Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories

Matilde Trevisani[a,*], Arjuna Tuzzi[b]

[a] Department of Economics, Business, Mathematics and Statistics (DEAMS), University of Trieste, Via Tigor 22, Trieste 34124, Italy
[b] Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA), University of Padova, Via M. Cesarotti 10/12, Padova 35123, Italy

ARTICLE INFO

Article history:
Accepted 31 January 2018

Keywords:
Diachronic corpora
Chronological textual data
Keyword retrieval
Curve clustering
Functional data analysis
Normalization

ABSTRACT

The growing availability of large diachronic corpora of scientific literature offers the opportunity of reading the temporal evolution of concepts, methods and applications, i.e., the history of disciplines involved in the strand under investigation. After a retrieval process of the most relevant keywords, bag-of-words approaches produce words × time-points contingency tables, i.e. the frequencies of each word in the set of texts grouped by time-points. Through the analysis of word counts over the observed period of time, main purpose of the study is, after reconstructing the "life-cycle" of words, clustering words that have similar life-cycles and, thus, detecting prototypical or exemplary temporal patterns. Unveiling such relevant and (through expert opinion) meaningful inner dynamics enables us to trace a historical narrative of the discipline of interest. However, different history readings are possible depending on the type of data normalization, which is needed to account for the fluctuating size of texts across time and the general problems of data sparsity and strong asymmetry. This study proposes a methodology consisting of (1) a stepwise information retrieval procedure for keywords' selection and (2) a functional clustering two-stage approach for statistical learning. Moreover, a sample of possible normalizations of word frequencies is considered, showing that the different concept of curve similarity induced in clustering by the type of transformation heavily affects groups' composition and size. The corpus of titles of scientific papers published by the American Statistical Association journals in the time span 1888–2012 is examined for illustration.

1. Introduction

A diachronic corpus is a collection of texts including information on the time period to which they relate, e.g. the publication date of a document. In many situations these texts are arranged into groups (subcorpora) that refer to the same time interval, thus generating a sequence of text sets corresponding to chronological points on the time-axis. Diachronic corpora represent the ideal ground for studying the history of a "language", e.g., when a corpus is able to reflect the relevant features of a text genre in a well-defined time period, the temporal evolution of word occurrences (frequencies) mirrors the historical development of the corresponding concepts [see, e.g., [1–5]]. In our work, the temporal course of a word occurrence is viewed as a proxy of the word diffusion and vitality, i.e. of the word "life-cycle". The idea of studying the words' "quality of life" is innovative in language studies [4,6–9] since the history of language has always endeavored to date the birth (or, sometimes, semantic changes) of individual words but paid little attention to their fortunes (or death), i.e. to recognize words whose presence has grown over time, or that have disappeared, or that were very successful for a limited period and then fell into oblivion, etc.

In bag-of-words approaches a diachronic corpus originates a words × time points contingency table that reports the frequency of each word at each time point. In this table, rows—each representing discrete observations of a word time trajectory—are the ideal basis to reconstruct words' life-cycles, hence cluster words having similar life-cycles and detect (as cluster prototypes) any prototypical or exemplary temporal patterns. If a corpus draws upon the scientific literature of a discipline, unveiling important (as they are shared by groups of words) and meaningful (as the latent content of word groups is interpreted through expert opinion as a consistent ensemble of topics, methods and research areas) in-

* Corresponding author.
E-mail addresses: matilde.trevisani@deams.units.it (M. Trevisani), arjuna.tuzzi@unipd.it (A. Tuzzi).

https://doi.org/10.1016/j.knosys.2018.01.035

ner dynamics enables us to identify temporal phases and processes hence to trace a historical narrative of the discipline under investigation.

Since we are interested in shaping word histories, we adopt a functional data analysis (FDA) approach under which observations through time are viewed as a realization of an underlying continuous function representing the temporal development of a word.

In order to compare time trajectories of counts, further processing is required. First, size of subcorpora (number of texts and their size in word-tokens) may vary greatly over time. Hence, a sort of data normalization by time point should be regarded as preliminary in order to adjust for the uneven document dimension across time. Second, total frequency or "popularity" of individual words in the entire corpus is greatly variable. As well as a strong asymmetry characterizes the frequency spectrum by time-point, showing the typical pattern of textual data where there are "few giants, many dwarves", namely a large number of word types having a quite low probability of occurring. From the foregoing, a sort of normalization by word is needed to compare word curves in terms of their phase variation (or synchrony) rather than of their amplitude variability (or height). Lastly, we just note that a common problem in FDA known as curve registration, i.e. the alignment of curves conceived as having the same underlain pattern but appearing with a lateral displacement/deformation, is not a concern to us. Rather, it is the timing or phase variation to be of essential interest for clustering, movements in different periods are not comparable when the "clock time" is that of the history, the object of analysis [10,11].

The most proper data transformation depends on the study aims and it is crucial for a consistent reading of results. In this work, we will examine how different data transformations affect clusters' generation.

1.1. Method details

From the foregoing overview of perspectives and issues, this paper aims at introducing a procedural method that, starting from a large database of scientific articles published by a selection of premier journals of the discipline (or, more in general, knowledge field) of interest, leads to the creation of a well-founded corpus of scientific literature (organized as a keywords $\times$ time-points matrix) and from this to a possible outline of the history of the discipline (or knowledge field) under investigation. As regards the latter, the detection of groups of words having a similar life-cycle and, thus, of prototypical (or exemplary) temporal patterns, allow us to uncover the inner dynamics of the latent (to the word groups) topics, methods and research areas, hence to trace a possible evolution of the discipline. The proposed method consists of (1) a stepwise information retrieval procedure for keywords' selection and (2) a functional clustering two-stage approach for statistical learning. Step (1) starts with a massive download of any references to all papers published by a mainstream scientific journal and ends with a contingency table including the frequencies of the most relevant keywords along a set of time-points. Step (2) can be divided into two stages: (a) a filtering step in which functional data (FD) are represented as smooth functions by a basis-expansion method, (b) a distance-based clustering in which the k-means algorithm is used combined with an opportunely chosen metric to measure distance between curves. In (a) B-splines are used as they consist in a very flexible basis system for non-periodic FD, then, in (b) distance is estimated on the evaluation points of the approximating curves.

Intrinsically connected to the underlying aim is the crucial choice of how to properly normalize word raw frequencies. Any decision on data transformation is likely to lead, through changing the similarity frame between words, to a different history reading. In this study, three different types of normalization will be in particular analyzed: one by time dimension and two by both time and word dimensions.

For illustration, we conduct an analysis of the life cycle of the most relevant keywords that occurred in the titles of papers published by the *Journal of the American Statistical Association* (JASA) and its predecessors in the time span 1888–2012.

1.2. Related work

In quantitative linguistics a number of textual features can be observed in form of linear sequences of linguistic units and/or their properties. The general problem of reading over time the evolution of a linguistic phenomenon is often tackled by resorting to linguistic laws [9,12,13], Fourier analysis (and similar) or time series analysis [14]. But, in our study, a word trajectory hardly shows a regular behavior (e.g. that fits a function) and is only apparently a matter of time series analysis: this last is focused on studying the correlation of observations over time and, normally, seek a model for prediction; the output is not a "shape" or a curve.

Another important research area which partially shares aims similar to our project is topic modeling. A leading reference work is by Griffiths and Steyvers [15] who conducted a co-word analysis of abstracts of papers published in PNAS from 1991 to 2001, and introduced a Latent Dirichlet Allocation (LDA) generative model to discover topics covered in the corpus. Analogies with [15] can be established thanks to the particular data they examine: documents can be referenced to time-points, hence co-occurrence is doubleface (in documents/times); (latent) topics resemble our clusters and topic number selection parallels our cluster number selection; topic dynamics is an elementary and summary version of our temporal patterns. Nevertheless, differences appear evident in the primary research object: unveiling topics (hence mapping science and possibly tracking its evolution) versus tracing life-cycles of words (hence dynamics of temporally homogeneous bundles of word curves to eventually decipher the history of a knowledge field). In other words, topic modeling produces clusters of words that should reflect a topic as they appear together in documents (but the shape of word trajectories is not relevant), our approach leads to clusters of words that should evolve similarly over time (but that might represent different topics, different approaches, different schools of thought). Still, LDA is a model-based approach whereas ours is an unsupervised learning methodology. Lastly, notwithstanding LDA is very popular in text mining applications and its aims are similar to ours, some experiments demonstrated that it does not represent the best approach to analyse corpora that include texts of limited length [cfr. 16, 17, 18], as in our application (see Section 2).

Topic modeling connects to scientometrics or, more in general, quantitative methods for mapping knowledge domains from scientific articles databases. Co-citation and co-word methods have long been used for designing knowledge maps. Some approaches propose to use both terms occurrences and references. Recently, many researchers have adopted generative probabilistic models to topic detection and tracking (TDT) or, in general, dynamic science mapping: LDA, yet some deficiencies undermined its role (it requires to specify the number of topics in advance and tends to an even distribution of topics) if our interest is in finding emerging topics and how they evolve over time; hierarchical Dirichlet process (HDP), a nonparametric Bayesian model which can automatically decide the number of topics, thus considered more competent than the former in dynamic topic analysis [19]. However, traditional approaches of topic analysis are relatively static; they ignore any possible change (in both the external representation and the internal content of a scientific topic) resulting from time. Two recent works facing with topic changes and emerging topic detection (ETD) are [5,20] which deal, in particular, with science, technology and inno-

vation research and, respectively, the journal *Knowledge-based Systems* (KnoSys) topics. Both use a term clumping process for core term retrieval, then: [20] applies a *k*-means-based clustering to obtain topics and finally produces a "roadmapping" blending historical analysis and expert-based forecasting; Zhang et al. [5] applies a LDA-based topic model to profile the topic landscape, then a model of scientific evolutionary pathways to detect the topic changes in sequential time slices and to indicate emerging topics, finally, a prediction model to foresee possible topic trends. Another approach to analyze the thematic evolution of a given research field is presented in [21] and carries out the following stages: co-word analysis is used in a longitudinal framework to create a science map showing by thematic networks the different themes or topics treated by the scientific field for each subperiod in a given time, strategic diagrams and thematic areas are used to study the thematic evolution of the research field, finally, performance analysis is used to quantify the impact of the research field and its conceptual subdomains. This approach has been incorporated into SciMAT [22] and used to perform a science mapping analysis of the scientific content of KnoSys from 1991 to 2014 [23], among other applications. Recently, the traditional topic evolution map based on text corpora has been extended to more complex subjects like cross-media data [24] and memes [25]. An interesting overview, from an epistemological perspective, of the literature on dynamic science mapping has been provided by Chavalarias and Cointet in [26]. The same authors have recently presented a methodology that reconstructs the dynamics of scientific fields and relies on an adaptation of the concept of the phylogenetic tree in analogy of the evolution of science with the evolution of living organisms [2]. Our work presents various connections with [2] for the assumption that scientific fields evolve over time like living organisms, for implementing a bottom-up unsupervised learning procedure, and (in common also with [5,20]) for an accurate keyword selection. However, science mapping research is based on co-occurrences in documents, possibly observed over time so allowing the reconstruction of science evolution, whilst our work considers term co-occurrence solely in time being the temporal evolution of terms the primary focus. Indeed, the original aim of shaping the life-cycle of words reflecting their quality of life continues to be a distinctive feature.

