An introduction to using graphical displays for analysing the editing of motion pictures

Nick Redfern

> The best single device for suggesting, and at times answering, questions beyond those originally posed is the graphical display.
>
> John Tukey

Since a film typically comprises several hundred (if not thousands) of shots describing its style clearly and concisely can be challenging. This is further complicated by the fact that editing patterns change over the course of a film. Numerical summaries are useful but limited in the amount of information they convey, and while two films may have the same median shot length or interquartile range they may have very different densities. They are also less effective for describing changes in style over time. These problems may be overcome by using graphical as well as numerical summaries to communicate large amounts of information quickly and simply. Graphs also fulfil an analytical role, providing insights into a data set and revealing its structure. A good graph not only allows the reader to see what is important about a data set the writer wishes to convey, but also enables the researcher to discover what is important.

The use of graphical methods should be common practice in the statistical analysis of film style, and there are several different types of simple graphs that can be used. The purpose of this paper is to introduce some graphical methods for analysing the editing of motion pictures. In the next section I introduce exploratory data analysis as a framework for using graphical methods. I then outline five methods for describing the distribution of shot lengths in a motion picture and for comparing the style of two or more films. Finally I describe some simple methods for analysing time-ordered shot length data in order to analyse film editing as a dynamic stylistic system.

The material presented in this paper is drawn from several of my blog posts and draft articles from the past four years. The purpose of representing this material is to make it easier to compare and contrast different methods rather than being forced to hunt through different web pages, and I have included some references that will provide more detail on the methods discussed. One way to think of this paper is as a compendium of methods that can be drawn upon depending upon the research you wish to conduct. The methods listed in this paper do not in any way exhaust the possibilities for the graphical analysis of film editing and so you should not feel constrained by them in any way – but they are a good place to start and the only way to learn is to get some data and try it for yourselves. All of the graphs in this paper were produced using Microsoft Excel, or two *freely available* statistics programmes – **R** (2.15.0) and PAST (2.17) – so there can be no excuses for failing to use graphical methods in the analysis of film style.

Exploratory data analysis

In the 1970s, John Tukey (1977) proposed *exploratory data analysis* (EDA) as a 'practical philosophy' of data analysis that minimizes the researcher's dependence on prior

assumptions and maximises insight into the phenomenon at hand (see Vellemen & Hoaglin 1981). Tukey described the role of the researcher as that of a detective:

> Exploratory data analysis is detective work – numerical detective work - or counting detective work - or graphical detective work. ... [It is] about looking at data to see what it seems to say. It is about simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights (Tukey 1977: 1).

The goal of EDA is to discover 'potentially explicable' patterns in data (Good 1983); with 'an emphasis on the substantive understanding of data that address the broad question of "what is going on here?"' (Behrens 1997: 131). This process depends on the researcher adopting a position of *scepticism* of methods that may obscure informative aspects of data and of *openness* to unanticipated patterns (Hartwig & Dearing 1979: 9). To this end, EDA employs a range of numerical and graphical techniques to maximise our insight into the data by revealing its underlying structure and extracting the relevant features, identifying outliers and anomalies, generating hypotheses, and testing underlying statistical assumptions. EDA places substantial emphasis on resistant and robust methods requiring few *a priori* assumptions about data and which are applicable in a wide range of circumstances.

EDA is a data-driven, bottom-up approach to film form. This is different to the way film scholars typically approach film style, which is usually a top-down, theory-driven form of analysis requiring the researcher to make *a priori* assumptions regarding the purposes and content of motion pictures that will direct which features of film form are to be considered relevant to whatever she interprets the function a film to be (see Carroll 2009). There will always be a place for top-down research on film style since the questions we wish to ask about the form of a motion picture will often be stimulated by our existing knowledge of modes of production, genres, the history of film style, other works by the same filmmakers, or by issues that interest us such as the representation of women, national cinemas, genre, and so on. However, such research will only result in a partial description of a film style, overlooking potentially interesting or confounding relations of film form. The logic of EDA is *abductive* and moves from data to hypothesis via a bottom-up process of pattern extraction (Behrens & Yu 2003). It not only allows us to answer the question we wish to ask but also suggests new questions based on what we discover. From an EDA perspective a descriptive account of a motion picture's form is *precisely* a guide to what is significant for understanding and appreciating the movie, but which avoids description as an end in itself. Beginning any study of film style with an exploratory stage both simplifies and amplifies the analytical process: 'Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step' (Tukey 1977: 3).

Graphical displays play a fundamental role in exploratory data analysis (Ellison 1993), and fulfil two functions (Jacoby 1997: 2):

- Analytical graphics: techniques employed as an integral part of the data analysis process
- Presentational graphics: displays used primarily communicating the results of the completes analysis

Far from merely reproducing a film's style, such displays make it possible to see the formal relationships across a film and differences in the formal relation between films so that we may extract interesting and unusual patterns:

> Effective visual presentations highlight interesting and unusual aspects of quantitative information under investigation. This encourages the researcher to pursue these features to identify their sources and implications for understanding the processes that are generating the data in the first place (Jacoby 1997: 7).

Graphical displays are not a substitute for the films we wish study: we use graphs to identify those features that could be interesting and then we go back to the film to see what those features are and to understand how they function in the film's formal scheme. For example, we might observe in a graph of the shot lengths of a news bulletin a handful of shots that are of much longer duration than others and wonder why this should be the case and, though the graph has brought this feature to our attention, it is only when we go back to bulletin itself that we can begin to understand what discourse elements these features are associated with and how they shape our experience of watching the news. 'Graphics *reveal* data' (Tufte 2001: 13, original emphasis), and provide an unparalleled method for open-mindedly discovering pertinent relations of film form.

Tukey (1977: v) summarised the goals of EDA in the following terms:

> A basic problem about any problem of data is to make it more easily and effectively handleable by minds – our minds, her mind, his mind. To this end:
>
> - anything that makes a simpler description possible makes the description more easily handleable
> - anything that looks below the previously described surface makes the description more effective
>
> So we shall always be glad (a) to simplify description and (b) to describe one layer deeper.

It is the philosophy of exploratory data analysis, and these two principles in particular, that informs the description of graphical methods for analysing shot length data below.

Graphics for examining shot length distributions

A shot length distribution is the complete data set produced by recording the duration of each shot in a motion picture in seconds. This distribution can be characterised using descriptive statistics such as the median shot length, the interquartile range, the five-number summary, and so on. While valuable, these numerical descriptions provide only a limited amount of information about the distribution of shot lengths in a motion picture. In this section I introduce five different graphical displays for exploring and describing data that can be produced by any standard statistical software. Using these methods we can provide a complete description of a data set with an economy unattainable using linguistic descriptions, identify key features in the data that may be explored at length, compare the distributions of two or more films to determine if they have similar styles, and to check statistical assumptions.

Histogram

A histogram is a simple nonparametric method of density estimation for a quantitative variable produced by sorting data into discrete intervals called *bins*. The frequency of data points occurring in each bin is represented by vertical columns against a scale. The values on the x-axis form a continuum and the point at which one bin ends is the point at which the

next bin begins. Gaps in a histogram indicate bins containing *no* values. To construct a histogram we do not need to calculate any descriptive statistics or make any assumptions regarding the underlying probability distribution of the data. The shape of a histogram depends only on the data, the choice of the location of the origin, and the width of the bins.

From a histogram we can identify the shape of a distribution (uni-, bi-, or multimodal, symmetrical or skewed, kurtosis); the range of the data; and the presence of outliers. Figure 1 presents the histogram of the 1830 news bulletin broadcast on ITV1 on 12 August 2011. This chart also includes a 1-D scatter plot drawn under the histogram, where each line shows the length of a shot. From this chart we can see the distribution of shot lengths is unimodal and skewed, with the majority of shot lengths less than 10 seconds in duration. There are number of shots in the range 10 to 25 seconds, with four outliers of much greater duration than others in the bulletin. Having indentified this structure in the data we can go back to the data collected to find out what features they are associated. In this example, we find that shots less than 10 seconds in duration are associated with the shots comprising the main part of a news item over which the reporter speaks; while shots greater than 10s tend to be associated people speaking directly to camera either as part of an interview or the 'kernel' that introduces each item in the bulletin, with the longest shots occurring when a reporter is speaking directly to camera, typically as part of a two-way interview (see Montgomery 2007). We could have discovered this by going through the raw data one data point at a time, but it is much quicker to identify the key features of a data set and far easier to communicate the results using a histogram.

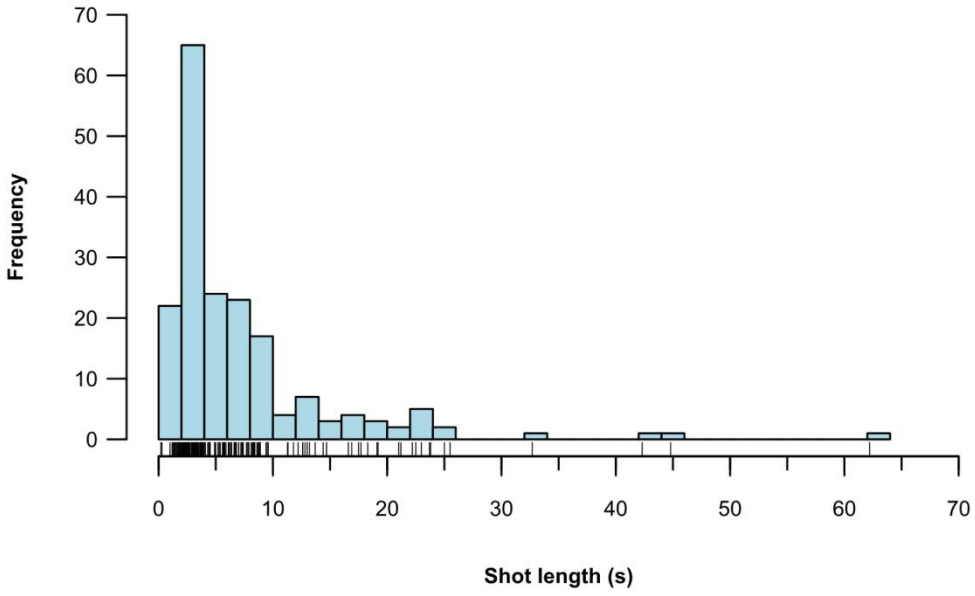

Figure 1 Histogram of shot lengths in the ITV1 1830 news bulletin on 12 August 2011

Histograms are useful for a first look at data but have their limitations:

- The shape of the histogram depends on the choice of the origin and the bin-width, and making the wrong choice can lead to flawed interpretations. Too many and too few bins obscure the structure of the data by providing too much noise or too little information.
- There is a lack of precision in describing data because the limits of the bins do not correspond to the limits of the data
- Two shot length distributions cannot be placed on the same histogram so that comparing the style of films is challenging.

