

Comparing cutting patterns – a working paper

Mike Baxter, 16 Lady Bay Road, West Bridgford, Nottingham, NG2 5BJ, U.K.
(e-mail: michaelj.baxter@btconnect.com)

1 Introduction

The aim of the present working paper is to explore ways in which patterns of cutting across a body of films might be averaged to obtain a ‘typical’ cutting pattern. The problem is addressed explicitly in a post by Keith Brisson on the Cinemetrics Discussion Board¹ and illustrated, with applications to bodies of films of D. W. Griffith and commentary by Yuri Tsivian in *Question 3: Looking for Lookalikes* and one of the Cinemetrics Labs².

The main problem in effecting this kind of averaging is that films have different lengths, whether measured as time or number of shots. What Brisson does is to discretize the data by partitioning each film into an equal number of intervals; measuring the average shot length (ASL) in each interval; and then averaging these ASLs across films for each interval to obtain a set of ‘mean partition ASLs’ that can be plotted against partition number to get a picture of the typical cutting profile. Because not every interval may contain a cut or complete shot, this idea has to be implemented by introducing the idea of a ‘fractional shot’ in order to provide a count for each interval on which an ASL calculation can be based.

As a number of commentators on the Cinemetrics site have noted, ASL is the inverse of the cut-rate or cutting density which, across the length of a film, is just the number of cuts per whatever unit of time one chooses to use. A fast (large) cut-rate corresponds to a short (small) ASL, ‘large’ and ‘small’ being relative concepts. Barry Salt, in his comments on Brisson’s post, exploits this relationship. Using Brisson’s idea of partitioning, for a single film all that is needed is to count the number of shots in each partition. To average across partitions these counts/densities need to be ‘normalised’ to take into account the differing number of shots per film; Salt does so by dividing counts by the average shot density across the partitions.

Salt uses this idea, in passing, to comment on the investigation by Cutting *et al.* (2011) into the hypothesis that films tend to have four-act structures. The idea here is that if films have four-act structure with acts of approximately equal length *and if* this is reflected in cutting patterns (not a logical necessity) then this should be reflected in clear changes in patterns at the act boundaries. Salt, *contra* Cutting *et al.*, found no such evidence, and the latter authors later revised their claims (Cutting *et al.*, 2012). The main problem was that the averaging procedures used unintentionally imposed structure on the pattern of the kind it was independently supposed to investigate.

Rather than estimating densities after discretizing the time interval it is possible to treat time as a continuous variable and apply some form of density estimation to the cut-points. Redfern (2013)³, and in earlier work on his research blog, used

¹http://www.cinemetrics.lv/topic.php?topic_ID=356

²<http://www.cinemetrics.lv/lab.php?ID=119>

³http://www.cinemetrics.lv/dev/on_statistics.php - his contribution to *Question 2: What do lines tell?*

kernel density estimates (KDEs) to compare the variation in cutting patterns across two Astaire/Rogers musicals, using two separate graphs. The trick here is to scale cut-points by dividing by the film length so that ‘time’ ranges over the interval $[0, 1]$ and is comparable for both films. Baxter (2012)⁴ developed the idea to compare densities for several films in one graph, looking more closely at the algorithmic details involved. The idea is developed further here to compute averaged profiles.

The Brisson methodology presents averaged profiles in terms of ASLs, whereas the other methods use the inverse representation involving cutting densities. If an ASL representation is preferred then a continuous-time analogue to the use of discretized time would be to model shot length variation directly using parametric or non-parametric regression models, an idea not explored further here.

All these different methods have their pros and cons. The main purpose of the present argument is to begin to investigate what these are. For illustration the sample of D.W. Griffith films from his Biograph period, listed in the Cinemetric Lab previously referenced, are used. There is only one 1908 film, with few shots, listed there, and this is omitted from analysis, as are two 1913 two-reelers with rather more shots than the remainder⁵. Of the remaining 58, two are 1913, with the others almost equally split between the years 1909-1912.

2 Methods

2.1 Notation

Divide a single film of n shots, $\{x_t\}$, with a total film length $T = \sum x_t$, into C intervals of width $I = T/C$. Let $t_i = \sum_{t=1}^i x_t$ be the i th cut-point, $i = 1, \dots, n$; let $b_c = cI$ be the upper boundary of the c th interval, $c = 1, \dots, C$.

2.2 Salt’s method

This is probably the simplest to explain of the methods that have been proposed. The basic idea is:

1. Count the number of cut-points, n_c , in the interval $(b_{c-1}, b_c]$.⁶
2. Convert to densities $p_c = n_c/n$, and ‘normalize’ as Cp_c .⁷
3. For a sample of R films and for each interval, average across the films to get $C\bar{p}_c$.⁸
4. Plot $C\bar{p}_c$ against c and smooth the plot using the method of choice.⁹

⁴Chapter 8 in <http://www.mikemetrics.com/#/cinemetrics-data-analysis/4569975605>

⁵*The Mothering Heart* and *The Battle of Elderbush Gulch*.

⁶That is, a cut exactly on an interval boundary goes to the left.

⁷The normalization isn’t necessary for a ‘once-off’ application, but is needed if results using different numbers of partitions are to be compared.

⁸More formally, let \mathbf{D} be the $R \times C$ matrix of densities with typical element p_{rc} , $r = 1, \dots, R$, $c = 1, \dots, C$, introducing the extra subscript to distinguish films. The \bar{p}_c are the column means.

⁹Salt illustrates the idea using 6th degree polynomial smoothing; an alternative method used here is non-parametric loess smoothing.

What results is a plot of (averaged and normalised) cutting density or cut-rate. Since the cut-rate is the inverse of the average shot length (ASL), reversing the y -axis of the plot should resemble a plot based on ASLs. This reversal is applied in the illustrations to follow unless otherwise stated. Figure 1 shows the results of applying Salt’s method – henceforth called ‘discrete density estimation’ – to the sample of Griffith’s films, for four different partitions.

