

Multivariate Analysis goes to the Movies

Barry Salt

email: barrysalt@postmaster.co.uk

Abstract

This article investigates the style of motion pictures using some of the tools of multivariate analysis, and looks for stylistic evidence of directorial authorship.

Introduction

Fifty years ago, I created Statistical Style Analysis of Motion Pictures, as I called it, with an article of that name in *Film Quarterly* v 28 n 1, Fall 1974 (**Salt 1974**). Nowadays, the discipline would be called film stylometry, but the latter word was not used back then. The statistical analysis of literature and music already existed in a small way, and that was part of my inspiration. I continued working in this area, and the results of this can be read in my book *Film Style and Technology: History and Analysis 3rd.Ed.* (**Salt 2009**) and *Moving Into Pictures*, which can be downloaded for free from the *starword.com* website, along with my subsequent papers on the subject.

A few other researchers eventually joined in on this enterprise, starting with Charles O'Brien, and some of his extensive work is contained in *Cinema's Conversion to Sound: Technology and Film Style in France and the U.S.* (**O'Brien 2004**).

In 2005 Yuri Tsivian created the Cinemetrics website which at first concentrated solely on the statistics of film shot lengths, but soon broadened out to be a collecting point for other film statistics. Others to be mentioned include Warren Buckland in parts of his book *Directed by Steven Spielberg: Poetics of the Contemporary Hollywood Blockbuster* (**Buckland 2006**) and James Cutting with *Movies on Our Minds: The Evolution of Cinematic Engagement* (**Cutting 2021**).

My endeavour is part of what I call 'Practical Film Theory', and a fuller description of this can be found in my *Film Style and Technology: History and Analysis* (**Salt 2009**). The essence of my analytical methods is the reversal of the way films are put together by their makers, and using the concepts and terms that they use. I establish norms for the contexts of individual

films, and the differences of the variables of a particular film from the relevant norms are what indicate that variable's special significance. I have measured an array of formal statistics for 385 films, and the tables of these can be found on the *starword.com* website.

Fifty years ago, computers were not readily available, and all this data had to be recorded with pencil and paper while examining the films on film editing tables. Then the tabulated results were inspected by eye to detect the correspondences and regularities in them. Since then, new methods of multivariate analysis have appeared, and been used in classification tasks in research in sciences such as biology. These methods have in recent times been adapted to investigate literature and other arts. They seem to work there, even though their subject is things created by human intentions, not by the "blind watchmaker" of biological evolution. So I will now use some of these tools of multivariate statistics to analyse my data, and so make observations about it that are slightly more objective than previously.

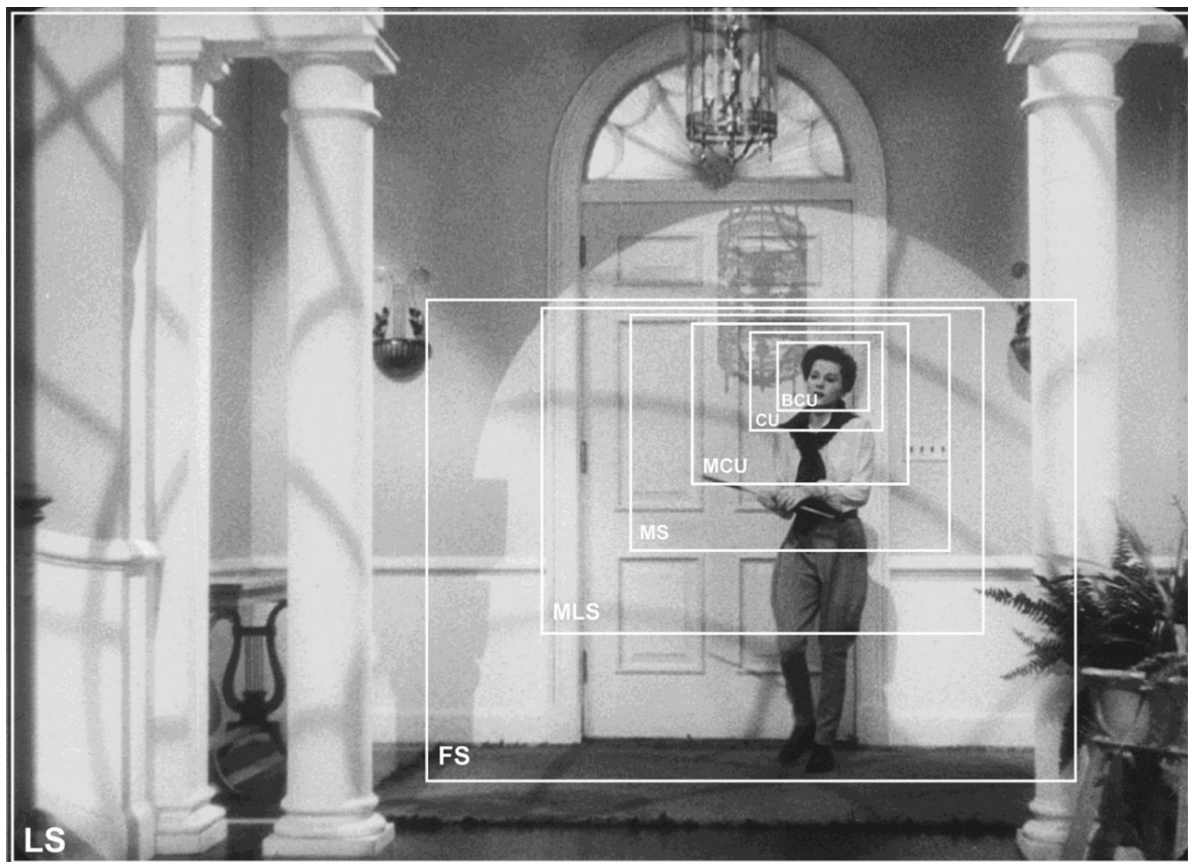
The Analysis

The corpus of films I will consider here comprises a varied collection of 140 films made from 1932 to 1959. The restriction to the period between 1932 and 1960 is to avoid films made under the distorting restrictions of early sound cinema, and films made after 1960, when the **Great Speed Up (Salt 2021)** began.

The style variables considered are firstly ASL, the average length (arithmetic mean) of the shots in seconds. Then RA, the number of reverse-angle cuts with an angle change greater than 90 degrees between shots per 500 shots. Next, POV, the number of cuts to or from a watcher and what they see, taken along their line of sight. INS is the number of shots of an object other than a recognizable person, again taken per 500 shots. Finally, there is a series of categories of Scale of Shot.

Scale of Shot is measured by the size of the human figure relative to the height of the frame. The scales I use are as follows: Big Close Up (BCU) shows head only, or even less than that, Close Up (CU) shows head and shoulders, Medium Close Up (MCU) includes body from the waist up, Medium Shot (MS) includes from just below the hip to above the head of upright actors, Medium Long Shot (MLS) shows the body from the knee upwards, Long Shot (LS) shows at least the full height of the body, and Very Long Shot (VLS) shows the actor small in the frame. As usual, the figures quoted are the number of each scale per 500 shots in the film concerned. An illustration of these scales can be found on the front cover of *Moving Into*

Pictures (Salt 2006) (the category of Full Shot (FS) is included within the Long Shot category in most of my data.) These scales are illustrated in the figure below. (Fig 1)



For a smaller number of films I have also measured the kinds of camera movements used in shooting them. These variables are Pan, Tilt, Pan with Tilt, Tracking, Tracking with a Free Head, Craning, Zooming, and Zooming with Pan and Tilt. (For my present purposes I have combined my previous categories of Tracking with Pan, and Tracking with Pan and Tilt into one variable called Tracking with a Free Head. This is the term used in the film industry.) Like the scale of shot variables, the values of these movement variables are normalized to the numbers per 500 shots.

