On modeling player fitness in training for team sports with application to professional rugby

Matthew Revie, Department of Management Science, University of Strathclyde
Kevin J Wilson, School of Mathematics and Statistics, Newcastle University
Rob Holdsworth, Scottish Rugby Union
Stuart Yule, Glasgow Warriors Rugby Club.

Abstract

It is increasingly important for professional sports teams to monitor player fitness in order to optimize performance. Models have been put forward linking fitness in training to performance in competition but rely on regular measurements of player fitness. As formal tests for measuring player fitness are typically time-consuming and inconvenient, measurements are taken infrequently. As such, it may be challenging to accurately predict performance in competition as player fitness is unknown. Alternatively, other data, such as how the players are feeling, may be measured more regularly. This data, however, may be biased as players may answer the questions differently and these differences may dominate the data. Linear Mixed Methods and Support Vector Machines were used to estimate player fitness from available covariates at times when explicit measures of fitness are unavailable. Using data provided by a professional rugby club, a case study was used to illustrate the application and value of these models. Both models performed well with R^2 values ranging from 60% to 85%, demonstrating that the models largely captured the biases introduced by individual players.

Keywords: Machine Learning; Predictive Modelling; Probability; Sports; Statistics

1 Introduction

The differences in performance between athletes in top level sports are often very small (Maughan 2002) and as such minor improvements to training and preparation can be the difference between victory and defeat. In recent years there have been examples of sports teams and organizations who have used scientific approaches to training and preparation and who have then dominated at the highest level such as Great Britain cycling and rowing teams and Team Sky cycling team (James 2012). Ensuring that athletes are at peak performance levels for the most important competitions appears to be a critical factor for success.

In order to achieve this it is necessary to be able to link performance or fitness measures from training to performance in competition. Calvert et al. (1976)

considered a system model to relate data, in the form of a profile, from training to a profile for competition. The model took four parts: endurance, skill, strength and psychological factors, and was based on the model by Banister et al. (1975). The authors simplified these factors down to two functions corresponding to fitness and fatigue. They implemented their model to data from swimmers and showed a good fit. Morton et al. (1990) further simplified this model to represent both fitness and fatigue in a single linear difference equation. They considered the case of individual runners and showed significant observed correlations between model predictions and actual athlete performance. Hellard et al. (2006) investigated the Banister model for 9 elite swimmers to assess its predictive powers. They found that the confidence intervals for model parameters were very wide and parameters were correlated. They suggested penalized models, ridge regression and Bayesian methods as possible solutions to these issues. The measures of training and performance are crucial to the Banister model. A power output, heart rate relationship and application of the Morton simplification of the model for two individual high level cyclists was considered by Scarf et al. (2013).

In sports in which athletes compete individually or as part of a small team, it will often be feasible to take measurements of fitness as inputs to such models fairly frequently. However, in large team sports such as association football, rugby, hockey, etc. in which squads will typically comprise 30-40 players this will usually not be the case and measurements for individual players may only be taken sporadically and unevenly across players. However, teams in such sports will typically collect information much more regularly which may be related to player fitness, both in the form of data and more subjective information.

This paper investigates two methods for estimating individual player fitness using data with reduced entropy but which is collected more regularly. We aim to demonstrate that this data, previously being viewed by both players and coaches as having limited information, can help coaches predict the outcome of tests that they are unable to regularly complete. These estimates can then be used as inputs into performance models and used to aid decision making for training schedules, etc. As each player in a team will only take the formal fitness tests a small number of times over a season, models will be fit to the data for all players. However, there will typically be differences between players, both through differences in the potential fitness levels for individuals and, if subjectively assessed information is used, in the different ways that individuals respond. We propose and evaluate two different approaches which are sufficiently flexible to be able to model these differences; Bayesian linear mixed effects models and Support Vector Machines.

The paper is motivated by work with a professional rugby club and the main contributions of the paper is the application of different quantitative methodologies to the problem faced by the club. We will identify the data available and fit both models to the data to estimate player fitness for the entire squad of players. We will then compare the fits and usefulness of each of the models for this case.

The rest of the paper is structured as follows. In Section 2 we outline the

problem faced and in Section 3 we discuss the application of linear mixed effects models and support vector machines to the estimation of player fitness in team sports generally. In Section 4 we apply the two approaches to the specific problem and in Section 5 we assess the fits of the models in this case. We conclude the paper and suggest areas for further work in Section 6.

