Chapter 1

Cinemetrics and R

1.1 Introduction

It started with an enforced period of convalescence and a DVD player, bought to while away time, that rekindled a dormant interest in German expressionist and other early films of the kind screened by the better university film societies when I was a student. Things got out of hand and, as academics do, it occurred to me that I ought to know more about the subject beyond just watching the films (noir by this stage also having engaged my interest). Beyond the copious information and mis-information available on the web something more seemed to be needed – a book.

You can't stop at one book, they're addictive, my tastes are eclectic, and everything from pulp biography, through serious (and in some cases seriously long) historical studies, and film theory including feminist readers on silent film, became grist to the mill. At some point I think I must have identified any understanding of film technology, and the way it might have conditioned the appearance of films I was watching, as one of many major gaps in my knowledge. Having seen several intriguing references to it, Barry Salt's *Film Style & Technology: History & Analysis* found its way onto my shelves. It wasn't quite what I was expecting.

Apart from the bracing views on film theory and theorists expressed in the opening chapters, it was the several chapters on statistical style analysis of motion pictures that intrigued me. This kind of unexpected intersection between leisure and professional interests is rare, though not unknown as a lot of my research activity as an academic statistician has been in the realm of archaeology, which began independently as a non-academic pursuit. My initial intrigue was piqued by the discovery that statistical analysis of 'filmic' data had a name, cinemetrics, an eponymous web-site, and that practitioners got quite excited (or is it excitable) about the 'correct' application of statistical methods to cinemetric data.

This, I thought, looks fun. I lead an unexciting life and as one of my ideas of fun is doing data analysis (out of and for interest, but not of the monetary kind) I began to play around with some of the data available, looking at issues that had been raised in the cinemetric literature and emulating analyses that had been done – not always reaching the same conclusions. These notes, undreamed of in the not too distant past, are one consequence.

One purpose of the notes is to gather together some of the ideas that have been put forward for statistical data analysis in cinemetrics. This seemed useful, since the material is scattered and much of it is web-based. There is an emphasis on how to implement ideas using the statistical software, `R`, an open-source, state-of-the-art, highly flexible package. That a lot might be done in cinemetric analysis using modern statistical software, but largely wasn't, was an initial impression – hence the emphasis. It also struck me that some standard approaches to data analysis and presentation could be done differently and possibly more efficiently. All that's really being said here is that modern statistical software, and some methodology, can usefully add to what is already done, sometimes routinely, and that there is nothing terribly demanding about either using the

software or implementing and understanding the methodologies.

Chapter 2 provides a foretaste of what I have in mind, without the distraction of computational details. These are provided in later chapters, where there is an initial focus on graphical data analysis. Some of the methods should be familiar, others less so. While the aim has been to collect together and organise material, a critical appraisal comparing the merits and limitations of different methods is provided where appropriate. I intend to provide 'case studies' on some of the more contentious issues that have emerged in cinemetric debate at a later stage; these are alluded to in what is currently on view, but the idea has been to concentrate on more 'bread-and-butter' topics as a vehicle for illustrating the merits of the R software and its implementation.

As a 'newcomer' to cinemetrics I was quite surprised by the combative nature of some of the writing on statistics related to it. Some of this could be described as opinionated and opinion is not absent here; I've tried, however, to make it clear where an opinion is being expressed, so apologise to literary purists for what may be regarded as over-use of the first person singular in places. This is not a frivolous point; readers without a grounding in statistics who have been exposed to some of the cinemetric literature might be misled by the assertive tone of some contributions into thinking they are reading something more authoritative than deeply held opinion.

My background, as indicated above, is that of an applied statistician with a long-standing interest in applications outside the 'scientific mainstream', particularly, but not exclusively, archaeology. That I have no prior background in any aspect of film scholarship has obvious disadvantages, with the possible compensation that it's possible to look at the way statistics has been used in cinemetrics with a dispassionate eye. To that eye an important body of technique looks a little 'old-fashioned', not in the use of long-established methodology, but more in its implementation, which does not exploit the enormous and accessible computational power now available. Spreadsheet packages like Excel have their uses, are readily available, and people are comfortable with them, but they are limited as far as statistical analysis goes, compared with R[1].

