Chapter 2

Examples

2.1 Preamble

The intention here is to illustrate the kind of thing I understand by 'pattern' by way of examples. These serve as an introduction to later chapters where more technical discussion, including computational aspects, is provided.

If forced to define 'pattern' it would be something rather vague, like an organisation of the elements of analysis that departs in a recognisable and describable fashion from what would be expected from 'random' organisation ('random' not always being easy to define); or a marked departure from the pattern exhibited by some hypothetical 'base' model, not itself of intrinsic interest.

Some patterns can be described qualitatively; for example, many SL distributions have a characteristic shape, singly peaked and skew, that differs from a hypothetical and uninteresting 'flat' distribution. A subset of these exhibit a stronger pattern, that they have in common and which potentially establishes a 'norm', in the sense that they can be described mathematically using the lognormal distribution (Section 2.2.3). This might be thought of as mathematically describable regularity. 'Describable regularity' is manifest in the general decline in ASLs of films over the last 60 years or more; it is less easy to characterise using a precise mathematical formula, but readily displayed using graphical statistical methods (Section 2.2.1).

Other forms of pattern require the use of external knowledge to identify them. For example, Figure 2.8 reduces shot-scale data for 24 of Fritz Lang's films to 24 points on a map. Unannotated, their disposition looks fairly random; add labelling to differentiate between the early German films (mainly silent) and later American ones and a pattern in the form of separation between the two is apparent. In this sort of application, if you are lucky, you may get clear white space between distinctive clusters of points. This is indicative of a form of pattern that can be identified without reference to external knowledge, but may be interpreted in the light of such.

This brief discussion is intended to be indicative rather than exhaustive. Issues raised in the identification, display, and comparison of the pattern of SLs *within* films, are illustrated by example in Sections 2.2.4 and 2.2.5. The point to reiterate is that much cinemetric data analysis is motivated by the idea of comparison which, in turn, is facilitated by the ability to recognise and contrast patterns within bodies of data.

2.2 Examples

2.2.1 ASLs over time

One of the more obvious and commented on examples of a 'pattern' in cinemetric data is the decline in average shot lengths (ASLs) over a period since about the early 1950s. This is illustrated, for example, in Figure 1 of Cutting *et al.* (2011a) using their sample of 150 films and data collected

by Barry Salt, and Salt's (2009, p.378) own similar demonstration for US feature films. This sort of pattern is illustrated for US films (1955-2005) in the left-hand plot of of Figure 2.1[1]. ASLs are shown for individual films, the scale being truncated at an ASL of 20 for display purposes), with a line joining the mean ASLs for each year.

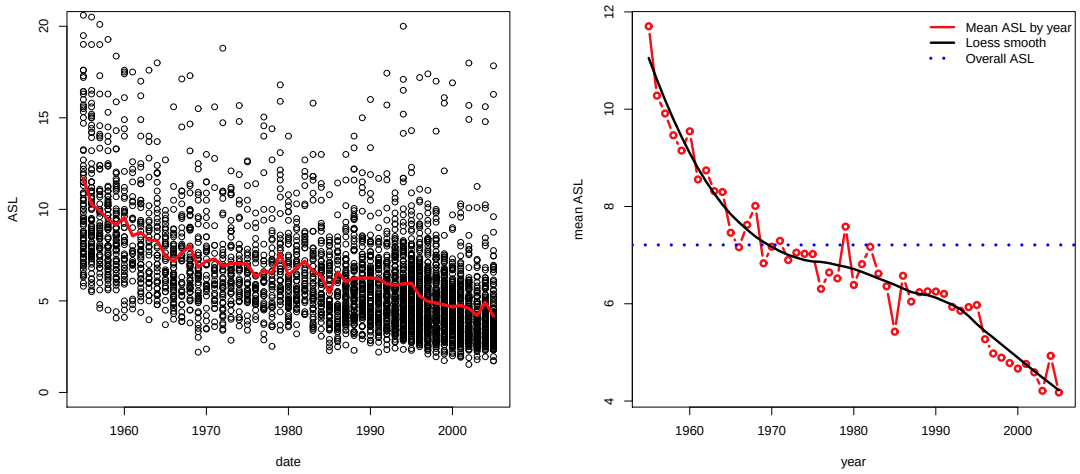

Figure 2.1: *The left-hand plot is of ASLs for American films (1955-2005), the line joining the mean ASLs for each year. This is repeated in the right-hand plot along with a smoothed version; the overall ASL for all films is also shown.*

The general decline in ASLs is clear enough. A magnified plot, omitting individual ASLs is shown to the right, where the decline in mean ASLs of about 7 seconds over a period of 50 years is more apparent. The general pattern is usefully summarised by fitting a smooth curve to the yearly mean ASLs, which better allows general trends to be discussed. Thus, there is a fairly sharp and steady decline from the mid-1950s to late 1960s, a gentler decline with some variation to the early 1990s, then a pick-up in the rate of decline. As ASLs can't go below zero, and there is a practical limit, it is obvious that the pattern will eventually 'bottom-out' at some limit above zero, unless the trend is reversed.

The smoothing method used is discussed in Section 7.5. The right-hand plot shows the mean ASL for all the films used in the period concerned. Pattern here is manifest as a regular and describable deviation from a model – that there is no variation in mean ASLs over time – that can be viewed as a baseline for comarison, which may or may not be of intrinsic interest in its own right.

2.2.2 Comparison of ASL distributions

The comparison effected here, in Figures 2.2 and 2.3, is between samples of American and European silent films for the period 1918-1923. The first figure emulates a comparison from Salt (2009, p.192); the data used are from Salt's database and are not identical to that he used, but the figures are close enough to his for the illustrative purposes intended.

Figure 2.2 compares histograms of ASLs for the two bodies of films. This is a fairly common approach to this kind of comparison; O'Brien's (2005, p.92) study of American and French films during the period of transition to sound provides another example. Pattern can be thought of in more than one way. At the level of qualitative description both sets of films are similar in that they have a single main peak, but the distribution is more spread out for European films.

[1]The data were extracted from Barry Salt's database on the Cinemetrics website.

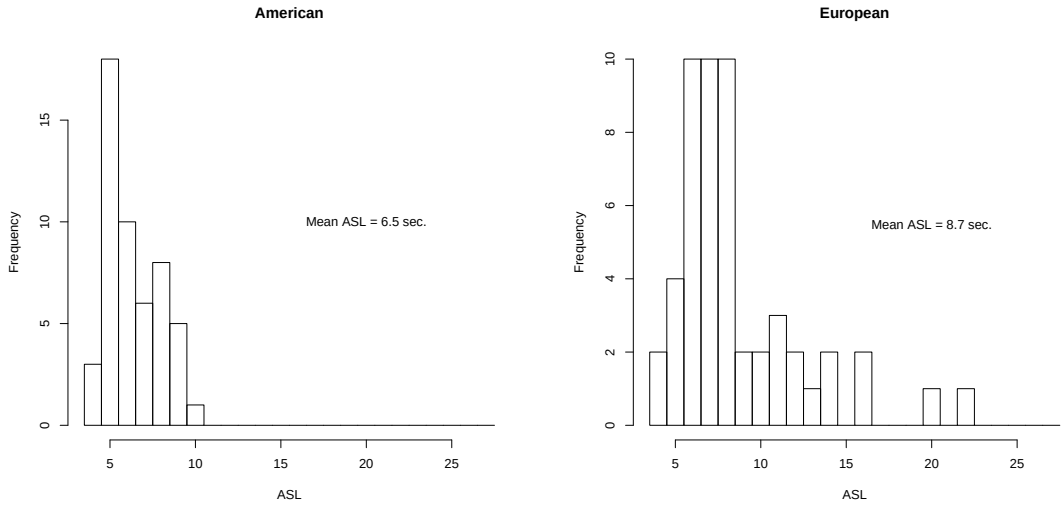

Figure 2.2: *Histograms for the ASLs of American and European silent films 1918-1923. This is after, but not identical to, Salt (2009, p.192).*

Here, and differently from the previous example, the focus is on comparison between two bodies of data. A quantitative comparison of differences in the qualitatively similar ASL distibutions can be effected by calculating statistics that summarise properties of the distribution. One such is the mean of the ASLs, indicated on the histograms for the two sets of films. That for the European films is larger than that for the American ones, suggesting its pattern is shifted to the right by comparison. The difference can be measured, giving a quantitative 'edge' to the qualitative analysis. The use of the mean, alternatives to it, and measures of spread is discussed in a later chapter.

