# Chapter 9
# Functional Data Analysis and Knowledge-Based Systems


**Matilde Trevisani**

## Contents



**Abstract** In the present study, the challenge is whether a distant reading of the history of a discipline can be achieved by analysing the temporal evolution of keywords retrieved from papers in the discipline's mainstream journals. This calls for the so-called knowledge-based system (KBS), i.e. a computer-based system that supports human learning not only by acquiring and manipulating large volumes of data and information, but also by integrating knowledge from different sources. In this chapter, we introduce a KBS that, starting from a large database of texts retrieved from scientific articles published over a lengthy period by a selection of the discipline's premier journals, leads to the construction of a well-founded corpus of scientific literature and from this to a possible outline of the discipline's history. Our work is based on the idea that the temporal course of a word occurrence is a

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



A. Tuzzi (ed.), *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, Quantitative Methods in the Humanities and Social Sciences,
https://doi.org/10.1007/978-3-319-97064-6_9

proxy of the word's life cycle. We then adopt a functional data analysis (FDA) approach under which we first reconstruct words' life cycles. Second, by clustering words with similar life cycles, we detect any prototypical or exemplary temporal patterns representing the latent dynamics of word micro-histories. The major dynamics uncovered at this stage are then submitted to subject matter experts for interpretation and guidance in decision-making, thus making it possible to trace a history of the discipline. Moreover, we propose several kinds of data normalisation which involve different concepts of life cycle similarity and hence a different reading of the history of the discipline under examination.

**Keywords** Distant reading · Trajectory normalisation · Curve clustering · Clustering validation · Clustering agreement

## 9.1 Introduction

Scientometrics studies the evolution of science quantitatively, by analysing publications. One of its main objectives is to develop information systems that can help explore the enormous number of scientific articles that are constantly being published. In the present study, the challenge is whether a "distant" reading of the history of a discipline (or, more generally, a field of knowledge) can be achieved by analysing the temporal evolution of keywords retrieved from papers in the discipline's mainstream journals. This calls for a so-called knowledge-based system (KBS), i.e. a computer-based system that supports human learning not only by acquiring and manipulating large volumes of data and information, but also by integrating knowledge from different sources. In this chapter, we introduce a KBS that, starting from a large database of texts retrieved from scientific articles published over a lengthy period by a selection of premier journals in a specific discipline, leads to the construction of a well-founded corpus of scientific literature (organised as a "keywords × time points" matrix) and from this to a possible outline of the discipline's history. Corpus creation is assisted by knowledge from linguistics experts captured in the system's knowledge base, while knowledge from experts in the specific domain being investigated assists the learning process in both interpretation and decision-making, potentially enabling it to culminate in a conclusive reading (or readings) of the history.

The underlying assumption of the research project presented in this book is that the temporal evolution of words (in terms of their occurrence) reflects the relevance of the corresponding concepts (ideas, themes, research problems) in the scientific discourse over time. In making this assumption, we are aware that the timeframes for methods and research fields yielded by our analysis reflect the moment when they became established in the scientific community and spread to the literature (which is necessarily later than the time these methods are introduced or interest is first expressed in these fields).

Our work is based on the idea that the temporal course of a word occurrence is a proxy of a word's diffusion and vitality, i.e. of the word's life cycle (Trevisani and Tuzzi 2012, 2013a, b, 2015, 2018; Tuzzi and Köhler 2015). We then adopt a functional data analysis (FDA) approach where the observations (occurrences) through time are viewed as a realisation of an underlying continuous function representing the temporal development of a word.

In the FDA approach, we first reconstruct words' life cycles. Second, by clustering words with similar life cycles, we detect any prototypical or exemplary temporal patterns representing the latent dynamics of word micro-histories. Examples of prototypical patterns include essentially increasing, decreasing or constant trends, trends with an isolated peak for briefly faddish words, or roughly bell-shaped trends for words which had a golden age and then disappeared, etc. The major dynamics uncovered at this stage are then submitted to subject matter experts for interpretation and guidance in decision-making, thus making it possible to trace a history of the discipline in question.

Intrinsically connected to the primary aim of this study is what type of information the time trajectory of a word occurrence should contain in order to correctly construct and compare words' life cycles and discover the important dynamics of the ideas latent in word groups. Should timing or synchrony be the sole determinant when assessing the similarity of words' life cycles, or should word popularity be considered significant when comparing different words? For FDA, in other words, should we compare word curves only on the basis of their phase variation, or also by accounting for their amplitude variability? This question arises from a typical feature of textual data, the so-called Large Number of Rare Events (LNRE) property, i.e. the presence of a large number of word-types whose probability of occurring is quite low, which implies data sparsity and high skewness. Regardless of any decision concerning these issues, preliminary data processing is advisable in order to adjust for the uneven size of subcorpora (number of texts and their size in word-tokens) over time and hence regularize the "signal". Further data transformation depends on the study's specific aims and is crucial for a consistent reading of results. In this chapter, we propose several kinds of data normalisation which involve different concepts of life cycle similarity and hence a different reading of the history of the discipline under examination.

### *9.1.1 An Outline of the Method*

Given a knowledge field of interest, the KBS consists of two main stages:

1. An information retrieval process that, starting from a large database of scientific articles published by a selection of premier journals in the field, leads to the creation of a well-founded corpus of scientific literature.

2. A statistical learning process that leads to the reconstruction of the important dynamics underlying word micro-histories and hence to an outline of the overall evolution of the knowledge field in four steps:

   (a) Normalisation of time trajectories of word (raw) frequencies, chosen according to aspects of life cycles that are considered substantive when comparing words.
   (b) Filtering time trajectories of word (normalised) frequencies, interpreted as functional data (FD) and thus represented as smooth functions.
   (c) Curve clustering (CC) to detect all important dynamics underlying the evolution of groups of word micro-histories.
   (d) Interpretation by expert opinion to decipher detected dynamics and thus compose a narrative of the evolution of the knowledge field as a whole.

We adopt a basis function approach to filtering with a B-spline basis system. 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. In the illustration, we use the Euclidean distance. 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.

We situate our methodological choices in the literature in Sect. 9.2, and describe the method in greater detail in Sect. 9.3.

## 9.2 Related Literature

The objective we pursue here has some analogies with other research areas though the approaches they propose are markedly different and cannot answer our particular question effectively. We list three major lines of research which are alternative to ours.

Quantitative linguistics often deals with textual data consisting of temporal sequences of linguistic units and, generally, addresses the problem of reading the evolution of a linguistic phenomenon over time by applying linguistic laws, Fourier analysis (and similar methods) or time series analysis. In our study, however, a word trajectory is very unlikely to show a regular behaviour (e.g. that fits a function) and is only apparently a matter of time series analysis. The latter focuses on studying the correlation of observations over time and, normally, seeks a model for prediction. By contrast, the word life cycle is the primary, indivisible unit of our analysis (the functional datum) and our primary goal is to recognise temporal shapes or curves from raw word trajectories.

