The Recent History of Statistics: Comparing Temporal Patterns of Word Clusters

Matilde Trevisani and Arjuna Tuzzi

Contents

Abstract The abstracts published by the *Journal of the American Statistical Association* in the time span 1946–2016 have been examined in order to identify relevant timings in the recent history of statistics and retrieve past and current topics that have drawn the attention of one of the most influential communities of statisticians in the world. The focus is on clusters of words that, over time, share a similar trajectory of occurrences in the issues of the journal and on the effect of different choices in the number of clusters. When arrangements in coarser and finer groupings have been compared and contrasted, an interesting nested structure has emerged. Moreover, results have highlighted the conjoint effect of word cycle synchrony and word popularity, which are two of the most important features to be accounted for by the researcher in reading the output of a curve clustering based on observations of word frequencies from a chronological perspective. The research

M. Trevisani (✉)
University of Trieste, Trieste, Italy
e-mail: matilde.trevisani@deams.units.it

A. Tuzzi
Department of Philosophy, Sociology, Education and Applied Psychology,
University of Padova, Padova, Italy

also shows that a knowledge-based system (a computer-based system that supports human learning, endowed with a knowledge-base, a statistical learning engine and a user interface) is able to achieve an effective representation of abstracts and that many elements of the history of statistics may be gleaned by reading the abstracts of a large number of papers and considering 'texts as data'.

Keywords History of statistics · Functional data analysis · Normalisation · Splines · Curve clustering · Cluster number validation

6.1 Introduction

Statistics is a young discipline that initially developed as an instrument to provide governments and public administrators with a reliable picture of the population and its needs. As a response to the administrative and accounting needs of the modern state, economic, political and health statistics have also developed in parallel with demographic and social statistics.

In a previous research study (Trevisani and Tuzzi 2015), we attempted to trace a history of the discipline which, rather than starting from handbooks of the history of statistics,[1] adopted a peculiar perspective. Starting merely from the scientific debate on a mainstream statistical journal, we endeavoured to identify past and current topics covered by statistics, from the perspectives of both methods and application fields. We examined a large set of keywords found in the titles of papers published in the period 1888–2012 by the *Journal of the American Statistical Association* (JASA) and its predecessors.

When we analysed the titles of articles as textual data and looked at the history of the discipline from this viewpoint, we found moments, events and timings that articulate the history of statistics into time spans that are similar to those adopted by Historical Sciences to periodise the history of Europe. Once we decided that the appearance on stage of modern scientific journals represented for statistics the same revolution as the advent of writing in the history of humankind (i.e. it marked the exit from *Prehistory*), we periodised the history of statistics in four phases and considered that:

1. The *Ancient History* of statistics begins at the end of the nineteenth century with the publication of the official journal of the *American Statistical Association* (ASA) in 1888: *Publications of the American Statistical Association* (PASA, 1888–1912). During this period, statistics has a scientific language that is not fully codified and standardised, and its main research topics are wide and diversified. Demography and social statistics are the most explored fields: statisticians collect and examine data with the aim of responding/solving the big

[1] The literature is rich in relevant handbooks used to study the history of Statistics: David and Edwards (2001), Hald (1986, 1998, 2007), Stigler (1986, 1999), Walker (1931), and Westergaard (1932), to cite only a few.

problems of humanity (*wealth*, *poverty*, *health*, *safety*, *cause of death*, *infant mortality*) and to study the living conditions and lifestyle of the population. At the end of this first era, PASA gives way to *Quarterly Publications of the American Statistical Association* (QASA, 1912–1921), and in 1922 the *Journal of the American Statistical Association* (JASA) is born.

2. In the 1920s statistics experiences its *Middle Age*, which spans the period of the two World Wars and ends at the end of the 1950s. In this period the economy, the Great Depression and post-war reconstruction dominate the scholars' research interests, and statistics appear to be a true 'political arithmetic'. At the end of this period, rudimentary mathematical instruments appear and give life to a new *Humanism* that is ready to prepare the way for a *Renaissance* of the discipline.
3. In the early 1960s, the *Modern History* begins. Statistics establishes itself as an autonomous discipline thanks to the dissemination of modern statistical tools and develops its own lexicon and specific method. In this period, statistics deals with theories, concepts and methods that are generally considered the essentials of the discipline (*distribution*, *probability*, *regression*, *sampling*, *test*) and nowadays no longer represent research objects anymore as they are well-established research tools. The development of new methods and techniques runs until the late 1980s, when technological revolutions and the diffusion of modern computers lead statistics to a new era.
4. The last part of the time span represents the *Contemporary History* of statistics, which has conducted statistics since the 1990s until today. In the early 2000s, new problems emerged: high-dimensionality, information complexity and the need for new computational-intensive algorithms and high-performing computers (*dimension reduction, smoothing, functional data analysis, neural networks, mixture model, hierarchical models*), but also interdisciplinarity and hybridisation with other disciplines and new fields of application, especially in the environmental, epidemiological, medical and biomedical fields.

The last two periods of this history, namely *Modern History* and *Contemporary History*, are the most relevant ones in order to study in depth the development of statistical methods 'as we understand them today', and for this reason, we decided to tackle a new study based on the analysis of abstracts published in the last decades. An effective and schematic representation of this storytelling is reported in the Appendix of this chapter.

The abstracts of the articles represent an interesting research object because they express the main contents in few words. Nevertheless, as they do not merely include keywords like titles, they envisage some challenges from the methodological and computational viewpoint.

We have decided to continue to examine the JASA because it still portrays the scientific debate of a large and prestigious community of statisticians in the world today, and it still supports the development of statistics in a broad sense (meetings, publications, education, accreditation, advocacy). Moreover, the JASA still represents one of the world's most relevant premier journals of statistical sciences, and it provides a genuine generalist's perspective and special attention to innovation.

With the appearance of other statistical journal and with the differentiation of statistical fields (economics, demography, epidemiology, etc.), JASA has become less generalistic than before. Nevertheless, it is still considered less specialised than other journals of the field.

6.2 Corpus Description and Trends

Abstracts were not available in the JASA until the 1930s (the only one from the 1930s appeared in 1933) are sporadic in the 1940s and 1950s and gradually become more regular and systematic after the 1960s (Table 6.1), that is, during years that, from this point of view, also mark the beginning of a standardisation of scientific writing.

Table 6.1 The JASA: volumes, issues and available abstracts

Name of the journal	Years	Volumes	Issues	Abstracts	*N*
PASA	1888–1990	1–2	10	0	0
PASA	1891–1895	2–4	16	0	0
PASA	1896–1900	5–7	18	0	0
PASA	1901–1905	7–9	18	0	0
PASA	1906–1910	10–12	20	0	0
PASA	1911–1915	12–14	20	0	0
PASA	1916–1919	15–16	16	0	0
QASA	1920–1921	17	8	0	0
JASA	1922–1925	18–20	16	0	0
JASA	1926–1930	21–25	22	0	0
JASA	1931–1935	26–30	24	1	91
JASA	1936–1940	31–35	21	0	0
JASA	1941–1945	36–40	20	0	0
JASA	1946–1950	41–45	20	143	16,694
JASA	1951–1955	46–50	20	63	7811
JASA	1956–1960	51–55	20	188	24,399
JASA	1961–1965	56–60	20	365	41,769
JASA	1966–1970	61–65	20	556	65,105
JASA	1971–1975	66–70	21	775	62,917
JASA	1976–1980	71–75	20	658	56,277
JASA	1981–1985	76–80	20	570	56,082
JASA	1986–1990	81–85	20	647	134,556
JASA	1991–1995	86–90	20	711	118,861
JASA	1996–2000	91–95	20	636	99,686
JASA	2001–2005	96–100	20	514	92,724
JASA	2006–2010	101–105	20	650	115,351
JASA	2011–2016	106–111	24	745	137,019

Size in word-tokens (*N*)

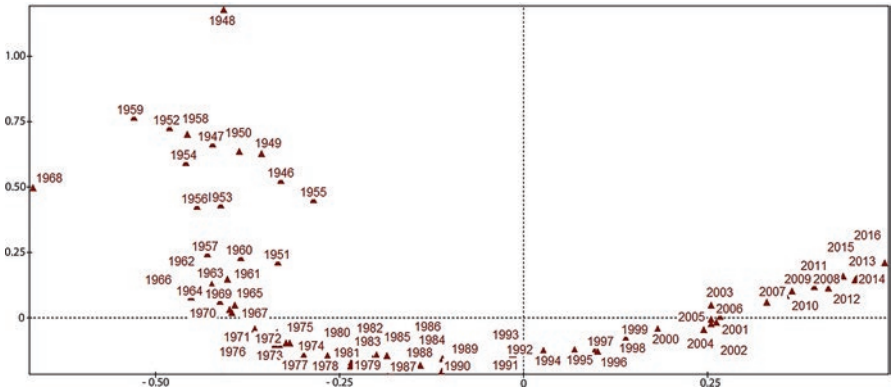

Fig. 6.1 First factorial plane of CA based on 4915 words × 71 years. Projection of years

The abstracts of the papers allow us to observe a more recent period (1946–2016, 71 years) in depth. It represents a little more than half of the time span studied by our previous work that considered only the titles (1888–2012). In the time span of 1946–2016, we retrieved 7221 abstracts. They constitute a large corpus that includes over a million word-tokens (N = 1,029,251) and nearly 27,000 word-types (V = 26,686). The mean length of an abstract (in terms of word-tokens) is 138 words (min: 10; max: 958; st. dev.: 77 words). For a first explorative data analysis (EDA), we selected words with frequency higher than or equal to 10 and worked with a contingency table of words × years (4915 × 71).

