NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Cornforth, David; Campbell, Piers; Nesbitt, Keith; Robinson, Dean; Jelinek, Herbert F
“Prediction of game performance in Australian football using heart rate variability
measures”. Originally published in International Journal of Signal and Imaging Systems
Engineering Vol. 8, Issue 1-2, p. 80-88 (2015)

Available from: <http://dx.doi.org/10.1504/IJSISE.2015.067072>

Accessed from: <http://hdl.handle.net/1959.13/1326009>

Prediction of Game Performance in Australian Football using Heart Rate Variability Measures

D. Cornforth*, P. Campbell and K. Nesbitt

Applied Informatics Research Group
School of DCIT
University of Newcastle
Callaghan 2308, NSW Australia
E-mail: David.Cornforth@newcastle.edu.au
E-mail: Piers.Campbell@newcastle.edu.au
E-mail: Keith.Nesbitt@newcastle.edu.au
*Corresponding author

Dean Robinson

Corio Bay Sports Medicine Centre, Geelong, VIC, Australia

Herbert F. Jelinek

Centre for Research in Complex Systems
School of Community Health
Charles Sturt University
P.O. Box 789, Albury, NSW 2640, Australia
Department of Biomedical Engineering
Khalifa University of Science, Technology and Research Abu Dhabi, UAE
E-mail: HJelinek@csu.edu.au

Abstract: In-match player performance, measured by data from Geographical Positioning System (GPS) devices, was predicted with a correlation coefficient of greater than 0.7. Predictions were based on Heart Rate Variability measures and used advanced regression techniques based on machine learning. These techniques included methods for the selection of variables to be included in the regression study. Results indicate that variable selection using a Wrapper Subset method with a Genetic Algorithm outperformed both Principal Components Analysis and the default method of using all variables. The success of prediction of match performance suggests a potential for new tools to assist the team coach in player selection and management of player training. This work also provides the possibility for a training program to be adjusted specifically to meet the challenges of the size of the playing field and the temperature likely to be encountered on the day of the match.

Keywords: Heart Rate Variability, Medical Data Analysis, Prediction, Regression, Game Performance, Australian Football

Biographical Notes: Dr David Cornforth received the B.Sc. degree in Electrical and Electronic Engineering from Nottingham Trent University, UK, in 1982, and the Ph.D. degree in Computer Science from the University of Nottingham, UK, in 1994. He has been an educator and researcher at Charles Sturt University, the University of New South Wales, and now at the University of Newcastle. He has also been a research scientist at the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Newcastle, Australia. His research interests are in health information systems, pattern recognition, artificial intelligence, multi-agent simulation, and optimization.

Dr Piers Campbell is a Lecturer in the Faculty of Science and also a member of the Applied Informatics Research (AIR) group and IT at the University of Newcastle, Australia. Piers received his B.Sc. (Hons) Computer Science in 2000, his M.Res. in Informatics in 2001 and his Ph.D. in 2005, all from the University of Ulster, UK. Piers' research interests are primarily in the application of data mining and machine learning techniques to real world problems particularly in the energy, production and medical domains.

Dr Keith Nesbitt completed his Bachelor's degree in Mathematics at Newcastle University in 1988 and his Masters in Computing in 1993. Between 1989 and 1999, Keith worked on applied computer research for BHP Research investigating business applications of Virtual Reality and Intelligent Agents. His PhD examined the design of multi-sensory displays and was completed at Sydney University in 2003. Outcomes from this work have received international recognition, and have been favorably reviewed in the prestigious journal of Science. In 2007 he completed a post-doctoral year in Boston, working at The New England Complex Systems Institute, visualizing health related data. He currently works at the University of Newcastle, Australia, and teaches computer programming, game design and game production. His ongoing research focuses on key issues of perception and cognition particularly related to Computer Games and the design of multi-sensory information displays.

Dean Robinson has a Master's degree in Physiotherapy from the University of Sydney. He was the Strength and Conditioning Coach at the New South Wales Institute of Sport between 2001 and 2004, then joined Manly Football Club as High Performance Manager. He joined the Geelong Football Club in 2006, then the Gold Coast Club in 2010. He joined Essendon Football Club before the 2012 pre-season and oversaw the players' strength and conditioning program for that year.

Associate Professor Herbert Jelinek holds the B.Sc. (Hons.) in human genetics from the University of New South Wales, Australia (1984), Graduate Diploma in neuroscience from the Australian National University (1986) and Ph.D. in medicine from the University of Sydney (1996). He is Clinical Associate Professor with the Australian School of Advanced Medicine, Macquarie University, and a member of the Centre for Research in Complex Systems, Charles Sturt University, Australia. Dr Jelinek is currently visiting Associate Professor at Khalifa University of Science, Technology and Research, Abu Dhabi, UAE. He is a member of the IEEE Biomedical Engineering Society and the Australian Diabetes Association.

