Higher Degrees by Research
Examiner's report on thesis

[redacted]

For

Andrew J Simmons

Overall comments

Thank you for the opportunity to review this thesis. Overall, it is a large piece of work, with detailed writing across a number of disciplines highlighting the process needed to align data, anonymise, work it over and add a hint of analytics at the end. As a thesis, it is a complete piece and flows well, so well done to the author.

The thesis has a number of published pieces within the work, and I am happy to pass the work subject to the remaining comments and corrections attached. There are a number of suggestions and corrections, and I believe if addressed shall make this work a satisfactory piece.

A few broad comments:

A lot of statements are made as if they are fact – especially around coaching and statistics. I believe the author has worked with a club, however coaches do have very broad ideas. I have worked with many coaches, and there are number haters, number lovers and a big mash in the middle. Ultimately, and as is well argued in pieces, if the data matches and outcome and is attributable to an action then we have something. One of the big consternations I have is the de and re identification of data. The concept present here makes a lot of sense for researchers and users outside of a club. However, the club itself will need something in return! So this is a rather delicate juggling act and I am not sure that I am convinced of this, especially given Chapter 7 details some nice broad statistics, but nothing down the level of detail I am aware of in clubland. So at times I shall criticise statements that I believe are too broad and sweeping and ask for some addition to tidy up the 'always and everybody' nature of these comments.

Second, numbers less or equal to ten in words. Easy fix. The use of 's' instead of 'z'.

Thirdly, there is a rather large contradiction with the de-identification debate in Chapter 5 and the work in Chapter 7. This is outlined in comments for Chapter 7, but I think the author is treading a little dangerously, if by accident, however needs to pick a side. The deletion of some sentences, and the tidying up of comments should make it all sound.

I will leave it up to the discretion of the Head of School as to whether a comment requires adhering to by the recommendation; or a salient rejoinder highlighting why that comment is not applicable, or otherwise. I am happy to be contacted for any points of clarification.

Well done to Andrew.

Chapter 1

Comments:

I really enjoyed the opening and was pleased with the layout, clarity and definition of purpose. The motivation is clear and is a nice teaser for the forthcoming chapters.

I was expecting some insight into the rationale for de-identification, as the author does make it a focal point. From a coaching/analyst's perspective this would be a rather big hurdle in a pragmatic sense, and I believe that there needs to be, even if brief, some mention of why this is an issue. I know it is coming, but upfront be specific about it. Else, the work could have been undertaken without the data be de-identified, and the work embargoed if need be, and still have achieved the outcome. I get it (am acutely aware of the data sensitivity around club and league land!) but a reader outside of AFL might think why? I know this covered in great detail in Chapter 5, but you could talk about it, or refer to it, here.

p.6 states "Common data operations required for spatial analysis…" introduces this idea of de-identifying without a detail of its merit. I am presuming that this operation is necessary for *research* purposes but not pragmatic needs.

Annotations:

AFL is mentioned first on page 10, in the Considerations, without its explanation, which appears first in Chapter 2. Needs to be detailed i.e. AFL (Australian Football League) or replaced with Australian Rules Football upon its first mention.

Chapter 2

Comments: This is a neat chapter that really ties up the aim of the work in a way that hinges on the perceived gaps in spatio-temporal analysis. Like the +/- discussion and the structure of the arguments. There are a few points that need address and I have some thoughts around some generalisations and *black and white* sounding type paragraphs that need some addressing.

p.13. (see Annotations too) Please add annotation++ as it is not always the case that a pass is always a player decision. Coaches can advise for specific connections to occur despite a player's intuition. Furthermore, tactics are often driven from players on the field – this point needs to made as I feel the author has a fairly compartmentalised approach to coach and player roll

p.14 – 'thus this form of communication…" I am not sure about this statement -but I understand the idea – I assume you mean the coach cannot say anything as it happens, however direction from a coach is possible in terms of things to do next. If this the case all good
p.14 – "Coaches provide video…their own style". This is a generalised perception and speaking from experience there are multiple utilities of video feedback. This needs expansion. I suggest the following as a guide to the use of video, as the author has put it into bins

- This set up was wrong – coach feedback on a play/set up of a structure
- Why did you do this – coach feedback to a player on their actions
- You should do this – coach feedback on what did work well
- Was this ok – player feedback
- Opposition vision – let's do this is as they do that

I feel this section needs expansion, as it reads like there are narrow ways to do video. Indeed, and as is stated, we want to encourage uniqueness (ie the flop) and if anything, over structured analysis leads to predictability. Don't assume it is always done this way. Clubs don't always disclose what they do as that is their competitive edge.

p.15 "Currently ***some*** coaches use a range of summary statistics…"

Statement as written is asserted as if law and this is definitely not true. I would be shocked if this was true for most professional clubs/leagues. This is a very general paragraph and you need to maybe provide a vignette as I simply don't agree. Summary statistics have their utility, however are you saying that there are only summary statistics available that is incorrect?
In my experience at clubland, summary statistics are broken into team based and individual based parameters and if the coach/players aren't aware that these are products of performance, not causative, then there is certainly a problem.