Topic modelling shares similar aims with our perspective, but only to a certain extent. When topic modelling is applied to documents referenced to time points, co-occurrence (of words within documents) analysis—on which topic modelling is based—can be transmuted to our analysis (Griffiths and Steyvers 2004).

Nevertheless, differences from our approach are evident from their primary aim: unveiling topics (hence mapping science and tracking its evolution) versus tracing life cycles of words (hence dynamics of temporally homogeneous bundles of words in order to decipher the history of a knowledge field). Topic modelling produces clusters of words that should reflect a topic when they appear together in documents (but the shape of word trajectories is not relevant), whereas our approach leads to clusters of words that should evolve similarly over time (but that might represent different topics, different approaches, or different schools of thought). Additionally, it has been shown that Latent Dirichlet Allocation (LDA)—the standard method used for topic modelling—is not the best approach for analysing corpora that include texts of limited length (e.g. titles of articles, Trevisani and Tuzzi 2018 and references therein).

Scientometrics (to which topic modelling connects) or, more in general, quantitative methods for mapping knowledge domains from scientific article databases, are based on term and/or citation co-occurrences in documents, possibly observed over time in order to reconstruct a field's evolution. Recently, many researchers have adopted generative probabilistic models for topic detection and tracking (TDT) or, in general, dynamic science mapping. These models include LDA (despite certain shortcomings that undermine its role—viz., it requires that the number of topics be specified in advance and tends to an even distribution of topics—if the focus is on finding emerging topics and how they evolve over time) and the hierarchical Dirichlet process (HDP), a nonparametric Bayesian model which can automatically decide the number of topics, and is thus considered more competent than LDA in dynamic topic analysis (Ding and Chen 2014). However, traditional topic analysis approaches are relatively static, as they ignore any changes (in both the external representation and the internal content of a scientific topic) that may occur over time. Two recent studies dealing with topic changes and emerging topic detection (ETD) are Zhang et al. (2016, 2017). Both use a term clumping process for core term retrieval, after which the first applies *k*-means-based clustering to obtain topics and finally produces a "roadmapping" that blends historical analysis and expert-based forecasting. The second applies an LDA-based topic model to profile the topic landscape, then a model of scientific evolutionary pathways to detect topic changes and to indicate emerging topics, and lastly, a prediction model to foresee possible topic trends. Another approach to analysing the thematic evolution of a given research field is presented in Cobo et al. (2011) and has been incorporated in SciMAT (Cobo et al. 2012). Recently, the traditional topic evolution map based on text corpora has been extended to more complex subjects like cross-media data (Zhou et al. 2017) and memes (Shabunina and Pasi 2018).

In conclusion, science mapping research is based on co-occurrences in documents possibly observed over time, while our work considers term co-occurrence solely in time, as our primary focus is the temporal evolution of terms. More importantly, our approach differs conceptually from the main alternatives that address the problem of knowledge evolution, such as those developed for TDT, ETD and, generally, for dynamic knowledge mapping in scientometric studies. Our analysis focuses on detecting important dynamics each of which represents the temporal

evolution of a group of words. Thus, on principle, different themes, research fields and approaches can be represented within the same group of words. Conversely, "topic-centered" methods focus first on the structure of science and detecting topics, and then on tracking their evolution. As a consequence, words that represent the same topic may have an irreconcilable temporal evolution. Moreover, topic evolution can only be a roadmap, i.e. an abstract description (the average evolution of words grouped by co-occurrence) of basic movements over time. Additionally, the abstract definition of topics is subjected to continuous destruction and reconstruction by time, making topic tracking a fragile and questionable artefact.

Finally, our choice of specific statistical tools is underpinned by the literature as follows.

The basis function approach is the most widely used for representing FD, and B-splines are a very flexible basis system for non-periodic FD (Ramsay and Silverman 2005). Moreover, B-splines enable us to recognise continuous and regular curves, and hence more easily interpretable shapes. Other systems, e.g. wavelets, can be better suited to the typical bumpy trend of word trajectories (Trevisani and Tuzzi 2015). Upstream, we decided for a distance-based approach, as one of our objectives was to set up an exploratory and mostly automated procedure. In fact, the procedure is called upon to look for interesting patterns—without prescribing any specific interpretation—to be submitted to experts who can potentially formulate new hypotheses and research questions. This eminently exploratory task requires the procedure to be fast and relatively easy to use and understand even by non-statisticians in interdisciplinary groups involved in research projects. The alternative or model-based approach is typically chosen for confirmatory analyses and is generally more demanding in terms of computing and inferential expertise. In a previous study (Trevisani and Tuzzi 2015), we used a functional mixed (normal mixture) model based on a wavelet-based decomposition which proved effective in accommodating the irregularity of word curves and the high inter-word variability, as well as being computationally efficient in a modelling context with high-dimensional data.

Once opted for distance-based methods, *k*-means type clustering algorithms have been widely applied to FD, especially when combined with the finite basis expansion approach. Other strategies which extend the classical *k*-means algorithm with FD are essentially based on functional principal components. However, they are recent extensions, rarely used and, thus, less justifiable as the basis for our explorative approach (some interesting overviews of strategies for clustering FD are provided by Jacques and Preda 2014, and Wang et al. 2016).

Lastly, we opted for the Euclidean distance ($L_2$ metric) for measuring distance between curves since conventional distances between raw data evaluating a one-to-one mapping of each pair of sequences meet our needs. In fact, one of our objectives is to compare curve profiles after data transformation. Accordingly, our strategy entails first transforming data and then seeing what this involves for clustering results by using a distance measure that can approximate the area between two curves as simply as possible. In our application, we used $L_2$ as it is the most popular metric though an equally simple alternative would be the $L_1$ or Manhattan distance.

The alternative way of directly choosing a dissimilarity measure which is invariant to specific distortions of the data is not suitable here, as filtering is to be performed on preprocessed data.

## 9.3 Method in Detail

The first stage of the KBS consists in compiling and constructing the corpus, as Chap. 6 describes in detail. Here, we will review only the main steps of the information retrieval procedure.

Corpus compilation involves a preliminary selection of data sources, i.e. choosing outstanding journals that can cover main topics and represent the temporal evolution of the knowledge field. Text harvesting follows, i.e. downloading information (all references, numbers, issues, volumes) from journal archives to make up the article database. At the end of this step, a diachronic corpus, i.e. a collection of texts including information on their time period, is created. Text under consideration consists of titles and/or abstracts and/or full texts of the articles. Moreover, a corpus is typically organised into subcorpora, or groups of texts sharing the same time reference, thus generating a sequence of text sets along a chronological sequence of time points.