2 Motivating Problem

The following data were collected. Between 30th January 2012 and 17th April 2012, 38 players completed a questionnaire on each day that they trained. Each player answered eight questions, marking each question on a ten point ordinal scale. The players were asked to rate, on a scale of 1-10, each of the following: Upper Body soreness (UB), Lower Body soreness (LB), Sleep Quality (SQ), Appetite (APP), Energy (EN), Mood (MD), Stress (STR) and Motivation (MOT). Completing the questionnaires were part of the daily routine of the players. The organization believed that the players' general physical and mental wellbeing could be assessed through the questionnaire.

Over the same period, players infrequently carried out counter movement jump (CMJ) tests. The CMJ was performed with arms fixed by holding a wooden broomstick across the shoulder behind the head. In all, 430 CMJ tests were carried out. The CMJ test is used to measure performance, including muscular strength and anaerobic power Castro-Piero et al. (2010), as well as maximal leg power Marek et al. (2005). To do this, measurements are made on many parameters, including peak power and jump height. Measuring these from the CMJ is a simple way to establish the physical potential of the player to perform explosive based movements utilized on the rugby field and is used by the organization to inform training intensity and team selection. Many measurements were taken using force plates; all of which give information about different aspects of the condition of the player in question. Larry (1998) Larry (1998) questioned the reliability of these devices; however, for the purposes of this paper, we ignored sampling error resulting from measurement inaccuracies. Two measurements were of primary interest to the organization: Peak Distance (height) and Peak Power.

Summary statistics for the 430 observations of Peak Distance and Peak Power are given in Table 1.

Summary	Peak Distance	Peak Power
Minimum	0.330	3313
Lower Quartile	0.435	5934
Median	0.477	6826
Mean	0.477	6776
Upper Quartile	0.522	7450
Maximum	0.633	9694

Table 1: Summary statistics for Peak Distance and Peak Power.

We see fairly symmetric distributions in each case, with the mean and median close together and no obvious outliers either at the top or bottom end of the data.

Despite the usefulness of the data provided by the CMJ, it was carried out infrequently due to time constraints. In contrast, the wellness test scores were collected from the players every day. The organization wished to predict the results of the CMJ based on the answers the players give to the wellness questionnaire. This would highlight each player's fitness and fatigue on days when the CMJ is not performed.

The organization previously used the mean of the eight questions to assess the wellness of each player. This was normalized as a percentage and then categorized as good, normal or bad. From this, in the absence of jump test data, different training schedules were developed for each player. However, upon examining the effect the mean of the eight questions had on the output of the CMJ, the strength of the relationship was found to be insignificant, i.e. $R^2 < 5\%$. In addition, a simple linear regression model was applied to the data. As with the mean, the strength of the relationship was found to be insignificant, i.e. $R^2 < 5\%$.

There were three possible reasons that the basic regression model may perform poorly. First, there may be no information contained in the players' answers to the wellness test which was useful for predicting Peak Power or Peak Distance in the CMJ. Second, each player may have a different potential for Peak Power or Peak Distance, and the simple regressive models used may not be able to capture this. Finally, the players may be answering the questions differently and these differences may be dominating the data. It seemed intuitive that, given that the questionnaire was subjective, while players were internally consistent, there may exist inconsistencies between players in how the questions were answered. Coaches intuitively believed that the signal in the data being captured may be small, but should still be able to inform the output of the CMJ. As such, an alternative method of modeling the questionnaire test that is capable of capturing the effect of the individual players to infer the CMJ score was required.

3 Methods

In this section we discuss the suitability of two standard modeling approaches, linear mixed effects models and support vector machines, to the estimation of player fitness in teams.

3.1 A linear mixed effects model

Mixed effects regression models are widely used tools to predict the values of unknowns of interest $\boldsymbol{Y}$ using fixed variables $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ and random effects, $\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_m}$, for which the values are known, or, in the case of prediction, will be known at some point in the future. In our case this form is useful as we define

$\boldsymbol{Y}$ to be the, possibly multivariate, measure of player fitness for the team and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ to be the information collected by the organization regularly which is related to player fitness. This then allows us to use the $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ to take into account the differences between the players in terms of their underlying ability in the fitness measure and, in the case of subjectively assessed information from the players, the random differences between how players answer questions.

The linear mixed effects model for the player fitness $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, can then be expressed in terms of a linear function of the information collected from the players and a further linear function taking into account the differences between the players. That is,

$$\boldsymbol{\mu} = \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{X} + \boldsymbol{\gamma}^{\boldsymbol{T}} \boldsymbol{Z} + \epsilon,$$

where ϵ are the residuals representing the error between the predictions and measurements of the fitness measures of interest and $\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_n}$ and $\boldsymbol{\gamma} \sim \mathrm{N}(0, \Sigma_\gamma)$ represent the influence of the information collected and individual players respectively.