Histograms, discussed further in Section 5.1, are quite a clumsy way of effecting comparison. An alternative is the use of kernel density estimates, KDEs (Section 5.2), which can be thought of, for present purposes, as a smoothed form of histogram. They are much more readily overlaid, for comparative purposes, as the left-hand panel of Figure 2.3 shows. It is immediately evident that the ASLs for the American films are much more concentrated; that the body of the European films is shifted to the right by comparison (i.e. ASLs tend to be larger); and that they spread well beyond the upper limit of ASLs for American films. Another feature of possible interest, that KDEs reveal better than histograms, is that there are two apparent peaks in the distribution for American films. This may turn out not to mean anything, but one is being alerted by the display to a less obvious aspect of pattern that might merit exploration.

The right-hand panel is another way of comparing the same data, using cumulative frequency diagrams (Section 6.3). The vertical axis shows proportions, so if you read across from some value then drop down when you hit a curve you get the ASL which has that proportion of films with smaller ASLs. Such diagrams are common enough in the statistical literature, though not, I think, as easily 'read' as histograms or KDEs. The present example shows, fairly starkly, that a greater proportion of the European films exceed almost any given ASL than the American films.

We have here three different ways of effecting a comparison. I'd make a distinction between the 'big idea', of comparison, and the choice of technique used to effect it [2]. Salt's early work is important at the level of ideas and I think some of the debate about some of his methodology, which I'd view as being at the level of technique, loses sight of this. I will pick up on this later in

[2]The analogy doesn't work exactly, but think of the comparison as a structure that can be erected using different building materials; the choice of technique corresponds to choice of material. And depending on design and intended function, not all materials are suitable for all structures.

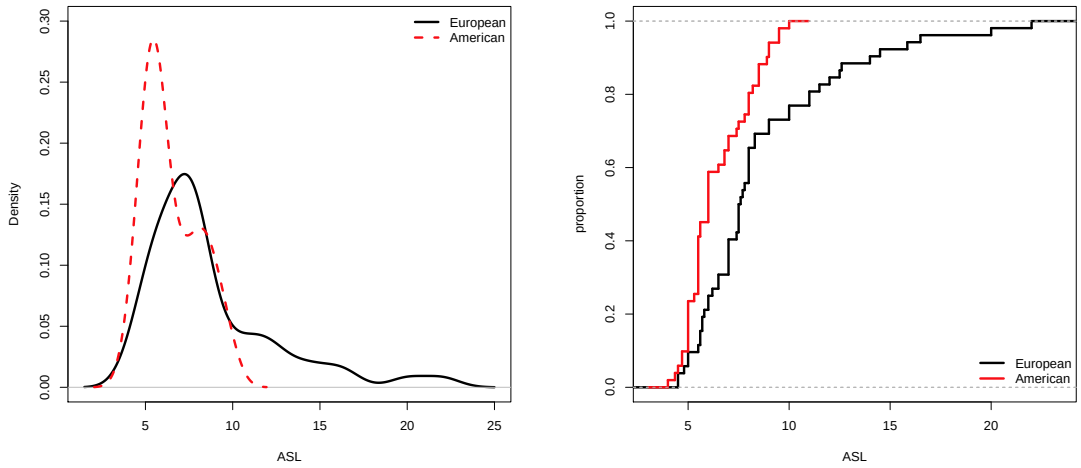

Figure 2.3: *Kernel density estimates (to the left) and cumulative frequency diagrams used to compare ASL distributions for American and European silent films 1918-1923.*

the notes.

2.2.3 Pattern and SL distributions

This example may look similar to the previous one, but there important differences, some of a statistically technical nature. In the last section the elements of which the pattern was composed were film ASLs, the interest being in comparison across national cinemas (genres or historical periods could equally well have been used for illustration). If the elements are SLs and the units of comparison are films identical methods of graphical display are available. SL distributions for two films, *Brief Encounter* (1945) and *Harvey* (1950), are illustrated using histograms in the left-hand panel of Figure 2.4.

Salt (1974, p.15) observed, of SL distributions, that 'a considerable similarity of overall shape is apparent'. In the language I'm using, shape can be equated with pattern and what is being observed is that a lot of films have a qualitatively similar pattern. Many SL distributions have a single obvious peak (or mode)[3] and are skewed with a long right tail, the bulk of the SLs being bunched up nearer the lower bound of zero. Salt went further.

Descriptive statistics, such as the ASL or median SL, can be used to differentiate between films with qualitatitively similar SL distributions. The tack Salt took was to suggest that many distributions could be modelled (approximately) by the same kind of 'mathematically' defined curve, eventually settling on the lognormal distribution (Salt, 2006, p.391; and see the Appendix). The important idea is that the (ideal) shape can be reconstructed perfectly if you know two numbers (parameters), which define the form of the distribution. Several consequences follow.

1. Given a large body of films it can be asserted that those which have a lognormal SL distribution exhibit the same shape, or pattern. That is, they are single peaked, skew and have (usually) a longish right tail.

2. A much stronger statement about pattern, a quantitative one, that they have the same mathematical form is possible. Differences between distributions that can be so categorised can be summarised solely in terms of the *two* parameters that define them[4].

[3] Strictly speaking, 'modal class' if talking about grouped data, as in a histogram.

[4] The dependence on two numbers is why discussions about whether the ASL or median SL is the better descriptor

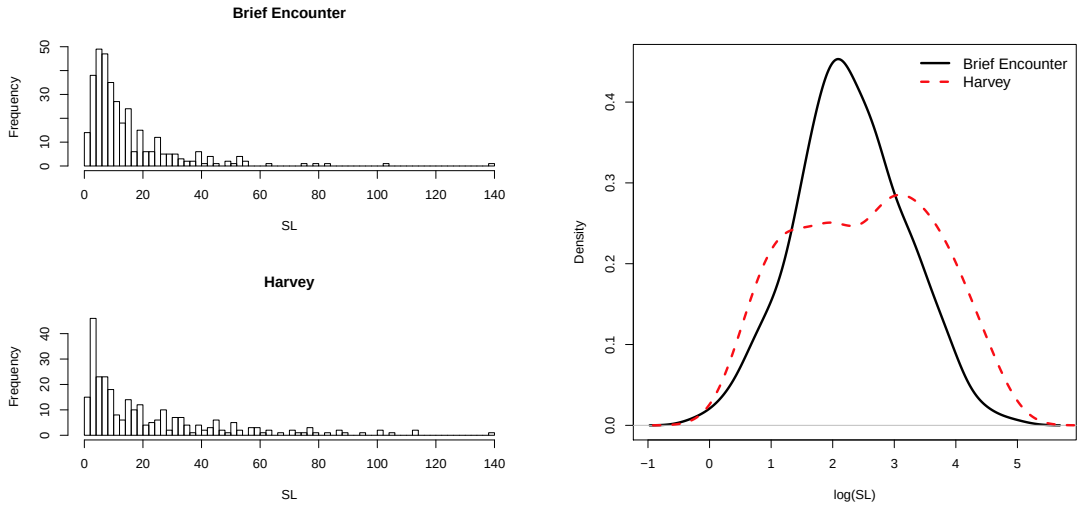

Figure 2.4: *Histograms for the SL distributions of* Brief Encounter *and* Harvey*, and kernel density estimates of the logarithms of the SLs.*

3. Identifying a body of films that exhibit the same strong pattern establishes a norm against which films that deviate from the norm can be identified. It then becomes interesting to ask what form the deviation takes and why.