Corpus construction and pre-processing (Chap. 7) involve identifying all words (in the tokenisation stage, words are sequences of letters isolated by means of separators), as well as other possible forms of tagging: stemming, or transforming words into stems; identifying (and ranking) stem-segments (or *n*-stem-grams, i.e. sequences of stems); tagging keywords, or identifying all words (stems and stem-segments) relevant to the specific knowledge field (e.g. by matching the corpus vocabulary with item lists of relevant glossaries for the knowledge field); and thresholding, or selecting all keywords with frequencies at least equal to an appropriate threshold. Finally, the corpus is represented by a words × documents/time points contingency table containing the frequencies of the selected keywords (by row) along the time points (by column) of the period considered.

The second stage of the KBS consists of a stepwise process of statistical learning that enables a distant reading of the diachronic corpus.

### *9.3.1 Normalisation of Word Trajectories*

A diachronic corpus is typically characterised by the following features.

1. Size of subcorpora (number of texts and their size in word-tokens) may vary greatly over time.
2. The LNRE property of textual data, i.e. a large number of word-types whose probability of occurring is quite low, which implies:

(a) The total frequency (or popularity) of individual words in the entire corpus is highly variable,
(b) The frequency spectrum by time point is highly asymmetric,
(c) Frequency sparsity, i.e. many cells of the contingency table have small counts or are empty.

In Sect. 9.4, features (1) and (2) are illustrated in Figs. 9.1 (subcorpora size) and 9.2 (original word trajectories), respectively.

As the foregoing considerations indicate, raw frequency normalisation is necessary to reconstruct and compare the temporal evolution of words. In particular, a form of normalisation by time point should be regarded as preliminary in order to adjust the uneven size of subcorpora across time and hence regularise the "signal". A further form of normalisation by word might be appropriate in order to adjust the great disparity in word popularity, thus making it possible to compare word trajectories by timing (synchrony) regardless of height (popularity). We envision several types of normalisation, of which Table 9.1 gives an excerpt.

Normalisation by column can be obtained, for example, from dividing raw frequencies by the number of texts (option $c_1$) or the total number of word-tokens in texts ($c_2$) for each subcorpus (/ time point), still, by the column sum ($c_3$) or the column maximum frequency ($c_4$) of the data table. Normalisation by row can be obtained, for example, from dividing the raw frequencies by the row sum ($r_1$) or the row maximum frequency ($r_3$) of data table, still, by computing the $z$-scores of word raw frequencies ($r_2$). Double (by both row and column) normalisation ($d$) serves to fix both (1) and (2). The calculation of specific double normalisations ($d_1$ and $d_2$) is illustrated in Sect. 9.4 (Figs. 9.3 and 9.4).

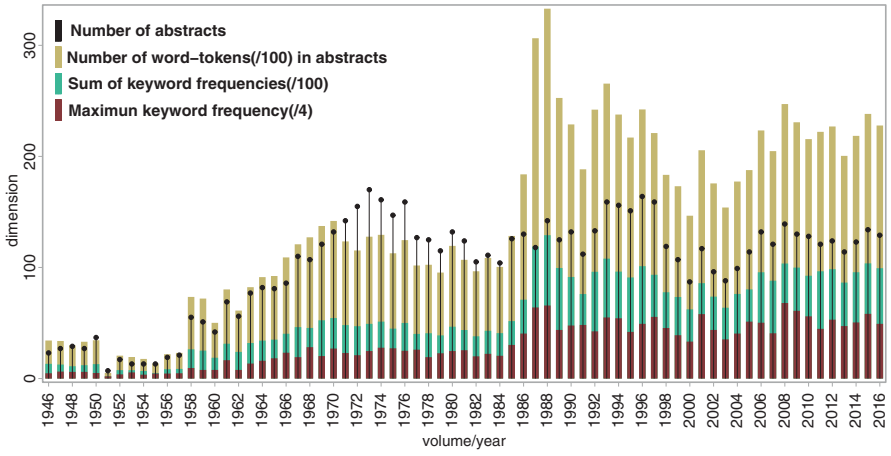


**Fig. 9.1** Subcorpora size: for each volume, number of abstracts (dot-line), total number of word-tokens in abstracts/100, sum of keyword frequencies in data table/100, maximum keyword frequency in data table/4

**Table 9.1** Excerpt of normalisation plan

| Normalisation: | By col | Subcorpus | | Words × documents table | | |
|---|---|---|---|---|---|---|
| By row | | # texts | #tokens | col sum ($\sqrt{\cdot}$) | col max freq | |
| Row sum | | $d$ | $d$ | $d_1$ | $d$ | $r_1$ |
| $z$-score by row | | $d$ | $d$ | $d$ | $d$ | $r_2$ |
| Row max freq | | $d$ | $d_2$ | $d$ | $d$ | $r_3$ |
| | | $c_1$ | $c_2$ | $c_3$ | $c_4$ | |

### *9.3.2 Word Trajectory Filtering*

From an FDA perspective, the functional observation $\mathbf{y}_i = \{y_{ij}\}$ of word $i$ consisting of the set of (normalised) frequencies at time points $t_j = t_1, \ldots, t_T$, for each $i = 1, \ldots, N$, is viewed as a realisation of an underlying continuous function $x_i(t)$—sufficiently smooth or regular—representing the word's temporal evolution. As $\mathbf{y}_i$ is a noisy observation of the underlying $x_i(t)$, an adequate model of their relationship is $\mathbf{y}_i = x_i(\mathbf{t}) + \boldsymbol{\varepsilon}_i$, where $\mathbf{t} = \{t_j\}$ and $\boldsymbol{\varepsilon}_i = \{\varepsilon_{ij}\}$ is a zero mean vector with dispersion matrix $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \Sigma_\varepsilon$. In the standard model, the $\varepsilon_{ij}$s, often termed "measurement errors", are assumed independent across $j$ and homoscedastic with $\sigma_{ij}^2 = \sigma^2$, but, in a more general case, $\Sigma_\varepsilon$ can be regarded as full and time dependent. The following choices are adopted for filtering $x_i(t)$ from $\mathbf{y}_i$ (see Trevisani and Tuzzi 2018, for a detailed description and rationale).

We adopt the basis function approach for representing FD as smooth functions where $x_i(t)$ is expressed as a finite linear combination

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \quad c_{ik} \in \Re, K < \infty$$

of real-valued functions $\varphi_k$ called basis functions (Ramsay and Silverman 2005).

We consider B-spline bases, the most popular basis system for building spline functions, which are piecewise polynomials joined smoothly at the interior nodes.

As regards the positioning of knots—the values of $t$ at which adjacent segments are joined, a direct and reasonable choice is that of placing knots at each point of observation $t_j$.

