LC)	1.03 ± 0.06	0.99 ± 0.10	0.98 ± 0.08
SPR (LC)	0.88 ± 0.03	0.87 ± 0.02	0.84 ± 0.04
MAX (LC)	0.93 ± 0.05	0.92 ± 0.03	0.92 ± 0.04
RHL (LC)	1.23 ± 0.05	1.01 ± 0.10	1.00 ± 0.10
AMP (MB)	0.60 ± 0.03	0.71 ± 0.07	0.66 ± 0.05
HML (MB)	1.03 ± 0.06	1.02 ± 0.06	0.96 ± 0.07
HEF (MB)	0.70 ± 0.03	0.66 ± 0.05	0.71 ± 0.04
EMD (MB)	0.72 ± 0.04	0.67 ± 0.03	0.69 ± 0.04
PER (MB)	0.79 ± 0.04	0.70 ± 0.12	0.70 ± 0.07
SPI (MB)	0.70 ± 0.03	0.67 ± 0.04	0.68 ± 0.05
FAI (MC)	0.72 ± 0.01	0.73 ± 0.01	0.79 ± 0.04
DSL (MC)	0.77 ± 0.05	0.83 ± 0.08	0.86 ± 0.08
LSL (MC)	0.99 ± 0.12	0.98 ± 0.02	0.99 ± 0.02
TLO (MC)	0.72 ± 0.04	0.79 ± 0.06	0.85 ± 0.05
ACC (MC)	0.63 ± 0.04	0.66 ± 0.04	0.68 ± 0.03
DEC (MC)	0.64 ± 0.03	0.68 ± 0.05	0.66 ± 0.05

cross validation approach was considerably low. The low variation of prediction can be associated with a high stability of the model and also validates the correctness of the parameter selection approach.

5.3.2 Variable Importance

Observing the results of regression analysis it becomes clear that Random Forest is a reasonable choice for variable importance calculation. In terms of NRMSE, Random Forest produced the best performing or second best performing models, in most

Table 5.4: Mean prediction error and standard deviation in NRMSE units among folds, for linear algorithms. Dark Gray cells indicate the best NRMSE, and Light Gray cells the models achieving under 0.75 NRMSE

Variable	Linear SVM			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.67 ± 0.08	1.86 ± 0.63	0.82 ± 0.18	0.82 ± 0.18
HSR (LC)	0.99 ± 0.09	1.92 ± 0.50	1.07 ± 0.23	1.10 ± 0.23
SPR (LC)	0.87 ± 0.03	2.05 ± 0.71	0.94 ± 0.07	0.95 ± 0.07
MAX (LC)	0.91 ± 0.03	2.89 ± 0.98	0.99 ± 0.12	1.00 ± 0.12
RHL (LC)	1.00 ± 0.08	1.84 ± 0.48	1.83 ± 0.06	0.97 ± 0.06
AMP (MB)	0.78 ± 0.18	1.37 ± 0.40	0.80 ± 0.06	0.66 ± 0.06
HML (MB)	1.02 ± 0.06	1.79 ± 0.39	0.99 ± 0.08	1.02 ± 0.08
HEF (MB)	0.66 ± 0.06	1.62 ± 0.07	1.02 ± 0.16	0.85 ± 0.16
EMD (MB)	0.67 ± 0.03	1.94 ± 0.42	0.98 ± 0.18	0.82 ± 0.18
PER (MB)	0.74 ± 0.08	0.74 ± 0.29	0.48 ± 0.29	0.82 ± 0.29
SPI (MB)	0.67 ± 0.04	1.62 ± 0.53	0.94 ± 0.18	0.88 ± 0.18
FAI (MC)	0.74 ± 0.01	1.17 ± 0.30	0.80 ± 0.02	0.75 ± 0.02
DSL (MC)	0.83 ± 0.08	1.01 ± 0.20	0.91 ± 0.16	0.90 ± 0.16
LSL (MC)	0.98 ± 0.02	1.19 ± 0.26	0.97 ± 0.05	0.95 ± 0.05
TLO (MC)	0.79 ± 0.06	1.37 ± 0.52	0.82 ± 0.09	0.82 ± 0.09
ACC (MC)	0.68 ± 0.03	1.36 ± 0.39	0.96 ± 0.18	0.84 ± 0.18
DEC (MC)	0.69 ± 0.05	1.83 ± 0.50	0.98 ± 0.06	0.84 ± 0.06

Variable	Linear Regression			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.84 ± 0.17	3.28 ± 0.84	1.13 ± 0.11	0.86 ± 0.16
HSR (LC)	1.13 ± 0.33	3.12 ± 1.33	2.14 ± 1.08	1.65 ± 0.72
SPR (LC)	0.95 ± 0.05	3.96 ± 2.41	2.10 ± 0.97	1.07 ± 0.24
MAX (LC)	1.05 ± 0.18	4.36 ± 2.52	1.72 ± 0.54	1.62 ± 0.64
RHL (LC)	1.00 ± 0.03	4.23 ± 1.85	2.95 ± 1.53	2.61 ± 1.38
AMP (MB)	0.82 ± 0.10	3.02 ± 0.52	1.05 ± 0.07	0.89 ± 0.13
HML (MB)	1.03 ± 0.12	3.04 ± 0.89	1.46 ± 0.46	1.42 ± 0.43
HEF (MB)	1.25 ± 0.65	2.90 ± 0.98	1.04 ± 0.11	1.35 ± 0.51
EMD (MB)	0.96 ± 0.24	2.78 ± 1.08	1.02 ± 0.11	1.04 ± 0.36
PER (MB)	1.04 ± 0.12	0.75 ± 0.24	0.47 ± 0.14	0.98 ± 0.14
SPI (MB)	0.86 ± 0.12	3.04 ± 0.43	1.03 ± 0.16	0.94 ± 0.23
FAI (MC)	0.78 ± 0.03	2.04 ± 1.07	1.01 ± 0.38	0.81 ± 0.04
DSL (MC)	0.91 ± 0.12	1.84 ± 0.88	0.92 ± 0.14	1.03 ± 0.17
LSL (MC)	1.00 ± 0.08	2.01 ± 1.13	1.16 ± 0.16	1.13 ± 0.18
TLO (MC)	0.89 ± 0.10	2.10 ± 0.36	0.98 ± 0.10	1.02 ± 0.31
ACC (MC)	0.74 ± 0.04	2.38 ± 1.15	1.12 ± 0.26	0.78 ± 0.08
DEC (MC)	0.91 ± 0.12	2.18 ± 1.43	1.24 ± 0.39	0.82 ± 0.10

cases. Figure 3.16, Figure 3.17 and Figure 3.18, present chord diagrams showing the influence of each the predictor variables in each of target physical variables, for each variable group. Variables in the bottom half of the diagram correspond to predictors while the ones at the top half correspond to target variables. The size of the incoming chords for each target are proportional to their influence in terms of mean increase error when they are absent. Just variables above 0.25 MIE are shown. The *3W* suffix of the predictors refer to the average value of that variable during matches in the last 3 weeks. The suffix *3W Tr.* is used instead for average value during last 3 week training sessions. Locomotor predictors are shown in blue, metabolic ones in red and mechanical predictors in green, while non physical variables are drawn in yellow.

From the three figures one can observe the influence of two or three types of variables in the top ranking predictors. Both for the locomotor and metabolic groups two main variables from each one function were selected as best predictors (3W AMP and 3W DIS). Given that the COR filter has been previously applied, these two variables are acting as representatives of the variables highly correlated with them in each group. In this sense, for example, 3W AMP can be used to explain or understand a large part of future SPI, EMD, HEF and AMP values. Similarly 3W SPI could be selected instead by the COR filter as surrogate of these variables and would have a similar predictive effect than 3W AMP. This brings the idea that, instead of requiring to analyse a high amount of variables for explaining player behaviour, the highly correlated variables could be substituted by one representative with a similar effect. For mechanical variables a similar effect is observed with 3W FAI, 3W DSL and 3W ACC. Is observed that wide majority of the better explaining predictors correspond to 3-week average of match physical variables instead of training information. Also, the player id and position play a relevant role for predicting most of the variables, providing the idea that the inherent differences between players and positions also determine the forecast of values, which is an expected result. For a level of over 0.25 MIE (in NRMSE), which is considered moderate, variables can be explained by 3 to 5 predictors in average.

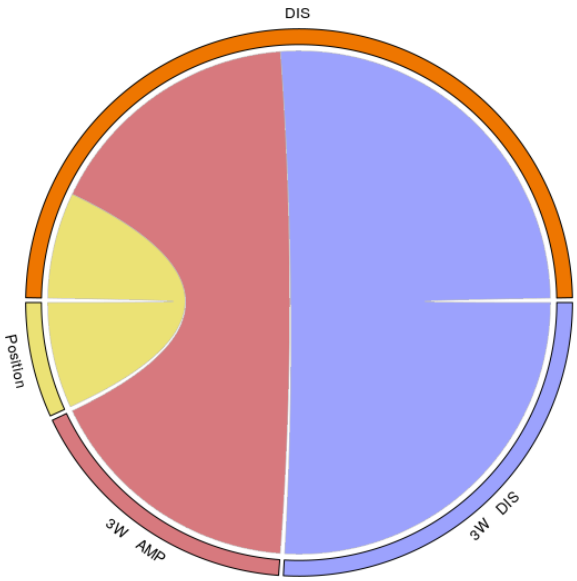


Figure 3.16: Chord diagrams of influence of variables with a MIE higher than 0.25, for Locomotor Variables

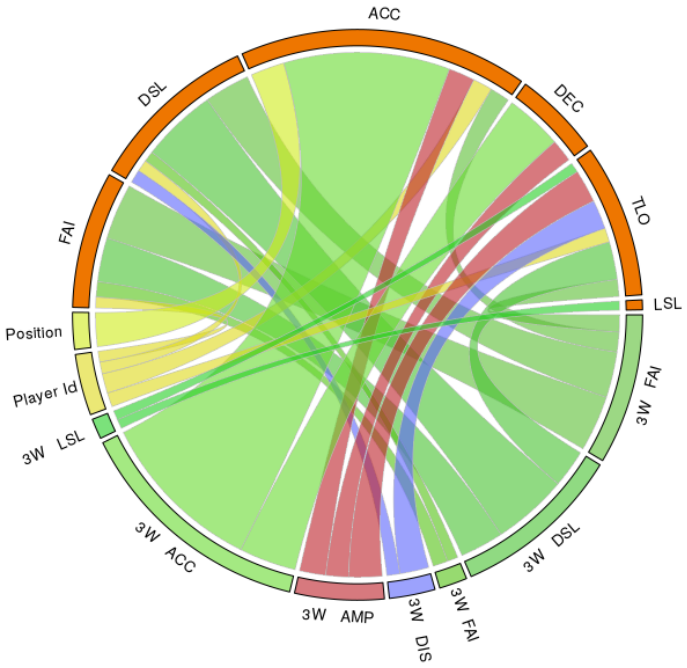


Figure 3.17: Chord diagrams of influence of variables with a MIE higher than 0.25, for Mechanical Variables

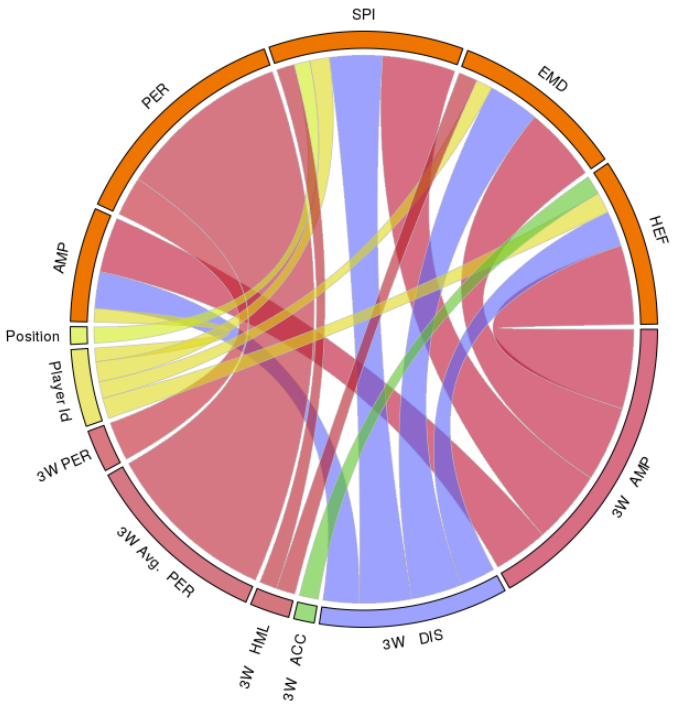


Figure 3.18: Chord diagrams of influence of variables with a MIE higher than 0.25, for Metabolic Variables

Chapter 6

Conclusions and Future Work

This section presents first a detailed set of conclusions reached in each phase of the study addressing the main research questions proposed. Then overall conclusions and future work suggestions are described.

