Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Quantification of Individual Rugby Player Performance through Multivariate Analysis and Data Mining

A thesis presented for the fulfilment
of the requirements for the degree of
Doctor of Philosophy
at Massey University, Albany,
New Zealand.

Paul J. Bracewell B.Sc M.Appl.Stat(Hons)

2003

OGS 10

MASSEY UNIVERSITY
APPLICATION FOR APPROVAL OF REQUEST TO EMBARGO A THESIS
(Pursuant to AC 98/168 (Revised 2), Approved by Academic Board 16.02.99)

Name of Candidate: Paul J. Bracewell ID Number: 95052126

Degree: PhD Dept/Institute/School: Statistics/IIMS

Thesis Title: Quantification of Individual Rugby Player Performance Through Multivariate Analysis and Data Mining

Name of Chief Supervisor: Denny H. Meyer Telephone Extn: 9495

As author of the above named thesis, I request that my thesis be embargoed from public access until (date) 28/02/05 for the following reasons:

☒ Thesis contains commercially sensitive information.

□ Thesis contains information which is personal or private and/or which was given on the basis that it not be disclosed.

□ Immediate disclosure of thesis contents would not allow the author a reasonable opportunity to publish all or part of the thesis.

□ Other (specify): ______

Please explain here why you think this request is justified:

This thesis details the construction of a commercial rating system and discusses the business processes that contribute to the calculation of such a rating system. Public dissemination of this information would remove the competitive edge the sponsoring company would otherwise possess in the commercial sector.

Signed (Candidate): ______ Date: 22/2/02

Endorsed (Chief Supervisor): D. H. Meyer Date: 22/2/02

Approved/~~Not Approved~~ (Representative of VC): ______ Date: 24/2/02

Note: Copies of this form, once approved by the representative of the Vice-Chancellor, must be bound into ***every*** *copy of the thesis.*

Abstract

This doctoral thesis examines the multivariate nature of performance to develop a contextual rating system for individual rugby players on a match-by-match basis.

The data, provided by Eagle Sports, is a summary of the physical tasks completed by the individual in a match, such as the number of tackles, metres run and number of kicks made. More than 130 variables were available for analysis. Assuming that the successful completion of observed tasks are an expression of ability enables the extraction of the latent dimensionality of the data, or key performance indicators (KPI), which are the core components of an individual's skill-set.

Multivariate techniques (factor analysis) and data mining techniques (self-organising maps and self-supervising feed-forward neural networks) are employed to reduce the dimensionality of match performance data and create KPI's. For this rating system to be meaningful, the underlying model must use suitable data, and the end model itself must be transparent, contextual and robust.

The half-moon statistic was developed to promote transparency, understanding and interpretation of dimension reduction neural networks. This novel non-parametric multivariate method is a tool for determining the strength of a relationship between input variables and a single output variable, whilst not requiring prior knowledge of the relationship between the input and output variables. This resolves the issue of transparency, which is necessary to ensure the rating system is contextual.

A hybrid methodology is developed to combine the most appropriate KPI's into a contextual, robust and transparent univariate measure for individual performance. The KPI's are collapsed to a single performance measure using an adaptation of quality control ideology where observations are compared with perfection rather than the average to suit the circumstances presented in sport.

The use of this performance rating and the underlying key performance indicators is demonstrated in a coaching setting. Individual performance is monitored with the use of control charts enabling changes in form to be identified. This enables the detection of strengths/weakness in the individual's underlying skill-set (KPI's) and skills.

This process is not restricted to rugby or sports data and is applicable in any field where a summary of multivariate data is required to understand performance.

Acknowledgements

During the course of this thesis I have received support, encouragement and advice from my supervisors, academics, sports-people, fellow postgraduate students and business associates. I thank all these people whole-heartedly for their assistance, understanding and contributions allowing this thesis to be completed.

I am especially grateful to the input provided by my supervisor, Associate Professor Denny Meyer, who inspired my pursuit of statistics, nurtured my passion for sport statistics, encouraged me to chase my ideas and provided me with the necessary direction and support required to complete this research. I cannot speak highly enough of Denny's wonderful influence upon my research and development as a student. Additionally, her assistance with defining the theoretical variance associated with the univariate parametric half-moon statistic was most helpful.

The encouragement and guidance provided by my co-supervisor, Dr Siva Ganesh, is also greatly appreciated. Additionally, I wish to thank Professor Jeff Hunter and Dr Paul Cowpertwait for their constructive comments for improving the heuristic test of independence introduced in Chapter Five

The input of business mentor, Mr Chris Lines, of the Eagle Technology Group was also extremely valuable. His assistance was instrumental in obtaining funding via the Graduate in Industry Fellowship from Technology New Zealand. Further, Chris provided excellent support relating to the feasibility of the models created. Additional to the academic and business contribution was the beneficial assistance provided by the numerous top-level coaches, selectors and players who made themselves available for discussion and freely voiced their opinions.

