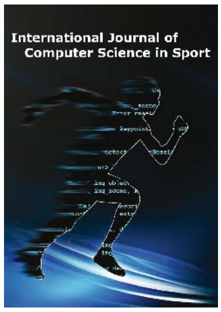
International Journal of Computer Science in Sport

Volume 18, Issue 1, 2019

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2019-0005



Modelling Match Outcome in Australian Football: Improved accuracy with large databases

Young, C.[1], Luo, W.[2], Gastin, P.[4], Tran, J.[1,3], Dwyer, D.[1]*

[1]Deakin University, Centre for Sport Research, Geelong, Australia,

[2]Deakin University, School of Information Technology, Geelong, Australia,

[3] Geelong Cats Football Club, Geelong, Australia

[4]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University, Melbourne, Australia

Abstract

Mathematical models that explain match outcome, based on the value of technical performance indicators (PIs), can be used to identify the most important aspects of technical performance in team field-sports. The purpose of this study was to evaluate several methodological opportunities, to enhance the accuracy of this type of modelling. Specifically, we evaluated the potential benefits of 1) modelling match outcome using an increased number of seasons and PIs compared with previous reports, 2) how to identify eras where technical performance characteristics were stable and 3) the application of a novel feature selection method. Ninety-one PIs across sixteen Australian Football (AF) League seasons were analysed. Change-point and Segmented Regression analyses were used to identify eras and they produced similar but non-identical outcomes. A feature selection ensemble method identified the most valuable 45 PIs for modelling. The use of a larger number of seasons for model development lead to improvement in the classification accuracy of the models, compared with previous studies (88.8 vs 78.9%). This study demonstrates the potential benefits of large databases when creating models of match outcome and the pitfalls of determining whether there are eras in a longitudinal database.

KEYWORDS: PERFORMANCE ANALYSIS, FEATURE SELECTION, COACHING, TEAM SPORTS, DATA MINING

Introduction

Performance analysis involves the investigation of sports performance during training or competition and it includes three major streams; the technical, tactical and physical assessments of performance (O'Donoghue, 2009). There are also other areas of interest in performance analysis that include biomechanics and psychology (O'Donoghue, 2009). An assessment of technical performance can be accomplished through an analysis of technical performance indicators (PIs) which can be acquired via notational analysis. The analysis of PIs has been conducted with traditional statistical approaches (e.g. multiple regression) and also mathematical modelling. The literature includes examples of these types of analyses applied to a variety of sports such as basketball (Gómez, Lorenzo, Barakat, Ortega, & José M, 2008) rugby and ruby sevens (Higham, Hopkins, Pyne, & Anson, 2014; Jones, Mellalieu, & James, 2004; Vaz, Van Rooyen, & Sampaio, 2010) and soccer (Association football) (Castellano, Casamichana, & Lago, 2012; Lago-Peñas, Lago-Ballesteros, & Rey, 2011).

Models that represent the relationships between the value of technical performance indicators (PIs) and the outcome of matches, can reveal which PIs are most important. The information contained in models of performance can be used to influence training and coaching strategies, although the certainty of this information is influenced by the accuracy of the model/s. Therefore, it is important to strive for optimal model accuracy which may be achieved in several ways.

Increasing the amount of data that is used to train a model may improve its accuracy, although this improvement usually plateaus at a certain point (Hall, Witten, & Frank, 2011). The inclusion of different variables (i.e. PIs) within a model has also been shown to improve model accuracy (O’Donoghue, Ball, Eustace, McFarlan, & Nisotaki, 2016). Australian (Rules) Football (AF) is a field-based, team invasion sport played between two teams of 18 players and it provides an opportunity for modelling much like other football codes (see Gray & Jenkins (2010) for a thorough description of AF). Despite this, there are only a few published reports that involve models of performance. The amount of data used for modelling in these reports amounts to 198 games (Robertson Gupta et al 2016), 396 games (Robertson Back et al 2016) and 990 games (Stewart et al 2007). Similar amounts of data have been used for related modelling in soccer such as 64 games (Liu, Gomez, Lago-Peñas, & Sampaio, 2015), 96 games (Moura, Martins, & Cunha, 2014) 380 games (Lago-Peñas, et al., 2011) and 1900 games (Gómez, Gómez-Lopez, Lago, & Sampaio, 2012). The models generated by the studies that used relatively small amounts of data, may have limited long term stability. Therefore, when there is more data available for modelling, the opportunity exists to determine whether the additional data improves the performance of models, which may also improve their value to coaches and athletes.