6.1 Conclusions

6.1.1 Initial Exploration

The phase of data exploration allowed to fulfil the expectations of gaining a deeper insight on the main characteristics of data, as well as validating some of the most critical believes of physical coaches. The alignment between the believes on physical conditioning and the observed patterns, supports a stronger believe that the processed data has sufficient quality to provide insightful information on player's physical state. Several questions where proposed in order to address this phase, which are focused on the distribution of data among session types and player positions, the consideration of relations among variables, a further validation of the proposed categories of variables, the relation between match performance and historical variables, and the expected tendency of presenting oscillatory patterns.

Clear differences where found between matchday (MD) and the different training

sessions. MD-3 and MD-4 where validated as the most similar to MD, while MD-3 tends to present higher variation of variables. This is validated both through individual variable observation and visual plots approximating relations between variables. Some differences where observed between different player's position, where center midfielders (MC), wingers (AM/WN) and lower backs (LB) presented expected higher values of locomotor and metabolic variables, while center backs where consistently presenting lower values and variations in comparison. Although these differences might provide more specific details for a fine-grained study of players by positions, the slight differences between positions where not considered enough so to discard a study considering information from all player positions at a time, which is preferred given the low amount of data (one season).

An observation of the linear correlation between variables evidenced a tendency to high correlation in several variables. Most importantly, the proposed categorization of three groups: locomotor, metabolic, and mechanical variables, showed to preserve correlations within groups and lower correlations between groups. This provided a better understanding of the proposed groups, and goes along with the observed similar patterns between some variables.

The use of historical information showed to provide interesting insights on players match performance. The Fatigue Index (FAI) during matches showed to have a direct linear relation with the accumulated fatigue in 6-week windows. Also when accumulated fatigue increases the registered physical values tend to decrease significantly. For a second level of fatigue, above the minimum and considerably below the maximum registered, players evidence the higher peaks in physical variables of the three groups. These observations seem to validate the idea of an rather optimal fitness state which is not reached with excessively high or low values of accumulated load or training. When contrasting with lower and higher number of match minutes played, highly loaded (or wearied) players tend to present even lower physical values in matches when they have played the most minutes. There is also some evidence of

the effect of matches in physical conditioning, that must be further validated with higher availability of data.

By last, it was validated the presence of an oscillatory tendency of most physical variables along a season time. There where no significant differences between mean values of the whole team and different positions, so a more organized and less chaotic variation of variables is observed. This is most probably associated to training design and the capacity of coaches to control effort and intensities up to some degree. This supports the idea of analysis of player variations without discriminating by players, by expecting a rather stable tendency in general, but slighter variational differences along time, inherent to each player's fitness condition.

6.1.2 Unsupervised Analysis

The presented approach of relating the periodical variations of physical variables during training with following matches allowed to observe considerable differences between groups of higher and lower magnitude of variation. The players presenting higher variations during training reflected in higher values for intensity, distance and effort related variables in the following matches. The same group presents lower values for three-weekly training minutes and accumulated training and match load. This approach might provide a way for analyzing the adaptation of players to training dynamics, and even to evaluate training design. The procedure follows a series of simplifications such as the selection of session type MD-3 exclusively and the use of daily average data instead of the complete time series of minute to minute registers during training, which might incur in lost of information. However, this type of calculations can be easily and directly integrated to daily routine performance analysis carried out by physical coaches, without the need of additional systems or requiring high processing times. It is unclear whether the observed relations are patterns related with specific conditions of the analyzed season or the club internals training structure. However, the findings provide sufficient evidence to suggest the incorporation of this calculation in daily routine and track its evolution along the rest of tasks that

conform the complete process of training and match analysis.

The summarized dataset allowed a more representative grouping and also more concluding results, which is also more convenient from a practical sense. This dataset allowed to cluster two groups with considerably clear difference in absolute average values of variation, which can be translated in practice into the selection of specific ranges for each variable, such as the presented in the results table. Also, time-window aggregated information is showing to add value for performance analysis and should be considered in future research. On the other hand, despite is common use in time series matching, DTW could not provide sufficiently clear results in this study, most probably due the the short-size characteristics of the analyzed time series and that exact match of variation patterns might be too strict for the few data available.

The player normalization seems to favor a cleaner and possibly more correct comparison between the physical registers of players. It can be shown that the use of absolute values in training provide even higher effect sizes in most cases but that would most probably be a consequence of the inherent differences between players and positions than the actual magnitude of variation. Notably, differences are not considerably large in absolute values. This has to do with the not strict separation between groups achieved by the clustering procedure, and follows an expected result by physical coaches.

6.1.3 Supervised Analysis

This part of the study showed that it is possible to predict physical variables based on training and match information from EPTS devices, for practical applications. Past match information provides critical value on predicting future match performance, possibly due to the idea that competition efforts are the highest demanding for players and where stimuli are not controlled such as in training sessions, thus leading to more challenging but also more representative information. Historical aggregates of

both match and training session physical variables showed a highly relevant influence within the predictive models. The prediction error achieved for 11 of 17 variables might allow its direct application in practice and is suggested to be incorporated as additional information for the physical coaches routinely evaluation. Future studies should also incorporate internal metrics such as the rate of perceived exertion (RPE) and heart rate exertion (HRE), as well as tactical information, for providing a more robust context of information. For the three groups of variables, both metabolic and mechanical ones showed to be more accurately predictable. Locomotor variables prediction were less well performing possibly due to a high dependency on match-specific and tactical conditions.

Both algorithms exploiting non-linear relations on physical variables performed considerably better than linear models, providing a glance of the complexity of this type of data. We observed the presence of highly correlated features whose fine-grained removal produced a considerable improvement for the predictions. Recursive feature elimination helped to improve the results only slightly while PCA did not produce much advantage for the predictions.

We introduce the use of NRMSE as an error metric for regression that can be more easily translated into practice. The approach for calculating NRMSE based on matches standard deviations of variables could allow sport professionals to have clearer and faster interpretation of the quality of the results, and the expectation of their future application in practice.

The observation of the importance of variables for prediction provided an insight on the influence of the three defined type of variables. The use of representative variables for highly correlated ones could provide a crucial simplification of the fast-paced analysis carried out by practitioners. These observations are relevant due to the increasing availability of new variables everyday which might obstruct the analysis if not properly acknowledged.

6.1.4 Overall Conclusions

The main question addressed by this study reached successful results. It was found that EPTS derived data of physical performance, can provide rich insights regarding the relation between training and match performance, in professional football. Some of these relations and patterns can be modelled through machine learning techniques, taking advantage of the joint information provided by multiple variables, and providing the possibility to improve with further availability of data. Most importantly, the three different approaches of the study, provided a way of explaining underlying knowledge that can be translated into practice. The initial exploratory analysis allowed to validate a series of believes of physical coaches, most of which proceed from expert observation and field knowledge, instead of statistical data analysis. Additionally, previously discarded variables, such as the Fatigue Index (FAI), showed interesting effects so to be considered within the daily review and analysis process of coaches. On the other hand, the unsupervised approach for relating training variability and adaptability with match performance provided new insights that can be translated into practice. Although higher or lower physical values during matches do not provide a direct indication of better or poorer performance, physical coaches were provided with new type of information for the analysis of physical conditioning. Specifically, the categorization of player variation can be directly incorporated as a new variable, which is a result of inter-variable relations and incorporates historical physical information. The supervised approach faced a more ambitious challenge, which is to estimate future physical performance of players. Beside the common understanding that match physical performance is a result of additional components beyond the strictly physical (such as tactical or psychological), the obtained results provide sufficiently good results so to gradually incorporate this predictions into player fitness state assessment. The proposal of normalized root mean square error (NRMSE) in the presented way, intends to provide a common framework for comparing predictive models on such wide variety of different variables, in particular for practical purposes.

NRMSE allow physical coaches to understand the quality of prediction in terms of deviations from the team mean values on each variable, which in case of F.C. Barcelona, are commonly used concepts. Additionally, the capacity of predicting several of the most important physical variables, allowed to get deeper insights on the most relevant variables that influences each other one. In today's practice, coaches are used to relate, either one variable at a time, or trying to grasp visually information that might arise from many variables. Both approaches have evident limitations in terms of the explanatory effects to obtain or available time to spend. The understanding of the relation among variables, from a predictive approach, might provide a way to reduce the amount of analysed variables, and guide to a more direct analysis. Also, in more general terms, we have proposed a classification of physical variables in three main groups, that will allow, hopefully, to simplify and structure the analysis and communication of physical information.

Overall, EPTS tracking data seems to provide valuable information for assessing physical performance through data analysis and statistical models. This is the first study, up to our knowledge, to relate training and match physical variables directly registered from player using EPTS devices, in professional football.

6.2 Future Work

In the following years, with higher availability of data, these remarks must be further validated. Regarding physical variability, the yearly knowledge of physical evolution of training dynamics and even specific players might provide new insights about the physical preparation of teams and the performance during competition. With the upcoming availability of data from different categories of teams, professional first and second teams, youth teams, and professional woman football, more generalizable characteristics of physical data might arise. Moreover, when having physical performance data of same player along several years, a more fine-grained understanding of its capabilities and fitness characteristics might allowed for a more refined training

design. This same type of information might provide better insights on the power of predictive models for future performance. In both cases, with further availability of data we propose to validate and expand the presented study.

The selection of training session MD-3 presents a considerable reduction of the available information of players. Further studies might approach the analysis of the fitness state of player, through the use of all the training session days available within micro-cycles. This includes matches, which are considered the most influential stimulus for player training. This also comes with the challenge of dealing with the expected high amount of noise that will derive from less demanding training sessions, and higher dimensionality for low amounts of examples.

One of the most desirable pieces of information for training design is the knowledge of each player's fitness profile. This is understood as the range of physical variability in which the player is considered to be in optimal conditions. This varies from player to player, and is most commonly unknown. The value provided by historical information, and the availability of larger data on training and matches, will definitely provide better knowledge on individual player characteristics and the effect of team training designs on her.

For this study session average and max values where used instead of the complete timely detailed information of performance during the whole sessions. Research in the area of complexity is used to deal with this kind of temporal and highly variable type of data, and its believe to provide much refined insights of actual adaptability and stability of players as a system. The application of machine learning approaches to exploit non-linear relations and make sense of this complex type of data, might add additional and insightful knowledge on physical behaviour.

Bibliography

[1] Lesne A. Chaos in biology. *Biology Forum. Rivista di Biologia*, 99:467–481, 2006.

[2] McCall A., Davison M., Carling C., Buckthorpe M., Coutts A.J., and Dupont G. Can off-field 'brains' provide a competitive advantage in professional football? *Br J Sports Med*, 50:710–712, 2016.

[3] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008.

[4] Arni Arnason, Stefan B Sigurdsson, Arni Gudmundsson, Ingar Holme, Lars Engebretsen, and Roald Bahr. Physical fitness, injuries, and team performance in soccer. *Medicine & Science in Sports & Exercise*, 36(2):278–285, 2004.

[5] Desgraupes B. Clustering indices. https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf, 2013.

[6] Andrew Borrie, Gudberg K Jonsson, and Magnus S Magnusson. Temporal pattern analysis and its applicability in sport: an explanation and exemplar data. *Journal of sports sciences*, 20(10):845–852, 2002.

[7] Markus Brandt and Ulf Brefeld. Graph-based approaches for analyzing team interaction on the example of soccer.

[8] Martin Buchheit, Hani Al Haddad, Ben M Simpson, Dino Palazzi, Pitre C Bourdon, Valter Di Salvo, and Alberto Mendez-Villanueva. Monitoring accelerations with gps in football: time to slow down. *Int J Sports Physiol Perform*, 9(3):442–445, 2014.

[9] Cummins C., Orr R., O'Connor H., and West C. Global positioning systems (gps) and microtechnology sensors in team sports: A systematic review. *Sports Mede*, 43:1025–1042, 2013.

[10] Gavin C, Talbot, and Nicola LC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

[11] Petersohn C. Model assesment and selection in the elements of statistical learning. *Springer series in statistics*, 1:245–2477, 2001.

[12] Strobl C., Boulesteix A., Kneib T., Augustin T, and Zeileis A. Conditional variable importance for random forests. *BMC bioinformatics*, 9:1, 2008.

[13] David Casamichana, Julen Castellano, Julio Calleja-Gonzalez, Jaime San Román, and Carlo Castagna. Relationship between indicators of training load in soccer players. *The Journal of Strength & Conditioning Research*, 27(2):369–374, 2013.

[14] Julen Castellano and David Álvarez. Uso defensivo del espacio de interacción en fútbol.(defensive use of the interaction space in soccer). *RICYDE. Revista Internacional de Ciencias del Deporte. doi: 10.5232/ricyde*, 9(32):126–136, 2013.

[15] Julen Castellano, David Casamichana, and Carlos Lago. The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of human kinetics*, 31:137–147, 2012.

[16] Paolo Cintia, Fosca Giannotti, Luca Pappalardo, Dino Pedreschi, and Marco Malvaldi. The harsh rule of the goals: data-driven performance indicators for football teams. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[17] Filipe Manuel Clemente, Micael Santos Couceiro, Fernando Manuel Lourenço Martins, and Rui Sousa Mendes. Using network metrics in soccer: A macro-analysis. *Journal of human kinetics*, 45(1):123–134, 2015.