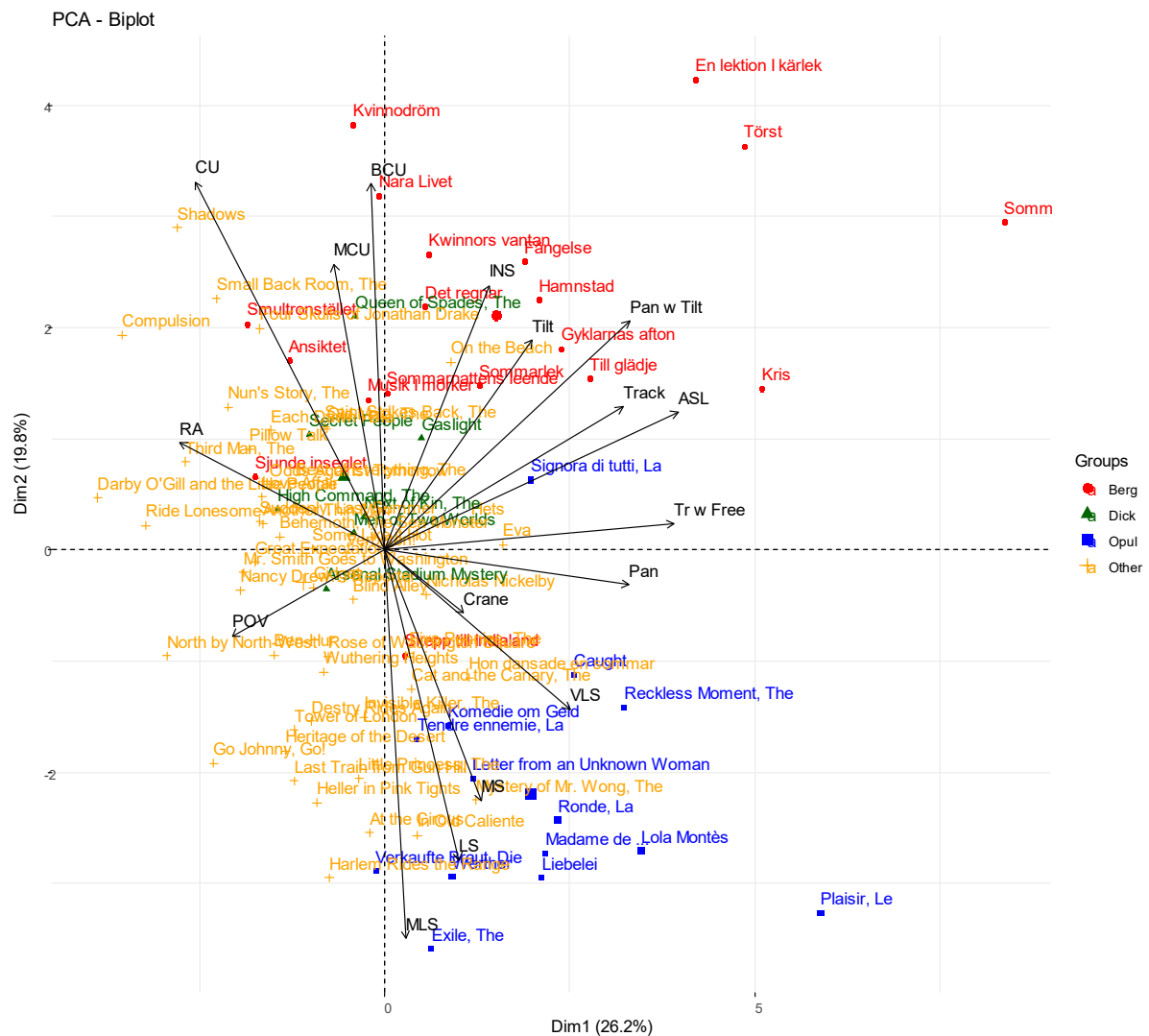
And here is a list of the 67 films in my first sub-corpus. (Table1). The “Graphic Title” is used for assigning colours to the titles *after* the analysis.

GRAPHIC TITLE	TITLE	YEAR	DIRECTOR
Other_At the Circus	At the Circus	1939	Buzzell, Edward
Other_Mr. Smith Goes to Washington	Mr. Smith Goes to Washington	1939	Capra, Frank
Other_Nicholas Nickelby	Nicholas Nickelby	1947	Cavalcanti, Alberto
Other_Nancy Drew ... Reporter	Nancy Drew ... Reporter	1939	Clemens, Wiliam
Other_Of Human Bondage	Of Human Bondage	1934	Cromwell, John
Dick_High Command, The	High Command, The	1937	Dickinson, Thorold
Dick_Arsenal Stadium Mystery, The	Arsenal Stadium Mystery, The	1939	Dickinson, Thorold

Dick_Gaslight	Gaslight	1940	Dickinson, Thorold
Dick_Next of Kin, The	Next of Kin, The	1942	Dickinson, Thorold
Dick_Men of Two Worlds	Men of Two Worlds	1946	Dickinson, Thorold
Dick_Queen of Spades, The	Queen of Spades, The	1948	Dickinson, Thorold
Dick_Secret People	Secret People	1951	Dickinson, Thorold
Other_Harlem Rides the Range	Harlem Rides the Range	1939	Kahn, Richard C.
Other_In Old Caliente	In Old Caliente	1939	Kane, Joseph
Other_Each Dawn I Die	Each Dawn I Die	1939	Keighley, William
Other_Tower of London	Tower of London	1939	Lee, Rowland V.
Other_Destroy Rides Again	Destroy Rides Again	1939	Marshall, George
Other_Love Affair	Love Affair	1939	McCarey, Leo
Other_Invisible Killer, The	Invisible Killer, The	1939	Newfield, Sam
Other_Mystery of Mr. Wong, The	Mystery of Mr. Wong, The	1939	Nigh, William
Other_Cat and the Canary, The	Cat and the Canary, The	1939	Nugent Elliott
Opul_Liebelei	Liebelei	1932	Ophuls, Max
Opul_Verkaufte Braut, Die	Verkaufte Braut, Die	1932	Ophuls, Max
Opul_Signora di tutti, La	Signora di tutti, La	1934	Ophuls, Max
Opul_Komedie om Geld	Komedie om Geld	1935	Ophuls, Max
Opul_Tendre ennemie, La	Tendre ennemie, La	1936	Ophuls, Max
Opul_Werther	Werther	1938	Ophuls, Max
Opul_Exile, The	Exile, The	1948	Ophuls, Max
Opul_Letter from an Unknown Woman	Letter from an Unknown Woman	1948	Ophuls, Max
Opul_Caught	Caught	1949	Ophuls, Max
Opul_Reckless Moment, The	Reckless Moment, The	1949	Ophuls, Max
Opul_Ronde, La	Ronde, La	1950	Ophuls, Max
Opul_Plaisir, Le	Plaisir, Le	1952	Ophuls, Max
Opul_Madame de ...	Madame de ...	1953	Ophuls, Max
Opul_Lola Montès	Lola Montès	1955	Ophuls, Max
Other_Rose of Washington Square	Rose of Washington Square	1939	Ratoff, Gregory
Other_Heritage of the Desert	Heritage of the Desert	1939	Selander, Lesley
Other_Another Thin Man	Another Thin Man	1939	Van Dyke, W.S.
Other_Blind Alley	Blind Alley	1939	Vidor, Charles
Other_Wuthering Heights	Wuthering Heights	1939	Wyler, William
Berg_Kris	Kris	1946	Bergman, Ingmar
Berg_Det regnar på vår kärlek	Det regnar på vår kärlek	1946	Bergman, Ingmar
Berg_Skepp till Indialand	Skepp till Indialand	1947	Bergman, Ingmar
Berg_Musik I mörker	Musik I mörker	1948	Bergman, Ingmar
Berg_Hamnstad	Hamnstad	1948	Bergman, Ingmar
Berg_Fängelse	Fängelse	1949	Bergman, Ingmar
Berg_Törst	Törst	1949	Bergman, Ingmar
Berg_Till Glädje	Till Glädje	1950	Bergman, Ingmar
Berg_Sommarlek	Sommarlek	1951	Bergman, Ingmar
Berg_Kvinnors vantan	Kvinnors vantan	1952	Bergman, Ingmar
Berg_Sommaren med Monika	Sommaren med Monika	1953	Bergman, Ingmar
Berg_Gycklarnas afton	Gycklarnas afton	1953	Bergman, Ingmar
Berg_En lektion I kärlek	En lektion I kärlek	1954	Bergman, Ingmar