Some of these issues can be resolved by using the kernel density estimate instead of a histogram, and this is discussed in the next section.

Finally, histograms should not be confused with *bar charts*, which are a similar graphical method for representing frequencies of qualitative variables (e.g. shot scales, camera movement, etc).

Kernel density estimation

The kernel density is a nonparametric estimate of the probability density function of a data set, and will show us the range of the data, the presence of outliers, the symmetry of the distribution (or lack thereof), the shape of the peak, and the modality of the data.

The kernel density is estimated by summing the kernel functions superimposed on the data at every value on the x-axis. This means that we fit a symmetrical function (the kernel) over each individual data point and then add together the values of the kernels so that the contribution of some data point x_i to the density at x depends on how far it lies from x. The kernel density estimator is

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

where n is the sample size, h is a smoothing parameter called the *bandwidth*, and K is the kernel function. There are several choices for K (Gaussian, Epanechnikov, triangular, etc.) though the choice of kernel is relatively unimportant, and it is the choice of the bandwidth that determines the shape of the density since this value controls the width of the kernel. If the bandwidth is too narrow the estimate will contain lots of spikes and the noisiness of the data will obscure its structure. Conversely, if the bandwidth is too wide the estimate will be over-smoothed and this will again obscure the structure of the data. The kernel density estimate is an improvement on the use of histograms to represent the density of a data set since the estimate is smooth and does not depend on the end-points of the bins, although a shared limitation is the dependence on the choice of the bandwidth. Another advantage of the kernel density is that two or more densities can be overlaid on the same chart for ease of comparison whereas this is not possible with a histogram.

Figure 1 illustrates this process for *Deduce, You Say* (1956), in which the density shows how the shot lengths of this film are distributed. Beneath the density we see a 1-D scatter plot in which each line indicates the length of a shot in this film (x_i), with several shots having identical values. The Gaussian kernels fitted over each data point are shown in red and the density at any point on the x-axis is equal to the sum of the kernel functions at that point. The closer the data points are to one another the more the individual kernels overlap and the greater the sum of the kernels – and therefore the greater the density – at that point.

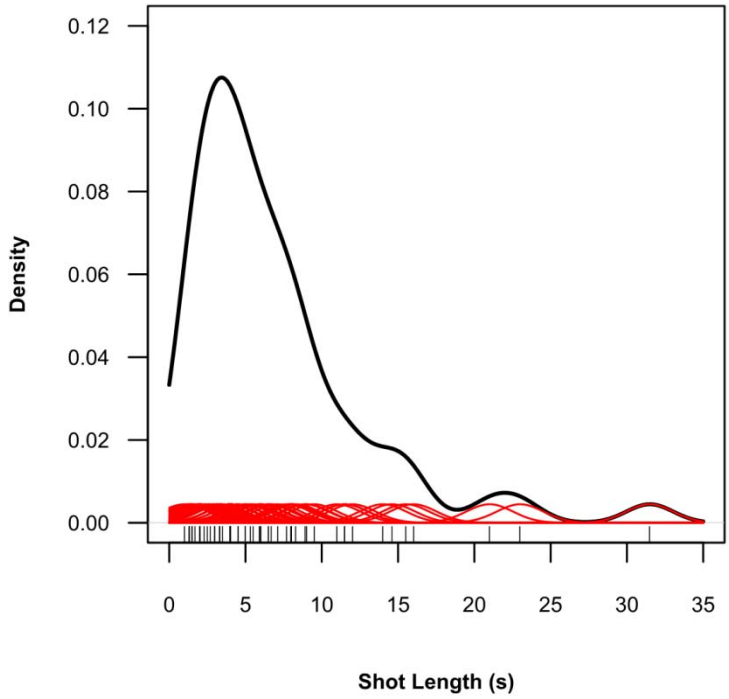

Figure 2 The kernel density estimate of shot lengths in *Deduce, You Say* (1956) showing the kernel functions fitted to each data point (N = 58, Bandwidth = 1.356)

Suppose we wanted to compare the shot length distributions of two films. Figure 2 shows the kernel density estimates of the Laurel and Hardy shorts *Early to Bed* (1928) and *Perfect Day* (1929). It is immediately that clear though both distributions are positively skewed, the shot length distributions of these two films are very different. The density of shot lengths for *Early to Bed* covers a narrow range of shot lengths while that for *Perfect Day* is spread out over a wide range of shot lengths. The high density at ~2 seconds for *Early to Bed* shows that the majority of shots in this film are concentrated at lower end of the distribution with few shots longer than 10 seconds, while the lower peak for *Perfect Day* shows there is no similar concentration of shots of shorter duration and the shot lengths are spread out across a wide range (from 20 to 50.2 seconds) in the upper tail of the distribution.

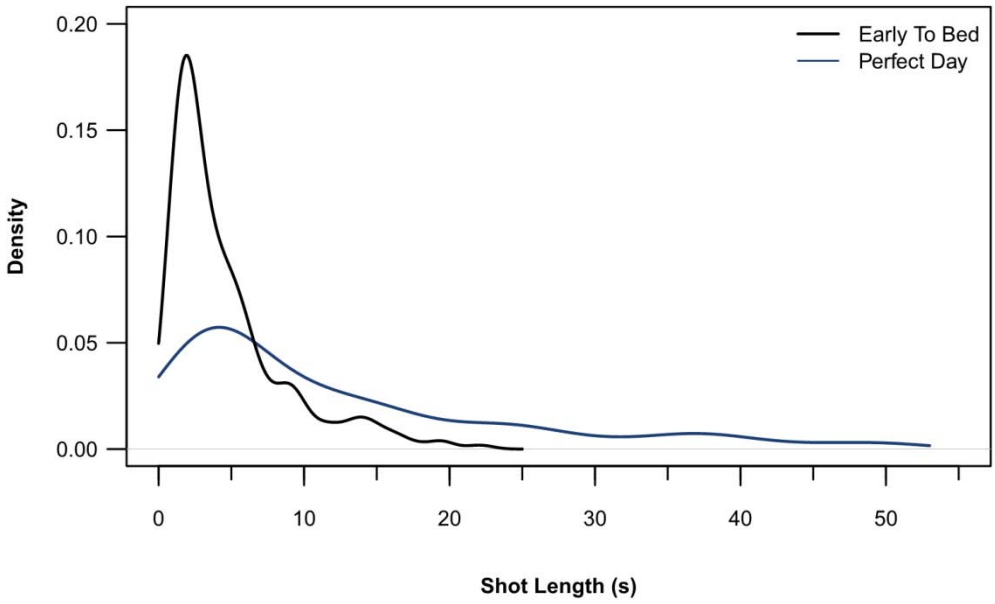

Figure 3 Kernel density estimates shot lengths in *Early to Bed* (1928) and *Perfect Day* (1929)

We can conclude that *Early to Bed* is edited more quickly than *Perfect Day* and that it shot lengths exhibit less variation; and though we could have come to these same conclusions using numerical summaries alone the comparison is clearer and more intuitive when represented visually.

Empirical Cumulative distribution function

The empirical cumulative distribution function (ECDF) gives a complete description of a data set, and is simply *the fraction of a data set less than or equal to some specified value*. The simplest method of calculating the ECDF is to count the number of shots (X) less than or equal to some value (x), and then divide by the sample size (N):

$$F(x) = \frac{\#(X \leq x)}{N}.$$

We can interpret this function in several ways: we can think of it as the probability of randomly selecting a shot with duration less than or equal to x seconds ($P[X \leq x]$); or we can think of it as the proportion of values less than or equal to x; or, if we multiply by 100, the percentage of values in a data set less than or equal to x.

For example, using the data set for *Easy Virtue* (1928) from the Cinemetrics website we can calculate the ECDF as illustrated in Table 1.

Table 1 Calculating the ECDF for *Easy Virtue* (1928) (N = 706)

Shot length (x)	Frequency (f)	f/N	ECDF
0.9	0	0.0000	0.0000
1.0	1	0.0014	0.0014
1.1	3	0.0042	0.0057
1.2	0	0.0000	0.0057
1.3	1	0.0014	0.0071
…	…	…	…
66.3	0	0.0000	0.9986
66.4	0	0.0000	0.9986
66.5	0	0.0000	0.9986
66.6	1	0.0014	1.0000
66.7	0	0.0000	1.0000

To start, look at the value of x in the first column and count the number of shots in the film with length less than or equal to that value. The first value is 0.9 but there are no shots this short in the film and so the frequency is zero. Divide this zero by the number of shots in the film (i.e. 706) and you have the ECDF when x = 0.9, which is 0 (because 0 divided by any number is always 0). Next, x = 1.0 seconds and there is 1 shot less than or equal to this value and so the ECDF at x = 1.0 is 1/706 = 0.0014. Turning to x = 1.1 we see there are three shots that are 1.1 seconds long AND there is one shot that is shorter in length (i.e. the one at 1.0s), and so the ECDF at X = 1.1 is 4/706 = 0.0057. This is equal to the frequency of 1.0 second long shots divided by N (0.0014) PLUS the frequency of shots that are 1.1 seconds long (3/706 = 0.0042) – and that is why it's called the cumulative distribution function. From this point you keep going until to reach the end: the longest shot in the film is given as 66.6 seconds long and so all 706 shots must be less than or equal to 66.6 seconds and so at this value of x the ECDF = 706/706 = 1.0. The ECDF is 1.0 for any value of x greater than the longest shot in the film.