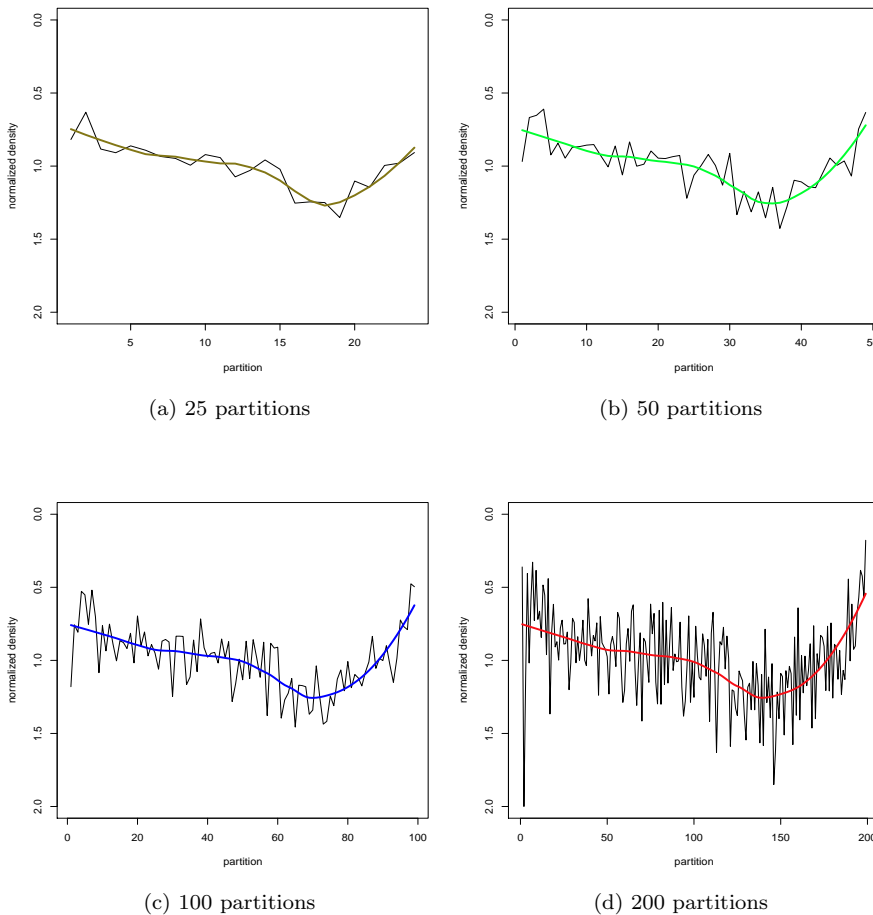


Figure 1: *Results of applying discrete density estimation to the Griffith sample for different numbers of partitions.*

The similarity of the smoothed fits is marked; if overlaid, and adjusting for the different partition sizes, they are almost identical (not shown)¹⁰. The choice of partition size would thus seem to have little impact on the smoothed pattern, but does affect the partition averages to which the smoothing is applied. This is most evident for the finest partition, of 200, where the second of the intervals used shows a higher density than elsewhere; the same phenomenon is apparent, though less

¹⁰If the end of the film is counted as a cut then the final interval in a partition includes a cut for every film and, particularly for the finer partitions, the interval has a higher average density than others. It has been omitted from the plots. On another technical note, loess smooths using the defaults in **R** were applied except that a span of $1/2$ was used – see Baxter (2013)¹¹ for a discussion of this.

obviously so, using 100 partitions. This phenomenon has previously been observed using the Brisson method, and is discussed in that context in Section 2.3.

Detail apart, the general picture is of a slow beginning, a gradual increase in pace to about two-thirds to three-quarters of the way through, then a slowing-up towards the end. It can be asked if the pattern is real. For a partition of 200 the green lines in Figure 2 give an indication of what might be expected if shots were randomly distributed through the film. The idea is an extension of a suggestion of Salt's¹². For each film, shot order was randomised and the same algorithm used to generate the patterns shown, this being repeated 10 times. It is clear that the real pattern departs significantly from what is to be expected from randomness¹³.

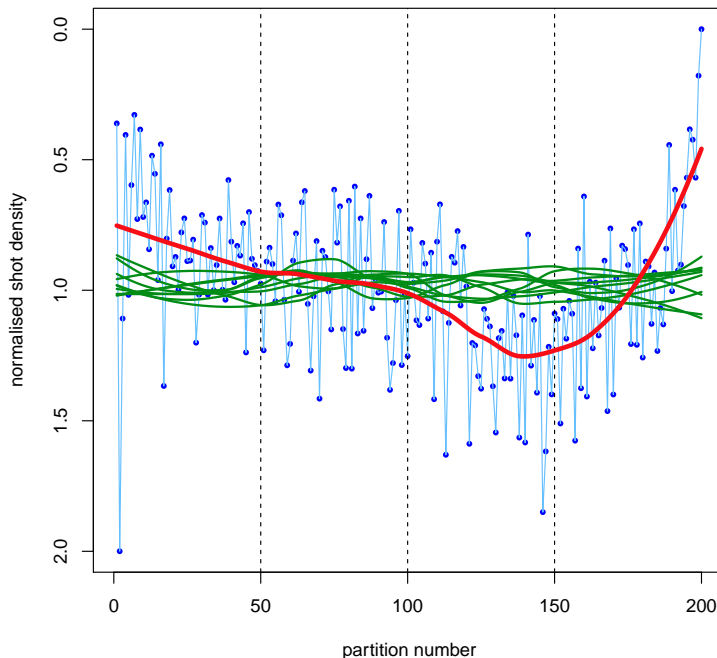


Figure 2: *An application of the discrete density estimation to the Griffith sample using a partition of 200. The green lines are obtained from a randomisation procedure described in the text*

There will obviously be considerable variation of individual films about the averaged pattern. It is possible, in principle, for no films to have a pattern that corresponds to the average. This is discussed further when looking at the use of (continuous) density estimates, but a ‘taster’ is provided by Figure 3 which, for a partition of size 100, shows the results for the calendar years 1909-1912 to which 56/58 films are dated.