A technical issue raises its head at this point, which is the use of logarithmic transformation of the SLs. This is discussed in more detail in the Appendix; the practical value of such a transformation is that it can make it a lot easier to see what's going on. It does this by eliminating what might be called visual biases that arise in comparing SL distributions, attributable to the uneven distribution of mass along the curve. Taking the logarithm of lognormal data produces new data that have a normal distribution, and it is much easier to visually assess normality, and departures from it, than lognormality from the untransformed data. This is best illustrated by returning to Figure 2.4.

Brief Encounter and *Harvey* were chosen for illustration because, according to some formal methods of making the judgment, they are among the most and least lognormal SL distributions you could wish to meet (Redfern, 2012a). Is this apparent from the histograms to the left of the figure? It depends on how experienced you are at looking at this kind of thing; close inspection reveals differences, but their importance is hard to judge, particulary given that appearances of histograms can be affected by the manner of their construction.

That there is a radical difference is immediately apparent if the data are transformed logarithmically, and displayed using KDEs as in the right of the figure. The log-transformation evens up the visual weight given to longer and shorter SLs. *Brief Encounter* looks comfortably normal, and hence lognormal on the original scale; *Harvey* doesn't, the main departure from normality being two modes. The larger of these occurs at just over 3 on the log-scale which translates to about 25 seconds on the original scale. Reference back to the histogram reveals a bump in this region that the logged analysis suggests is genuine. The histogram for *Brief Encounter* is also uneven in this region but the logged analysis suggests it is not important variation. The statistical analysis here indicates a clear difference in patterns for the two films, that for *Brief Encounter*

of 'film style' miss the point(see Baxter, 2012a). Call the parameters that define the exact shape of the distribution μ and σ – they govern the position of the bulk of the data and its spread. The mathematics is in the Appendix, but briefly, the median is $\exp(\mu)$, where exp() is the exponential function. It involves only one parameter. The ASL is $\exp(\mu + \sigma^2/2)$, which combines both parameters but doesn't allow them to be separately distinguished. Using the ASL and median *jointly* is preferable to using either separately if a numerical summary is required.

being representative of the norm. Interpretation of what these difference mean in 'filmic' terms, if anything, is a substantive matter, outwith the remit of statistics[5].

2.2.4 Internal pattern - individual films

The examples so far focus on what I shall call 'external' pattern or structure. Statistics or graphs that characterise individual films can be studied for their own interest or, more interestingly, compared, with the aim of seeing if broader pattern, treating the individual patterns as elements to be ordered, can be discerned. This might be thought of as an attempt to discern 'global' patterning on the basis of external structure. Internal patterning or structure, the arrangement of shots within a film, is of equal interest, but quantitative approaches to this have perhaps received less systematic attention than external patterning. Chapters 7 and 8 explore possibilities, so the present example only touches on the possibilities.

Data on SLs submitted to the Cinemetrics database come in two versions, 'simple' where essentially only the SLs are recorded, and 'advanced' where shot type is recorded as well according to whetever categories the analyst chooses. The Paramount version of von Sternberg's *The Blue Angel* (1930), submitted by Charles O'Brien, is used for illustration in Figure 2.5

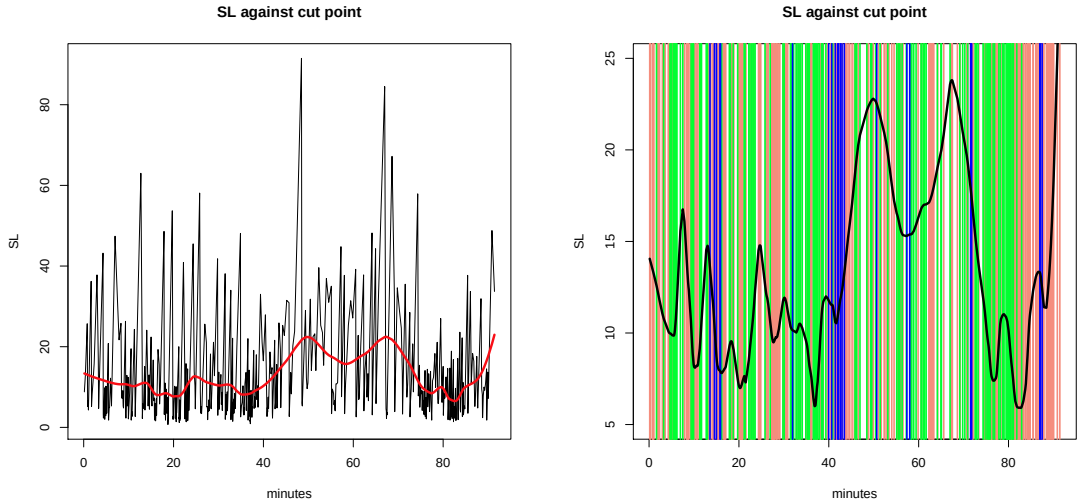

Figure 2.5: *Smoothed SL data for* The Blue Angel*, superimposed on backgrounds showing the actual SLs (left) and 'wallpaper' with colour-coded shot type (right), blue being 'singing', green 'dialog'; and salmon 'action'. (see text for more detail).*

The left-hand panel shows the SL data plotted against cut-point. As is typical the data are rather 'noisy' and a (loess) curve has been fitted through the data to smooth out the detail. An over-smoothed curve has been used to highlight the general trend in SLs where, for example, the faster cutting over the first 40 minutes or so can be contrasted with the slower and more variable cutting rate over the following 30 minutes. The range of SLs dictates the scale of the plot which can sometimes make it difficult to see the finer detail of the trend.

In the right-hand plot the full scale is dispensed with; a less smoothed curve is used to show more of the detailed variation; and shots are coded by colour according to the categorisation used

[5]Lognormality is the dominant paradigm, but it has been challenged (Redfern (2012a). I'd regard debate about this sort of thing as at the level of 'technique', the important idea being that a majority of SL distributions display a regularity capable of characterisation. I belive this can be done, but a detailed discussion is beyond the scope of these notes.

by O'Brien. Shots are coloured at the cut-points, so white spaces correspond to the longer shots.[6].

The smoothing chosen shows the same distinction between the earlier and later parts of the film as the smooth to the left, while also indicating more of the nature of variation in the first half of the film. The degree of smoothing chosen depends on the purpose of illustration and need not be confined to a single choice. 'Over-smoothed' curves are useful for showing very general trends in the patterning of SLs; considerably less smoothing is needed if the aim is to illustrate variation between 'scenes' (Salt, 2010). Cutting *et al.* (2011) have used SL data to investigate whether internal patterning is generally consistent with a four-act structure, following ideas of Thompson (1999), but have conceded, following an intervention by Salt, that their initial positive confirmation of the hypothesis was methodologically flawed (Cutting *et al.* (2011)).