You can get a simple graph of $F(x)$ by plotting x on the x-axis and the ECDF on the y-axis. More usefully, you can plot the ECDFs of two or more films on the same graph so that you can compare their shot length distributions. Figure 4 shows the empirical cumulative distribution functions of *Easy Virtue* and *The Skin Game* (1931), and to make it easier to see the structure of the data I plotted the x-axis on a logarithmic scale.

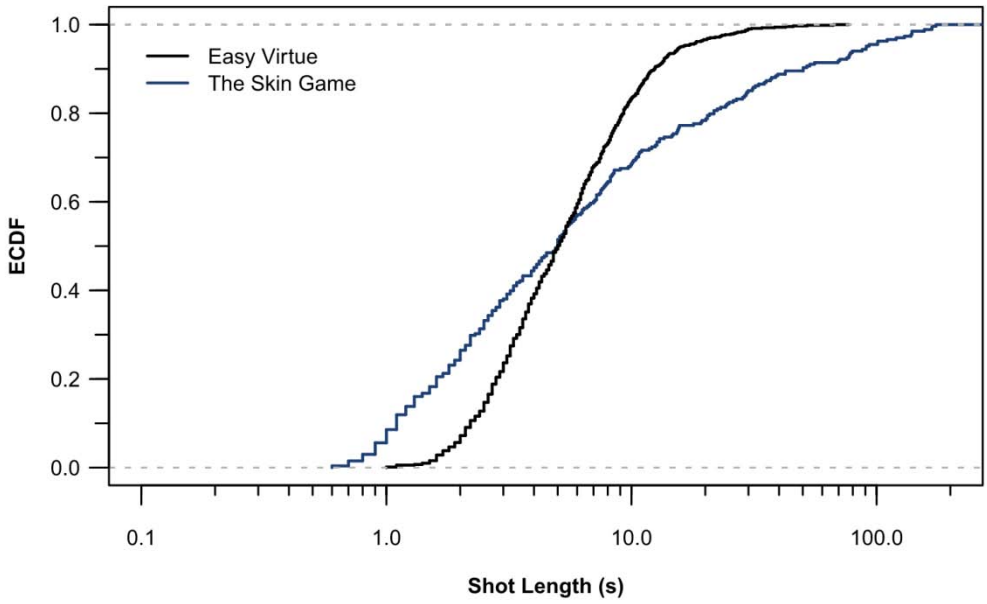

Figure 4 The empirical cumulative distribution functions of *Easy Virtue* (1928) and *The Skin Game* (1931) on a log-10 scale

We can learn a great deal about the stylistic differences between these films by comparing their respective distribution functions. First, it is clear that these two films have same median shot length because the probability of randomly selecting a shot less than or equal to 5.0 seconds is 0.5 in both films. The definition of the median shot length is the value that divides a data set in two so that half are less than or equal to x and greater than or equal to x (i.e. $P[X \leq x] = 0.5$). We might therefore conclude that they have the same style. However, these two films clearly have different shot length distributions and it is easier to appreciate this when we combine numerical descriptions with a plot of the actual distributions. A basic rule for interpreting the plot of ECDFs for two films is that if the plot for film A lies to the right of the plot for film B then film A is edited more slowly. Obviously this is not so clear cut in Figure a. Below the median shot length, the ECDF of *The Skin Game* lies to the left of that of *Easy Virtue* indicating that at those shot lengths it has a greater proportion of shots at the low-end of the distribution: for example, 25% of the shots in *The Skin Game* are less than or equal to 2.0 seconds in length compared to just 6% of the shots in *Easy Virtue*. This would seem to indicate that *The Skin Game* is edited more *quickly* than *Easy Virtue*. At the same time we see that above the median shot length that the ECDF of *The Skin Game* lies to the right of that of *Easy Virtue* indicating that it has a lower proportion of shots at the high-end of the distribution: for example, 75% of the shots in *Easy Virtue* are less than or equal to 8.3 seconds compared to 66% of the shots in *The Skin Game*. This would appear to suggest that *The Skin Game* is edited more *slowly* than *Easy Virtue*. Clearly there is something more interesting going on than indicated by the equality of the medians, and the answer lies in how spread out the shot lengths of these two films. The ECDF of *Easy Virtue* is very steep and covers only a limited range of values; where as the ECDF of *The Skin Game* covers a much wider range of shot lengths. The interquartile range of *Easy Virtue* is 5.2 seconds (Q_1 = 3.1s, Q_3 = 8.3s) indicating the shot lengths of this film are not widely dispersed; while the IQR of *The Skin Game* is 12.7s (Q_1 = 2.0s, Q_3 = 14.7s). This example is an excellent demonstration of why it is important to always provide a measure of the dispersion of a data set when describing film style. It is not enough to only provide the

average shot length since two films may have the same median shot length and completely different editing styles.

Box plots

Box-plots are an excellent method for conveying a large amount of information about a data set quickly and clearly, and do not require any prior assumptions about the distribution of the data. Analysing the box-plots of shot lengths in motion pictures we compare the centre and variation of the data, and identify the skew and the presence of outliers. They are also an efficient method of comparing multiple data sets, and placing the box-plots for two or more films side-by-side allows us to directly compare the centre and variation of shot length distributions intuitively.

The box plot provides a graphical representation of the five-number summary, which includes

- the minimum: the lowest value in a dataset,
- the lower quartile: the value that cuts off the lowest 25% of values in a dataset,
- the median: the value that divides the data set into,
- the upper quartile: the value that cuts off the highest 25% of values in a dataset, and
- the maximum: the largest value in a dataset.

The core of the data is represented by the box which includes the middle 50 per cent of the data. The difference between the upper and lower quartiles is the *interquartile range*, and is a numerical description of the dispersion of shot lengths in a motion picture. Outliers are plotted individual points.

Figure 5 presents the box plots of the three main daily news bulletins broadcast on the BBC from 11 to 15 April 2011. The box plots have been adjusted to take into account the skewed nature of the data. From Figure 5 we see that these fifteen bulletins have similar median shot lengths and similar interquartile ranges and so they have similar shot length distributions. We can also see the distribution of shot lengths in each bulletin is positively skewed, and that every bulletin has a number of outliers. This is a very efficient way of presenting a large amount of information that would be impossible to achieve using histograms; and, if not impossible, then certainly very difficult to understand putting fifteen data sets into a single graph using the kernel density estimate or ECDF.

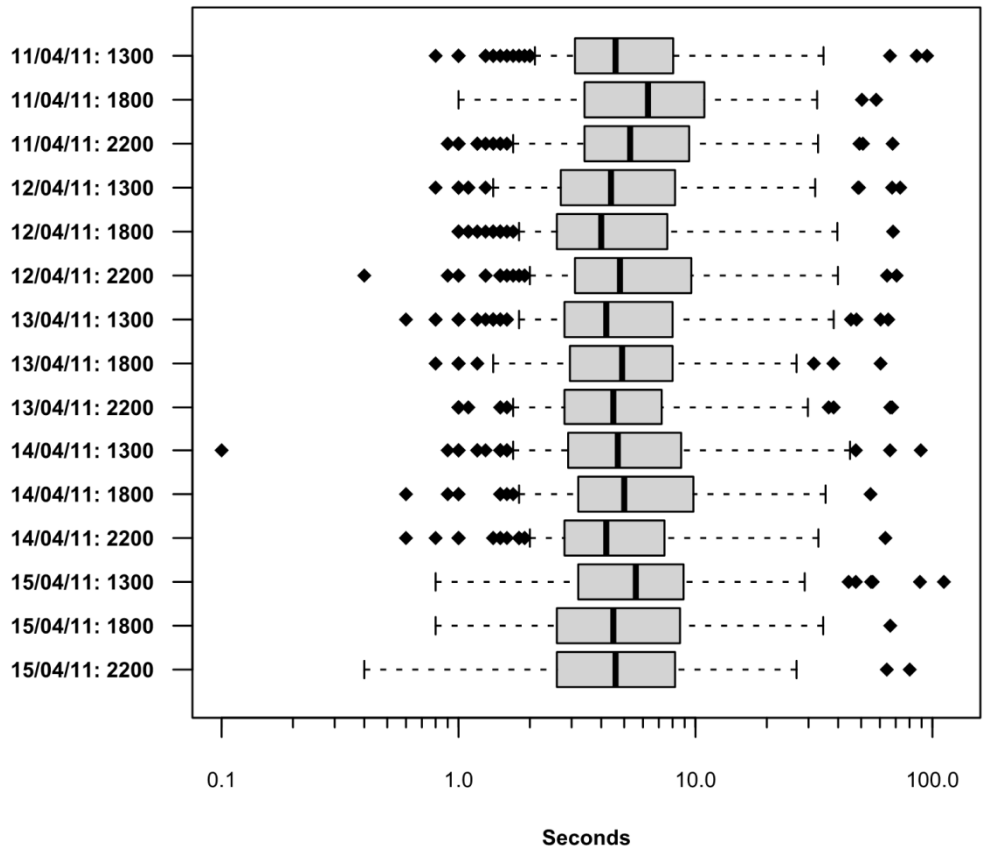

Figure 5 Adjusted box plots of shot length for the three main daily news bulletins broadcast on the BBC from 11 to 15 April 2011.