1. Introduction

This work makes a new contribution to data analytics for Australian Football by allowing prediction of in-game performance, measured by data collected from Geographical Positioning System (GPS) devices during a professional game. Due to the fast paced nature of an AFL game, there would be some expectation of a correlation between player fitness and the overall physical effort made during the game. Predictions were based on weekly ECG recordings collected from football players, playing field dimensions and temperature for the day of the game. A correlation model was able to account for over 70% of the variation in *Player Load*, a measure of effort required and obtained from GPS data collected during matches. This study is the first time these advanced regression methods have been applied to relate pre-match physiological data to match performance. The ability of sophisticated models to predict player performance during matches may assist the team coach in selection for matches based on physiological data, and may also have bearing on the organisation of individual training programs to maximise match performance.

This study uses data collected from elite Australian Football players during training and daily data collection at rest to predict in-game performance recorded for 12 games of a football season. Regression analysis is a common method used for prediction and may be in a parametric or nonparametric form. All the methods used here employ advanced regression algorithms based on machine learning methods. In this paradigm, a model of the form $y = f(\mathbf{x})$ is fitted to the data, where y is a match performance variable, and $\mathbf{x}$ is a vector of measures derived from data collected prior to the game. Regression is a process whereby the parameters of the model are tuned to minimise some error measure. A number of methods were attempted and some promising results provide evidence that such predications are possible. The potential value of regression modelling is the possibility of predicting the likely performance of individual players before each match. This can be of assistance to the coach in choosing players to be fielded for a particular game, or in crafting the pre-match training load.

Performance analysis of athletes has been the focus of research for decades with an emphasis on motor skills and other physical factors (Seashore, 1930). The application of

the information sciences to sport goes back several decades starting with statistical analysis and calculations based on biomechanical data (Baca, 2007). Since then the concept of sports informatics has grown to include modelling and simulation, databases, soft computing, pattern recognition and visualisation. Recently the advent of light weight mobile technological devices such as heart rate monitors, global positioning system units and other sensor devices have led to increased opportunities for the partnering of Information Technology with Medical and Sports Science. A number of recent studies have explored the application of these technologies to a range of sports including rugby and soccer, in order to more accurately understand player performance, conditioning and recovery (McLellan, Lovell and Gass, 2011; Burgess, Naughton and Norton, 2006).

Australian Rules Football (also known as AFL) is played with an oval ball on an oval field with a free flowing pace. Fig. 1 shows a diagram of the field along with the most important features, including the long and short dimension of the field.

Figure 1 Diagram of playing field for Australian Football, showing oval shaped field and main dimensions.

AFL stands in contrast to more structured styles of football, where a game is divided into well-defined phases, marked by a particular formation (O'Shaughnessy, 2006) AFL is a fast moving game where speed and good ball handling and kicking skills determine the outcome. For this reason, it may be possible to link player fitness to the amount of walking, running at speed, during a game, or to the distance covered. The use of AFL as a subject for study requires extra variables to the analysis of fitness data because there is no standard size for the playing field, although the length is generally 135 to 185 metres and the width 110 to 155 metres. In this work the actual dimensions of the field for each game were recorded and used as one of the potential predictors of athlete

performance. This is an important point because athletes have to make a different effort depending on the field that they find themselves playing on at any particular occasion. The prediction of athlete performance at any match is potentially confounded by this variable, so it should be included in regression models.

Player performance has also been reported to change throughout a playing season in the game of soccer (Rampinini, 2007). This variability may be related to training status, perceived wellness and environmental conditions (Gastin, 2012). With respect to the latter, there have been reports of a greater incidence in injuries in football played in warmer and/or drier conditions (Orchard, 2002). Changes in core body temperature may also be a factor in performance but has only been studied under warmer conditions (Duffield, 2009). Our study includes match temperature as an additional variable. There are numerous physiological variables measures used in high performance sport. The actual measures used depend upon the club, and many of these are proprietary and therefore are closely guarded by the club. Heart rate is one such measure that is both commonly used, and known to be associated with fitness and performance.

2. Heart Rate Variability

A standard ECG signal is shown in Fig. 2. This type of signal has been extensively studied and the diagnostic value of the different features is well established. ECG features are referred to using letters. The large spike is referred to as the QRS complex, with R being the peak of the wave or fiducial point. The smaller peak to the left is the P wave and the peak to the right of the QRS is the T wave.

Figure 2 Normal ECG recording showing RR interval.

The signal is degraded by the presence of noise, so that the most reliable feature that can be obtained from low quality recordings (and therefore the most easily obtained measurement) is the interval between successive R peaks, known as the RR interval or heart rate. This can be expressed as beats per minute and varies considerably between individuals, but a typical resting heart rate is 60-80 beats per minute.