The example illustrated above, and many of those in the relevant chapters to follow, are concerned with the detection and illustration of internal structure for individual films. What Cutting *et al.* attempt is something altogether more ambitious and challenging, the detection of global patterning on the basis of internal structure. This is challenging and some aspects of the challenge are illustrated in the next example. (The explanation is also more complex that other examples, so some readers might prefer to skip it on a first reading.)

2.2.5 Internal pattern - an example of global analysis

The problem with global comparison based on internal SL patterns is that the latter are not easily reduced to a form readily admitting synthesis of some kind. It is possible to look at the patterns for a small body of films and say how and why they are different, but how to express this in quantitative form that allows further statistical analysis is not obvious. The process is complicated by the fact that internal patterns may be viewed at different scales.

Shot lengths measured at a cut-point are an example of a time series. Aspects of time series can be quantified and compared across series. This is the kind of approach adopted in Cutting *et al.* (2010), one of whose analyses is described here by way of illustrating the complexity involved.

Salt (2010) expresses the view 'that variation in cutting rate is ordinarily used as an expressive device of a conventional kind - more cuts for sections where there is more dramatic tension or action, and less for less of the same', and that 'in general there is a conventional idea about alternating scenes with different dramatic character in plays and films, so that things like cutting rate and closeness of shot which are used expressively should change on the average from scene to scene'. That is, if correct, short SLs will tend to occur in 'clusters' with other short SLs, and similarly for long SLs. There are several ways in which this might be investigated, one such being the *partial autocorrelation function* (PACF). The left-hand panel of Figure 2.6 emulates, with embellishment, an example given in Figure 1 of Cutting *et al.* (2010) for *King Kong* (2005).

The solid vertical lines indicate the size of the partial autocorrelation at lags of up to 20 – that is, they measure the strength of relationship between the SLs of shots, 1, 2 , ..., 20 positions apart in the shot sequence. It is expected, if Salt's view is correct, that the first few of these will be 'significantly' positive. The statistical significance is judged by the horizontal dashed line (at $2/\sqrt{n}$) – if this line cuts a vertical line the partial autocorrelation is significant at that lag. What Cutting *et al.* (2010) call the *autoregressive index* (AR index) is determined by the position of the first lag that fails to be significant. This is at lag 6 (just) for *King Kong* so its AR index is 5. The larger the index, as Cutting *et al.* interpret it, the more clustered a film is into packets of shots of similar length.

As described above the AR index can only take on the values 0, 1, 2 .... The cut-off point for determining significance depends on the sample size (via $2/\sqrt{n}$) and Cutting *et al.* express the view that 'films with fewer shots are penalized; their bounds are higher, which tends to generate smaller AR indices'. They thus prefer to work with a modified autoregressive (MAR) index. This is obtained by fitting a smooth decay curve through the partial autocorrelations – the solid red curve in the figure – and defining the MAR index as the point of intersection with the solid

[6]Choice of background colour is interesting. If you choose (in)appropriately and omit the smoothed curve then, depending on the film, op-art effects that sometimes resemble the paintings of Bridget Riley can be achieved.

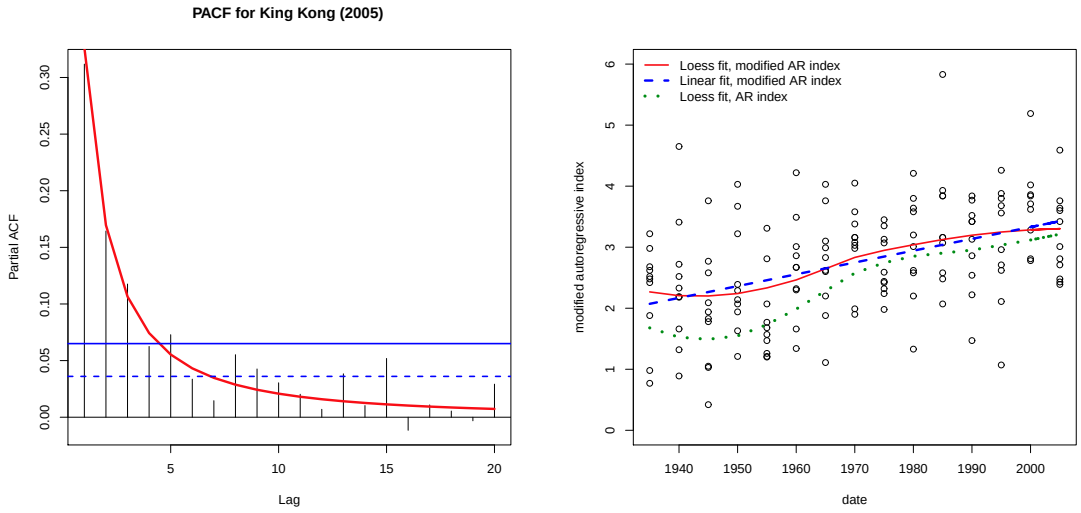

Figure 2.6: *To the left the partial autocorrelation function for the SLs of* King Kong *(2005), with a fitted curve. This is used to estimate a modified autoregressive index, plotted for 150 films in the right-hand plot. See the text for further details.*

horizontal line at 0.065, which is based on the mean number of shots in all films. For *King Kong* an MAR of 4.59 results[7].

I have some technical reservations about the procedures used which will be confined to a footnote; they might be summarised by saying that I think the analytical lily is being rather over-gilded[8]. Subsequent discussion, of the right-hand panel of Figure 2.6 is based on the results in the supplement to Cutting et al. (2010) and on their terms.

The MARs are plotted against date in the right-hand panel of Figure 2.6; the dashed (blue) straight line is that fitted to the data in Cutting *et al.* (2010) (the upper-left panel in their Figure 2). From it they conclude that the results suggest 'that Hollywood film has become increasingly clustered in packets of shots of similar length' and that in this and other matters 'film editors and directors have incrementally increased their control over the visual momentum of their narratives, making the relations among shot lengths more coherent over a 70-year span'. A more nuanced interpretation is possible.

The solid (red) curve (a loess smooth) does not presume linearity. There are subtle differences, disguised a little by the scale on which the curves are plotted. There is very little change over about 20 years from 1935 to about 1955. For the next 35 years or so there is a gradual increase, but also a point of inflection at about 1975 where the rate of increase seems to decrease, with a flattening out at about 1990. Either interpretation shows an increase over time, albeit a fairly modest one; that based on the loess smooth suggests that the increase mainly occurred over a period between abot 1955 to 1990 and may have flattened out. You really need more, and more representative data[9], and a longer time period, to be assertive about any of this; the conclusions

[7]Here and for other films I get slightly different results. This is, I presume, a result of algorithmic differences in the fitting process; I used the `nls` funtion in `R` with the same model described in Cutting *et al.*. Their results are used in the right-hand panel of the figure.

[8]If a fixed cut-off point is used to determine the MAR, basing it on all the data is sensible but it still introduces an arbitrary element into the definition. No account is taken of the imprecision of estimation of the fitted curves to the data which, to my eye, look as if they will be non-trivial in some cases. There will also be a dependency on the model used and cut-off at lag 20 (though the latter effect is probably small). I suspect – I've not checked – you'd get similar results to those in the paper by using a cut-off point to determine 'significance' from the original PACFs. You can live with the fact that these take on integer values only; the MAR, quoted to two decimal places, allows a 'finer' discrimination between films, but whether this is necessary might be questioned.

[9]Of the 150 films used those from 1980 on were selected to be high-grossing ones, and earlier films selected from

are interesting, but the more 'nuanced' interpretation possibly has less of a 'wow' factor than that in the original publication.