We could then fit such a model using, for example, restricted maximum likelihood. However, coaches in sports teams will often have knowledge of the relationships between player fitness and other information collected and also of the underlying player potentials in terms of the performance or fitness measure. Thus, we propose a Bayesian approach to inference. For an explanation of Bayesian methods see Lee (2012).

In the Bayesian analysis we represent the beliefs of the coaches on the relationships which exist between the fitness measure and the other information and players using a probability distribution named the prior distribution. We then use the data to update those beliefs and the result is the posterior distribution which represents the beliefs of the coaches considering all of the information at their disposal.

In a Bayesian linear mixed effects model the Gibbs sampler can be used to compute the posterior distribution if conditional distributions for the parameters can be found up to proportionality. We sample from the conditional distributions, discarding samples until the Markov chain has converged. This can be achieved by iterating the following steps

$$\begin{aligned}
L(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\gamma}, \Sigma, \Sigma_\gamma) &\propto \phi(\boldsymbol{y} - \boldsymbol{Z\gamma}; \boldsymbol{X\beta}, \Sigma)\pi(\boldsymbol{\beta}) \\
L(\boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{\beta}, \Sigma, \Sigma_\gamma) &\propto \phi(\boldsymbol{y} - \boldsymbol{X\beta}; \boldsymbol{Z\gamma}, \Sigma)\phi(\boldsymbol{\gamma}; 0, \Sigma_\gamma) \\
L(\Sigma|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \Sigma_\gamma) &\propto \phi(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Z\gamma}; 0, \Sigma)\pi(\Sigma) \\
L(\Sigma_\gamma|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \Sigma) &\propto \phi(\boldsymbol{\gamma}; 0, \Sigma_\gamma)\pi(\Sigma_\gamma),
\end{aligned}$$

where $\phi(\boldsymbol{a}; \boldsymbol{b}, \boldsymbol{c})$ represents the probability density function of the Normal distribution at $\boldsymbol{a}$ with mean $\boldsymbol{b}$ and variance matrix $\boldsymbol{c}$ and $\pi(\cdot)$ represents a prior density.

Having found the posterior distributions, we are able to make predictions of how players will perform on the fitness test given the information collected by the organization.

3.2 A Support Vector Machine

Recently, Support Vector Machines (SVM) (Cristianini & Shawe-Taylor 2000, Vapnik 1999) have been the subject of intensive research (Scholkopf et al. 1996, Vapnik 1999) and have been applied successfully to many classification and regression studies in many disciplines, e.g. energy (Moulin et al. 2004), particle identification (Barabino et al. 1999), face identification (Guo et al. 2000, Osuna et al. 1997), text categorization (Drucker et al. 1999) and bioinformatics (Brown et al. 2000). SVM, based on the work of Vladimir Vapnik in statistical learning theory (Vapnik 1999) have become one of the most popular data mining algorithms within the computing science domain (Wu et al. 2008). Training involves optimization of a convex cost function, and in comparison with other machine learning models, there are no false local minima to complicate the learning process.

A simplistic view of SVM is that the data x are mapped onto a higher dimensional space $\mathcal{F}$ via a non-linear mapping and subsequently, linear regression is carried out in $\mathcal{F}$. The linear regression model in a high dimensional space can be viewed as equivalent to nonlinear regression in a low dimensional space. To do this, margin maximization and kernels are considered. The margin is defined as the smallest distance between the hyperplane and any of the data. The purpose of an SVM is to choose a hyperplane such that it separates the closest members of different groups by as much as possible, i.e. it maximizes the margin. A kernel function is used to map the data to a higher dimension. One challenge of kernel mapping is choosing the most appropriate kernel. While a potentially infinite number of kernel functions exist, a relatively small number of functions have been demonstrated to work well across many problem domains. The most widely used kernel function is the Gaussian Radial Basis Function (RBF), $k(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$ for $\gamma > 0$. In order to use the RBF kernel, we must specify γ.

For those wishing to implement SVM, there are two key elements that need to be controlled: overfitting and underfitting. By transforming the data into a new dimension space, we can avoid underfitting. This is particularly true when we believe that the data do not follow a linear relationship. Overfitting is more challenging to address. Overfitting is common when the data are limited (as patterns may actually be noise), the data have a lot of noise, or when there are many variables and the underlying relationship is not well understood. Clearly in our case there is a risk of overfitting due to the sparse nature of observations of the fitness of individual players in teams. We can control for overfitting and underfitting primarily through the modification of two parameters; C, and γ; where C determines the trade off between allowing errors and enforcing strict margins, while γ is the Gaussian parameter in our RBF kernel. The above only considers SVM to classification problems. To apply SVM to regression problems, a penalty function is adapted such that a penalty is not applied if the predicted value is within a given distance of the actual value, denoted by ξ.