A larger longitudinal database of technical performance indicator data may also include temporal changes in dominant styles of play or the rules of the game. Models may be more accurate when relationships are stable for a prolonged period of time (i.e. several seasons). This is analogous to the concept of stationarity which is an issue that has been examined in time series forecasting (Levendis, 2018). Therefore, there may be a need to determine whether eras exist in a large longitudinal database prior to modelling. Recent work in AF used multivariate analysis to identify periods of stability (i.e. eras) within the 2001–2015 AFL (Australian Football League) seasons (Woods, Robertson, & Collier, 2017). An ordination plot using non-metric multidimensional scaling of a distance matrix calculated from 18 team PIs identified three eras in a 20-year period. Another study in soccer (Jacklin, 2005) explored the temporal changes in home advantage over 57 seasons and found three eras between 1946–2003.

The eras identified in the previous AF study (Woods, Robertson, & Collier, 2017) were relatively short (i.e. 3-4 years) which may limit the opportunity (i.e. small sample size) to create highly accurate models in each era. Consequently, there is a need to establish a balance between utilising the benefits of a larger databases, if the benefits exist, against the need to detect instability in the data.

Previous work in AF has modelled match outcome using 20 PIs (Stewart, Mitchell, & Stavros, 2007) and 17 PIs (Robertson, Back, & Bartlett, 2016). Modelling work in soccer has used 16 (Castellano, et al., 2012), 18 (Lago-Peñas, et al., 2011) and 19 PIs respectively (Fernandez-Navarro, Fradua, Zubillaga, Ford, & McRobert, 2016). The number and type of PIs that are recorded in AF has increased in recent years and this increase may have also occurred in other major sports. A major sports statistics company based in the UK collects nearly 200 PIs in professional soccer matches (personal communication with Opta Sport). This increase in the variety of PIs that can be included in models may also provide an opportunity to improve the accuracy of models. However, with the increased number of PIs and additional seasons of data, it is important to identify which of the PIs are valuable for modelling and maintain parsimony and interpretability of models. This can be achieved using effective feature selection methods which are able to quantify the relative contribution of a PI to model accuracy.

It is likely that in the future, analysts will continue to use models to identify aspects of individual and team performance that are associated with success. As the amount and variety of data available to analysts increases over time, there is a need to understand both the opportunities and challenges associated with the analysis of large longitudinal databases. Therefore, the present study sought to evaluate the benefits of 1) increasing the number of seasons and PIs used for modelling, 2) the identification of eras where technical performance characteristics were stable and 3) the application of a novel feature selection method. The context for this study was AF, as we had access to a database that was relatively large in terms of the number of seasons and the number of PIs. The intent of this study was to provide information that is useful for those who model team performance.

Methods

Data

Ethics approval was granted by the institutional research ethics committee prior to the commencement of the study. A set of team PIs from 2001 through to 2016 AFL home-and-away seasons were provided via licence from the official supplier of match statistics to the AFL (Champion Data, Southbank, Australia). This consisted of 54 match aggregate raw team PIs from 3145 AFL matches in total. The reliability and validity of the data collected by Champion Data has been evaluated previously and deemed to be acceptable (O'Shaughnessy, 2006; Robertson, Gupta, & McIntosh, 2016). The authors derived a number of new PIs from the existing database, which provides additional information as these PIs are not just counts of an event. They include PIs that were divided by the value of another PI (Kick:Handball ratio, Contested:Uncontested Possession ratio, Ratio of Turnovers to Turnovers Forced) or the sum of two different PIs, resulting in a final database of 103 PI for analysis. The relative form of each PI was calculated as the difference in the value of that PI for each match. The relative form of PIs have been previously reported as valuable when modelling match outcome (Robertson, Back, et al., 2016). This is an example of a descriptive conversion which is proposed to best describe the nature of a sport (Ofoghi, Zeleznikow, MacMahon, & Raab, 2013). The PIs were used as independent variables, and match outcome (i.e. win-loss and score margin) were used as the dependent variables.

Drawn matches (n=25) were removed from the analysis. Any score-related variable was deemed not to provide insight on how a team needs to perform to have successful match outcomes. Therefore, Goals, Behinds, Goal Assists, Score Assists and Goal Conversion were removed from the analysis as these were considered to be mathematically related to the dependent variables.

Statistical Analysis

An iterative, explorative approach to data analysis was used to determine the most appropriate methods for each part of the analysis. The analysis process consists of four phases: data preparation, era detection, feature selection and modelling (Figure 1).