Finally, we incorporate our choices of spline bases and *k*-means algorithm for functional clustering in the related literature. In a previous study, within a model-based approach to curve clustering, we applied a wavelet-based decomposition which proved successful in recognizing the typical bumpy trend of word trajectories (see Fig. 1), besides being more computationally efficient in a modeling context [4]. This time our objectives are recognizing continuous hence more easily interpretable shapes—leading us to opt for the splines—and setting up an exploratory and mostly automated procedure (equipped with an R package)—making us decide for a distance-based approach. This work connects to a project (Section 6) involving an interdisciplinary research group (linguistics, philosophy, psychology, sociology, statistics), whose aim is to construct corpora of scientific literature by extracting important keywords of a discipline from main specialized journals, and, hence, to investigate whether a discipline history can be traced from analyzing the keywords' temporal pattern. Then, the procedure is asked to look for "interesting patterns", without prescribing any specific interpretation, to be submitted to experts who potentially formulate new research questions and hypotheses and drive to research insights. This eminently exploratory task requires the procedure to be fast and relatively easy to use and understand even by non-statisticians of the research group. The alternative approach to functional clustering, i.e. that model-based, is typically chosen for confirmatory analyses and is generally more demanding in terms of computing and inferential expertise. Functional model-based clustering is standardly based on finite mixture models and classically assumes Gaussian processes for the mixture distributions [27–29]. More recently mixed effects models have been introduced [30,31] and either non-Gaussian distributions [32] or, within the Bayesian framework, Dirichlet processes [33–35] were assumed for mixture components. In [4] we used a functional mixed (normal mixture) model based on a wavelet-based decomposition adapting the approach developed in [31] to the analysis of chronological corpora.

Within distance-based methods, *k*-means type clustering algorithms have been widely applied to FD especially when combined with the (most widely used) finite basis expansion approach to FD. Other choices, which extend the classical *k*-means algorithm with FD, are: *k*-means algorithm on the functional principal components (FPC) scores (it performs dimensionality reduction and clustering simultaneously), FPC subspace-projected *k*-centers functional clustering approach (the subspace spanned by the FPC define cluster centers) [see 36,37, for a survey], functional principal points (FPP, are the equivalent to *k*-means cluster centers when FD are defined as random variables, [38,39]). Though, they are more recent extensions, rarely used, thus less justifiable as the basis for our explorative approach. Indeed, several authors argue that the type of algorithm is not all that important when the use of clustering is data exploration and results are used in a qualitative context (such as ours). It is the design of the human computer interaction interface, visualization and data manipulation capabilities of the system to be more responsible for success or failure of the attempt to discover structure in the data [40].

Finally, we mention other strategies for clustering FD: raw data methods (clustering is performed directly on the discretely observed raw data, hence is not performed on the signal but on noisy data); filtering methods (clustering is performed on the finite set of parameters resulting from the filtering step); adaptive methods (functional representation and clustering are performed simultaneously); distance-based methods (adaptation of geometric clustering algorithms for FD with a proper definition of distance between curves). Our method falls in the last category and connects to the second, as we use a geometric clustering combined with a (classical) distance between curves which is approximated by using the discretely observed evaluation points of the estimated curves [see details in 36].

2. Material: The corpus

As far as corpora collection is concerned, the (ex-ante) selection of the data sources is always crucial as we need outstanding journals able to cover main topics and represent the temporal evolution of a specific knowledge field or discipline.

The *American Statistical Association* (ASA) represents the world's largest community of statisticians and the JASA has long been considered the world's premier review in its field, besides of being the oldest (and still available) statistical journal.

We are aware to assume that the "shape" of the keywords' trajectories in terms of occurrences in JASA reflects the relevance of the corresponding concepts and topics in the scientific discourse. We believe that JASA is a good source to represent Statistics as a whole and to mirror the history of Statistics.

Established in 1922 by the ASA, JASA has inherited a long tradition from two predecessors: *Publications of the ASA* (1888–1912) and *Quarterly Publications of the ASA* (1912–1921). Taking into account only the texts of titles including content words and disregarding items that not refer to research papers (e.g., *List of publications, News, Comment, Rejoinder*), the corpus includes 10,077 titles (out of 12,557) of articles published in the period 1888–2012 (125 years, from Volume No. 1, Issue No. 1 to Volume No. 107, Issue No. 500, since at the very beginning the volumes of the ASA's journals were biennial). The corpus is composed of 87,060 word-tokens and

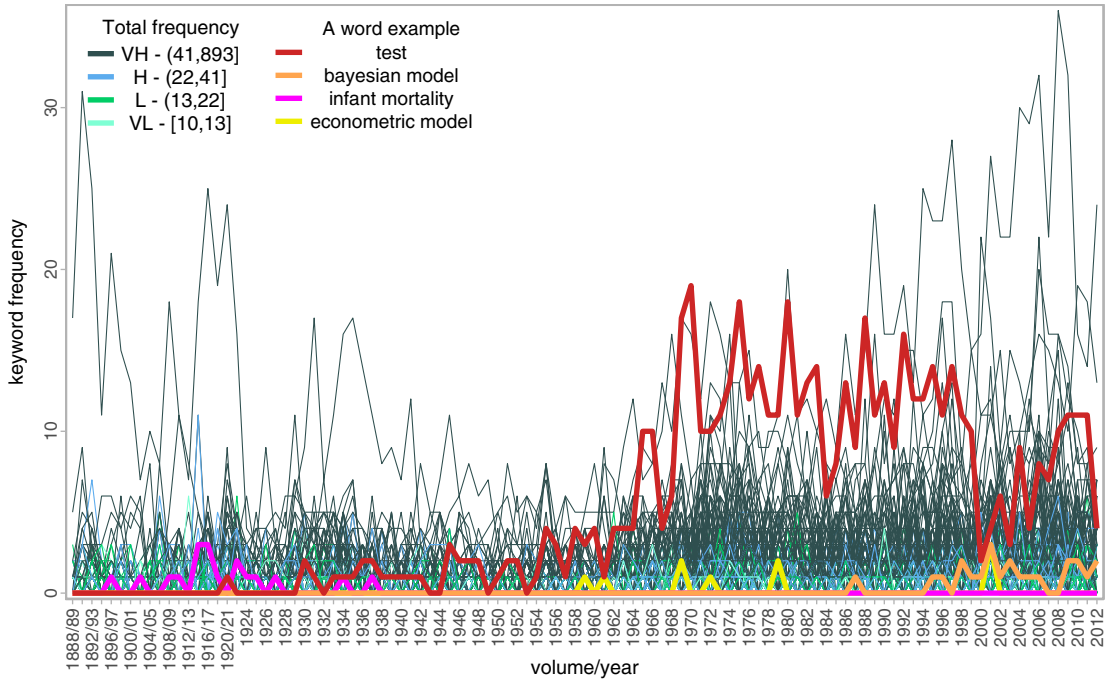

Fig. 1. Keyword trajectories (original data): *y*-axis represents the keyword frequency for each volume; line color identifies the keyword frequency class (Very Low, Low, High and Very High denote equal-frequency intervals of keyword total frequency in the entire corpus). An example of word trajectory has been superimposed for each frequency class.

7,746 word-types. To solve the problem of identifying a set of keywords that prove relevant for the study of the history of Statistics, we adopt a stepwise procedure:

1. to overcome some of the limitations of analyses based on simple word-types, we replace words with stems by means of the popular Porter's stemming algorithm [41,42];
2. to take into account compounds, multi-words and sequences of words which have different meanings when they are considered in their context of use and together with adjacent words, we identify n-stem-grams (sequences of stems occurring at least twice and composed of a minimum of two and a maximum of six consecutive stems);
3. to identify the most relevant statistical keywords, we match the vocabulary with popular statistics glossaries available on-line (over 12,700 unique entries obtained from merging six sources): ISI-International Statistical Institute; OECD-Organisation for Economic Cooperation and Development; Statistics.com-Institute for Statistics Education; StatSoft Inc.; University of California, Berkeley; University of Glasgow;
4. to reduce low frequency keywords we adopt a threshold and select keywords with frequencies equal to or higher than 10.

The final contingency table includes the frequencies of 900 keywords over 107 time-points.

