MethodsX 5 (2018) 1576–1587

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Method Article

Chronological corpora curve clustering: From scientific corpora construction to knowledge dynamics discovery through word life-cycles clustering

Matilde Trevisani*, Arjuna Tuzzi

Department of Economics, Business, Mathematics and Statistics (DEAMS) of University of Trieste, Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA) of University of Padova, Italy

A B S T R A C T

Aim of this procedural method is to construct well-founded corpora of scientific literature, and, hence, to track the evolution of knowledge fields from the reconstruction and clustering of words' life-cycles. The method contains:

- an original selection process of relevant keywords involving the identification of relevant stems and stem *n*-grams through a matching with item lists of relevant glossaries;
- several types of normalization of temporal trajectories of word raw frequencies
- a properly customized clustering of word life-cycles, with a graphical extensive investigation of the best candidates for cluster number, to unveil the important dynamics and decipher the history of a scientific field.

A R T I C L E I N F O

Method name: Chronological corpora curve clustering

Keywords: Diachronic corpora, Functional data analysis, Normalization, Cluster number selection

Article history: Received 23 March 2018; Accepted 10 November 2018; Available online 19 November 2018

Specifications Table

Subject area	*Computer Science*
More specific subject area	*Computational linguistics*
Method name	*Chronological corpora curve clustering*

* Corresponding author.
E-mail address: matilde.trevisani@deams.units.it (M. Trevisani).

https://doi.org/10.1016/j.mex.2018.11.010

Name and reference of original method	M. Trevisani, A. Tuzzi [1] Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories, Knowledge-Based Systems 146 (2018) 129-141.

Method details

Given a knowledge field of interest, the procedural method consists of two main phases:

I An information retrieval process that starting from a large corpus of texts retrieved from scientific articles published over a lengthy period by a selection of premier journals of the field, leads to an effective representation of the corpus by a lexical contingency table reporting the frequencies over time of all relevant keywords.

II A statistical learning process that through four stages
- normalization of time trajectories of word (raw) frequencies, chosen according to the different aspects of word life-cycles to be highlighted;
- filtering time trajectories of word (normalized) frequencies, interpreted as functional data (FD) and thus represented as smooth functions;
- curve clustering (CC) to discover important macro-dynamics latent to word micro-histories;
- interpretation by expert opinion to decipher detected dynamics,

leads to a reading (or readings) of the history of the knowledge field.
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.

Related work

The method aims at composing an history of a field of knowledge by a distant reading of scientific literature available through an articles database. The objective situates our method within the various approaches for science mapping which has drawn much attention in the recent years. However, the main methodologies developed in bibliometrics, scientometrics, informetrics and related fields, though partly sharing similar purposes, are substantively different from our proposal and cannot answer our particular question effectively.

Topic modelling aims at detecting topics, i.e. thematic groups, in collections of documents. Moreover, when documents exhibit a temporal ordering, it enables the discovery of topic trends. Latent Dirichlet Allocation (LDA), the most widespread topic model, is a probabilistic generative process that models each document as a mixture of topics where each topic corresponds to a multinomial distribution over words [2]. Topics over time can be detected by modelling time jointly with word co-occurrence patterns for topic discovery [3,4]. A further extension of LDA incorporates both the temporal ordering and the authorship information of documents to improve topic discovery process [5]. Topic modelling connects to scientometrics or, more in general, to quantitative methods for mapping knowledge domains from scientific article databases. They are based on term and/or citation co-occurrences in documents, possibly observed over time in order to reconstruct a field's evolution [6,7]. Recent developments of co-citation network-based analyses build a dynamic scientific map via overlapping authors across fields [8] or via communities of authors working on semantically related topics at the same time [9].

Moreover, recently, generative probabilistic models (like LDA and the hierarchical Dirichlet process) have been exploited for topic detection and tracking (TDT) or for emerging topic detection (ETD), both that can be framed in dynamic science mapping [10,11]. After this brief overview of 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, the differences from our approach are evident. First of all, science mapping research is based on co-occurrences in documents

possibly observed over time, while our work considers word co-occurrence solely in time, as our primary focus is the temporal evolution of words. Then, more importantly, topic-centered methods focus first on the structure of science and on detecting topics and then on tracking their evolution, whereas our approach focuses first on tracing life cycles of words and then on detecting important dynamics of temporally homogeneous groups of words in order to decipher the history of a knowledge field. As a consequence, in topic-centered methods, words that represent the same topic (as they appear together in documents) may have an irreconcilable temporal evolution, whereas, in our approach, different themes, research fields and schools of thought can on principle be represented within the same group of words. Moreover, in topic-centered methods, 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. Conversely, in our approach, the detected dynamics really represent temporal patterns of words, e.g., 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.