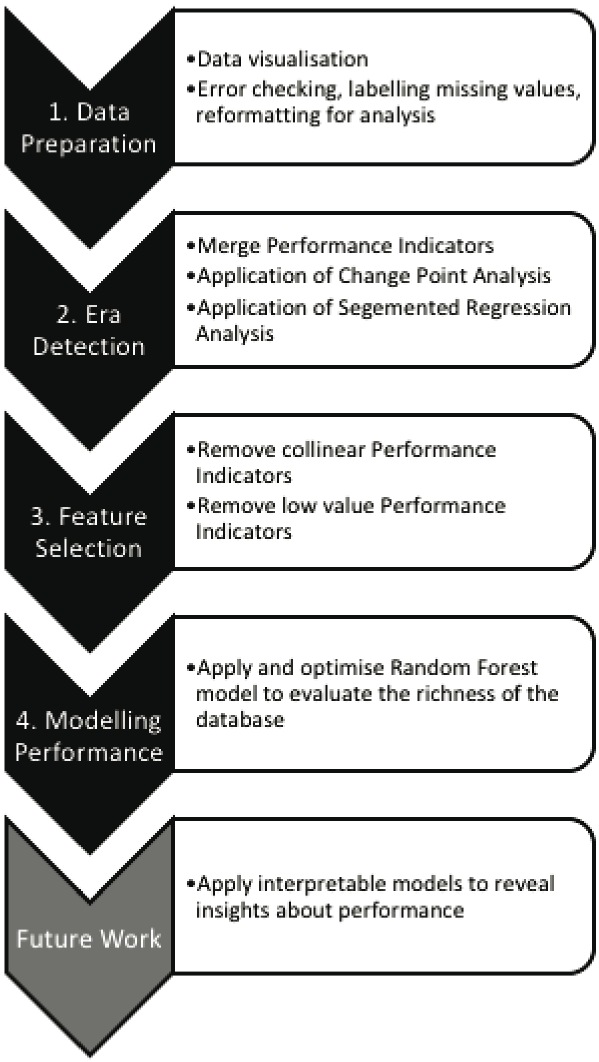


Figure 1. The data analysis process. This diagram represents the order of the processes and analysis used.

Prior to the first phase the data was explored using visualisation software (Tableau 10.0.7, Tableau software company, Washington, USA) to gain an understanding of the nature of the dataset prior to the identification of eras within the 16-year database. A check for collinearity was undertaken prior to any analyses. A correlation matrix was generated displaying the *Pearson* (r) between all 103 PIs. If the correlation value was equal to or greater than 0.95, the attribute with a weaker correlation to score margin was removed, this is supported by recent work (O'Brien, 2017).

This resulted in a total of 91 PIs in the database. Repeated measures analysis of variance (RMANOVA) of 38 PIs available in their absolute form indicated that most seasons were significantly different ($p<0.001$) from each other the 16 years. The RMANOVA results for each PI were not uniform, but it appeared there were some general trends in the dataset. However, due to these inconsistencies and the complexity of the interpretation of such a large number of results, this method was not used. To simplify the era detection process, the value of each PI for each team in each match was converted to a z-score, that we refer to as a universal PI. The z-scores were calculated from the average and standard deviation of each PI for all years combined. The mean universal PI value for all PIs was calculated for each year. Therefore, the value of the mean universal PI for a year represented its relative difference to other years. Change-point analysis (Taylor, 2000) and Segmented Regression analysis (Muggeo, 2003) were subsequently implemented in an attempt to identify distinct changes in the value of the universal PIs across the time span of 2001 to 2016.

The next phase was feature selection, which required a different approach for each outcome measure, as one was binomial (win-loss) and the other was continuous (score margin). Feature selection stability has been suggested to improve through the application of an iterative ensemble approach (Abeel, Helleputte, Van de Peer, Dupont, & Saeys, 2009; Yang, 2013). Therefore, feature selection for the outcome measure of win-loss, was a combination four different techniques; Information Gain, Information Gain Ratio, Gini Index and Correlation. For each PI, the weighting value from each of the four techniques were normalised to a range of 0 to 1, which allowed us to add them together. This calculation of the weight of all PIs was repeated 100 times, using bootstrap sampling of the training data. This process generated 100 weights for each PI, allowing us to determine the mean and variance of the weight of each PI. The PIs were then ranked based on their mean weight and variance provided an indication of the stability of the weight (See Yang, 2013 for a more detailed description). Feature selection for the outcome measure of score margin was not based on an ensemble approach, rather the weight of all PIs was based only on Correlation (Gini-Index, Information Gain Ratio and Information Gain are not compatible with continuous variables). The calculation of weight was also repeated 100 times using bootstrap sampling of training data, allowing us to rank the PIs by average weight. The top 45 PIs selected by the greatest average ‘weight’ (mean and variance), represented the most valuable half of the initial 91 PIs. The entire process was completed for each of the identified eras, as well as for both outcome measures, providing six separate datasets.