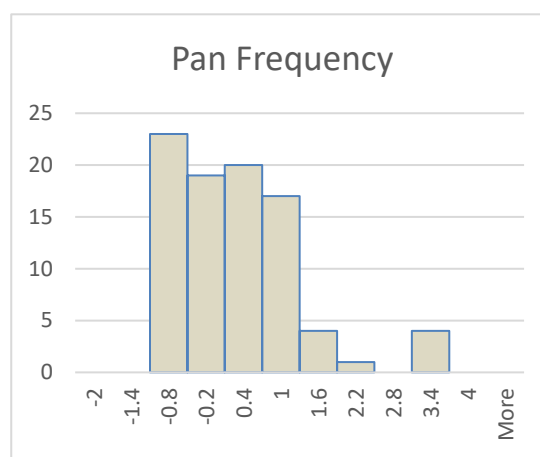
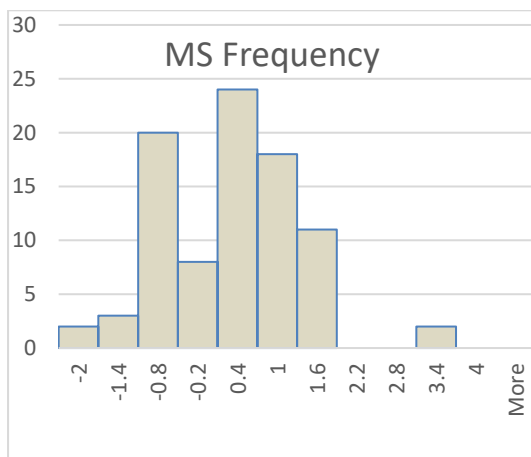
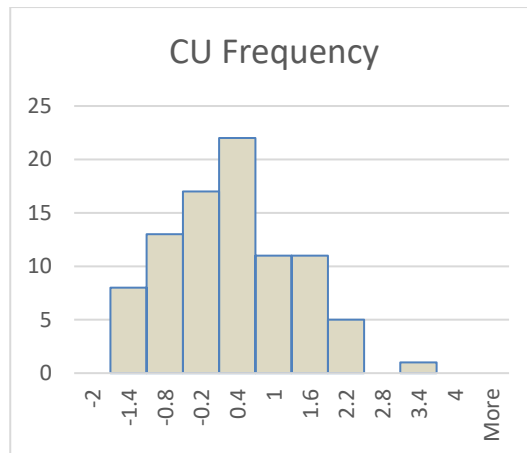
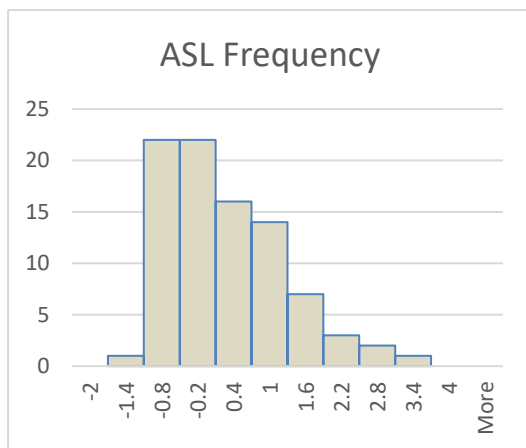
Berg_Kvinnodröm	Kvinnodröm	1955	Bergman, Ingmar
Berg_Sommarnattens leende	Sommarnattens leende	1955	Bergman, Ingmar
Berg_Det sjunde inseglet	Det sjunde inseglet	1957	Bergman, Ingmar
Berg_Smultronstället	Smultronstället	1957	Bergman, Ingmar
Berg_Nara livet	Nara livet	1958	Bergman, Ingmar
Berg_Ansiktet	Ansiktet	1958	Bergman, Ingmar
Other_Nicholas Nickelby	Nicholas Nickelby	1947	Cavalcanti, Alberto
Other_Heller in Pink Tights	Heller in Pink Tights	1960	Cukor, George
Other_Great Expectations	Great Expectations	1946	Lean, David
Other_Small Back Room, The	Small Back Room, The	1949	Powell & Pressburger
Other_Third Man, The	Third Man, The	1949	Reed, Carol
Other_Hon dansade en sommar	Hon dansade en sommar	1951	Mattsson, Arne
Other_Eva	Eva	1948	Molander, Gustav
Other_Hets	Hets	1944	Sjöberg, Alf

The style data for these films is in the supplementary tables to this article. (My previous stylistic analysis of the style of Ingmar Bergman's films (Salt 2020), which contains a list of the many different English titles they were given by distributors, can be found on the *starword.com* website.) A Principal Components analysis (PCA) of the style data produces the following graph (**Fig 2**):



The first principal component (Dim 1) is heavily weighted for camera movement, which produces the separation between the Bergman and Ophuls films to the right, and everything else to the left. The Average Shot Length (ASL) also contributes to the right-left separation and shows up the general tendency of longer takes to be associated with more camera movement. The second principal component (Dim 2) is largely influenced by the Scale of Shot – whether the shots are closer or more distant to the actors. This is what produces the separation of the Ophuls films and the Bergman films from one another. The Thorold Dickinson films cluster close to the general norm in among the other films. In other words, they are very typical for the period I am considering. The furthest outliers are obviously the Bergman films *En lektion i kärlek*, *Törst*, and *Sommaren med Monika* and Max Ophuls' *Le plaisir*, pushed there by their excessive amount of camera movement.

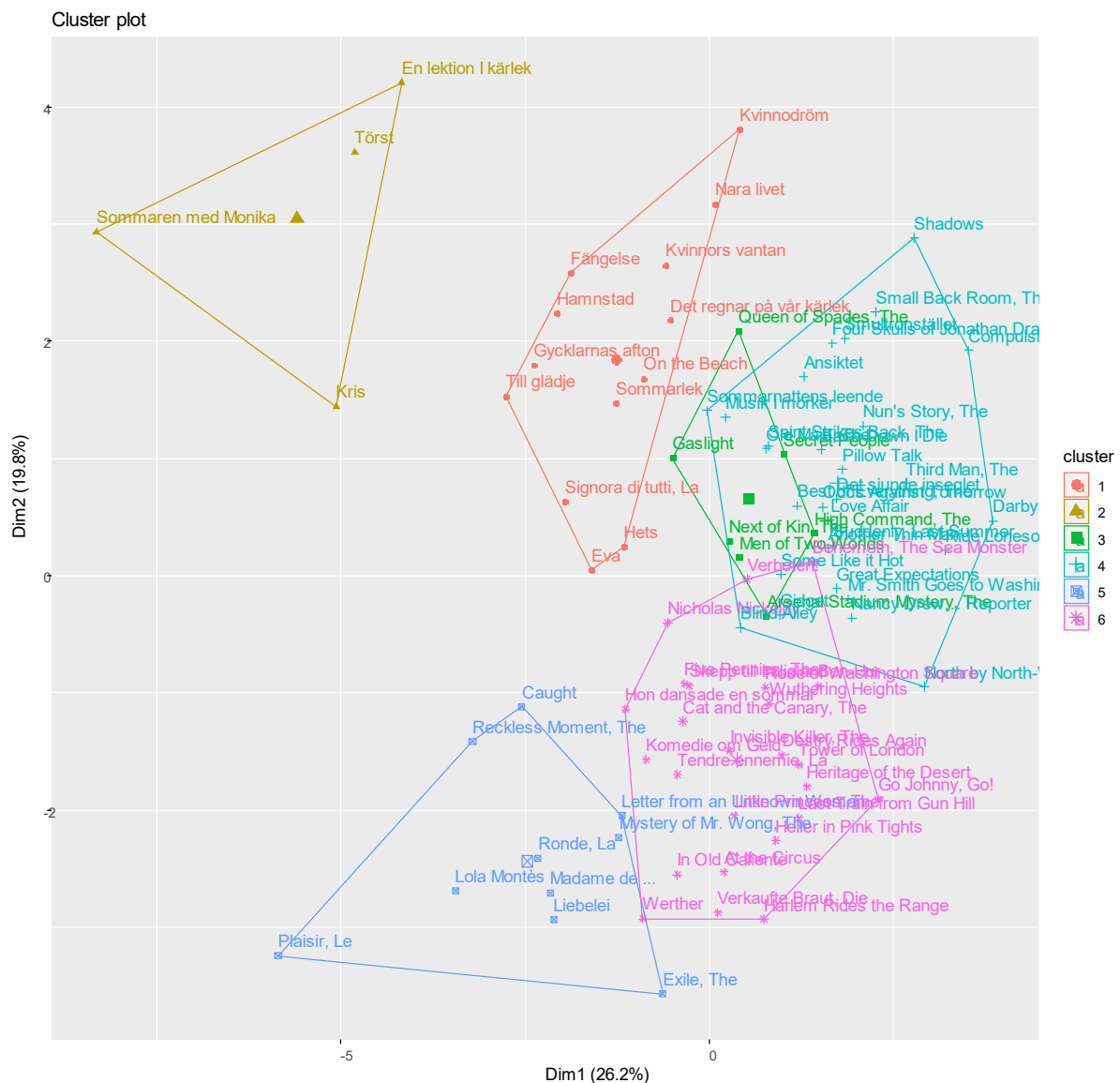
The denser clustering towards the centre of the graph reflects the underlying distribution of the values of each of the style variables in their own space. This can be seen in histograms for some of the distributions of the style variables below. (Fig. 3)



