NLP

Course of DSSC Master degree - University of Trieste

Matilde Trevisani

DEAMS

2020/05/25 (updated: 2020-06-19)

Chronological Corpora

Outline

1. Diachronic corpora
2. Case study
3. CA
4. FDA approach
5. Sparsity
6. Wavelets and model-based CC
7. Splines and distance-based CC

Diachronic corpora

A **diachronic corpus** is a collection of texts accompanied by a temporal reference.

Usually these texts are organized in sub-corpora, that is in collections of texts that share the same time reference.

Many corpora contain **texts** organized **in chronological order**.

Temporal evolution of textual units of interest (words, key-words, sequences of words, bigrams, trigrams, etc.) and themes (topics = sets of words) is crucial to highlight the distinctive features characterizing texts, time periods or groups of words that show similar temporal trends.

Examples:

- institutional speeches delivered by public figures over the years
- articles retrieved from newspaper archives
- literary works written by authors during their life
- essays written by students at different stages of their educational experience
- documents of historical interest
- social network (Facebook, Twitter, blogs, ...)
- scientific literature ->

Constitution of a Corpus of scientific literature

A corpus is not simply a "collection of texts". A corpus must have characteristics of **breadth**, **coherence**, **homogeneity** and, in some cases, **exhaustiveness** that make it suitable for research purposes.

A corpus is a set of texts that responds to the needs of a specific **research question**.

Criteria for evaluating the quality of the corpus (some):

- **sizes** of texts
- **textual genre**
- **language**, **theme**, **style**, author (Muller & Brunet, 1988)

Five dimensions of **variation** of language (Berruto, 1987) that represent the research object:

1. **diachronic** (variation due to chronological differences),
2. **diatopic** (variation due to differences in geographical location),
3. **diaphasic** (variation due to differences in the communicative situation, e.g. formal vs. informal),
4. **diastratic** (variation due to differences in the social groups of reference, e.g. educational qualification),
5. **diamesic** (variation linked to the medium of transmission, typical opposition: oral vs. written text).

About the diachronic variation

The language needs long periods of time (centuries) to show significant changes.

Diachronic variations observable in the short term (decades) almost always concern the lexical level (typical examples are neologisms and foresters), which represents the most superficial part of the language.

A change at the syntactic level would represent a profound change in the language.

Bag-of-words

In a typical bag-of-words approach, the lexical table

Lexical table or frequency table or **vocabulary (of frequency)** is the list of word types with relative frequencies.

is of the type words per unit of time.

Lexical table: keyword year

	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

Excerpt of a lexical table (keywords by volumes / years)
[data from the corpus of titles of articles published by the ASA journals in the period 1888-2012]

An odd idea?

The idea of studying a word history, of "shaping" a word history is unusual in linguistic studies.

Starting from the studies of the Sixties by Bruno Migliorini, the main objective of research on the history of Italian language has always been dating the **birth of words** (first attestation) or, in some cases, detecting semantic changes across time.

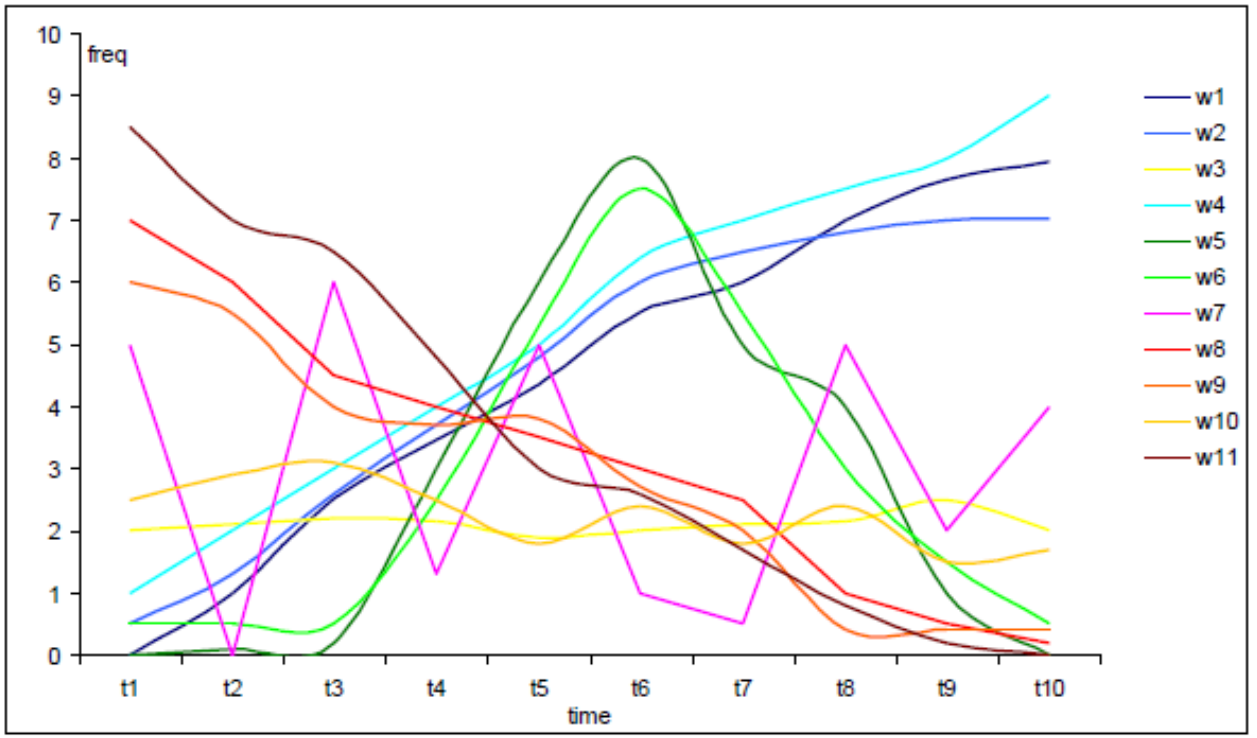
Beyond dating, little attention has been paid to history, fortunes or disappearance of words.