There is an interesting interchange between Salt and Cutting about the paper under discussion here on the Cinemetrics website. Start with http://www.cinemetrics.lv/salt_on_cutting.php and follow the thread. Cutting's second contribution explains what the PACF is in more detail than attempted here[10], and both commentators discuss the relationship of AR and MAR indices to other forms of representing internal structure covered in Salt (2010). Salt usefully discusses how structure of the kind both authors deal with, and changes over time, can emerge from film-making practice at the level of individual films.

2.2.6 Shot-scale analysis

Table 2.1 is an example of shot-scale data, for a sample of Fritz Lang films, extracted from Barry Salt's database on the Cinemetrics website. The analysis of such data was introduced in Salt (1974), and such analyses are a dominant feature of Chapters 12, 16, 19, 24 and 26 of Salt (2009). Shots are classified by scale as big close-up (BCU), close-up (CU), medium close-up (MC), medium shot (MS), medium long shot (MLS), long shot (LS) and very long shot (VLS). Illustrations of what is to be understood by these terms are provided in Salt (2009, p.156) and Salt (2006, p.253). The row numbers in the tables are scaled to add to 100%; Salt scales numbers to add to 500.

Title	Year	BCU	CU	MCU	MS	MLS	LS	VLS
Spinnen, Die (1)	1919	1	7	9	13	20	45	5
Spinnen, Die (2)	1920	6	13	7	9	32	30	3
Müde Tod, Der	1921	4	10	9	8	22	43	4
Dr. Mabuse der Spieler (1)	1922	6	12	14	14	25	29	1
Dr. Mabuse der Spieler (2)	1922	2	6	8	10	26	47	2
Kriemhilds Rache	1924	1	4	6	9	17	60	4
Siegfried	1924	1	6	7	8	17	56	5
Metropolis	1926	1	10	12	15	14	39	9
Spione	1928	6	9	15	21	25	23	1
M	1931	6	6	8	20	24	32	3
Testament des Dr. Mabuse, Das	1933	7	5	8	16	28	33	4
Fury	1936	3	13	15	23	24	20	2
You Only Live Once	1937	9	16	14	24	22	14	0
You and Me	1938	4	15	22	23	19	15	2
Hangmen Also Die	1943	4	8	16	19	26	26	1
Ministry of Fear, The	1945	3	11	14	22	22	27	2
Woman in the Window, The	1945	3	12	19	28	16	20	1
Cloak and Dagger	1946	2	10	17	30	17	20	3
Secret Beyond the Door, The	1948	3	5	17	26	24	24	1
Human Desire	1954	1	16	14	22	22	19	4
Beyond a Reasonable Doubt	1956	6	9	23	28	20	11	1
While the City Sleeps	1956	6	16	27	21	16	12	1
Indische Grabmal, Das	1959	2	12	11	13	20	41	2
Tiger von Eschnapur, Der	1959	0	9	11	14	20	42	4

Table 2.1: *Shot scale data (%) for films of Fritz Lang.*

The data are represented as *bar charts* in Figure 2.7 and something like this is the standard way of presenting such data. Salt (2009) provides about 180 examples in the chapters cited above[11].

Presented in this way, comparisons of whatever kind one wishes to effect – between specific films, directors, time periods etc. – requires visual comparison of the shapes of the bar charts, which is subjective and can be (unconciously) affected by the differential 'weight' given to absolute

those with high ratings on the Internet Movie Database.

[10] As it is relevant to other aspects of these notes, that Cutting's discussion of normality and lognormality (in his first contribution) is potentially misleading, needs mention. Data standardization, to zero mean and unit variance, and normalization (transformation to a normal distribution) are treated as the same thing. The confusion is common, but the distinction is an important one. Some of the examples Cutting uses to illustrate the prevalence of lognormality, based on histograms of log-transformed data, are also arguably misleading because the interval widths used for the histograms are too large to make useful judgments (Section 6.4.5).

[11] Notwithstanding what is in some of the cinemetric literature, 'bar charts', rather than 'histogram', is the appropriate form of diagram (and terminology) that should be used for the graphical representation of shot-scale data, and the bars should have gaps between them to make it clear that one is dealing with counts for categorised data, rather than grouped continuous data (e.g., SLs) for which histograms would be appropriate.

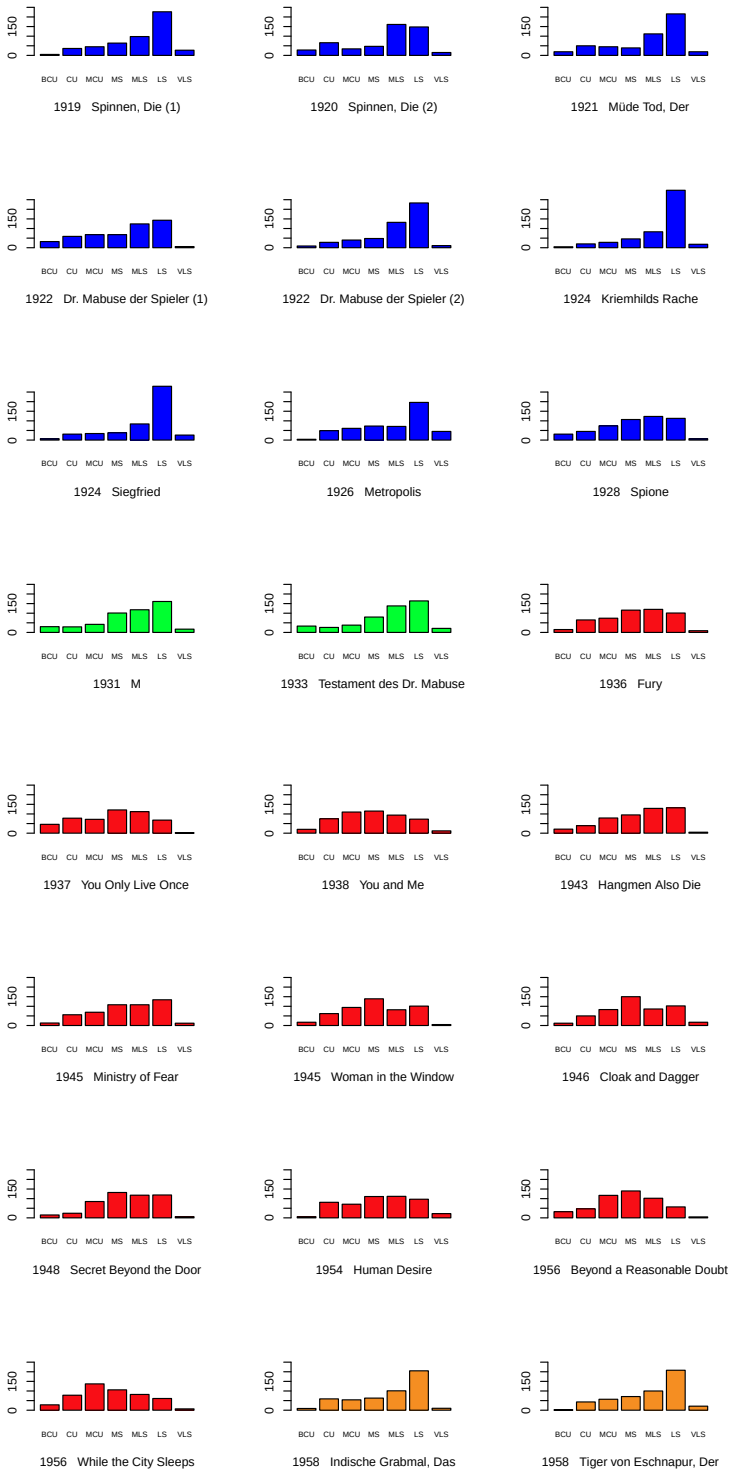

Figure 2.7: *Bar charts of shot-scale data for 24 films of Fritz Lang. Blue shading for early German silent films, green for early German sound, red for American and orange for late German.*

or relative differences in the heights of bars for different categories, or where a more 'holistic' appraisal of differences the general shape of the charts is attempted. Comparison can also rendered cumbersome by what may be an unavoidable lack of concision of presentation; some of Salt's illustrations are spread over five pages, and it can be difficult for a reader to follow an argument if reference to charts on several different pages is required.