Data from each era was partitioned into model training and model testing sets using a 70:30 ratio in accordance with published recommendations (Hall, et al., 2011). For Eras 1 & 2 combined (2001–2016), the years 2001–10 were used for training and 2011–2016 for testing. For Era 1 (2001–2008), the years 2001–2005 were used for training and 2006–2008 for testing. For Era 2 (2009–2016), 2009–2013 was used for training and 2014–2016 for testing. Ten-fold cross-validation was applied to each eras’ training set. The testing set for each era were bootstrapped with replacement and iterated 100 times (Luo et al., 2016). Random forest models were created for Eras 1, 2 and 1 & 2 combined, using all 91 PIs. The criterion measures used to train the models were gini index for win-loss and least squares for score margin. The maximum depth ($n=7$) was chosen based on the performance of the random forest algorithm on the training data and comparing this with the performance of the trained model on the test data. A tree depth of 7, achieved relatively high performance in both data sets and no decrease in performance on the test set compared to the training set. The performance of the model of score margin, for the combined eras, was assessed by examining the residual errors, calculated as observed minus predicted values. The residual values were created by subtracting fitted value from the observed score margin.

Analyses of era identification used R Studio Computing Environment (Version 1.0.153, R Studio, 2016; (R Studio Team, 2015) and the 'segmented' package (Muggeo and Muggeo, 2017) for segmented regression, Microsoft Excel (Microsoft Excel for Mac, version 15.40, 2017) was utilised for Change-point analysis and Rapidminer Studio was utilised for the feature selection process and random forest modelling (Rapidminer Studio. Version 7.6.001. Dortmund, North Rhine-Westphalia, Germany).

Results

Segmented Regression analysis identified 2009.9 (standard error of 1.2 years) as the breakpoint between two eras. Whereas, Change-point analysis identified the breakpoint as being between the 2006 and 2007 AFL seasons. Continuing these analyses identified additional breakpoints which created eras that were small (e.g. ≤ 4 years) which may have limited the opportunity for modelling within these eras (i.e. because they were relatively small). The two methods (change point and segmented regression analysis) used to identify eras produced similar but not identical findings, consequently we had to reconcile the difference. Therefore, the mid-point between the findings of each method was used - the break between 2008 and the 2009 AFL seasons. Subsequent analyses were performed separately on seasons 2001–2008 (Era 1), 2009–2016 (Era 2), and 2001–2016 (Eras 1 & 2 combined).

Feature selection was completed for each era, using both dependent variables; win-loss and score margin. This resulted in six unique datasets for subsequent modelling. Results from Era 2 for win-loss and score margin are presented in Table 1. Despite the different feature selection processes required for each dependent variable, there was reasonable agreement between the PIs that were selected for each era. Of the most valuable 45 PIs selected for each era, 29 are common to all eras. There were an additional six PIs that appeared in five of the six sets, eight that appeared in four sets and five that appeared in three sets, a further four that appeared in two sets and five PIs that appeared in only one dataset. The top 10 selected PIs for each dependent variable and era were largely in common, including; Time in possession difference, Metres Gained relative, Marks Inside 50m, Kicks relative, Inside 50m relative and Disposals relative. Inside 50m per Shot on Goal, Metres Gained and Turnovers Forcing a Score relative each had one ranking outside the top 10, with rankings of 11, 21 and 23 respectively. The relative form of the PIs appears to be more closely related to match outcome than the raw form, as they represent 7-8 of the top 10 ranked PIs (see Table 1.).

Random Forest models were created to explain both win-loss and score margin for each era. The classification accuracies for win-loss in Era 2 performed best (88.8%), closely followed by Eras 1 & 2 combined (88.5%) and Era 1 (85.3%). The score margin based models achieved root mean squared errors of 17.8±0.1 for Era 2, 19.8±0.2 for Era 1 & 2 combined and 21.4±0.2 for Era 1 respectively. The figure below (Figure 2), is a residual plot of the combined eras. The model slightly under-estimates the magnitude of the score margin, for very large positive and negative observed score margins.

Discussion

The key aims of this study were to evaluate the benefits of 1) increasing the number of seasons and PIs used for modelling, 2) the identification of eras where technical performance characteristics were stable and 3) the application of a novel feature selection method. These aims were intended to collectively provide a basis for the methodology that could be used for modelling of match outcome, using sport databases that had relatively large numbers of PIs and seasons.

Table 1. Feature Selection. The 45 PIs with the highest weighting (i.e. relationship to match outcome) for the dependent variables, win-loss and score margin. Weight is the mean of weight determined from 100 iterations of the feature selection process. These results relate to Era 2 (2009–2016).