I am also grateful to the support provided by my family. Some of the philosophies pursued in this thesis were sown at a young age by listening to my father and his brothers debating their theories on sports coaching and performance – all with first class playing experience, in cricket and/or rugby. Given that they are still involved in high level coaching, their ideas proved to be relevant, challenging and thought provoking.

Table of Contents

Embargo iii

Abstract v

Acknowledgements vii

Table of Contents ix

List of Illustrations xiii

List of Tables xiv

1 Introduction 1

1.1 Prologue 1
1.2 Ideology 3
1.3 Framework 4
1.3.1 Previous Ideals *4*
1.3.2 Relevance of Research *6*
1.4 Research Questions 7
1.5 Process 9
1.6 Summary of Methodologies 11
1.6.1 Methods *11*
1.6.2 Philosophies *12*
1.7 Summary of Results 14
1.8 Layout 16

Part One: Introducing The Eagle Rating

2 Literature Review – Statistics In Sport 21

2.1 Statistics and Complex Team Sport 21
2.2 Definition of Sporting Ability and Performance 24
2.3 Quantification of Sporting Performance in Team Sports 28
2.3.1 Data Collection *29*
2.3.2 Selection of Numerical Measures *30*
2.3.3 Variety of Measures Between and Within Positions *32*
2.3.4 Combined Measures *34*
2.3.5 Validation of Measures *36*
2.3.6 Rating Systems for Individuals in Team Sports *37*
2.3.7 Commercial Systems *40*
2.4 Quality Control as Individual Performance Monitoring in Sport 42

3 Multivariate Analysis, Quality Control, and the Birth of the Eagle Rating........ 45

3.1 Overview........ 45
3.2 Data........ 47
3.3 Methods for Data Analysis........ 54
3.3.1 Cluster Analysis........ 54
3.3.2 Factor Analysis........ 56
3.4 Cluster Analysis........ 59
3.4.1 Clustering Philosophy........ 60
3.4.3 Coaching Perspective........ 63
3.4.4 Analysis........ 69
3.4.5 Application........ 70
3.4.6 Discussion........ 76
3.5 Factor Analysis........ 78
3.5.1 Underlying Philosophy........ 78
3.5.2 Applying Factor Analysis........ 80
3.5.3 Results........ 83
3.5.4 Discussion........ 86
3.6 Multivariate Quality Control........ 87
3.6.1 Standard Quality Control and Performance Monitoring in Sport........ 88
3.6.2 Existing Multivariate Control Methodology and the Perfection Method... 90
3.7 Summary of the Eagle Rating........ 97
3.8 Implementation........ 100
3.9 Limitations........ 106
3.10 Conclusion........ 111

Part Two: Improving The Eagle Rating

4 Literature Review – Data Mining and Neural Networks........ 115

4.1 Data Mining........ 116
4.2 Neural Networks........ 121
4.2.1 Relation to Biological Networks........ 121
4.2.2 Brief Historical Overview........ 122
4.2.3 Applications........ 128
4.2.4 Cautions........ 129
4.3 Supervised Neural Networks........ 136
4.3.1 Types of Learning........ 137
4.3.2 Supervised Neural Networks Structure........ 138
4.3.3 Neural Network Architecture........ 142
4.3.4 Training Algorithm for Neural Networks........ 144
4.3.5 Activation Functions........ 147
4.4 Self-Supervised Neural Networks in Dimension Reduction........ 149
4.5 Self-Organising Maps in Dimension Reduction........ 151
4.5.1 Architecture........ 152
4.5.2 SOM Mechanical Properties........ 155
4.6 Interpreting Networks........ 157
4.7 Neural Networks and Sport........ 158

5 Investigation and Development of a Heuristic Test for Independence 163

5.1 The Introduction of a Heuristic Test for Independence 163
5.1.1 Covariance, Correlation and the Coefficient of Determination 165
5.1.2 Other Tests of Independence 167
5.2 Underlying Philosophy for the Half-Moon Statistic 169
5.2.1 Mathematical Background for Bivariate Case 170
5.3 Parametric Test for Bivariate Case 172
5.3.1 Disadvantages of Parametric Test 174
5.4 Non Parametric Test for Bivariate Case 175
5.4.1 Effect of Sample Size 177
5.4.2 Effect of Skewness and Kurtosis 178
5.4.3 Examining Non-Linear Functions 183
5.5 Expansion to the Multivariate Case (Multiple Inputs) 189
5.6 Implementation of MHM to Calculate Relative Influence 192
5.7 Application 194
5.8 Conclusion 196