Through the correspondence analysis (CA) of the corpus of abstracts (Fig. 6.1), a new temporal trace unfolds and confirms the existence of a clear temporal pattern in the data. The second axis splits the Cartesian plane into two main half-planes (right and left) around the 1990s, and this division has been already identified in the analysis of titles as a moment of transition from *Modern History* to *Contemporary History*. This analysis, based on abstracts, confirms the previous analysis made on the titles.

In the time span 1888–2012, titles showed a more pronounced reduction in variability over time (Trevisani and Tuzzi 2015), but in more recent times, we can also observe that the scientific language is becoming more and more specialised. The scatter of years on the left side of the graph shows that the range of topics is broader and the lexicon seems richer before the 1960s and 1970s, while the research areas become more limited and the lexicon becomes more technical in recent times, especially in the early years of the twentieth century. Variety is reduced in the name of standardisation, which is also the effect of a process of learning and sharing a 'special language'.

6.3 From Content Mapping to Curve Clustering

In quantitative linguistics, the problem of exploring the temporal evolution of a linguistic phenomenon has often been studied by resorting to linguistic laws (Köhler 2011; Popescu 2009; Tuzzi and Köhler 2015) and time series analysis (Pawlowski

et al. 2010). Nevertheless, when trajectories do not portray a regular pattern, these approaches are not able to find a satisfactory solution in terms of goodness of fit and thus achieve results that provide an effective reading of the temporal evolution of word occurrences.

The existence of a latent temporal pattern in word occurrences can be explored by means of CA, which, in our study, reveals a clear time dimension in abstracts and shows that much of the history of statistics may be gleaned by simply reading the abstracts of papers through an EDA. The CA based on the lexical contingency table (words × time-points) is a well-known and established statistical tool in the literature on textual data analysis. When CA is exploited from an exploratory perspective in order to position years and words on a Cartesian plane, it is useful to represent meaningful relationships among words, among time-points and between words and time-points. CA reveals most preeminent timings although it does not highlight how single concepts evolved over time and which words shared the same temporal evolution.

To reconstruct the micro-history of each word and identify words that portray similar temporal patterns, we resort to a functional data analysis (FDA) approach and, within this, to curve clustering (Trevisani and Tuzzi 2015, 2018). A knowledge-based system (KBS) is then proposed to first reconstruct words' life cycles and second, by clustering words with similar life cycles, detect any exemplary temporal patterns representing the latent dynamics of word micro-histories. The major dynamics thus uncovered are then submitted to subject matter experts for interpretation and guidance in the learning process, potentially enabling this to culminate in a conclusive reading (or readings) of the history of the discipline (see Chap. 9 for an extensive description).

In particular, the statistical learning stage of the KBS consists of four steps:

1. Normalising time trajectories of word (raw) frequencies, the type of transformation being chosen according to aspects of life cycles that are considered substantive when comparing words.
2. Filtering time trajectories of word (normalised) frequencies, interpreted as functional data (FD) and thus represented as smooth functions.
3. Curve clustering (CC) to detect all important dynamics underlying the evolution of groups of word micro-histories.
4. Interpretation by expert opinion to decipher detected dynamics and thus compose a narrative of the evolution of the discipline as a whole.

In this study, we have chosen a double normalisation, in particular, d_2 (see Table 9.1, Chap. 9), in order to adjust the uneven document dimension across time (number of texts and their size in word-tokens may vary greatly over time; see Fig. 9.1, Chap. 9) as well as to remedy the great disparity in word popularity (total frequency of individual words in the entire corpus is greatly variable; see Fig. 9.2, Chap. 9), thus enabling the comparison of word curves by timing or synchrony rather than by amplitude. We adopt a basis function approach to filtering with a B-spline basis system (Ramsay and Silverman 2005). Moreover, we take a distance-based approach to CC and use a *k*-means algorithm for FD combined with an appropriate metric for

measuring distance between curves (Jacques and Preda 2014; Wang et al. 2016). In this study, we use the Euclidean distance. Lastly, while interpreting, experts can formulate new research questions that may lead to further insights. If CC yields concurrent solutions, the experts can decide on one or more historical narratives for the knowledge field in the period examined.

6.4 Illustration of Cluster Generation and Reading

To achieve a good representation of the trajectories of relevant keywords, we adopted a procedure for the pre-processing of the corpus that envisages an automatic recognition of multiword expressions (see Chap. 8) and then the intersection of the word list with available glossaries for statistical sciences (Trevisani and Tuzzi 2015). Moreover, to reduce the number of items that refer to the same keyword and overcome some of the limitations of an analysis based on word-types, we replaced words with stems. We considered only stems (*estimation*, *model*, *statistics*) and stem segments (e.g. *likelihood estimation*, *mean square error*, *gene expression*) that occur in the corpus of abstracts at least 50 times. The contingency table includes 1351 rows (keywords) and 71 columns (years/volumes). Table 6.2 provides an excerpt of the matrix that reports the occurrences of each keyword in the corpus as a whole and in each time-point. The temporal evolution of a keyword is drawn from the sequence of its occurrences over time, that is, each row of this table portrays a trajectory that from an FDA perspective represents a realisation of an underlying continuous function.

The KBS applied to the corpus produces the best partitions corresponding to the candidates to cluster number that emerged from the pooled validation approach proposed therein (see Fig. 9.7, Chap. 9). In particular, the set of cluster numbers 4, 5, 7 and 15 are subjected to scrutiny (Fig. 6.2).

From the analysis of the partitions selected for the four cluster numbers, it emerges that the more refined groupings are somehow nested in the coarser ones. We then take the four-cluster partition (see top-left panel of Fig. 6.2 and the individual clusters in Fig. 9.8, Chap. 9) as reference as it shows the four fundamental temporal patterns which are gradually more detailed in the finer partitions, namely:

- Pattern ‘A’: words with decreasing trend from the beginning of the period, generally of low total frequency.
- Pattern ‘B’: words with increasing trend after 1960 or emerging more recently.
- Pattern ‘C’: words exhibiting a period of culminant popularity around 1960–1980/1985 and then slowly decreasing.
- Pattern ‘D’: most popular words everlasting or even with a slowly increasing trend.

In the following paragraphs, each basic cluster will be analysed: first, by examining the nested structure across the four increasingly refined groupings; second, by showing the relevant groups of the finest partition constituting the basic cluster; and

Table 6.2 Excerpt of the contingency table (stemmed) keywords × years (1351 × 71)

Keywords (stemmed)	Occurrences (corpus)	1946	1947	1948	…	1991	1992	…	2014	2015	2016
estim	7524	5	23	9	:	193	153	:	140	158	137
model	7215	1	1	0	:	129	170	:	202	233	197
data	5107	13	11	10	:	118	139	:	128	156	140
method	4776	19	14	11	:	93	102	:	126	162	134
test	4770	7	16	1	:	82	120	:	44	85	45
distribut	3764	4	9	3	:	59	70	:	34	41	58
propos	3434	0	0	1	:	51	72	:	141	181	153
sampl	3339	19	25	7	:	54	75	:	46	47	41
articl	3109	2	1	8	:	61	79	:	130	151	171
base	2954	8	4	1	:	45	81	:	83	90	77
gener	2692	12	5	3	:	61	82	:	54	41	59
statist	2556	6	9	22	:	48	44	:	45	50	37
studi	2528	3	8	2	:	39	58	:	110	95	82
result	2473	6	14	6	:	35	52	:	43	36	53
procedur	2243	8	5	0	:	40	56	:	45	44	38
function	2169	4	0	1	:	54	35	:	65	48	67
time	2162	7	5	7	:	43	54	:	50	54	51
analysi	2145	6	4	4	:	54	64	:	53	79	61
effect	2089	4	5	5	:	39	40	:	70	65	63
problem	2075	5	8	10	:	47	58	:	29	54	37
regress	2073	0	0	0	:	46	59	:	48	65	41
likelihood estim	602	0	0	0	:	18	11	:	11	9	6
exist	595	1	5	2	:	12	15	:	23	32	18
robust	590	0	0	0	:	19	11	:	8	23	19
limit	590	2	4	4	:	17	11	:	7	18	12
work	588	3	3	2	:	10	16	:	21	16	20
bia	587	4	6	0	:	4	12	:	8	11	5
mean squar error	285	1	0	0	:	8	7	:	1	1	0
statement	50	0	1	0	:	0	1	:	2	0	0
gene express	50	0	0	0	:	0	0	:	1	6	5

lastly, by illustrating the typical temporal pattern of these groups through a selection of group words. In establishing the relevance and order of the groups and words presented, we will consider the degree of stability as measured by the multiple Rand index per word and the derived index, on average, per group (see Sect. 9.3.3, Chap. 9). A final summary of the reconstruction of the history of statistics in the period examined is presented in the Appendix of this chapter.