The majority of films are one-reelers with a median length of 14 minutes. The 1909 films have 19–44 shots; 1910 has 30–66, all but one of 44–66; 1911 has 26–118; 1912 has 72–135. Given the relative constancy of length, this increase over time

¹²To be found in his comments on Brisson at http://www.cinematics.lv/topic.php?topic_ID=356

¹³This idea can be, and has been, applied, to all the kinds of graph shown in this paper. The general picture is the same, and randomised results are not shown in other plots to avoid ‘clutter’.

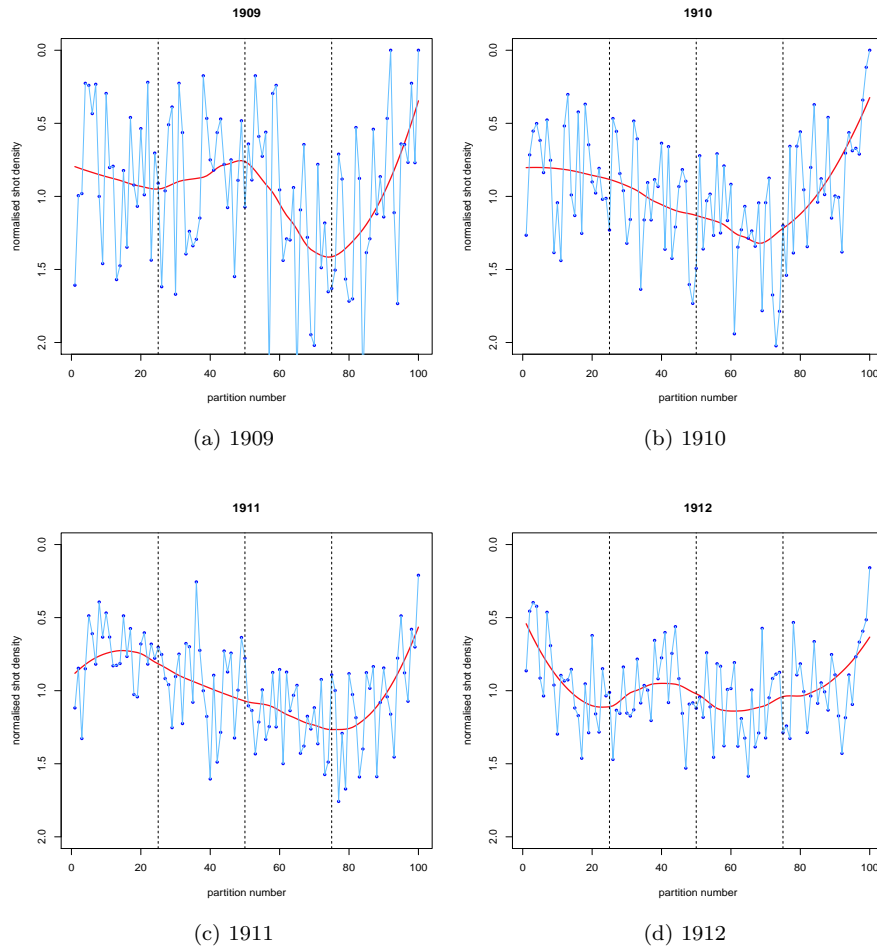


Figure 3: Discrete density estimation applied to films from different years, using a partition of 100.

in the number of shots is reflected in a general decrease in ASLs. Comment on variation between individual films is deferred, except to note that many of the 1909 films have patterns that seem unusual.

What is evident is that, even allowing that division of time into calendar years is a little arbitrary, the pattern for the later two years seems a little ‘tighter’ in terms of variation about the smoothed average. The 1912 pattern differs from others in that there is some evidence of ‘speeding-up’ about a quarter of the way through, and none of the ‘fast’ start, in the first 2% or so, of the films for which there is some suggestion in other years. Two caveats need to be mentioned. The obvious one is that only a sample of Griffith’s vast output is represented. The other, to be examined in more detail later, is the possibility that the smoothing methods used ‘interact’ with aspects of film structure (e.g., number of shots) to produce artefacts that have a mathematical rather than filmic interpretation. End effects are a potential problem here.

2.3 Brisson’s method

This works in a similar kind of way to the Salt method (and preceded it) but is based on ‘partition’ ASLs rather than cut-rates (densities). These are obtained as interval width divided by a fractional count of the number of shots in an interval. The fractional count, f_c , is the number of shots contained entirely within the interval plus the fractions of the, at most, two shots either side of these that are also included in the interval¹⁴. It is of the form $f_c = n_c + e_{lc} + e_{uc}$, where e_{lc} and e_{uc} are the fractions contributed by shots included, but not entirely so, at the lower and upper ends of the interval.

The ‘Salt method’ is based on the counts, n_c of cuts in an interval, scaled to normalized densities which are then averaged across films within each partition of width I . The density is the inverse of the ASL so in principle an analysis in terms of ASLs can be based on the I/n_c . This doesn’t work in general because a lot of the n_c are zero; hence the need to introduce the idea of fractional counts. If cut-points are sufficiently dense and/or interval widths sufficiently large, so that each interval for each film contains at least one cut-point, then averaging partition ASLs, calculated as I/n_c , should give sensible results.