We list some notes relative to the above steps or as general remarks. With regard to step 1. we exploited Porter stemmer as a preprocessing phase. In a text the nature of some lexical choices is contingent, e.g., verb tenses and the plural forms of nouns. Then, we carry out step 1. (e.g., word-types "model", "models", "modeling", "modelling" are replaced with the same stem "model"). Porter's algorithm is a popular stemmer in information retrieval applications, an alternative is Lancaster stemmer but it seems more aggressive and often leads to results that are not as readable as they should be. Related to step 2. is the problem of changing concepts in scientific literature, i.e., that the same words denote quite different concepts/meanings in different periods. On this regard, we note that in any text we have sequences of words that have different meanings if they are considered alongside adjacent words (KeyWord In Context or KWIC perspective). Observation of the frequency of multi-words (keywords that include two or more words) increases the amount of information conveyed by words because a sequence of words reduces "noise" and disambiguate the meaning. Semantic changes and semantic shifts of a word over time should envisage also the arrival/appearance of new collocations and compounds and, when these new "objects" become relevant in a language (and in a scientific language), their frequency increases (and are taken into consideration for our analyses). A word that is polisemic in potential can reduce its ambiguity by means of the context of use (KWIC) that is mirrored by a sequence of words. When possible, i.e. when the frequency was higher than the threshold, we included and preferred multiwords rather than words. Lastly, related to both step 3. and 4., we are aware that there are many ways to identify "relevant" words in a text. Nevertheless, we are working with titles of scientific papers that are extremely short, thick and concise. They include technical words, nouns (e.g. research objects), names (e.g. authors), keywords and nothing else. When we achieve our vocabulary, we select nouns, names, and multiwords with frequency higher than 10. After matching with statistical glossaries no relevant word should get lost.

As a final remark, we choose to work with titles as they provided us with a series of 125 years going back to 1888, whereas longer texts, such as abstracts (needed e.g. for LDA, see Section 1.2) would have resulted in a considerable loss of information (our archival analysis shows that abstracts did not appear until the 1930s–there is one in 1933, were sporadic in the 1940s and 1950s, and became increasingly regular and systematic after the 1960s [cfr. 4]).

3. Method: A functional clustering two-stage approach

From a FDA perspective, discrete observations $\mathbf{y}_i = \{y_{ij}\}$ of the frequency of a keyword $i(= 1, \ldots, N)$ in the volumes $j = 1, \ldots, T$ are viewed as a realization of an underlying continuous function

$x_i(t)$—sufficiently smooth or regular—representing the word temporal development. As $\mathbf{y}_i$ is a noisy observation of the underlying $x_i(t)$, an adequate model of their relationship is $\mathbf{y}_i = x_i(\mathbf{t}) + \boldsymbol{\epsilon}_i$, where $\mathbf{t} = \{t_j\}$ is the finite set of time-points associated with volumes' publication and $\boldsymbol{\epsilon}_i = \{\epsilon_{ij}\}$ is a zero mean vector with dispersion matrix $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \Sigma_\epsilon$. In the standard model, the ϵ_{ij}s, often termed "measurement errors", are independent across j and homoscedastic with $\mathrm{Var}(\epsilon_{ij}) \equiv \sigma^2$, but, in a more general case, Σ_ϵ can be regarded as full and time dependent.

In our application, $N = 900$, $T = 107$ and time-points t_js correspond to years of volumes' publication in the time span 1888–2012, generally occurring yearly except for the earliest period (up to 1923) when the volumes were biennal.

3.1. B-spline smoothing

For representing FD as smooth functions one method is the basis function approach where $x_i(t)$ is represented by a finite-dimensional linear combination

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

for sufficiently large K, of real-valued functions ϕ_k called basis functions.

The most common basis systems are monomial, Fourier and B-spline bases. Other useful systems are, e.g., wavelets, exponential and power bases. In this study we consider B-splines as they consist in a very flexible basis for non-periodic FD. Technically, B-splines are a particular basis for building spline functions which consist of piecewise polynomials joined together "smoothly" at the interior nodes.

More in detail, a polynomial spline of order m consists of piecewise polynomials of order m (degree $m-1$) defined on a sequence of subintervals partitioning the domain $[t_1, t_T]$, with smoothness condition on each interior breakpoint or *knot* defined by $m-1$ constraints that correspond to continuity conditions for $x(t)$ and its $m-2$ derivatives.

As regards the positioning of breakpoints, a direct and reasonable choice is placing knots at each time-point t_j. It entails that dimension K in (1) equals $T + m - 2$ (that is, a saturated or supersaturated basis as long as $m \geq 2$), hence basis expansion $x_i(t)$ is an exact interpolator of data. Clearly we seek a model that offers an image of the data having fewer degrees of freedom than in the original data.

We adopt the *roughness penalty* or *regularization* approach for estimation under which the estimate $\hat{x}_i$ is that function that minimizes the penalized residual sum of squares, $\mathrm{PENSSE}(x_i) = \mathrm{SSE}(x_i) + \lambda \cdot \mathrm{PEN}_r(x_i)$, where $\mathrm{SSE}(x_i) = \{\mathbf{y}_i - x_i(\mathbf{t})\}^T W \{\mathbf{y}_i - x_i(\mathbf{t})\}$ (W being the reciprocal of Σ_ϵ) is the residual sum of squares, $\mathrm{PEN}_r(x_i)$ the penalty term and λ a smoothing parameter. In particular, the penalty term measures a function roughness by the integrated squared r-th derivative over the observation time, $\mathrm{PEN}_r(x) = \int [D^r x(s)]^2 ds$. Thus, the parameter λ measures the rate of exchange between fit to the data, as quantified by the residual sum of squares, and roughness of the function x. In other words, λ regulates the bias-variance tradeoff: as $\lambda \to 0$ the curve x approaches an interpolant to the data, satisfying $x(t_j) = y_j$ for all j; as $\lambda \to \infty$ the condition $\mathrm{PEN}_r(x) = 0$ implies the fitted curve x to be a spline of order r. Choosing λ is part of the model selection issue.

A standard practice for choosing λ is to use cross-validation (CV) which is a resampling method for assessing a model generalization performance. The basic idea is holding out a set of data (the validation set) from the fitting process, and then seeing how well the model fitted on the remainder of the data (the training set) predicts those held out observations. A widely used approach is K-fold CV in which data are partitioned into K equally (or nearly equally) sized folds and then K iterations of model training (on $K-1$ combined folds) and validation (on the held-out fold) are performed. Model performance is typically assessed using mean squared error in the case of a quantitative outcome and K-fold CV returns the average of K such errors as the overall performance metric for the model. When tuning a smoothing parameter a common choice is $K = 1$ or the leave one-out CV, however, it may be computationally intensive expecially for large sample sizes and lead to under-smoothing. Generalized cross validation (GCV), $\mathrm{GCV}(\lambda) = \frac{T}{(T - df(\lambda))^2}\mathrm{SSE}(\hat{x}_i)$, 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 regularization, which is monotone decreasing in λ 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 [43,44].

We smooth the data by trying different spline orders (m from 1 to 8), combined with various roughness penalties, and varying the smoothing parameter over an opportune range of values ($\log_{10}\lambda$ from -6 to 9). Different penalty forms were tried for each spline order, in particular we set the derivative order of PEN_r to be $r = m-2, 2, 1, 0$, i.e. let the function to be increasingly bumpy (notice that for $m = 3$ the highest possible order is $r = 1$, for $m \leq 2$ is $r = 0$ that is the function itself).

3.2. Distance-based curve clustering

Classical clustering concepts for vector-valued multivariate data can typically be extended to FD, where, however, various additional considerations arise, such as dimension reduction of the infinite-dimensional FD objects and discrete approximations of distance measures [36,37].

In this study we set up a distance-based approach, in particular the k-means algorithm is used combined with the L_2 metric to measure distance between curves. This is estimated on the evaluation points of the approximating curves previously obtained from the B-spline smoothing of FD. The evaluation points are the discretely observed points common to all FD.

Besides the L_2 metric (Euclidean distance) other measures of proximity can be considered, such as the L_1 metric (Manhattan distance), the adaptive dissimilarity index (with either euclidean distance or Dynamic Time Warping, DTW), and the correlation-based dissimilarity. They are just a possible set of options, suitable to the clustering objectives under study, which are taken from the broad range of dissimilarity measures set out to perform clustering of time series [see 45, and references therein]. A guideline for choosing a dissimilarity is what are the specific distortions of the data to which it must be invariant [an interesting review is in 46]. In practice the option of a proper distance and that of a preprocessing to remove distortions from data can be logically equivalent. For example, DTW can be seen as a more robust distance measure, or it can be seen as using the Euclidean distance after removing warping from data. Still, the correlation-based distance is invariant to curve amplitude, but it would be equivalent to using the z-score row-normalization (r_2 in Table A.2) and then apply L_2 norm.

Here the objective is to compare curve profiles once a transformation has been performed in accordance with the clustering purpose (Section 4), then conventional distances between raw data (Euclidean or Manhattan, among others) evaluating a one-to-one mapping of each pair of point sequences can produce satisfactory results.

Cluster validation, that is assessing the quality of a clustering on a dataset, is an essential step in the cluster analysis process. Beyond being of interest in its own right, quality assessment of a single clustering is the basis for comparing different clusterings, particularly with different numbers of clusters (when clustering

method and distance measure have been chosen), to help decide what the most appropriate number is in a certain application [47].

Within the approaches to cluster validation [47], the use of external information is a valuable and ultimately necessary tool. Here, external information consists of an informal assessment of subject matter experts who can decide to what extent a clustering is meaningful to them. Note that expert opinion should be examined in turn: it may provide useful information to improve the clustering methodology, but it may also reflect expert's preconceptions. On the other side, a large number of indexes has been proposed in the literature for a validation based on the clustered data alone. It is well known that "one index does not fit it all", rather, the many existing indexes can be grouped into different types each measuring a different aspect of clustering quality. Given this, and given the exploratory and evocative task of our clustering, we have gathered a large basket of indexes for not favoring any criterion, each in principle equally valid. These include measures of within-cluster homogeneity, e.g., *Ball-Hall, Banfeld-Raftery, C-index, Marriot* (*Ksq_DetW*), *Scott-Symons, Trace_W*; of between-cluster separation, e.g., *Det_Ratio, Scott* (*Log_Det_Ratio*), *Ratkowsky-Lance*; and of their combination, e.g., *Calinski-Harabasz, Hartigan* (*Log_SS_Ratio*), *Dunn* and its generalizations, *Davies-Bouldin, Ray-Turi, Xie-Beni, S_Dbw validity, Friedman, SD, Silhouette, Tau*; besides of measures of similarity between the empirical within-cluster distribution and distributional shapes such as the Gaussian distribution, e.g., *BIC, AIC* and their variants [see 48,49, and references therein]. In this study, the idea is that of pooling the ratings from a large number of internal validation indexes, without integrating subject matter knowledge in a first instance, so as to let the data bring out a ranking of the best candidates to cluster number rather than guide to a unique solution. Our clustering procedure is thought of as a tool of thorough investigation before submitting the results to experts who possibly will guide towards other analyses (see details of the procedure in Section 5).