A general overview of the groups in the finest partition, which have been ordered according to the chronological sequence of the four basic patterns ('A', 'C', 'D', 'B', that is, from the cluster of words that have tended to disappear to the cluster of

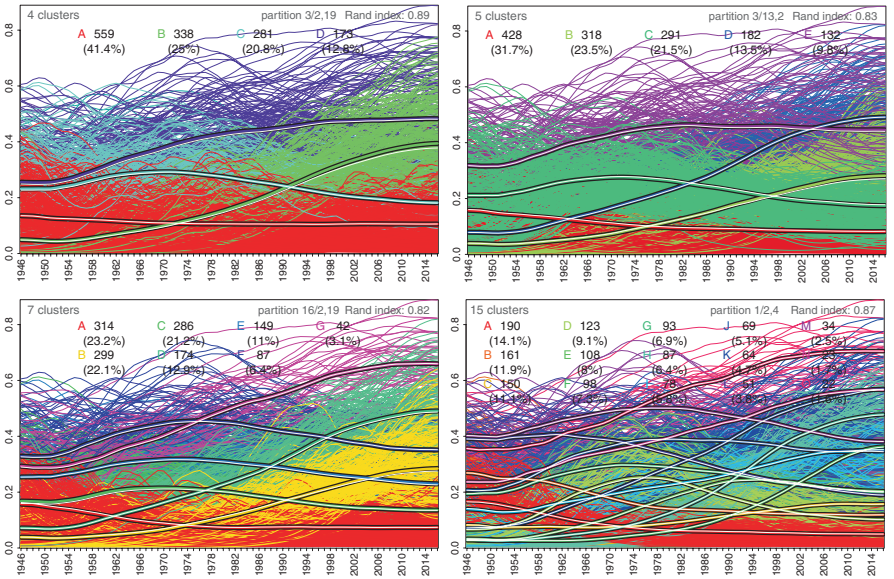

Fig. 6.2 Best partitions corresponding to the set of cluster number candidates in the clustering on d_2 normalised data: the overall 4, 5, 7 and 15 groups

emerging words in the period 1946–2016), is illustrated in Fig. 6.3. For each group, the basic pattern is identified by considering both the direct nesting of the finest partition into the coarsest one and the indirect nesting as derived from the full nesting structure across the four cluster numbers. Whenever a cluster of a finer partition is not entirely nested in a cluster of a coarser partition, the nesting cluster is identified as the one that contains the largest portion of the considered cluster (with a difference from the second largest portion not smaller than 10%, otherwise two nesting clusters are identified). An intermediate pattern may arise wherever the two types of nesting (direct and indirect) give a different basic pattern, as well as wherever two nesting clusters are identified, thus indicating a phase shift (e.g. 'A/C', 'A/B' and 'D/B' in Fig. 6.3).

6.4.1 Pattern 'A': Cluster of Words with Decreasing Trend

Cluster 'A' of the four-group partition contains words that are less and less frequently present, some that eventually disappear, or that were popular quite discontinuously over the period. It produces a 'matrioska'-like structure across the four groupings (Fig. 6.4). The cluster core, which corresponds to A of the 15-group partition (Fig. 6.5), consists mostly of low-frequency words (see panel A in Fig. 6.3) having in general a peak for a short period (right at the beginning, 1946–1950, or during the decade 1950–1960) and a slow decline afterwards, generally

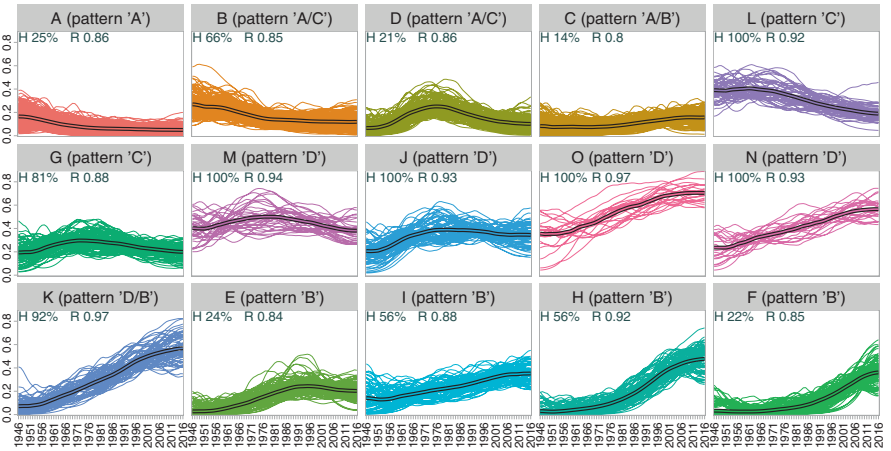

Fig. 6.3 The finest partition into 15 groups ordered according to the chronological sequence of the four basic patterns emerged from the four-cluster partition ('A', 'C', 'D', 'B' and some intermediates). The percentage of high-frequency words (H) and the multiple Rand index (R) are also indicated per group

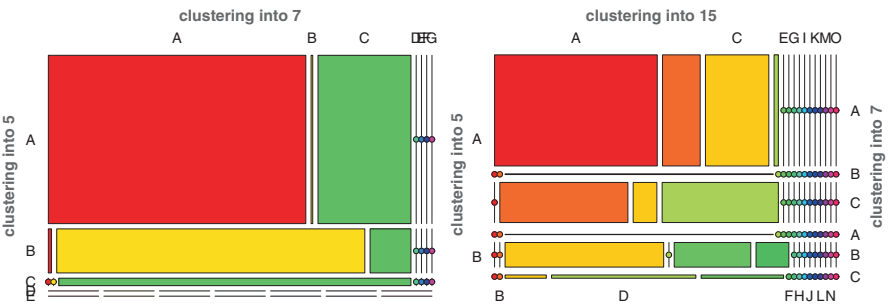

Fig. 6.4 The nested structure of cluster 'A' of the four-group partition: the core A of overall partitions (left and right panels); the rest split in B of both five- and seven-group partitions and C of seven-group partition (left); B, C, D, beyond the core A, of the 15-group partition make up cluster 'A' (right)

disappearing after about 1975 (Fig. 6.6, label A/A). Constituent words refer to the different declinations of statistics before it became an established discipline (demography: *death*, *mortality*, *fertility*, *insurance*, *demography*; social statistics: *social*, *life*, *household*, *interview*, *school*, *labour*, *migration*, *city*, *women*, *familiar*; institutional statistics: *policy*, *bureau*, *institution*, *country*, *administration*; economic statistics: *employment*, *expenditure*, *manufacture*, *firm*, *earning*, *investment*, *agriculture*, *consumption*, *wage*) and to first tools and technical words of statistics (data collection and design of experiments: *interview*, *universe*, *block*, *stratification*, *sampling design*; descriptive statistics: *quality, chart*, *tabular*, *row*, *column*, *cumulative*, *percentage*, *summary*; inferential tools: *method of estimation*, *confidence limit*, *sign test*, *fisher*, *distribution free, point estimation*, *probit*, *logit*, *failure rate*, *number*

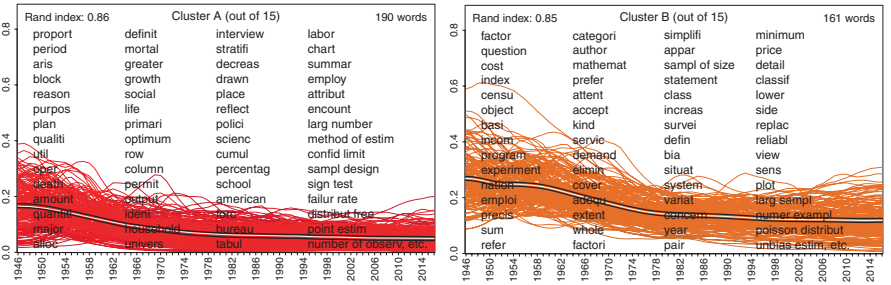

Fig. 6.5 Some clusters of the 15-group partition representing pattern 'A': the core A (left) and the transient (to pattern 'C') B (right)