These are indicative of the general nature of these distributions. The Scale of Shot distributions are symmetrical distributions, roughly like the Gaussian (or Normal) distribution, whereas the others are positively skewed. The distribution of ASL values is fairly like a LogNormal distribution. The numbers of values are too small to produce ideally smoothed distributions.

My analysis, which uses 17 style variables, works differently, but analogously, to literary stylistic analysis where a thousand or more different words are usually the variables.

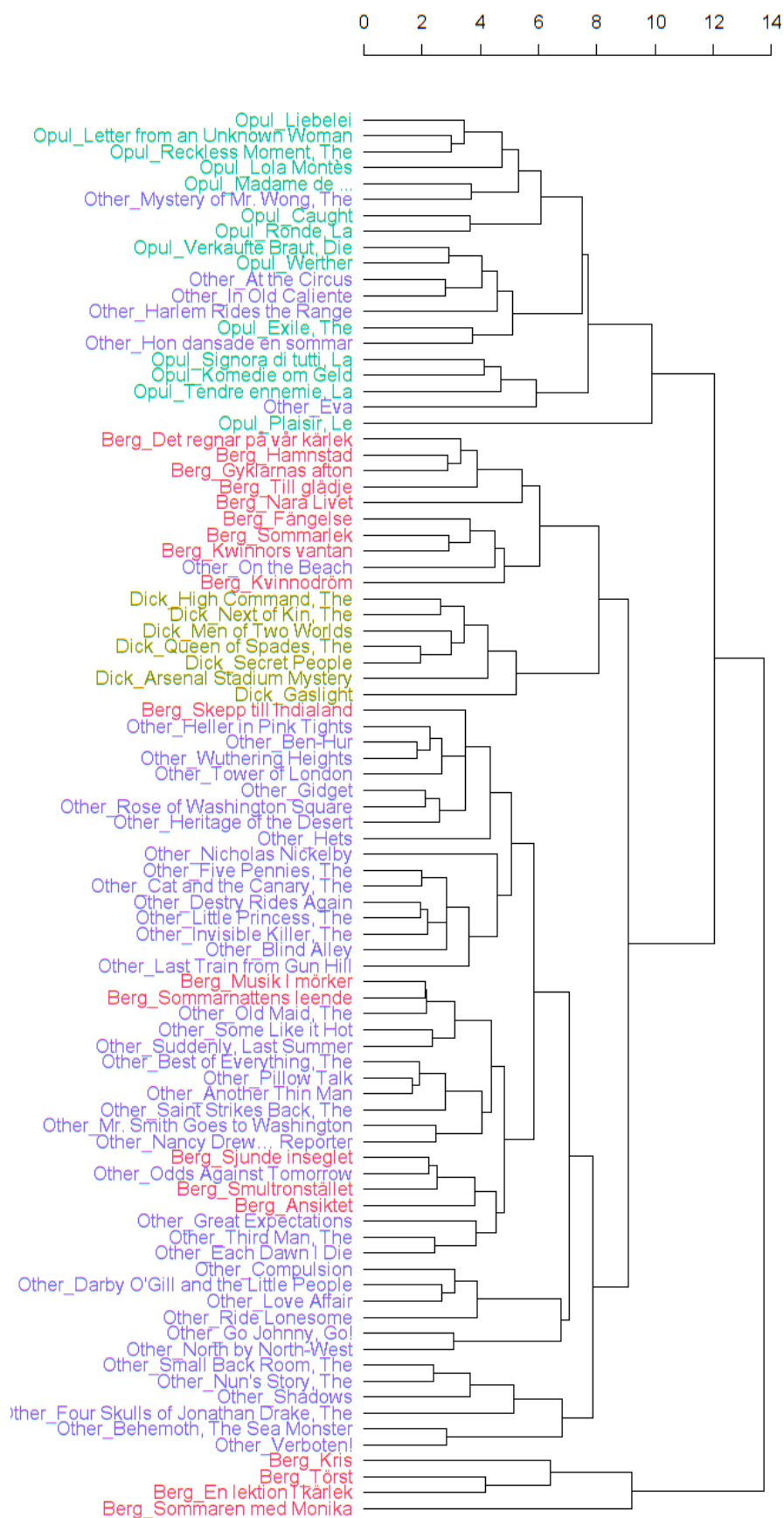
We can also use K-means cluster analysis on this corpus, and in this case the “silhouette” test clearly suggests the use of six clusters. **(Fig. 4)**



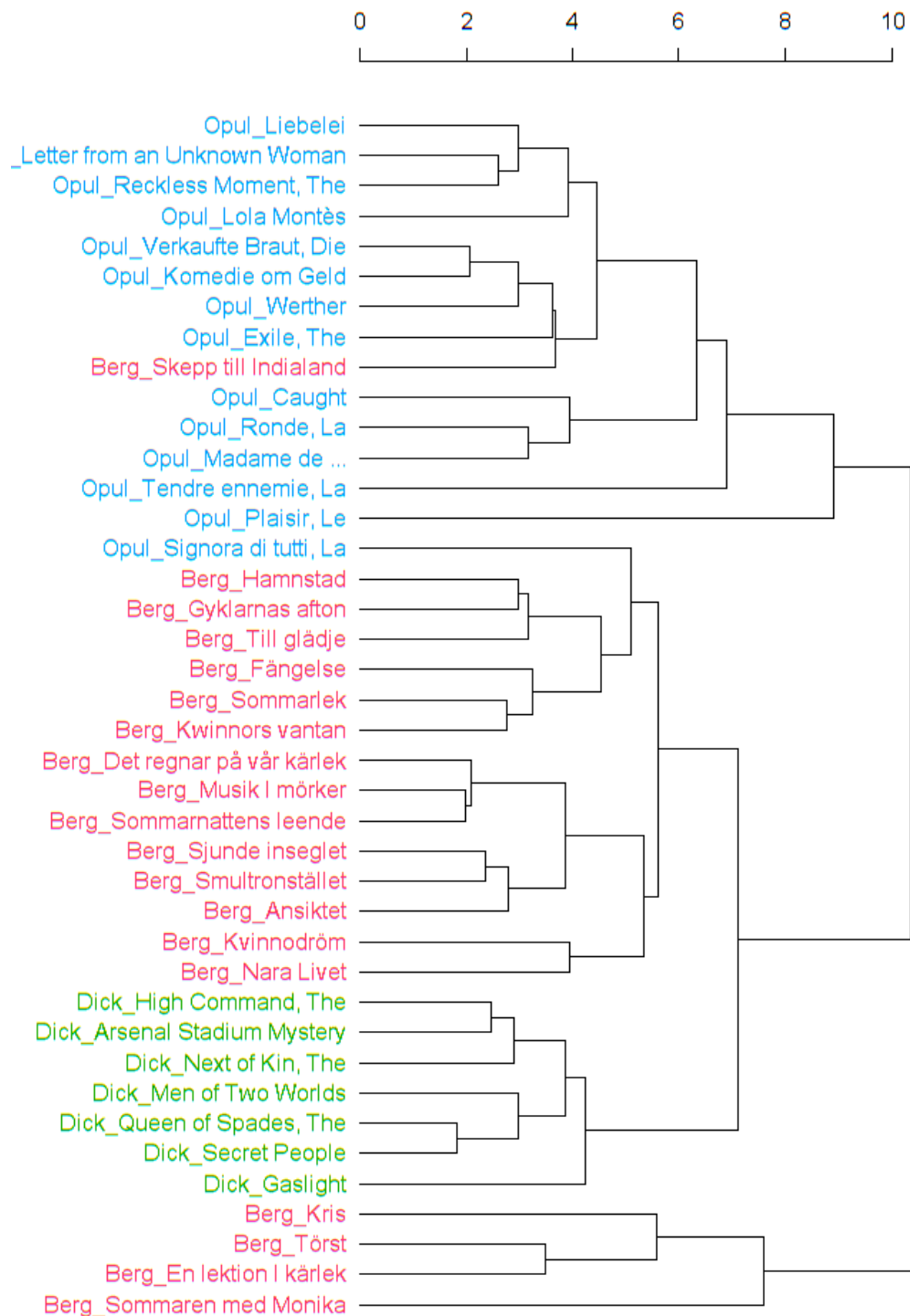
(The graphical space in which these results of the K-means method is displayed is the same as that of the principal components analysis, except that it is left-right inverted.)