4 Results: Case Study

4.1 The Linear Mixed Effects Model

This section provides an overview of applying the linear mixed effects model to the problem of predicting jump test scores using questionnaire data. All of the modeling for both approaches is implemented using R, (R Development Core Team 2005).

For the jump test we have observations of a number, $i = 1, \ldots, n$, of players, and for each player we have a number of observations from individual tests $j = 1, \ldots, m_i$. That is we have, for each response of interest, observations Y_{ij}. These responses are Peak Power and Peak Distance respectively. For each of these observed responses we also have recorded values for each of the questionnaire variables, $X_{1ij}, \ldots, X_{8ij}$. These are Upper Body, Lower Body, etc.

There is substantial between player variability in both the responses to the questions from the questionnaire and in the two jump test response variables. In addition, as new players come into the squad, we may wish to predict their jump test scores based on no previous data. As a result of these two factors, the following linear mixed model is deemed suitable to represent the relationship between the two jump test responses and the questionnaire answers. The response is $Y_{ij} \sim \mathrm{N}(\mu_{ij}, 1/\tau_{ij})$, where

$$\mu_{ij} = \beta_0 + \beta_1 X_{1ij} + \ldots, \beta_8 X_{8ij} + \gamma_i U_{ij} + \epsilon_{ij},$$

and μ_{ij} is the mean of either Peak Distance or Peak Power for player i on jump j, $X_{1ij}, \ldots, X_{8ij}$ are the values of the questionnaire answers for that player on the day of that jump and $U_{ij} = 1$ so that $\gamma_i \sim \mathrm{N}(0, 1/\tau_\gamma^{(i)})$ is the random effect associated with player i. We find Bayesian estimates of the model parameters as they allow us flexibility in the relationships we define between the model parameters. In order to do so, we first require prior distributions for $\beta_1, \ldots, \beta_8, \tau_{ij}, \tau_\gamma^{(i)}$. Those chosen must be suitable for the range of values the relevant parameter can take. In particular, we take

$$\begin{aligned} \beta_k &\sim \mathrm{N}(m_\beta, v_\beta^{(k)}), \\ \tau_\gamma^{(i)} &\sim \mathrm{gamma}(r_\gamma^{(i)}, \theta_\gamma^{(i)}), \\ \tau_{ij} &\sim \mathrm{gamma}(r, \theta). \end{aligned}$$

It seems reasonable that many of the fixed regression parameters for the responses to the questionnaires, such as Upper Body and Lower Body, will not be independent of one another. That is, knowing one of the covariate values will provide information about the likely values of the other covariates. We can incorporate this dependence in the model by adding a third level of hierarchy to the model. This is

$$m_\beta \sim \mathrm{N}(\mu_\beta, 1/\tau_\beta).$$

The prior, and therefore the model, is fully specified once values have been chosen for $(v_\beta^{(k)}, r_\gamma^{(i)}, \theta_\gamma^{(i)}, r, \theta, \mu_\beta, \tau_\beta)$. In practice, these values have been chosen to define approximate non-informative prior distributions. This was primarily to allow comparison with the SVM approach. An important area of future work will be to elicit informative prior distributions from the coaches, though this will not be a simple task as the model parameters are not directly observable.

We have two linear mixed effects models as detailed above; one for Peak Distance and one for Peak Power. Both models were run for 10,000 iterations of the Gibbs sampler as burn in, after which the Markov chains demonstrated good convergence based on trace plots. The posterior distributions were then calculated using a further 100,000 iterations. There were no indications of problems with autocorrelation in either model and multiple chains mixed well.

The posterior distributions for the model considering Peak Distance are given in Figure 1. The posterior means and variances are given in Table 1 in the Appendix.

Figure 1: The posterior densities of each of the fixed effect parameters for Peak Distance

We can see from the plots that the posterior distributions for all of the fixed effect parameters for Peak Distance follow approximately symmetric distributions. All of the densities contain zero so we cannot say that the coefficients are non-zero. However, for each of $\beta_1, \beta_2, \beta_5$, zero is in the tails of the density indicating that UB, LB and EN are the covariates which are most clearly having an effect on a player's Peak Distance.

We give a similar posterior density plot for all of the fixed effect parameters in the regression model for Peak Power in Figure 2. Again the table of the posterior means and variances, Table 2, is given in the Appendix.