We adopt the roughness penalty or regularisation approach for smoothing FD, whereby the estimate of $x_i$ is the function minimising the penalised residual sum of squares $\mathrm{PENSSE}(x_i) = \mathrm{SSE}(x_i) + \lambda\ \mathrm{PEN}_r(x_i)$ where $\mathrm{SSE}(x_i)$ is the residual sum of squares measuring the fit to the data, $\mathrm{PEN}_r(x)$ is the penalty term measuring a function roughness (by the integrated squared $r$th derivative over the observation time, $\mathrm{PEN}_r(x) = \int [D^r x(s)]^2 \mathrm{d}s$) and $\lambda$ is a smoothing parameter. Thus, $\lambda$ measures the tradeoff between fit to the data and roughness of the function $x$: as $\lambda \to 0$ the fitted curve approaches an interpolant to the data, as $\lambda \to \infty$ the condition $\mathrm{PEN}_r(x) \to 0$ means the fitted curve is a spline of order $r$.

Choosing $\lambda$ is part of the model selection issue. A standard practice for choosing $\lambda$ is to use cross-validation (CV). When tuning a smoothing parameter, a common choice is the leave-one-out CV, however, it may be computationally intensive especially for large sample sizes and lead to under-smoothing. Generalised cross-validation (GCV), $\mathrm{GCV}(\lambda) = T / \left(T - \mathrm{df}(\lambda)\right)^2 \mathrm{SSE}(\hat{x}_i)$, provides a convenient approximation to leave-one-out CV for linear fitting under squared error loss. df($\lambda$) is the effective degrees of freedom under regularisation, which is monotone decreasing in $\lambda$ with maximum equal to $K$ when $\lambda = 0$. GCV can sufficiently remedy the tendency to under-smoothing unless the sample size is small or moderate (Lukas et al. 2016; Ramsay and Silverman 2005).

We smooth the data by varying

- Spline order $m$ (from 1 to 8);
- Roughness penalty order $r$: Besides the standard $r = m - 2$, $r = 2$ for $m > 3$, $r = 1$ for $m > 2$, $r = 0$:
- Smoothing parameter $\lambda$ over an appropriate range of values ($\log_{10}\lambda$ from $-6$ to 9).

The GCV criterion is used to select the optimal smoothing.

Calculation is carried out in the R software environment using the *fda* library and an ad hoc routine (R core team 2017). Optimal smoothing selection is illustrated for the case of $d_2$ normalised data in Sect. 9.4 (Figs. 9.5 and 9.6).

### *9.3.3 Curve Clustering*

In a distance-based approach to CC, we apply a $k$-means algorithm for FD where the distance between curves is approximated by using the discretely observed evaluation points of the estimated curves $x_i(t)$ (Jacques and Preda 2014). We use $L_2$ metric, though several options for distance besides the conventional ones can be taken from the broad range of dissimilarity measures available for time series clustering (Montero and Vilar 2014). Moreover, for each cluster number ($k$ from 2 to an appropriate range maximum), we re-run the algorithm starting from 20 different initial configurations set through the $k$-means++ seeding method.

At this step of our KBS, clustering validation is performed by using the internal information of the clustering process to evaluate the goodness of a clustering structure without reference to external information. External validation is postponed to the next step and consists of an informal assessment by subject matter experts who can decide to what extent a clustering is meaningful to them (see subsection below). Internal validation can be also used to decide what the most appropriate number of clusters is in a certain application. In our research context, no "natural" or "true" clusters exist in the available data, so there is not even a "true" number of clusters. Since we only have to be reasonably confident that nothing important is left unexplored, the idea is that of identifying a set of best candidates for cluster number by pooling the ratings from a large number of clustering quality criteria (about 50,

see Desgraupes 2016, and Genolini et al. 2015). It is well known that one index does not fit all situations, rather, the many existing indexes can be grouped into different types each measuring a different aspect of clustering quality. Accordingly, and given the exploratory and evocative task of our clustering, we have gathered a large basket of indexes in order not to favour any single criterion, as each in principle is equally valid. These include measures of within-cluster homogeneity, e.g. *Ball-Hall, Banfeld-Raftery, C-index, Gap*, *Krzanowski-Lai*, *Marriot*, *Scott-Symons*; of between-cluster separation, e.g. *Rubin, Scott*, *Ratkowsky-Lance*; and of their combination, e.g. *Calinski-Harabasz*, *Davies-Bouldin*, *Dunn* and its generalisations, *Gamma*, *Hartigan*, *McClain, PBM, Point-Biserial, Ray-Turi, SD, Silhouette, Friedman, Xie-Beni, Tau*; as well as measures of similarity between the empirical within-cluster distribution and distributional shapes such as the Gaussian distribution, e.g. *BIC, AIC* and their variants.

In detail, cluster number selection includes the following steps, in order:

- A cluster number ranking is computed for each quality index.
- All the rankings are pooled and, for each cluster number, the frequency of being ranked first (top-1), second (top-2), third (top-3) and fourth (top-4) is calculated.
- An ordered set of best candidates for cluster number is retrieved from a qualitative inspection of the graphical representation of the frequencies of being in the top four positions for each cluster number (an R code that essentially mimics the visual selection was developed in order to make the procedure automated without the need for human "eye").

An example of this procedure is illustrated in Sect. 9.4 (Fig. 9.7).

For each candidate for cluster number, the best partition between the 20 replications must be chosen. But, in our approach to cluster number selection, there may be multiple criteria that ranked the candidate in the top positions. Then, we compare, for each cluster number, all the distinct partitions resulting from these multiple criteria by concordance measures. In particular, we consider the Rand index (Rand 1971) as measure of agreement between two clusterings and propose a generalisation of it for comparing more than two clusterings, thus obtaining a measure of concordance between multiple clusterings. Moreover, this "multiple" Rand index can be computed at several levels (of individual words, of single clusters as well as of the overall partition), thus offering a measure of stability of clustering results for each of these levels. The standard Rand index calculates the rate of pairs of units that are classified in the same way (i.e. pairs that are in the same cluster or in different clusters, respectively) in both clusterings. We use the standard version for choosing the best partition for each cluster number as the one that maximises the average of the agreement measures with each of the other partitions selected for the cluster number. Namely, the best partition for each cluster number is the one that best mediates between all the partitions of the cluster number. In addition, we use the multiple Rand index to provide a measure of individual agreement for each word and thus investigate whether a particular word is consistently grouped or separated from other words by different partitions. The information on individual words can help to

screen out "wird" words with very low agreement measures. The average of such measures of individual agreement over a cluster gives a measure of agreement per cluster; the average over the entire corpus coincides with the multiple Rand index of global agreement between multiple clusterings here proposed.