In this classification of the films the Bergman films are split into two groups (Cluster1 and Cluster 2), and they cannot be reunited by reducing the number of clusters. The split corresponds to the markedly long ASLs (above 20 seconds) of *Sommaren med Monika*, *En lektion i kärlek*, *Kris* and *Törst* compared to the other Bergman films. The Dickinson films are all in Cluster 3 on their own, and most of the Ophüls films are in Cluster 5. The other contextual films are broken down into two clusters, Nos. 4 and 6. These last two clusters of the “Other” films are mostly separated by where they stand on the Scale of Shot statistic. A few films made by the three named directors are in the clusters of contextual films. They include Bergman’s *Ansiktet*, *Musik i mörker*, *Det sjunde inseglet*, and *Smultronstället*, and are put there by their relative lack of camera movement and shorter ASL. Five Ophüls films are in the lower contextual film cluster for the same reason.

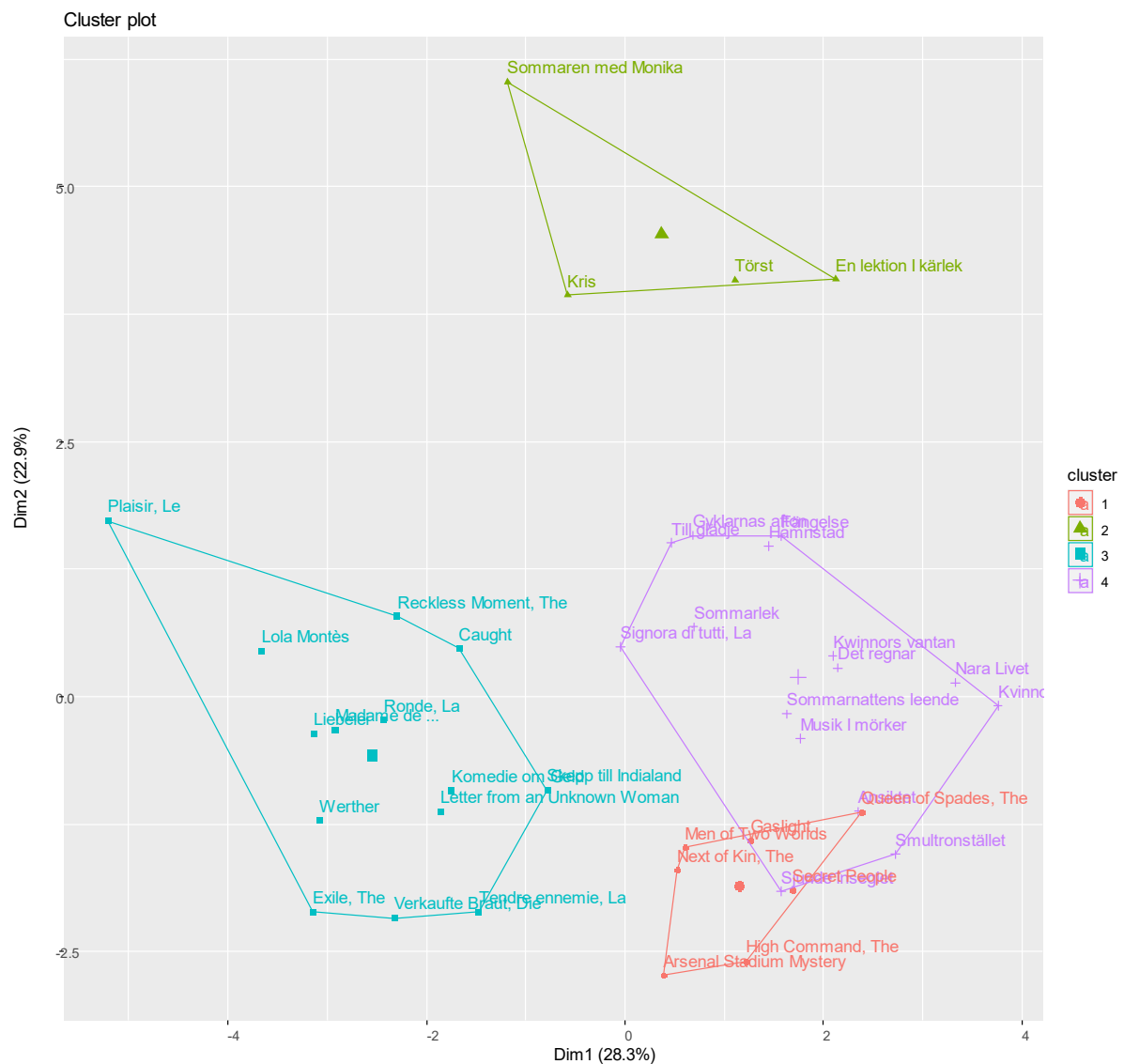
Another major form of cluster analysis uses dendrograms to display the results. In this area I choose agglomerative hierarchical clustering using the DIANA method to test out the data. Here the contextual films are labelled with an “Other_” prefix, and are coloured blue. **(Fig. 5)**



The classification is rather similar to that produced by K-means, with the Dickinson films still hanging closely together, and the Bergman films split into several groups. The Ophuls films are almost a group, but penetrated by a scattering of “Other” films. Eliminating the “Other” films from the Diana analysis gives the following dendrogram (**Fig. 6**)