[18] Jacob Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.

[19] Chad Cook. Predicting future physical injury in sports: it's a complicated dynamic system. *British Journal of Sports Medicine*, pages bjsports–2016, 2016.

[20] David Cornforth, Piers Campbell, Keith Nesbitt, Dean Robinson, and Herbert F Jelinek. Prediction of game performance in australian football using heart rate variability measures. *International Journal of Signal and Imaging Systems Engineering*, 8(1-2):80–88, 2015.

[21] A. Coutts. Working fast and working slow: The benefits of embedding research in high performance sport. international journal of sports physiology and performance. *Acciónmotriz*, 1:1–2, 2016.

[22] Medina D. Are there potential medical legal issues concerning the safe usage of electronic personal tracking devices? the experience of a multi-sport elite club. May 2016.

[23] Memmert D., Lemmink KA., and Sampaio J. Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, pages 1–10, 2016.

[24] Jordi Duch, Joshua S Waitzman, and Luís A Nunes Amaral. Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937, 2010.

[25] Alberti G. Gaudino P. and Iaia M. Estimated metabolic and mechanical demands during different small-sided games in elite soccer players. *Elsevier*, 36:123–133, 2014.

[26] Laszlo Gyarmati, Haewoon Kwak, and Pablo Rodriguez. Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*, 2014.

[27] Folgado H., Duarte R., Fernandes O., and Sampaio J. Plos ones. *Sports Medicine*, 9(5):e97145, 2015.

[28] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *ACM Sigmod Record*, 31(2):40–45, 2002.

[29] T Haugen and S Seile. Physical and physiological testing of soccer players: why, what and how should we measure. *Sport Sci*, 19:10–26, 2015.

[30] IFAB. 129th annual general meeting the football association. http://resources.fifa.com/mm/document/affederation/ifab/02/60/90/85/2015agm_minutes_v10_neutral.pdf, 2015.

[31] Arjol J. La planificación actual del entrenamiento en fútbol. análisis comparado del enfoque estructurado y la periodizaciÃşn táctica. *Acciónmotriz*, 8:27–37, 2012.

[32] Fernandez J., Medina D., Gomez A., Arias M, and Gavaldà R. Does training affect match performance? a study using data mining and tracking devices. Presented at European Conference of Machine Learning (ECML-PKDD), Sports Analytics Workshop, 2016.

[33] Fernandez J., Medina D., Gomez A., Arias M, and Gavaldà R. From training to match performance: A predictive and explanatory study on novel tracking data. To be presented at International Conference on Data Mining (ICDM), Data Mining for the Analysis of Performance and Sucess Workshop, 2016.

[34] Mallo J. Complex football: from seirul.lo's structured training to frade's tactical periodisation. *Editorial Topprosoccer S.L.*, 1:65–116, 2015.

[35] Mallo J. Complex football: from seirul.lo's structured training to frade's tactical periodisation. *Seirul.lo's Structured Training. Editorial Topprosoccer S.L.*, 1:65–116, 2015.

[36] Herbert F Jelinek, Andrei Kelarev, Dean J Robinson, Andrew Stranieri, and David J Cornforth. Using meta-regression data mining to improve predictions of performance based on heart rate dynamics for australian football. *Applied Soft Computing*, 14:81–87, 2014.

[37] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[38] Hader K., Mendez-Villanueva A., Palazzi D., Ahmaidi S., and Bucheit M. Metabolic power requirement of change of direction speed in young soccer players: Not all is what it seems. *PLoS One*, page e0149839, 2016.

[39] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[40] Gyarmati L. and Hefeeda M. Estimating the maximal speed of soccer players on scale. *In Proc. of Machine Learning and Data Mining for Sports Analytics Workshop*, 2015.

[41] STATSports Technologies Ltd. Statssports viper metrics version 1.2, 2012.

[42] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[43] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[44] Sandbakk Ø., Cunningham D., Shearer D., Drawer S., Eager R., Taylor N., Cook C., and Kilduff L. Movement demands of elite u20 international rugby union players. *Plos One*, 11:e0153275, 2016.

[45] Cristian Osgnach, Stefano Poser, Riccardo Bernardini, Roberto Rinaldo, and Pietro Enrico Di Prampero. Energy cost and metabolic power in elite soccer: a new match analysis approach. *Med Sci Sports Exerc*, 42(1):170–178, 2010.

[46] Chotirat Ann Ratanamahatana and Eamonn Keogh. Everything you know about dynamic time warping is wrong. In *Third Workshop on Mining Temporal and Sequential Data*. Citeseer, 2004.

[47] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[48] Zeileis A. Strobl C., Boulesteix A. and Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8:1, 2007.

[49] Hisham Talukder, Thomas Vincent, Geoff FosterB, Camden HuB, Juan HuertaA, Aparna KumarA, Mark MalazarteA, Diego SaldanaA, and Shawn SimpsonA. Preventing in–game injuries for nba players.

[50] Hopkins WG, Marshall SW, Batterham AM, , and Hanin J. Progressive statistics for studies in sports medicine and exercise science. *Medicine and science in sports and exercise*, 41:3–13, 2009.

[51] Wikipedia. Analysis of variance, 2006. [Online; accessed 3-October-2016].

[52] Saeys Y., Inza I, , and Larra naga P. A review of feature selection techniques in bioinformatics. *Oxford Univ Press*, 23(19):2507–2517, 2007.

[53] Yuji Yamamoto and Keiko Yokoyama. Common and unique network dynamics in football games. *PloS one*, 6(12):e29638, 2011.

Does Training Affect Match Performance? A Study Using Data Mining And Tracking Devices

Javier Fernández[1,2], Daniel Medina[1], Antonio Gómez[1], Marta Arias[2], and Ricard Gavaldà[2]

[1] Fútbol Club Barcelona, Ciudad Deportiva Joan Gamper, Barcelona, Spain
[2] Universitat Politècnica de Catalunya, Campus Nord, Barcelona, Spain

Abstract. FIFA has recently allowed the use of electronic performance and tracking systems (*EPTS*) in professional football competition, providing teams with novel and more accurate data. Physical performance has not yet taken much attention from the research community, due to the difficulty of accessing this information with the same devices during training and competition. This study provides a methodology based on machine learning and statistical methods to relate the physical performance variation of players during time-framed training sessions, and their performance in the following matches. The analysis is carried out over F.C. Barcelona B, season 2015-2016 data, and makes emphasis on exploiting the design characteristics of the *structured training* methodology implemented within the club. The use of summarized physical variation data has provided a remarkable relation between higher magnitudes of variation in 3-week time frames during training, and higher physical values in the following matches. With increased data availability this and new approaches could provide a new frontier in physical performance analysis. This is, up to our knowledge, the first study to relate training and matches performance through the same *EPTS* devices in professional football.

Keywords: GPS, tracking devices, football physical performance, sports analytics, dtw, cluster analysis

1 Introduction

Professional football has attracted the attention of the data science community in the last decade due to the increasing availability of quantitative data. The latest technology has provided the possibility of gathering different kinds of specific metrics, from team statistics to in-game detailed events, contributing to the improvement of typical and critical tasks such as team tactics evaluation, opponent analysis, player scouting and training design. The idea that exploiting data-related analysis can become a competitive advantage within professional sports is increasingly supported [1]. However, it should be noted that few of the current studies are devoted to the analysis of physical information of the players [2]. This has to do mainly with the difficulty of having access to this data

through training and competition, which is considered highly valued by football clubs [2]. Typically, such information is gathered through the use of electronic performance and tracking systems (EPTS) which include GPS and microsensor technology such as accelerometeres, gyroscopes and magnetometers. Such is the case of professional sections at F.C. Barcelona where these tools are used for monitoring load and many other physical variables. Despite the existing concerns regarding its reliability, they have increasingly being adapted and accepted in sports such as Rugby, Australian football, Cricket and Hockey [3]. Recently, the Football Association Board (IFAB) has authorized the use of these devices during official football competition for the 2015-2016 season [4], opening the doors for novel research regarding physical performance of players during the season.

At F.C. Barcelona, EPTS devices have been recently used to aid the evaluation of the applied training methodology, the *structured training*, a system that sets the baselines for the planning and adaptation of the training activities along the season, providing the novelty of incorporating competition activities in this design. This involves the idea of providing a schema in which the player is promoted to adapt to the training demands and evolve in each of its structures, beyond the strictly physical conditions [5]. A *player optimization* is sought through the application of training situations that cause imbalance in one of the subject's structures in order to promote its adaptation, so forcing a continuous auto-organization process in sets of 3 weeks periodization [5]. This methodology considers not only training as a stimulus to induce adaptation but also competition as the most relevant stimulus to optimize the athlete capabilities. This implies that physical demands for players during training are structured within consecutive cycles but are not strictly defined, so the measured physical player values can provide uncertainty and richness in its analysis. Also, given the idea of *deterministic chaos* present in biological systems [6], players are expected to evidence different adaptational behaviors along the season trainings. Based on this, it is plausible to think that periodical variation of physical values could provide valuable information regarding the adaptability and fitness of the player.

The main objective of this study is to find whether there exist significant relations between physical performance of players during training and the measured performance in subsequent matches, for F.C. Barcelona B data from season 2015-2016. Machine learning algorithms are used in order to exploit the contribution of the high amount of measured variables as a whole, all of which are expected to contribute explaining the player's dynamic up to some extent. The study is structured in three main stages. A data preparation stage in which data is pre-processed and normalized, and two datasets are created. An exploration stage where dynamic time warping and cluster analysis is applied in order to obtain representative natural groups from data. And finally, a validation stage, where the matches associated with clustered series are extracted and statistical tests are performed to determine the existence of significant differences. Final conclusions and future work suggestions are detailed, regarding the usefulness

of this approach and the finding of moderate standardized differences between groups presenting high and low variations of physical values from week to week.

2 Methodology

2.1 Data Collection

F.C. Barcelona B has collected both training and matches physical performance measurements, for season 2015-2016, using the *StatsSports GPS Viper Pod* devices. The resulting tracking information is manually segmented by physical coaches, and further visualized through a software integrated with the devices which outputs several variables. From this set of variables, we have selected 15 along physical coaches, described in Table 1, which summarize the considered most relevant performance information. Variables are structured in three main groups: locomotor, metabolic and mechanical. Locomotor variables refer to simple direct measurements of travelled distance and speed, that are obtained solely through GPS. Metabolic variables are associated with energy expenditure and exertion, while mechanical variables relate with intensity changes and impacts [7]. For these last two groups variables are calculated by a combination of GPS and accelerometers. The data consists of 153 training sessions and 34 matches, which adds up to 2478 training rows and 473 match rows among all the 42 different players throughout the season 2015-2016. The season information is queried from the central database containing the total 2951 rows, where each one contains the measured variables for a single player in a specific session and additional variables that contextualize the information such as player id, position, name, total session time, the session id and session type.

2.2 Data Processing

The dataset is initially processed, adding additional contextualization variables and performing several types of normalizations. Within F.C. Barcelona training structure, training days are labelled in strict relation with the following match day, where match is labelled as MD, the following two days MD+1 and MD+2, and the previous days MD-1 up to MD-4. Each day-type follows specific design rules for training drills. For simplicity of the study, only day MD-3 sessions are used, due to they similarities to match days in terms of number of players, playing spaces and opposition level. Additionally, day MD-3 involves the highest differences between physical values. Goalkeepers are deleted from the database since they face considerably different physical challenges than field players. A new variable, load percentage (PER) is added in order to reflect the session load, which is calculated as a ratio of the average AMP from matches. All the measured values are normalized by dividing by the total time of duration of the session. Variables that already represent averages or maximums are kept

Table 1: Description of selected physical variables splitted in three groups: locomotor, metabolic and mechanical.

Locomotor Variables	
Name and Acronym	**Description**
Travelled Distance (DIS) [8]	Total distance travelled during session drills or matches
Sprints (SPR) [8]	Number of times over $5.5m/s$ during > 1
High Speed Running (HSR) [8]	Travelled meters when speed $> 5.8m/s$
Max Speed (MAX) [8]	Maximum speed reached by the player

Metabolic Variables	
Name and Acronym	**Description**
Average Metabolic Power (AMP) [8]	Energy expended by the player per second per kg, measured in W/Kg
High Metabolic Load Distance (HML) [8]	Distance travelled by a player when the metabolic power is $> 25.5W/Kg$
High Metabolic Efforts (HEF) [9]	The number of separate movements/efforts undertaken in producing HML distance
Load Percentage (PER)	Proportion of AMP with respect to an average 9.5 AMP in matches

Mechanical Variables	
Name and Acronym	**Description**
Fatigue Index (FAI) [8]	Accumulated DSL from the total session volume, in terms of speed. (DSL/SPI)
Dynamic Stress Load (DSL) [8]	Total of the weighted impacts, based on accelerometer values over 2g
Lower Speed Loading (LSL) [8]	Load associated with the low speed activity alone
Total Loading (TLO) [8]	The total of the forces on the player over the entire session based on accelerometer data alone
Accelerations (ACC) [8]	Number of increases of speed during at least 0.5 s ($> 3m/s^2$)
Decelerations (DEC)[8]	Number of decreases of speed during at least 0.5 s ($< 3m/s^2$)
Step Balance (STE) [8]	Ratio of left step impact to the sum of the left step impact and right step impact

as originally measured, such as AMP, FI, PER, STE and MAX. Additionally, summarized information is added to matches data such as the average training minutes, average fatigue and total (training plus match) load in the previous three weeks. An additional normalization is applied where absolute values are transformed into the number of standard deviations of each particular player in the given day label type. This transformation is performed in order to avoid differences that arise due to player physical characteristics instead of a response to training. Finally, a last transformation performed over training data seeks to quantify the degree of variability from week to week on each physical value. The idea is to measure the difference between registered values from two consecutive weeks, as presented in Figure 1.