The R software environment contains several *k*-means implementations as well as libraries for computing clustering quality criteria. Our procedure uses the kml routine (Genolini et al. 2015) which is designed specifically for longitudinal data and provides various efficient methods of *k*-means initialisation. The *clusterCrit*, *cclust*, *clusterSim* and *kml* libraries are used to source the quality criteria considered by our method. Ad hoc functions have also been developed for specific criteria and for calculating the multiple Rand index of local and global agreement between partitions.

### *9.3.4 Substantive Expertise*

Clustering results obtained with the cluster numbers selected as the best candidates are then presented to subject matter experts. To facilitate the comparison between different groupings (partitions with different number and composition of groups), we assess the congruence of different partitions by calculating their indexes of agreement (Wagner and Wagner 2007) and visualising set overlaps (see mosaic plots in Chap. 6). Experts try to interpret the latent content of word groups as a consistent ensemble of topics, methods and research areas, and to identify temporal phases and processes from group dynamics in order to reconstruct a historical narrative of the knowledge field. Where possible, they will be instrumental in suggesting other analyses.

## 9.4 Illustration

For illustration, we apply the KBS to the corpus of abstracts of scientific papers published by the *Journal of the American Statistical Association* (JASA) in the time span 1946–2016 in order to trace a history of statistics. The extensive analysis is presented in Chap. 6. Here, we will track the main steps to provide an exemplification of the theoretical method outlined above.

### *9.4.1 Corpus Collection and Construction*

JASA is the oldest statistical journal and has long been considered the world's premier review in its field. We downloaded from online resources all references, abstracts and metadata of articles published in the period 1946–2016 (71 years,

from Volume No. 41, Issue No. 234, to Volume No. 111, Issue No. 516). Abstracts of articles constitute the text corpus considered in this study. The corpus includes 7221 abstracts, 1,029,251 word-tokens (word occurrences) and 26,686 word-types (distinct words). After stemming, all potentially relevant stem-segments are identified. Relevant statistical keywords (stemmed words and sequences of stemmed words) are then tagged.

Lastly, fixing the threshold at 50, 1351 keywords are selected. At the end, the corpus yields a 1351 (words) × 71 (time points/volumes) contingency table (see an excerpt in Table 6.2, Chap. 6).

### 9.4.2 *Normalisation*

For illustration, we choose to transform data (Fig. 9.2) by the double normalisations $d_1$ and $d_2$ (Table 9.1). Let $n_{ij}$ be the raw frequency of word $i$ at time point/volume $j$, $n_{i.}$ the $i$-row sum, $n_{.j}$ the $j$-column sum and $n$ the matrix total of the corpus table. Then, the $d_1$ normalised frequency is computed as $y_{ij} = n_{ij}/(n_{i.}\sqrt{n_{.j}/n})$ and is equivalent to calculating a $\chi^2$ distance between original word profiles if the Euclidean distance is used as measure of dissimilarity ($n_{.j}/n$ is the $j$-column mass in correspondence analysis). Note that this double normalisation produces a somewhat

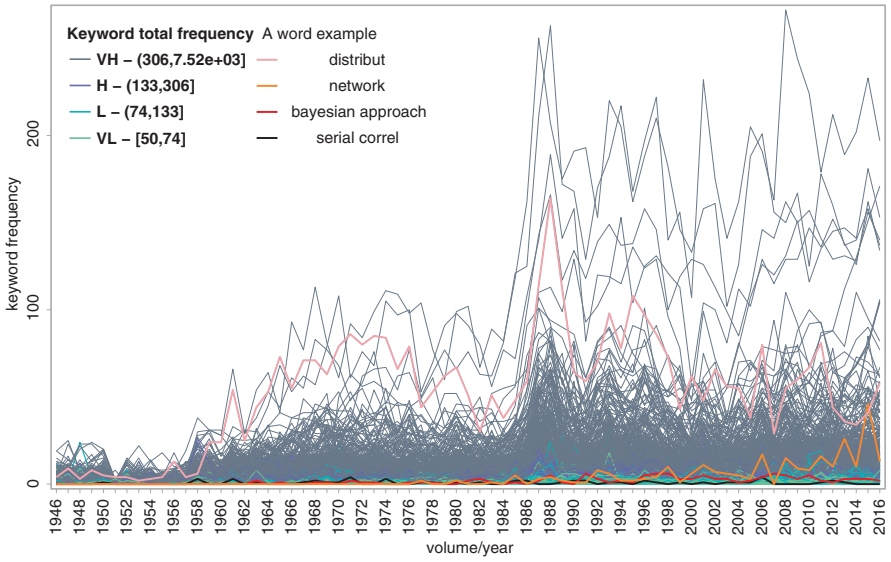


**Fig. 9.2** Word trajectories (original data): $y$-axis represents the raw word frequency for each volume; $x$-axis represents the volume publication year; line colour identifies the word frequency class (Very Low, Low, High and Very High denote equal-frequency intervals of total word frequency). A word example for each class is superimposed