6 The Rebirth of the Eagle Rating Through Neural Networks 199

6.1 Data Pre-Processing 200
6.1.1 Sub-Group Identification 200
6.1.2 Data Limitations 202
6.1.3 Data 206
6.1.4 Defence/Attack Philosophy 210
6.2 Dimension Reduction Neural Network 212
6.3 Factor Analysis 229
6.3.1 Condensed Factor Analysis Module using Key Attributes 229
6.3.2 Player Performance Indices 235
6.3.3 Standard Factor Analysis using Summary Variables 235
6.4 Dimension Reduction Self Organising Map 238
6.5 Quartile Method Based Method for Discrete Ranking 245
6.6 Summary of Potential Rugby Ratings 248

7 Comparison, Selection and Implementation of Suitable Methods 251

7.1 Methods for Comparison 252
7.1.1 Original Eagle Rating 254
7.1.2 Factor Analysis for Condensed Data 255
7.1.3 Neural Network 257
7.1.4 1 × 10 Self Organising Map 258
7.1.5 4 × 4 Self Organising Map 260
7.1.6 Paired Self Organising Map 261
7.1.7 Augmented Indices 262
7.1.8 Quartile Model 264
7.2 Comparison of Methods 265
7.2.1 Model Context 266
7.2.3 Model Validity 269

7.2.3 Model Robustness *271*
7.3 Discussion 273
7.4 Implementation of Quartile Model 275
7.4.1 Implementation in the Media *275*
7.4.2 Implementation in the Coaching Environment *276*
7.5 Implementing Statistics in a Diagnostic Coaching Structure 276
7.5.1 Basic Philosophy *276*
7.5.2 Performance Applicability *279*
7.5.3 Creation of Stable Statistics *281*
7.5.4 Incorporating 'Quality Control' Into a Diagnostic Structure *282*
7.5.5 Application of Control Charting Procedures *283*
7.6 Conclusion 285

8 Conclusion 287
8.1 Overview 287
8.2.1 Research Questions *288*
8.2.2 Research Answers *288*
8.3 Implications 292
8.3.1 Limitations *293*
8.4 Contributions 294
8.5 Future Research 295

Appendix A: Data Description 299

Appendix B: Results of Factor Analysis 307

Appendix C: Job Requirements 317

Appendix D: Commercial Applications 321

Appendix E: Glossary 327

Appendix F: Half-Moon Related Methods 331

References 333

List of Illustrations

2.1: Schematic Representation Of Performance 25
3.1: Boxplot Showing Openside Flanker Domination of Tackle Counts 69
3.2: Dendogram Indicating Potential Positional Clusters 72
3.3: Boxplot for Metres Kicked Per Game by Cluster 74
3.4: Two Dimensional Plot Of Perfection Method Explanation 92
3.5: Relationship Between The Magnitude Of Perfection And Normality 93
3.6: Histogram Of The Eagle Rating 96
3.7: Normal Probability Plot For Eagle Ratings From The 2000 Super 12 96
3.8: Comparison Of The Eagle Rating And Wynne Gray 102
3.9: Shewhart Control Chart Monitoring Cullen's Super 12 Performance 105
4.1: Mcculloch-Pitts Neuron For The Logical And Function 124
4.2: A Simple Neuron 139
4.3: Multilayer Neural Network For Output K 140
4.4: Self-Supervised Learning 150
4.5: Self Organising Map Architecture For A 2-D SOM 153
4.6: Output Layer For A 1×10 Self Organising Map 154
5.1: Geometrical Representation Of The Half-Moon Statistic 170
5.2: Boxplot Of Parametric HM Statistic For Varied Radii 171
5.3: Histogram Of Parametric HM 173
5.4: Plot Of Parametric HM Statistic Against The Coefficient Of Determination 174
5.5: Histogram Of Non-Parametric HM 175
5.6: Plot Illustrating The Effect of m on the SD For The HM Statistic 176
5.7: Relationship Between Sample Size And Half Moon Variability 178
5.8: Relationship Between The HM Statistic And Skewness 181
5.9: Relationship Between The HM Statistic And Kurtosis 181
5.10: 3D Plot Demonstrating Relationship Between HM, Skewness And Kurtosis 181
5.11: HM Statistic Versus R-Sq For Polynomial Functions 184
5.12: Type II Errors For Non-Parametric Hm Statistic By Varied Sample Size 184
5.13: Mean DHM Statistic Versus R-Sq For Polynomial Functions 186
5.14: Scaled Chi-Square Statistic Versus R-Sq For Polynomial Functions 186
5.15: Mean DHM Statistic Versus Scaled Chi-Square Statistic 186
5.16: DHM Statistic Versus Observed R-Sq For Linear Functions 188
5.17: Scaled Chi-Square Statistic Versus Observed R-Sq For Linear Functions 188
5.18: Boxplot Illustrating Unscaled Z Distances For n=1,2,3,10 And 20 191
5.19: Boxplot Illustrating Scaled Z Components For n=1,2,3,10 And 20 191
6.1: Draftsman Plot Of First Defensive 'Output' Node And Input Variables 217
6.2: Draftsman Plot Of Second Defensive 'Output' Node And Input Variables 217
6.3: Draftsman Plot Of First Attack 'Output' Node And Input Variables 218
6.4: Draftsman Plot Of Second Attack 'Output' Node And Input Variables 218
6.5: Implication Of Principal Planes And Surfaces 219
6.6: 1×10 Self Organising Map for Midfield Backs 243
6.7: 4×4 SOM Referenced Against Attack And Defence Indices For Midfield Backs 243
6.8: Boxplot Of Attack Measure By SOM Column For Midfield Backs 244
6.9: Box Plot Of Defensive Measure By SOM Row For Midfield Backs 244
6.10: One-Dimensional 4×4 SOM Referenced Against Attack And Defence Indices 245
6.11: Structure Of Forced Operator Model Scaling For Attack And Defence 247
7.1: Flow Chart Applying An Individualistic Rating System As A Diagnostic Tool 278
7.2: Shewhart Chart Of Kelleher's Overall Rating By Match In The Super 12, 2000 284
7.3: EWMA Of Overall Rating For Kelleher By Match In The Super 12, 2000 285