Figure 2: The posterior densities of each of the fixed effect parameters for Peak Power

Many of the densities contain zero in the tails of the distribution. In particular, the covariates which are the most significant in terms of predicting the Peak Power of a player's jump are those associated with $\beta_1, \beta_3, \beta_4$ and β_5. These correspond to the covariates UB, SQ, APP and EN. Once again, all of the posterior densities are approximately symmetric around their modes.

4.2 Support Vector Machines

The following steps are carried out to develop the SVM model. First, the explanatory variables are scaled between [0,1]. Next, in order to assess overfitting

noise in small datasets, for each C and γ, cross-validation is carried out on the data. The data is divided into k equal sized groups. From this, k different experiments are run, ensuring that each data point is included in the test data exactly once. The accuracy of the model is then evaluated across the k experiments rather than a single experiment. Alternative procedures to cross-validation have been proposed by Joachims (2000), Campbell et al. (1999), Chapelle et al. (2002) and more recently by Bishop (2006), Wu & Wang (2009). When developing the training set, in order to capture the effect of each player within each experiment, at least one data point is included for each player.

Once the k experiments are carried out, the mean residual standard deviation and R^2 for the test data is calculated. This is then repeated for different C and γ ranging from $C \in \{1, 1\times10^2, \ldots, 1\times10^6\}$ and $\gamma \in \{1\times10^{-5}, 1\times10^{-4}, \ldots, 1, 1\times 10^2\}$. The mean R^2 for each combination of C and γ for both the overall dataset and the test set for Peak Power are captured in Table 3 in the Appendix.

For certain values, e.g. $\gamma = 1$ or $\gamma = 10$, the model overfits the training data set and performs poorly for the test data. Using the information in Table 3, values of $C = 1000$ and $\gamma = 0.01$ are initially chosen to develop the model. The regions of C and γ close to these values are further explored. From this analysis, $C = 800$ and $\gamma = 0.01$ are chosen to further develop the model. The model developed for these values is evaluated in Section 5. This analysis is repeated for Peak Distance. Details are omitted.

5 Comparison of Modeling

We evaluate the models using two different measures: the residual standard deviation and the coefficient of determination, R^2. Other measurements of accuracy, such as the deviance, the Akaike Information Criteria and the Schwarz Information Criteria (Congdon 2004), could have been used to assess model fit. However, such more complex model selection criteria typically include a calculation about the value of adding extra parameters to models and, as such, the resulting model fit is dependent on the number of parameters in the model. Since the approaches we are considering in this paper are very different and, in particular, utilize parameters in very different ways, we prefer to assess the models using the relatively simple and absolute measures we have defined above.

5.1 Quantitative Evaluation

Each model is used to predict the values of Peak Distance and Peak Power for individual players on each of the days in which they took the jump test based on their questionnaire answers for that day. The linear mixed effects model for Peak Distance produces a residual standard deviation of $s_\epsilon = 0.027$ and $s_\epsilon = 403.11$ for Peak Power, while the SVM produces a residual standard deviation for Peak Distance of $s_\epsilon = 0.035$ and $s_\epsilon = 685.93$ for Peak Power. For Peak Distance, the mean observation is 0.477 and the standard deviation is 0.0601. For both models, an average error between the observed modeled values of approximately

0.001-0.002 is fairly small. In the case of Peak Power, the average error of about 19 and 33 is compared with observations which have a mean of 6775.75 and standard deviation of 1030.567; once again this represents a small average error between the model and the observations. For the jump test results the R^2 values are 79.1% for Peak Distance and 84.7% for Peak Power using the linear mixed effects model. The SVM produces an R^2 value of 66.1% for Peak Distance and 55.7% for Peak Power. For both models, each represents a marked improvement over the basic models as discussed in Section 1 in which the R^2 values were of the order of 5%.

We plot the observed values for each test for each of the players against the predicted values from the model for both Peak Distance and Peak Power to illustrate the model fit. A plot of these quantities for both response variables is given in Figure 3 for the linear mixed effects model and in Figure 4 for SVM.

Figure 3: Observed versus predicted values for Peak Distance (on the left) and Peak Power (on the right) for linear mixed effects model.

Figure 4: Observed versus predicted values for Peak Distance (on the left) and Peak Power (on the right) for SVM.

Figure 3 and Figure 4 illustrate that the observations are fairly evenly scattered around the line $y = x$. This implies that both models are working fairly well for the range of jump test scores of each variable. We can also plot the densities of the model predictions for each of the 430 tests in comparison to the densities of the observed values for Peak Distance and Peak Power. Plots of these quantities are given in Figure 5 and Figure 6.

Figure 5: Densities of the predicted (dashed) and observed (solid) values for Peak Distance and Peak Power using the linear mixed effects model.