The Brisson method, though simply enough described, is not simply programmed and I have not yet done so. Applications of the methods to data similar to that used here can, however, be seen at the Cinemetrics Lab previously noted and in Yuri Tsivian’s *Question 3: Looking for Lookalikes?*. The general pattern is the same as in Figure 2 and in particular what Tsivian has termed the ‘curiously long beak’, indicating apparently faster cutting at the start of the film, is apparent in both analyses (looking at the detailed rather than smoothed plot in the figure). It has been suggested that this might be attributed to a tendency to use short titles at the start of a film; Tsivian presents analyses bearing on this in *Question 3: Looking for Lookalikes?*. An alternative explanation is outlined below, based on the density rather than ASL analysis.

In Figure 2 the ‘beak’ is most pronounced in the second of the 200 intervals, that is, within the first 1% of the film and more specifically the 0.5-1% interval. There are six films with just one shot in the first interval, the rest having none. There are 31 films with shots in the second interval; one with two and the rest with one. Two of these films also have a shot in the first interval, so that for 29 films (half the total) the first shot occurs in the second interval while 35 films have their first shot occurring in the first two intervals (first 1%) of the film.

A very heuristic argument is as follows. Assume all films are 14 minutes long, or 8400 seconds, the first 1% of a film therefore occupies 16.8 seconds, and the second interval is between 8.4 and 16.8 seconds. The ASL of a film (the expected shot length for a single shot) varies between 6.4 and 43.3 for the films in the sample. Of the 58 films, 39 have an ASL less than 16.8 while 30 have ASLs in the interval 8.4 and 16.8 seconds. These numbers are fairly consistent with those observed in the first two intervals.

Imagine the films as participants in a race which starts at the same point and ends in a dead-heat for all participants. The average ‘step-size’ is the ASL and if

¹⁴This might be zero or one, rather than two shots, depending on where cut-points, t_i , lie in relation to shot boundaries, b_c .

this corresponds to the first step there will be some bunching in the second interval, as observed. All competitors start at the same point but their second step does not; allowing for random variation and differing step sizes it is to be expected that ends of the second and third steps etc. will be spread out over more intervals with less bunching.

This is, as stated, a very heuristic argument designed simply to suggest that the observed ‘beak’ might be at least partly a function of the fact that all films start at the same point and have different ASLs which, however, lie within the first 1% of a film in over half of the envisaged sample of equal film lengths above. A more refined argument based on actual film lengths is possible, but not pursued in detail here. Earlier films tend to have longer ASLs and tend to be a little shorter, though the latter effect is – I think – less pronounced. This would lead one to expect that the incidence of early bunching is greater for later films and this is the case. For each of the years 1909-1912 the percentages of films with the first cut in the first 1% of the film are 46%, 53%, 64% and 79% (with sample sizes of 13, 15, 14, 14).

This does not preclude the possibility that ‘fast’ beginnings are also associated with introductory titles, as has been suggested in discussion of analyses of the Griffith data. Since 44/58 of the films have an introductory title of some sort, scope for detecting ‘significant’ associations is limited. There is, nevertheless, evidence of one. If the presence or absence of an introductory title is cross-tabulated with whether or not it occurs in the first 1% of the film a chi-squared test rejects the hypothesis of no significance at the 2% level (p -value = 0.013). Only 4/14 films with no introductory title have a shot in the first 1%, compared with 31/44 of shots with a title that does so occur. There is also an association with date; only just over half of the 1909 films have introductory titles, whereas most of the films from other years do have such a title¹⁵.

2.4 Continuous density estimation methods

The two approaches described so far ‘discretize’ time by dividing the lengths of films into an equal number of intervals and average either cutting densities or ASLs within intervals across films. The intervals differ in length from film to film, but correspond to equal proportions of a film.

An alternative approach is to treat time as a continuous variable and apply a density estimation method to the cut-points. The idea is illustrated in Redfern’s (2013)¹⁶ Figure 11 using kernel density estimates (KDEs). These are used to visually compare cutting patterns for two films across two graphs, standardizing for lengths by dividing by T . Baxter (2012, Section 8.3)¹⁷ adapts the idea to compare cutting patterns across several different films on the same graph. The idea is extended in this paper to the construction of averaged patterns.

KDEs are obtained by associating each cut-point with a density defined by a probability distribution that ‘smears’ the density in a symmetrical fashion about the point. The spread of the smear is dictated by the *bandwidth* chosen for the

¹⁵The 58 films used here were recorded by different people who classified shots differently. The only distinction made here is between titles of some kind and other shots.

¹⁶http://www.cinemetrics.lv/dev/on_statistics.php - his contribution to *Question 2: What do lines tell?*

¹⁷<http://www.mikemetrics.com/#/cinemetrics-data-analysis/4569975605>

distribution, and will extend beyond the limits of the actual film for points close to those limits extending, in particular, into negative values for shots near the start of a film.

In principle the KDE is obtained by summing the height of the smears at all points between the minimum and maximum values obtained. In practice these sums are obtained at a finite number of points which, in the software used here, R, can be controlled by the user¹⁸. This does not eliminate the problem caused by ‘overspill’ at the ends. The user can control the boundaries of the plotting positions so that the KDE is plotted between 0 and the length of the film at a chosen number of points. The problem of biased estimates near the lower ‘physically determined’ boundary remains, and is problematic to deal with. For present purposes localised polynomial density estimation was used, which is much less sensitive to end effects (Venables and Ripley, 2002, p.132).