Potter (2006) provides a detailed survey of the methodology of constructing and interpreting box plots, and a discussion of extensions of and alternatives to box plots (e.g. vioplots).

Q-Q plots

A Q-Q plot (also known as a normal probability plot) allows us to visually inspect how well a data set fits a hypothesised normal distribution by plotting the theoretical quantiles against the sample quantiles of the data. If the data come from a normal distribution then the data will lie along an approximately straight line; while a markedly curved line indicates deviations from the normal distribution. A Q-Q plot can also be used to identify the presence of outliers, which will show up as data points distant from the rest of the distribution. Q-Q plots can also be used to check goodness-of-fit for the two-parameter lognormal distribution (because X is lognormally distributed if $\log(X)$ is normally distributed). This method also has an associated test statistic – the Shapiro-Francia test – which is equal to the square of the correlation between the theoretical and observed quantiles. It is important to note that because the data are not independently and identically distributed interpretation of this test statistic should be in the context of *generalized least squares* in order to avoid erroneous conclusions (see Looney & Gulledge 1985; Redfern 2012a).

For example, De Long, Brunick, and Cutting (in press) claimed that the shot length data of *A Night at the Opera* was well-fitted by a lognormal distribution. To test the veracity of this statement we perform the Shapiro-Francia test and find that a lognormal distribution is not a good fit ($p =< 0.01$). The Q-Q plot is an excellent example of how using graphical methods

allow us to make sense of numerical statistics by showing us why we get this result. The Q-Q plot in Figure 6.a shows that the hypothesised lognormal distribution (represented by the red reference line) underestimates the proportion of shots in the lower tail while at the same time overestimating the proportion of shots in the upper tail. In other words, even after a logarithmic transformation has been applied, the distribution of the shot length data for *A Night at the Opera* is still positively skewed, as can be seen from the kernel density estimate of the log-transformed data in Figure 6.b.

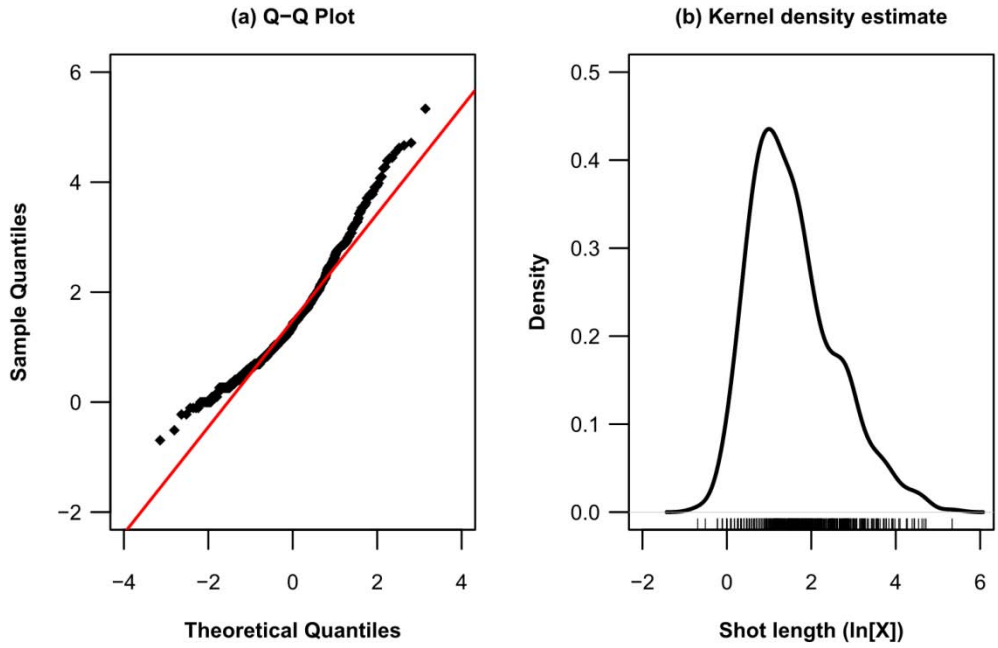

Figure 6 Q-Q plot and kernel density estimate of log-transformed shot lengths of *A Night at the Opera* (1935)

Q-Q plots are simple to produce and easy to interpret. By employing them at an exploratory stage in research they can help us in two ways. First, we avoid making baseless claims that the data is well-fitted by a particular distribution when it does not. In the case of *A Night at the Opera*, the data is not even close to being lognormally distributed and even the merest of glances at a Q-Q plot of the data would have revealed this. Second, many statistical methods assume the data is well-fitted by a particular distribution and can be inefficient, lacking in power, or simply misleading when this is not the case. By checking distributional assumptions using a Q-Q plot we can make informed decisions about which methods to use so that when reporting our results we can be sure they accurately describe the relationships in the data and do not fall down on the basis of unchecked assumptions.

Time-ordered graphical displays

Klevan (2011: 74) writes that 'film – visual, aural *and* moving – is a particularly slippery art form' (original emphasis) that 'sets up peculiar problems for analysis and description because it is tantalisingly present and yet always escaping.' The use of time-ordered graphical displays allows us to overcome these problems to look at the dynamic structure of style in a simple, quick, and informative manner that encourages us to interact with the data

so that we may perceive formal relations across a whole film whilst also identifying relevant features at smaller scales. Since our perception of time is subjective, graphical representations of the dynamics of film style afford a degree of precision in discovering when differences in the pace and tempo of a motion picture occur and in characterising the size and nature of those differences not attainable by other methods, and to bring to our attention subtle dynamic variations that otherwise may pass unnoticed. Graphical displays enable us to reveal the editing structure of a film by identifying the underlying trend, the presence of cycles and clusters of shots, or bringing interesting features to our attention; and in this section I outline five simple time series methods.

Moving averages

A simple moving average shows the average value of data within a time window (which usually has an odd number of data values):

$$MA = \frac{x_1 + \cdots + x_n}{n},$$

where n is the number of data points in the window. Windows can either be centred, with the average value plotted at the centre of the window with an equal number of observations on either side (hence the need for an odd number of observations in a window), or lagged, so that the average value is estimated for last observation in the window. The larger the window used in calculating the moving average the smoother the trendline. The simple moving average is the most commonly used smoothing method but is not robust to the presence of outliers in a time series resulting in misleading descriptions of a time series. The value of the moving average falls as a high value is dropped from the window and rises as a low value is dropped but neither of these changes are related to the actual state of the time series. These problems are evident in Figure 7.a, which shows the time series of shot length data for the 1830 ITV1 news bulletin from 12 August 2011 with a moving average calculated using a 9 shot centred window. The trendline from shots 25 to 50, in particular, doesn't provide a good description of the data.

An alternative method is a weighted moving average which assigns weights to the values in a window so that shots furthest from the centre of window contribute the least to the trendline. The trendline in Figure 7.b is a moving average of the same time series also using a 9 shot centred window with the data points in the window weighted using a Gaussian kernel and provides a markedly better fit than the simple moving average. However, this method is also not resistant to the influence of outliers and does not provide a good description of this time series when using larger windows (using this method with a 21 shot window, for example, is much less effective).

The simple moving median works in the same way as a moving average replacing the mean value of the observations in a window with their median:

$$SMM = median(x_1, \ldots, x_n)\,.$$

The simple moving median is robust to outliers and will not be pulled around in a manner similar to the moving average. This is evident in Figure 7.c with the trendline providing a much better description of the time series, albeit one that it is a little Spartan and which may lead us to overlook potentially interesting features. An oddity evident in this trendline is the peak at shots 57 and 58 that doesn't seem to correspond to any of the peaks in the time series, but this is in fact an artefact of the window size used in this example and disappears when using a larger window size and illustrates the problem of choosing a window of the

right size. These issues can be resolved by re-smoothing the trendline using a second moving average, but there is no guarantee this will work.