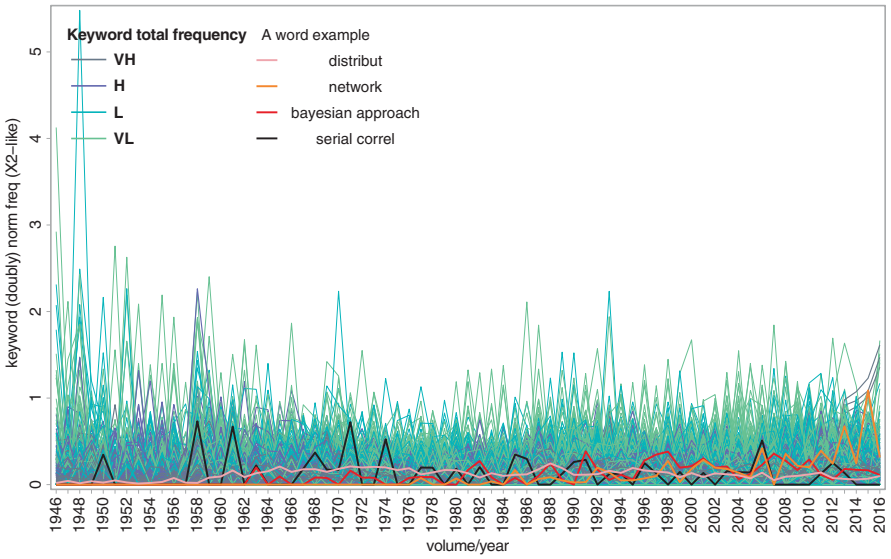


**Fig. 9.3** Keyword trajectories (doubly normalised data, $d_1$ or $\chi^2$-like)

reversed asymmetry (low-frequency words tend to dominate being the associated curves larger in amplitude; see Fig. 9.3). This is mainly due to a greater sparsity of low-frequency words across time. The problem of asymmetry is instead substantially reduced by $d_2$, thus allowing a comparison between curves mainly in terms of horizontal or phase variation (Fig. 9.4). Let $N_j$ be the total number of word-tokens in the subcorpus $j$ and $M_i$ the $i$-row maximum frequency of the column-normalised frequencies $n_{ij}/N_j$. Then, the $d_2$ normalised frequency is computed as $y_{ij} = n_{ij}/(M_i N_j)$. However, $d_2$ cannot completely remedy the problem of sparsity. Rare words tend to have sparse trajectories (i.e. to have zero or almost zero frequency for relatively long stretches of the period, either continuous or intermittent) and very high differences of amplitude along the trajectory (being frequency values little spaced; see, e.g. panels of *semiparametric model* and *realistic*—at left-most, third and sixth rows, respectively, in Fig. 6.15, Chap. 6). On the contrary, words with a high or very high popularity 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. panels of *nonparametric* and *simulation*, left-most panels of first row in Fig. 6.15).

### *9.4.3 Filtering*

Optimal smoothing for $d_2$ normalised data is achieved with spline order $m = 7$ and smoothing parameter $\lambda = 10^{1.5}$ (df = 6.56) under a roughness penalty of order $r = 1$ (Fig. 9.5).

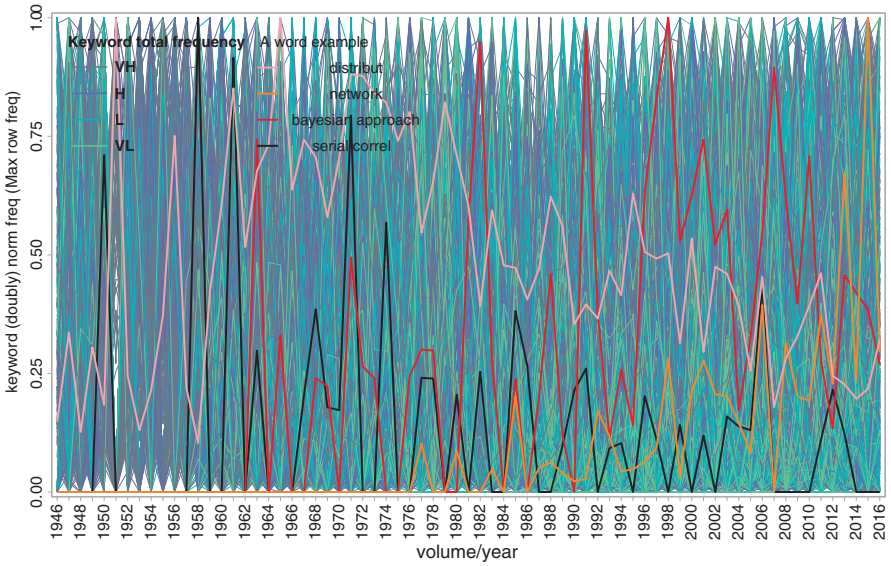


**Fig. 9.4** Keyword trajectories (doubly normalised data, $d_2$)

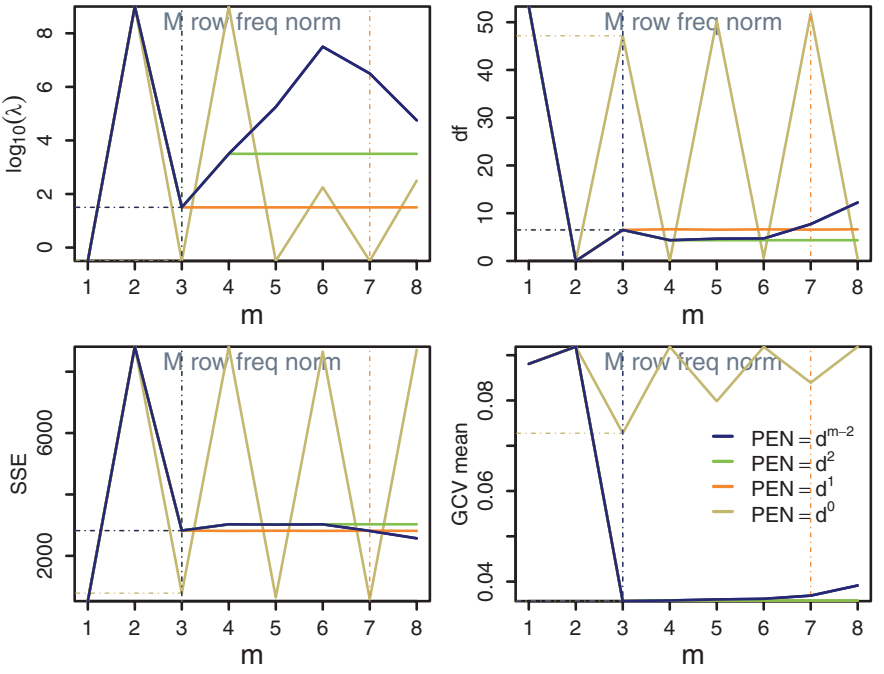


**Fig. 9.5** Smoothing selection: optimal $\lambda$ (top-left) and corresponding effective degrees of freedom (df, top-right), sum of square errors (SSE, bottom-left) and GCV (bottom-right) by varying spline order $m$ and roughness penalty order $r$ ($PEN_r$). Optimal smoothing is obtained by minimising GCV. $d_2$ normalisation