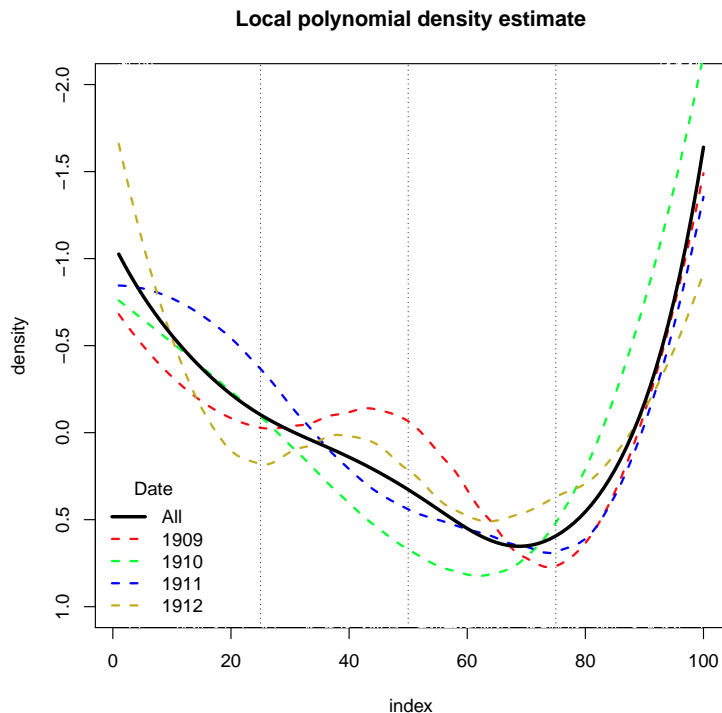


Figure 4: *Standardised continuous density estimates for the Griffith sample and subsets of it, based on cut-points.*

Since the user can control the number of equally spaced plotting positions, using 100 with estimates restricted to the range $[0, T]$ is similar to using 100 intervals in the discrete density estimation method¹⁹. In particular, densities can be estimated for each film separately and then averaged across films. Figure 4 is obtained for the Griffith films, where separate estimates for the films from 1909-1912 are also shown. The estimated densities have been standardised to a mean of 0 and variance of 1

¹⁸The R default is 512 points.

¹⁹The choice of number of plotting positions has little effect on the results; choice of bandwidth is much more important. Where defaults in R were not used this is noted in the text

for each film to remove differences attributable to the differing numbers of shots.

Compared to previous figures it can be seen that the general pattern, overall and for separate years, is broadly similar. The choice of number of plotting position, within reason, has no effect on the plot, so 100 is convenient; it is analogous to using 100 intervals in density analyses based on discretization and the scale can be read as corresponding to the elapsed percentage of a film. Previous analyses averaged across intervals and then smoothed the averaged values. In the present example smoothing takes place for individual films and averaging is across plotting positions without further smoothing. This means that fine detail is lost because of the initial smoothing, which eliminates the appearance of a ‘fast’ beginning; comparisons also being affected by the boundary problem to which the density estimate is subject.

3 Further analyses

As far as I can tell there is a general view that a lot of films start off relatively slowly and slow down a bit at the end, so that the faster ‘action’ occurs in the middle of the film, often in the later stages to stop the less dedicated spectator from falling asleep. This is obviously an over-simplification and depends a bit on genre and suchlike factors. The idea, if it is one, does raise the spectre that the pattern observed thus far for the Griffith films is a ‘generic’ one to be expected of any decent film that aspires to maintain audience interest²⁰.

To obtain some reassurance that this is not the case it is therefore useful to examine other bodies of films. Salt illustrated the idea of averaged discrete density estimation using the 1935-1955 films from the 1935-2005 database of ‘Hollywood style’ films used in Cutting *et al.* (2010). This is emulated here, except that a small number of films with SLs recorded as zero or negative have been excluded, leaving 44 films of all genres in total. Figure 5 shows the result of the discrete density estimation analysis, the loess smooth being obtained using with the `loess` function in R with defaults, apart from a `span = 1/3`. The estimates were obtained using 200 partitions in this and later analyses, the x-axis being scaled to express this in terms of the elapsed percentage of a film. For comparison with Salt’s analysis a 6th degree polynomial smooth is also shown.

Salt interpreted his analysis as showing little obvious structure, contradicting the idea explored in Cutting *et al.* (2011) that their sample of films provided support for the hypothesis of a common four act structure with acts of equal length. Intriguingly the loess smooth, which is less constrained than the polynomial smooth, does offer more scope for interpretation. There are turning points at the quarters – the kind of phenomenon Cutting *et al.* were looking for – though whether this can be interpreted in terms of act structure of the kind they envisaged is beyond the scope of present discussion. For present purposes it suffices to note that the pattern, however interpreted, is distinct from that of the Griffith films.

Pursuing analysis of the Cutting *et al.* data for its own sake, it might be objected

²⁰To be a bit more precise about this, there is considerable variation between films both in the amount of ‘excitement’ – equated here with faster cutting – where it occurs exactly, and how often. We are, however, talking about *averaged* patterns where individual variation is smoothed out, so if we allow that it exists but that film-makers have a common urge to ‘speed things up’ in its later stages, is the ‘Griffith pattern’ inevitable?