List of Tables

3.1: Summary Of Eagle Sports Variables 50
3.2: Correlation Matrix Indicating Correlation Between Selected Variables 51
3.3: Perceived Positional Clusters....67
3.4: Cross Tabulation Indicating Relationship Between Position And Cluster 73
3.5: Cross Tabulation Indicating Relationships Between Clusters And Position 75
3.6: Summary Of Factor Labels By Positional Cluster....84
3.7: Varimax Rotated Factor Loadings And Communalities For Midfield Backs 85
3.8a: Correlation Matrix of Ratings With Different Magnitudes Of Perfection 94
3.8b: Correlation Matrix of Ratings With Different Magnitudes Of Perfection....94
3.9: Top 22 Performers In The 2000 Super 12 100
3.10: Allocation Of Tri-Nation Representatives Ranked By The Eagle Rating 101
3.11: Anova Differentiating Between Internationals And Non-Internationals 104
3.12: Anova Demonstrating Differences Between Winners And Losers 104
5.1: Skewness And Kurtosis Data....182
5.2: MHM Comparable Radii 190
5.3: Calculation Of Key MHM Components Using A Stepwise Procedure 193
5.4: HM Procedure Output For Attack KPI By Summarised Key Inputs....195
6.1: Summary Variables For Modelling 207
6.2: Correlation Matrix Of Summarised Variables....209
6.3: Optimal Architecture For Midfield Defence Neural Networks 215
6.4: Optimal Architecture For Midfield Attack Neural Networks....215
6.5: Numerical Measures For Defensive Parameter Importance 220
6.6: Numerical Measures For Attacking Parameter Importance....224
6.8: Summary Of Self-Supervising Neural Results For Attack 226
6.9: Summary Of Self-Supervising Neural Results For Defence 226
6.10: Summary Of Important Variables In Neural Network Module 228
6.11: Rotated Factor Loadings And Communalities For Midfield Defence Variables....230
6.12: Rotated Factor Loadings And Communalities For Midfield Attack Variables 230
6.13: Coefficients For Midfield Attacking Variables 231
6.14: Factor Analysis Module Summary 234
6.15: Appropriate Methods For Positional Cluster Performance Indices 235
6.16: Rotated Factor Loadings And Communalities For Midfield Backs 236
6.17: Summary Of Factor Analysis Results....237
6.18: Differences In Variable Importance Between SOM Topologies....241
6.19: 4×4 Som Co-Ordinates And Output ID 243
7.1: 4×4 Som Topology 261
7.2: Correlations Between Input Variables And Obtained Ratings 266
7.3: Correlations Between Match Outcome And Meta-Data....268
7.4: Inter-Rating Correlations....269
7.5: Top 22 Performers in the 2000 Super 12 According to the Augmented Indices 270
7.6: Comparative Performance Measures Of Ratings....271
A.1: Summary Of The Eagle Sports Coding System....300
B.1: Labels And Codes For Factor Analysis 307