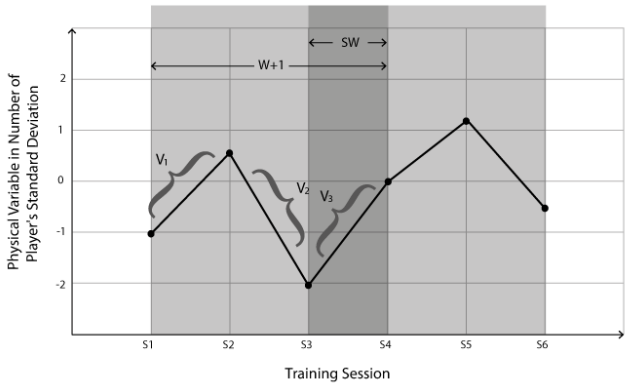


Fig. 1: Representation of a series of measured values of a particular variable during weekly training sessions (x axis). V_i values refer to the difference of values registered at sessions S_{i+1} and S_i. W is the size of the sliding window, used to build time-series and summarized datasets. SW refers to the amount of weeks to slide each time.

Each value V_i represents the absolute difference between a value registered at sessions S_{i+1} and S_i. Two datasets were built: the first one consists of time-series of W window size. A sliding window approach is followed by using a fix-sized (W) window of consecutive weeks. The time-series dataset is conformed by groups of W rows containing the 15 physical variables, corresponding to a player in a specific period of the season. Selected windows sizes during experiments are 3 and 6 in order to match the methodology of the club. Windows are moved SW steps each time, so to control the degree of coincidence of values between windows. The value of SW was selected following Equation (1) to avoid an excessive overlap between windows and to avoid a too strict separation that would reduce significantly the amount of data. Another dataset is built which summarizes each group of W rows in each variable, by calculating the average of absolute differences. Equation (2) describes the performed calculations, where

P_{jvd} corresponds to the absolute average of window differences of a variable v of a player j, measured in the window frame d, substracted by the mean of P_{ivd} for every other player i. P corresponds to the set of all possible players.

$$SW = W - (W/3) \tag{1}$$

$$P_{jvd} = \frac{\sum_{i=2}^{W+1} \|S_i - S_{i-1}\|}{W} - \frac{\sum_{i \neq j}^{|P|} P_{ivd}}{|P|} \tag{2}$$

2.3 Data Exploration

Visual Exploration. Specific differences of physical variables where assessed visually through boxplots and analytically through one-way *ANOVA* and Post Hoc tests observing the differences between type-days (i.e MD-4, MD-3, MD-2, etc.). A *PCA* analysis was also performed, and the two principal components where plotted accounting for 69% of variance and observing the acknowledged differences. On the other hand, different plots over the time-series and summarized datasets allowed to visualize oscillatory patterns along the season that respond to cycles design. Also, it is observed how players tend to oscillate in similar patterns due to the training design. There exist, however, several cases in which certain players magnitude of variations starts differing considerably from the mean variation. The results of these observations coincided with the understanding of physical responses in training from the club's physical coaches. For space restriction reasons, the graphical results are omitted from this section.

Calculating Series Similarities through Dynamic Time Warping. Dynamic time warping (DTW) is a highly used method that allows to measure the similarity between two temporal series, while being less sensitive to signal transformations such as shifting, uniform amplitude scaling or uniform time scaling [12]. DTW was applied over the time-series dataset in order to calculate similarity between windowed variations along the season on different players. The idea is to find variation patterns that are more similar to each other, independently from the specific player or position. A distance or dissimilarity matrix is found for each pair of series in the dataset. Euclidean distance was used, in order to prioritize vectors magnitude over angles since the degree of variation is believed to be more informative than the actual followed pattern, in order to approximate the physiological response. Once the dissimilarity matrix is found, the *k-mediods* algorithm is applied for finding a natural clustering of the time series.

Cluster Analysis For both datasets cluster analysis is applied to find natural groupings of variation. It is critical to observe that the clustering procedure is applied to multidimensional data, aiming to incorporate the relation between each of the variables. For the time-series dataset the k-mediods algorithm is used, since its capability of being applied to distance matrices and the flexibility

of controlling the number of clusters. For the summarized dataset, *k-means* is used instead. The selection of number of clusters is performed by calculating five internal indices and selecting the number of clusters picked by the majority. These indices are: C-index, C-H index, DB index, Silhouette index and the Ratkowsky-Lance index [13]. Also, the dimensionality reduction technique T-Stochastic Neighbor Embedding *t-SNE* [14] was applied to visually asses the quality of clusters. Once the training sessions information is clustered, each of the window-frames is associated with next upcoming match, generating a cluster-labelled dataset containing the absolute values of matches physical variables.

3 Results

Results for both the time-series and the summarized datasets are presented together since they follow an identical approach in its evaluation. For both cases, the selected number of clusters was 2 by four of the five different indices, the sample size of the training sessions dataset is 112, and the sample size of associated matches dataset is 82. Only the results for 3-week window are presented, since no statistically significant relation was found with 6-week window frames. For each of the variables conforming the two groups (in each dataset) the standardized difference of means was calculated to describe the effect size. The limits of the effect sizes are those suggested by Hopkins [16] which are recommended in sports related data and for practical applications (trivial effect: < 0.2, small effect: $0.2 - 0.6$, moderate effect: $0.6 - 1.2$, large effect $1.2 - 2.0$ and, very large: > 2.0), with a confidence interval of 90%.

Detailed results are presented in Table 2. It can be clearly observed that for the summarized dataset almost every variable in training registered a moderate to large effect size when comparing groups. So, we are observing the detection of two groups: one where the average magnitude of variations of each variable is higher (*high variation group*), and one where is lower (*low variation group*). It is critical to observe that separation among groups is not absolute, and there exist ranges of values which overlap. This has to do with multivariate nature of the clustering procedure, and coincides with the original expectation of this study. It can also be observed that for the timeseries dataset few variables where able to stand out just with a small size effect. Even with the selection of Euclidean distance to favor magnitudes, the cluster analysis over the DTW procedure was not able to found a clear separation between groups. The procedure over the summarized dataset, instead, did find a considerably separation between training groups so the analysis over associated matches is easier to interpret and translate to practice. Figure 2 presents the effect sizes for the associated matches in both datasets. It can be observed for both cases that variables registering high intensity efforts, energy consumption and distance travelled appear with higher magnitude in the *high variation group* consistently, while the total load percentage and training minutes in the previous three weeks are considerably low in this same group. HML, AMP and DIS present moderate effect size in

the summarized dataset, variables belonging to metabolic group (the first two) and locomotor group. For the timeseries dataset only HML presents a moderate effect size, toward the same tendency. A small effect size is also observed in other locomotor (MAX), metabolic (PER and HSR) and mechanical variables (DSL, DEC) toward the same tendency. Three-weekly PER and training minutes show also a moderate effect in differences, towards lower values. Sample size for associated matches allows to conclude with certainty about moderate size effects. Small effects should be taken into account, but must be further validated with the future increase of availability of data.

Table 2: Mean and standard deviation for each physical variable in each of the clustered groups. For both training data and the associated matches, values obtained in both summarized and timeseries datasets are presented. The standardized difference of means *SDM* is presented for each case. Training results refer to the absolute average of variation while matches results refer to the actual measured physical values.

	Training (mean ± SD)						Matches (mean ± SD)					
	Summarized			Timeseries			Summarized			Timeseries		
Variable	Cluster 1	Cluster 2	SDM	Cluster 1	Cluster 2	SDM	Cluster 1	Cluster 2	SDM	Cluster 1	Cluster 2	SDM
DSL p/m	0.89 ± 0.29	0.58 ± 0.28	*Moderate*	0.41 ± 0.198	0.41 ± 0.18	*Trivial*	3.60 ± 1.30	3.26 ± 1.37	*Small*	3.37 ± 1.17	4.29 ± 1.84	*Small*
ACC p/m	1.011 ± 0.45	0.67 ± 0.31	*Moderate*	0.45 ± 0.26	0.49 ± 0.23	*Trivial*	0.56 ± 0.22	0.52 ± 0.24	*Trivial*	0.65 ± 0.23	0.561 ± 0.21	*Small*
DEC p/m	0.94 ± 0.38	0.51 ± 0.23	*Large*	0.43 ± 0.25	0.38 ± 0.19	*Trivial*	0.80 ± 0.27	0.69 ± 0.26	*Small*	0.82 ± 0.29	0.785 ± 0.307	*Trivial*
SPR p/m	0.74 ± 0.35	0.59 ± 0.28	*Small*	0.36 ± 0.17	0.40 ± 0.20	*Trivial*	0.37 ± 0.11	0.311 ± 0.09	*Small*	0.37 ± 0.12	0.38 ± 0.20	*Trivial*
HSR p/m	0.98 ± 0.46	0.53 ± 0.26	*Large*	0.39 ± 0.17	0.43 ± 0.28	*Trivial*	13.30 ± 4.73	10.78 ± 4.31	*Small*	12.58 ± 5.27	10.686 ± 3.837	*Small*
AMP	0.62 ± 0.36	0.29 ± 0.14	*Large*	0.28 ± 0.10	0.25 ± 0.13	*Small*	10.35 ± 1.08	9.65 ± 1.17	*Moderate*	10.27 ± 1.09	10.26 ± 1.50	*Trivial*
HML	0.58 ± 0.24	0.40 ± 0.20	*Moderate*	0.33 ± 0.14	0.25 ± 0.13	*Small*	39.12 ± 10.05	32.05 ± 10.01	*Moderate*	36.09 ± 9.11	30.892 ± 7.95	*Moderate*
HEF p/m	0.74 ± 0.31	0.40 ± 0.22	*Large*	0.34 ± 0.17	0.29 ± 0.19	*Small*	2.23 ± 0.52	1.94 ± 0.5	*Small*	2.31 ± 0.5	2.20 ± 0.64	*Trivial*
FAI	0.93 ± 0.39	0.71 ± 0.34	*Moderate*	0.47 ± 0.24	0.48 ± 0.21	*Trivial*	0.62 ± 0.19	0.64 ± 0.25	*Trivial*	0.602 ± 0.18	0.75 ± 0.29	*Moderate*
DIS p/m	0.58 ± 0.36	0.27 ± 0.15	*Large*	0.25 ± 0.10	0.21 ± 0.20	*Small*	111.5 ± 10.98	104 ± 11.3	*Moderate*	109.5 ± 10.72	110.71 ± 15.41	*Trivial*
TLO p/m	0.83 ± 0.25	0.39 ± 0.21	*Large*	0.35 ± 0.16	0.32 ± 01.9	*Small*	1.59 ± 0.23	1.48 ± 0.32	*Small*	1.56 ± 0.22	1.67 ± 0.38	*Small*
MAX	0.80 ± 0.39	0.63 ± 0.32	*Trivial*	0.40 ± 0.198	0.42 ± 0.22	*Trivial*	29.6 ± 2.11	28.79 ± 2.32	*Small*	29.612 ± 2.57	29.28 ± 1.61	*Trivial*
STE	1.05 ± 0.57	0.97 ± 0.53	*Trivial*	0.61 ± 0.35	0.596 ± 0.317	*Trivial*	0.006 ± 0.03	0.008 ± 0.026	*Trivial*	0.012 ± 0.022	0.002 ± 0.029	*Small*
PER	0.57 ± 0.35	0.27 ± 0.15	*Large*	0.23 ± 0.12	0.22 ± 0.19	*Trivial*	0.96 ± 0.25	0.85 ± 0.29	*Small*	0.86 ± 0.31	0.73 ± 0.34	*Small*
3W Training PER	–	–	–	–	–	–	5.26 ± 1.24	8.29 ± 1.68	*Moderate*	7.04 ± 1.09	7.45 ± 1.79	*Moderate*
3W Training Minutes	–	–	–	–	–	–	796 ± 171	964 ± 201	*Moderate*	725 ± 186.50	873.30 ± 202.87	*Moderate*
3W Total PER	–	–	–	–	–	–	7.71 ± 1.08	8.29 ± 1.68	*Small*	7.04 ± 1.09	7.45 ± 1.79	*Small*
3W Average FAI	–	–	–	–	–	–	0.65 ± 0.14	0.67 ± 0.18	*Trivial*	0.629 ± 0.14	0.77 ± 0.22	*Moderate*