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 [12]. Moreover, B-splines enable us to recognise continuous and regular curves, and hence more easily interpretable shapes. Upstream, we decided for a distance-based approach to CC, 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 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. 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 [13] and [14]).

Procedure

I – Compiling and pre-processing the corpus

Corpus design and compilation

0 Selection of data sources, i.e. choice of outstanding journals able to cover main topics and represent the temporal evolution of the knowledge field.

1 Text harvesting, i.e. downloading of available information on articles (authors, title/abstract/full text, number, issue, volume) from journal archives, to constitute the corpus. Texts under consideration may consist of titles or abstracts or full texts of the articles. The corpus is typically organized into subcorpora, i.e. collections of texts sharing the same time reference, thus generating a sequence of text sets associated with chronological points on the time axis.

2 Tokenization of the corpus, i.e. identification of all words (sequences of letters isolated by means of separators). The corpus contains a finite set of different words (i.e. word-types) that represents the vocabulary (or word list) of it. A word-token is a particular occurrence of a word-type and the number of occurrences is the word-type frequency.

Preparation of textual data

3 Stemming, i.e. transformation of words into stems by means of the Porter's stemming algorithm [15].

4 Identifying stem-segments, i.e. identification of all sequences of stems (or stem n-grams) occurring in the corpus at least twice and composed of a minimum of two and a maximum of six consecutive stems. In order to select "content sequences" (e.g. nouns like *generalized linear model*) and disregard "empty sequences" (e.g. grammatical sequences like *such as the*) as well as incomplete sequences (e.g. *President of the*), stem-segments are ranked according to Morrone's IS indexes [16].

Table 1
Excerpt of the normalization plan from Table A.2 in [1].

Normalization: by col	Subcorpus		Matrix		
by row	# titles	#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
max row freq	*d*	*d*	*d*	*d*	r_3
	c_1	c_2	c_3	c_4	

5 Tagging keywords, i.e. identification of all relevant statistical keywords (stems and stem-segments) by matching the (stemmed) vocabulary of the corpus with the (stemmed) list of items retrieved from relevant glossaries of the knowledge field. The tagging procedure assigns a label to all vocabulary items that are included in glossaries.
6 Thresholding, i.e. selection of all keywords with frequencies at least equal to an opportunely fixed threshold.

Finally, the corpus is represented by a keywords × documents/time-points contingency table containing the frequencies of the selected keywords (by row) along the time-points (by column) of the considered period.

Stemming can be carried out by the Porter Stemmer available online (http://textanalysisonline.com/nltk-porter-stemmer) or, alternatively, within the R software environment [17] by the *wordstem* routine of the *snowballC* library. We use Taltac software [18] for tagging though it can be equivalently performed by any software enabling the comparison between two lists (e.g., Excel).

II – Statistical learning

Normalization

A chronological corpus is typically characterized by the following features.

(i) Size of subcorpora (number of texts and their size in word-tokens) may vary greatly over time.
(ii) The *large number of rare events* (LNRE) property of textual data, i.e. a large number of word-types having a quite low probability of occurring. This property implies:
- total frequency (or popularity) of individual words in the entire corpus is greatly variable
- frequency spectrum by time-point is highly asymmetric,
- sparsity, i.e. many cells of the contingency table have small counts or are empty.

In the section Method validation, features (ii) are evident from the plot of the original word trajectories (Fig. 2). Classification of words according to their popularity highlights the great disparity of curve amplitude between high-frequency and low-frequency words (VH, H, L and VL classes are identified by colour intensity in Fig. 2) and the 0-level curve sections characterizing rare words.

From the foregoing, normalization of raw frequencies is necessary to properly reconstruct and compare the temporal evolution of words.

Several types of normalization are showed in the table below (which is an excerpt of Table A.2 in [1]).