You are probably attempting to state that stats out of context, or without some spatial reference, are somewhat misleading. That is certainly true, and to put this all away, I think it is the knowledge of what the stat is describing and doing that is a potential problem, not that the stat itself is poor. Please rewrite this paragraph to assert this point, or rejoinder my argument.

p.16 “Teams head in …” Define that a bit better. (ie play towards differing ends). You could easily use some visuals to help clarify this section.

You also receive 1 point if you hit the post, or rush a behind…these little details are completed nicely in the Ryall PhD thesis if you need ideas.

Also, please calculate the occurrence of draws – shouldn’t be hard to evaluate.

p.20 the goal assists paragraph – I am again uncertain I agree. The aim of a game is to score goals, and indeed the pass before a goal is a back-credit for that goal. However no goals no score no win. It is not possible *not* to pass the ball and get a goal? So how can this be a bad incentive? How can a goal assist not provide an advantage to a team? Please rewrite or remove

p.22 and p.29 The Clarke model and commentary around gambling etc is important. The odds of a team offered are effectively a public perception model of the outcome of a game, which could be approximated as the true odds of a game. Beating a market model is typically possibly if you introduce a new variable that maps against winning that others do not have. Furthermore, the market itself may simply be inefficient, or have certain holes in odds estimates. Beating a gambling market really is not important to teams

p.40 GPS v Video indoors – This is not quite true. RF systems indoors do provide detailed positional based info too, and are often suitable over video systems due to occlusion. However, video systems have the distinct advantage of not requiring the athletes to wear anything which is a big deal in codes that do not have rules around everyone wearing the same tech (ie netball) and information sharing.

Annotations:

p.13 – optimizing (inconsistent use of s and z)
p. 13 – (a decision **traditionally/typically**…) ++
p.25 - 2 minute – change to two minute

Chapter 3

This chapter develops the model framework for the thesis through the definition of solutions. This was a well-developed chapter and was impressive in its breadth and care.
I thought 3.3 was excellent; and 3.3.1 was nice in its idea around unknowns.

p.66 the talk of holistic approach – I was a little confused here. The knowledge of a team performance would help inform player feedback if the system was holistic? So is the proposal inclusive of performance variables?

p.68 I thought AFL used T6/clearsky which is RF &/or GPS, but I could be wrong! You should say this is at time of publishing too. T6 are 100 Hz?

Annotations/Corrections
p.56 – "This is represented…. Actions take ***place***"

p.56 - "…stochastic process". Please refer to the PhD thesis by Don Forbes – "Dynamic prediction of Australian Rules football using real time performance statistics"

p.62 – AFL equity ratings – reference?

Chapter 4

Section 4.2.3. Was excellent. This is closely tied to how PA work. Well done. It is worth noting that the Workflow Provenance example of a player rating is a real consequence of a PA's job. One must have the rating tied to tangible and identifiable causes before you throw numbers at them. This was nicely done and relates closely to PA practice

This chapter was polished and a nice intro into the methods needed to achieve the desired outcomes expressed in chapter 3. A lot of this chapter was new to me so I was quite impressed by the W3C idiom

Chapter 5

P126. Interesting but this is one thing that I sort of grapple with

I will talk from experience – we collect data and have a blanket allowance to research (internally) and exclude identification of individuals in the knowledge that the re-coupling of information to the data is complex, if not impossible, without the data in the beginning.

I am a bit perplexed (and this might be at the heart of the problem!) as to how this can be managed. If you obtain consent from players etc. and have an open environment then it is case closed!

We share all our data with the players and some data is provided live on TV during training and games. Henceforth, the transparency of data and its use forms a part of the club, culture and its treatment is still one that is considered private. If the data isn't released, and the analysis is completed in house, then there is a layer of protection?

If the raw data is obtained it is rather trivial to align vision with tracking data. I am not sure how it is of use if it is not paired with other information.

However, you (p128) go on to outline the true difficulty in this task nicely.

p.148-151 Greenham paper – Did they acknowledge Champion Data as there are other places one could get that data (ie ProWess, AFL Tables, or independently)
Also if a paper is approved by a University committee, then is the consequence theirs, the researchers, or both? Nicely done.

Section 5.6 is really well done. I like the process and it makes heaps of sense. Again whilst I am grappling with the idea of this being a problem, this is a nice way of attacking the problem. The mixed sets do clearly make sense, as does a downsizing in resolution, however again you are losing huge amounts of information. This is probably ok in an AFL context as it is an open sport with anyone allowed to do anything; however, this would not work that well for say Netball with specific roles and zones endemic to the rules of the game.
Was happy with the way you argued the limitations.
The proposed interaction model in 5.7.1 is very nice and tidy, and the model proposed is excellent. It kind of reminds me of the current catapult openfield/cloud system, whereby you have differing access levels with different outcomes (although this is much better!)