None of this is to suggest that the conclusions derived from this kind of analysis are flawed – quite the opposite – but it can be time-consuming to arrive at them. What is now presented is an alternative approach to analysis, *correspondence analysis*, that operates in exactly the same spirit, but results in a much more concise representation of the data (one graph) and easier interpretation.

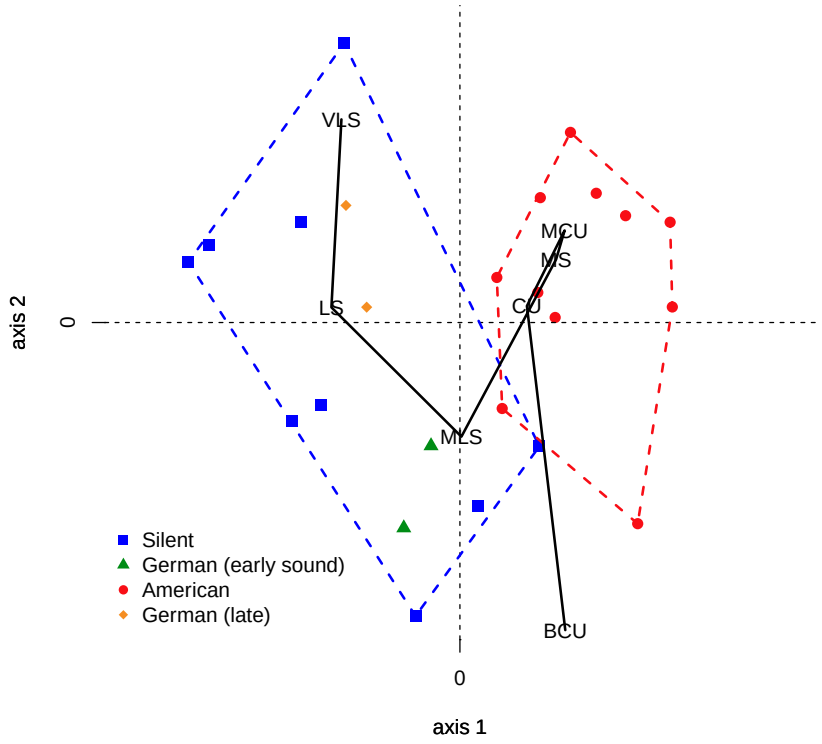

Figure 2.8: *A correspondence analysis for the shot-scale data of Table 2.1.*

Details of how correspondence analysis works are left till later. For current illustration it is enough to know that the rows of Table 2.1 can be reduced to points on a map. The distance between any two points on the map (i.e. their proximity) is an indicator of how similar their shot-scale profiles are. For example, in Figure 2.8 the two films furthest to the left in the upper-left quadrant, are very close, suggesting that they have very similar shot-scale profiles. The films in question are the *Die Niebelungen* films, *Siegfried* (1924) and *Kriemhilds Rache* (1924). The reasons for their similarity, and comparative difference from other films, can be confirmed by inspection of the bar charts or, for preference, the numbers in the table. It can be seen that the two films have fewer close-ups (of any kind and as a proportion) and a rather higher proportion of long shots than other films.

In the figure, films are labelled by type (silent/sound) and origin (German/American) and the German silents and American Films have been 'fenced in' by joining up the dots that define the

outer limits of their scatter[12]. The two sets of films are fairly clearly stylistically distinct – the only film corralled by both fences, and then only just, is *Hangmen Also Die* (1943). The cluster for the American films is a bit 'tighter' than that for the German silents.

The early German sound films, *M* (1931) and *Das Testament des Dr. Mabuse* (1933), sit more comfortably with the German silents than they do with the later American sound films. The late 'Indian' films, *Das Indische Grabmal* (1959) and *Der Tiger von Eschnapur* (1959), made after Lang returned to Germany in the late 1950s, are interesting in that they sit comfortably in the midst of the silent German films made 30-40 years earlier[13].

Having established that there is pattern in the data, a difference between the silent and American films in this case, it's then of interest to ask what features of the data are responsible for it. The correspondence analysis provides information on this as well. Markers corresponding to the columns, in this case shot-scales, can be plotted on the same graph as the row markers (films). Think of them as landmarks, defining the nature of the terrain in which they sit. The films partake, as it were, of more of the terrain corresponding to the landmarks to which they are closest and less of that of distant landmarks[14]. The closest marker to the *Niebelungen* films, for example, is that for long shots, and the furthest markers are those associated with the various degrees of close-up. This confirms the observations about these films, three paragraphs ago, based on the table and bar charts.

As further illustration of interpretation, the six films closest to the BCU marker are among the eight with percentages of 6% or more for BCUs in Table 2.1. These six are the two early German sound films, *M* (1931), *Das Testament Des Dr. Mabuse* (1933), three silents (from left to right) *Die Spinnen (1)* (1920), *Dr. Mabuse der Spieler (1)* (1922), *Spione* (1928), and one American sound film *You Only Live Once* (1937). The two American films with a relatively high proportion of BCUs that plot differently, late and the last Lang made before his return to Germany, are *Beyond a Reasonable Doubt* (1956) and *While the City Sleeps* (1956). They are the two films that plot furthest to the right in the top-right quadrant (on the same side of the graph as the BCU marker) and are distinguished from the other six films by having a noticeably larger proprtion of MCUs, to which they plot closely and fewer MLSs and LSs.

Nothing new is claimed for any of the interpretation here; Salt (2009, pp 242-243) covers the main points, for example. This does provide confirmation (or reassurance) that the correspondence analysis is producing sensible results and I'd argue its a more effective way of seeing what the data have to tell you.

A second illustration of correspondence analysis is provided for 33 sound films of Alfred Hitchcock (14 British and 19 American) from 1929 to 1963, 13 early German and Anerican sound films of Fritz Lang (1931-1956) and 18 films of Max Ophuls (1931-1955). The data are from Barry Salt's database, and have been selected to make some additional interpretive points. Labelled output from the correspondence analysis is shown in Figure 2.9.

In the previous illustration 'fencing-in' the Lang silents and American films was a fairly natural thing to do. The same might be done here, but is less helpful because the spread of some of the

[12]Called a *convex hull* in respectable statistical parlance.

[13]I'm certain I've seen the phrase 'reverted to his earlier style', or something very similar, used in connection with these films, but can't relocate it. It ought to occur round about pages 242-243 of Salt (2009) in his discussion of much the same films as here, but I can't find it there. Eisner (1976) notes that the films were based on a scenario by Lang and Thea von Harbou, also credited with the screen play for *Das Indische Grabmal* (1921) when the idea was that Lang would direct it, before it 'had been taken over by Joe May on the pretext that Lang was too young to direct this subject'. Eisner futher notes, of the later films, that occasionally 'the stylisation and deliberate abstraction recall the *Niebelungen*'. I don't know if this is relevant at all.