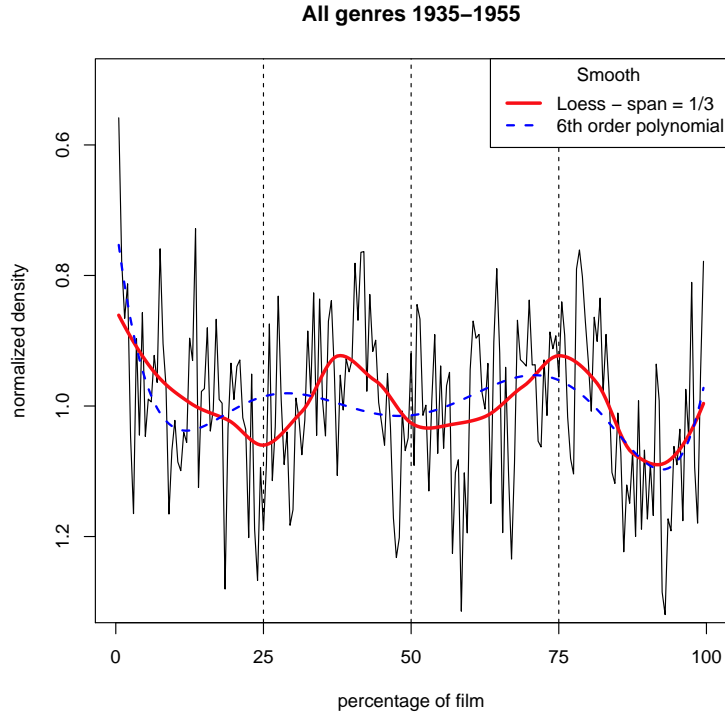


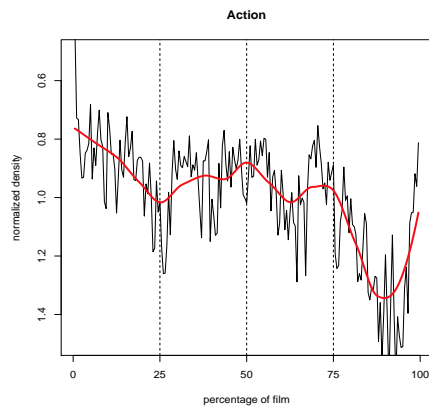
Figure 5: Averaged and smoothed discrete density estimates, using two different smoothers, for films from the years 1935-1955 (all genres).

that it does not make sense to lump markedly different genres together in one analysis. This issue is partially addressed in Figure 6 where results for four different genres are shown for the years 1935-2005.

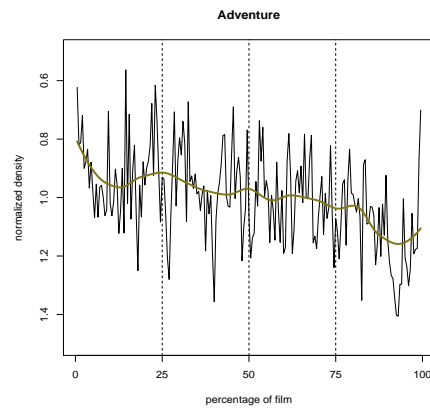
That there are differences between genres is evident. Action films stand out as distinct, the increased shot density in the final quarter being particularly evident. This is not a great surprise, the features exhibited having previously been commented on in, for example, Cutting *et al.* (2011). The other three genres are less obviously distinct from each other but that there are some differences is suggested if the smoothed plots are overlaid, as in Figure 7. Again, not too much ought to be read into the finer differences since, within genres, there may be temporal differences. The distribution of genres across years is also not uniform, since action films are over-represented in the later years and drama in the earlier ones²¹.

As an example of what can happen if some control is exercised over temporal factors Figure 8 is presented, for action and drama films classified as ‘early’ and ‘late’. The 26 action films have been split to provide roughly equal numbers in the two periods used (12 and 14). There are rather more drama films, 47, which permits greater temporal separation into two groups of 20 and 18, omitting films from the mid-1950s and 1960s. There is little difference in the pattern for action films across time. Drama is more interesting in that later films show a flatter profile

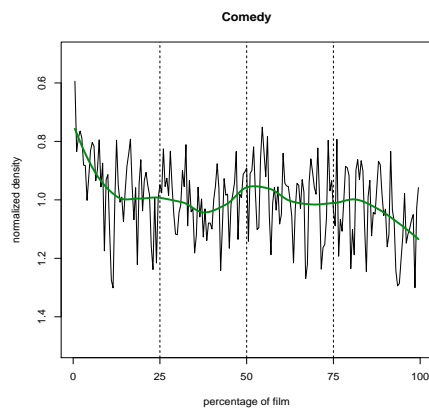
²¹‘Over-representation’ is relative to what would be expected if an attempt had been made to select the sample such that genres were equally spread across the years sampled, and this was not attempted. The actual sample may be more representative of the changing popularity of genres across the years.



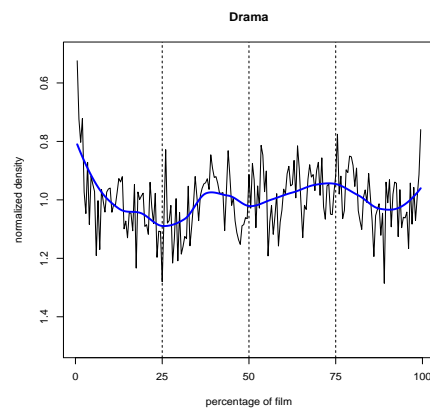
(a) Action films 1935-2005



(b) Adventure films 1935-2005



(c) Comedy films 1935-2005



(d) Drama films 1935-2005

Figure 6: *Discrete density estimates, averaged and smoothed, for different genres for 1935-2005.*

after about the first third of the film. The earlier films repeat the pattern seen in Figure 5 and are likely to be responsible for that pattern. One would really like a larger sample to work with, but the suggestion is that if SL patterns can be used to infer particular forms of act structure it may be limited to particular genres and periods of time.