Rank	Match Outcome Dependent Variables			
	Win-Loss		Score Margin	
	Performance Indicator	**Weight**	**Performance Indicator**	**Weight**
1	Metres Gained relative	1.000	Metres Gained relative	1.000
2	Time In Possession difference	0.702	Kicks relative	0.814
3	Kicks relative	0.631	Time In Possession difference	0.822
4	Marks In 50m relative	0.522	Inside 50 Metres relative	0.794
5	Inside 50 Metres relative	0.508	Marks Inside 50 Metres relative	0.787
6	Disposals relative	0.491	Turnovers Forced Score relative	0.785
7	Inside 50 Metres Per Shot relative	0.480	Disposals relative	0.742
8	Kicks	0.475	Metres Gained	0.704
9	Turnovers Forced Score relative	0.444	Inside 50 Metres	0.696
10	Metres Gained	0.430	Kicks	0.704
11	Turnovers relative	0.419	Inside 50 Metres Per Shot relative	0.674
12	Inside 50 Metres	0.397	Rebound 50 Metres to Inside 50 Metres Percentage relative	0.676
13	Disposal Efficiency Percentage relative	0.374	Turnovers relative	0.656
14	Marks relative	0.373	Ball Gains relative	0.642
15	Rebound 50 Metres to Inside 50 Metres Percentage relative	0.362	Contested Possessions relative	0.612
16	Disposals Per Shot on Goal	0.354	Marks Inside 50 Metres	0.612
17	Disposals Per Turnover	0.339	Disposal Efficiency Percentage relative	0.610
18	Disposals	0.330	Disposals	0.585
19	Marks Inside 50 Metres	0.327	Disposal Per Shot on Goal	0.603
20	Ball Gains relative	0.319	Marks relative	0.602
21	Contested Possessions relative	0.316	Disposals Per Turnover	0.580
22	Inside 50 Metres Kick Retain Percentage relative	0.306	Turnovers Forced Score	0.591
23	Turnovers Forced Score	0.280	Inside 50 Metres Kick Retain Percentage relative	0.561
24	Ball Losses relative	0.273	Ball Losses relative	0.540
25	Inside 50 Metres Per Shot	0.272	Ball Gains	0.466
26	Marks	0.228	Inside 50 Metres Per Shot	0.477
27	Ball Gains	0.216	Rebound 50 Metres to Inside 50 Metres Percentage	0.465
28	Uncontested Possessions	0.212	Inside 50 Metres Kick Mark Percentage relative	0.425
29	Rebound 50 Metres to Inside 50 Metres Percentage	0.208	Uncontested Possessions	0.417
30	Inside 50 Metres Kick Mark Percentage relative	0.196	Clearances Effective relative	0.366
31	Marks from Opposition Kick relative	0.192	Marks	0.421

32	Inside 50 Metres Kick Retain Percentage	0.191	Rebound 50 Metres relative	0.395
33	Ratio of Turnovers to Turnovers forced	0.184	Turnovers Forced relative	0.421
34	Marks Contested relative	0.179	Clangers relative	0.403
35	Rebound 50 Metres relative	0.176	Inside 50 Metres Kick Retain Percentage	0.391
36	Clangers relative	0.175	Clearance Win Percentage relative	0.323
37	Turnovers Forced relative	0.174	Marks Contested relative	0.368
38	Disposal Efficiency	0.171	Handballs relative	0.365
39	Handballs relative	0.162	Ratio of Turnovers to Turnovers forced	0.403
40	Contested Possessions	0.150	Contested Possessions	0.335
41	Clearances Effective relative	0.143	Marks from Opposition Kick relative	0.365
42	Bounces relative	0.140	Clearance Win Percentage	0.274
43	Rebound 50 Metres	0.136	Disposal Efficiency	0.323
44	Clearance Win Percentage relative	0.136	Rebound 50 Metres	0.314
45	Turnovers	0.133	Bounces relative	0.299