A sort of normalization by column (c_1, c_2, c_3 or c_4) is necessary to adjust the uneven document dimension across time (i). A sort of normalization by row (r_1, r_2 or r_3) allows to compare word trajectories by timing (synchrony) regardless of height (popularity) (ii). A double (both by row and column) normalization (*d*) serves to fix both (i) and (ii).

In the section Method validation, the calculation of a specific double normalization (d_1) is showed.

Filtering

In our method, the time trajectory of word frequencies is viewed as a proxy of word diffusion and vitality, i.e. of word life-cycle. Then, we adopt a functional data analysis (FDA) approach under which the time trajectory of word frequencies constitutes a functional datum assumed to be a realization of an underlying continuous function representing the word temporal evolution.

Let $\mathbf{y}_i = \{y_{ij}\}$ the functional observation of word *i* consisting of the set of (normalized) frequencies at time-points $j = 1, \ldots, T$, for each $i = 1, \ldots, N$, and $x_i(t)$ the underlying continuous function representing the word temporal development. The following choices are taken for filtering $x_i(t)$ from $\mathbf{y}_i$.

We adopt the basis function approach for representing FD as smooth functions where $x_i(t)$ is expressed as a finite linear combination of basis functions [12]. We consider B-spline bases which are piecewise polynomials joined smoothly at the interior nodes. Lastly, we place knots – the values of t at which adjacent segments are joined – at each time-point of observation.

As regards the estimation, we adopt the roughness penalty approach for smoothing FD where the estimate of xi is the one optimizing the bias-variance trade-off by tuning the smoothing parameter λ. We consider the generalized cross validation (GCV) criterion for selecting the optimal smoothing by varying spline order *m* (*m* from 1 to 8) as well as roughness penalty order *r* (besides the standard $r = m-2$, $r = 2$, for $m > 3$, $r = 1$, for $m > 2$, finally, $r = 0$) [1].

In the section Method validation, the optimal smoothing selection is illustrated for the case of d1 normalized data (Fig. 4).

The calculation is carried out within the *fda* library in R and an ad-hoc developed routine.

Curve clustering

We adopt a distance-based method to CC where the distance between curves is approximated by using the discretely observed evaluation points of the estimated curves $x_i(t)$ [13].

The following choices are taken for clustering:

- k-means algorithm
- several options for distance: besides the conventional distances (Euclidean or Manhattan, between others), other options can be taken from the broad range of dissimilarity measures set out to perform clustering of time series [19].
- for each cluster number (*k* from 2 to an opportune range maximum), 20 re-runs from different initial configurations set through the k-means++ seeding method.
- the best candidates to cluster number are identified by pooling the ratings from a large number of clustering quality criteria (about 50, see [20] and [21]). More in detail, in the order:
  - a ranking of cluster number is computed for each quality index,
  - all the rankings are pooled and, for each cluster number, the frequencies of being ranked first (top-1), second (top-2), third (top-3) and fourth (top-4) are 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 first four top positions for each cluster number (see Fig. 6; in section Method validation for illustration).

Clustering results obtained with the cluster numbers selected as the best candidates are then presented to experts (of the subject matter) who possibly will guide towards other analyses.

R contains several k-means implementations as well as libraries for computing clustering quality criteria. Our procedure uses the *kml* routine [21] which is designed specifically for longitudinal data and which provides various efficient methods of k means initialization. The *clusterCrit* [20] and *kml* [21] are the packages used to gather the large basket of quality criteria considered by our method. These include measures of within-cluster homogeneity, e.g., *Ball-Hall, Banfeld-Raftery, C-index, 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 generalizations, *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.

Method validation

For illustration, we apply the procedural method to the corpus of titles of scientific papers published by the American Statistical Association (ASA) journals in the time span 1888–2012 in order to trace a history of Statistics.