4 Conclusions and Future Work

The presented approach allowed to observe considerable relation between training variations and match performance. The players presenting higher variations during training reflected in higher values in 11 of the 15 analyzed variables for locomotor (4/4), metabolic (4/4) and mechanical (3/7) groups in the next matches, and also lower training minutes and accumulated load during training. This approach might provide a way for analyzing the adaptation of players to training dynamics, and even to evaluate training design. The procedure follows a series of simplifications such as the selection of day-type MD-3 which might incur in loss of information. However, this type of calculations can be easily integrated to daily routine performance analysis carried out by physical coaches, without the need of additional systems or requiring high processing times. The findings provide sufficient evidence to suggest the incorporation of this calculation in daily analysis and track its evolution in order to further measure is

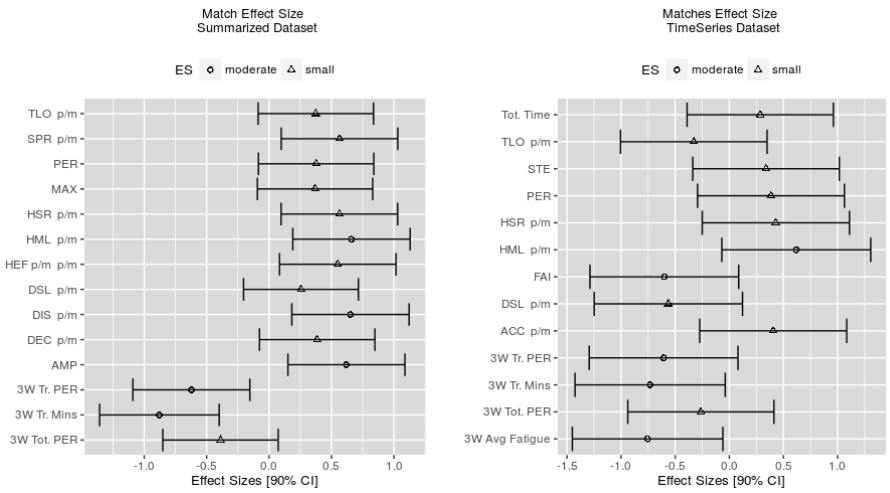


Fig. 2: Effect size differences in group mean values in standardize units for matches groups found through the summarized dataset (left) and the timeseries dataset (right). Trivial effect sizes are not shown.

effectiveness on relating with match performance.

The summarized dataset allowed a more representative grouping and more conclusive results. In practice, high and low variations can be found directly by using the ranges found by the clustering procedure for each variable. Also, time-window aggregated information is showing to add value for performance analysis and should be considered in future research. On the other hand, DTW could not provide sufficiently clear results in this study, most probably due the the short-size characteristics of the analyzed time series and that exact match of variation patterns might be too strict for the few data available. Also, the player normalization seems to favor a cleaner comparison between players, instead of using absolute values which could lead to differences that are more related to physical characteristics than actual adaptation patterns.

This is the first study, up to our knowledge, to relate training and match physical values directly registered from player using EPTS devices during training and matches for a whole season. In the following years, with higher availability of data these remarks must be further validated. Future work should incorporate new day-types in the analysis and factors beyond the physical such as tactical information and variables related with psychological information such as the rate of perceived exertion (RPE). The yearly knowledge of physical evolution of training dynamics and even specific players might provide new insights about the physical preparation of teams and the performance during competition.

References

1. McCall A., Davison M., Carling C., Buckthorpe M., Coutts A.J., Dupont G., Can off-field 'brains' provide a competitive advantage in professional football?. Br J Sports Med, vol. 50, pp. 710-712 (2016)
2. Gyarmati L., Hefeeda M., Estimating the Maximal Speed of Soccer Players on Scale, In Proc. of Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal, (2015).
3. Cummins C., Orr R., OConnor H., West C. Global Positioning Systems (GPS) and Microtechnology Sensors in Team Sports: A Systematic Review. Sports Med vol. 43, pp. 10251042 (2013)
4. IFAB. 129th Annual General Meeting The Football Association, http://resources.fifa.com/mm/document/affederation/ifab/02/60/90/85/ 2015agm_minutes_v10_neutral.pdf (2015)
5. Arjol J., La planificación actual del entrenamiento en fútbol. Análisis comparado del enfoque estructurado y la periodizacin táctica. Acciónmotriz, vol. 8, pp. 27-37 (2012)
6. Lesne A., Chaos in Biology. Biology Forum / Rivista di Biologia. vol. 99, issue 3, pp. 467-481 (2006)
7. Gaudino P., Alberti G., and Iaia M. Estimated metabolic and mechanical demands during different small-sided games in elite soccer players. Elsevier, vol. 36, pp. 123-133. (2014)
8. STATSports Technologies Ltd. STATSSports Viper Metrics. version 1.2 (2012)
9. Sandbakk Ø., Cunningham D., Shearer D., Drawer S., Eager R., Taylor N., Cook C., Kilduff L. Movement Demands of Elite U20 International Rugby Union Players. Plos One vol. 11, issue 4, pp. e0153275 (2016)
10. Talukder, Hisham and Vincent, Thomas and Foster, Geoff and Hu, Camden and Huerta, Juan and Kumar, Aparna and Malazarte, Mark and Saldana, Diego and Simpson, Shawn Simpson. Preventing in-game injuries for NBA players. MIT Sloan Sports Analytics Conference 2016. http://www.sloansportsconference.com/wp-content/uploads/2016/02/1590-Preventing-in-game-injuries-for-NBA-players.pdf
11. Bangsbo J, Mohr M., Krustrup P. Physical and metabolic demands of training and match-play in the elite football player. Journal of Sports Sciences, vol. 24 , pp. 665-674 (2006)
12. Ratanamahatana, Chotirat A., and Eamon k. Everything you know about dynamic time warping is wrong Third Workshop on Mining Temporal and Sequential Data. (2004)
13. Desgraupes B., University Paris Ou. (2013) https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf
14. Van der Maaten L., Hinton G. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, vol. 9. pp. 2579-2605 (2008)
15. Cohen, J. A power primer. Psychological Bulletin, vol. 112, pp. 155-159 (1992)
16. Hopkins WG, Marshall SW, Batterham AM, and Hanin J. Progressive statistics for studies in sports medicine and exercise science. Medicine and science in sports and exercise vol. 41, pp. 3-13. (2009)

From Training to Match Performance: A Predictive and Explanatory Study on Novel Tracking Data

***Abstract*—The recent FIFA approval of the use of Electronic Performance and Tracking Systems (EPTS) during competition, has provided the availability of novel data regarding physical player performance. The analysis of this kind of information will provide teams with competitive advantages, by gaining a deeper understanding of the relation between training and match load, and individual player's fitness characteristics. In order to make sense of this physical data, which is inherently complex, machine learning algorithms that exploit both non-linear and linear relations among variables could be of great aid on building predictive and explanatory models. Also, the increasing availability of information brings the necessity and the challenge for successful interpretation of these models in order to be able to translate the findings into information that can be quickly applied by fast-paced practitioners, such as physical coaches. For season 2015-2016 <*club−name*> [1] has collected both physical information from both training sessions and matches using EPTS devices. This study focuses primarily on evaluating up to what extent is possible to predict match performance from training and match physical information. Different machine learning algorithms are applied for building predictive regression models, in combination with feature selection techniques and Principal Component Analysis (PCA) for dimensionality reduction. Physical Variables are segmented into three groups: Locomotor, Metabolic and Mechanical variables, reaching successful prediction rates in 11 out of 17 total variables, based on a threshold determined by expert physical coaches. A normalized root mean square error metric is proposed that allows better understanding of results for practitioners. The second part of this study is focused on understanding the predictor variables that better explain each of the 17 analyzed match variables. It was found that specific variables can act as representatives of the set of highly correlated ones, so reducing greatly the amount of variables needed in the periodical physical analysis carried out by coaches, passing from 17 to 4 variables in average.**

I. Introduction

The recent availability of all kinds of quantitative data in professional sports, from general statistics to in-game detailed events, is currently attracting the interest from the data science community and is believed to provide a competitive advantage in the following years [1]. The application of statistical analysis has provided developments in critical tasks such as team tactics evaluation, opponent analysis, player scouting and training design [2], [3]. However, few of the current studies are devoted to the analysis of physical information of the players, mainly due to the difficulty of having access to this data through training and competition, which is considered highly valued by football clubs [4]. Typically, such information is gathered through the use of electronic performance and tracking systems (EPTS) which include GPS and microsensor technology such as accelerometers, gyroscopes and magnetometers. Collecting this information was not allowed during official football competition until the recent authorization of the Football Association Board (IFAB), for the 2015-2016 season [5]. These devices have been increasingly adapted and accepted in sports such as Rugby, Australian football, Cricket and Hockey [6]. Despite some concerns over the reliability of GPS measurement of accelerations, especially at low sample rates, it has been an important parameter for analyzing the activity profile in team sports [7]. Such is the case of professional sections at <*club−name*> where these tools are used for monitoring load and many other physical variables.

At <*club−name*>, EPTS devices have been recently used to aid the evaluation of the applied training methodology, the *structured training* [8], a system that sets the baselines for the planning and adaptation of the training activities along the season. Within 3 weeks periodization frames, physical coaches design strategies to induce player adaptation taking into account training activities and the competition, considering the latter the most relevant stimulus to optimize the athlete's capabilities. The information that is provided by EPTS devices becomes then highly important to analyze the physical demands of the sessions and the performance of both individual players and the team as a whole. However, this also presents to coaches a wide set of new variables, most of which were not previously quantified, that need to be understood and incorporated within the weekly design and analysis process. Also, the availability of matches data provides the opportunity to relate physical performance during competition and training, guiding a more fine-grained design of player adaptation, and adding information for better understanding of each player's fitness profile.

Beyond the availability of new data, it becomes essential that efforts to analyze and make sense of this data can be translated into practice. As proposed by Aaron J. Coutts, the laborious and slow-paced research effort based on robust and detailed analysis, must be able to produce findings and results that can be applied by fast-working practitioners [10], which commonly act (and need to act) quickly, intuitively and emotionally. Latest EPTS devices provide over a hundred variables that aim to quantify the different physical efforts and responses of players. However, this amount of information makes unfeasible for physical coaches to perform

[1]Name withheld to comply with the double-blind submission policy. It is one of the major clubs of the European scene.

a one-to-one variable analysis in a frequent basis and be able to reach conclusions quickly. This opens the door for statistical analysis for exploring the relations among variables, understanding which are more informative, and providing mechanisms for simplifying the fast-paced periodical analysis.

The main purpose of this study is to analyze up to what extent is possible to predict the values of 17 physical variables in upcoming matches, and understanding which other variables contribute to that information. $<club{-}name>$ second team data from season 2015-2016 is used, which contains 153 training sessions and 34 matches from 42 different players. Machine learning algorithms that exploit either linear or non-linear relations among variables are applied, within regression analysis. Also, two different feature selection strategies are evaluated with the aim of reducing the noise caused of highly correlated variables which occur with high frequency, facilitating variable analysis and increasing prediction accuracy. Random forests are further used for obtaining the importance of the predictor variables for each of the target variables. Mean Square Error (MSE) is used for evaluating the quality of the models. A second metric, Normalized Root Mean Squared Error (NRMSE), is introduced that allows assessing the results in more practical terms. The original data is also expanded with aggregated historical variables in order to evaluate its influence in explaining future outcomes. This papers presents a detailed description of the proposed methodology and the results of applying it to a full season data.

II. Methodology

This section presents the different phases of the applied methodology, from the collection and preparation of data for the construction of regression models for explaining upcoming matches physical performance.

A. Data Collection

$<club{-}name>$ has collected both training and matches physical performance measurements, for season 2015-2016, using the *StatsSports GPS Viper Pod* devices, which are carried by individual players. The resulting tracking information is manually segmented by physical coaches, which cut parts of the session where the player was not involved in specifics drills. During this process, a software integrated with the device allows to obtain the overall and segmented results of the session distributed over a hundred variables. From this set of variables, physical coaches have selected 17, described in Table I, which summarize the physical information considered most relevant performance information. The data consists of 153 training sessions and 34 matches, which adds up to 2478 training rows and 473 match rows among all the 42 different players throughout the season 2015-2016. The season information is queried from the central database containing the total 2951 rows, where each one contains the measured variables for a single player in a specific session and additional variables that contextualize the information such as player id, position, name, total session time, the session id and session type.

TABLE I
DESCRIPTION OF SELECTED PHYSICAL VARIABLES SPLIT IN THREE GROUPS: LOCOMOTOR, METABOLIC AND MECHANICAL.