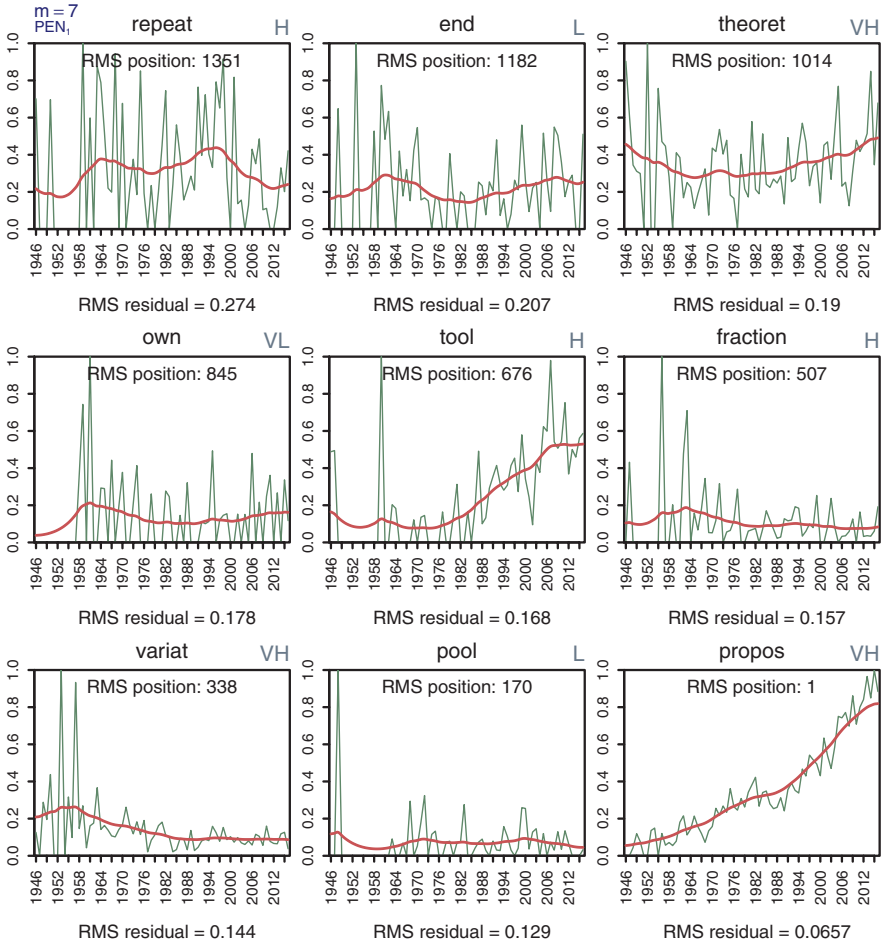


**Fig. 9.6** Optimal smoothing: fit a selection of fitted curves ordered according to decreasing root mean square (RMS) residual. Fit of a smoothing spline of order $m = 7$, with $PEN_1$, to $d_2$ normalised data

A sample of curves fitted by the optimal smoothing is shown in Fig. 9.4, from the word with highest root mean square (RMS) residual (*repetition*) to the word with lowest RMS residual (*proposal*) (Fig. 9.6).

### *9.4.4 Curve Clustering*

Curves are partitioned by means of the *k*-means algorithm combined with the $L_2$ metric with cluster number *k* ranging from 2 to 26 and 20 re-runs for each *k*.

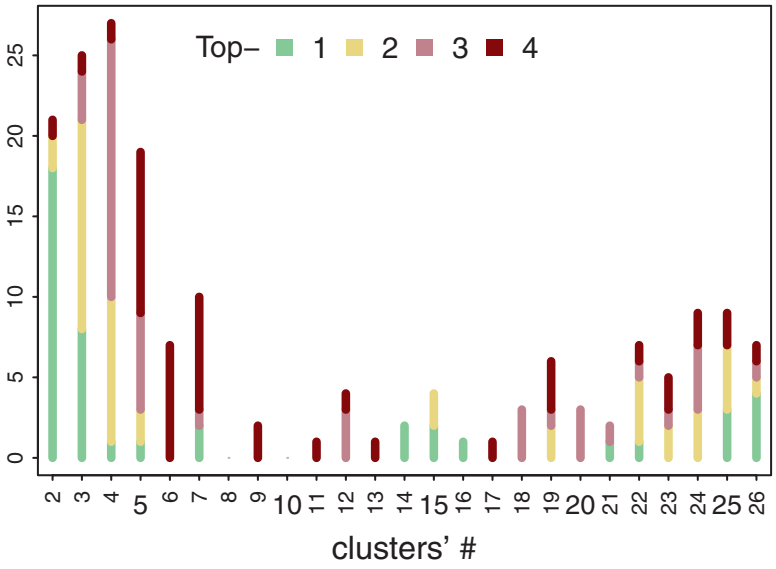


**Fig. 9.7** Cluster number selection: frequency of being ranked first (top-1), second (top-2), third (top-3) and fourth (top-4) for each cluster number by pooling rankings from the overall quality criteria. $d_2$ normalisation

A set of more than 50 quality criteria are then computed in order to identify a set of best candidates for cluster number. Visual representation of the cluster number rating (Fig. 9.7) shows that: (1) partitions into two/three clusters are the best rated; (2) partitions with a cluster number close to the maximum of the considered range have also been frequently selected in the highest positions; (3) in the range of more interesting cluster numbers (neither too low nor too high), the most selected in the top four positions are 4/5 and secondarily 7/15 (in reading the figure, note that bar height corresponds to the cumulated frequency of being in the top four, and colour indicates the position level). This ranking is the output of an R code that essentially mimics a qualitative rating based purely on a graphical inspection.

Discarding the less interesting solutions (1) as well as (2) (which on the one hand may reflect the lack of a defined structure and parsimonious grouping, but on the other may be a failure due to the standard assumption underlying many quality criteria of normally distributed data and hence of compact and convex clusters), the method produces the best partitions corresponding to cluster numbers that emerged at (3) in order to subject them to the scrutiny of experts. In particular, given the set of quality criteria that ranked the cluster number in the top positions and the partitions selected by these criteria, the partition which maximises the average Rand index of agreement with all the other selected partitions is chosen as the best partition for each cluster number.

### 9.4.5 *Substantive Expertise*

Here, we illustrate the best partition found with the cluster number ranked first, i.e. $k = 4$. It corresponds to partition 3, out of the 20 replications, as it maximises the average Rand index of agreement with the other partitions (2, 19) selected by multiple criteria (top panel of Fig. 9.8). The graphical output shows the groups, with the cluster mean patterns superimposed, together and individually (Fig. 9.8). Individual clusters are chronologically ordered, from the cluster of words that have tended to disappear to the cluster of emerging words in the period 1946–2016 (A, C, D, B in Fig. 9.8), in order to facilitate the identification of subsequent stages in the knowledge field's evolution.

The multiple Rand index of agreement for the overall partition and for single clusters can provide a measure of the stability of the results. The transcribed words—which are just a subset of group words—are ordered according to both their