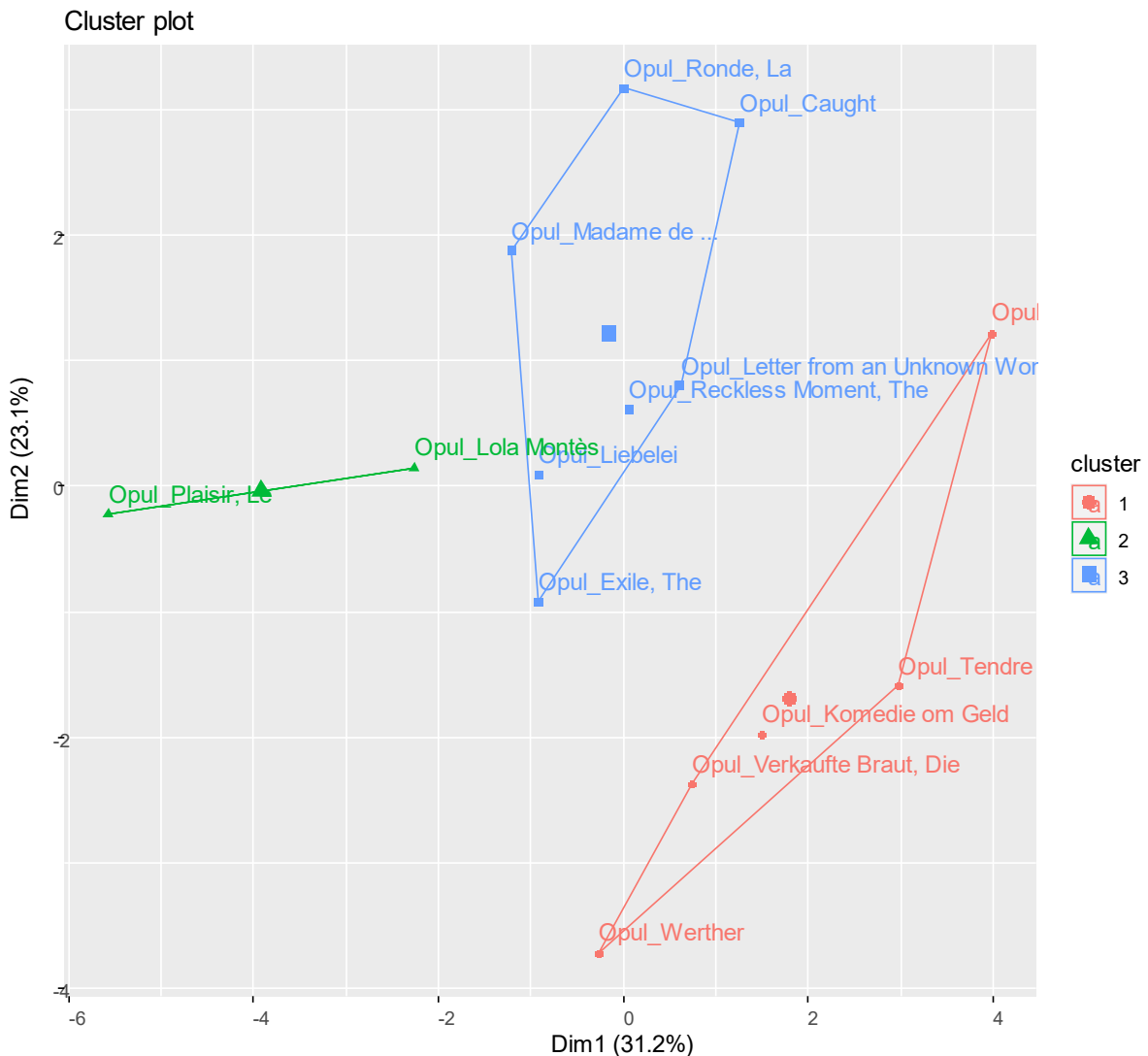
Using K-means to classify the same set of films produces this: (Fig. 7)



Which is almost exactly the same as the result of the Diana algorithm and its dendrogram, with the Bergman films breaking up into two groups as before. The very long take films *Sommaren med Monika*, *En lektion i kärlek*, *Kris* and *Törst* are separated off again, and *Skepp till Indialand* is still in the Ophüls group.

Skepp till Indialand is not shot in the same way as the other Bergman films, as most of the action is shot in and on a small salvage barge, giving very limited opportunities for camera movement.

Cluster analysis can also be used to investigate any style changes over time in the work of a film director. For instance, doing a K-means analysis of the films of Max Ophüls included in my sample produces the following: (Fig. 8)



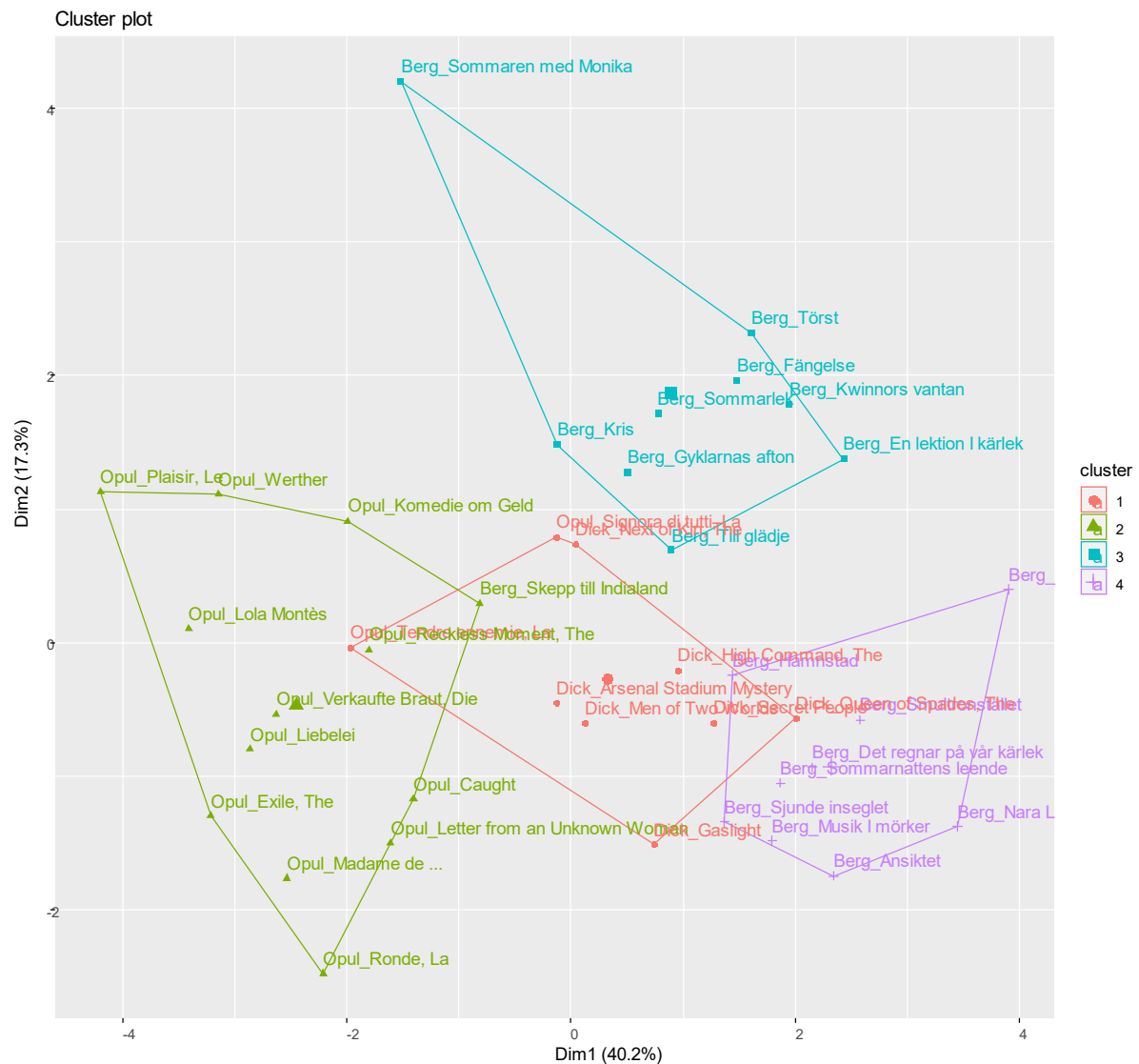
This classification is slightly more objective confirmation of my analysis in Chapter 26 of Ophüls' films in *Films Style and Technology* (Salt 2009).

