2016 IEEE 16th International Conference on Data Mining Workshops

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

Javier Fernández
Sports Science and Health Department
Fútbol Club Barcelona
Barcelona, Spain
javier.fernandez@pl.fcbarcelona.cat

Daniel Medina
Sports Science and Health Department
Fútbol Club Barcelona
Barcelona, Spain
daniel.medina@fcbarcelona.cat

Antonio Gómez
Sports Science and Health Department
Fútbol Club Barcelona
Barcelona, Spain
antonio.gomez@fcbarcelona.cat

Marta Arias
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain.
marias@cs.upc.edu

Ricard Gavaldà
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain.
gavalda@cs.upc.edu

***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 F.C. Barcelona 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 F.C. Barcelona where these tools are used for monitoring load and many other physical variables.

At F.C. Barcelona, 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

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.18

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 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. F.C. Barcelona 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

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. 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.

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, 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. 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.

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 ($< 3m/s^2$)

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.

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 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

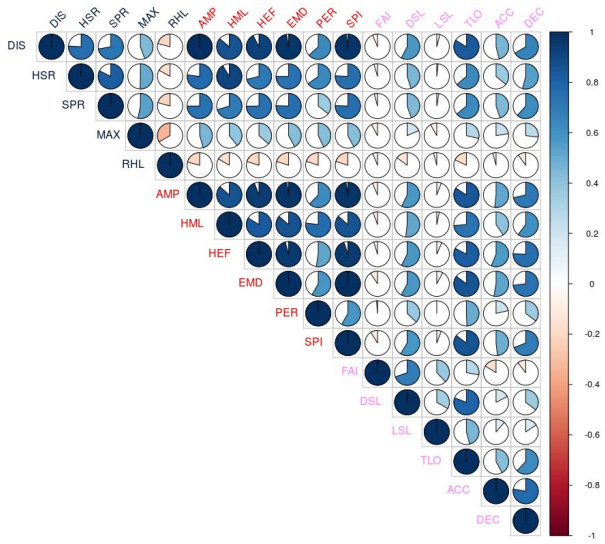
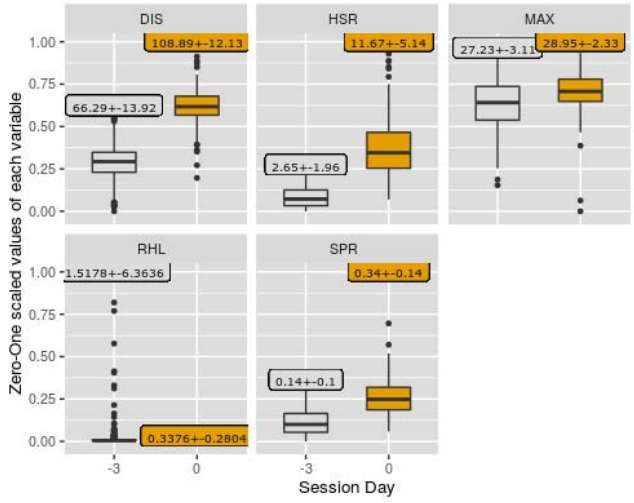


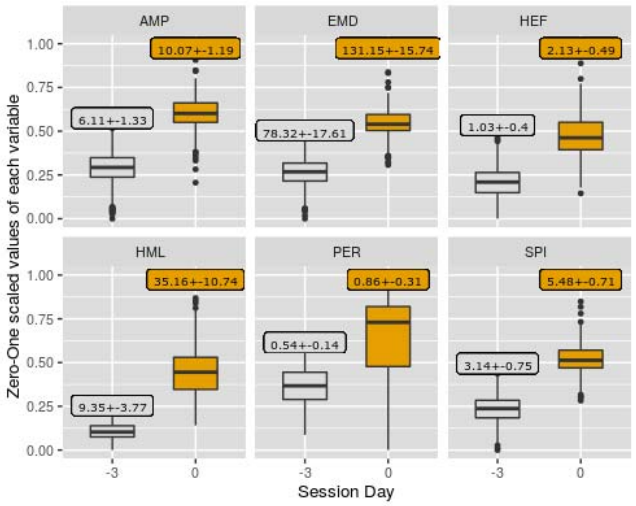
Fig. 2.1. Pairwise Pearson correlation of the target variables from both training and matches data. Variables are organized following the three structured groups from top to bottom: locomotor (blue or dark grey), metabolic (red or medium dark grey), mechanical (pink or light grey). A filled circle refers to full correlation, where blue and red colors refer to positive or negative correlation respectively.

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.