The natural rhythm of the human heart is subject to variation that is believed to indicate the health of the cardiovascular system. RR intervals are obtained from the recorded ECG and subjected to further analysis through a variety of algorithms in order to yield variables with good discriminant power (TFESC, 1996). However, the RR

interval can also be obtained from a simple and cheap chest strap device, which makes it particularly suitable as a non-invasive test which allows data to be easily collected from players.

As the electrical system of the heart is a complex and adaptive system, variation between beats is not only common but an indicator of the general health of the cardiovascular system. Heart rate variability (HRV) is an indicator of the regulation of the heart (TFESC, 1996). It is commonly used to assess the health of elite athletes, for example as reported in (Javorka et al., 2003). The response of the cardiac autonomic nervous system can be assessed non-invasively using a number of measures based on HRV (Buchheit et al. 2007), all of which can provide useful information regarding the functional adaptations to a given training stimulus. As heart rate increases due to exercise, HRV tends to decrease. This is known as the sympathetic effect and is manifest as lower frequency changes in heart rate. At the same time, there is a withdrawal of the parasympathetic activity (Javorka et al., 2003). Conversely, during rest the heart rate slows and higher frequency variation predominates. High and low frequency domain parameters are obtained by Fast Fourier Transform from the RR intervals of the biosignal. The extent of variability in the frequency domain can therefore be used as an indicator of cardiac modulation by the autonomic nervous system and an indicator of general health of the athlete or level of stress.

HRV has been conventionally analysed with time- and frequency-domain methods, which measure the overall magnitude of fluctuations around the mean or the magnitude of fluctuations in some predetermined frequencies. More recent nonlinear analysis has shown an increased sensitivity for characterising beat-to-beat fluctuations. These methods use fractal and entropy based measures to extract further information from HRV. For example, Detrended Fluctuation Analysis (DFA) has been shown to be effective in detection of diabetic autonomic neuropathy (Jelinek and Cornforth, 2008) and the Renyi entropy measure has provided good results in the detection of abnormal cardiac rhythms (Jelinek, Tarvainen, and Cornforth, 2012).

Out of many measures that can be derived from HRV, it is not known which ones affect the match performance, or by how much. This work involves analysis methods that are the domain of data mining, which are able to identify important factors that can predict match performance, and to identify variables that have the most influence over performance. The authors were able to identify correlations between player measures prior to game and game performance measures. The questions addressed in this work are:

1. What is the best correlation that can be obtained using any method?
2. Which game performance measures can be predicted with the highest accuracy?
3. Which method of selecting variables provides the highest correlation result?
4. Which of the regression algorithms provide the best accuracy?

3. The Data Set

Participants in the study were professional AFL players, with the following description:

- Number of participants: 44
- Average Age: 20
- Average Body mass: 85.7 kg

Players participated in a minimum of 15 hours per week of training during the twelve-week study. HRV data were collected at the same time each day to avoid any circadian changes in heart rate (Reilly et al., 1984). Data collection was made using the Polar heart rate chest strap connected to an iPhone and sent by email to a server for processing and extraction of HRV measures. Heart rate data was collected from players on waking in the morning. Measures were derived from HRV using time and frequency domain, and non-linear methods. Some of the measures used are described below.

Time domain measures include the minimum, maximum and mean of the RR interval recorded, as well as other measures. The number of pairs of successive intervals that differ by more than 50ms divided by the total number of intervals is a parasympathetic measure known as pNN50

The Poincaré plot is a visual representation of the time series and is constructed by plotting each consecutive RR interval as a point where $y = RR(t)$ and $x = RR(t-1)$. Fig. 3 illustrates this. An ellipse is fitted to the resulting cloud of points and the major and minor axes of the ellipse are estimated as $SD1$ and $SD2$. These were included in the set of independent variables used. $SD1$ and $SD2$ have been shown to be related to the short-term and long-term variability, respectively, but are fundamentally linear measures (Brennan et al., 2001).

Figure 3 Poincaré plot for a sequence of RR intervals allows the estimation of $SD1$ and $SD2$.

The HRV triangular index is based on estimating the density distribution of RR intervals. The maximum value of this curve is the Mode Amplitude (AMo), which indicates the

stability of the heart rhythm. The area under the curve (integral) provides the number of all intervals and is divided by the maximum value of the distribution (TFESC, 1996). The triangular interpolation of the interval histogram (TINN) is the estimated width of the density distribution. This is also known as the Index of Control Systems Pressure and is believed to be sensitive to physical and emotional load, or to intensifying of the sympathetic nervous system tone.