My discussion on page 396 of that book of the question as to whether Ophüls has a single style is not completely clarified by the present series cluster analyses. The K-means analysis of the whole corpus in **Fig. 4** has eight of the Ophüls films separated out in one cluster, leaving five Ophüls films mixed in with one cluster of the contextual films. On the other hand, in the restricted subset without the contextual films clustered by k-means (**Fig. 7**), all the Ophüls films are in one separate cluster, only contaminated by the inclusion of Bergman's *Skepp till Indialand*. So you could say that Max Ophüls has a distinct style on this evidence. Nevertheless, on the basis of (**Fig. 8**) above, his films do break down into four style periods through his career, as I claimed in Chapter 26 of *Film Style and Technology*. *Liebelei* has a style that Ophüls dropped for the rest of the 'thirties, as his other 'thirties films are in a separate cluster. His American films, and subsequent European films, return to the style of *Liebelei*, except that, at the end of his career, he turned to a new and separate style with *Le Plaisir* and *Lola Montès*. These two films are shot much further back than the previous films,

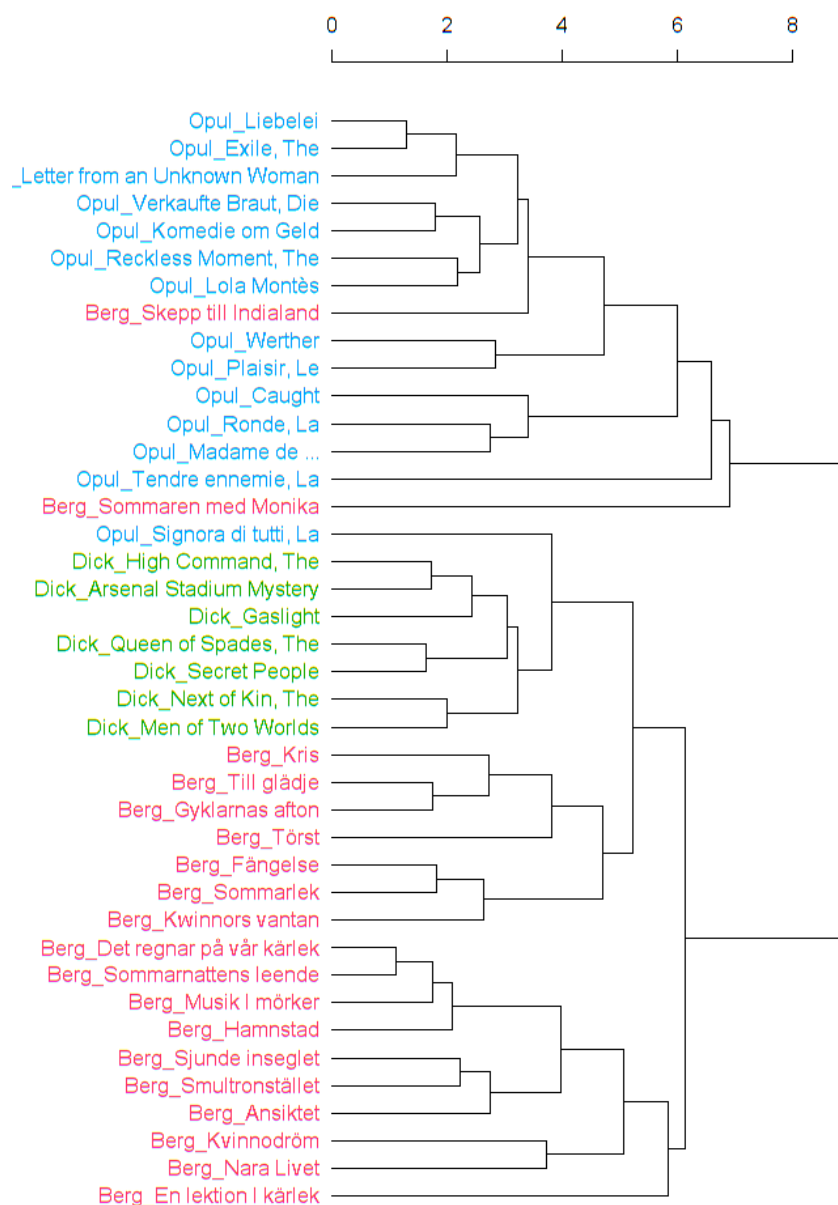
with longer takes and more camera movement. In *Film Style and Technology* Chapter 26, I also included *Madame de ...* in this group, but I now see that I was wrong.

Reducing the Variables

It is worth investigating the result of reducing the number of variables descriptive of the films from 17 to 11, by excluding the camera movement variables. Then we use this restriction in the K-means classification on the present group of films we get: **(Fig. 9)**



Applying the DIANA cluster analysis to this same sub-set of films gives a very slightly better classification. (**Fig. 10**)



Both of these classifications using 11 variables are almost identical to the cluster analysis using 17 variables, which is surprising.

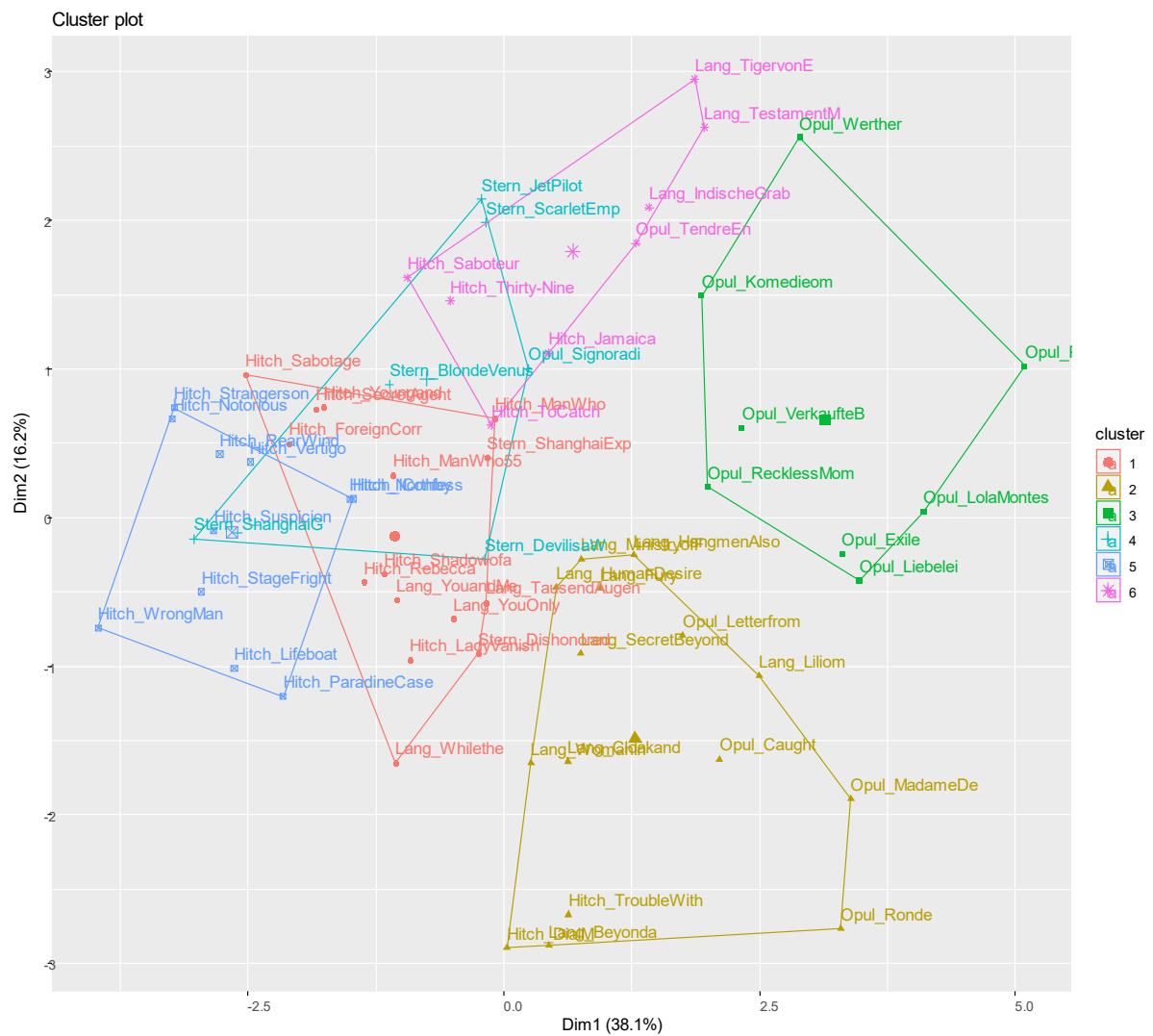
I have a much larger collection of other style data that only uses this 11 variable analysis. From that I can extract a sample of 60 films from 1932-1960 including the following films made by directors, mostly working in the American film industry, who are considered to be *auteurs* – that is sufficiently in control of their films to make them on subjects they chose, and also make them the way they want to.