	keyword	v001	v002	v003	v004	v005	..	..	v098	v099	v100	v101	v102	v103	v104	v105	v106	v107
1	statist	17	31	25	11	21	..	..	15	13	10	22	11	15	4	5	5	2
2	model	0	0	0	1	0	..	..	22	30	29	32	22	36	32	16	14	24
3	test	0	0	0	0	0	..	..	3	9	4	8	7	10	11	11	11	4
4	data	0	0	1	0	0	..	..	10	10	13	16	15	13	10	19	18	13
5	distribut	1	0	4	1	0	..	..	9	6	6	11	1	6	5	1	2	2
6	analysi	0	0	0	0	0	..	..	8	10	10	20	16	16	14	8	9	3
7	sampl	0	0	0	0	0	..	..	2	2	5	5	3	3	4	4	5	1
8	method	0	0	1	0	0	..	..	11	7	12	7	3	12	3	4	8	2
9	popul	0	7	3	3	5	..	..	1	1	2	1	2	2	2	2	5	1
10	regress	0	0	0	0	0	..	..	5	4	7	6	11	2	6	1	7	5
...	...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
...	...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
...	...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
...	...	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
891	smooth spline	0	0	0	0	0	..	..	1	1	0	0	0	0	1	0	0	0
892	curv fit	0	0	0	0	0	..	..	0	0	0	0	0	0	0	0	0	0
893	t test	0	0	0	0	0	..	..	0	0	0	0	0	0	0	0	0	0
894	estim function	0	0	0	0	0	..	..	0	0	0	0	1	1	0	1	0	0
895	high breakdown	0	0	0	0	0	..	..	0	0	0	1	0	0	0	1	0	0
896	normal variabl	0	0	0	0	0	..	..	0	0	0	0	0	0	0	0	0	0
897	unit root	0	0	0	0	0	..	..	0	1	1	0	0	0	0	0	0	0
898	british	0	0	0	0	1	..	..	0	0	0	0	0	0	0	0	0	0
899	metropolitan	0	0	0	0	0	..	..	0	0	0	0	0	0	0	0	0	0
900	census	0	0	0	0	0	..	..	0	0	0	0	0	0	0	0	0	0

Fig. 1. Excerpt of the 900 (words) × 107 (volumes) table from the corpus of titles of papers published by the ASA's journals 1888–2012.

Corpus design and compilation

0 The ASA represents the world's largest community of statisticians and the Journal of the ASA (JASA) has long been considered the world's premier review in its field. Established in 1888, JASA, which has two predecessors (Publications of the ASA, 1888–1912, Quarterly Publications of the ASA, 1912–1921) is one of the oldest and prestigious statistical journals.

1 Download from journal archives of available information for all issues that refer to 12,577 items 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). Titles of articles are the text considered in this study.

2 After discarding items that are not articles (e.g., List of publications, News) or do not include content words (e.g., Comment, Rejoinder), the corpus includes 10,077 titles and is composed of 7746 word-types and 87,060 word-tokens.

Preparation of textual data

3 After stemming, 4834 different stems are obtained (e.g., the word-types: *model*, *models*, *modeling*, and *modelling* are replaced with the same stem *model*).

4 All potentially relevant stem-segments are identified (e.g., *model select*, *hierarch model*, *log linear model*) and included in the word list.

5 Relevant statistical keywords (e.g. stemmed words: *statist*, *model*, *test*, *distribut*, *analysi*, *regress*, *probabl*; and sequences of stemmed words: *time seri*, *regress model*, *conting tabl*, *confid interv*, *maximum likelihood estim*, *analysi of varianc*, *normal distribut*) are tagged by matching the stemmed vocabulary of the corpus with a stemmed list (over 12,700 unique entries) including all non-redundant entries of six Statistics glossaries:

1 ISI - International Statistical Institute;
2 OECD - Organisation for Economic Cooperation and Development;
3 Statistics.com - Institute for Statistics Education;
4 StatSoft Inc.;
5 University of California, Berkeley;
6 University of Glasgow.