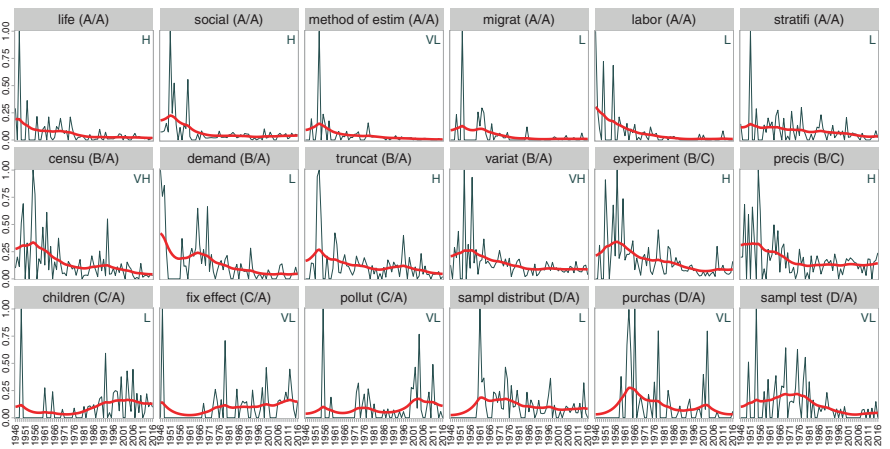

Fig. 6.6 Instances of keywords of cluster 'A' for each nested cluster of the 15-group partition: A, B, C and D. Normalised frequency trajectories and fitted curves. The frequency class is indicated for each keyword: very high (VH), high (H), low (L), very low (VL) (The same reading applies for Figs. 6.9, 6.12, 6.15, and 6.16)

of observations; common words: *purpose*, *reason*, *definition*, *utility*, *circumstances*, *judge*, *efforts*, *wish*). The remaining part of 'A' is constituted by clusters B, C and D of the 15-group partition (Fig. 6.4). However, both B (to a minor degree) and D show a progressive shift toward 'C', the next pattern in chronological order, whereas C even recalls pattern 'B' (Fig. 6.3). Note also a lower multiple Rand index for C that indicates to some extent cluster instability. As a side note, we add that C and D are mostly composed of low-frequency words, thus generating very discontinuous trajectories (Fig. 6.6, labels C/A and D/A). Then, we omit a detailed analysis for C, postpone it for D and dwell here briefly on B. It consists of a mix of popular and less popular words which were more frequent during 1946–1970, but that declined afterwards though they have never vanished (Fig. 6.6, label B/A). We can recognise words that refer to economic statistics (*income*, *demand*, *price*, *index*, *cost*) and

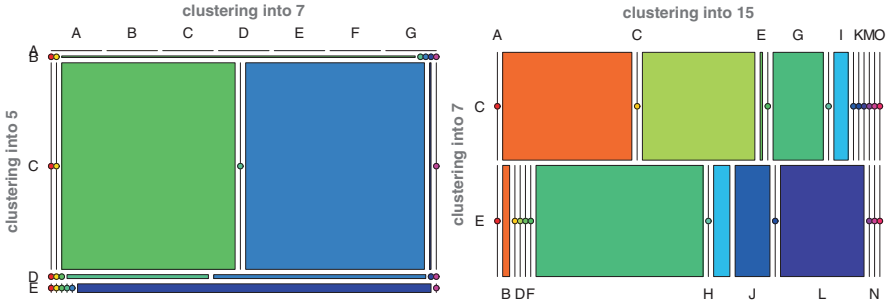

Fig. 6.7 The nested structure of cluster 'C' of the four-group partition: it practically coincides with C of the five-group partition, while it is split into C and E of the seven-group partition (left); couples (B, D) and (G, L) of the 15-group partition make up C and, respectively, E of the seven-group partition (right)

basics of statistics (surveys and sampling: *census*, *survey*, *sampling of size*, *reliable*; design of experiments: *experiment*, *factorial*; principles of estimation: *bias*, *unbiased estimation*, *large sample*, *accuracy*; *categorial*; elements of probability: *binomial*, *poisson distribution*; *truncation*; common words: *statistician*, *numerical example*, *question*, *statement*, *situation*, *concept*, *principle*, *systematic*, *instance*, *agreement*, *support*).

6.4.2 *Pattern 'C': Cluster of Words of the Classic Era of Statistics*

Cluster 'C' of the four-group partition contains words that had a period of peak popularity around 1960–1980 and since have seen a constant, more or less rapid, decline. As regards its nesting structure, it practically coincides with cluster C of the five-group partition, whereas it is split into C and E of the seven-group partition. Lastly, at the innermost level, couples (B, D) and (G, L) of the 15-group partition are the constituting groups of C and, respectively, E of the seven-group partition (Figs. 6.7 and 6.2). B and D of the 15-group partition are connecting groups between patterns 'A' and 'C' (Fig. 6.3). In particular, D is mostly composed of low-frequency words which exhibit a culminant popularity over 1960–1985 (Fig. 6.9) and refer to theory of estimation and hypopthesis testing (*problem of estimation*, *unknown parameter, mean square error*, *minimax, ML*, *asymptotic efficiency*, *sample mean*, *trimmed*, *weighted average*, *significance level, null distribution*, *Tukey*, *Student*, *Wald*, *Stein*, *goodness-of-fit test*; *Monte Carlo study*—as simulation tools), linear regression (*linear function*, *sum of squares, linear estimation, linear unbiased estimation, dependent variable, disturbance*) and other ordinary tools (*contingency*, *matrices*, *order statistics*; probability distributions: *identical/empirical/sampling distribution, multivariate normal*, *skew*, *compound*, *gamma*, *beta*, *Bernoulli*, *normal*

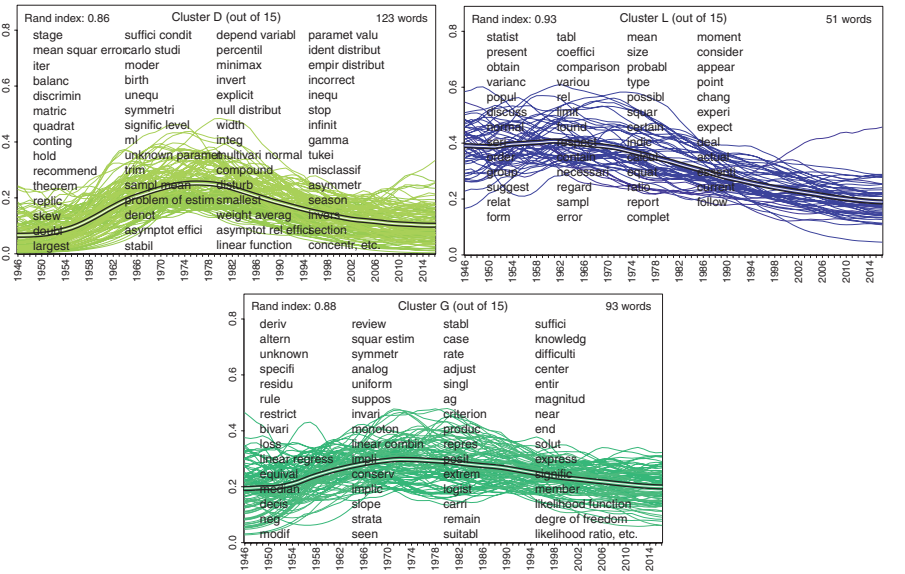

Fig. 6.8 Clusters of 15-group partition representing pattern 'C': D (transient from pattern 'A'), L and G

approximation; time series analysis: *econometrics*, *autocorrelation*, *serial correlation*, *season*, *lag*; common words: *theorem*, *replication*, *hold*, *recommend*) of classic statistics (Fig. 6.8).

A similar temporal pattern is found for L and G, which, however, are almost entirely composed of high-frequency words (Fig. 6.9). These refer to the founder concepts of descriptive statistics and probability (in L, e.g. *statistics*, *population*, *normal*, *series*, *order*, *variance*, *mean*, *sample*, *size*, *error*, *probability*, *moment*; common words: *obtain*, *discuss*, *suggest*, *comparison*, *contain*, *table*) as well as of the bases of inference and linear models (in G, e.g. *linear regression*, *square estimation*, *residual*, *linear combination*, *bivariate*, *univariate median*, *significance*, *sufficiency*, *accuracy*, *efficiency*, *logistic*, *likelihood function*, *likelihood ratio*, *degree of freedom*; common words: *rule*, *restrict*, *loss*, *decision*, *equivalent*, *suppose*, *invariant*, *implication*).