At the other pole are publications which do make use of modern software and recently developed statistical methodology. These don't always supply sufficient detail to allow easy emulation of what is done on the computational front, and don't always explain methodology at a level readily comprehensible to a reader without statistical training. As far as computation goes these notes provide what I hope is sufficient detail of R code to allow ready emulation. The appendix apart, I've avoided 'mathematical' detail, including notation, as much as possible and tried to provide 'English-language' explanations. The mathematics is provided where it is an essential component of the methodology as presented in the source article, but I've also tried to explain the *ideas* involved at a simplistic level, with analogies if useful, not always attempted in the original. This may or may not help; an apology is needed once again, this time to statistical purists who may deplore the compromises involved in the 'simplification'.

My understanding of what cinemetrics is, with some of the highlights of its development, is outlined in Section 1.2. Salt's (1974) seminal paper is important in several respects. From this, and his other writings, and in the context of these notes, I want to highlight what I see as his emphasis on comparison, involving pattern recognition, and the way this can be approached using quantified data. Section 1.2.2 elaborates on this and Chapter 2 provides some examples of what I view as pattern recognition. The final section of this chapter, Section 1.3, outlines the structure of the notes as they currently stand.

1.2 Cinemetrics

1.2.1 The idea

Cinemetrics in these notes is understood to mean the statistical analysis of quantitative data, descriptive of the structure and content of films that might be viewed as aspects of 'style'. For

[1] Purpose built commercial software of the kind used for teaching statistics in universities, such as SPSS and MINITAB, is better than Excel, but still inflexible compared to R. They are typically menu-driven and you are stuck with what's on the menu, even if it's not quite what you want

practical purposes this usually means the analysis of data that describe shots in some way, for example their length or a characterisation of their type. Shot lengths (SLs) have attracted most attention in the literature; examples of type include classification by shot scale, content (e.g., action/dialogue), camera movement and so on.

Everything must have a beginning, and Barry Salt's article *The Statistical Style Analysis of Motion Pictures* published in *Film Quarterly* (Salt, 1974) is a good place to start. The paper is usefully discussed in Buckland's (2008) review of Salt's book *Moving Into Pictures* (2006). Salt's earlier book *Film Style & Technology: History & Analysis*, first published in 1983 and now in its third edition (Salt, 2009), is important. It was fairly recent exposure to this that alerted me to the existence of cinemetrics and aroused my interest.

It cannot be claimed that the subject took the world by storm. Finding relevant publications prior to the last few years is not easy. One obvious reason for this is that the data collection needed to make cinemetric analysis possible is both time-consuming and (I assume – I'm an 'end-user' and have not done this myself) tedious. In easing this problem the development of the Cinemetrics web-site (http://www.cinemetrics.lv/), headlined as a 'Movie Measurement and Study Tool Database' on its front page, seems to me to be a major landmark. It simultaneously facilitates data collection and acts as a repository for data collected and articles published on cinemetric data analysis. Quantified information is now available for thousands of films, so the absence of data is no longer a reason for not engaging in cinemetric studies.

Two other bodies of work merit attention. One is the diverse set of postings related to cinemetrics and statistics on Nick Redfern's research blog (http://nickredfern.wordpress.com/). Where possible I've preferred to reference his more structured articles – in a format associated with conventional journal publications – available as pdf files on the web. The blog, though, is the only source I've seen for the application of some statistical ideas (both old and relatively new) to cinemetric data analysis, and some of this deserves both notice and evaluation.

The other body of work I have in mind is that of James Cutting and his colleagues, mostly Jordan DeLong and Kaitlin Brunick, published in, mostly, journal articles in the period 2010-2012. I have seen some of this work labelled as 'cognitive film theory'. From a statistical point of view it is technically more demanding than most of what is covered in these notes. The applications involve fairly complicated approaches to pattern recognition in cinemetric data. The idea, as I interpret it, of pattern recognition facilitated by the quantification of 'filmic' entities is one of the major innovations in Salt's 1974 paper and later work.