As a major conclusion, we point out that the analyst/user is asked to choose from a number of options at every step of a cluster analysis, here particularly when faced with the decisions on data processing , on dissimilarity definition, and of the number of clusters. If we start from the assumption that no "natural" or "true" clusters exist in the available data, it is the definition of clustering aim, hence of cluster concept and corresponding appropriate clustering methodology, that will produce a grouping structure, either "real" (that is, a meaningful structure in the observer-independent reality) or "constructive" (a split-up for pragmatic use) or "useful" (context-dependent). See [40,47,50] for a thorough discussion on these different situations of clustering.

4. Theory: Corpus data transformation

The decision about what data to use is an important part of the clustering process, and often has a fundamental impact on the resulting clusters. In this study we examine how different data representations affect clusters' generation.

If we consider the keywords × time-points table by row, a typical feature of a word trajectory is a sharp peak-and-valley trend, mainly due to the sparsity affecting frequency data of a corpus (Fig. 1). On the other hand, if we look at data by column they appear strongly asymmetrical, in particular for the marked disparity of frequency classes between the most popular words and all of the others (in Fig. 1 color is increasingly darker with the keyword frequency class). This is a typical feature of word-type frequency distributions, also known as *large number rare events* (LNRE) property, consisting in a large number of word-types occurring very rarely. A result is the aforementioned sparsity, i.e., many cells of the contingency table have small counts or are empty. Lastly, the size of subcorpora (number of documents and their size in word-tokens) may vary greatly over time (Fig. 2). Again, this reinforces data sparsity.

In our research, we envisage several transformations which, generally speaking, address two different objectives: whether, in assessing two curves as similar, we should consider height (word popularity) and timing (synchrony) jointly, or timing only. In the first case, we just need to adjust the uneven document dimension across time, then normalize data by column somehow. In the other case, in order to effectively gather the synchrony of word histories, we need to normalize data by row, or better still, since a sort of column-normalization should be regarded as preliminary (to remedy the signal irregularity over time), to resort to some double normalization.

The normalization step (Table A.2) of our procedure provides several transformations: by column, obtained from dividing each keyword frequency on a time-point, n_{ij}, by the total number of documents (c_1) or the total number of word-tokens (c_2) in the subcorpus referring to the time-point, or else the column sum (c_3) or the column maximum frequency (c_4) of the lexical table at the time-point, lastly, calculating the cumulated frequency (c_5) at the time-point weighted by a time dimension; by row, obtained from dividing each n_{ij} by the row sum (total frequency) for the keyword (r_1), calculating the z-score of the keyword frequencies (r_2), dividing by the row maximum frequency for the keyword (r_3), calculating a nonlinear transformation of the keyword frequencies (r_4, r_{4b}), finally, calculating the relative (to the row sum) difference of consecutive keyword frequencies (r_5). In our study, documents are the titles (or articles) and a subcorpus is a volume.

Within column normalizations, c_1, c_2 and c_3 options are practically equivalent in order to adjust the uneven document dimension across time (Fig. 2). Whilst c_4 seems to be the least appropriate for being weakly correlated with document length over time (for construction, it retraces the time profile of the most popular keywords). Option c_5 has been conceived for data featured by LNRE and consists, for each keyword, of a dynamic probability of occurring converging to the overall relative frequency in the entire corpus.

Row normalizations from r_1 to r_4 are increasingly effective at reducing the asymmetry of frequency spectrum. We note that r_1 produces a sort of "reversed" asymmetry, i.e., low frequency words tend to dominate (higher level curves) because of their greater sparsity. Methods r_4 and r_{4b} consist of a more general nonlinear transformation (of which the standard normalization, r_3, is a special case) aimed at sterilizing asymmetry due to different variation ranges of keyword frequencies (here we refer to eq. 12 in [51] with smoothing coefficient $p_{x(1)}$ and $p_{x(2)}$ of eq. 15–16). Method r_5 has been designed with the different perspective of describing relative change rather than the absolute one.

Crossing a column- by a row-normalization generates a double normalization. Our comprehensive study examines all the transformations specifically indexed in the Table A.2. Here we present a small subset to illustrate the procedure and highlight some of the results arising from the decision on transformation.

4.1. Calculation: Normalizations at comparison

In this work we compare three different types of transformations, one by time dimension and two double normalizations. From the foregoing (Section 4), we choose c_2 for a preliminary column normalization as this is the most customary way to transform raw data into relative frequencies. After that, we consider d_1 which is equivalent to calculate a χ^2 distance between word profiles if we use L_2 as measure of dissimilarity. This choice is in homage to the correspondence analysis, one of the mostly used multivariate methods in the linguistic field. Finally, d_3 is chosen for it, although simple, substantially reduces the problem of asymmetry

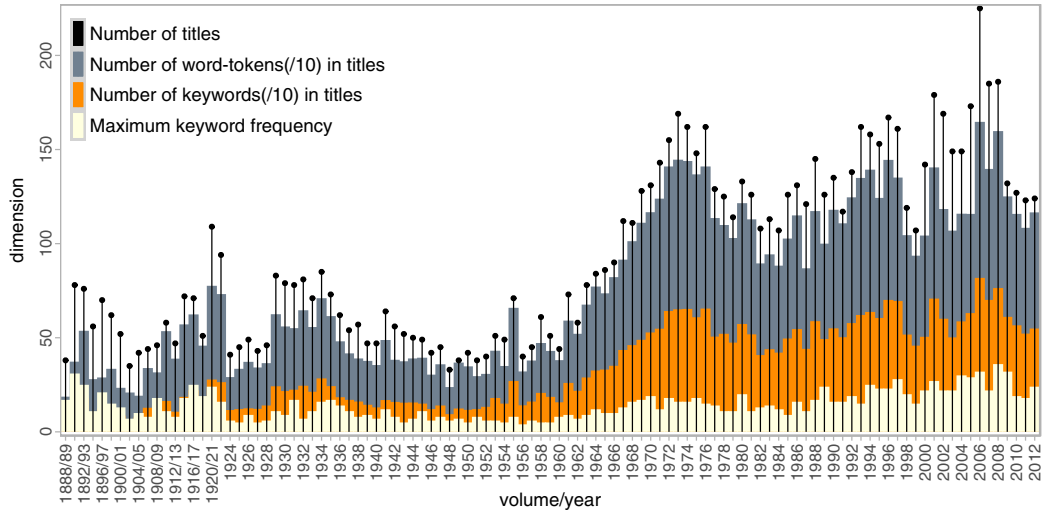

Fig. 2. Subcorpora dimension: for each volume, number of titles/articles, total number of word-tokens in titles/10, total number of keywords in titles/10, maximum keyword frequency.

allowing to compare curves in terms of horizontal or phase variation. The computing is as follows: c_2) $y_{ij} = n_{ij}/N_j (\times 1000)$, where N_j is the total number of word-tokens in the subcorpus j; d_1) $y_{ij} = n_{ij}/(n_{i.}\sqrt{n_{.j}/n})$, where $n_{i.}$ is the row-i sum, $n_{.j}$ the column-j sum and n the matrix total ($n_{.j}/n$ is the column-j mass in correspondence analysis); d_3) $y_{ij} = n_{ij}/(M_i N_j)(\times 1000)$, where M_i is the row-i maximum frequency.

5. Results

With the roughness penalty approach to estimation, the smoothing selection is carried out by varying roughness penalty form and, over opportune value ranges, spline order m and smoothing parameter λ (Section 3.1). According to the GCV criterion, optimal smoothing for c_2 normalized data is achieved with $m = 5$ and $\lambda = 10^3$ ($df = 7.7$) after setting a PEN_2 roughness penalty, whereas for both d_1 and d_3 normalized data the criterion lead to $m = 3$ and $\lambda = 10^{1.75}$ ($df = 7.375$) under a PEN_1 roughness penalty.

Curves are then partitioned by the k-means algorithm on the basis of the Euclidean distance between the trajectories evaluated on the observed sequence of time-points. As argued above (Section 3.2), in the current study the L_2 distance is a proper choice within the set of possible alternatives. R software environment [52] contains several k-means implementations. Our procedure uses the `kml` routine [49] which is designed specifically for longitudinal data and which provides, inter alia, various efficient methods of k means initialization. In the present illustration, the algorithm is re-run, for each k from 2 to 26, 20 times from different initial configurations set through the k-means++ seeding method [53].

A set of 49 quality criteria are then computed in order to identify the best partition/number(s) of clusters (see Section 3.2).