Well done a great chapter. Again, I am still coming to terms with the removal of the levels of player-based information, however I get it if this is for researchers outside of a clubland situation wishing to play with team data.

Annotations

p.136 two instead of 2
p.138 as above
p.152 2 of 3 – two of three

Chapter 6

This chapter looks at the transformation of data into an accurately referenced set without the need to be a GIS expert. Mention is made of systems such as clearsky, and the author outlines an algorithm which undertakes this process. This was an interesting chapter, which I found a little more instructional than innovative, although there was the algorithm process detailed towards the end. I particularly enjoyed the 6.2.3 section.

Section 6.3 needs a tidy up as it reads like a paper rather than a part of the thesis – this is covered in the annotations

Annotations/Corrections

p.194 footnotes precede the text on the next page
p.199 sport analyst then sport performance analyst – changing titles? Please pick one
p.199 Use of the word 'we'
p.199 "In this paper…"

Chapter 7

This chapter really was the one I was looking forward too. It has some nice applied examples, and I was heartened by some of the analysis. I must say though, I was expecting a lot more analysis here, and given the length of the thesis, maybe I was being overly optimistic. Nonetheless, there has to be some practicable outcomes from this work to sell the idea.

Firstly, I was nicely surprised by the use of the de-identified data and thought that it was obvious to me for the first time that it could be utilised in a meaningful manner. Nonetheless, it is essentially very basic inferential statistics and I was left salivating the idea of much more detailed and insightful analysis could be achieved. I will suggest some ideas; however I hope the author utilises more of the outcomes in future research, or at least mentions that they could be used in a variety of methods.

p.217 Now I have just read about the process and importance of de-identification yet you have done precisely that yourself. You mention the football club and the club numerous times, yet you de-identify the team as Geelong (I think!) though induction of when the game was missed due to the death of the Adelaide coach. Then on p 220 you identify Kardinia Park, and the chapter continues with numerous Geelong references and examples. Now this is something you spent so much time on de-identifying, then you go about analysis without the same protocols in your writing! This whole chapter talks about the club etc. Either identify it as Geelong (or whichever club it is) or remove the information that leads me at least to think it is Geelong. It is hypocritical to be critical of other works in Chapter 5 for doing this when this occurs in this chapter – so please make a choice either way.

On a technical note, the sampling is very poor in the end. To have only six matches of data is a big concern! Speaking from my experience, we use a *fickle* T6 system and have never lost data despite plenty of drama with the tech – this seems like either the users aren't adept at using it or the tech is inadequate or the players aren't wearing it properly or a combination as you outline nicely in Fig 7.6. Whilst not the issue for the author, it is a concern if one is to do any form of analysis and this level of data loss occurs. How can a PA be taken seriously with this level of data error? They would be out of their job in a week. It would be interesting to note if there are places that are better than others to receiving signals, or contingencies for data loss.

The synchronised vision and pipeline sections were well done.

7.2 This was nice with ideas thrown about at a team level. From experience, wouldn't it be better to use end-on footage rather than broadcast? Most coaches I have worked with rarely to never use broadcast

7.2.6 Quantitative Analysis

The team spread is interesting – I can't help but think that you could look at the outcomes as a function of the matched paired spreads too.

So something like goal scored proportions as a function of overlapping spreads

		Attack					
		0-10	10-20	20-30	30-40	40-50	50+
Def	0-10	x%	y%				
	10-20						
	20-30						
	30-40						
	40-50						
	50+						

This would describe the interactive nature of this index as it happens, rather than a group-based approach you have taken. Whilst differences were found, I wonder if this is not of value given that this is not considered for. The location as a reference starting point is value, however as the data is detailed one should be able to deduce this.

Game speed and Team Spread

This one too shows us some nice results. It is probable that as the speed of the game is quicker, so is the spread as one would not think that everyone would speed into each other, rather away from each other. What is notable is that whilst both are linear, the defence gradient is gentler than attack. Also, the quantity of goals scored is more of a cluster than a correlation? Maybe you could address this in the text.

p.250 First paragraph is pretty close to the mark. I think that the How is now really important. You have demonstrated it is possible, however I wonder if there is still a way to go to make this accessible, dynamic and interactive as or near-as the play goes.

Annotations/Corrections

p.214 for 2015 matches … rewrite this as it sounds like a quantity not a year

p.217,8 you use numbers for words when you should use words – anything 10 or less should be written in words – please correct throughout this chapter.

p.241 during the goal kick – should that be goal attempt?

Chapter 8

A nice tie together. Nicely done.

p.256 Sounds good, but I am sure the enemy doesn't wear GPS and share

Appendix C

p.310 – the mean group being taller sounds a stretch in de-identification. That is a clunky conclusion but yes we could think it is ruckmen

p.313 – That is a shocker! But one could draw similar conclusions about this thesis, Deakin being aligned with Geelong so be careful