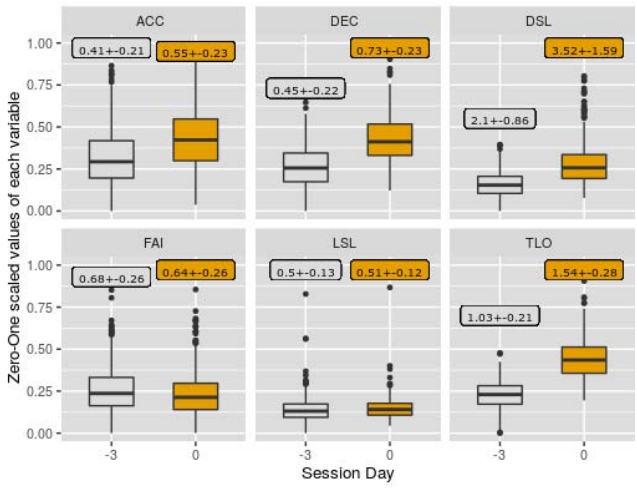
For the parwise-correlation filter a threshold of 0.8 Pearson linear correlation was chosen, while for the RFE procedure the set of features achieving the lowest MSE where selected. There exists plenty additional techniques for selecting the optimal number of features, however the described methods were 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.



(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.

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

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

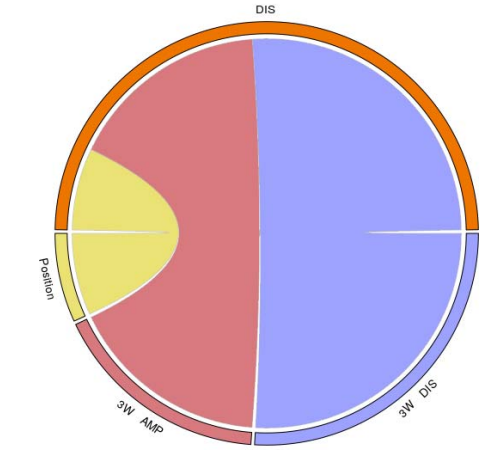
Variable	Linear Regression			
	PLAIN	COR	COR+PCA	COR+RFE
DIS (LC)	0.84 ± 0.17	3.28 ± 0.84	1.13 ± 0.11	0.86 ± 0.16
HSR (LC)	1.13 ± 0.33	3.12 ± 1.33	2.14 ± 1.08	1.65 ± 0.72
SPR (LC)	0.95 ± 0.05	3.96 ± 2.41	2.10 ± 0.97	1.07 ± 0.24
MAX (LC)	1.05 ± 0.18	4.36 ± 2.52	1.72 ± 0.54	1.62 ± 0.64
RHL (LC)	1.00 ± 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

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 physical variables are drawn in yellow.

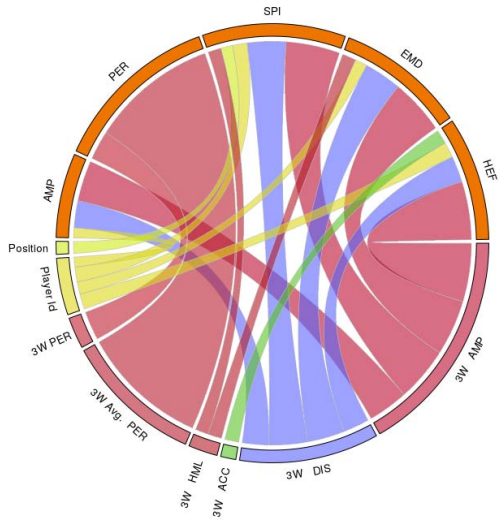
From the three figures one can observe the influence of two or three types of variables in the top ranking predictors. Both for the locomotor and metabolic groups two main variables from each one function were selected as best predictors (3W AMP and 3W DIS). Given that the COR filter has been previously 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. 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.

IV. Practical Applications

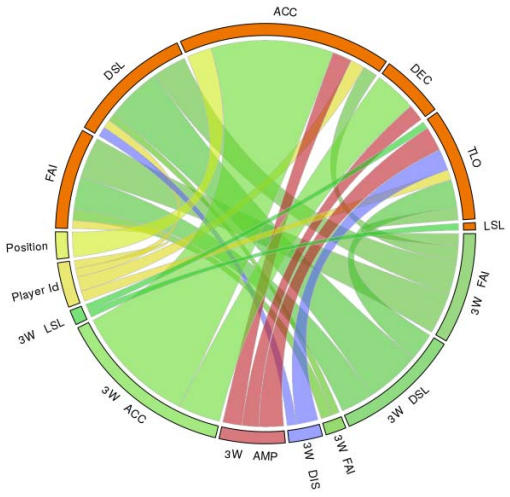
The results of this study provide two specific practical applications. First, the capacity of predicting future variables allows physical coaches to evaluate the fitness state of a player (up to a limit), and also to analyze the effects of training and match load on players. Instead of using hand-designed threshold for variables and performing univariate analysis, the relation among multiple variables can be assessed to more



(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

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.

Acknowledgments

The work of Marta Arias and Ricard Gavaldà is partially funded by the MACDA project of the Generalitat de Catalunya (SGR2014-890) and the APCOM project of MINECO (TIN2014-57226-P).

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)