Frequency domain analysis of heart rate was introduced by Akselrod (1981) and these methods divide the spectral distribution into regions labelled as very low, low or high frequency (Fig. 4). Low frequency power (LF) is believed to be indicative of sympathetic activity or sympathovagal balance. High frequency (HF) is indicative of parasympathetic activity. Very Low Frequency (VLF) amplitude is closely connected with psycho-emotional state and functional condition of the brain (Baevisky and Berseneva, 2008; Haspekowa, 1996). The fraction of the power spectrum falling in each of these three ranges was included in the set of independent variables used.

Figure 4 Power spectrum of RR intervals showing VLF, LF and HF regions.

Fleishman et al. (1999) have shown the important meaning of VLF range analysis, and that the capacity of VLF fluctuations of HRV is a sensitive indicator of management of metabolic processes and reflects deficit energy states. VLF is also used as a reliable marker of blood circulation and provides information on the degree to which the autonomous (segmental) level and supra-segmental regulation levels are functionally synchronised via neuronal output. To be normal, VLF is between 15 and 30% of the total spectrum power and can be measured from the RR intervals associated with the ECG, which are modulated by supra-segmental and segmental neural modulatory systems including the cardiac autonomic nervous system innervation.

The peak value of the frequency distribution, the frequency at maximum power, is expressed as Hz while the power is expressed as ms^2 . The logarithm of HF and LF is also calculated, as it allows the definition of a new index based on HF and LF energy which appears relevant to measure HR regulation and corrects for data not normally distributed (Khalfa, 2012). The ratio of low to high frequency components is also calculated for our analysis as well as the total power (TFESC, 1996). Other measures used in this study are proprietary to the club, but are derived from commonly available ECG data.

In addition, the Club collects detailed quantitative data on player performance in terms of number of passes and receipts of the ball during every game. Additional physiological performance data were collected from each match, using GPS sensors worn by players during the game. From these data, 11 measures were used to assess match performance, including two measures of time, six measures of distance covered at various speeds, as well as a total player load and maximum speed. These are listed in Table 1 and are the dependent variables of the study, which will be predicted. The distance covered during the match and the total distance covered at the event were recorded. Total distance includes data before the match and continues during breaks.

Additional measures were obtained from knowledge of the game location for the twelve games included in this study:

- Long Dimension of the playing field
- Short Dimension of the playing field
- Minimum Temperature (24Hr)
- Maximum Temperature (24Hr)
- Average Temperature (24Hr)

These are added to the other independent variables, to be used as predictors of the performance measures provided in Table 1.

Table 1 Measures derived from GPS data collected during matches. These are the dependent variables of the regression analysis.

Abbreviation	Description
B4 Time	Time that the player spent running at more than 24 km/h
B3 Time	Time that the player spent running at a speed of 18 to 24 km/h
Total Dist.	Total distance moved by the player during the entire event
Player Load	A derived measure that is proprietary to the club
Max Speed	The maximum speed reached by the player during the game
Walk	The total distance covered by the player at a speed of 3 to 7 km/h
Jog	The total distance covered by the player at a speed of 7 to 12 km/h
Cruise	The distance covered by the player at a speed of 15 to 28 km/h
B4 Dist	The distance covered by the player at a speed of more than 24 km/h
B3 Dist	The distance covered by the player at a speed of 18 to 24 km/h
Match Dist	Total distance travelled by the player during the game

4. Regression Analysis

Advanced regression techniques include partial least squares regression (PLSR), which allows previously unidentified relationships between the independent and dependent

variables to be identified as well as derivations of support vector machine and neural network based regression models (SVM) regression analysis (Stone and Brookes, 1990; Mukherjee et al., 1997). This work used the data mining toolkit known as Weka, as it is free, open source, and widely accepted in the machine learning community (Witten and Frank, 2005). A selection of regression algorithms is available on Weka and these were used to discover relationships between the measures extracted from ECG data and player performance as measured by in-game quantitative data. This is the first study which uses these techniques to relate ECG data to match performance. The following seven Weka algorithms were used for prediction of continuous variables:

- *Gaussian Processes* is a generalisation of Gaussian functions to model possible functions relating dependent to independent variables (Williams, 2002).
- *Linear Regression* is an implementation of linear regression which uses the Akaike criterion for model selection, in order to minimise over fitting caused by the use of too many parameters (Akaike, 1974).
- *LeastMedSq* is an implementation of linear regression using least median squared as the error measure (Witten and Frank, 2005).
- The *Multilayer Perceptron* is an implementation of an artificial neural network trained using the back-propagation algorithm (Rumelhart and McClelland, 1986).
- *PLS Classifier* is a wrapper for the Partial Least Squares Regression filter that attempts to decompose the independent variables in a manner similar to PCA (Wold et al., 2001).
- *RBF Network* uses a k-means clustering algorithm to determine a number of centres for basis functions. At each centre it fits a multivariate Gaussian function and uses linear regression. The resulting basis functions are combined into a single model (Broomhead and Lowe, 1988).
- *SMOreg* is another regression method using kernel functions, but based on the Support Vector Machine. The algorithm is known as Sequential Minimal Optimisation, and breaks the task of training the SVM into a series of smaller problems to improve the training time (Platt, 1999).