For both Peak Distance and Peak Power, the linear mixed effects model is capturing the approximate shape of the density. However, for Peak Power, the model overreacts to multi-modality in the density. The SVM produces a smoother density and as such, is unable to sufficiently adapt to changes in the shape of the density function. Both models slightly under-represent the spread of the response variable for both variables. As a result the mode of the predicted values in each plot has a higher density than that of the observed values. However, in both cases the modes of observed and predicted values are in good agreement.

Figure 6: Densities of the predicted (dashed) and observed (solid) values for Peak Distance and Peak Power using SVM.

It is surprising that the SVM performs noticeably more poorly than the linear mixed effects model. There are a number of potential reasons why this may be the case. First, the models utilise the data in subtly different ways. The linear mixed effects model uses the entire data set to develop the model while the SVM uses a subset of the data to develop the model parameters and then tests that model on the actual dataset[1]. Second, while the linear mixed effects model explicitly captures the intra-subject correlations, this was captured by the SVM using binary variables to identify each player. Due to the way SVM learn the model parameters, it is possible that these variables are not included in the final model. More importantly, however, is that an 'off-the-shelf' SVM was used, e.g. RBF kernel and default regression algorithm. These were chosen to explore how current analysts could quickly apply SVMs. With further training and experience, more complex models could be developed. For example, when using the ν-SVM regression algorithm (Schlkopf et al. 2000), the SVM produces similar results to the linear mixed effects model.

5.2 Providing Managerial Insight

We evaluate the output of the model from the perspective of the rugby club to consider it's impact. The CMJ is an important source of information as coaches believed that the results provide information regarding power and explosive strength potential of the players. Studies, such as those by (Gathercole, Sporer & Stellingwerff 2015) and (Gathercole, Stellingwerff & Sporer 2015) have evaluated the links between the outcome of CMJ and other performance parameters such as fatigue. As discussed above, based on resources, time and equipment, and other training priorities, it is not feasible to do the CMJ test more frequently. As a result there is limited knowledge of the power potential of the players pre-game, how they may have 'peaked' leading into a game, or how they may have recovered after a game. It is important therefore to establish whether the physical wellbeing of the player, which can be gathered quickly and non-intrusively, relates to the power they can express through the jump.

To illustrate the output from the model, we consider two separate days on which the jump test and questionnaire were administered to a specific player, a back. The measurements were made in consecutive weeks. The player in question has a random effect (posterior mean) of -0.0678, which will reduce the prediction of peak distance as compared to other players. In the first week, the recorded scores for (UB,LB,SQ,APP,ENERGY,MOOD,STR,MOT) were $(6, 4, 8, 7, 6, 8, 7, 8)$ and in the second week they were $(7, 5, 9, 9, 8, 8, 8, 9)$.

[1]Note that the linear mixed effects model was separately implemented using a training and test dataset with similar results.

We see that, in the second week, this player's condition had improved according to the questionnaire. This is reflected in the (posterior) predictions of peak distance from the model, which were 0.378 and 0.404 respectively, and also the observed peak distances on the jump test, which were 0.382 and 0.408 respectively.

The models developed in this paper support the coaches in multiple ways. First, the models allow coaches to identify players that would perform poorly on the CMJ on any given day. This supports the coaches in tailoring individual training sessions for players prior to game, and with pre-game decision making. Second, the model identifies those measurements taken during the wellbeing test that are well correlated with the CMJ. By identifiying the covariates that are strongest, coaches can better monitor players development and recovery. Finally, the model identifies those players who are 'optimistic' and 'pessimistic' about their condition, i.e. players. This provides coaches with additional information and ensures players cannot 'cheat' the wellbeing test.

In conclusion, this paper demonstrates that simple subjective wellbeing questions, previously viewed as having no predicitve power, can indicate the physical potential of the players and from this, the training process leading into a game can be adjusted accordingly to maximize the rugby performance goals.

6 Discussion and Conclusions

The paper illustrates how different models can be used to gain insight from data currently collected by sports organizations and to solve novel problems. As organizations gather more data, it is imperative that they empower their employees to take advantage of the insight buried within these data. This is particularly important for organizations where resources are limited and where a small competitive advantage may have a large impact on results. As more and more organizations move into the realm of "big data", modelers require different skills to extract insight from these complex, non-linear, noisy data sets. In this paper, we give an overview of two different approaches for analysts working with data containing non-linear relationships and illustrate how each method could be used.