Locomotor Variables	
Name and Acronym	**Description**
Travelled Distance (DIS) [11]	Total distance travelled during session drills or matches
Sprints (SPR) [11]	Number of times over $5.5m/s$ during $> 1s$
High Speed Running (HSR) [11]	Travelled meters when speed $> 5.8m/s$
Max Speed (MAX) [11]	Maximum speed reached by the player
Ratio HI/LI (RHL)	The ratio of travelled distances at high intensity ($> 5.8m/s$) and low intensity ($< 5.8m/s$)

Metabolic Variables	
Name and Acronym	**Description**
Average Metabolic Power (AMP) [11]	Energy expended by the player per second per kg, measured in W/Kg
High Metabolic Load Distance (HML) [11]	Distance travelled by a player when the metabolic power is $> 25.5W/Kg$
High Metabolic Efforts (HEF) [13]	The number of separate movements/efforts undertaken in producing HML distance
Equivalent Metabolic Distance (EMD) [11]	Distance in metres that an athlete would need to cover at a constant speed to expend the total amount of energy.
Load Percentage (PER)	Proportion of AMP with respect to an average 9.5 AMP in matches
Speed Intensity (SPI) [11]	Total exertion of a player in a session based on time spent at each speed values.

Mechanical Variables	
Name and Acronym	**Description**
Fatigue Index (FAI) [11]	Accumulated DSL from the total session volume, in terms of speed. (DSL/SPI)
Dynamic Stress Load (DSL) [11]	Total of the weighted impacts, based on accelerometer values over 2g
Lower Speed Loading (LSL) [11]	Load associated with the low speed activity alone
Total Loading (TLO) [11]	The total of the forces on the player over the entire session based on accelerometer data alone
Accelerations (ACC) [11]	Number of increases in speed during at least 0.5 s ($> 3m/s^2$)
Decelerations (DEC) [11]	Number of decreases in speed during at least 0.5 s ($< 3m/s^2$)

B. Data Processing

The dataset is initially processed, adding additional variables that allow further contextualization of each row of data. Each training day is labelled in strict relation with the following match day $<club{-}name>$ training structure. Match day is labelled as MD, the following two days MD+1 and MD+2, and the previous days MD-1 up to MD-4. Each day-type follows specific design rules for training drills. Sessions MD-4 and MD-3 are oriented to strength and resistance, respectively, and also are the more demanding, presenting the

higher differences in absolute values and distribution among players. For simplicity of the study, only MD-3 sessions are used, due to their similarities to match days in terms of number of players, playing spaces and opposition level. Additionally, MD-3 involves the highest differences between physical values. Goalkeepers are deleted from the database since they face considerably different physical challenges than field players. A new variable, load percentage (PER) is added in order to reflect the session load, which is computed as a ratio of the average metabolic power (AMP) from matches. All the measured values are normalized by dividing by the total time of duration of the session. Variables that already represent averages or maximums are kept as originally measured, such as AMP, FI, PER, STE and MAX. Additionally, for each of the physical variables two additional variables are added to dataset, representing the average value of that variable shown by a player in the last 3-week matches and training sessions (MD-3), respectively. We refer to this last two set of variables as historical matches and historical training information. The selection of only MD-3 training information allows to avoid the issue of having historical variables repeated among rows with the same target variable, which will tend to greatly bias the trained model and provide erroneous results.

C. Structure of Data

The different variables presented in Table I are structured in tree main groups regarding the origin of measurement and their nature: metabolic, mechanical and locomotor. The first two groups follow the classification used in a recent paper where metabolic-related variables are associated with energy expenditure and exertion, and mechanical variables relate with intensity changes and impacts [12]. The first two groups contain variables which are calculated in most cases with a combination of GPS and accelerometer with higher influence of GPS in the first one and higher influence of the accelerometer in the second one. The third group, locomotor, refers to calculations associated to simple direct measurements of travelled distance and speed, that are obtained solely through GPS. The relation between the different variables conforming these groups is better detailed in Figure 2.1 where the correlation between each of the predictor variables in MD-3 is presented. It can be observed that metabolic and locomotor variables tend to present high pairwise linear correlation. Also there is a moderate to high correlation between some of the locomotor and the metabolic variables. This is expected since most of the metabolic variables are created through calculations that take into account locomotor variables. Each of this variables is used as a target variable for prediction, thus implying the generation of 17 different datasets, which contain the same predictor variables but different targets. Figure 2.2 presents the boxplot distribution of the different variables for MD-3 and MD. The range of each variable is constrained to the $[0..1]$ range by substracting the minimum and dividing by the difference between the minimum and the maximum values. This is applied to facilitate the visual comparison of variables since their inherent differences in units and magnitudes. Above each boxplot the mean and standard deviation of the original data is presented. It can be observed that the average of all training variables is lower than that of the corresponding match variables, and that the distance among most variable pairs is approximately the same. The exception are some mechanical variables, which exhibit smaller distance than the others. This follows the training design idea of MD-3 which is intended to be as similar as possible to MD but with a proportionally lower load. Following the selection of MD-3 for training data and since the use of match variables as target for prediction, each dataset is reduced to contain strictly the training sessions and aggregated information of players that played the next match. The resulting datasets consists of 217 observations, where the target variable in each case corresponds to one of the 17 match variables to predict. This transforms the original task into 17 different prediction tasks. After adding the historical training and matches variables, and the additional context variables the number of predictors raises up to 71.

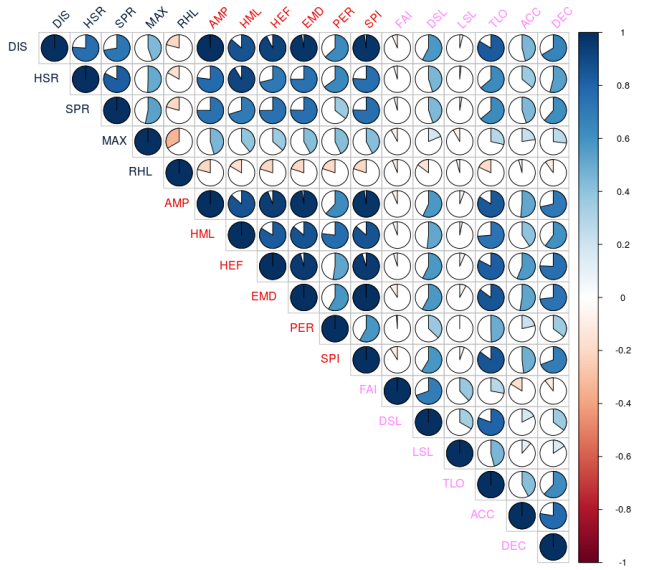
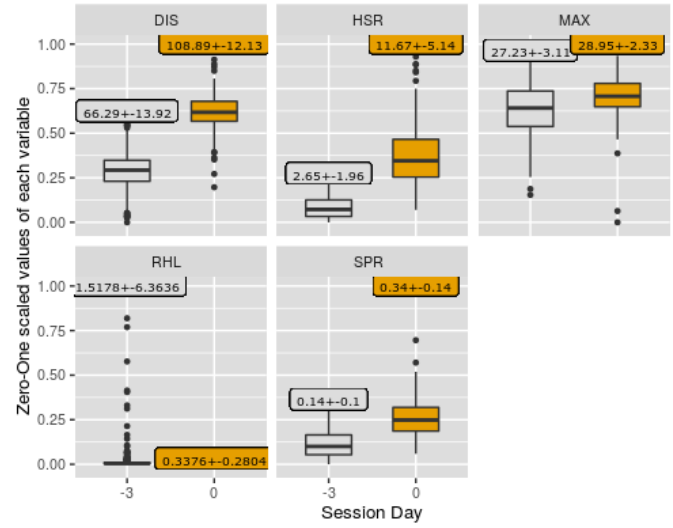


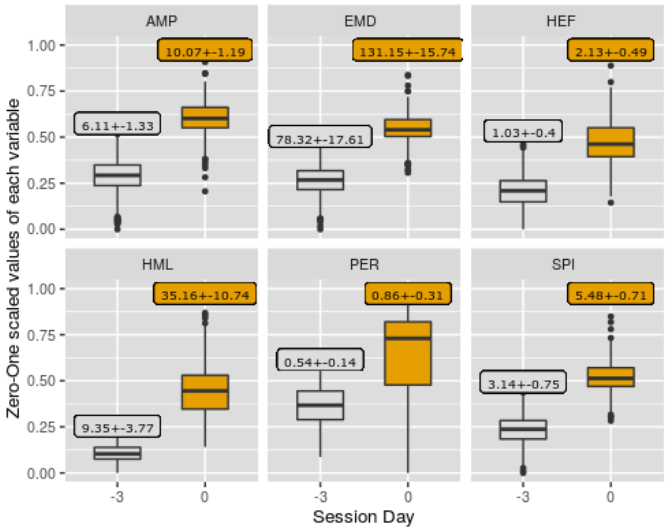
Fig. 2.1. Pairwise Pearson correlation of the target variables from both training and matches data. Variables are organized following the three structured groups from top to bottom: locomotor (blue or dark grey), metabolic (red or medium dark grey), mechanical (pink or light grey).

D. Feature Selection

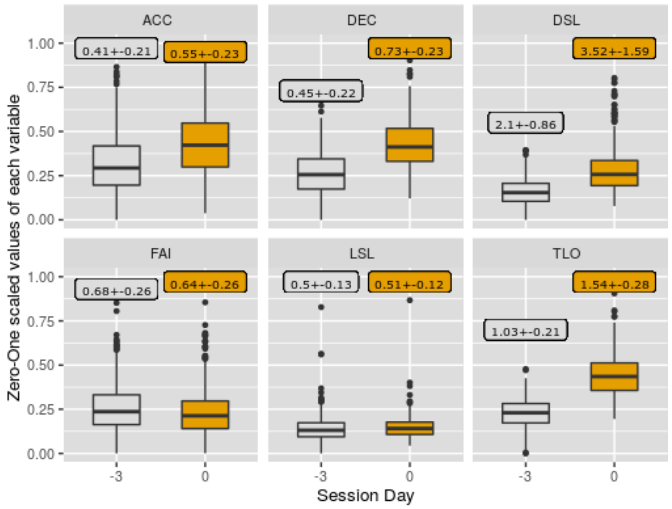
Considering the high number of predictor variables (71) in relation to the number of observations (217), and given the high correlation among some of these variables, feature selection seems like highly desirable. The main advantages of these methods are avoiding overfitting while improving model performance, building faster and cost-effective models, and most importantly, allowing to build more interpretable models by preserving the semantic of original variables [17]. These advantages come at the price of adding additional complexity to the model-building procedure and the possible loss of information that may get unnoticed by the method used. Literature refers to three main types of feature selection



(a) Locomotor variables



(b) Metabolic variables



(c) Mechanical variables

Fig. 2.2. Boxplot distribution of 17 physical variables. Y-axis values are normalized to [0..1] range. Over each boxplot the original mean and standard deviation is presented.

methods: filter methods, which exploit intrinsic properties of the data; wrapper methods which embed the model hypothesis search with the feature subset search; and embedded methods where the search of features is mixed with the model building procedure [17]. For this study we have considered two feature selection approaches: pairwise-correlation selection (COR) and recursive feature elimination (RFE). The first approach, which can be roughly considered a filter method, consists on finding the pairwise Pearson correlation among the predictor variables and removing variables that are above a certain threshold. The second approach was applied by using Random Forest variable importance ranking, which have shown high performance in multiple types of problems, especially those where variables do not vary greatly in their scale of measurements [16]. The COR procedure becomes relevant given the high correlation among some of the predictor variables, as shown in Figure 2.1, which is known to impact negatively on final regression (or classification) error in most machine learning tasks. The COR procedure is always applied before the RFE, since high correlation of predictor variables has been shown to bias the selection of features by wrapper methods, and particularly in the case of random forest [20]. Also, RFE is performed using cross-validation, where average feature ranking is used in order to obtain an unbiased estimator of importance.

For the parwise-correlation filter a threshold of 0.8 Pearson linear correlation was chosen, while for the RFE procedure the set of features achieving the lowest MSE where selected. There exists plenty additional techniques for selecting the optimal number of features, however the described methods were considered sufficient to explore the effect of feature selection in this problem. For both techniques data is previously standardized by transforming each data column to have mean 0 and unit variance.

E. Regression Analysis

A regression analysis procedure is carried out that seeks to evaluate how predictable these variables are, with the given data. For each target variable multiple combinations of pre-processing steps are applied to also multiple different algorithms. Random Forest (RF) and Radial Basis Function Kernel Support Vector Machines (KSVM) were selected as the set of algorithms that exploit non-linear relations among variables. On the other hand Linear Support Vector Machines (LSVM) and Linear Regression (LREG) were used as methods that are based on exploiting linear spaces. Also a set of pre-processing procedures were applied such as the previously described COR and RFE, and principal component analysis (PCA). For each algorithm the set of pre-processing combinations were the following. First the COR filter was either applied or not. For the cases where the filter was applied, the following combinations were also applied: COR+RFE and COR+PCA. PCA was not applied to KSVM since the kernel function is already transforming the feature space. This approach provides 4 different combinations for

each algorithm, except only 3 in the case of KSVM. For each algorithm a parameter selection phase is carried out by testing different parameter combinations. For Random Forest both number of trees ($[50, 100, 250, 500, 750]$) and the number of variables sampled as candidate at each split are tried ([ncol/3,ncol/4] where $ncol$ refers to the total number of predictor variables). For KSVM the tested parameters are the gamma parameter of the Gaussian kernel ($[0.0001, 0.001, 0.01, 0.1, 1, 10, 100]$) and the cost of missclassifications ($[0.0001, 0.001, 0.01, 0.1, 1, 10, 100]$). The same cost of missclassifications list is used for LSVM.