6.4.3 *Pattern 'D': Cluster of the Most Popular and Evergreen Words*

Cluster 'D' of the four-group partition contains only words of very high frequency (panels M, J, O and N in Fig. 6.3), some reaching their peak of popularity over the period 1965–1985, stabilising afterwards to a level only slightly inferior, and others

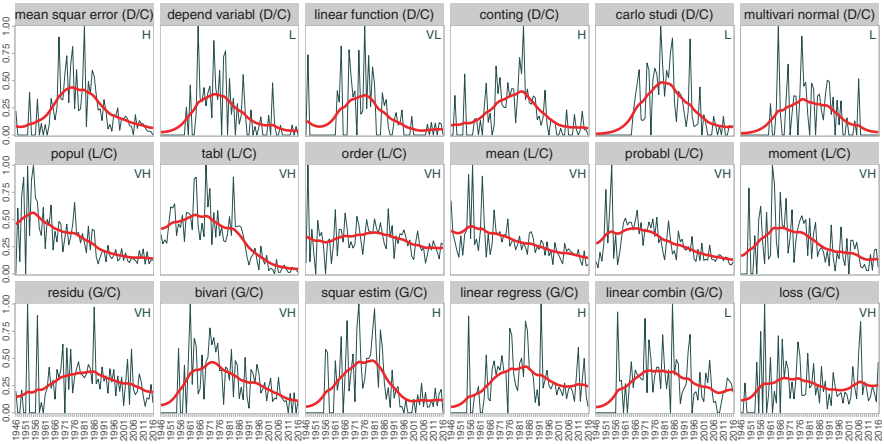

Fig. 6.9 Instances of keywords of cluster 'C' for each nested cluster of the 15-group partition: D, L and G

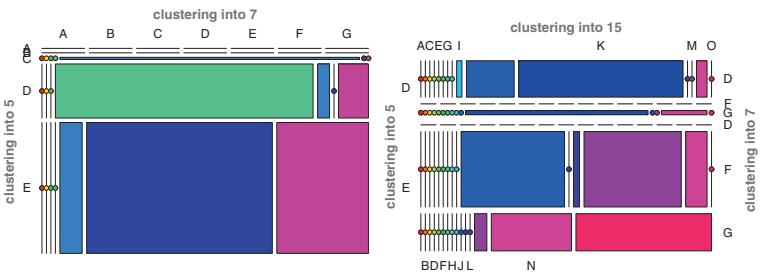

Fig. 6.10 The nested structure of cluster 'D' of the four-group partition: it is mostly represented by E of the five-group partition, which is split into F and G of the seven-group partition (left) [Note that the core D of the seven-group partition (left) is a smaller component relatively to the core D of cluster 'B' (Fig. 6.13)]; couples (J, M) and (N, O) of the 15-group partition make up F and, respectively, G of the seven-group partition (right)

being in constant increase, thus replacing the classic lexicon of statistics in the contemporary age.

As regards its nesting structure, it is mostly represented by cluster E of the five-group partition, which, in turn, is split into F and G of the seven-group partition; lastly, at the innermost level, couples (J, M) and (N, O) of the 15-group partition are the constituting groups of F and, respectively, G of the seven-group partition (Figs. 6.10 and 6.2). Note that cluster K of the 15-group partition is a concatenating group between patterns 'D' and 'B', as will be illustrated in the next paragraph (Figs. 6.13–6.15).

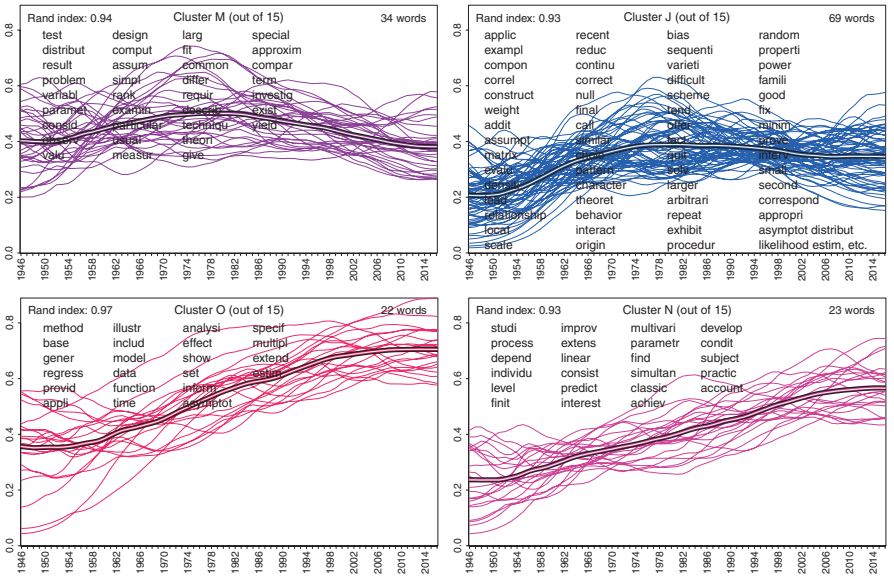

Fig. 6.11 Clusters of 15-group partition representing pattern 'D': M, J, O and N

M contains basic words which were dominant over 1965–1985, stabilising after that time at a lower level or currently decreasing, such as founding terms of traditional lexicon (*test*, *distribution*, *variable*, *parameter*, *design*, *rank*, *observation*, *fit*, *measure*) and common terms (*problem, result*, *assume, compute*, *examine*, *theory, technique, require*, *differ*, *approximation, compare*, *investigation*); J features words which peaked at around 1975/80 but have not lost their vitality over time (*correlation, matrix, weight*, *relationship, similar, density, continuous*, *location, scale, choice, pattern*, *character*, *interaction*, *bias*, *correct*, *additive, repeat, correspondence*, *independence*, *asymptotic distribution, likelihood estimation*; as for the language: *application, example*, *component*, *construct*, *assumption*, *evaluate*, *lead*, *reduce*, *theoretical*, *pattern*, *behaviour, scheme*, *tend*, *offer*, *arbitrary*, *exhibit*, *procedure*, *prove*); O represents fundamental words continuously increasing up to 2000 and then stabilising (*model*, *data*, *regression*, *function*, *asymptotic*, *multiple*, *effect*, *estimation*; as for the language: *method*, *analysis*, *apply*, *illustration*, *provide*, *inform*, *extend*); lastly, N includes words that, although classical, have not lost vital force and are still unattainable when composing a statistical text (*process, dependent, individual*, *level*, *linear, parametric*, *consistent, predict, multivariate, simultaneous*, *subject*; as for the language: *study, improvement, extension*, *interest*, *find*, *achieve*, *develop*, *practice*, *account*) (see Figs. 6.11 and 6.12).

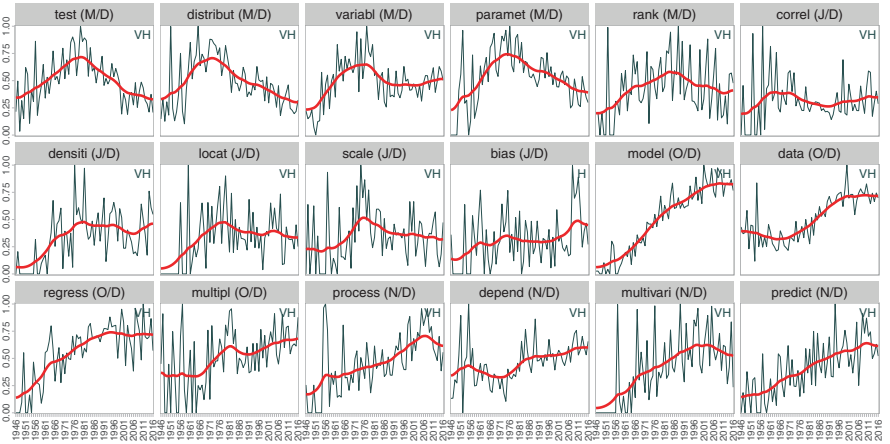

Fig. 6.12 Instances of keywords of cluster 'D' for each nested cluster of the 15-group partition: M, J, O and N

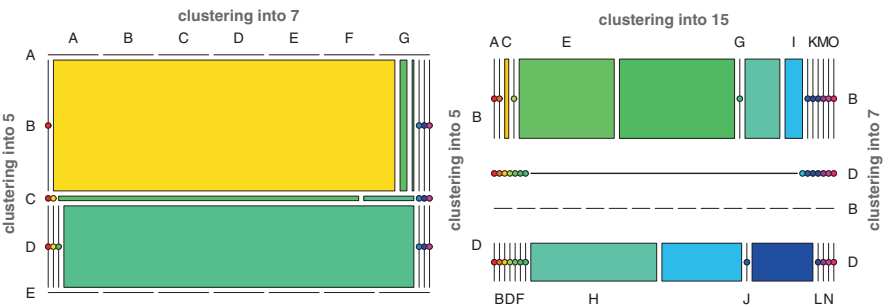

Fig. 6.13 The nested structure of cluster 'B' of the four-group partition: it is split into (B, D) of both the five-group and seven-group partitions (left); (E, F) and (H, I, K) of the 15-group partition make up B and D, respectively, of the seven-group partition (right)

6.4.4 Pattern 'B': Cluster of Words with Increasing Trend and Emerging

Cluster 'B' of the four-group partition contains the contemporary bag of words of statistics: words that have increased their popularity especially after 1990, that have already begun the descending parable after 2000 or that are emerging in the last 10 years. As regards its nesting structure, it is split into B and D of both the five-group and seven-group partitions. Then, B and D, which almost match in these last two groupings, are roughly split into (E, F) and, respectively, (H, I, K) of the 15-group partition (Fig. 6.13).