1.2.2 Pattern recognition in cinemetrics

It is convenient to begin by quoting, at a little length and with comment, from Buckland (2008), a sympathetic reviewer of Salt's work. References are from pages 22-24 of his paper. Buckland suggests that for Salt

> style designates a set of measurable patterns that significantly deviate from contextual norms. He, therefore, has to establish those norms before he can define how a director deviates from them

and

> The purpose of tables and charts is not only to display the data itself, but also to show patterns and exceptions in the data. ... Whereas tables report actual figures, charts are designed to leave a general impression, rather than communicate exact numbers. ... The representation of data in charts is more visual, for it immediately reveals patterns and exceptions in the data.

In context, Buckland is referring to Salt's (1974) focus on directorial style, relating it to auteur criticism. The important idea, of more general application, is that of pattern recognition and comparison. The idea of objective *comparison* is central to Salt's thinking and quantification is central to this. Buckland recognises that

> the data [Salt] collects on individual directors can be recontextualized into broader frameworks to characterize the style of films genres, national cinemas, or historical periods, which Salt carried out in *Film Style and Technology.*

Salt's (1974, p.14) original thoughts are worth remembering

> To establish the existence of an individual formal style in the work of a director, it is necessary to compare not only a sufficient number of his films with each other, but also – which is always forgotten – to compare his films with

films of similar genre made by other directors at the same time. ... An even more absolute norm for any period is really needed as well, to give a standard of comparison that reflects the technical and other constraints on the work of filmmakers at that time and place ...

This was written with reference to directorial style but, in Buckland's word, can be 'recontextualised'. The strong emphasis on comparison is there, as is the need to establish norms, which I'm interpreting as 'pattern'. Comparison can operate at various levels, most simply in the comparison of some aspect of a small body of films, such as their shot length distributions. A more challenging problem is to identify patterns across a large body of films.

While I find it convenient to think that a lot of cinemetric data analysis can be viewd as pattern regognition, it is difficult to define precisely what I mean by 'pattern'. This is dealt with, at least partially, in Chapter 2 which provides a collection of examples, inspired by what has been done in the literature where what I'd regard as 'pattern recognition' is going on.

1.2.3 Lies, damned lies and statistics (and fear)

The first six chapters (and the preface) of Salt (2009), are largely occupied with an entertaining if possibly over-the-top polemic against the iniquities of 'film theory', where it militates against (I think this is the argument) the pursuit of objective 'scientific' study and analysis of film. It reminded me, in some ways, of arguments that raged in archaeology between the 1960s and 1990s.

About the time that Salt (1974) was published, archaeology was in the throes of a 'quantitative revolution' that had begun in earnest in the 1960s[2]. The ideas of quantification in archaeology were initially labelled the 'New Archaeology' and came to be known as processualism. The more philosophically minded protagonists explicitly linked their ideas to (logico-deductive) positivism as expounded, for example, in the work of Hempel (1965), attempting to explain variation in the archaeological record in terms of law-like relationships. Some protagonists were rather evangelical in their promotion of 'scientific' archaeology and made seriously inflated claims about what quantification could achieve. As Brandon, in the preface to Westcott and Brandon (2000), noted 'during its heyday, statistics had been waved above archaeologists' heads as an "answer" to dealing with a multitude of archaeological problems'.

As might be expected there was a theoretical reaction, from the 1980s on, often labelled post-processualism, a generic term for a diverse and incompatible set of 'philosophies'. This embraces different forms of interpretive archaeology and some of the same -isms (structuralism, post-structuralism, Marxism etc.) and thinkers that Salt (2009) castigates. On both sides of the argument as much heat as light, if not more, was generated, but things settled down. As Brandon concluded, 'after much yelling and arm-waving, most agreed that statistics were not an answer in themselves but [were] an extremly important tool available for archaeological use'.

There are still those who regard the application of scientific methodology, and more specifically statistical analysis, to archaeological understanding as de-humanising, demeaning, and inappropriate for 'humanistic' study. This is an extreme position but not a caricature. That this view can be deeply felt is not in doubt; the suspicion exists that 'ideology' can be worn as a 'principled' philosophical cloak behind which fear and ignorance of matters mathematical and statistical is concealed.