In order to reduce bias caused by over training or over fitting of the model to the data, the predictive ability of the model must be tested on independent data not seen by the algorithm during training. This requires splitting the available data into two sets, the training set and the holdout set. The regression model is built using the first dataset then evaluated by its performance on the holdout set. Obviously the holdout set must be chosen carefully to prevent a source of bias. The most popular way of testing involves the cross validation method where the dataset is divided into a number of subsets (Efron, 1983). A number of tests are performed where each subset in turn becomes the holdout set. In this way the algorithm is tested against all available data. The algorithms used in this work use ten-fold cross validation, where one

tenth of the data is used as the holdout set, and ten models are built and tried, each using a different one tenth to test against.

5. Methods

Data on player training and daily physiological data collection at rest as well as in game performance measures were obtained as de-identified data from the Club. Prior to this research the study was approved by The Australian Institute of Sport Ethics and Research Committee and informed consent obtained from all players participating in this research. Data analysis was performed using freely available software such as Weka (distributed by the University of Waikato) and bespoke software prepared under the supervision of the authors.

Correlation measures were calculated based on the collected GPS match data and the two sources of data were integrated by matching the dates of collection of ECG data with the dates of matches played. Temperature data for locations and dates of matches was sourced from the Australian Bureau of Meteorology and added to create a unified data set. The data were thoroughly checked for consistency and records with inconsistent values were discarded. The data were converted to the “arff” format suitable for analysis using Weka.

In the first attempt to predict dependent variables, the entire dataset (all independent variables) was used as input to the automated regression methods. For each dependent variable, a dataset was prepared with one dependent variable and all the independent variables. The seven regression algorithms were run on each of these datasets.

In the second attempt to predict the dependent variables, PCA was applied to the set of independent variables only. The outcome was a transformed data set consisting of weighted combinations of the independent variables. For each dependent variable, a dataset was prepared with one dependent variable and all the transformed new variables. The seven regression algorithms were run on each of these datasets.

In the third attempt to predict the dependent variables, the dataset was reduced by selecting variables using Wrapper subset analysis that maximises performance on each regression algorithm, using a Genetic Algorithm (GA) to search through the list of independent variables. This resulted in a large number of datasets, where each dataset included one dependent variable and a selection of independent variables. All of these new datasets were run on all seven of the regression algorithms.

In order to answer question 3 above, the three approaches to regression were compared, by selecting all results achieving a correlation of greater than 0.6. The mean and 95% confidence intervals (CI) were calculated for each method.

6. Results

The result of the automated regression using all variables is shown in Table 2. Each row represents the use of a different regression algorithm. Each column represents the attempt to fit a regression model to the variable given at the top of the column. Figures in bold indicate the best result for that

variable (column). In order to predict the variable *B4 Time*, the best result was obtained from the *PLS classifier*, with a correlation of 0.3463 achieved using cross validation. Some of the other variables were predicted with a higher correlation. *Player Load* was predicted with a correlation of 0.6394 by the *PLS classifier*. The best result obtained was the prediction of *Match Distance* with a correlation of 0.6612 by the *SMOreg* algorithm.

PCA produced 19 new variables that were linear combinations of the HRV measures. The results of the automated regression using PCA are shown in Table 3. The best result obtained was a correlation of 0.6676 for predicting *Match Distance* using *SMOreg*, the same as in the previous case, with approximately 1% improvement. Overall it seems that PCA enables the choice of variables that can enhance the results of the regression analysis. However, not all individual results are substantially improved.

The results from the regression using variables selected by Wrapper subset with GA are shown in Table 4. The variables with correlation coefficient greater than 0.7 are *Player Load*, *Cruise*, *Jog*, *Walk* and *Match Distance*. These results are much improved over those using all variables, and over those using transformed variables from PCA. *Player Load* was predicted with a correlation of 0.7166 by the *Gaussian Process* algorithm. *Cruise* was predicted with a correlation of 0.7346 by the *SMOreg* algorithm (Table 4). *Jog* was predicted with a correlation of 0.7456 by the *SMOreg* algorithm. *Walk* was predicted with a correlation of 0.7565 by the *SMOreg*. *Match Distance* was predicted with a correlation of 0.7253 by the *Gaussian Process* algorithm. Overall the best performing algorithm was the *SMOreg* method.

Figure 5 compares the three methods used to find the best correlation of variables. Data are included for the best 5 dependent variables (*Load*, *Cruise*, *Jog*, *Walk*, *Match Distance*).