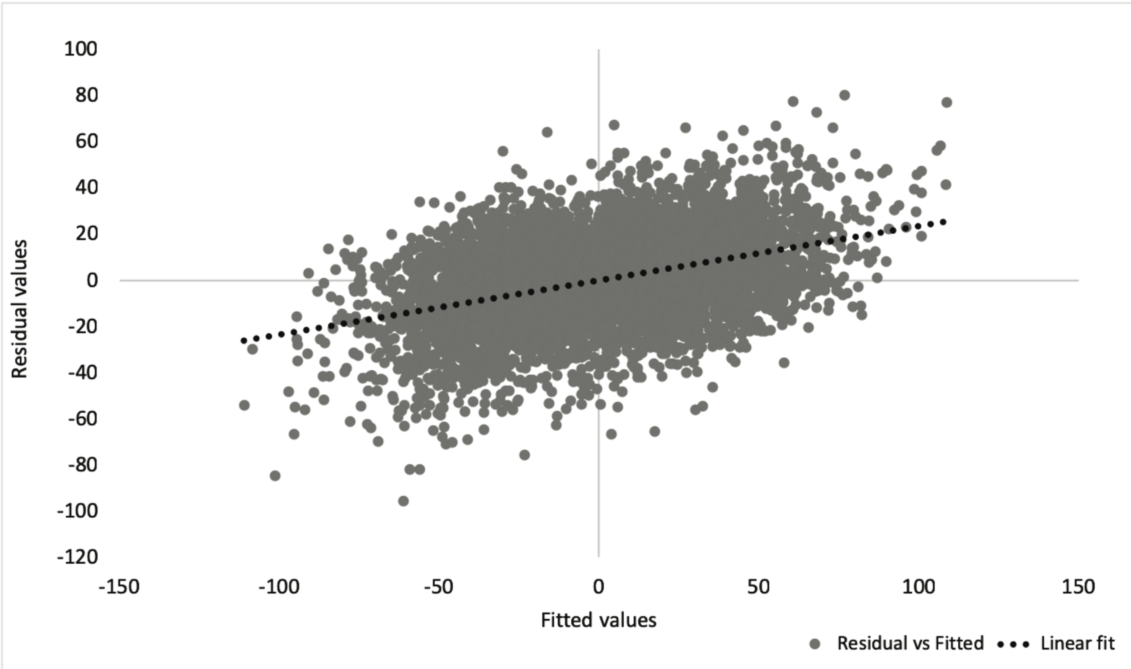


Figure 2. Plot of residual errors for the model that predicts score margin. The results are for the combined eras and the model overestimates the score margin for very large positive and negative observed values.

The identification of eras is an important initial step when analysing PI databases, as it can help to create models of performance that account for changes in the game and therefore create more accurate models. Eras were identified in the data using two methods, although the breakpoint/s between eras were not consistent between the methods. This inconsistency may be due to the multifaceted nature of the identification of eras historically (Berthelot et al., 2010). We could not identify a reason to choose the results from one method over the other, although this may have been possible if the results were very different. Nor did we want to continue the analyses in order to detect additional and therefore smaller eras. In the absence of a clear choice, we used the mid-point of the results for each method, dividing the dataset into only two equal sized eras with a break point between 2008 and 2009. A recent article by Woods, et al., (2017) investigated the evolution of game-play in AF across the 2001–2015 timespan and reported a period of change from 2005–2009, before levelling off in 2010. This period of change roughly aligns with the breakpoint between the eras we have used. Woods, et al., (2017) analysed a very similar time span of data to that used in the current work, although the number of PIs used for era detection were different (18 vs. 38). Given there were different PIs used in the current study, it is not surprising the current study identified slightly different breakpoints between eras.

The process of feature selection is an important step in the development of models that explain match outcome and the method utilised in this work has a number of strengths. Combining four methods and iterating the process 100 times provided an opportunity to determine the variance (and the confidence) in the ranking of the PIs. Feature selection identified the PIs that were likely to be most useful for modelling and these were similar to those identified in previous reports, even though our feature selection methods are novel (Robertson, Back, et al., 2016; Stewart, et al., 2007; Woods, 2016). We are in agreement with the PIs identified as being important by other authors including Kicks, Inside 50's, Marks, Marks Inside 50m and Contested Possessions (Robertson, Back, et al., 2016; Stewart, et al., 2007; Woods, 2016). Kicks and Goal Conversion, and Kicks and Inside 50s were the two most important PIs for modelling match outcome by Robertson, Back, et al., (2016) and Stewart, et al., (2007) respectively. These findings agree with the current study despite the inclusion of a much larger number of PIs, three of these four PIs were prominent in the top 45 PIs for each era and dependent variable. The exception was Goal Conversion, which we also found to be closely related to match outcome, but it was excluded from the current analysis as it was considered to be too closely related to scoring.

The PIs identified by the feature selection process demonstrated some variation between eras. An explanation of this could be that some PIs are not available in all seasons, including; Metres Gained relative, Time in Possession, Turnover Forced Score and Marks from Opposition Kicks. One of the most valuable PIs in the 2009–2016 era was Metres Gained relative, which is defined as 'the net distance gained with the ball by a player in possession, by kicking, handballing or running, combining measures towards attacking goal and away from defensive goal' (ChampionData, 2017). Time in Possession difference was also identified as an important PI, which may be considered an obvious conclusion.