There may or may not be analogies here with the place of cinemetrics in film scholarship. You can't do much about ideology; fear, if acknowledged, can be confronted and dealt with. My general attitude to the use of statistics by non-specialists is that you should dip your toe in the water and at some point jump in. Most will learn to swim well enough for their purposes (though it has to be acknowledged that 'collateral damage' in the form of drowning, can occur). You need the motivation to learn to swim; not everyone has it and they are usefully employed doing other things. This, so long as swimming is tolerated, is fine.

From the perspective of the 2010s I don't share the pessimistic view that Salt (2012) had about what his intended audience in 1974 could reasonably be expected to cope with. The mathematics needed for some statistical methodology can be complex but, conceptually, not the idea that

[2] A review of some of the statistical highlights is provided in Chapter 1 of Baxter (2003).

often underpins an applied statistical technique. Computational ease is the key to the successful implementation of (possibly) mathematically complex, conceptually simple ideas and this is not the problem it once was. These notes are based on the belief that if you understand an idea, and it is easy to apply it, then it is straightforward. You don't need to worry about the mathematical detail, and any complex computer programming necessary has been done for you.

1.3 Structure of the notes

Chapter 2 highlights some applications of what I regard as applications of pattern recognition in cinemetrics. For the most part a more detailed exposition of the ideas involved, and discussion of computational matters, is reserved for later chapters; this is just a 'taster'.

Practical implemention is based on the open-source statistical software `R`, introduced in Chapter 3. `R` has almost the status of an industry standard in the recent applied statististical literature, and a vast amount of material on it, much web-based, is available. All that's attempted here is to provide enough detail to allow readers unfamiliar with the software, get their data into it, and access useful 'add-ons' in the form of analytical tools not bundled with the package.

It is explained how to import data from an Excel file, since this is likely to be familiar to most readers. Acquaintance with the Cinemetrics web-site (http://www.cinemetrics.lv/) is also assumed. This give access to an enormous amount of data that is easily downloaded as Excel files that, with minimal editing, can be imported into `R`. If you know what you're looking for it can take less than five minutes from going into the web-site to beginning data analysis in `R`.

The Cinemetrics site is also the source for a number of articles referenced in the notes. Where these are dated they are listed, in a conventional way, in the text (e.g., Salt (2010)) and bibliography; if not dated the URL is given in the text. It may sometimes be necessary to search through several contributions with the same URL to find that intended. Fairly frequent reference is made to the 'Cinemetrics debate', at http://www.cinemetrics.lv/dev/on_statistics.php, without repeating the URL in every case. This consists of a collection of articles on aspects of cinemetric data analysis that have previously aroused occasionally contentious debate.

Chapter 4 is a first introduction to how `R` can be used to produce descriptive statistics of the kind listed when you access the details for a film in the cinemetrics database. This includes statistics like the median and average shot length (ASL), the 'proper' use of which has attracted a surprising amount of comment in the cinemetrics literature, as the Cinemetrics debate testifies. I have views on this, but detailed commentary is deferred to a later section (not written at the time of writing this), so the section is really just a brief listing of statistics that have been used in the literature.

The main focus in the notes as written so far, Chapters 5 to 7, is on graphical methods of analysis. Others will follow, including one on shot-scale analysis. There is an attempt to be reasonably comprehensive – meaning I've included most of the things I've seen which admit reasonably uncomplicated treatment. In a sense it's a 'how to' guide, providing enough information on how to produce the graphs in `R` to admit emulation.

At one level the aim of comprehensiveness involves a lack of discrimination; several of the methods examined attempt the same kind of thing and, in practice, will often give similar results. Where an idea, in its cinemetric context, seems novel and possibly not well-known I hope I've given credit to the originator and presented the idea as intended. Some passages, paticularly Section 7.8, adopt a more evaluative approach. It's possible some of the ideas presented here are, in cinemetric terms, new, but whenever I think this I usually find similar ideas have been investigated in some previously unexplored corner of the web.

Notation and mathematical formulae are unavoidable is some instances, particularly when discussing the work of others, but as I'm claiming that a user can effectively apply statistics without needing a deep understanding of the more complex mathematics involved I've tried to keep it to a minimum. The main exceptions are confined to the appendix, of where the lognormal distribution, and the idea of logarithmic transformation of data, are of most importance.