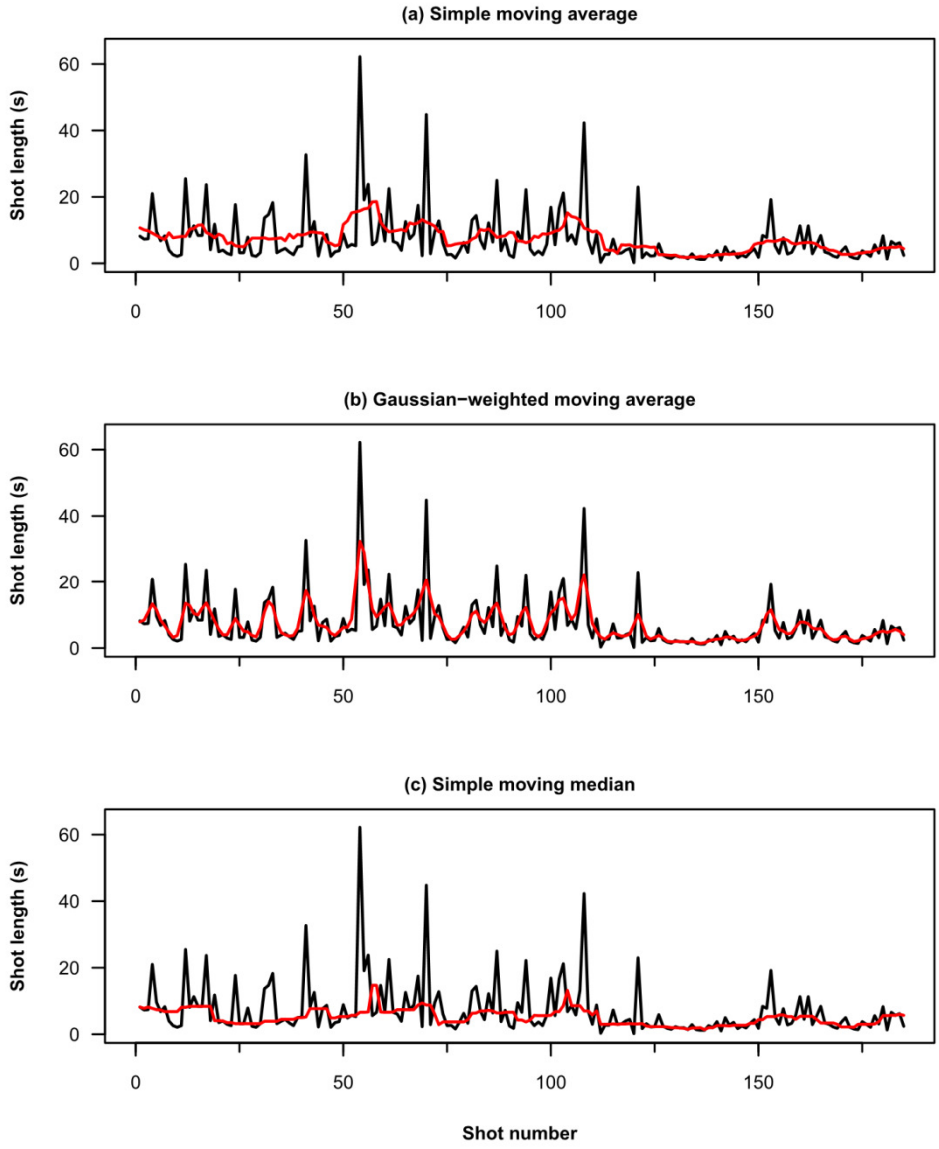

Figure 7 Time series of the 1830 ITV1 news bulletin broadcast on 12 August 2011 with three simple smoothers using a 9 shot centred window

Running averages are simple to calculate and easy to understand, and of the three methods illustrated here the Gaussian-weighted moving average or the simple moving median (preferably re-smoothed) should be used if an appropriately sized window can be identified. The simple moving average should not be used. However, these methods all suffer from the same drawbacks. First, they use a single window of fixed size even though the features of a time series of a motion picture will not be of a single fixed size. Second, choosing a window of the wrong size will either over- or undersmooths the data resulting in key features of a time series being wrongly identified or missed altogether. Third, moving averages are not an efficient way to look at several time series alongside one another.

Regression trendlines

An alternative to moving averages is to use regression-based methods to find a trendline to describe the data. Regression is a statistical technique for estimating the relationship between variables. In applying regression to the time series of motion pictures we are looking at the relationship between the position of a shot in the film and its duration. Therefore, we are going to apply regression to the relationship between the number of a shot and its duration. There are several methods for applying regression to time series analysis of motion pictures, but I will focus on three examples here.

Figure 8 shows the time series data for the 1830 ITV1 news bulletin from 12 August 2011 with trendlines produced using three different regression methods. The trendline in Figure 8.a is based on a fifth-order polynomial and is nowhere a good description of the time series. A polynomial trendline is based on an equation with a finite number of terms and the coefficients are estimated from the whole data set. This is not desirable in analysing the time series of motion pictures because the editing of one section will be very different to the editing of another and we do not want the style of one sequence to affect our interpretation of another sequence, especially when they serve different functions. For example, the cluster of short takes in Figure 8.a from shot 122 to shot 150 is part of a news report on a cricket match and is formally unrelated to the reports on the London riots of 2011 that account for shots 12 to 67. We do not want the estimation of the trend of either of these sections of the bulletin to affect the other. Higher-order polynomials may fit the data better but there are simpler methods to do the same job, while some software programmes won't calculate very high order polynomials (for example, PAST will only calculate up to the fifth order and Excel only up to the sixth).

We can achieve far better results by using locally estimated regression methods. LOESS regression is a nonparametric method that uses a low-order polynomial to fit a trendline to a time series by using only those data points in the neighbourhood of a specific point in time series rather than fitting the trendline globally. This neighbourhood is called the *span* and is a fraction of the whole data set used to estimate the trend at a given point. LOESS regression also uses weights to estimate the trendline. Figure 8.b fits a LOESS trendline using a span of 0.1 and a second-order polynomial with Gaussian weighting, and gives a much better description of the data than the globally-estimated polynomial. A robust method of performing the same task uses Least Quartile Difference (LQD) regression to fit a trendline to the data using a moving window, estimating the trendline at a specific location using only those data points in the window (see Bernholt, Nunkesser, & Schettlinger 2007 for an introduction). Figure 8.c shows a trendline fitted using LQD regression using a centred 11 shot window and produces a similar trendline to the simple moving median. Using methods such as LOESS and LQD regression allows us to compare the predicted values of the model with the observed values of the data set, and by examining the residuals (i.e. the difference between the predicted and observed values) we can identify the presence of outliers in a time series and take steps to deal with these issues as they arise (see, for example, Redfern 2012b).

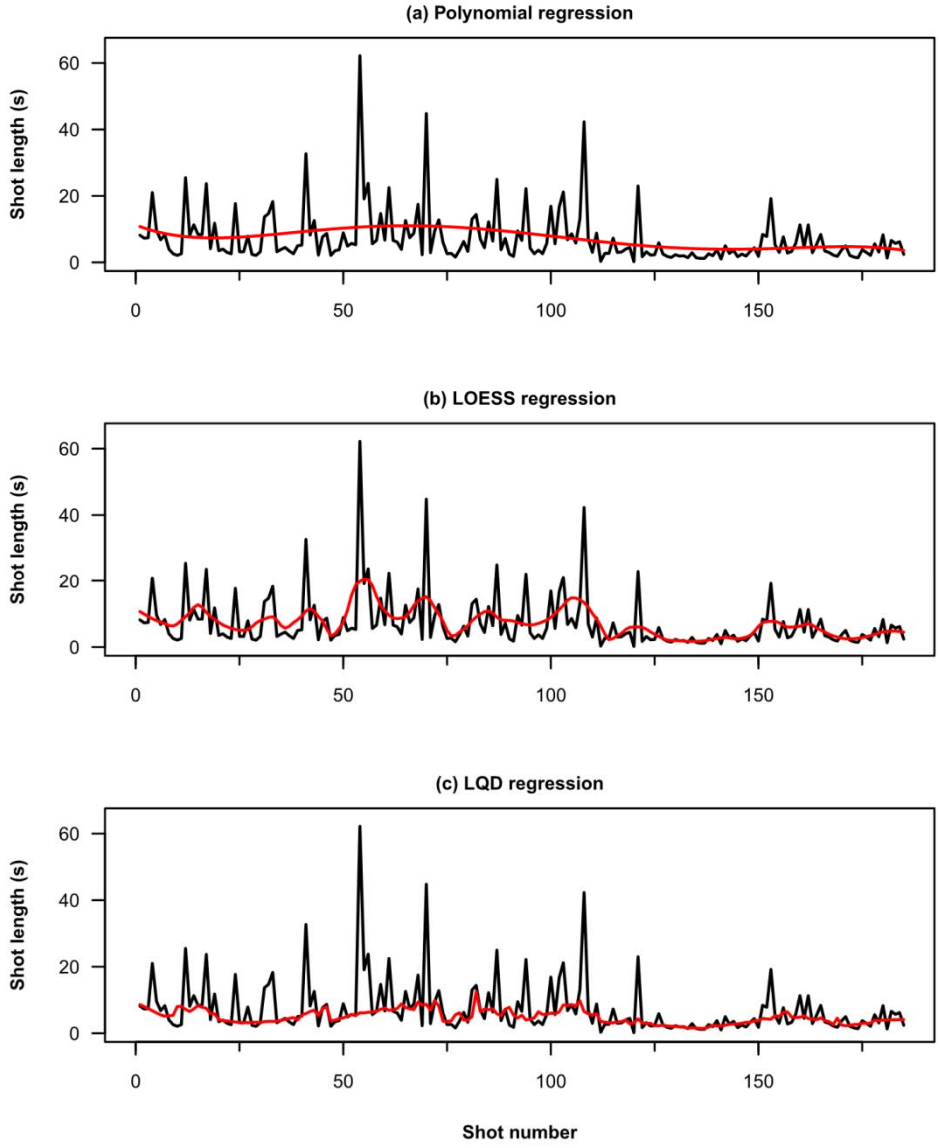

Figure 8 Time series of the 1830 ITV1 news bulletin broadcast on 12 August 2011 with three regression trendlines

These methods suffer the same problem as moving averages in that it is not very easy to put several trendlines on the same graph in order to compare the time series of several films, while the LOESS and LQD methods still require us to use a single window of fixed size in fitting the trendline.

Running Mann-Whitney Z statistics

Running Mann-Whitney Z statistics provide an alternative to the moving averages and regression trendlines described above that combine resistance to outliers, the opportunity to use multiple windows of different sizes, and the ability to compare several time series on the same scale (Mauget 2011).

The first step in the analysis is to rank the N shots in a motion picture, with the smallest x_i assigned rank 1 and the largest x_i assigned rank N. Shots of equal length are assigned the average of the ranks they would have been assigned if there were no ties: if x_2 and x_3 are

observations with equal values, the average rank assigned to each is $\frac{2+3}{2} = 2.5$, and the next highest value will be assigned a rank of 4. The rankings are then sampled with a moving window of size n_1, and converted to *U* statistics by

$$U = R_1 - \frac{n_1}{2}(n_1 + 1)\,,$$

where R_1 is the sum of ranks in the window of size n_1. When the sample size is large ($n_1 \geq 10$), the distribution of U is approximately normal with mean $\mu = (n_1 \times n_2)/2$, and standard deviation,

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}\,,$$

where $n_2 = N - n_1$. Statistical significance can therefore be determined by calculating a Z statistic,

$$Z = \frac{U - \mu}{\sigma}\,,$$

which is compared to a standard normal distribution.