6 After fixing the threshold at 10, 900 keywords are finally selected.

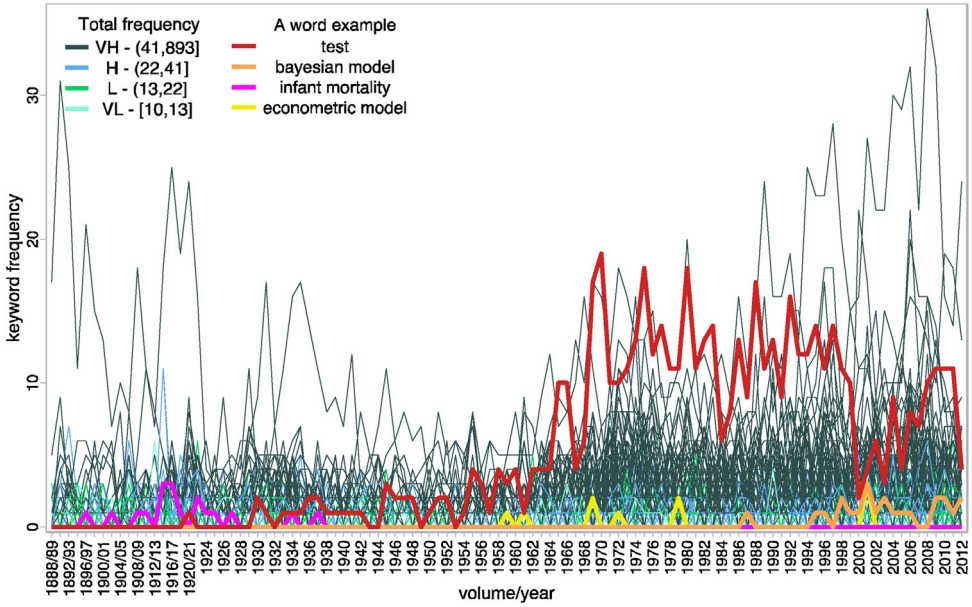

Fig. 2. Word trajectories (original data): *y*-axis represents the word raw frequency for each volume; *x*-axis represents the volume publication year; line color identifies the word frequency class (Very Low, Low, High and Very High denote equal-frequency intervals of word total frequency in the entire corpus). An example of word trajectory has been superimposed for each frequency class.

At the end, the corpus originates a 900 (words) × 107 (time-points/volumes) contingency table (Fig. 1).

Normalization

For illustration, we choose to transform data (Fig. 2) by the double normalization d_1 (Table 1) which is equivalent to calculate a χ^2 distance between original word profiles if the Euclidean distance is used as measure of dissimilarity.

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 normalized frequency is computed as:

$$y_{ij} = \frac{n_{ij}}{n_{i.}\sqrt{n_{.j}/n}}$$

($n_{.j}/n$ is the j-column mass in correspondence analysis).

Note that this double normalization produces a somewhat reversed asymmetry (low-frequency words tend to dominate in amplitude on high-frequency words, see the inversion of color intensity in Fig. 3). This is mainly due to a greater sparsity of low-frequency keywords across time.

Filtering

Optimal smoothing for d_1 normalized data is achieved with spline order $m = 3$ and smoothing parameter $\lambda = 10^{1.75}$ ($df = 7.4$) under a roughness penalty of order $r = 1$ (Fig. 4).

A sample of curves fitted by the optimal smoothing are shown in Fig. 5, from the word with highest root mean square (RMS) residual (*rural*) to the word with lowest RMS residual (*model*).

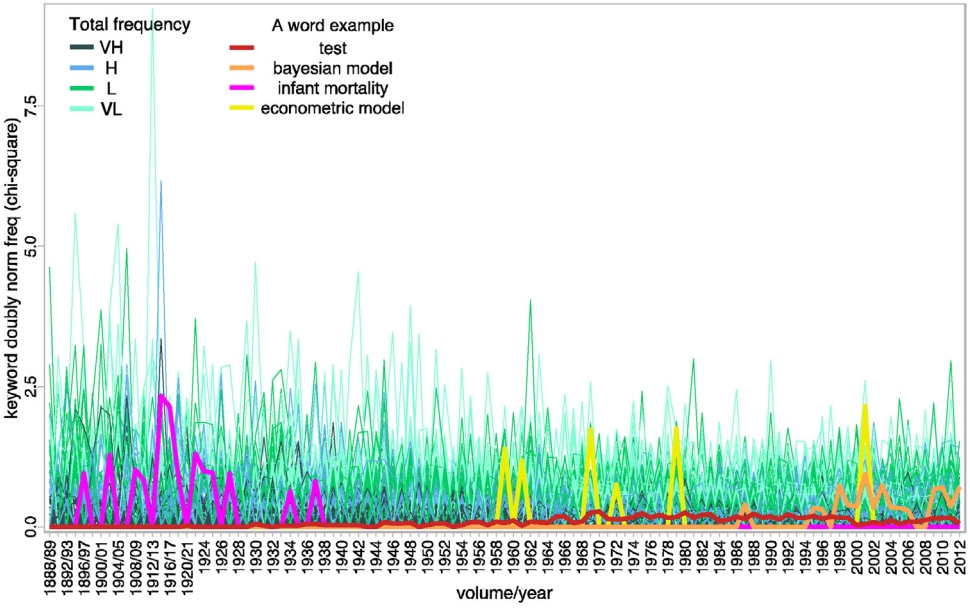

Fig. 3. Keyword trajectories (doubly normalized data, d_1 or χ^2-like).