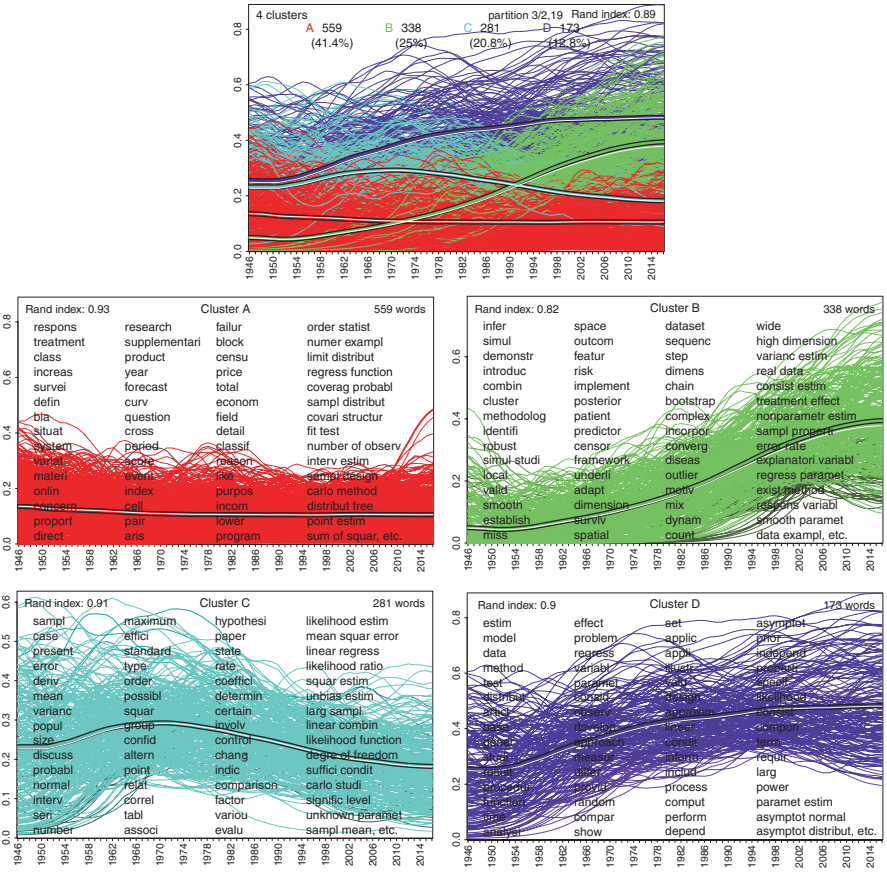


**Fig. 9.8** Clustering on $d_2$ doubly normalised data: all four groups and individual clusters

popularity (from highest to lowest) and their individual multiple Rand index (from highest to lowest). However, a portion of these is dedicated to multiwords whose total frequency is often low or very low (if they are not among the most popular words, they are generally added in the last column of the cluster graph).

The dynamics thus found are then examined and, if considered interesting, are interpreted by subject matter experts. A possible reading of the history of statistics on the basis of the illustrated findings is offered in Chap. 6 where finer partitions—corresponding to the other candidates for cluster number, namely, 5, 7, and 15—are also examined for a more in-depth and detailed interpretation. An interesting nested structure will be found for this particular case of study.

## 9.5 Concluding Remarks

Normalisation of raw frequencies is critical to reconstruct and compare words' temporal evolution appropriately. The choice of normalisation depends on the interplay between cyclical synchrony and popularity level of words, which underpins the concept of word similarity and ultimately leads to word clustering. This point is discussed in Sect. 9.3.1 and illustrated in Sect. 9.4.

In this study, word trajectories interpreted as FD have been filtered by a basis expansion approach whereby the infinite-dimensional FD are projected onto a low-dimensional space of a set of basis functions. Here, we have chosen B-splines as pre-specified basis functions. A data-driven approach to basis function specification is also possible: functional principal component analysis (FPCA) is a dimension reduction tool that can be used as a method for constructing an optimal orthogonal basis of fixed dimensionality as well. Indeed, functional principal components (FPCs) are often referred to as empirical basis functions. We intend to extend the KBS by providing it with this alternative method of FPC expansion that, among all basis expansions that use $K$ components for a fixed $K$, explains most of the variation in the FD.

In this study, a two-stage CC via functional basis expansion has been presented. Taking up the finite approximation FPCA approach mentioned above, an alternative can consist of a two-stage CC via FPCA where a $k$-means algorithm is used on the FPC scores (Peng and Muller 2008). An even more refined method is the FPC subspace projected $k$-centres functional clustering algorithm (Chiou and Li 2007) whereby cluster centres are identified as subspaces, which account for both the means and the modes of variation differentials between clusters, rather than as cluster means only (like for the $k$-means algorithm).

Lastly, contrary to two-stage methods, in which filtering is done previously to clustering, we intend to explore strategies performing these two tasks simultaneously (like with model-based techniques). For example, to identify optimal subspaces for clustering and optimal clusters of functions simultaneously, Yamamoto (2012) developed an alternate algorithm which optimises an objective function defined as the sum of the distances between the observations and their projections plus the distances between the projections and the cluster means.

## References

Chiou, J. M., & Li, P. L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B, 69*, 679–699.

Cobo, M., López-Herrera, A., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory Field. *Journal of Informetrics, 5*(1), 146–166.

Cobo, M., López-Herrera, A., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology, 63*(8), 1609–1630.

Desgraupes, B. (2016). clusterCrit: Clustering indices, R package version 1.2.7.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science and Technology, 65*(10), 2084–2097.

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software, 65*(4), 1–34.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl 1), 5228–5235.

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

Lukas, M. A., de Hoog, F. R., & Anderssen, R. S. (2016). Practical use of robust GCV and modified GCV for spline smoothing. *Computational Statistics, 31*(1), 269–289.

Montero, P., & Vilar, J. (2014). Tsclust: An R package for time series clustering. *Journal of Statistical Software, 62*(1), 1–43.

Peng, J., & Muller, H. G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Annals of Applied Statistics, 2*, 1056–1077.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*(336), 846–850.

Shabunina, E., & Pasi, G. (2018). A graph-based approach to ememes identification and tracking in social media streams. *Knowledge-Based Systems, 139*(Suppl C), 108–118.

Trevisani, M., & Tuzzi, A. (2012). Chronological analysis of textual data and curve clustering: preliminary results based on wavelets. In Società Italiana di Statistica (Ed.), *Proceedings of the XLVI Scientific Meeting* (pp. 1–4). Padova: Cleup.

Trevisani, M., & Tuzzi, A. (2013a). Shaping the history of words. In I. Obradovic, E. Kelih, & R. Köhler (Eds.), *Methods and applications of quantitative linguistics: Selected papers of the VIIIth international conference on quantitative linguistics* (pp. 84–95). Belgrad: Academic Mind.

Trevisani, M., & Tuzzi, A. (2013b). Through the JASA's looking-glass, and what we found there. In Proceedings of the *28th International Workshop on Statistical Modelling* (vol. 1, pp. 417–422). Istituto Palermo: Poligrafico Europeo.

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). New York: 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

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

Yamamoto, M. (2012). Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification, 6*, 219–247.

Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting and Social Change, 105*, 179–191.

Zhang, Y., Chen, H., Lu, J., & Zhang, G. (2017). Detecting and predicting the topic change of knowledge-based systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowledge-Based Systems, 133*(Suppl C), 255–268.

Zhou, H., Yu, H., Hu, R., & Hu, J. (2017). A survey on trends of cross-media topic evolution map. *Knowledge-Based Systems, 124*(Suppl C), 164–175.