The first method, using all variables, had a mean correlation coefficient of 0.49 with lower and upper 95% CI of 0.44 and 0.54. The second method, using PCA to create new variables, had a mean correlation coefficient of 0.53 with lower and upper 95% CI of 0.49 and 0.57. This appears to be an improvement over using all variables, but was not significant

at the 95% CI level. The third method, using a Wrapper algorithm with GA to select variables, had a mean correlation coefficient of 0.60 with lower and upper 95% CI of 0.58 and 0.62. This was higher than the others and the difference was significant at the 95% CI level. It should be emphasised that these comparisons relate to the mean correlation coefficient and not to be the best results, as these have been discussed above. This figure merely provides evidence that the Wrapper subset approach, coupled with a GA search, provides significantly better results than either models using all variables or variables produced using PCA.

Conclusion

A regression method based on the *Support Vector Machine* performed best, with a correlation of 0.76 between a selection of pre-match features based on Heart Rate Variability, and the distance walked during matches. Correlation measures obtained also suggested a link between other match performance measures and data collected before the match. This suggests a potential for the development of advanced data analytics that could provide new tools for a team coach in terms of player selection and design of the training program.

In previous work, we investigated playing experience, player age and aerobic / anaerobic fitness on player performance, using multilevel modelling. Player experience and aerobic fitness had positive predictive value, whilst age and anaerobic fitness was the opposite (Gastin et al., 2012). The outcome of that work was the realisation that prediction of player performance might be feasible. This current work poses four questions in relation to the prediction of in-game performance based on data collected over the period of training for each game, in particular ECG data, and extends previous work by including playing field dimensions and match day temperature.

The first question concerns the best correlation that can be found between pre-match training data and match outcomes. The highest correlation found using any method was 0.7565 for predicting *Walk* using the *SMOreg* regression algorithm. However, other match performance measures were predicted with similar, but slightly smaller, correlation figures.

The second question relates to the game performance that can be predicted. The game performance measures that can be predicted with the highest accuracy are (in descending order) *Walk* (0.7565), *Jog* (0.7471), *Cruise* (0.7346), *Match Distance* (0.7106) and *Load* (0.7114). The results indicate that player behaviour during the match can be deduced from information available before the match, including analysis of ECG data. But it also suggests that the fitness of an individual player is a good predictor of player performance on the match day.

The third question relates to the method used to select variables to include in the regression models. Three methods were attempted to select variables in order to maximise the correlation coefficient. The first method attempted to build regression models using all variables, and to leave the regression model to weight variables so as to effectively ignore those not contributing to an accurate model. The second method used PCA to generate 19 new variables, and

results using this method were on average 9% better than simply using all the variables without transformation. However when compared using the five most successful variables, this difference was not significant at the 95% CI level. The third method used a more sophisticated approach. A Wrapper subset selection using a GA search produced an average 23% improved performance over using all variables, and this was significant at the 95% CI level. This suggests that although PCA can produce better results, the Wrapper method can greatly improve the results of correlation in this type of study.

The fourth question relates to the best performing regression algorithm for this problem. Out of seven regression algorithms studied, the *SMOreg* algorithm provided the best accuracy in terms of correlation coefficient after cross validation.

This work has taken a new approach to attempt the prediction of match performance data using pre-match HRV information. Sophisticated regression techniques based on Machine Learning have shown that regression results of more than 0.7 can be achieved. This has the potential to assist the coach in player selection and training options, and could result in a competitive advantage over the opposing team. Once the coach is able to predict the match performance, the training load can be adjusted to take into account ground size and temperature on the match day. In this way the coach can adapt training to the likely match conditions.