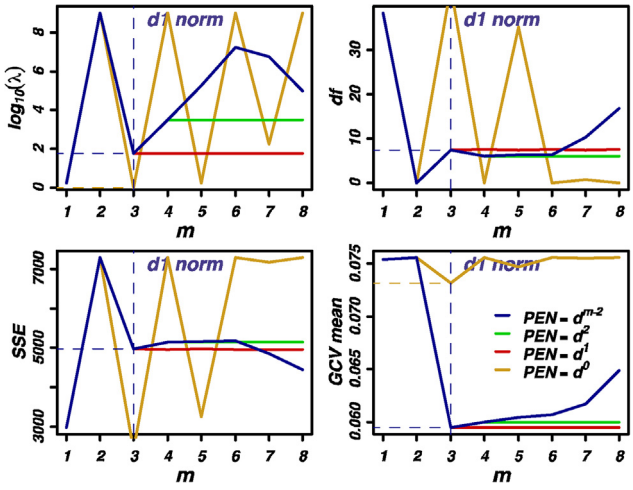

Fig. 4. Smoothing selection: overview of $\log_{10}\lambda$, effective degrees of freedom (*df*), sum of square errors (SSE) and GCV by varying order *m* and roughness penalty order *r* (PEN_r). Optimal smoothing is obtained by minimizing GCV. d_1 normalization.

Curve clustering

Curves are partitioned by means of the k-means algorithm combined with the Euclidean distance with cluster number *k* ranging from 2 to 26 and 20 reruns for each *k*.

A set of 49 quality criteria are then computed in order to identify the best candidates to cluster number.

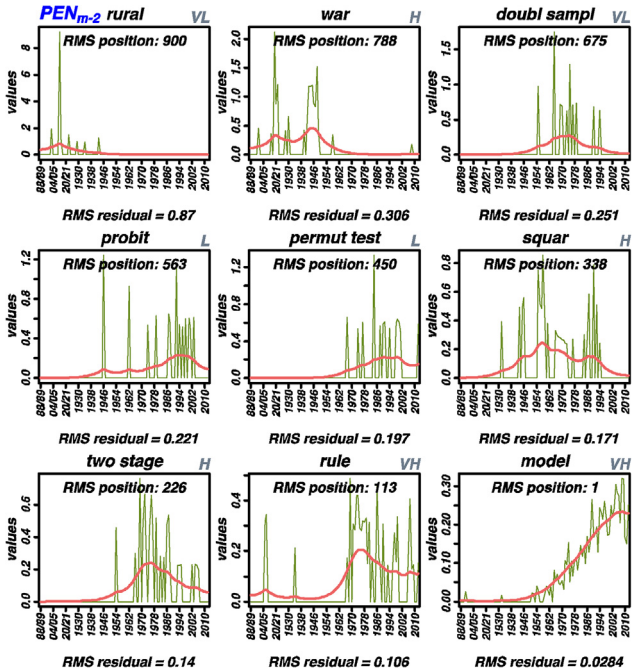

Fig. 5. Optimal smoothing fit: a selection of fitted curves ordered according to decreasing root mean square (RMS) residual. Fit of a smoothing spline of order $m = 3$, with PEN_1, to d_1 normalized data.

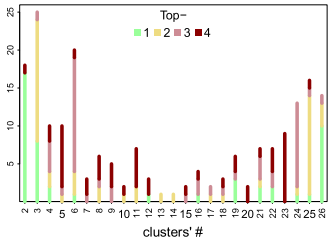

Fig. 6. Cluster number selection: frequency of being ranked first (top-1), second (top-2), third (top-3) and fourth (top-4) for each cluster number by pooling rankings from the overall quality criteria. d_1 normalization.

Visual representation of the rating for the cluster number shows (Fig. 6) that:

(i) partitions into two/three clusters are the best rated,
(ii) partitions with a cluster number close to the maximum of the considered range (24–26) have also been frequently selected in the highest positions,
(iii) in the range of more interesting cluster numbers (neither too low nor too high), the most selected in the top four positions is 6, second is 4, third is 19 (the eye should be guided both by the bar height, corresponding to the cumulated frequency of being in the top four, and by the color composition, informing on the position level).