Nevertheless, the identification of these PIs as important is novel and suggests there may be more useful PIs available that have not been used in this area of research to date. It has also been suggested the use of PIs in their relative form is more insightful (Robertson, Back, et al., 2016), the current work further supports this inclination. When all PIs were pooled together in their absolute and relative forms, the relative form of PIs were ranked more highly by the feature selection process and are more insightful for the modelling of match outcomes.

We have demonstrated that an ensemble feature selection process that has provided improvements to modelling in another area of science(Abeel, et al., 2009), can also be applied to the sport context. Although the process is slightly more complicated than more commonly used methods (i.e. a single algorithm) it should be relatively easy to implement in many types of data science software. Preliminary modelling of our dataset using random forest achieved model accuracies of between 85.3% and 88.8% for win-loss and a root mean squared error of between 17.8 ± 0.1 – 21.3 ± 0.2 points for score margin. Although, the random forest model had some difficulty in accurately predicting score margins of a large magnitude (positive or negative). These model accuracies are higher than the accuracy of 78.9% reported by Robertson, Back, et al., (2016), although some of the difference may be due to the different model types (random forest versus CHAID decision tree). It is likely that the improvement in accuracy could be attributed to the number and type of PIs. Era 2 held the most accurate models using both win-loss and score margin, closely followed by Era 1 and 2 combined. The current study has demonstrated an increased number of PIs results in more accurate models. The work of Stewart, et al., (2007) had the most extensive list of PIs previously with 52 PIs from 5 AFL season, including both dependent and independent variables, whereas the current study had a dataset of up to 110 PIs from 16 AFL seasons.

Previous authors have used a variety of methods to validate their models, which may have been limited by the amount of data available to them. It is common practice in the field of data mining that approximately two-thirds of available data should be used in training a model and one-third for testing purposes (Hall, et al., 2011), In the present study, five of eight seasons in each era (Era 1 and Era 2) and 10 of 16 seasons (Era 1 and 2 combined) were preserved for training our models and three/six seasons were held out for model testing. We also adhered to the recommendations that models are trained using a cross-validation approach and the determination of model accuracy on the test set, employs bootstrapped sampling with replacement (Hastie, Tibshirani, & Friedman, 2013; Luo, et al., 2016).

Some limitations of the current work include the outcome of our era detection. The compromise in era identification was a necessary decision given the lack of consensus between methods, which suggest that this is still an area for improvement in future research. Another limitation was the choice of which dependent variable to use. It has been acknowledged (Robertson, Back, et al., 2016; Robertson, Gupta, et al., 2016; Stewart, et al., 2007) that score margin is the preferred choice but there are limitations in the models available for dependent variables of a continuous nature and therefore both types of dependent variables were used. There has been little consensus on which feature selection method provides the best outcome, which model types are most appropriate and how models should be validated. We feel the current work has made a contribution to address some of these challenges.

Conclusions

The use of two methods to identify eras in our database provided different breakpoints, therefore a compromise was reached that maximised our opportunity for subsequent modelling of match outcome in each era.

Our novel feature selection process generated a ranked list of the most valuable PIs for modelling, which includes PIs that have not been previously identified as being related to match outcome. Furthermore, the results of the feature selection process indicate that the relative form of PIs are more valuable than the absolute form. The development of models of match outcome demonstrate that analysing a larger range of PIs can improve modelling accuracy. In addition, the relationships between PIs and match outcomes varied between eras, which confirms the importance of identifying eras in the first place. The findings from this work will be of particular interest to those employing data analytics in team-invasion sports. This work is not limited to guiding practitioners, as the key findings can be used by coaches and athletes in team sports. Future work could make use of the methodology used in this study to model match outcomes using data mining techniques.

Acknowledgements

Thank you to Champion Data Holdings for providing the raw data for this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. J. B. (2009). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *26*(3), pp. 392-398.

Berthelot, G., Tafflet, M., El Helou, N., Len, S., Escolano, S., Guillaume, M., . . . Desgorces, F. D. (2010). Athlete atypicity on the edge of human achievement: performances stagnate after the last peak, in 1988. *PloS One, 5*(1), p e8800.

Castellano, J., Casamichana, D., & Lago, C. (2012). The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of human kinetics, 31*, pp. 137-147.

Champion Data. (2017). *AFL Prospectus: The Essential Number-Cuncher For Season 2017* (12th ed.): Champion Data Pty Ltd.

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P. R., & McRobert, A. P. (2016). Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. *Journal of Sports Sciences, 34*(24), pp. 2195-2204.

Gómez, M. A., Gómez-Lopez, M., Lago, C., & Sampaio, J. (2012). Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football. *European Journal of Sport Science, 12*(5), pp. 393-398.