The objective of this analysis is to obtain the best possible model in terms of minimizing prediction error. In order to approximate as much as possible the generalization error, nested cross-validation is used. It is critical to observe that recent studies have shown that when parameter selection is involved within a cross-validation procedure for model building, the average fold error will be biased to the model selection procedure, and thus the obtained error will be lower than the actual generalization capabilities of the model, leading to erroneous results [18]. We deal with this problem using nested cross-validation, where the outer cross-validation estimates the generalization error of a model, while the inner cross-validation optimizes its parameters. As a consequence, different outer fold models will possibly use different parameters. The variance of the errors among the outer folds will also provide an idea of how good or valid the parameter selection procedure is for each algorithm.

The amount of data available is considered insufficient for building a separate Test set beside the Training and Validation sets build during cross-validation. This is why the whole dataset is used during the nested-cross validation procedure (split in subsequent training and validation sets) which, as explained before, is expected to provide a performance error close to the true generalization power of the model, on similar data. For the outer and inner cross-validations 5 and 2 folds are used respectively. It should be noted that feature selection is applied to each of the folds, since these processing steps depend from the training data. Not doing so, would lead to data leakage and thus to an optimistically biased error estimation [21]. Also, standardization is applied to each of the folds.

For evaluating the performance of regression as well as for RFE, the mean square error (MSE) is used and minimized; see Equation 1. From this error we derive and additional error metric: normalized root mean square error (NRMSE), described in Equation 2. NRMSE is used as the ratio of root mean square error and the standard deviation of the target variable. This expresses the magnitude of the obtained error in terms of number of standard deviations of the target variables. Depending on the variable, an expert practitioner can assess if the provided error is acceptable or not for her analysis objectives.

$$MSE = \frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n} \tag{1}$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}}}{\sigma(y)} \tag{2}$$

F. Variable Importance

For each of the independent variables, predictor variable importance is calculated in order to provide a much clear and practical interpretation of their effect. The variable importance ranking from Random Forest is used, as well as in the RFE feature selection procedure described earlier. This approach is based on calculating the mean increase error (MIE), as an analogous to most typical mean decrease accuracy, which is obtained when predictor variables are randomly permuted. Variables are ranked based on the impact they have in final prediction error when removed. The parameters of the best performing model for Random Forest during the regression stage are selected and a new model is built using 5-fold cross validation analogously. Variable importance in each of the folds is averaged to produce a final variable importance ranking that is expected to provide the most reliable representation of the influence of the predictor variables. The choice of Random Forest derives from the results presented in the following section where the algorithm shows stable results and close to the best (or even the best) in most cases. Thus, the selection of one specific approach simplifies the overall explanation of the importance of variables, since the objective is to grasp the general influence of the different variables among the three defined groups.

Recent studies have shown that variable importance ranking through Random Forests can be biased in presence of highly correlated variables [20]. In order to deal with this, the COR procedure is applied to data before following the model fitting and variable importance calculation. Alternative methods have been proposed in order to approach this problem in a more elegant way [20], although at the price of higher computational costs. These more expensive methods are left out for future work.

In order to visualize the importance of variables a chord diagram is used where the proportional influence of each of the predictor variables is observed. This is further explained in the results section.

III. Results

A. Variable Prediction

The results from a applying the different mentioned algorithms, feature selection, and dimensionality reduction methods are presented in Table II, using the NRMSE metric described earlier. Values under 0.75 NRMSE are considered good results in the sense that they can be translated into practice. This threshold was arbitrarily selected together with physical coaches. The desired threshold was achieved in 11 out

of 17 target variables, mostly distributed among metabolic and mechanical groups. From the locomotor variables group it can be seen that only DIS was able to be successfully predicted, but results were below threshold for the other 4 variables (HSR, SPR, MAX, RHL). This situation might respond to a high association of these variables with specific match dynamics beyond the current fitness state of the player such as the opposition team's tactical game, the score or any other variable beyond the strictly physical performance.

The algorithms exploiting non-linear relations among the variables such as Random Forest and RBF-Kernel SVM showed significantly better results that the linear approaches, and achieved a successful threshold in most of the combinations. Also, the feature selection method based on removing highly correlated variables (COR) showed to be a critical resource in this set of combinations, helping to achieve the best result in each of the successfully predicted variables. Recursive feature elimination (RFE) allowed to improve slightly most of the results, however its high computational cost provides doubt regarding its usefulness in this context. Principal Component Analysis (PCA) did not provide a considerable improvement with the exception of few isolated cases. It is noticeable that, for most of the models performing under 0.75 NRMSE, the variation of prediction among folds of the outer loop from the nested cross validation approach was considerably low. The low variation of prediction can be associated with a high stability of the model and also validates the correctness of the parameter selection approach.

B. Variable Importance

For assessing the variable importance on each of the target variables, Random Forest was used, by applying the COR filter within an analogous nested cross-validation procedure where the average best ranking features among folds were selected. This is a reasonable choice since the obtained results for Random Forest produced the best performing or second best performing models, in most cases, in terms of NRMSE. Also, Random Forest variable importance metrics have been extensively used in literature. The mean increase error (MIE) obtained by the variable importance ranking is expressed in terms of NRMSE. So, the impact of variables is measured in terms of how many standard deviations of the target variable would be added to the prediction error if the variable was missing. Figure 3.3 presents a chord diagram showing the influence of each the predictor variables in each of target physical variables. Variables in the bottom half of the diagram correspond to predictors while the ones at the top half correspond to target variables. The size of the incoming chords for each target are proportional to their influence in terms of mean increase error when they are absent. Just variables above 0.25 MIE are shown. The *3W* suffix of the predictors refer to the average value of that variable during matches in the last 3 weeks. The suffix *3W Tr.* is used instead for average value during last 3 week training sessions. Locomotor predictors are shown in blue, metabolic ones in red and mechanical predictors in green, while non

TABLE II
MEAN PREDICTION ERROR AND STANDARD DEVIATION IN NRMSE UNITS AMONG FOLDS. DARK GRAY CELLS INDICATE THE BEST NRMSE, AND LIGHT GRAY CELLS THE MODELS ACHIEVING UNDER 0.75 NRMSE

Variable	Random Forest			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.74 ± 0.07	0.64 ± 0.05	0.80 ± 0.10	0.66 ± 0.06
HSR (LC)	0.97 ± 0.02	0.99 ± 0.05	0.99 ± 0.06	1.03 ± 0.06
SPR (LC)	0.94 ± 0.04	0.87 ± 0.04	0.89 ± 0.05	0.88 ± 0.03
MAX (LC)	1.12 ± 0.22	0.86 ± 0.07	1.09 ± 0.12	0.93 ± 0.05
RHL (LC)	1.33 ± 0.25	1.15 ± 0.05	1.39 ± 0.22	1.23 ± 0.05
AMP (MB)	0.71 ± 0.02	0.62 ± 0.05	0.72 ± 0.03	0.60 ± 0.03
HML (MB)	1.02 ± 0.07	1.01 ± 0.05	1.04 ± 0.00	1.03 ± 0.06
HEF (MB)	0.77 ± 0.02	0.69 ± 0.02	0.76 ± 0.07	0.70 ± 0.03
EMD (MB)	0.79 ± 0.03	0.70 ± 0.05	0.79 ± 0.02	0.72 ± 0.04
PER (MB)	0.92 ± 0.06	0.80 ± 0.06	0.95 ± 0.04	0.79 ± 0.04
SPI (MB)	0.76 ± 0.03	0.67 ± 0.03	0.80 ± 0.04	0.70 ± 0.03
FAI (MC)	0.72 ± 0.03	0.71 ± 0.01	0.85 ± 0.06	0.72 ± 0.01
DSL (MC)	0.68 ± 0.02	0.80 ± 0.05	0.93 ± 0.06	0.77 ± 0.05
LSL (MC)	0.98 ± 0.11	0.96 ± 0.05	1.03 ± 0.08	0.99 ± 0.12
TLO (MC)	0.69 ± 0.03	0.77 ± 0.04	0.87 ± 0.02	0.72 ± 0.04
ACC (MC)	0.64 ± 0.04	0.65 ± 0.04	0.80 ± 0.04	0.63 ± 0.04
DEC (MC)	0.70 ± 0.01	0.64 ± 0.02	0.79 ± 0.05	0.64 ± 0.03

Variable	RBF-K SVM		
	PLAIN	COR	COR+RFE
DIS (LC)	0.66 ± 0.06	0.67 ± 0.08	0.67 ± 0.04
HSR (LC)	1.03 ± 0.06	0.99 ± 0.10	0.98 ± 0.08
SPR (LC)	0.88 ± 0.03	0.87 ± 0.02	0.84 ± 0.04
MAX (LC)	0.93 ± 0.05	0.92 ± 0.03	0.92 ± 0.04
RHL (LC)	1.23 ± 0.05	1.01 ± 0.10	1.00 ± 0.10
AMP (MB)	0.60 ± 0.03	0.71 ± 0.07	0.66 ± 0.05
HML (MB)	1.03 ± 0.06	1.02 ± 0.06	0.96 ± 0.07
HEF (MB)	0.70 ± 0.03	0.66 ± 0.05	0.71 ± 0.04
EMD (MB)	0.72 ± 0.04	0.67 ± 0.03	0.69 ± 0.04
PER (MB)	0.79 ± 0.04	0.70 ± 0.12	0.70 ± 0.07
SPI (MB)	0.70 ± 0.03	0.67 ± 0.04	0.68 ± 0.05
FAI (MC)	0.72 ± 0.01	0.73 ± 0.01	0.79 ± 0.04
DSL (MC)	0.77 ± 0.05	0.83 ± 0.08	0.86 ± 0.08
LSL (MC)	0.99 ± 0.12	0.98 ± 0.02	0.99 ± 0.02
TLO (MC)	0.72 ± 0.04	0.79 ± 0.06	0.85 ± 0.05
ACC (MC)	0.63 ± 0.04	0.66 ± 0.04	0.68 ± 0.03
DEC (MC)	0.64 ± 0.03	0.68 ± 0.05	0.66 ± 0.05

Variable	Linear SVM			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.67 ± 0.08	1.86 ± 0.63	0.82 ± 0.18	0.82 ± 0.18
HSR (LC)	0.99 ± 0.09	1.92 ± 0.50	1.07 ± 0.23	1.10 ± 0.23
SPR (LC)	0.87 ± 0.03	2.05 ± 0.71	0.94 ± 0.07	0.95 ± 0.07
MAX (LC)	0.91 ± 0.03	2.89 ± 0.98	0.99 ± 0.12	1.00 ± 0.12
RHL (LC)	1.00 ± 0.08	1.84 ± 0.48	1.83 ± 0.06	0.97 ± 0.06
AMP (MB)	0.78 ± 0.18	1.37 ± 0.40	0.80 ± 0.06	0.66 ± 0.06
HML (MB)	1.02 ± 0.06	1.79 ± 0.39	0.99 ± 0.08	1.02 ± 0.08
HEF (MB)	0.66 ± 0.06	1.62 ± 0.07	1.02 ± 0.16	0.85 ± 0.16
EMD (MB)	0.67 ± 0.03	1.94 ± 0.42	0.98 ± 0.18	0.82 ± 0.18
PER (MB)	0.74 ± 0.08	0.74 ± 0.29	0.48 ± 0.29	0.82 ± 0.29
SPI (MB)	0.67 ± 0.04	1.62 ± 0.53	0.94 ± 0.18	0.88 ± 0.18
FAI (MC)	0.74 ± 0.01	1.17 ± 0.30	0.80 ± 0.02	0.75 ± 0.02
DSL (MC)	0.83 ± 0.08	1.01 ± 0.20	0.91 ± 0.16	0.90 ± 0.16
LSL (MC)	0.98 ± 0.02	1.19 ± 0.26	0.97 ± 0.05	0.95 ± 0.05
TLO (MC)	0.79 ± 0.06	1.37 ± 0.52	0.82 ± 0.09	0.82 ± 0.09
ACC (MC)	0.68 ± 0.03	1.36 ± 0.39	0.96 ± 0.18	0.84 ± 0.18
DEC (MC)	0.69 ± 0.05	1.83 ± 0.50	0.98 ± 0.06	0.84 ± 0.06