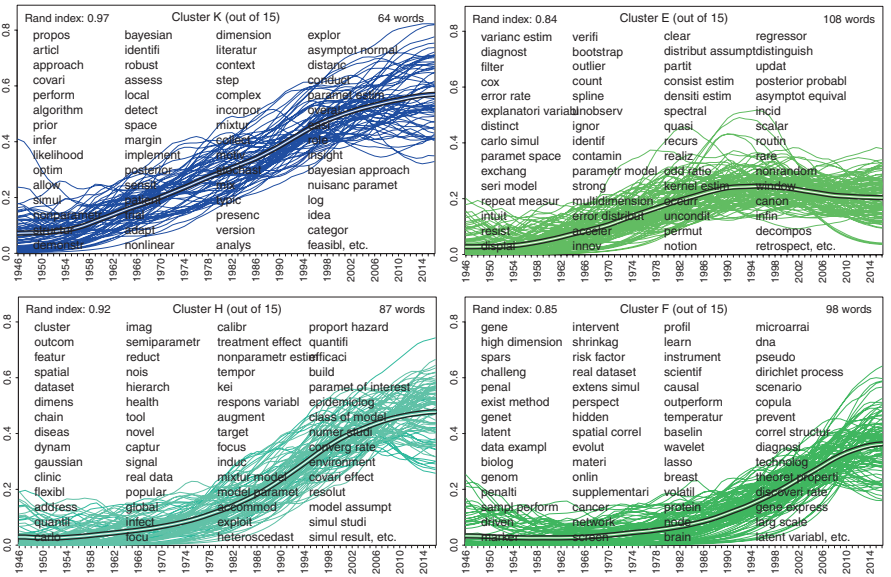

Fig. 6.14 Clusters of 15-group partition representing patterns 'B': K (transient from 'D'), E, H and F

Clusters K and H consist mostly of high frequency words with a continuous markedly increasing trend of popularity (Fig. 6.3). K includes very high-frequency words that suggest the evolution of approaches (e.g. *likelihood*, *bayesian*, *nonparametric*, *robust*, *local*, *adaptive*, *nonlinear*, *sensitivity* analysis, *distance*-, *algorithm*- and *simulation*-based), of modelling problems (e.g. *dimension*, *complex*, *space*, *mixture*, *mixed*, *survival*) and of language (e.g. *propose*, *perform*, *assess*, *detect*, *implement*, *explore*) since 1960, while H features words that gained popularity generally at the end of 1990 and translate those ideas into new models (*flexible*, *hierarchical*, *heteroscedastic*, *dynamic*, *temporal*, *spatial*, *capture*, *proportional hazard* and *mixture model*), methods (Monte *Carlo*, *augmentation*, *signal* processing, *smoothing*, *nonparametric*, *semiparametric*, *kernel*, *reduction*, *clustering, matching*, *calibration*, *image* and *longitudinal* data, *tree* models) and studies (*health* statistics, *clinic* trials, *infectious*, *disease*, *environmental* statistics, *dose*, *exposure*, *epidemiology*) (Fig. 6.14). Note that K is a concatenation group between patterns 'D' and 'B' being a cluster of words born with the consolidation of statistics that have acquired a high level of popularity over time (pattern 'D') but also constituting the specialised terms for dealing with the new themes and applications of the contemporary age (pattern 'B'). We omit a detailed description of cluster I, which consists of words having a more fluctuating popularity although stabilised around 2000, as it features less substantial words with respect to K and H.

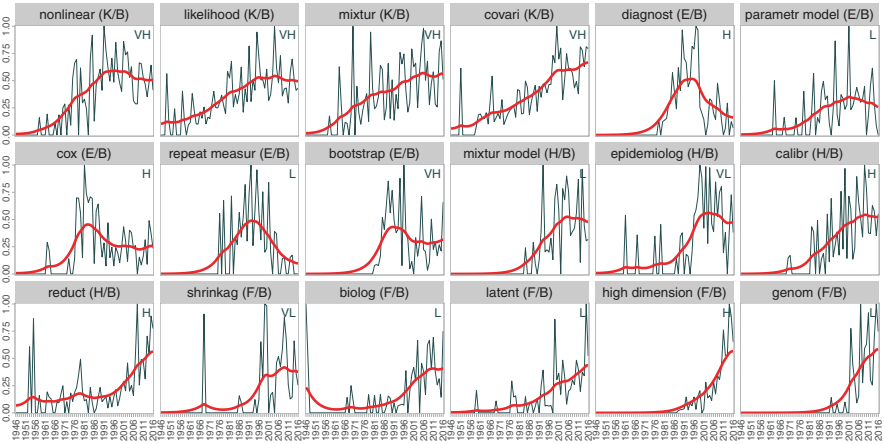

Fig. 6.15 Instances of keywords of cluster 'B' for each nested cluster of the 15-group partition: K, E, H and F

Clusters E and F consist mostly of low-frequency words with an increasing trend. Those contained in E suggest themes (*bootstrap*, *jacknife*, Monte C*arlo simulation* and *method*, *density* and *kernel estimation, variance estimation*, *consistent estimation*, *filtering*, *splines*, *smoother*, *outlier* detection, *diagnostic* tools, *missing values*, *nonrandom*, *quasi*-likelihood, *contamination* methods, *parametric* models, *error distribution*, *distributional assumption*, *asymptotic equivalence*, *multidimensional* analysis, *count* data, *odds ratio*, *spectral* analysis, time *series model*, *repeated measures*, *exchangeable* model, *cox*, *hazard* function, *retrospective* studies) that gained popularity until the late 1990s, after which they began to decline, while those in F represent issues (*high dimension*, *sparsity*, *latent* and *hidden* process, *heterogeneity*, *correlation structure*, *volatility*, *spatial correlation*, *risk factors*, *mapping*, *microarray*), approaches (nonparametric, e.g. *wavelets*; simulation-based, e.g. *sampler*, *slice* sampling, *extensive simulation*; machine *learning, data-driven*), estimation methods (*shrinkage*, *penalisation, penalty, instrumental variables, regularisation, sparse estimation, lasso*), models (e.g. *latent variable*, *causal*, *semiparametric model*, *copula*, *trait*, *trajectory*, *dirichlet process*, graphical models and *network* analysis), research areas (medicine, e.g. *virus*; public health, e.g. *prevention*, *intervention*; environmetrics, e.g. *environment, temperature*; *biology*; epidemiology, e.g. *prevalence*; biomedicine, e.g. *cancer*, *breast*, *brain, gene*, *dna*, *genetics*, *marker*, *genoma*, *protein*, *gene expression*) and 'common saying' (*challenge*, *goal*, *task*, *profile*, *perspective*, *scenario*, *realistic*) that have a surge after 2000 since they have a rebirth in recent times or are emerging (Figs. 6.14 and 6.15).

6.5 Some Remarks on Normalisation with a Focus on the Cluster of Emerging Words

For a better understanding of clustering results, let us examine some effects of the word frequency normalisation adopted in this study. We recall that a transformation of raw frequencies of words is necessary to correctly reconstruct and compare the temporal evolution of words. A form of normalisation by time-point should be regarded as preliminary in order to adjust the uneven size of subcorpora across time (see Fig. 9.1, Chap. 9). A further form of normalisation by word might be appropriate in order to regulate the great disparity in word popularity, which produces very strong asymmetry of frequency spectrum by time-point and sparseness of low-frequency word trajectories (see Fig. 9.2, Chap. 9).

In this study, we have chosen a double normalisation (d_2, see Table 9.1, Chap. 9) that normalises both by time-point and by word, in particular, from dividing the raw frequency of a word at each time-point/volume, both by the total number of word-tokens in each volume and by the maximum frequency of the word trajectory. In particular, this last normalisation (by word) is able to substantially reduce the high skewness featuring the bundle of word trajectories. However, it cannot completely remedy the problem of sparsity. As a result, a trace of word popularity remains and continues to influence the comparison between trajectories as we describe below.

The criterion chosen in the illustration of temporal patterns (both in subsection sequence of Sect. 6.4 and in Fig. 6.3, namely 'A', 'C', 'D', 'B') is the chronological one: clusters are ordered according to the exemplary life cycle of group words from the one that has already been concluded ('A') to the one that has recently begun ('B'). As well, clusters of the 15-group partition are chronologically ordered within each of the four basic temporal patterns (Fig. 6.3). However, a chronological reading is not sufficient to discriminate the four patterns and the more analytical patterns of the 15-group partition that compose them. A second key to reading is the level of popularity featuring the words of the considered cluster. In fact, note how some patterns have a relatively parallel gait and are mostly distinguished by the height of the curves (see B-L, D-G, C-I, F-H-K in Fig. 6.3). This result is due to the effect that the chosen normalisation has on the filtering of curves and therefore on their grouping on the basis of similarity. Namely, words with a low or very low total frequency tend to have sparse trajectories, i.e. to have zero or almost zero frequency for relatively long stretches of the period, either continuous (in the case that the word concludes or begins its life cycle; see, e.g. *semiparametric model* in Fig. 6.16, third row, left-most panel) or intermittent (giving rise to peak-and-valley trajectories; see, e.g. *realistic*, sixth row, left-most panel), and very high differences of height along the trajectory (being frequency values little spaced). This involves that the smoothing of the trajectory tends to produce a flattened curve downwards. On the contrary, words with a high or very high total frequency tend to have non-negligible frequencies for most of the period and trajectories with lower differences of height (being the grid of frequency values finer) (see, e.g. *nonparametric* and *simulation*, left-most panels of first row).