References

- Akaike, H. (1974) 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* Vol.19, No. 6, pp.716-723.
- Akselrod S., Gordon D., Ubel F.A., Shannon D.C., Barger A.C. and Cohen R.J. (1981) 'Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat to beat cardiovascular control', *Science* Vol. 213, No. 4504, pp.220-222.
- Baca, A. (2007) 'Computer Science in Sport – History, research areas and fields of application', *IACSS Symposium*, pp. 16-22.
- Baevsky, R.M. and Berseneva, A.P. (2008) *Use of Kardivar System for Determination of the Stress Level and Estimation of the Body Adaptability*, Technical Report, Kardia.Tel, [Online], Available: www.ehrlich.tv/Kardivar_Methodical_Eng.pdf, accessed Dec 2012.
- Brennan, M., Palaniswami, M. and Kamen, P. (2001) 'Do Existing Measures of Poincaré Plot Geometry Reflect Nonlinear Features of Heart Rate Variability?', *IEEE Transactions On Biomedical Engineering*, Vol. 48, No. 11, pp.1342-1347.
- Broomhead, D. S. and Lowe, D. (1988) *Radial basis functions, multi-variable functional interpolation and adaptive networks*, Technical report 4148, Royal Signals and Radar Establishment.
- M. Buchheit, Y. Papelier, P.B. Laursen, S. Ahmaidi, Noninvasive assessment of cardiac parasympathetic function: post-exercise heart rate recovery or heart rate variability?, *American Journal of Physiology: Heart & Circulation Physiology* 2293 (2007) H8-H10.
- Burgess, D.J., Naughton, G., and Norton, K.I., (2006), 'Profile of movement demands of national football players in Australia', *Journal Sci Med Sport*, Vol. 9, No. 4, pp. 334–341.
- Duffield, R., Coutts, A.J. and Quinn, J. (2009) 'Core temperature responses and match running performance during intermittent-sprint exercise competition in warm conditions', *Journal of Strength and Conditioning Research*, Vol. 23, pp. 1238-1244.
- Efron, B. (1983) 'Estimating the error rate of a prediction rule: improvement on cross-validation', *Journal of the American Statistical Association*, Vol. 78, No. 382, 316-330.
- Fleishman, A.N. (1999) *Slow hemodynamic oscillations. The theory, practical application in clinical medicine and prevention*, Novosibirsk: "Nauka", Siberian enterprises RAS, 1999.
- Gastin, P.B., Fahrner, B., Meyer, D., Robinson, D., and Cook, J., (2012) 'Influence of physical fitness, age, playing experience and weekly training load on match performance in elite Australian football', *Journal of Strength & Conditioning Research*, Vol. 27, No. 5, pp.1272-1279, Lippincott Williams & Wilkins, Philadelphia, Pa.
- Javorka, M., Zila, I., Balhárek, T., and Javorka, K. (2003) 'On- and off-responses of heart rate to exercise - relations to heart rate variability', *Clin Physiol Funct Imaging*, Vol. 23, No. 1, pp.1-8.
- Jelinek, H.F. and Cornforth, D.J. (2008) 'Computational methods for the early detection of diabetes', in Nilmini Wickramasinghe and Elie Geisler (eds.), *Encyclopaedia Of Healthcare Information Systems*, IDEA books.
- Jelinek, H.F., Tarvainen, M. and Cornforth, D.J. (2012) 'Renyi entropy in identification of cardiac autonomic neuropathy in diabetes', *Conference on Computing in Cardiology*, Krakow, pp.909-912.
- Khalifa, N., Bertrand, P.R., Boudet, G., Chamoux, A., and Billat, V. (2012) 'Heart Rate Regulation Processed Through Wavelet Analysis and Change Detection: Some Case Studies', *Acta Biotheoretica*, Vol. 60, No. 1-2, pp.109-129.
- McLellan, C., Lovell, D., and Gass, G. (2011) 'Performance Analysis of Elite Rugby League Match Play Using Global Positioning Systems', *Journal of Strength and Conditioning Research*, Vol. 25, No. 6, pp.1703-1710, Lippincott Williams & Wilkins; ISSN: 1064-8011.
- Mukherjee, S., Osuna, E., and Girosi, F., (1997) 'Nonlinear prediction of chaotic time series using support vector machines', *Neural Networks for Signal Processing, IEEE Workshop*, pp.511-520.
- Orchard, J. (2002) 'Is there a relationship between ground and climatic conditions and injuries in Football?' *Sports Medicine*, Vol. 32, pp. 419-432.
- O'Shaughnessy, D.M. (2006) 'Possession Versus Position: Strategic Evaluation in AFL', *Journal of Sports Science and Medicine*, Vol. 5, pp. 533-540.
- Platt, J.C. (1999) 'Fast Training of Support Vector Machines Using Sequential Minimal Optimization', in B. Schölkopf, C Burges, and A Smola (eds) *Advances in Kernel Methods – Support Vector Learning*, MIT press pp. 185-208.

- Rampinini, E., Coutts, A.J., Castagna, C., Sassi, A., and Impellizzeri, F. (2007) 'Variation in top level soccer match performance', *International Journal of Sports Medicine*, Vol. 28, pp.1018–1024.
- Reilly, T., Robinson, G., and Minors, O.S. (1984) 'Some circulatory responses to exercise at different times of day', *Medicine and Science in Sports and Exercise*, Vol. 16, No. 5, pp. 477–482.
- Rumelhart, D.E., McClelland, J.L., and the PDP Research Group (1986) *Parallel distributed processing. Exploration in the Microstructure of Cognition*, vol 1-2. MIT Press, Cambridge.
- Seashore, R.H. (1930) 'Individual Differences in Motor Skills', *Journal of General Psychology*, Vol. 3, No. 1, pp. 38-66.
- Stone, M. and Brooks, R.J. (1990) 'Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression', *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 52, No. 2, pp. 237-269.
- Task Force of the European Society of Cardiology (TFESC), North American Society of Pacing Electrophysiology (1996) 'Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use', *Circulation* Vol. 93, pp. 1043-1065.
- Williams, C.K.I. (2002) 'Gaussian Processes', In Arbib, M.A. (ed.) *Handbook of Brain Theory and Neural Networks*, pp. 466-470, MIT Press, Second Edition.
- Witten, I.H., and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco.
- Wold, S., Sjöström, M. and Eriksson, L. (2001) 'PLS-regression: a Basic Tool of Chemometrics', *Chemometrics and Intelligent Laboratory Systems*, Vol. 58, No. 2, pp. 109-130.