This *Z* statistic is compared to a critical value and when *Z* is less than or equal to this value we will identify a window that represents a cluster of low-ranking (i.e. shorter) shots; and, when *Z* is greater than or equal to this value we will identify a cluster of high-ranking (i.e. longer) shots. A set of time series of running Mann-Whitney *Z* statistics can be generated for each motion picture using multiple moving windows of different sizes and then screened for the most significant clusters in order to remove redundant significant values resulting from the overlapping windows. The most significant non-overlapping windows of shots with high and low rankings can then be colour-coded and plotted on a single horizontal axis.

Applying this method to the formal organization of a motion picture allows us to identify trends over the course of its running time, to identify clusters of takes of long or short duration, to identify the points at which the style changes, and to determine if any intermittent cyclical patterns are present. By going back to a film having identified significant clusters we can determine which features these clusters are associated with in order to discover if there is a relationship between the presence of a *type* of sequence in a film and its style.

The Mann-Whitney *U* test is the equivalent of applying a *t*-test to the ranks of the data (Connover & Iman 1981), and so calculating the running Mann-Whitney *Z* statistic for a window of length *k* is the equivalent of calculating the running average of length *k* for the ranks of the data. The advantage of the *Z* statistic method is that it lets you use several windows of different sizes and then choose between them in order to find the window size that fits that part of the data best. This allows us to overcome the problem of being stuck with a single moving window size that doesn't fit any of the features of the data exactly by using the right-sized window at the right time while also retaining the robustness of rank-based methods. A drawback of this method is that it is computationally intensive, but it has the advantage of being able to place large numbers of time series alongside one another to look for similarities and differences of film style.

Figure 9 presents the data for the fifteen major news bulletins broadcast on ITV1 from 8 to 12 August 2011 using a critical Z value of ±1.64 (See Redfern 2011). To make it easier to

compare the time series of these bulletins the running time have been normalized to a unit length.

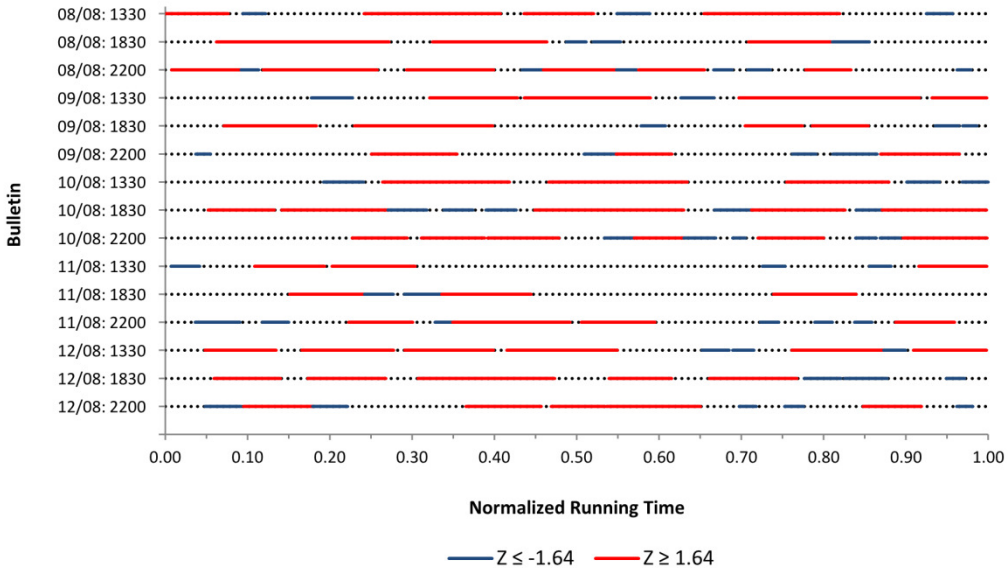

Figure 9 Side-by-side comparisons of the most significant non-overlapping regimes of short and long shots based on running Mann-Whitney Z statistics using multiple windows ($n_1 = 10 - 15$) in ITV News bulletins, 8 August 2011 to 12 August 2011

Although the discourse structure of these bulletins is governed by a strict set of constraints, there is no overall pattern to the formal structure. The number of significant clusters ranges from a low of five to a high of twelve, but shows no pattern by time or day of broadcast. There is no order in which the significant clusters of long or short takes occur: clusters of long shots may be followed by clusters of short shots and clusters of short shots may be followed by clusters of long shots, with numerous occasions when there appear to be runs of similar clusters. There are no trends or cycles evident over the course of the bulletins, and there are no clusters of shorter or longer takes common to the time series of all the bulletins.

While the results show that there is no particular pattern, we can identify some of the structural elements of the discourse of television news as being associated with clusters of longer of shorter takes. Clusters of long takes occur when there are several shots of people talking on screen in proximity to one another. This includes the kernel of a news item, press conferences, interviews with named key figures in news items, the reporter talking directly to camera at the beginning and/or end of a report, and the live 2-way interviews. Clusters of shorter shots are associated with four different elements of news discourse: montages with a voice-over provided off-screen by a reporter, sports reports, sequences of actualities that cover several items in quick succession without differentiating kernels, and of shorter takes are associated with footage that is not produced by ITN news, including footage from other broadcasters, library footage, and clips from feature films. The overall formal structure of a news bulletin is therefore determined by the presence of and sequence in which discourse elements are presented.

Kernel density estimation for a point process

So far I have at looked time series methods based on the duration of shots in a motion picture, but it is also possible to think of the time series as a *point process*. Rather than focussing on the length of a shot (L) as the time elapsed between two cuts, we are interested in the timing of the cuts (C) themselves. There is a one-to-one correspondence between cuts and shot lengths, and the time at which the jth cut occurs is equal to the sum of the lengths of the prior shots:

$$C_j = \sum_{i=1}^{j} L_i \, .$$

Figure 10 shows the one-to-one nature of this relationship clearly.

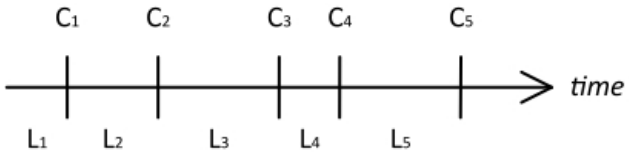

Figure 10 The one-to-one relationship between shot lengths (L_i) and the timing of a cut (C_j)

A point process is a stochastic process whose realizations comprise a set of point events in time, which for a motion picture is simply the set of times at which the cuts occur. We can use kernel density estimation to describe the cutting rate of a film, applying the same method used to describe shot length distributions to the point process to produce a density estimate of the time series. The density is greatest when one shot quickly follows another and, therefore, the shorter the shot lengths are at that point in the film. Conversely, low densities indicate shots of longer duration as consecutive shots will be distant from one another on the x-axis. Using kernel density estimation to understand the cutting rate of a film as a point process is advantageous since it requires no assumptions about the nature of the process. Salt (1974) suggested using Poisson distributions as a model of editing as a point process described by the rate parameter λ, but this method is unrealistic since homogenous Poisson point processes are useful only for applications involving temporal uniformity (Streit 2010: 1). For a motion picture the probability distribution of a cut occurring at any point in time is not independent of previous cuts, and the time series will often be non-stationary over the course of a film while also demonstrating acceleration and deceleration of the cutting rate because different types of sequences characterised by different editing regimes.

To illustrate the use of kernel densities in time series analysis we compare the editing of two films featuring Fred Astaire and Ginger Rogers: *Top Hat* (1935) and *Shall We Dance* (1937). In order to make a direct comparison between the evolutions of the cutting rates the running time of each film was normalised to a unit length by dividing each shot length by the total running time. In this case we treat slow transitions (e.g. fades, dissolves, etc) as cuts, with the cut between two shots marked at the approximate midpoint of the transition. Figure 11 shows the resulting densities.

From the plot in Figure 11 of *Top Hat* we can see the density for this film comprises a series of peaks and troughs, but that there is no overall trend in the density. The low densities in this graph are associated with the musical numbers, while the high densities occur with scenes based around the rapid dialogue between Astaire and Rogers. The first

musical number is 'No Strings (I'm Fancy Free)', which begins at ~0.07. Astaire is then interrupted when Rogers storms upstairs to complain about the racket, and we have a scene between the two in which both the dialogue and the editing are rapid. This occurs at the peak at ~0.11 to ~0.13, and is then followed by a reprise of 'No Strings,' which is again shot as a long takes. The next section of the film follows on the next day as Astaire takes on the role of a London cabby and drives Rogers across town and as before this dialogue scene is quickly edited resulting in a high density of shots at ~0.19. This sequence finishes with 'Isn't This a Lovely Day (to be Caught in the Rain),' which accounts for the low density of shots at ~0.21 to ~0.27 since this number again comprises long takes. The rapid cutting rate during dialogue scenes is repeated when Rogers mistakes Astaire for a married man at the hotel, and is again followed by the low density of a slow cutting rate for the scenes between Astaire and Edward Everett Horton at the theatre and the number 'Top Hat, White Tie and Tails' at ~0.4. After this number the action moves to Italy and there is much less variation in the density of shots in the first part of these scenes, which are focussed on dialogue and narrative. There is no big musical number until 'Cheek to Cheek' and this sequence accounts for the low density seen at ~0.66, being made up of just 13 shots that nonetheless run to 435.7 seconds. The density increases again as we move back to narrative and dialogue until we get to the sequence between in which Horton explains the mix-up over who is married and who is not to the policeman and 'The Piccolino' which begins at ~0.89 and runs until ~0.96.