Note that the final set of best candidates for cluster number is the output of an R code that essentially mimics a qualitative rating purely based on a graphical inspection.

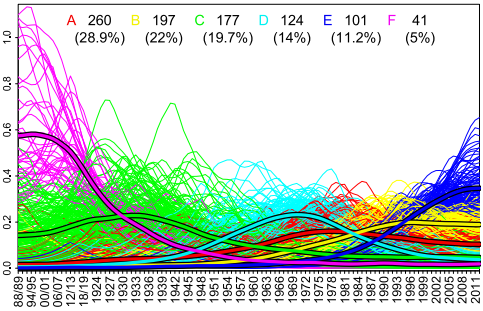

Fig. 7. Clustering: best partition into 6 groups. d_1 normalization.

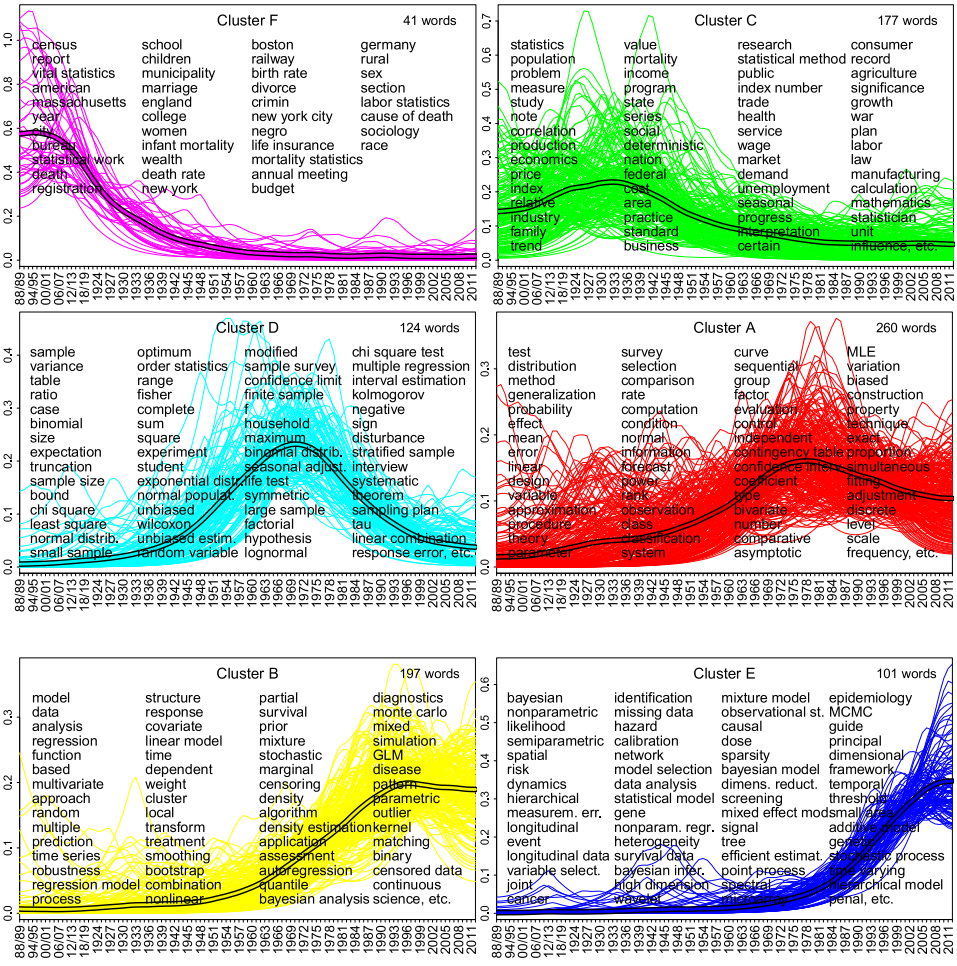

Fig. 8. Clustering: individual clusters of the 6 group-best partition. d_1 normalization.