(Table 2)

GRAPHIC TITLE	FILM TITLE	YEAR	DIRECTOR
Hitch_ManWho	Man Who Knew Too Much, The	1935	Hitchcock, Alfred
Hitch_Thirty-Nine	Thirty-Nine Steps, The	1935	Hitchcock, Alfred
Hitch_SecretAgent	Secret Agent, The	1936	Hitchcock, Alfred
Hitch_Sabotage	Sabotage	1937	Hitchcock, Alfred
Hitch_Youngand	Young and Innocent	1938	Hitchcock, Alfred
Hitch_Jamaica	Jamaica Inn	1939	Hitchcock, Alfred
Hitch_LadyVanish	Lady Vanishes, The	1939	Hitchcock, Alfred
Hitch_ForeignCorr	Foreign Correspondent	1940	Hitchcock, Alfred
Hitch_Rebecca	Rebecca	1940	Hitchcock, Alfred
Hitch_Saboteur	Saboteur	1942	Hitchcock, Alfred
Hitch_Suspicion	Suspicion	1942	Hitchcock, Alfred
Hitch_Shadowofa	Shadow of a Doubt	1943	Hitchcock, Alfred
Hitch_Lifeboat	Lifeboat	1944	Hitchcock, Alfred
Hitch_Notorious	Notorious	1947	Hitchcock, Alfred
Hitch_ParadineCase	Paradine Case, The	1947	Hitchcock, Alfred
Hitch_StageFright	Stage Fright	1950	Hitchcock, Alfred
Hitch_Strangerson	Strangers on a Train	1951	Hitchcock, Alfred
Hitch_IConfess	I Confess	1952	Hitchcock, Alfred
Hitch_DialM	Dial M for Murder	1954	Hitchcock, Alfred
Hitch_RearWind	Rear Window	1954	Hitchcock, Alfred
Hitch_ManWho55	Man Who Knew Too Much, The	1955	Hitchcock, Alfred
Hitch_ToCatch	To Catch a Thief	1955	Hitchcock, Alfred
Hitch_TroubleWith	Trouble With Harry, The	1956	Hitchcock, Alfred
Hitch_WrongMan	Wrong Man, The	1957	Hitchcock, Alfred
Hitch_Vertigo	Vertigo	1958	Hitchcock, Alfred
Hitch_Northby	North by North-West	1959	Hitchcock, Alfred
Lang_TestamentM	Testament des Dr. Mabuse, Das	1933	Lang, Fritz
Lang_Liliom	Liliom	1934	Lang, Fritz
Lang_Fury	Fury	1936	Lang, Fritz
Lang_YouOnly	You Only Live Once	1937	Lang, Fritz
Lang_YouandMe	You and Me	1938	Lang, Fritz
Lang_HangmenAlso	Hangmen Also Die	1943	Lang, Fritz
Lang_MinistryofF	Ministry of Fear, The	1945	Lang, Fritz
Lang_WomanIn	Woman in the Window, The	1945	Lang, Fritz
Lang_Cloakand	Cloak and Dagger	1946	Lang, Fritz
Lang_SecretBeyond	Secret Beyond the Door, The	1948	Lang, Fritz
Lang_HumanDesire	Human Desire	1954	Lang, Fritz
Lang_Whilethe	While the City Sleeps	1956	Lang, Fritz
Lang_Beyonda	Beyond a Reasonable Doubt	1956	Lang, Fritz
Lang_IndischeGrab	Indische Grabmal, Das	1958	Lang, Fritz
Lang_TigervonE	Tiger von Eschnapur, Der	1958	Lang, Fritz
Lang_TausendAugen	Tausend Augen des Dr.Mabuse, Die	1960	Lang, Fritz

Opul_Liebelei	Liebelei	1932	Ophuls, Max
Opul_VerkaufteB	Verkaufte Braut, Die	1932	Ophuls, Max
Opul_Signoradi	Signora di tutti, La	1934	Ophuls, Max
Opul_Komedieom	Komedie om Geld	1935	Ophuls, Max
Opul_TendreEn	Tendre ennemie, La	1936	Ophuls, Max
Opul_Werther	Werther	1938	Ophuls, Max
Opul_Exile	Exile, The	1948	Ophuls, Max
Opul_Letterfrom	Letter from an Unknown Woman	1948	Ophuls, Max
Opul_Caught	Caught	1949	Ophuls, Max
Opul_RecklessMom	Reckless Moment, The	1949	Ophuls, Max
Opul_Ronde	Ronde, La	1950	Ophuls, Max
Opul_Plaisir	Plaisir, Le	1952	Ophuls, Max
Opul_MadameDe	Madame de ...	1953	Ophuls, Max
Opul_LolaMontes	Lola Montès	1955	Ophuls, Max
Stern_Dishonored	Dishonored	1931	Sternberg, Josef
Stern_BlondeVenus	Blonde Venus	1932	Sternberg, Josef
Stern_ShanghaiExp	Shanghai Express	1932	Sternberg, Josef
Stern_ScarletEmp	Scarlet Empress, The	1934	Sternberg, Josef
Stern_DevilisaW	Devil is a Woman, The	1935	Sternberg, Josef
Stern_ShanghaiG	Shanghai Gesture, The	1942	Sternberg, Josef
Stern_JetPilot	Jet Pilot	1957	Sternberg, Josef

The Principal Components analysis of this *auteur* corpus of films looks like this (**Fig. 11**):



The dendrogram of this data produces some slight differences in the clustering, but no real extra sharpening of the classification *by auteur*.

Finally, it is worth having a look at the separation between the Lang films and the Hitchcock films when taken alone.

The principal components analysis of this sub-group of films produces a graph like **(Fig. 13)**:

graphically indicated in **figures 4 and 5** above, where the statistical tests are required to separate particular directors' work from a general mass of other directors' work. There is also a question how far a directors' style can change over their career, and the separate analysis of the films of Max Ophuls (**Fig. 8**) shows that this can happen. All of the directors selected for analysis work almost exclusively in a single genre, so my results have nothing to say about the possible influence of genre on the form of films. However, this could be investigated in the same way with data from a suitable collection of films.

References

- Buckland 2006** Buckland, Warren 2006 *Directed by Steven Spielberg: Poetics of the Contemporary Hollywood Blockbuster* (2006) Continuum
- Cutting 2021** Cutting, James 2021 *Movies on Our Minds: The Evolution of Cinematic Engagement* (2021).Oxford University Press
- O'Brien 2004** O'Brien Charles 2004 *Cinema's Conversion to Sound: Technology and Film Style in France and the U.S.* 2004 Indiana University Press.
- Salt 1974a** Salt, Barry 1974a Let a hundred flowers bloom : film form, style, and aesthetics *Sight & Sound* (v 43 n 2 p 108 1974).
- Salt 1974b** Salt, Barry 1974b Statistical Style Analysis of Motion Pictures, *Film Quarterly* v 28 n 1 Fall 1974
- Salt 2009** Salt, Barry 2009 *Film Style and Technology: History and Analysis – 3rd. Ed.* Starword
- Salt 2020** Salt, Barry 2020 *The style of Ingmar Bergman's films 127-150* [New Review of Film and Television Studies](#) Volume 18, 2020 - [Issue 2](#)
- Salt 2021** Salt, Barry *The end of the Great Speed-Up—and after* *Digital Scholarship in the Humanities*, Volume 37, Issue 1, April 2022, Pages 218–228, <https://doi.org/10.1093/lc/fqab049>

Software

- R Core Team** (2021). R: A language and environment for statistical computing.
- R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering.

Bioinformatics. DOI: 10.1093/bioinformatics/btv428

Alboukadel Kassambara and Fabian Mundt (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7.

<https://CRAN.R-project.org/package=factoextra>

Lê, Sebastien, Josse Julie, Husson, François and Mazet (2008). “FactoMineR: A Package for Multivariate Analysis.” *Journal of Statistical Software*, **25**(1), 1–18. [doi:10.18637/jss.v025.i01](https://doi.org/10.18637/jss.v025.i01)