[14]Anyone familiar with the statistical literature on CA will know that a lot of effort has gone into discussing appropriate ways of plotting column markers in relation to row markers. One style of plotting places the column markers on the periphery of the plot some way from the row markers. It is often emphasised that you *cannot* interpret the difference between a row and column marker as a 'distance' in the way mathematicians understand the term. The interpretation offered here, therefore, is not a mathematically rigorous one. It is how interpretation of correspondence analysis plots is often done, however, and often works well, but its best to check conclusions make sense by reference back to the original data. For preference, partly to avoid this issue and partly to reduce over-crowding of plots, I'd usually plot row and column markers separately, but joint plotting is the default in much software.

data, and consequent overlapping of the convex hulls, obscures the dominant patterns in the data. What has been done is to show boundaries that encompass the bulk of the Hitchcock American, Lang and Ophuls films, excluding 3, 2 and 3 films respectively, to emphasise that these bodies of films, in terms of their shot-scale distributions, are largely stylistically different. The British Hitchcock films have been left as individual points as they are quite scattered and overlap with each of the groups that have been isolated here.

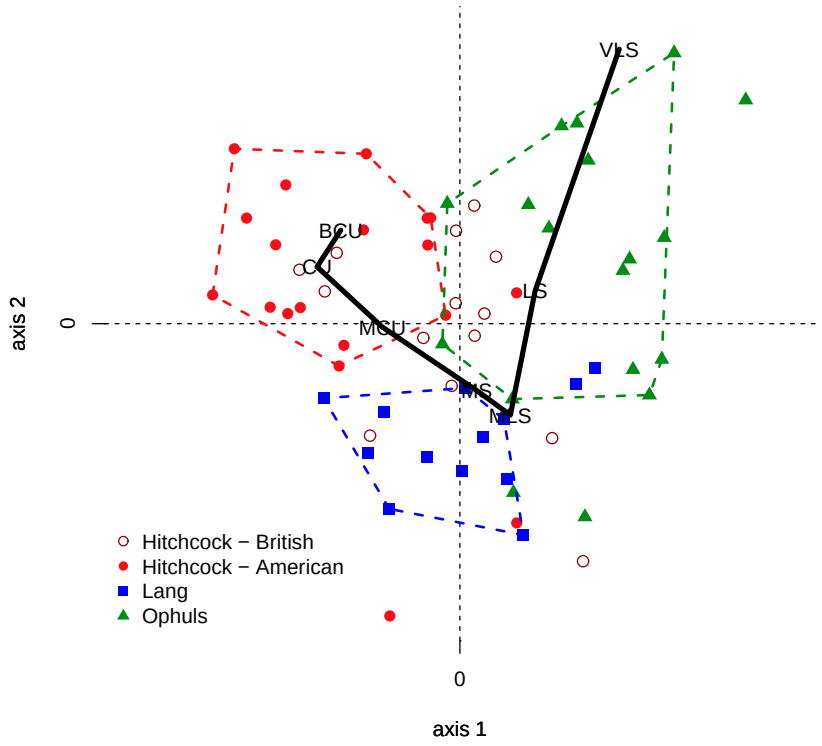

Figure 2.9: *A correspondence analysis of shot-scale data for sound films of Lang, Hitchcock and Ophuls.*

The two unenclosed Lang films are the earliest and only German ones in the sample used, *M* (1931) and *Das Testament Des Dr. Mabuse*. This separation from Lang's American films was also evident in Figure 2.8. Two of the three Ophuls films that sit closest to the pair, *Lachende Erben* (1932) and *Liebelei* (1932), are of comparable date. Their similarities can be confirmed by looking at the relevant bar charts in Figure 2.11, comparison across pages not being needed in this instance.

Three of the American Hitchcock's sit outside the main body. Reading from most to least distant (bottom to top on the plot) the 'outlying' films are *Dial M for Murder* (1954), *The Trouble with Harry* (1954) and *To Catch a Thief* (1955). Why these films differ from the bulk highlighted in the plot can be investigated using the bar charts, but from the correspondence analysis itself in can be inferred that the first two films make rather more use of MLSs and rather less use of CUs and BCUs than films in the main body (see shortly). The correctness of this inference is readily enough confirmed.

The three Ophuls films not enclosed by the convex hull are among his four latest films in the sample. Two, the lowest of the Ophuls films on the plot, *La Ronde* (1950) and *Madame de ...* (1953) are distinguished by having a somewhat higher ratio of MLSs to LSs and VLSs than the

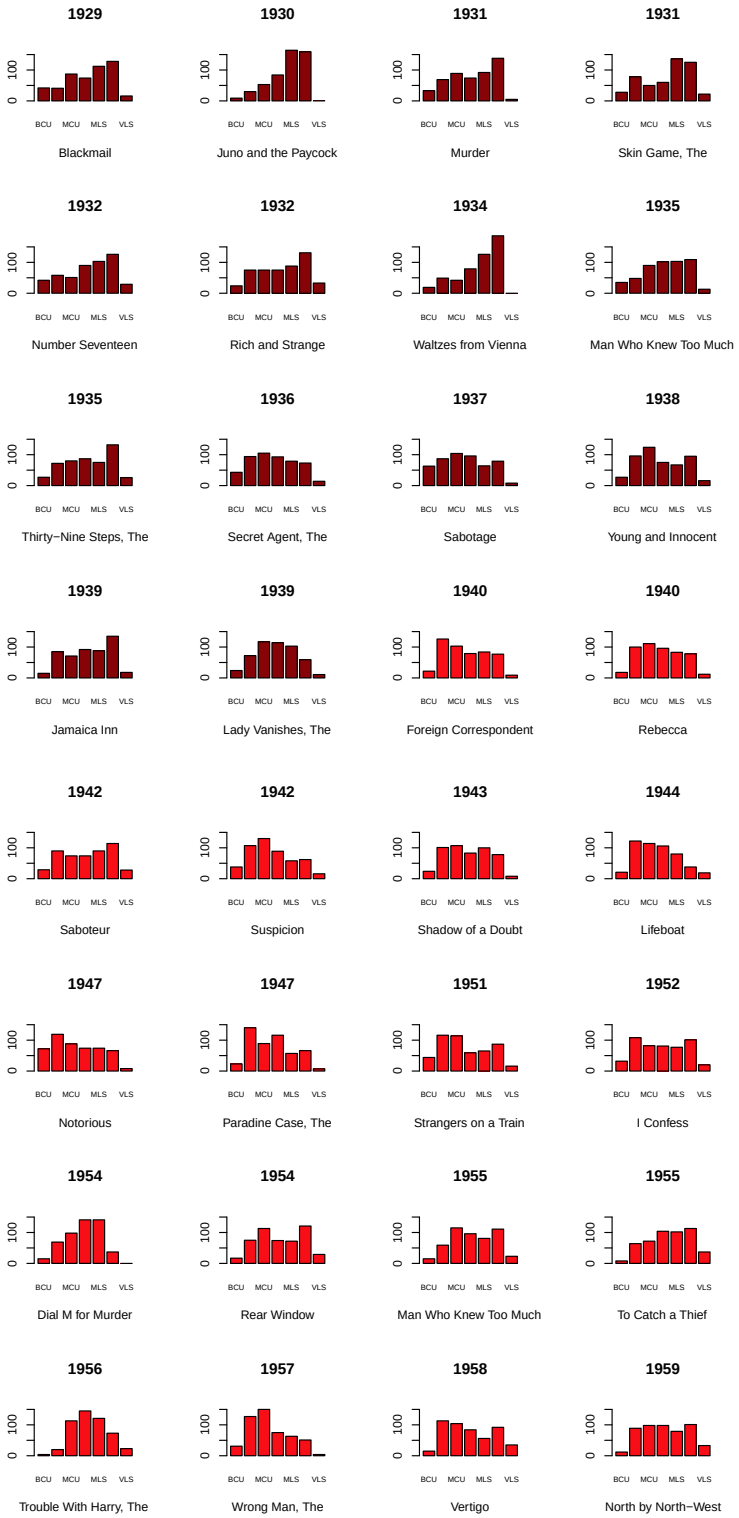

Figure 2.10: *Bar charts of shot-scale data for films of Hitchcock (dark red for British, red for American).*