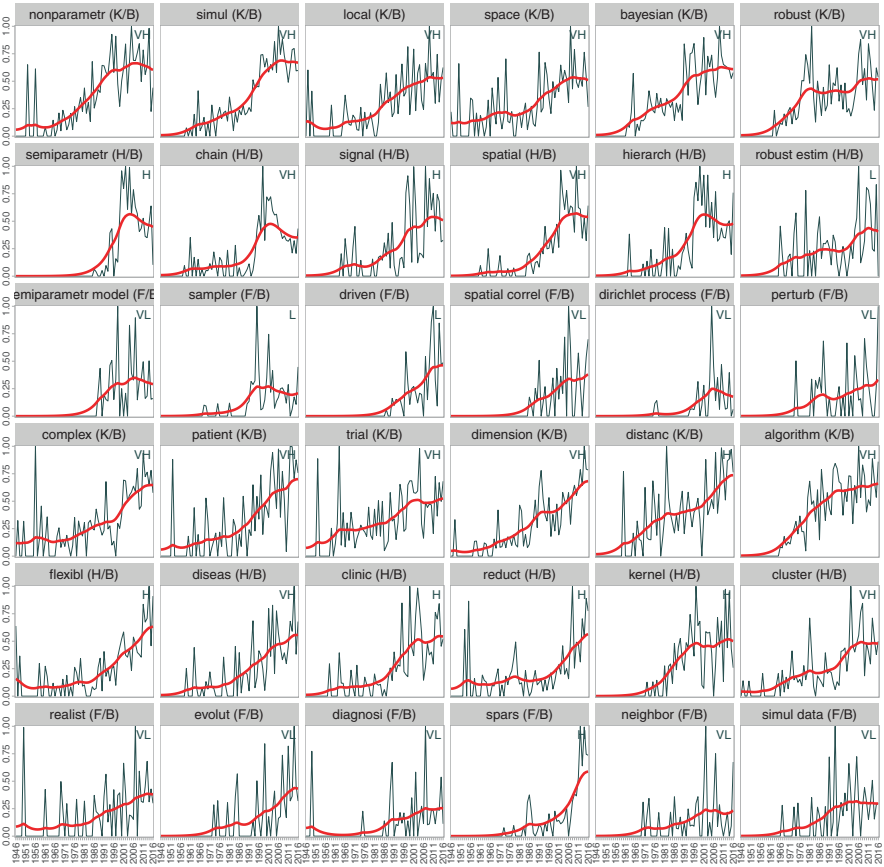

Fig. 6.16 Instances of words of cluster 'B' from clusters K, H and F of the 15-group partition; 12 themes illustrated by 12 (vertical) trios of words taken, in order, from K, H and F. For example, the first top-left trio *nonparametric* (K), *semiparametric* (H), *semiparameric model* (F) refers to theme 'nonparametric methods'

Let us illustrate the point considering pattern 'B', that is, the cluster of words that constitute the contemporary lexicon of statistics and that evoke themes emerging in more recent times. In particular, clusters K, H and F of the 15-group partition (Fig. 6.14) feature a relatively synchronised pattern, which is essentially distinguished by a different height. They comprise, respectively, 8%, 44% and 78% low-frequency words (Fig. 6.3). Figure 6.16 shows a sample of words which refer to 12 themes, each exemplified by a trio of words taken, in order, from clusters K, H and F, to compare the effect of popularity level on curve smoothing. The themes are nonparametric methods, simulation-based estimation, data-denoising and data-driven approach, spatial models, Bayesian models, robust estimation, complex and flexible models, epidemiology, statistical medicine, high-dimension and sparsity, local estimation and machine learning. We can take any of the 12 trios (K, H, F) to

see how the word popularity affects the height of smoothed curve, the shape being similar. For example, compare sparsity and bumpiness of trajectories for *space* (K), *spatial* (H) and *spatial correlation* (F) keywords (the fourth vertical trio in the first row, Fig. 6.15), which are, the first two, very-high (VH) and, the third, very-low (VL) frequency words: the smoothed curves are somehow synchronised although they are different in level.

6.6 Discussion and Conclusion

In a previous study, we observed the existence of a clear temporal pattern in articles of JASA, and we showed that a large share of relevant elements can be retrieved through the statistical analyses of the titles of papers published by JASA (Trevisani and Tuzzi 2015). In this new study, we had the opportunity to elaborate on more recent history by a (distant) reading of the abstracts. Through CA results, a clear temporal pattern emerges (Fig. 6.1), even if, beyond this pattern, it is difficult to understand how individual concepts evolved over time and, above all, which keywords share the same temporal development. In brief, CA proves useful to obtain a general overview of the main contents of the corpus, but curve recognition and clustering are necessary to trace the individual life cycles of keywords and find common dynamics latent to word micro-histories.

The proposed KBS leads to the identification of a number of possible partitions of the corpus in word clusters (Fig. 6.2). Our procedure for cluster number selection, in fact, produces a set of candidates to cluster number: in this study, 4, 5, 7 and 15. From an analysis of the agreement between the concurrent partitions (Wagner and Wagner 2007), it emerges that, for this corpus, the finer partitions are essentially nested in the coarser ones. From this finding, the reading of results is based on the four basic temporal patterns of the coarsest partition, each of which is analysed in depth by the examination of the nested clusters in the 15-group partition.

The reading follows the chronological order, that is, from the cluster of words that have tended to disappear to the cluster of emerging words in the period 1946–2016, both in the analysis of the sequence of the four fundamental patterns and in that of the sequence of the 15 nested sub-patterns (Fig. 6.3).

However, the chronological reading key must be integrated with the information on the popularity level of cluster words. In fact, on the one hand, the adopted double normalisation substantially solves the problem of strong asymmetry of the frequency distribution due to the enormous difference between popular and rare words, but on the other hand, it does not remedy the problem of sparseness in the trajectories of unpopular words. Therefore, a trace of the cluster words' popularity remains in the reconstruction of the temporal pattern. This reflects, on the one hand, the synchrony of curves, and on the other hand, the popularity level of cluster words. An example of this effect of the normalisation is offered by the case of the almost parallel temporal patterns, but of different heights, that feature the clusters of words evoking emerging themes (see Sect. 6.5).

Appendix

Reconstruction of the history of statistics in the period 1946–2016

Middle age	Until about 1960	Pattern	Group	Chronology	Popularity	Themes	Instances of keywords
		'A'	A	Peak during 1946–1950 or 1950–1960, slow decline afterwards, generally disappearing after about 1975	VH/H 3/22% L/**VL** 33/42%	Demography, social/ institutional/economic statistics; data collection and design of experiments; first tools of descriptive and inferential statistics; (common words)	*death, mortality, fertility, insurance, demography; social*, *life*, *household*, *interview*, *school*, *labour*, *migration*, *city*, *women*, *familiar*; *policy*, *bureau*, *institution*, *country*, *administration*; *employment*, *expenditure*, *manufacture*, *firm*, *earning*, *investment*, *agriculture*, *consumption*, *wage*; *interview*, *universe*, *block*, *stratification*, *sampling design*; *quality, chart*, *tabular*, *row*, *column*, *cumulative*, *percentage*, *summary*; *method of estimation*, *confidence limit, sign test*, *fisher*, *distribution free*, *point estimation*; *probit*, *logit*, *failure rate*; (*purpose*, *reason*, *definition*, *utility*, *circumstances*, *judge*, *efforts*, *wish*)
		'A/C'	B	Peak during 1946–1970, decline afterwards though never vanishing	VH/**H** 23/43% L/VL 24/10%	Economic statistics; surveys and sampling, design of experiments, elements of probability, principles of estimation	*income*, *demand*, *price*, *index*, *cost*; *census*, *survey*, *sampling of size*, *reliable*; *experiment*, *factorial*; *categorial*; *binomial*, *poisson distribution*, *truncation*; *precision*, *bias*, *unbiased estimation*, *large sample*, *accuracy*; (*statistician*, *numerical example*, *question*, *statement*, *situation*, *concept*, *principle*, *systematic*, *instance*, *agreement*, *support*)
Modern history	1960–1990		D	Dominant over 1960–85 since then rapid decline	VH/H 1/20% **L**/VL 44/35%	Classic theory of estimation and hypothesis testing, linear regression, probability distributions, descriptive tools, time series analysis	*problem of estimation*, *unknown parameter, mean square error*, *minimax, ML*, *asymptotic efficiency*, *sample mean*, *trimmed*, *weighted average*, *significance level, null distribution*, *Tukey*, *Student*, *Wald*, *Stein*, *goodness-of-fit test, population mean*; *linear function*, *sum of squares, linear estimation, linear unbiased estimation, dependent variable, disturbance*; *Monte Carlo study*; *identical distribution, empirical distribution, sampling distribution, multivariate normal*, *skew*, *compound*, *gamma*, *beta*, *Bernoulli*, *normal approximation*; *contingency*, *matrices, order statistics*; *econometrics*, *autocorrelation*, *serial correlation*, *season*, *lag*; (*theorem*, *replication*, *hold*, *recommend*)