Gómez, M. A., Lorenzo, A., Barakat, R., Ortega, E., & José M, P. (2008). Differences in game-related statistics of basketball performance by game location for men's winning and losing teams. *Perceptual and Motor Skills, 106*(1), pp. 43-50.

Hall, M., Witten, I., & Frank, E. (2011). Data mining: Practical machine learning tools and techniques. *Kaufmann, Burlington*

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*: Springer New York.

Higham, D. G., Hopkins, W. G., Pyne, D. B., & Anson, J. M. (2014). Performance indicators related to points scoring and winning in international rugby sevens. *Journal of Sports Science & Medicine, 13*(2), p 358.

Jacklin, P. B. (2005). Temporal changes in home advantage in English football since the Second World War: What explains improved away performance? *Journal of Sports Sciences, 23*(7), pp. 669-679. Retrieved from http://ezproxy.deakin.edu.au/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=s3h&AN=17472456&authtype=sso&custid=deakin&site=eds-live&scope=site

Jones, N. M. P., Mellalieu, S. D., & James, N. (2004). Team performance indicators as a function of winning and losing in rugby union. *International Journal of Performance Analysis in Sport, 4*(1), pp. 61-71. Retrieved from http://www.ingentaconnect.com/content/uwic/ujpa/2004/00000004/00000001/art00007

Lago-Peñas, C., Lago-Ballesteros, J., & Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. *Journal of human kinetics, 27*, pp. 135-146.

Levendis, J. D. (2018). Stationarity and Invertibility *Time Series Econometrics: Learning Through Replication* (pp. 81-99). Cham: Springer International Publishing.

Liu, H., Gomez, M.-Á., Lago-Peñas, C., & Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 Brazil FIFA World Cup. *Journal of Sports Sciences, 33*(12), pp. 1205-1213.

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., . . . Ho, T. B. (2016). Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of Medical Internet Research, 18*(12)

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences, 32*(20), pp. 1881-1887.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine, 22*(19), pp. 3055-3071.

Muggeo, V. M., & Muggeo, M. V. M. (2017). Package 'segmented'. *Biometrika, 58*, pp. 525-534.

O'Brien, R. M. (2017). Dropping highly collinear variables from a model: Why it typically is not a good idea. *Social Science Quarterly, 98*(1), pp. 360-375.

O'Donoghue, P. (2009). *Research methods for sports performance analysis*: Routledge.

O'Donoghue, P., Ball, D., Eustace, J., McFarlan, B., & Nisotaki, M. (2016). Predictive models of the 2015 Rugby World Cup: Accuracy and application. *15*(1), pp. 37-58.

O'Shaughnessy, D. M. (2006). Possession versus position: strategic evaluation in AFL. *Journal of Sports Science and Medicine, 5*(4), pp. 533-540.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: a review and a framework. *Measurement in Physical Education and Exercise Science, 17*(3), pp. 171-186.

Robertson, S., Back, N., & Bartlett, J. D. (2016). Explaining match outcome in elite Australian Rules football using team performance indicators. *Journal of Sports Sciences, 34*(7), pp. 637-644.

Robertson, S., Gupta, R., & McIntosh, S. (2016). A method to assess the influence of individual player performance distribution on match outcome in team sports. *Journal of Sports Sciences*, pp. 1-8.

Stewart, M., Mitchell, H., & Stavros, C. (2007). Moneyball applied: Econometrics and the identification and recruitment of elite Australian footballers. *International Journal of Sport Finance, 2*(4), pp. 231-248.

Taylor, W. A. (2000). Change-point analysis: a powerful new tool for detecting changes. Retrieved Date from http://www.variation.com/cpa/tech/changepoint.html.

R Studio Team. (2015). RStudio: integrated development for R. *RStudio, Inc., Boston, MA URL http://www.rstudio.com*

Vaz, L., Van Rooyen, M., & Sampaio, J. (2010). Rugby game-related statistics that discriminate between winning and losing teams in IRB and Super twelve close games. *Journal of Sports Science & Medicine, 9*(1), p 51.

Woods, C. T. (2016). The use of team performance indicator characteristics to explain ladder position at the conclusion of the Australian Football League home and away season. *International Journal of Performance Analysis in Sport, 16*(3), pp. 837-847.

Woods, C. T., Robertson, S., & Collier, N. F. (2017). Evolution of game-play in the Australian Football League from 2001 to 2015. *Journal of Sports Sciences, 35*(19), pp. 1879-1887.

Yang, P. Z., Bing Yang, Jean Zomaya, Albert (2013). Stability of feature selection algorithms and ensemble feature selection methods in bioinformatics. *23*, p 333.