Visual representation of the rating for the cluster number can help in the analysis of results. In particular, we computed a ranking of cluster number for each quality index, then pooled all the rankings, and calculated for each cluster number the frequency of being ranked first, second, third and fourth (Fig. 3). First we note that in general partitions into two/three clusters are the best rated. This reflects the substantial bifurcation of the historical period around the sixties, when the birth of Statistics as an autonomous and established discipline can be placed [4]. Indeed, the contour plot of the correlation function, $\rho_x(s,t) = \sigma_x(s,t)/(\sigma_x(s,s), \sigma_x(t,t))$ where $\sigma_x(s,t) = \Sigma_{i=1}^{N}(\hat{x}_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))/N$ with $\bar{x}(t) = \Sigma_{i=1}^{N} x_i(t)/N$, on smoothed data (Fig. 4) shows, transversely to the transformations, a narrowing of contours around the '60s suggesting a caesura in the historical period of reference (in general, a more rapid fall-off of correlation is observed concurrently with rather erratic uncoupled curves). Moreover, we can notice several time windows of slower decline of correlation (typical of periods in which curves are tightly connected and form bundles related temporally), four periods are more evident in all three panels: 1888/89-1918/19, '20/21-'48, at the turn of '50/'60s-early '80s, after 2000. Moreover, partitions with a number of clusters close to the maximum of the considered range (25, 26) have also been frequently selected. This result, on one hand, may reflect the lack of a defined structure and parsimonious grouping, but, on the other, it may be a failure due to the standard assumption—underlying by many quality criteria as well as by k-means algorithm itself and typically by model-based clustering—of data normally distributed hence of clusters being compact and convex. From the foregoing, our extensive internal evaluation suggests that, once discarded the solutions picking the extremes, the most selected cluster numbers (considering the ranking obtained from the cumulated frequency of being in the first four positions—given by the bar height in Fig. 3, but balanced with position importance—detected from the bar color composition) are: 5 for c_2 transformation (second best: 4, third: 9/22/6); 6 for d_1 (second best: 19/4, third: 24/21/22) and 5 for d_3 (second best: 22/24, third: 20/23). These rankings are the output of an R code that essentially mimics a qualitative rating purely based on a graphical inspection. To compare some aspects of how the three transformations affect clustering, we consider the best partition found with the cluster number ranked first, i.e., 5 for both c_2 and d_3, 6 for d_1 (Figs. 5–7).

When data are solely column normalized, word "popularity" plays a dominant role: clusters are primarily determined by high-level curves (high frequency words) leaving the majority of low frequency words in one or more fuzzy groups. In the instance of five-group clustering on c_2 normalized data (Fig. 5), three clusters—which account for only about 10% of the total words—look interesting, the rest being a singleton (*statistics*, the most frequent word making group E) and an indistinct agglomerate of low frequency words (the massive group A). Conversely, a double normalization allows for a more balanced partitioning where the shape and level of curves play a role on a par. In both the clusterings into six and

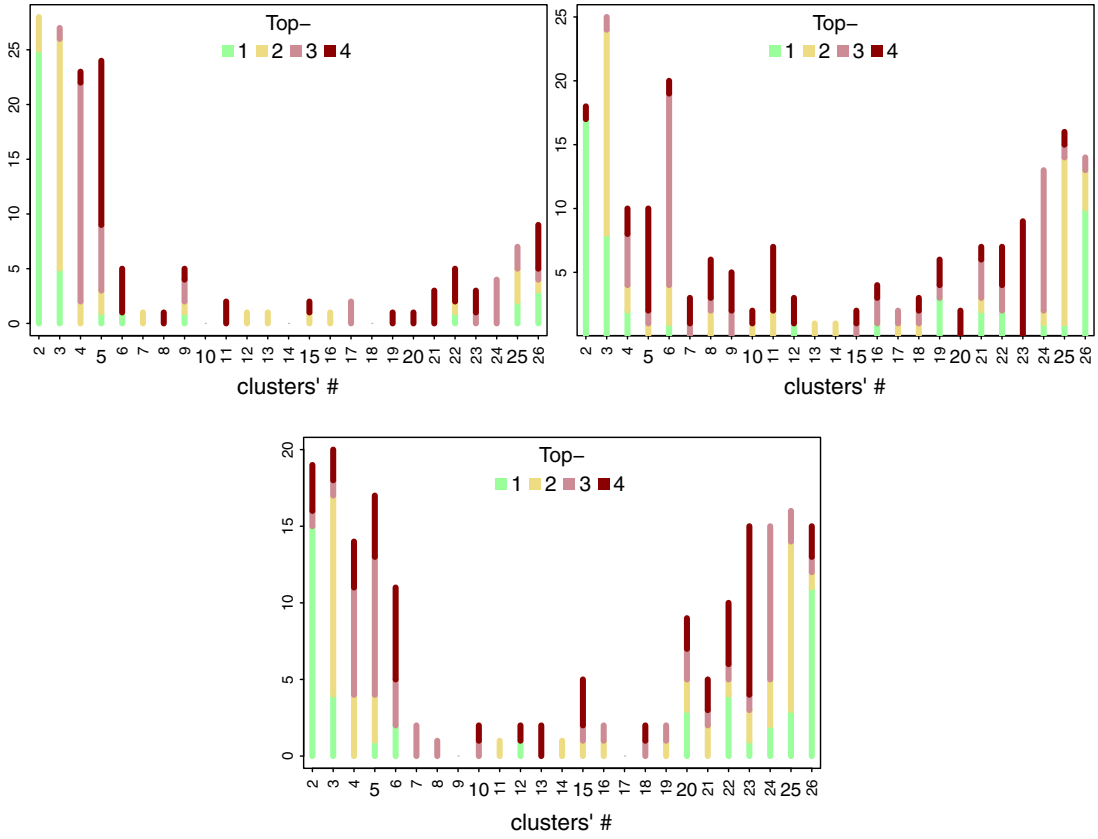

Fig. 3. Frequency of being ranked first (top-1), second (top-2), third (top-3) and fourth (top-4) for each cluster number by pooling rankings from the overall quality criteria: c_2 (top-left), d_1 (top-right), and d_3 (bottom) normalizations.

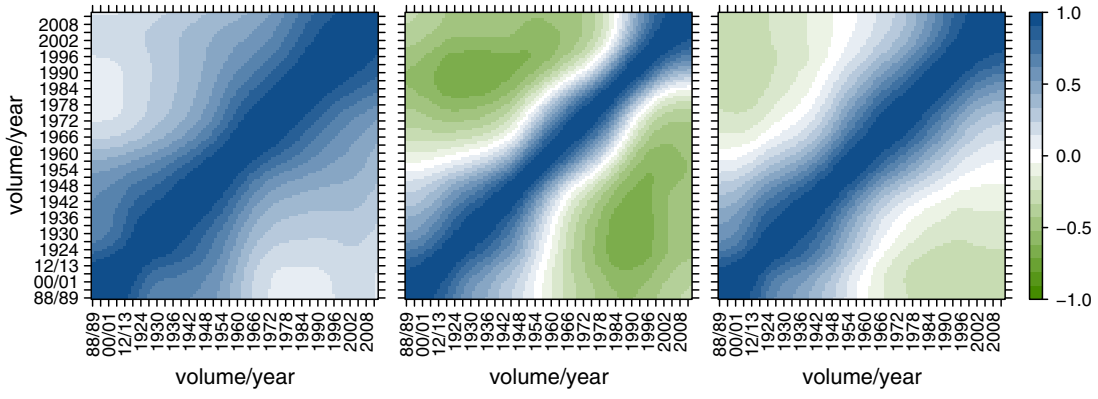

Fig. 4. Temporal correlation of word curves: c_2 (left), d_1 (center), and d_3 (right) normalizations.

five groups for, respectively, d_1 and d_3 normalized data (Figs. 6 and 7), some patterns which have already appeared for c_2 transformation are now confirmed and better structured by far more numerous groups: the long time span is clearly divided by clusters of words sharing similar life cycles. For instance, C group for c_2, representing the aforementioned period prior to the birth of Statistics as an autonomous discipline, is temporally structured in F (demography, population studies and official statistics) and C (economic and public statistics) groups for d_1 and is embodied in B group for d_3; B group for c_2 is temporally divided in A (statistics is born and states with its own lexicon), D (the "golden age" of classic statistics, 60s to 80s), B (contemporary statistics) and E (statistics of the new millenium) groups for d_1. Note that clustering from d_3, while presenting strong similarities to that from d_1, appears to split the time span in more blurring periods by creating transversal groups. A possible interpretation is that, while clustering with d_1 is mainly determined by the alternating moments of a word vitality or absence along time, with d_3 the peculiarities of a word life, that is, the pure form of its trajectory, primarily makes up the groups. Finally, D group for c_2 containing the most frequent "basic" words of statistics—that are distributed in various clusters for d_1—is enhanced by a larger set of keywords (E group) for d_3.

Let us now examine some aspects of clustering, in the three cases of normalization, by varying the number of clusters (Table 1): how much groups are balanced (or the cluster sizes are uniform); how many groups are singletons; how much groups are hetero-