For the coaches at the club, there are a few key aspects of the model that they deem important. First, the model is being built by analysts to better understand the system and not just for the predictive power. As such, the mixed linear effects model is attractive as they are typically building models as communication tools as well as predictive tools. The impact of the different variables can be extracted from the SVM, however, the kernel chosen here transforms these variables into different dimensions that are difficult to attach an operational interpretation to. While there is a desire for the model to be as accurate as possible, there is a need to understand the mechanics of the model in order to tailor feedback. At this time, the SVM is limited in terms of its decision support for this problem. However, the performance analysts were aware of the wide range of problems that they could apply advanced analytical methods to, and

they are keen to gain expertise on modeling approaches that could be applied to large number of datasets. As such, it is likely that SVM can be applied to a wider range of problems than the mixed effects model.

Future work could focus in a number of different areas. First, the data was gathered over a 10-week period in the middle of the season. While investigation did not indicate any time impact, a fuller longitudinal analysis of the data may indicate seasonal effects. Second, using the same problem described above, the modeling could be expanded. For example, the SVM could be further developed to consider different kernels or other machine learning methods such as Neural Networks (Bishop 2006) or Ensemble learners (Valentini & Masulli 2002). For the linear mixed effects model, an important step would be to capture informative-priors, or to try to separate the effect of subjective difference in questionnaire answers from the players' potential for peak distance and peak power.

Within the problem domain, the above models assume that previous measurements have been made for each player. Future work could focus on assessing the jump test score for players who have completed the questionnaire, but have yet to carry out a jump test. In this paper we have assumed sampling error in the jump test equipment to be negligible. There is evidence from the sport science literature that this may not be the case (Larry 1998). In particular, if equipment measurement bias is present in short time frames, and if the bias is correlated, then it would be necessary to include this in any model of jump test scores. Future work could investigate this.

References

Banister, E., Calvert, T., Savage, M. & Bach, T. (1975), 'A system model of training for athletic performance', *Australian Journal of Sports Medicine* **7**(3), 57–61.

Barabino, N., Pallavicini, M., Petrolini, A., Pontil, M. & Verri, A. (1999), Support vector machines vs multi-layer perceptrons in particle identification, *in* 'Proceedings of the European Symposium on Artifical Neural Networks' 99', pp. 257–262.

Bishop, C. (2006), *Pattern recognition and machine learning*, Vol. 4, springer New York.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. & Haussler, D. (2000), 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proceedings of the National Academy of Sciences* **97**(1), 262–267.

Calvert, T., Banister, E., Savage, M. & Bach, T. (1976), 'A systems model of the effects on training on physical performance', *IEEE Transactions on systems, man, and cybernetics* **SMC-6**, 94–102.

Campbell, C., Cristianini, N. & Shawe-Taylor, J. (1999), 'Dynamically adapting kernels in support vector machines', *Advances in neural information processing systems* **11**, 204–210.

Castro-Piero, J., Ortega, F. B., Artero, E. G., Girela-Rejn, M. J., Mora, J., Sjstrm, M. & Ruiz, J. R. (2010), 'Assessing muscular strength in youth: usefulness of standing long jump as a general index of muscular fitness', *The Journal of Strength & Conditioning Research* **24**(7), 1810–1817.

Chapelle, O., Vapnik, V., Bousquet, O. & Mukherjee, S. (2002), 'Choosing multiple parameters for support vector machines', *Machine learning* **46**(1), 131–159.

Congdon, P. (2004), *Applied Bayesian modelling*, Wiley, Chichester.

Cristianini, N. & Shawe-Taylor, J. (2000), *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press.

Drucker, H., Wu, D. & Vapnik, V. N. (1999), 'Support vector machines for spam categorization', *Neural Networks, IEEE Transactions on* **10**(5), 1048–1054.

Gathercole, R. J., Stellingwerff, T. & Sporer, B. C. (2015), 'Effect of acute fatigue and training adaptation on countermovement jump performance in elite snowboard cross athletes', *The Journal of Strength & Conditioning Research* **29**(1), 37–46.

Gathercole, R., Sporer, B. & Stellingwerff, T. (2015), 'Countermovement jump performance with increased training loads in elite female rugby athletes.', *International journal of sports medicine* **36**(9), 722–728.

Guo, G., Li, S. Z. & Chan, K. (2000), Face recognition by support vector machines, *in* 'Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on', IEEE, pp. 196–201.

Hellard, P., Avalos, M., Lacoste, L., Barale, F., Chatard, J. & Millet, G. (2006), 'Assessing the limitations of the banister model in monitoring training', *Journal of Sports Sciences* **24**, 509–520.

James, T. (2012), 'Science in search of gold', *Engineering and Technology* **7**, 44–47.

Joachims, T. (2000), Estimating the generalization performance of a svm efficiently, Technical report, Universität Dortmund.

Larry, D. I. (1998), 'Comparison of the vertec and just jump systems for measuring height of vertical jump by young children', *Perceptual and motor skills* **86**(2), 659–663.