The density plot of the point process for *Shall We Dance* differs from that of *Top Hat* showing a trend over the running time of the film from higher to lower densities of shots, indicating the cutting rate in this film slows over the course of the film. Nonetheless we see the same pattern of troughs and peaks, and as in *Top Hat* these are associated with musicals and comedy scenes, respectively. This film features numerous short dancing excerpts in its early scenes, but there is no large scale musical number until well into the picture. In fact, these early scenes are mostly about stopping Astaire dancing (e.g. when Horton keeps turning off the record) and the dialogue scenes that establish the confusion over Astaire's married status as the ship departs France. These scenes are based around a similar narrative device to that used in *Top Hat* and are again edited quickly. The first big number in the film is 'Slap that Bass' and coincides with the low density section of the film beginning at ~0.17, indicating that this part of the film is edited more slowly that the first section. The cutting rate slowly increases until ~0.37, and this section includes the 'Walking the Dog' and 'I've Got Beginner's Luck' numbers but is mostly made up of dialogue scenes between Astaire and Rogers. After this point the film exhibits a trend from higher to lower densities and there are a number of smaller cycles present between 0.37 and 0.64. This section includes 'They All Laughed (at Christopher Columbus)' and the subsequent dance routine, which begins at ~0.48 and includes the trough at ~0.54. The low density section beginning at 0.64 is the scene between Astaire and Rogers in which they try to avoid reporters in the park, and comprises a number of lengthy dialogue shots and the film's most famous number 'Let's Call the Whole Thing Off.' The editing then picks up during the dialogue scenes until we reach the next drop in the density at ~0.74 which coincides with the scenes on the ferry to Manhattan as Astaire sings 'They Can't Take That Away From Me.' The next low density section begins at ~0.9, and is the big production at the end of the film with the distant framing and static camera completing the long takes in showing off the 'Hoctor's Ballet' sequence, which then gives way to a more rapidly cut section featuring numerous cut-ways from the dancers to Rogers' arriving at the theatre with the court order for Astaire only to discover him on stage with dancers wearing masks of her face. The cutting rate then slows once Rogers insinuates herself into the 'Shall We Dance' routine and the film reaches its finale.

Comparing the two plots we note some of the low density periods coincide with one another. This is most clearly the case at around 0.2 and 0.64 in both films. This indicates that a musical number occurs at approximately the same points in both films even though the two films have different running times (*Top Hat*: 5819.9s, *Shall We Dance*: 6371.4s). This raises some interesting questions regarding the structure of other musicals featuring Astaire and Rogers; is there always a musical number about a fifth of the way into an RKO musical featuring this pair? Is there always a major number about two-thirds the way through picture? And does the finale always occupy the last 10 per cent of the picture?

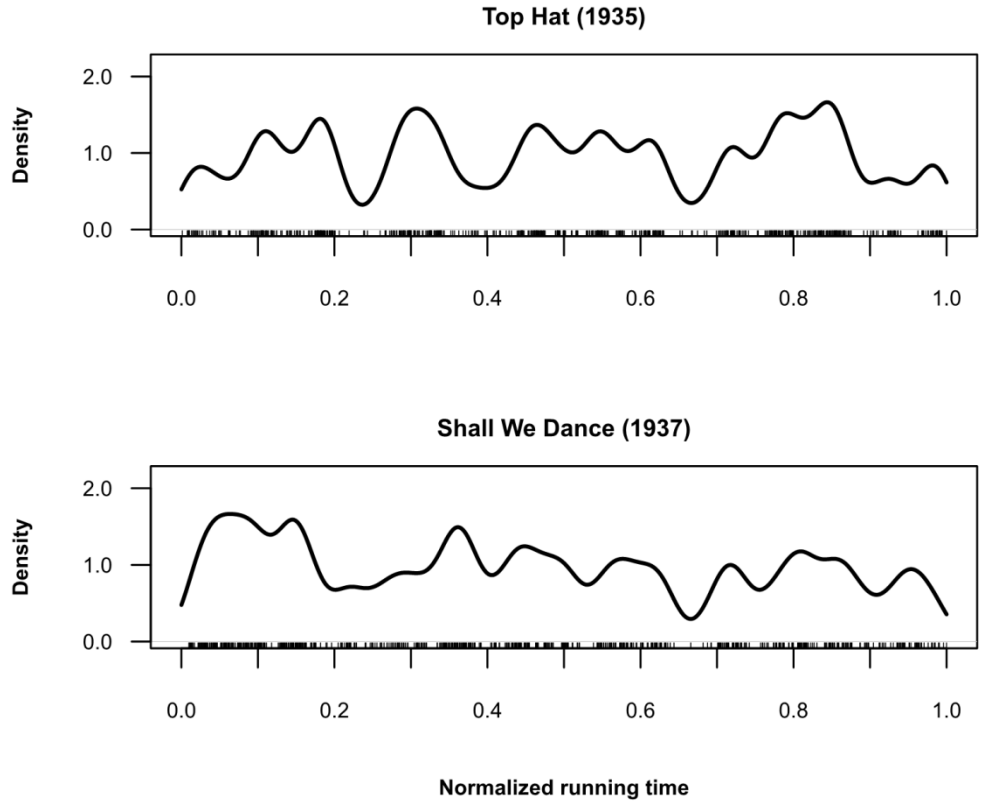

Figure 11 Kernel density estimates of the point processes for two RKO musicals with normalised running times

The Order Structure matrix

The order structure matrix (Bandt 2005) is a conceptually simple and robust method that provides a global description of the structure of a time series, $X_1, \ldots, X_T$. It is appropriate for the exploratory analysis of shot length data, allowing a researcher to identify clusters of longer or shorter takes, the presence of intermittent cyclical patterns, and to locate change points in the editing of a film that can then be analysed in more detail. The matrix ($\mathbf{M}$) compares the values of pairs of points at times s and t, where $1 \leq s, t \leq T$, and to construct the matrix we assign a value of 1 if $x_s \geq x_t$and a value of 0 if $x_s < x_t$. Assigning colours to these values (1 = black, 0 = white) we obtain a graph that makes it easy to visualise the editing structure of a motion picture. The matrix is reflected in the main diagonal and we use the transpose of the order matrix ($\mathbf{M}^{\mathrm{T}}$) to more easily distinguish editing patterns,

representing shorter shots which tend to cluster as white columns and longer takes that may occur in isolation as black columns.

Figure 12 shows the order structure matrix of *Friday the Thirteenth* (1980) and shows this film to be comprised of six different narrative segments, each characterized by a different editing style with transitions between these sequences are associated with changes in mood (see Redfern 2012c).

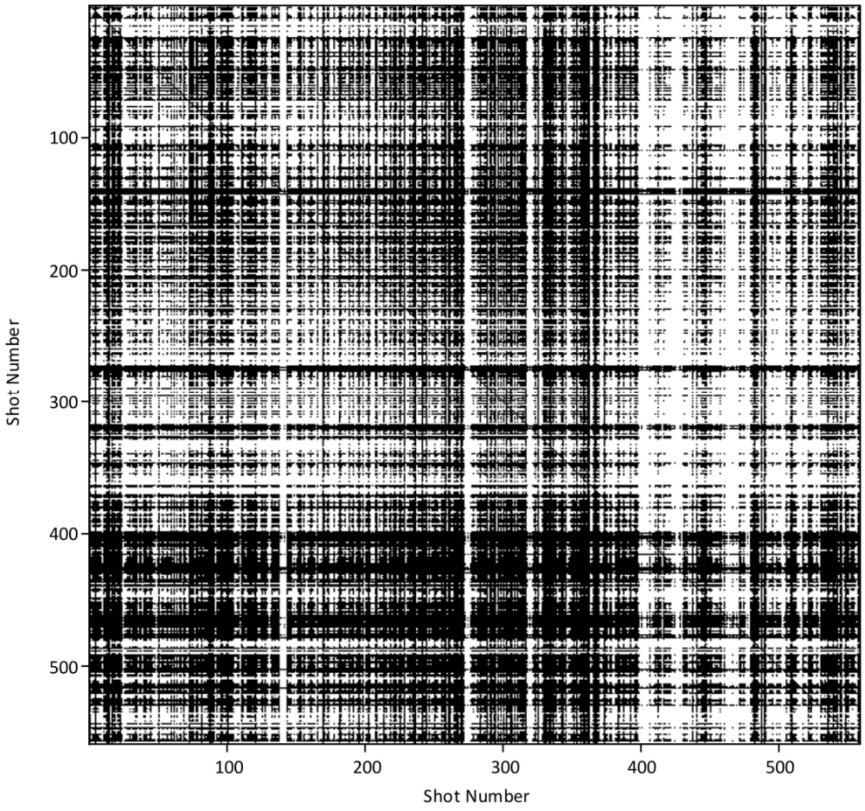

Figure 12 Order structure matrix of *Friday the Thirteenth* (1980)

The first section of the film is the originating event of the murder of two counsellors at Camp Crystal lake in 1958 (shots 1-17, Σ = 294.1s, median = 6.7s, IQR = 24.0s). This sequence includes several tracking shots from the point-of-view of the unseen killer, with six shots running to more than 19 seconds accounting for the dark column to the extreme left of the matrix. The second section of the films is set in the present, introducing the main characters and establishing the (shots 18-144, Σ = 965.5s, median = 5.5s, IQR = 7.9s). Within this sequence there are two narrative threads – Annie hitchhiking to the camp (shots 18-74) and the arrival of the new counsellors at the camp (shots 75-122) – with a slight tendency to longer takes in the latter as this part of the sequence is organised around a series of group shots compared to the shot-reverse shot pattern of singles that characterise Annie's interactions with the locals as they warn her about the past of the camp. This sequence ends with the murder of Annie by an unknown assailant, including a series of very short takes that feature as the thin white band at shots 139-144 in Figure 12 as the violence of this assault reaches its peak. From shot 145 shot lengths become increasingly longer until shot 262 (Σ = 1222.1s, median = 5.5s, IQR = 8.8s). This can be seen in the matrix in Figure 3 as the density

of the black columns increases over this portion of the film, and the trend towards higher ranked shots is particularly noticeable in the ranks run chart. This third sequence does not feature any violence and focuses on the creation of sense of foreboding through the near drowning and the snake in the cabin, the warnings of the police officer and 'Crazy Ralph,' and the appearance of a mysterious figure in an oil slicker. This is achieved formally as the film builds tension by progressively lengthening the duration of takes so that over the course of this sequence the pace of the film slows down.