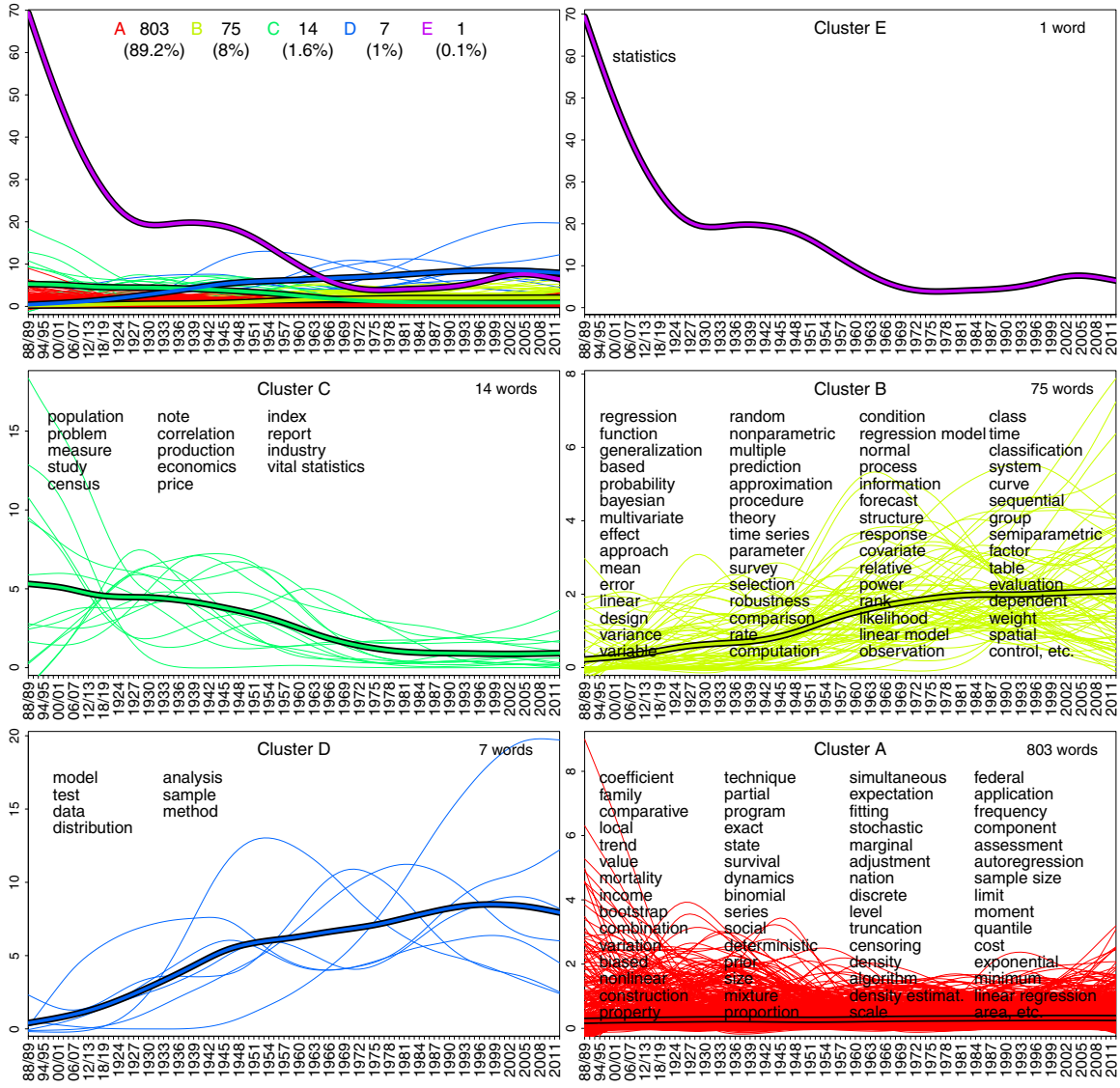

Fig. 5. Clustering on c_2 column-normalized data: all five groups and individual clusters. Stems have been replaced with the singular noun or, in case this is not present in the corpus, with the typical word related to the stem.

Table 1
Balance, presence of singletons and heterogeneity of frequency classes.

		cluster#											
	normalization	2	3	4	5	6	7	8	9	10	15	20	25
Quality of balancing	c_2	.00	.12	.26	.29	.44	.49	.56	.59	.63	.80	.86	.91
	d_1	.72	.93	.90	.90	.94	.96	.95	.97	.97	.98	.98	.99
	d_3	.84	.88	.92	.93	.93	.95	.95	.96	.97	.97	.98	.99
Number of singletons	c_2	1	1	1	2	2	3	3	3	3	7	10	11
	d_1	0	0	0	0	0	0	0	0	0	1	1	1
	d_3	0	0	0	0	1	0	0	1	0	3	5	5
Heterogeneity of frequency classes	c_2	1	.50	.06	.09	.02	.02	.05	.09	.09	.11	.05	.12
	d_1	1	1	1	.99	.99	.99	.98	.98	.97	.96	.95	.94
	d_3	.90	.95	.95	.93	.81	.85	.80	.82	.80	.80	.78	.77

geneous in being composed of words of different frequency class or popularity. Both the balance and frequency class heterogeneity metrics are measured by the normalized Gini index (considering the median of values calculated on the 20 replications for each cluster number); the number of singletons is the maximum found in the 20 replications. For c_2 case, it is evident a severe imbalance in cluster size together with a high presence of singletons (which partly creates the first). Moreover, a dominance of one/a few frequency classes within groups is highlighted and we already know that, generally, the sole class of very high frequency words constitutes the dominant class. (Note also that the perfect uniformity, 1, when the cluster number is 2 is due to the impossibility of cal-

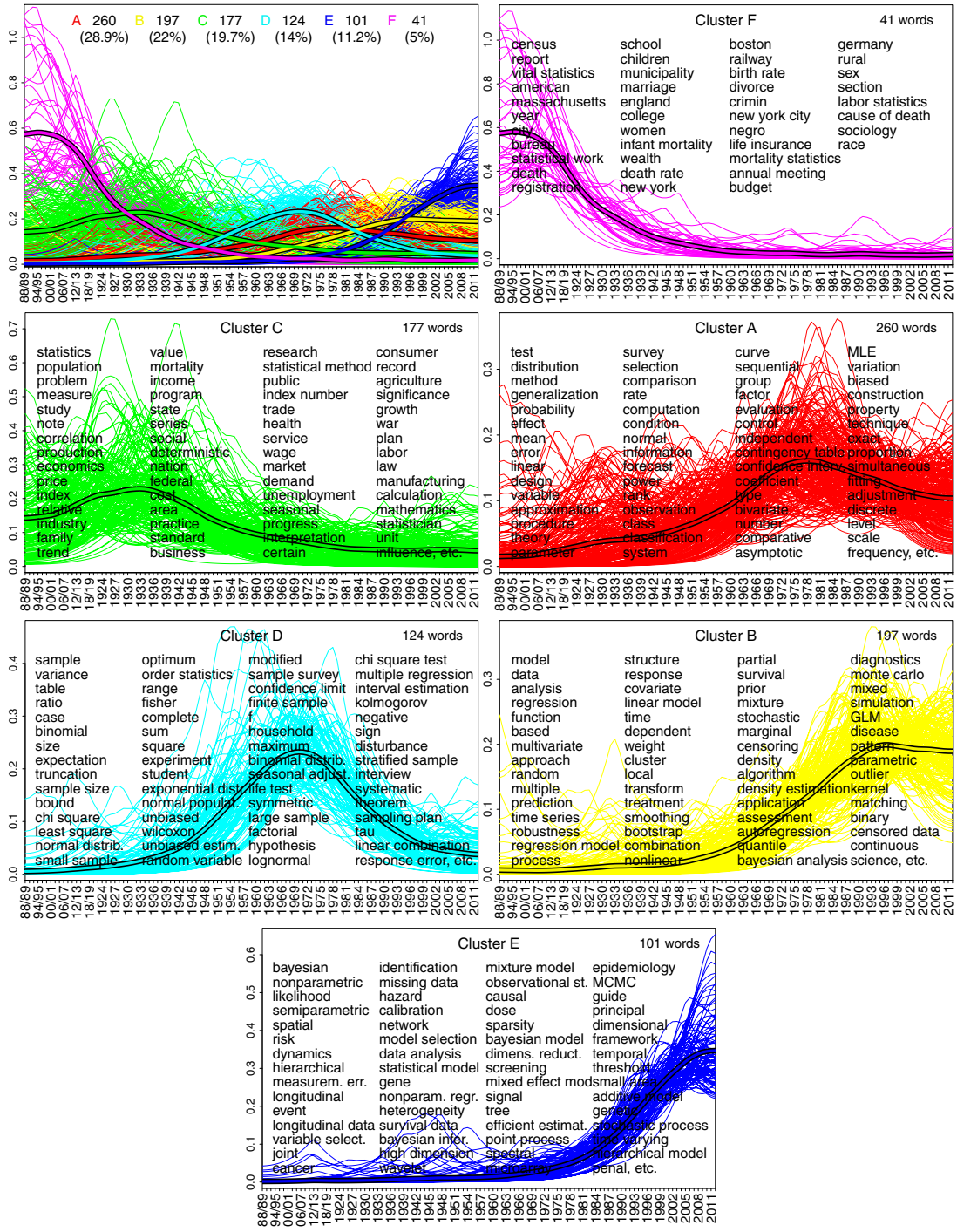

Fig. 6. Clustering on d_1 doubly normalized data: all six groups and individual clusters.

culating the Gini index on singleton clusters.) Conversely, for both double normalizations, groups appear well balanced in cluster size, very rarely of singleton type, and uniformly composed of words from all the four frequency classes.

5.1. A possible history reading

A narrative of the history of Statistics can be learned from the inner dynamics of its methods, topics and research areas as reconstructed by our clustering procedure. The normalization that best enables to separate the entire historical period into subsequent temporal phases is d_1 (see point 3 in Section 6). Thus, we consider the clustering obtained from data transformed according to this normalization, giving some reference to the results produced with the alternative transformations, better suited to discover other aspects of temporal evolution, as we summarize in Section 6. It is worth remembering that this paper was not intended to develop a thorough and in-depth reconstruction of the history of Statistics (as in [4]), then, here, we offer a synthetic draft, moreover limited by the cluster number previously selected for the primary aim of comparing results from different nornalizations (that is, nor too low neither too large and of about the same size).