Lee, P. (2012), *Bayesian Statistics: An Introduction*, Wiley.

Marek, S. M., Cramer, J. T., Fincher, A. L., Massey, L. L., Dangelmaier, S. M., Purkayastha, S., Fitz, K. A. & Culbertson, J. Y. (2005), 'Acute effects of static and proprioceptive neuromuscular facilitation stretching on muscle strength and power output', *Journal of Athletic Training* **40**(2), 94.

Maughan, R. (2002), 'The athlete's diet: nutritional goals and dietary strategies', *Proceedings of the Nutrition Society* **61**, 87–96.

Morton, R., Fitz-Clarke, J. & Banister, E. (1990), 'Modeling human performance in running', *Applied Physiology* **69**, 1171–1177.

Moulin, L., Da Silva, A. A., El-Sharkawi, M. & Marks, R. J. (2004), 'Support vector machines for transient stability analysis of large-scale power systems', *Power Systems, IEEE Transactions on* **19**(2), 818–825.

Osuna, E., Freund, R. & Girosit, F. (1997), Training support vector machines: an application to face detection, *in* 'Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on', IEEE, pp. 130–136.

R Development Core Team (2005), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Scarf, P., Shrahili, M., Jobson, S. & Passfield, L. (2013), 'Modelling and optimisation of the sport and exercise training process', *4th International Conference on Mathematics in Sport* .

Schlkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. (2000), 'New support vector algorithms', *Neural computation* **12**(5), 1207–1245.

Scholkopf, ., Simard, P., Vapnik, V. & Smola, A. (1996), Improving the accuracy and speed of support vector machines, *in* 'Advances in Neural Information Processing Systems 9: Proceedings of The 1996 Conference', Vol. 9, MIT Press, p. 375.

Valentini, G. & Masulli, F. (2002), 'Ensembles of learning machines', *Neural Nets* pp. 3–20.

Vapnik, V. (1999), *The nature of statistical learning theory*, springer.

Wu, K. & Wang, S. (2009), 'Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space', *Pattern Recognition* **42**(5), 710–717.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B. & Yu, P. (2008), 'Top 10 algorithms in data mining', *Knowledge and Information Systems* **14**(1), 1–37.

Appendix

Table 2: The posterior means and variances of each of the fixed effect parameters for Peak Distance

Covariate	$E[\beta_i \mid y]$	$Var(\beta_i \mid y)$
UB	-0.0025800	1.901641e-06
LB	0.0022640	1.638400e-06
SQ	-0.0008536	1.550025e-06
APP	0.0003494	3.129361e-06
ENERGY	0.0029600	5.475600e-06
MOOD	-0.0007191	4.318084e-06
STR	-0.0018240	2.975625e-06
MOT	0.0025050	3.783025e-06

Table 3: The posterior means and variances of each of the fixed effect parameters for Peak Power

Covariate	$E[\beta_i \mid y]$	$Var(\beta_i \mid y)$
UB	20.0700	257.9236
LB	-10.6500	224.4004
SQ	16.7600	238.3939
APP	25.4700	389.6676
ENERGY	25.8900	443.1025
MOOD	1.5880	436.3921
STR	3.7590	369.4084
MOT	0.7422	423.1249

Table 4: R^2 for different combinations of C and γ with test set R^2 in brackets

C \ γ	1×10^{-4}	1×10^{-3}	1×10^{-2}	1×10^{-1}	1	1×10^{2}
1	0.0002 (0.0045)	0.0022 (0.0288)	0.1478 (0.1381)	0.6836 (0.5842)	0.5705 (0.1138)	0.4088 (0.004758)
1×10^{2}	0.0026(0.0110)	0.1555 (0.1440)	0.5610 (0.5620)	0.7863 (0.5227)	0.6840 (0.1456)	0.6240 (0.0048)
1×10^{3}	0.1615 (0.1630)	0.5231 (0.5402)	0.6179 (0.6537)	0.7884 (0.5286)	0.6969 (0.1169)	0.6023 (0.01638)
1×10^{4}	0.5484 (0.5321)	0.6064 (0.6061)	0.6491 (0.5688)	0.7521 (0.6067)	0.6963 (0.1327)	0.6158 (0.0095)
1×10^{5}	0.6368 (0.5570)	0.5974 (0.6886)	0.6721 (0.5689)	0.7760 (0.6097)	0.6752 (0.1464)	0.6187 (0.0152)
1×10^{6}	0.6260 (0.5779)	0.6170 (0.6452)	0.6989 (0.5917)	0.7234 (0.5504)	0.7067 (0.1424)	0.6016 (0.0036)