Changes in editing style between these first three segments of the film occur when there are large differences in mood between consecutive scenes. Thus, the transition between the first and second segments of the film occurs with the shift from the violence of the original murders at the camp to the sunny optimism of the present day as Annie sets off to be a counsellor at the camp. (This transition is also separated by the opening titles, which are not included in the data set). Similarly, the shift between the second section ending with Annie's murder and the next section with the other counsellors swimming at the camp is based on a shift between extreme violence and the counsellors relaxing at the lake. These shifts are also marked by the use of white flash transitions that do not appear at any other point in the film, and which are used to cover shifts in time (past/present) and location (Annie/the camp). The use of white flash transitions to cover shifts in location occurs only at the transition *between* different narrative segments and the onset of a new editing pattern, and they do not feature *within* narrative sections even though multiple locations are used (e.g. the shifts between the counsellors at the camp and Annie in the second part of the film).

In the later parts of the film, transitions between editing regimes occur within scenes and are not associated with any particular optical effect as narrative space and time are continuous. The fourth sequence is the 'stalk-and-slash' segment of the film and includes the murders of Jack, Marcie, Brenda, and Steve (shots 263-399, Σ = 2095.0s, median = 8.4s, IQR = 15.9s). The sequence begins in the middle of a scene with the sudden murder of Jack, and from this point longer takes dominate as the emotional tone built up in the preceding segment is sustained. The use of long takes emphasises the anticipation of violence and creates for the viewer a continuing sense of dread that round every corner lurks a new terror. Though this is the slowest segment of the film, but also contains two clusters of very short takes at shots 273-280 and shots 318-323 associated with the murder of Marcie and the terrorising of Brenda, respectively. These sudden outbursts of shocking violence realise the fears of the viewer and represent a more aggressive form of terror in counterpoint to the dominant mood of the sequence. However, no such clusters are associated with the murders of Jack or Steve which are shot as long takes, and there appears to be a clear difference in the way in which violent scenes are edited according to the gender of the victim.

The shift from the stalk-and-slash section to the final girl sequence also occurs within a single scene when Mrs. Voorhees's demeanour suddenly changes when recollecting Jason's death; and again a change in the dominant mood is accompanied by a change in style. The final girl sequence runs from shot 400 to shot 534 and is edited much more quickly than other parts of the film (Σ = 647.7s, median = 2.8s, IQR = 2.9s). This sequence is characterised by the heightened emotional intensity of Mrs. Voorhees's psychosis, creating a mood of aggressive tensity that continues through the sustained violence she inflicts on the panicked Alice. This segment breaks Alice's fight for her life into three rapidly edited violent scenes divided by slower cut clusters as the final girl runs and hides from the killer; and these clusters can be clearly seen as the dark columns in this section in Figure 12. The final sequence of the film comprises shots 535 to 559 (Σ = 269.8s, median = 9.3s, IQR = 12.9s), and includes Alice's dream of encountering Jason as she floats across the lake in a canoe and waking in the hospital. This sequence begins with a shot of the moon over the camp recalling the opening shot of the film, and like the opening sequence is edited slowly. This coda

reintroduces the idea that a threat remains out there, and returns – emotionally and stylistically – to the pervading sense of unease and disquiet of the earlier narrative sequences.

The order structure matrix cannot be used to compare the style of two films, but it does have a key advantage over moving average, regression, or kernel density methods by preserving the abruptness of changes in editing style that is lost by smoothing the data.

Conclusion

The use of graphical methods both simplifies and amplifies the process of analysing film style, allowing us to identify interesting features, to compare the style of films, and test the assumptions that underpin our research. They make possible the analysis of film style as a dynamic system. Instead of dealing with one or two key scenes we can analyse the style of whole films at different scales. Clayton and Klevan characterize film criticism as 'a form of writing which addresses films as potential achievements and wishes to convey their distinctiveness and quality (or lack of it)' (2011: 1). That is all well and good for film *criticism* as a literary genre, but the *analysis* of cinema requires a visual dimension and for that we need the economy and the expansiveness only graphical representations can provide. As the use of methods such as those described here become commonplace research in film studies will increasingly become a visual rather than a verbal activity.

The methods described in this may be extended to include additional variables thereby enriching our understanding of how the different elements of film style are related. For example, by adding shot scale data to the cutting we can look at the time evolution of film style via a *marked point process* that combines both sets of data. There are whole areas of graphical analysis of shot length data that I have not touched upon (e.g. short-time Fourier transforms, wavelet analysis), as well as function-based methods for analysing film tempo that make full use of computers as research tools (Adams, Dorai, & Venketesh 2000). There are also network-based methods to identify higher-order cinematic constructs that can be used to search for relationships between shots and to reveal the editing structure of scenes, acts, and whole films (Truong, Venkatesh, & Dorai 2005: 280-281).

Again, I stress that the only way to learn how to produce and use graphical representations of film style is to get some data and try it; but whichever method you select for analysing film style it should meet the requirements set out by John Tukey: it must make the data simpler to work with and it must takes us deeper into the data so that we can discover what is going on.

References

Adams B, Dorai C, and Venketesh S 2000 Role of shot length in characterizing tempo and dramatic story sections in motion pictures, in L Guan and RLK Liu (eds.) *Proceedings of the First IEEE Pacific Rim Conference on Multimedia, 13-15 December 2000, Sydney, Australia*. Sydney: University of Sydney: 54-57.

Bandt C 2005 Ordinal time series analysis, *Ecological Modelling* 182 (3-4): 229-238.

Behrens JT 1997 Principles and practices of exploratory data analysis, *Psychological Methods* 2 (2): 131-160.

Behrens JT and Yu C-H 2003 Exploratory data analysis, in JA Schinka and WF Velicer (eds.) *Handbook of Psychology: Volume 2 – Research methods in Psychology*. Hoboken, NJ: John Wiley & Sons: 33-64.

Bernholt T, Nunkesser R, and Schettlinger K 2007 Computing the least quartile difference estimator in the plane, *Computational Statistics & Data Analysis* 52: 763 – 772.

Carroll N 2009 Style, in P Livingstone and C Plantinga (eds.) *The Routledge Companion to Philosophy and Film*. London: Routledge: 268-278.

Clayton A and Klevan A 2011 Introduction: the language and style of film criticism, in A Clayton and A Klevan (eds.) *The Language and Style of Film Criticism*. Abingdon: Routledge: 1-26.

Conover W J and Iman RL 1981 Rank transformations as a bridge between parametric and nonparametric statistics, *The American Statistician* 35 (3): 124-129.

De Long J, Brunick KL, and Cutting JE in press Film through the human visual system: finding patterns and limits, in JC Kaufman and DK Simonton (eds.) *The Social Science of Cinema*. New York: Oxford University Press. An online version of this paper is available at http://people.psych.cornell.edu/~jec7/pubs/socialsciencecinema.pdf, accessed 18 January 2012.

Ellison AM 1993 Exploratory data analysis and graphic display, in SM Scheiner and J Gurevitch (eds.) *Design and Analysis of Ecological Experiments*. New York: Chapman & Hall: 14-45.

Good IJ 1983 The philosophy of exploratory data analysis, *Philosophy of Science* 50 (2): 283-295.

Hartwig F and Dearing BE 1979 *Exploratory Data Analysis*. Newbury Park, CA: Sage.

Jacoby WG 1997 *Statistical Graphics for Univariate and Bivariate Data*. Thousand Oaks, CA: Sage.

Klevan A 2011 Description, in A Clayton and A Klevan (eds.) *The Language and Style of Film Criticism*. Abingdon: Routledge: 70-86.

Looney SW and Gulledge TR 1985 Use of the correlation coefficient with normal probability plots, *The American Statistician* 39 (1): 75-79.

Mauget SA 2011 Time series analysis based on running Mann-Whitney Z statistics, *Journal of Time Series Analysis* 32 (1): 47–53.

Montgomery M 2007 *The Discourse of Broadcast News: A Linguistic Approach*. London: Routledge.

Potter K 2006 Methods for presenting statistical information, in H Hagen, A Kerren, and P Dannenmann (eds.) *Visualization of Large and Unstructured Data: Lecture Notes in Informatics GI-Edition* S-4: 97–106.

Redfern N 2011 Time series analysis of ITV news bulletins, http://nickredfern.files.wordpress.com/2011/11/nick-redfern-time-series-analysis-of-itv-news-bulletins.pdf, accessed 7 January 2013.

Redfern N 2012a The log-normal distribution is not an appropriate parametric model for shot length distributions of Hollywood films, *Literary and Linguistic Computing*, Advance Access published December 13, 2012, doi:10.1093/llc/fqs066.

Redfern N 2012b Robust estimation of the modified autoregressive index of film style, http://nickredfern.files.wordpress.com/2012/10/nick-redfern-robust-estimation-of-the-modified-autoregressive-index-of-film-style.pdf, accessed 7 January 2013.

Redfern N 2012c Exploratory data analysis and film form: the editing structure of slasher films, http://nickredfern.files.wordpress.com/2012/05/nick-redfern-the-editing-structure-of-slasher-films.pdf, accessed 7 January 2013.

Truong BT, Venkatesh S, and Dorai C 2005 Extraction of film takes for cinematic analysis, *Multimedia Tools and Applications* 26 (3): 277–298.

Tufte ER 2001 *The Visual Display of Quantitative Information*, second edition. Cheshire, CT: Graphics Press.

Tukey JW 1977 *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Velleman PF and Hoaglin DC 1981 *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press.