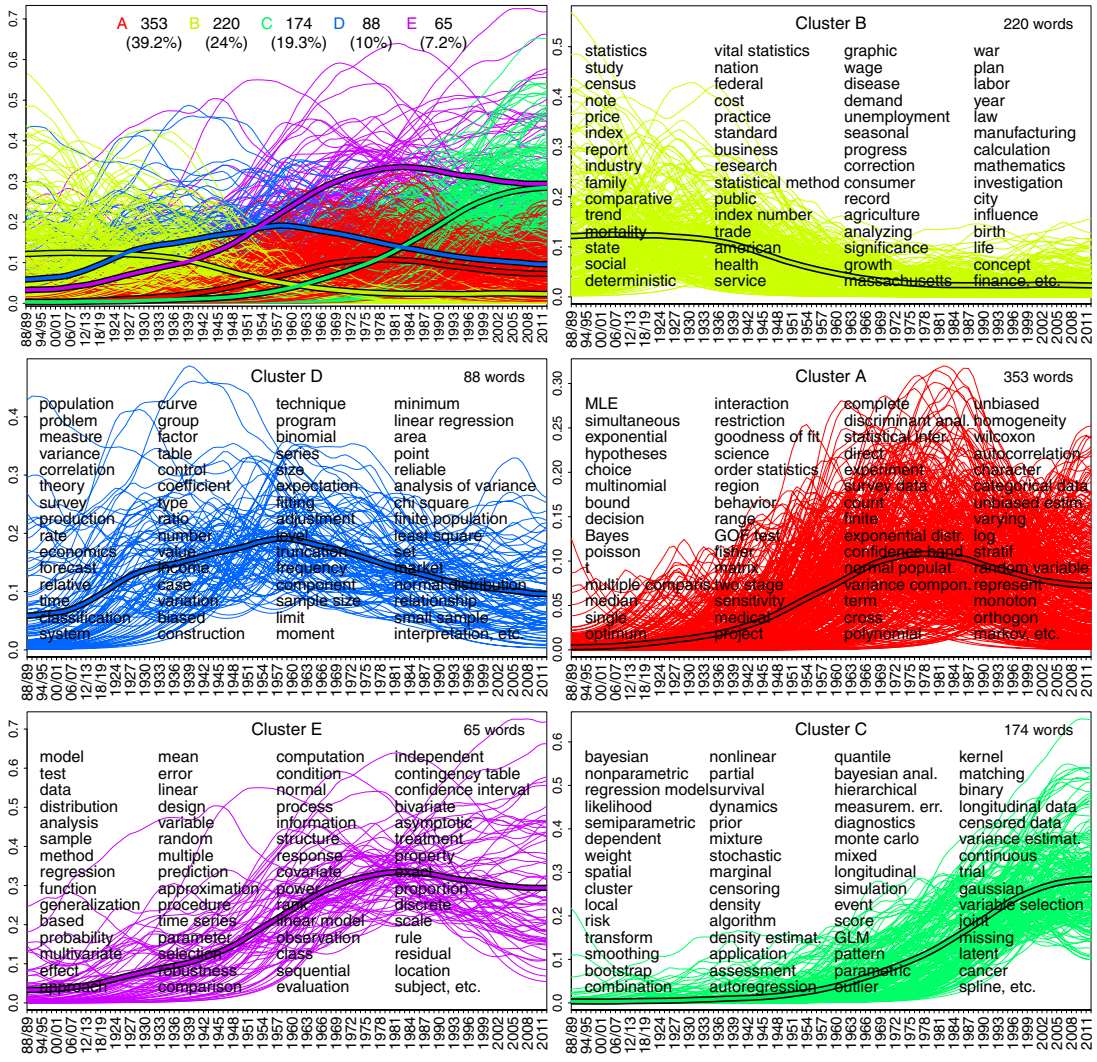

Fig. 7. Clustering on d_3 doubly normalized data: all five groups and individual clusters.

The birth of Statistics as an autonomous discipline should be placed at the turn of '50/'60s. Before this date, Statistics was primarily ancillary to other subjects:

- up to about 1920 (*Ancient History*), it was mainly social statistics (*census, registration, school, children, marriage, women, divorce, crimin, negro, race*) and demography (*vital statistics, infant mortality, death rate, birth rate, life insurance, mortality statistics, cause of death*), being an instrument of government (*bureau, municipality, railway, budget*) (see clusters F for d_1, Fig. 6, partly B for d_3, Fig. 7, and C for c_2, Fig. 5);
- then, during 1920–1950 (*Middle Ages*) it was public statistics (*population, state, social, nation, federal, public, service, war, law*) and economic statistics (*production, economics, price, industry, trend, income, cost, business, trade, health, wage, market, demand, unemployment, progress, consumer, manifacturing*) and developed first rudimental mathematical instruments (*measure, correlation, statistical method, index number, significance*) (clusters C for d_1, partly B/D for d_3 and C for c_2).
- From late '50s, Statistics became established as a separate discipline, affirming its own lexicon and basic tools (*distribution, probability, mean, error, linear, design, variable, parameter, survey, independent, contingency table, test, confidence interval, asymptotic, MLE, linear regression*) (clusters A for d_1, partly E/A for d_3 and B for c_2).
- Late '50s-early '80s (*Modern History*) represents the golden age of "classical" Statistics (*variance, expectation, sample, sample size, chi square, least square, normal distribution, binomial distribution, exponential distribution, small sample, large sample, order statistics, fisher, experiment, factorial, student, unbiased estimation, sample survey, finite sample, stratified sample, sampling plan, wilcoxon, f, chi square test, kolmogorov, sign test, multiple regression, response error*) (clusters D for d_1, partly A for d_3 and D/B for c_2).
- Since late '50s, with a plateau in 2000 followed by a slow decline, methods for data analysis are refined more and more (*model, data, analysis, function, algorithm, multivariate, multiple, time series, robustness, regression model, process, linear model, GLM, cluster, pattern, local, smoothing, nonlinear, survival, mixture, censoring, density estimation, autoregression*) and technological revolutions stimulated new modes of statistical computation, unthinkable before the arrival of modern computers (*bootstrap, monte carlo, simulation, bayesian analysis*) (clusters B for d_1, partly E/C for d_3 and B for c_2).
- Since late '90s and still on the rise (*Contemporary History*), approaches (*bayesian, likelihood, nonparametric, semiparametric*), classical issues (*variable selection, model selection, missing data, calibration*), modeling classes (*mixture model, mixed effect model, additive model, hierarchical model, longitudinal data, survival*

data, causal, dose, event, spectral, tree, point process, wavelet) specialize and the new millenium heralds new topics and new challenges such as the dimensionality and complexity of information and the need for hybridization and interdisciplinarity (*spatial, risk, dynamics, network, genetic, epidemiology, small area, cancer, gene, microarray, high dimensional, sparsity, dimension reduction*) (clusters E for d_1, partly C for d_3 and B for c_2).

6. Conclusions

In this study, we have showed how the history of a discipline can be traced from analyzing the temporal evolution of relevant keywords retrieved from articles of mainstream scientific journals in the field. The history is reconstructed on the basis of the most relevant inner dynamics as obtained from an opportunely set up clustering of words' life-cycles. The word's life-cycle is the primary, indivisible unit of our analysis (the functional datum, the curve-imprint in time). Conceptually, our approach differs from the main alternatives addressing the problem of knowledge evolution, like those developed for TDT, ETD and, generally, for dynamic knowledge mapping in scientometric studies. Focus of our analysis is the detection of important dynamics representing, each, the temporal evolution of a group of words. Thus, on principle, different topics, research areas, schools of thought can be represented within the same group of words. On the contrary, "topic-centered" methods cited above focus, first, on the detection of topics, then, on tracking their evolution. To help in choosing between these two perspectives, we just point to some practical consequences. In "topic-centered" methods, words that represent the same topic may have unconcilable temporal evolution; topic evolution can only be a "roadmapping", an abstract description (being the average evolution of words grouped by co-occurrence criterion) of basic movements over time; the abstract definition of topics is subjected to continuous destruction and reconstruction by time, making topic tracking a fragile and questionable artifact.

As regards the crucial choice of data normalization, the analysis of three examples of transformation of word frequencies has highlighted what influence each type can have on curve clustering results. In what follows we summary the main findings.

1. Normalization by column maintains the level of word popularity differentiated and produces a dominance of high frequency words on the clustering results. Significant imbalance in cluster size, large presence of singletons, lack of heterogeneity of frequency classes in group composition and, finally, the presence of one or more "amorphous" groups (including almost exclusively low frequency words) are some of the most evident effects of this type of transformation.
2. Conversely, the double normalization produces better balanced groups both in cluster size and frequency classes, rare singletons, and almost never amorphous groups, though, it does lose information on word popularity.
3. Type-d_1 normalization is better able to recognize any group of words having "sparse"trajectories, i.e., which have experienced birth and/or death over the period considered, while the d_3 variant, which more properly "normalizes" the frequency , builds the groups primarily looking at the curve shape, i.e., at if the "relative popularity" of a word has been constant over time or if has fluctuated (and how) during its life cycle.

Future work plan will address the study of other types of normalizations, especially some transformations dealing with the problem of zero excess due to the LNRE feature of textual data. Moreover, with regard to clustering, we intend to deepen the discussion either on a technical side, such as studying the effect of different types of distance that measure the similarity between trajectories, either on a more methodological level, in particular with reference to what line of thought should be adopted in the final choice of cluster number. Lastly, in parallel to the study of distance-based curve clustering, we will continue to review and propose model-based approaches, where the interweaving of data pre-processing and model assumptions becomes even more complex.

Acknowledgments

Funding This study was supported by the University of Padova, fund CPDA145940 "Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature" (P.I. Arjuna Tuzzi, 2014).

Appendix A. Overview of normalizations

Table A1
Normalization plan.

	normalized by column / normalized by row	(corpus logic) subcorpus #titles	(corpus logic) subcorpus #tokens	("table" logic) column sum ($\sqrt{\cdot}$)	("table" logic) column max. freq.	(LNRE) dynamic density	
Strong asymmetry	row sum	d	d	d_1 (χ^2)	d	d_{1b}	r_1
	z-score by row	d	d_2	d	d	d	r_2
	maximum row frequency	d	d_3	d	d	d	r_3
	nonlinear transformation: $p_{x(1)}$	d	d_4	d	d	d	r_4
	nonlinear transformation: $p_{x(2)}$	d	d_{4b}	d	d	d	r_{4b}
	relative difference	d	d_5	d	d	d	r_5
		c_1	c_2	c_3	c_4	c_5	

References

[1] M. Maggioni, T. Uberti, F. Gambarotto, Mapping the evolution of clusters: a meta-analysis, in: G.I. Ottaviano (Ed.), Working Papers Series, 2009 of Global Challenges Series, 74, Fondazione Eni Enrico Mattei, 2009, pp. 1–39.

[2] D. Chavalarias, J.-P. Cointet, Phylomemetic patterns in science evolution - the rise and fall of scientific fields, PLoS ONE 8 (2) (2013) e54847–1304, doi:10.1371/journal.pone.0054847.

[3] O. Popescu, C. Strapparava, Time corpora: epochs, opinions and changes, Knowl. Based Syst. 69 (2014) 3–13, doi:10.1016/j.knosys.2014.04.029.