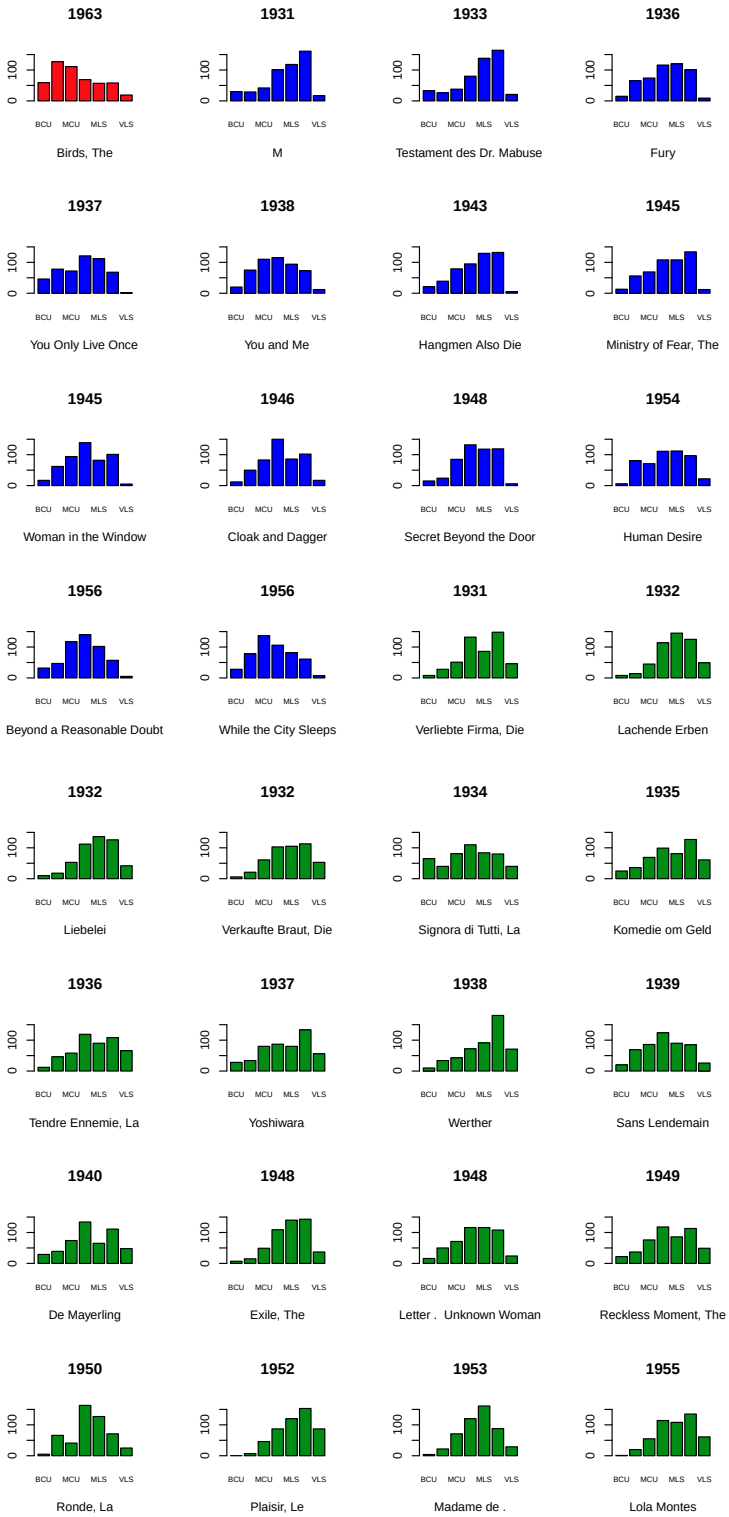

Figure 2.11: *Bar charts of shot-scale data for films of Hitchcock (red), Lang (blue) and Ophuls (green).*

typical Ophuls film. At the other extreme, sitting outside the convex hull, and the furthest right film on the plot, *Le Plaisir* (1953) has much more emphasis on LSs and VLSs compared to MLSs (precisely twice as many, compared to 0.83 and 0.73 for the other outlying Ophuls).

Detailed discussion of the British Hitchcock films is deferred to a later chapter. It suffices here to note that it is possible to see a stylistic development from early to late Hitchcock that is better appreciated if his films are examined in isolation from that of the other two directors. Salt (2009, p.243) comments on this in connection with a comparison with Lang, and their respective moves to Hollywood in the 1930s.

Detailed discussion of the interpretation of the shot-scale markers has been deferred until now. Correspondence analysis is widely used in quantitative archaeological applications (Baxter, 203, pp.136-146), often applied to tables of a form similar to Table 2.1 where rows correspond to archeological contexts, such as graves, and columns to artefact types (pottery, jewellery etc.) buried with the dead, table entries corresponding to counts of the artefact type, or simply 1 or 0 indicationg presence or absence. Fashions change, so if the contexts range over any great time span early graves would not be expected to share any artefact types in common with later graves. Graves of a similar period are expected to share some artefact types in common.

One use of correspondence analysis is to re-order the rows and columns of the table to reflect both the temporal sequence of burials and the associated developments in fashion of the artefact types. If subjected to a correspondence analysis, and if successful, the resultant plot – for both row and column markers – typically shows a fairly clear 'horseshoe' pattern that, subject to checking against other forms of evidence, can be interpreted as a temporal gradient[15]. This comes under the heading of *seriation* in archaeology.

From what I've seen, cinemetric data doesn't quite behave like this. You don't get the tight clustering of row markers (films) about a horseshoe that is hoped for in archaeological seriation. The shot-scale markers in Figure 2.9 behave as you might hope, and can be interpreted as a stylistic gradient. This is made more evident if you join up the dots corresponding to the shot-scale markers in their natural sequence from BCU to VLS. The horseshoe is possibly not as close to ideal as one might wish, but is as satisfactory as often occurs in reality. The configuration in Figure 2.8, isn't as satisfactory, which was why discussion was deferred at that point. It does, though, do a good job at separating scales in the MLS to VLS range from those closer up.

For those who prefer linearity (see the last footnote) it is possible to imagine straightening the horseshoe, the different ends carrying with them the points closest to them. Do this and you get a continuum with American Hitchcock dominating one end, Ophuls the other and Lang in the middle. British Hitchcock is spread around more, mainly in the middle and at the American Hitchcock end. Any evidence for a temporal gradient *within* the Hitchcock films, alluded to above, is not displayed with the labelling used here, and is better dealt with in a separate analysis.

[15]The horseshoe curve arises for mathematical reasons and is a non-linear gradient. This has troubled some, who would prefer their gradients linear (in vegetation science, for example). Techniques exist, such as detrended correspondence, that attempt to 'unbend' the horseshoe, but this deosn't usually trouble archaeologists, who are delighted to see the appearance of a horseshoe and vindication of their search for temporal structure.