Solution (i) (the two-group partition) reflects the substantial bifurcation of the historical period around the sixties when Statistics was born as an autonomous discipline (see [1] and [22] for explanation). Solution (ii) (25/26-group partition), 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 many quality criteria of data normally distributed hence of compact and convex clusters. That premised, we choose to investigate the most interesting solution (iii), that is the set of cluster numbers neither too small nor too large, and to subject them to the scrutiny of experts.

Here, we illustrate the best partition found with the cluster number ranked first, that is $k = 6$. The graphical output shows the groups all together with the cluster mean patterns (Fig. 7), and individually (Fig. 8). Note that, in order to make the reading easier, 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. Moreover, in order to make the identification of possible subsequent phases in the knowledge field evolution easier, individual clusters have been chronologically ordered. The found dynamics are then examined and - whether considered interesting - eventually interpreted by subject matter experts. A possible reading of the history of Statistics on the basis of the illustrated findings is offered in [1].

Acknowledgements

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. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.mex.2018.11.010.

References

[1] M. Trevisani, A. Tuzzi, Learning the evolution of disciplines from scientific literature: a functional clustering approach to normalized keyword count trajectories, Knowl. Based Syst. 146 (2018) 129–141, doi:http://dx.doi.org/10.1016/j.knosys.2018.01.035.
[2] D.M. Blei, A.Y. Ng, M. Jordan, Latent Dirichlet Allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (Suppl. 1) (2004) 5228–5235.
[4] D.M. Blei, J.D. Lafferty, Dynamic topic models, Proceedings of the 23rd International Conference on Machine Learning, (2006) , pp. 113–120.
[5] L. Bolelli, Ş. Ertekin, C.L. Giles, Topic and trend detection in text collections using latent dirichlet allocation, Advances in Information Retrieval, ECIR 2009, Lect. Notes Comput. Sci. 5478 (2009) 776–780.
[6] D. Chavalarias, J.-P. Cointet, Phylomemetic patterns in science evolution – the rise and fall of scientific fields, PLoS One 8 (2) (2013)e54847.
[7] K.W. Boyack, R. Klavans, Including cited non-source items in a large-scale map of science: What difference does it make? J. Informetr. 8 (3) (2014) 569–580.
[8] X. Sun, K. Ding, Y. Lin, Mapping the evolution of scientific fields based on cross-field authors, J. Informetr. 10 (3) (2016) 750–761.
[9] F. Osborne, G. Scavo, E. Motta, Identifying diachronic topic-based research communities by clustering shared research trajectories, European Semantic Web Conference (2014) 114–129.
[10] 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:http://dx.doi.org/10.1002/asi.23134.
[11] 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 (Suppl. C) (2017) 255–268, doi:http://dx.doi.org/10.1016/j.knosys.2017.07.011.
[12] J. Ramsay, B.W. Silverman, Functional Data Analysis (Springer Series in Statistics), Springer, 2005, doi:http://dx.doi.org/10.1007/b98888.
[13] J. Jacques, C. Preda, Functional data clustering: a survey, Adv. Data Anal. Classif. 8 (3) (2014) 231–255, doi:http://dx.doi.org/10.1007/s11634-013-0158-y.
[14] J.L. Wang, J.M. Chiou, H.G. Mueller, Functional data analysis, Annu. Rev. Stat. Appl. 3 (1) (2016) 257–295, doi:http://dx.doi.org/10.1146/annurev-statistics-041715-033624.
[15] M.F. Porter, An algorithm for suffix stripping, Program 14 (3) (1980) 130–137.

[16] A. Morrone, Temi generali e temi specifici dei programmi di governo attraverso le sequenze di discorso, in: M. Villone, A. Zuliani (Eds.), L'attività dei governi della Repubblica italiana (1948–1994), Il Mulino, Bologna, 1996, pp. 351–369.
[17] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.
[18] S. Bolasco, Taltac2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi, LED, Milano, 2010.
[19] P. Montero, J. Vilar, Tsclust: An R package for time series clustering, J. Stat. Softw. 62 (1) (2014) 1–43, doi:http://dx.doi.org/10.18637/jss.v062.i01.
[20] B. Desgraupes, clusterCrit: Clustering Indices, R package version 1.2.7, (2016) .
[21] 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:http://dx.doi.org/10.18637/jss.v065.i04.
[22] M. Trevisani, A. Tuzzi, A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal, Qual. Quant. 49 (3) (2015) 1287–1304, doi:http://dx.doi.org/10.1007/s11135-014-0050-7.