[4] M. Trevisani, A. Tuzzi, A portrait of JASA: the history of statistics through analysis of keyword counts in an early scientific journal, Qual. Quantity 49 (3) (2015) 1287–1304, doi:10.1007/s11135-014-0050-7.

[5] Y. Zhang, H. Chen, J. Lu, G. Zhang, Detecting and predicting the topic change of knowledge-based systems: a topic-based bibliometric analysis from 1991 to 2016, Knowl. Based Syst. 133 (Supplement C) (2017) 255–268, doi:10.1016/j.knosys.2017.07.011.

[6] M. Trevisani, A. Tuzzi, 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, Cleup, Padova, 2012, pp. 1–4.

[7] M. Trevisani, A. Tuzzi, 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, Academic Mind, Belgrad, 2013, pp. 84–95.

[8] M. Trevisani, A. Tuzzi, Through the JASA's looking-glass, and what we found there, in: Proceedings of the 28th International Workshop on Statistical Modelling, 1, Istituto Poligrafico Europeo, Palermo, 2013, pp. 417–422.

[9] A. Tuzzi, R. Köhler, Tracing the history of words, in: A. Tuzzi, M. Benesová, J. Macutek (Eds.), Recent Contributions to Quantitative Linguistics, DeGruyter, 2015, pp. 203–214, doi:10.1515/9783110420296-017.

[10] L. Slaets, G. Claeskens, M. Hubert, Phase and amplitude-based clustering for functional data, Comput. Stat. Data Anal. 56 (7) (2012) 2360–2374, doi:10.1016/j.csda.2012.01.017.

[11] J.S. Marron, J.O. Ramsay, L.M. Sangalli, A. Srivastava, Functional data analysis of amplitude and phase variation, Stat. Sci. 30 (4) (2015) 468–484, doi:10.1214/15-STS524.

[12] R. Köhler, Laws of languages, in: P.C. Hogan (Ed.), The Cambridge Encyclopedia of the Language Science, Cambridge University Press, Cambridge, 2011, pp. 424–426.

[13] I.-I. Popescu, Word Frequency Studies, Mouton De Gruyter, Berlin, 2009.

[14] A. Pawlowski, M. Krajewski, M. Eder, Time Series Modelling in the Analysis of Homeric Verse, 97, Eos, 2010, pp. 79–100.

[15] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (Suppl. 1) (2004) 5228–5235.

[16] L. Hong, B.D. Davison, Empirical study of topic modeling in Twitter, in: Proceedings of the First Workshop on Social Media Analytics, SOMA '10, ACM, New York, NY, USA, 2010, pp. 80–88, doi:10.1145/1964858.1964870.

[17] Z.-G. Zhao, X.-G. Guo, C.-T. Xu, B.-R. Pan, L.-X. Xu, Bibliometric analysis on retinoblastoma literatures in PubMed during 1929 to 2010, Int. J. Ophthalmol. 4 (2) (2011) 115–120, doi:10.3980/j.issn.2222-3959.2011.02.01.

[18] R. Recuero, R. Araujo, On the rise of artificial trending topics in twitter, in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12, ACM, New York, NY, USA, 2012, pp. 305–306, doi:10.1145/2309996.2310046.

[19] W. Ding, C. Chen, Dynamic topic detection and tracking: a comparison of HDP, C-word, and cocitation methods, J. Assoc. Inf. Sci. Technol. 65 (10) (2014) 2084–2097, doi:10.1002/asi.23134.

[20] Y. Zhang, G. Zhang, H. Chen, A.L. Porter, D. Zhu, J. Lu, Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research, Technol. Forecast. Soc. Change 105 (2016) 179–191, doi:10.1016/j.techfore.2016.01.015.

[21] M. Cobo, A. López-Herrera, E. Herrera-Viedma, F. Herrera, An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the fuzzy sets theory field, J. Informetr. 5 (1) (2011) 146–166, doi:10.1016/j.joi.2010.10.002.

[22] M. Cobo, A. López-Herrera, E. Herrera-Viedma, F. Herrera, SciMAT: a new science mapping analysis software tool, J. Am. Soc. Inf.Sci. Technol. 63 (8) (2012) 1609–1630, doi:10.1002/asi.22688.

[23] M. Cobo, M. Martínez, M. Gutierrez-Salcedo, H. Fujita, E. Herrera-Viedma, 25 years at knowledge-based systems: a bibliometric analysis, Knowl. Based Syst. 80 (Supplement C) (2015) 3–13. 25th anniversary of Knowledge-Based Systems. doi: 10.1016/j.knosys.2014.12.035.

[24] H. Zhou, H. Yu, R. Hu, J. Hu, A survey on trends of cross-media topic evolution map, Knowl. Based Syst. 124 (Supplement C) (2017) 164–175, doi:10.1016/j.knosys.2017.03.009.

[25] E. Shabunina, G. Pasi, A graph-based approach to ememes identification and tracking in social media streams, Knowl. Based Syst. 139 (Supplement C) (2018) 108–118, doi:10.1016/j.knosys.2017.10.013.

[26] D. Chavalarias, J.-P. Cointet, Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study, Scientometrics 75 (1) (2008) 37–50, doi:10.1007/s11192-007-1825-6.

[27] G.M. James, C.A. Sugar, Clustering for sparsely sampled functional data, J. Am. Stat. Assoc. 98 (2003) 397–408.

[28] J. Jacques, C. Preda, Funclust: a curves clustering method using functional random variables density approximation, Neurocomputing 112 (2013) 164–171.

[29] J. Jacques, C. Preda, Model-based clustering for multivariate functional data, Comput. Stat. Data Anal. 71 (2014) 92–106, doi:10.1016/j.csda.2012.12.004.

[30] N. Coffey, J. Hinde, E. Holian, Clustering longitudinal profiles using p-splines and mixed effects models applied to time-course gene expression data, Comput. Stat. Data Anal. 71 (Supplement C) (2014) 14–29, doi:10.1016/j.csda.2013.04.001.

[31] M. Giacofci, S. Lambert-Lacroix, G. Marot, F. Picard, Wavelet-based clustering for mixed-effects functional models in high dimension, Biometrics 69 (1) (2013) 31–40, doi:10.1111/j.1541-0420.2012.01828.x.

[32] S.X. Lee, G.J. McLachlan, Model-based clustering and classification with non-normal mixture distributions, Stat. Methods Appl. 22 (4) (2013) 427–454, doi:10.1007/s10260-013-0237-4.

[33] C. Angelini, D.D. Canditiis, M. Pensky, Clustering time-course microarray data using functional Bayesian infinite mixture model, J. Appl. Stat. 39 (1) (2012) 129–149, doi:10.1080/02664763.2011.578620.

[34] A. Rodríguez, D.B. Dunson, A.E. Gelfand, Bayesian nonparametric functional data analysis through density estimation, Biometrika 96 (1) (2009) 149–162.

[35] S. Ray, B. Mallick, Functional clustering by Bayesian wavelet methods, J. R. Stat. Soc. 68 (2) (2006) 305–332, doi:10.1111/j.1467-9868.2006.00545.x.

[36] J. Jacques, C. Preda, Functional data clustering: a survey, Adv. Data Anal. Classif. 8 (3) (2014) 231–255, doi:10.1007/s11634-013-0158-y.

[37] J.-L. Wang, J.-M. Chiou, H.-G. Mueller, Functional data analysis, Annu. Rev. Stat. Appl. 3 (1) (2016) 257–295, doi:10.1146/annurev-statistics-041715-033624.

[38] N. Shimizu, M. Mizuta, Functional clustering and functional principal points, in: B. Apolloni, R.J. Howlett, L. Jain (Eds.), Knowledge-Based Intelligent Information and Engineering Systems KES 2007, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2007, pp. 501–508, doi:10.1007/978-3-540-74827-4_63.

[39] T. Tarpey, K.K.J. Kinateder, Clustering functional data, J. Classif. 20 (1) (2003) 93–114, doi:10.1007/s00357-003-0007-3.

[40] U. von Luxburg, R.C. Williamson, I. Guyon, Clustering: science or art? JMLR 27 (2012) 6579.

[41] K. Jones, P. Willett, Readings in Information Retrieval, Morgan Kaufmann series in multimedia information and systems, Morgan Kaufman, San Francisco, 1997.

[42] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[43] M.A. Lukas, F.R. de Hoog, R.S. Anderssen, Practical use of robust GCV and modified GCV for spline smoothing, Comput. Stat. 31 (1) (2016) 269–289, doi:10.1007/s00180-015-0577-7.

[44] J. Ramsay, B.W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer, 2005, doi:10.1007/b98888.

[45] P. Montero, J. Vilar, Tsclust: an R package for time series clustering, J. Stat. Softw. 62 (1) (2014) 1–43, doi:10.18637/jss.v062.i01.

[46] G.E.A.P.A. Batista, E.J. Keogh, O.M. Tataw, V.M.A. De Souza, CID: an efficient complexity-invariant distance for time series, Data Min. Knowl. Discov. 28 (3) (2014) 634–669, doi:10.1007/s10618-013-0312-3.

[47] C. Hennig, M. Meila, F. Murtagh, R.E. Rocci, Handbook of Cluster Analysis, Chapman & Hall, 2016.

[48] B. Desgraupes, clusterCrit: Clustering Indices, 2016. R package version 1.2.7

[49] C. Genolini, X. Alacoque, M. Sentenac, C. Arnaud, kml and kml3d: R packages to cluster longitudinal data, J. Stat. Softw. 65 (4) (2015) 1–34, doi:10.18637/jss.v065.i04.

[50] C. Hennig, What are the true clusters? Pattern Recognit. Lett. 64 (2015) 53–62, doi:10.1016/j.patrec.2015.04.009.

[51] L. Grilli, M.A. Russo, R. Gismondi, Methodological proposals for a qualitative evaluation of italian durum wheat varieties, J. Appl. Econ. Sci. 7 (2(20)) (2012) 103–122.

[52] R Core Team, R: alanguage and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.

[53] D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding, in: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007, pp. 1027–1035.