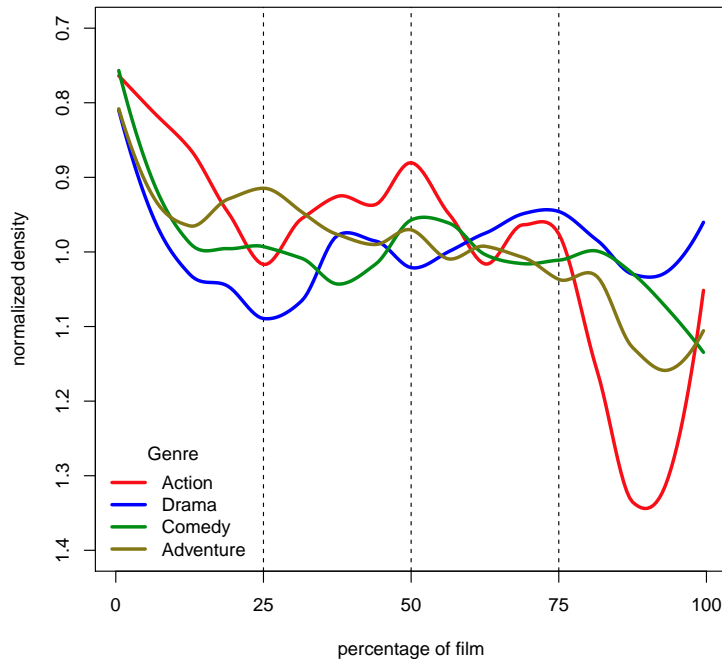


Figure 7: A comparison of the averaged and smoothed discrete density estimates across genres from Figure 6.

4 Other analytical possibilities

The different approaches investigated above have in common the structure that each film is reduced to a set of numbers (corresponding to the number of partitions or plotting positions) that are averaged to get a picture of the overall cutting or ASL pattern. It has already been noted that in principle no single film need resemble the overall averaged pattern, though the results seem sensible as broad summaries. It remains the case, though that many films will not be close to the overall pattern.

Since films are reduced to comparable sets of numbers it is possible to measure mathematically the ‘distance’ between the profiles of films, so that similar films are ‘close’ to each other in terms of distance. There are many ways in which distance might be measured. There are also a lot of statistical methods (multidimensional scaling techniques; cluster analysis) around that, given a set of distances, attempt to visualise how similar or different films are from each other. Thus, do films separate into groups where films are similar withing groups and distinct from others? Do

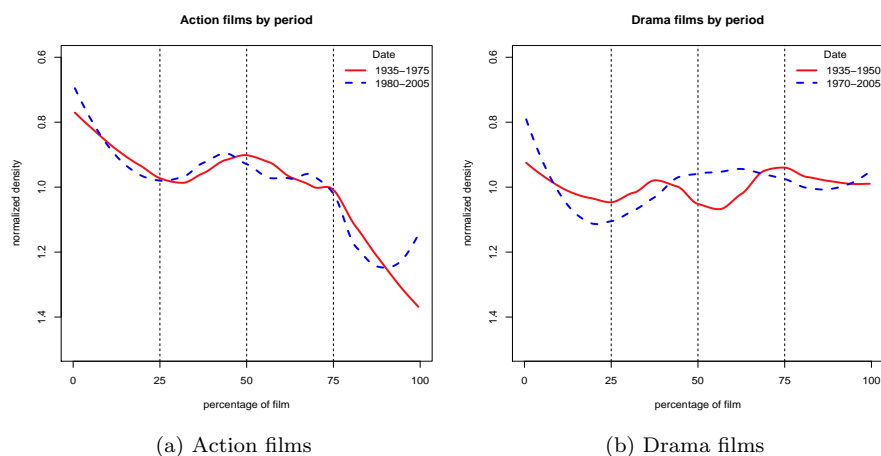


Figure 8: *Discrete density estimates, averaged and smoothed, for action and drama films genres separated into ‘early’ and ‘late’ films.*

some films exhibit reasonably ‘unique’ cutting patterns, and so on?

One way of thinking about this is that measurements for a film are equally spaced samples from a curve that defines the cutting/ASL profile of a film. Methods of *functional data analysis* exist for comparing collections of such curves. I have little practical experience of these, but application involves a lot of considerations. It is probably best to base analysis on smoothed curves of some kind, which raises the issue of what degree of smoothing should be applied. My gut feeling is that, in the first instance at least, a reasonably high degree of smoothing should be applied to avoid problems that might arise in comparing ‘noisy’ under-smoothed data.

Scaling issues are also important. Standardizing data to control for variation about some mean (density or ASL) value can be done in more ways than one but is (more-or-less) standard practice. This can be thought of as a ‘vertical’ adjustment of the scale (thinking of density or ASL plotted against time). ‘Horizontal’ adjustment, which involves transforming, or ‘shifting’, the time scale to match up – for example – major peaks and troughs in a profile, is more complicated, and it is questionable if it should be done with cutting patterns. The idea is to align major features so that the underlying pattern adjusted for shift differences is as similar as possible, but the differences the shift is intended to compensate for may be of interest in their own right.

This is not pursued here (as I have not thought this through). Figures 14-18 in Tsivian’s *Question 3: Looking for Lookalikes?*, which I had not seen when I first drafted this paper, are intriguing, though. The averaged Figure 18 is clearly interpretable but, as Tsivian notes, loses information that looks interesting. Figure 14 is rather different from Figures 15-17. These latter have 4-5 noticeable but differently positioned ‘peaks’ which ‘average out’ and disappear in Figure 18. The idea suggested in the last paragraph is that you might try and align the peaks in some way to obtain a more ‘peaky average’ that reflects the qualitatively apparent patterns. You’d want separate out films that have distinctly different patterns before doing this, and would need to define what is meant by a ‘noticeable’ peak, before

looking at the non-trivial technicalities of obtaining a useful average. The other question is, and for illustration, if you had two films each with two distinct ‘peaks’ occurring at rather different points in the films, do you want to use methodology that treats them as ‘similar’?