		'C'	L	Dominant up to 1970 since then slow decline	**VH**/H 84/16%	Founding concepts of descriptive statistics and probability	*statistics*, *population*, *normal*, *series*, *order*, *variance*, *mean*, *sample*, *size*, *error*, *probability*, *moment*; (*obtain*, *discuss*, *suggest*, *comparison*, *contain*, *table*)
			G	Dominant over 1960–1980/1990 since then slow decline	VH/**H** 35/45% L/VL 18/1%	Basics of inference and linear models	*linear regression*, *square estimation*, *residual*, *linear combination*, *bivariate*, *univariate median*, *significance*, *sufficiency*, *accuracy*, *efficiency*, *logistic*, *likelihood function*, *likelihood ratio*, *degree of freedom*; (*rule*, *restrict*, *loss*, *decision*, *equivalent*, *suppose*, *invariant*, *implication*)
		'D'	M	Dominant over 1965–1985 since then slow decline or stabilised at a lower level	**VH** 100%	Founding concepts of classic statistics	*test*, *distribution*, *variable*, *parameter*, *design*, *rank*, *observation*, *measure*, *fit;* (*problem, result*, *assume, compute*, *examine*, *theory, technique, require*, *differ*, *approximation, compare*, *investigation*)
			J	Peak at ~1975/80, stabilised afterwards	**VH**/H 68/32%	Toolbox of statistical analysis up to contemporary age	*correlation, matrix, weight*, *relationship, similar, density, continuous*, *location, scale, choice, character*, *interaction*, *bias*, *correct*, *additive, repeat, correspondence*, *independence*, *asymptotic distribution, likelihood estimation;* (*application, example*, *component*, *construct*, *assumption*, *evaluate*, *lead*, *reduce*, *theoretical*, *pattern*, *behaviour, scheme*, *tend*, *offer*, *arbitrary*, *exhibit*, *procedure*, *prove*)
Contemporary history	1990-nowadays		O	Steady increase up to 2000, stabilised afterwards	**VH** 100%	Fundamental words of statistics even more in the contemporary age	*model*, *data*, *regression*, *function*, *asymptotic*, *multiple*, *effect*, *estimation*; (*method*, *analysis*, *apply*, *illustration*, *provide*, *inform*, *extend*)
			N	Steady increase up to 2005/10, stabilised afterwards	**VH**/H 96/4%	Classical although unattainable words in a statistical text	*process, dependent, individual*, *level*, *linear, parametric*, *consistent, predict, multivariate, simultaneous*, *subject;* (*study, improvement, extension*, *interest*, *find*, *achieve*, *develop*, *practice*, *account*)

		'D/B'	K	Marked increase since 1960	**VH**/H 70/22% L 8%	Approaches and modelling problems of contemporary statistics	*inference*, *likelihood, bayesian, prior*, *posterior*, *bayesian approach*, *nonparametric, robust, local*, *adaptive*, *nonlinear, distance*, *algorithm*, *simulation*, *asymptotic normality*, *sensitivity; dimension*, *complex*, *space*, *mixture*, *mixed*, *survival*, *patient*, *trial*; (*propose*, *approach*, *perform*, *assess*, *detect*, *implement*, *analyse*, *explore*)
Contemporary history	1990-nowadays	'B'	H	Marked increase since after 1990	VH/**H** 13/43% L/VL 23/21%	Models and methods of contemporary statistics, medical/health/ environmental statistics, epidemiology	*flexible*, *hierarchical*, *heteroscedastic*, *dynamic*, *temporal*, *spatial*, *capture*, *proportional hazard*, *mixture model*, Monte *Carlo*, *chain, augmentation*, *signal*, *smoothing*, *nonparametric*, *semiparametric*, *kernel*, *reduction*, *clustering, matching*, *calibration*, *image*, *longitudinal* data, *trees*; *health*, *clinic* trials, *infectious*, *disease*, *environmental*, *dose*, *exposure*, *epidemiology*; (*address*, *novel*, *focus*, *target*, *accommodate*, *exploit*, *build*, *enhance*, *highlight*, *suffer*, *monitor*, *real data*, *package*, *software*)
			E	Increase from late 1980 until late 1990, decline afterwards	VH/H 4/18% L/**VL** 35/43%	Research areas and models losing vitality in the new millennium	*bootstrap*, *jacknife*, Monte *Carlo simulation* and *method*, *density* and *kernel estimation, variance estimation*, *consistent estimation*, *filtering, splines, smoother*, *outliers*, *diagnostics*, *missing values, nonrandom*, *quasi*-likelihood, *contamination* methods, *parametric model*, *error distribution*, *distributional assumption, asymptotic equivalence*, *multidimensional* analysis, *count* data, *odds ratio*, *spectral* analysis, *time series model*, *repeated measures*, *exchangeable* model, *cox*, *hazard* function, *retrospective* studies; (*unobserved*, *ignorance*, *occurrence*, *realisation*, *unconditional*, *notion*, *display*, *verify*, *identify*, *distinguish*)
			F	Surge after 2000 as a rebirth or since emerging	VH/H 4/20% L/**VL** 32/44%	Issues, approaches, estimation methods, models, research areas of the new millennium, machine learning, medical/health/ environmental statistics, epidemiology, biostatistics	*high dimension*, *sparsity*, *latent*, *hidden* process, *heterogeneity*, *correlation structure*, *volatility*, *spatial correlation*, *risk factors*, *mapping*, *microarray; wavelets*, *sampler*, *slice*, *extensive simulation*, *learning,* data-*driven*; *shrinkage*, *penalisation, penalty, instrumental variables, regularisation, sparse estimation, lasso*; *latent variable, causal*, *semiparametric model*, *copula, trait, trajectory*, *dirichlet process*, *network*, *segmentation*; *virus*, *prevention*, *intervention*, *environment, temperature*, *biology*, *prevalence*, *cancer*, *breast*, *brain, gene*, *dna*, *genetics*, *marker*, *genoma*, *protein*, *gene expression*; (*challenge*, *goal*, *task*, *profile*, *perspective*, *scenario*, *framework*, *realistic*)

References

David, H. A., & Edwards, A. W. F. (2001). *Annotated readings in the history of statistics*. New York: Springer-Verlag.

Hald, A. (1986). *A history of probability and statistics and their applications before 1750*. New York: Wiley.

Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. New York: Wiley.

Hald, A. (2007). *A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935*. New York: Springer.

Jacques, J., & Preda, C. (2014). Functional data clustering: A survey. *Advances in Data Analysis and Classification, 8*(3), 231–255.

Köhler, R. (2011). Laws of languages. In P. C. Hogan (Ed.), *The Cambridge encyclopedia of the language science* (pp. 424–426). Cambridge: Cambridge University Press.

Pawlowski, A., Krajewski, M., & Eder, M. (2010). Time series modelling in the analysis of homeric verse. *Eos, 97*(2), 79–100.

Popescu, I. I. (2009). *Word frequency studies*. Berlin: Mouton De Gruyter.

Ramsay, J., & Silverman, B. W. (2005). *Functional data analysis (Springer series in statistics)*. New York: Springer.

Stigler, S. M. (1986). *The history of statistics. The measurement of uncertainty before 1900*. Cambridge: The Belknap Press of Harvard University Press.

Stigler, S. M. (1999). *Statistics on the table: The History of statistical concepts and methods*. Cambridge: Harvard University Press.

Trevisani, M., & Tuzzi, A. (2015). A portrait of JASA: The History of Statistics through analysis of keyword counts in an early scientific journal. *Quality and Quantity, 49*(3), 1287–1304.

Trevisani, M., & Tuzzi, A. (2018). Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based Systems, 146*, 129–141.

Tuzzi, A., & Köhler, R. (2015). Tracing the history of words. In A. Tuzzi, M. Benesová, & J. Macutek (Eds.), *Recent contributions to quantitative linguistics* (pp. 203–214). Berlin: DeGruyter.

Wagner, S., & Wagner, D. (2007). Comparing clusterings: An overview. Universitat Karlsruhe, Fakultat fur Informatik Karlsruhe. Retrieved from https://publikationen.bibliothek.kit.edu/1000011477/812079

Walker, H. M. (1931). *Studies in the history of statistical method with special reference to certain educational problems*. Baltimore: The Williams and Wilkins.

Wang, J. L., Chiou, J. M., & Mueller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application, 3*(1), 257–295.

Westergaard, H. (1932). *Contributions to the history of statistics*. London: P. S. King and son.