Variable	Linear Regression			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.84 ± 0.17	3.28 ± 0.84	1.13 ± 0.11	0.86 ± 0.16
HSR (LC)	1.13 ± 0.33	3.12 ± 1.33	2.14 ± 1.08	1.65 ± 0.72
SPR (LC)	0.95 ± 0.05	3.96 ± 2.41	2.10 ± 0.97	1.07 ± 0.24
MAX (LC)	1.05 ± 0.18	4.36 ± 2.52	1.72 ± 0.54	1.62 ± 0.64
RHL (LC)	1.00 ± 0.03	4.23 ± 1.85	2.95 ± 1.53	2.61 ± 1.38
AMP (MB)	0.82 ± 0.10	3.02 ± 0.52	1.05 ± 0.07	0.89 ± 0.13
HML (MB)	1.03 ± 0.12	3.04 ± 0.89	1.46 ± 0.46	1.42 ± 0.43
HEF (MB)	1.25 ± 0.65	2.90 ± 0.98	1.04 ± 0.11	1.35 ± 0.51
EMD (MB)	0.96 ± 0.24	2.78 ± 1.08	1.02 ± 0.11	1.04 ± 0.36
PER (MB)	1.04 ± 0.12	0.75 ± 0.24	0.47 ± 0.14	0.98 ± 0.14
SPI (MB)	0.86 ± 0.12	3.04 ± 0.43	1.03 ± 0.16	0.94 ± 0.23
FAI (MC)	0.78 ± 0.03	2.04 ± 1.07	1.01 ± 0.38	0.81 ± 0.04
DSL (MC)	0.91 ± 0.12	1.84 ± 0.88	0.92 ± 0.14	1.03 ± 0.17
LSL (MC)	1.00 ± 0.08	2.01 ± 1.13	1.16 ± 0.16	1.13 ± 0.18
TLO (MC)	0.89 ± 0.10	2.10 ± 0.36	0.98 ± 0.10	1.02 ± 0.31
ACC (MC)	0.74 ± 0.04	2.38 ± 1.15	1.12 ± 0.26	0.78 ± 0.08
DEC (MC)	0.91 ± 0.12	2.18 ± 1.43	1.24 ± 0.39	0.82 ± 0.10

TABLE III
NUMBER OF VARIABLES EXPLAINING EACH OF THE TARGET PHYSICAL VARIABLES, WHERE THE MEAN INCREASE ERROR (MIE) IS ABOVE THREE DIFFERENT THRESHOLDS.

Locomotor Variables			
Variable	**MIE > 0.5**	**MIE > 0.25**	**MIE > 0.10**
DIS	0	3	15

Metabolic Variables			
Variable	**MIE > 0.5**	**MIE > 0.25**	**MIE > 0.10**
AMP	1	5	16
HEF	1	5	14
EMD	1	3	15
PER	1	3	12
SPI	0	3	17

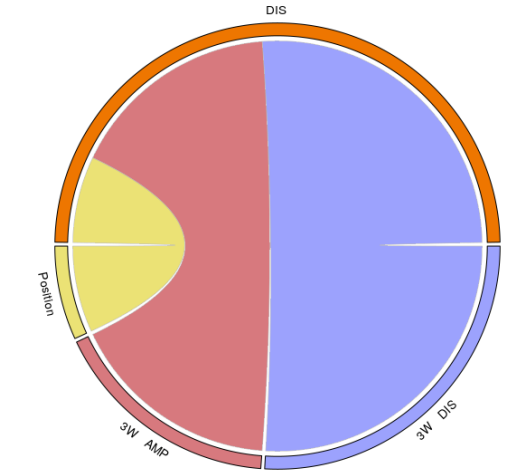
Mechanical Variables			
Variable	**MIE > 0.5**	**MIE > 0.25**	**MIE > 0.10**
FAI	2	4	13
DSL	2	5	13
LSL	0	1	16
TLO	1	8	19
ACC	1	4	16
DEC	1	5	16

physical variables are drawn in yellow.

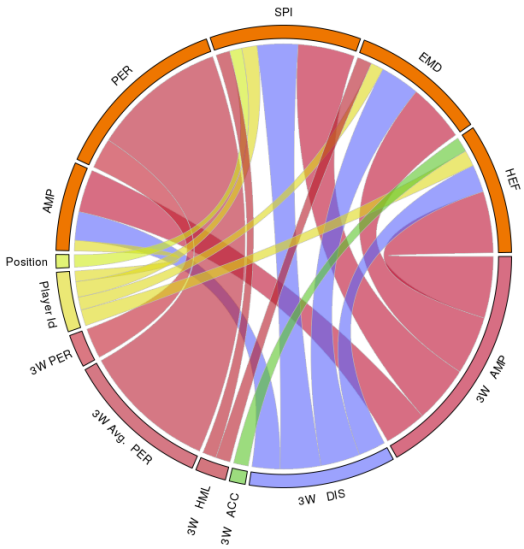
From the three figures one can observe the influence of two or three types of variables in the top ranking predictors. Both for the locomotor and metabolic groups two main variables from each one function were selected as best predictors (3W AMP and 3W DIS). Given that the COR filter has been previously applied, these two variables are acting as representatives of the variables highly correlated with them in each group. In this sense, for example, 3W AMP can be used to explain or understand a large part of future SPI, EMD, HEF and AMP values. Similarly 3W SPI could be selected instead by the COR filter as surrogate of these variables and would have a similar predictive effect than 3W AMP. This brings the idea that, instead of requiring to analyze a high amount of variables for explaining player behavior, the highly correlated variables could be substituted by one representative with a similar effect. For mechanical variables a similar effect is observed with 3W FAI, 3W DSL and 3W ACC. Is observed that wide majority of the better explaining predictors correspond to 3-week average of match physical variables instead of training information. Also, the player id and position play a relevant role for predicting most of the variables, providing the idea that the inherent differences between players and positions also determine the forecast of values, which is an expected result. Table III presents the number of predictors with a level of importance of over 0.5, 0.25 and 0.10 MIE for each target variable. It can be seen that for moderate level of over 0.25 MIE (in NRMSE) variables can be explained by 3 to 5 predictors in average.

IV. PRACTICAL APPLICATIONS

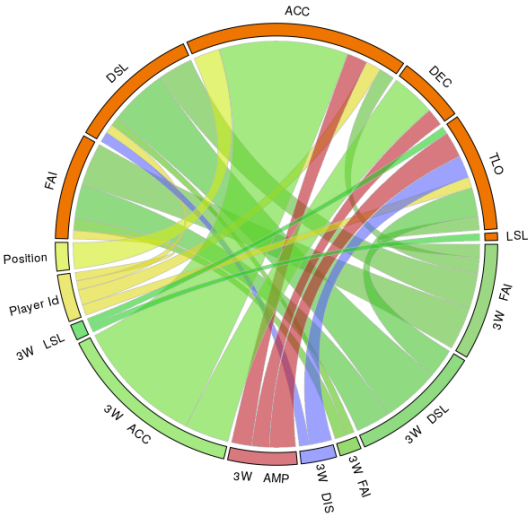
The results of this study provide two specific practical applications. First, the capacity of predicting future variables



(a) Locomotor variables



(b) Metabolic variables



(c) Mechanical variables

Fig. 3.3. Chord diagrams of influence of variables with a MIE higher than 0.25

allows physical coaches to evaluate the fitness state of a player (up to a limit), and also to analyze the effects of training and match load on players. Instead of using hand-designed threshold for variables and performing univariate analysis, the relation among multiple variables can be assessed to more accurately predict or explain future behaviour. The second practical application is the use of a widely shorter amount of variables in the fast-paced daily analysis, by acknowledging which variables explain others and the use of representative variables.

V. Conclusions and Future Work

This study shows that it is possible to predict physical variables based on training and match information from EPTS devices. Past match information provides critical value on predicting future match performance, possibly due to the idea that competition efforts are the highest demanding for players and where stimuli are not controlled such as in training sessions, thus leading to more challenging but also more representative information. Historical aggregates of both match and training session physical variables shown a highly relevant influence within the predictive models. The prediction error achieved for 11 of 17 variables might allow its direct application in practice and is suggested to be incorporated as additional information for the physical coaches routinely evaluation. Future studies should also incorporate internal metrics such as the rate of perceived exertion (RPE) and heart rate exertion (HRE), as well as tactical information, for providing a more robust context of information. For the three groups of variables, both metabolic and mechanical ones showed to be more accurately predictable. Locomotor variables prediction were less well performing possibly due to a high dependency on match-specific and tactical conditions.

Both algorithms exploiting non-linear relations on physical variables performed considerably better than linear models, providing a glance of the complexity of this type of data. We observed the presence of highly correlated features whose fine-grained removal produced a considerable improvement for the predictions. Recursive feature elimination helped to improve the results only slightly while PCA did not produce much advantage for the predictions. We introduce the use of NRMSE as an error metric for regression that can be more easily translated into practice.

The observation of the importance of variables for prediction provided an insight on the influence of the three defined type of variables. The use of representative variables for highly correlated ones could provide a crucial simplification of the fast-paced analysis carried out by practitioners. These observations are relevant due to the increasing availability of new variables everyday which might obstruct the analysis if not properly acknowledged.

References

[1] McCall A., Davison M., Carling C., Buckthorpe M., Coutts A.J., Dupont G., Can off-field 'brains' provide a competitive advantage in professional football?. Br J Sports Med, vol. 50, pp. 710-712 (2016)

[2] Memmert D., Lemmink KA., Sampaio J. Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. Sports Medicine. pp.1-10 (2016)

[3] Folgado H., Duarte R., Fernandes O., Sampaio J. Competing with lower level opponents decreases intra-team movement synchronization and time-motion demands during pre-season soccer matches. PLoS One. vol. 9, n. 5, pp. e97145 (2014)

[4] Gyarmati L., Hefeeda M., Estimating the Maximal Speed of Soccer Players on Scale, In Proc. of Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal, (2015).

[5] IFAB. 129th Annual General Meeting The Football Association, http://resources.fifa.com/mm/document/affederation/ifab/02/60/90/85/2015agm_minutes_v10_neutral.pdf (2015)

[6] Cummins C., Orr R., OConnor H., West C. Global Positioning Systems (GPS) and Microtechnology Sensors in Team Sports: A Systematic Review. Sports Med vol. 43, pp. 10251042 (2013)

[7] Hader K., Mendez-Villanueva A., Palazzi D., Ahmaidi S., Bucheit M. Metabolic Power Requirement of Change of Direction Speed in Young Soccer Players: Not All is What It Seems. PLoS One. pp. e0149839. (2016)

[8] Mallo J. Seirul.lo's Structured Training in Editorial Topprosoccer S.L., Spain. *Complex Football: from Seirul.lo's Structured Training to Frade's tactical Periodisation*. vol. 1, pp. 65-116. (2015)

[9] Arjol J., La planificación actual del entrenamiento en fútbol. Análisis comparado del enfoque estructurado y la periodizacin táctica. Acciónmotriz, vol. 8, pp. 27-37 (2012)

[10] Coutts, A. Working Fast and Working Slow: The Benefits of Embedding Research in High Performance Sport. International journal of sports physiology and performance, vol. 11, pp. 1-2 (2016)

[11] STATSports Technologies Ltd. STATSSports Viper Metrics. version 1.2 (2012)

[12] Gaudino P., Alberti G., and Iaia M. Estimated metabolic and mechanical demands during different small-sided games in elite soccer players. Elsevier, vol. 36, pp. 123-133. (2014)

[13] Sandbakk Ø., Cunningham D., Shearer D., Drawer S., Eager R., Taylor N., Cook C., Kilduff L. Movement Demands of Elite U20 International Rugby Union Players. Plos One vol. 11, issue 4, pp. e0153275 (2016)

[14] Talukder, Hisham and Vincent, Thomas and Foster, Geoff and Hu, Camden and Huerta, Juan and Kumar, Aparna and Malazarte, Mark and Saldana, Diego and Simpson, Shawn Simpson. Preventing in-game injuries for NBA players. MIT Sloan Sports Analytics Conference 2016. http://www.sloansportsconference.com/wp-content/uploads/2016/02/1590-Preventing-in-game-injuries-for-NBA-players.pdf

[15] Bangsbo J, Mohr M., Krustrup P. Physical and metabolic demands of training and match-play in the elite football player. Journal of Sports Sciences, vol. 24 , pp. 665-674 (2006)

[16] Strobl C., Boulesteix A., Zeileis A., and Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics, vol. 8, p.1. (2007)

[17] Saeys Y., Inza I,, and Larrañaga P. A review of feature selection techniques in bioinformatics. Oxford Univ Press, vol.23,num. 19, pp 2507-2517 (2007)

[18] Cawley, Gavin C and Talbot, Nicola LC, On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, vol.11, pp 2079-2107 (2010)

[19] Petersohn C. Training and Testing Strategies. In Jörg Vogt Verlag. *Temporal video segmentation*. pp 32-34 (2010)

[20] Strobl C., Boulesteix A., Kneib T., Augustin T, and Zeileis A. Conditional variable importance for random forests. BMC bioinformatics, vol. 9, pp 1 (2008)

[21] Petersohn C. Model Assesment and Selection. Springer series in statistics Springer, Berlin. *The elements of statistical learning*. vol. 1, pp 245-247 (2001)