Table 2 Correlation coefficients from automated regression using all variables. Figures in bold are the highest score for each variable predicted.

Algorithm	B4 Time	B3 Time	Total Dist	Load	Speed	Cruise	Jog	Walk	B4 Dist	B3 Dist	Match Dist
Gaussian Processes	0.2806	0.206	0.3899	0.6036	-0.0991	0.5387	0.528	0.5575	0.3012	0.5722	0.5959
LeastMedSq	0.2911	0.1481	0.375	0.4389	0.0762	0.2182	0.4903	0.1347	0.1061	0.2876	0.3999
Linear Regression	0.3099	0.1336	0.4463	0.6114	0.0762	0.5091	0.6233	0.5257	0.1469	0.346	0.6195
Multilayer Perceptron	0.1898	0.2364	0.1396	0.3288	0.083	0.4368	0.48	0.3285	0.2162	0.3149	0.3982
PLSClassifier	0.3463	0.2667	0.4822	0.6394	0.1613	0.5849	0.6154	0.5551	0.3125	0.3653	0.6256
RBFNetwork	-0.0253	0.0672	0.074	0.4544	-0.1463	0.3478	0.3501	0.2641	-0.0371	-0.0995	0.1646
SMOreg	0.3307	0.2417	0.4809	0.6131	0.1293	0.5414	0.6374	0.6612	0.2184	0.3485	0.6296

Table 3 Correlation coefficients from automated regression using transformed variables from PCA. Figures in bold are the highest score for each variable predicted.

Algorithm	B4 Time	B3 Time	Total Dist	Load	Speed	Cruise	Jog	Walk	B4 Dist	B3 Dist	Match Dist
Gaussian Processes	0.3203	0.2912	0.4807	0.6358	0.0571	0.6157	0.614	0.5943	0.2499	0.5745	0.6249
LeastMedSq	0.1501	0.297	0.4897	0.4621	0.2053	0.5977	0.6091	0.5437	0.1167	0.4529	0.5462
Linear Regression	0.347	0.2708	0.451	0.5459	-0.1299	0.6108	0.5646	0.5151	0.1682	0.5616	0.5922
Multilayer Perceptron	0.079	0.0994	0.2676	0.3078	-0.0406	0.4248	0.4238	0.3651	0.1507	0.4726	0.2676
PLSClassifier	0.3586	0.2761	0.4826	0.5896	0.1791	0.6216	0.6487	0.6499	0.2755	0.5231	0.6225
RBFNetwork	0.0043	0.0806	0.0626	0.2559	-0.1656	0.498	0.494	0.4451	-0.2009	0.052	0.1925
SMOreg	0.344	0.3125	0.5408	0.6019	0.1261	0.6289	0.6608	0.6676	0.2047	0.5268	0.576

Table 4 Correlation coefficients from automated regression using variables selected using the wrapper subset with GA. Only the best result is shown for each variable and regression algorithm. Figures in bold are the highest score for each variable predicted.

Algorithm	B4 Time	B3 Time	Total Dist	Load	Speed	Cruise	Jog	Walk	B4 Dist	B3 Dist	Match Dist
Gaussian Processes	0.3853	0.387	0.4839	0.7166	0.3475	0.6851	0.7054	0.6937	0.452	0.6229	0.7253
LeastMedSq	0.3947	0.414	0.5545	0.6949	0.2887	0.6675	0.7031	0.702	0.3575	0.5543	0.6553
Linear Regression	0.4265	0.3954	0.5369	0.7088	0.2996	0.7344	0.7471	0.7041	0.3263	0.6095	0.7143
Multilayer Perceptron	0.1751	0.2475	0.369	0.5604	0.3078	0.5468	0.5153	0.6355	0.2372	0.4308	0.6224
PLSClassifier	0.3813	0.3756	0.5549	0.6743	0.3545	0.6972	0.7209	0.6694	0.369	0.5975	0.7119
RBFNetwork	0.3139	0.2612	0.406	0.655	0.0784	0.6347	0.7002	0.7	0.2543	0.618	0.6761
SMOreg	0.4225	0.412	0.5596	0.7114	0.3434	0.7346	0.7456	0.7565	0.3768	0.6271	0.713