A crude analysis would ignore the possibility of this kind of adjustment and employ ‘standard’ methods of multivariate analysis to see if there is some kind of ‘subset’ pattern in a body of data. Full technical details are not given here, but Griffith’s 1911 films will be taken as a ‘playground’ for illustrative purposes. On the basis of a principal component analysis (PCA) the 14 films were divided into four groups as follows²².

- Group 1: *The Adventures of Billy, His Trust, The Indian Brothers, The Miser’s Heart*
- Group 2: *Fighting Blood, His Daughter, His Trust Fulfilled, Last Drop of Water, Swords and Hearts*
- Group 3: *The Battle, A Country Cupid, Enoch Arden, Heartbeats of Long Ago,*
- Group 4: *What Shall We Do With Our Old*

The singleton Group 4 has somewhat fewer shots (32) than any other film apart from *His Trust Fulfilled*. This group and Group 1 stand out fairly clearly in the analysis that defined them; the separation between Groups 2 and 3 is not as great.

Figure 9 shows the results, using discrete density estimation. There seem to be clear differences. For Group 3 the ‘action’ occurs in the middle of the film rather than nearer the end (Groups 1, 4); Group 2 has a rather flat profile. Whether this analysis is reflecting real differences would need to be judged by someone who knows the films. There is no obvious association with the number of shots and median/ASLs.

5 Discussion

This paper is precisely what it says in the title – a ‘Working Paper’ – and analysis is obviously incomplete. It raises several issues that require more detailed thought and analysis, so any conclusions are highly provisional. My current preference is for an approach based on discretizing the data. This is mainly because of the problems with boundaries associated with continuous density estimates.

If the data are discretized, then in principle patterns can be investigated by averaging either cutting densities or ASLs across partitions. The former seems a lot easier since it avoids the need to calculate fractional ASLs. The applications illustrated above focus on pattern, so that densities are standardized in various ways to eliminate effects due to the number of shots, so this means patterns can’t be interpreted in terms of ASL differences, for example. If such differences are of

²²For the record, 100 partitions and a span in the loess smooths of 2/3 were used. For the PCA the densities for each film were standardized (row standardization) to zero mean and unit variance, the columns were standardized in a similar way before applying the PCA. I haven’t looked at this in detail yet, but other ways of standardizing (e.g., omitting either the row or column standardization) gave broadly similar groupings.

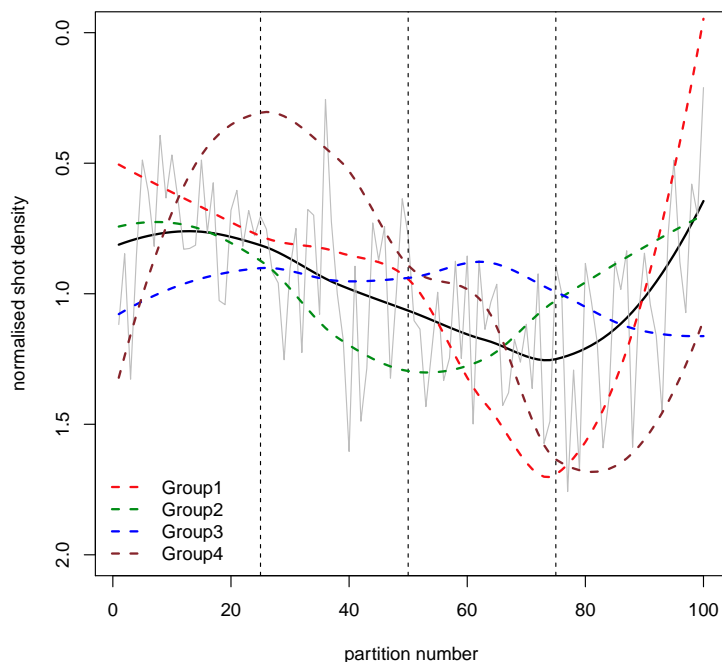


Figure 9: A comparison of the averaged and smoothed discrete density estimates across groups for Griffith films from 1911. See the text for details of grouping.

interest they can be captured by removing some of the standardization, but direct interpretation in terms of ASLs is still not straightforward.

In principle, variation in SLs (i.e. localised ASLs) can be modelled using local polynomial smoothers such as loess curves. This is straightforward enough for individual films. Averaging across films requires – I think – estimating the fitted curves at an equal number of comparable points. I can think of ways of doing this but have run into end-effect problems so far – mainly with films that have few shots and/or where the first shot is rather long. In the first instance it might be simplest to remove those films where the problem occurs to see how results compare with other approaches, but this needs to be pursued.

There is considerable variation between films, many of which do not conform well to an averaged pattern. This possibly raises questions about what such a pattern actually captures – the obvious answer is very broad trends from which there is a lot of deviation that is presumably of interest in its own right. Disaggregating the data and comparing patterns between the subsets so defined is an obvious line of development. Disaggregation by time and/or genre or other well-defined ‘filmic’ variables is also an obvious way of thinking, illustrated above. A more purely statistical/data analytic approach, broached briefly in the last section, is to use (multivariate) methodologies to try and identify sub-groups of films that are structurally similar. If this can be done – and I think doing so formally might be a difficult problem – the ‘filmic’ reasons for the similarities can be sought. I suspect,

also, that this may not be straightforward.

References

Cutting J.E., DeLong J.E. and Nothelfer C.E. (2010) Attention and the evolution of Hollywood film. *Psychological Science* **21**, 440-447.

Cutting J.E., Brunik K.L. and DeLong J.E., (2011) How act structure sculpts shot lengths and shot transitions in Hollywood film. *Projections* **5**, 1-16.

Cutting J.E., Brunik K.L. and DeLong J.E., (2012) On shot lengths and film acts: A revised view. *Projections* **6**, 142-145.

Venables, W.N and Ripley, B.D. (2002) *Modern Applied Statistics with R: Fourth Edition*, Springer, New York.