From a theoretical point of view, analyzes of historical textual data want to promote among scholars a reflection on concepts such as **quality of life** and **life cycle** of words.

In other areas, especially in the so called bibliometric disciplines, studies of scientific literature for **topic detection** or for **network** identification are more widespread and consolidated.

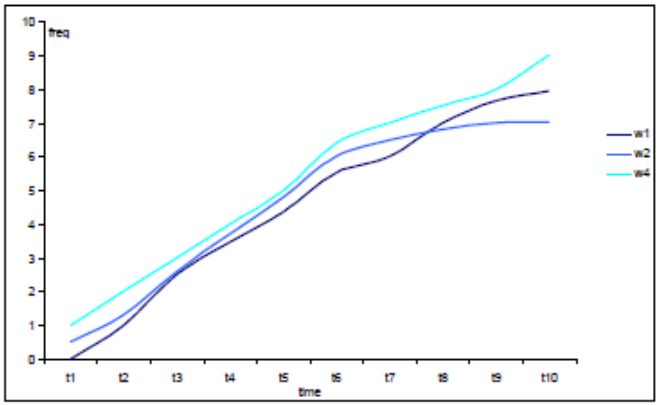
In the "ideal world"

In the analysis of phenomena that depend on time, we wish to find trends "easy to read"

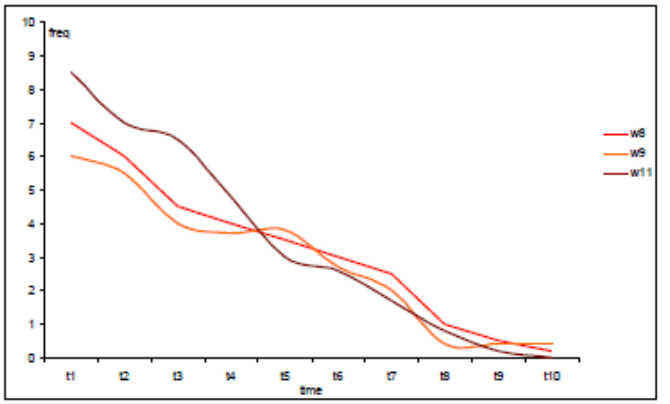


Synthetic data
11 words = trajectories / curves

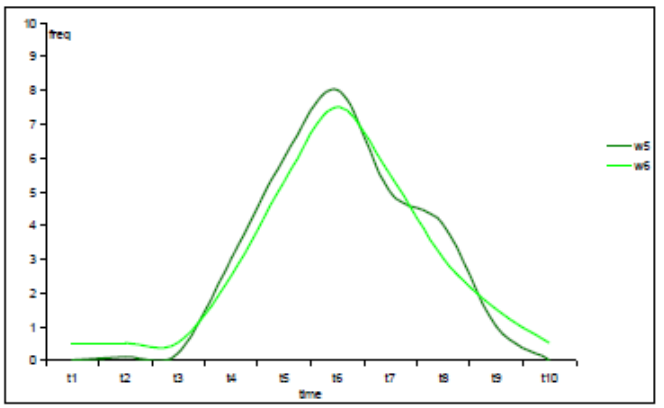
From trajectories to clusters that group words with similar time course:



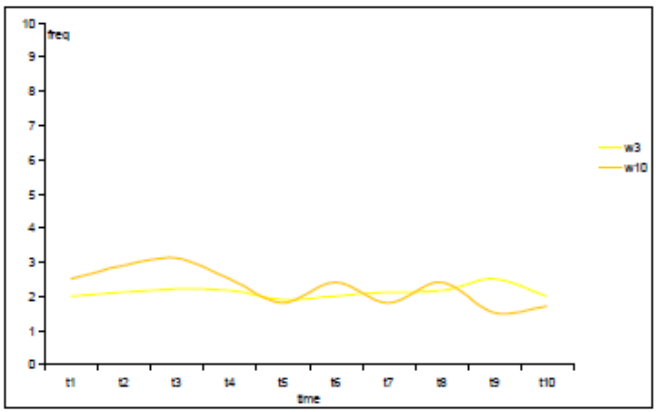
increasing trend



decreasing trend

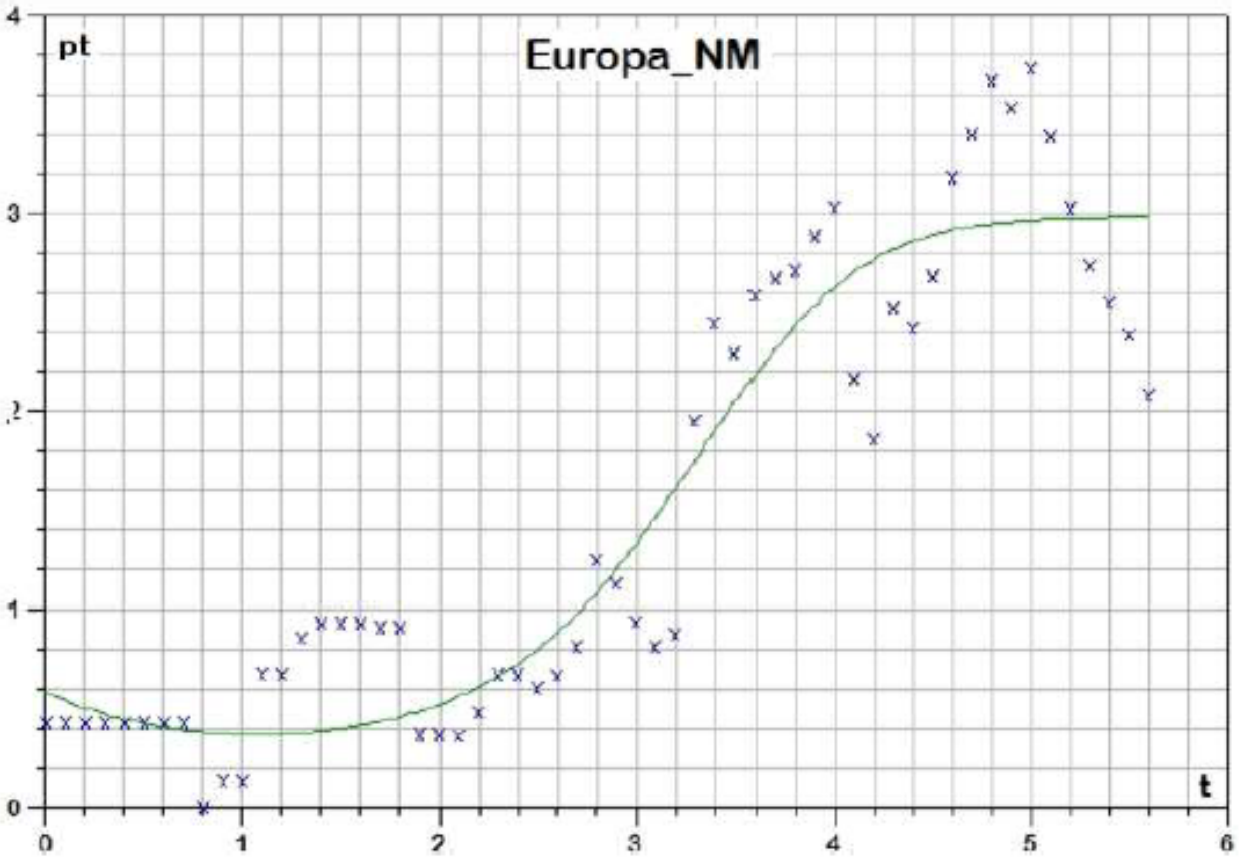


metheor-like



constant trend

In the analysis of phenomena that depend on time, we wish to find trends "easy to read"

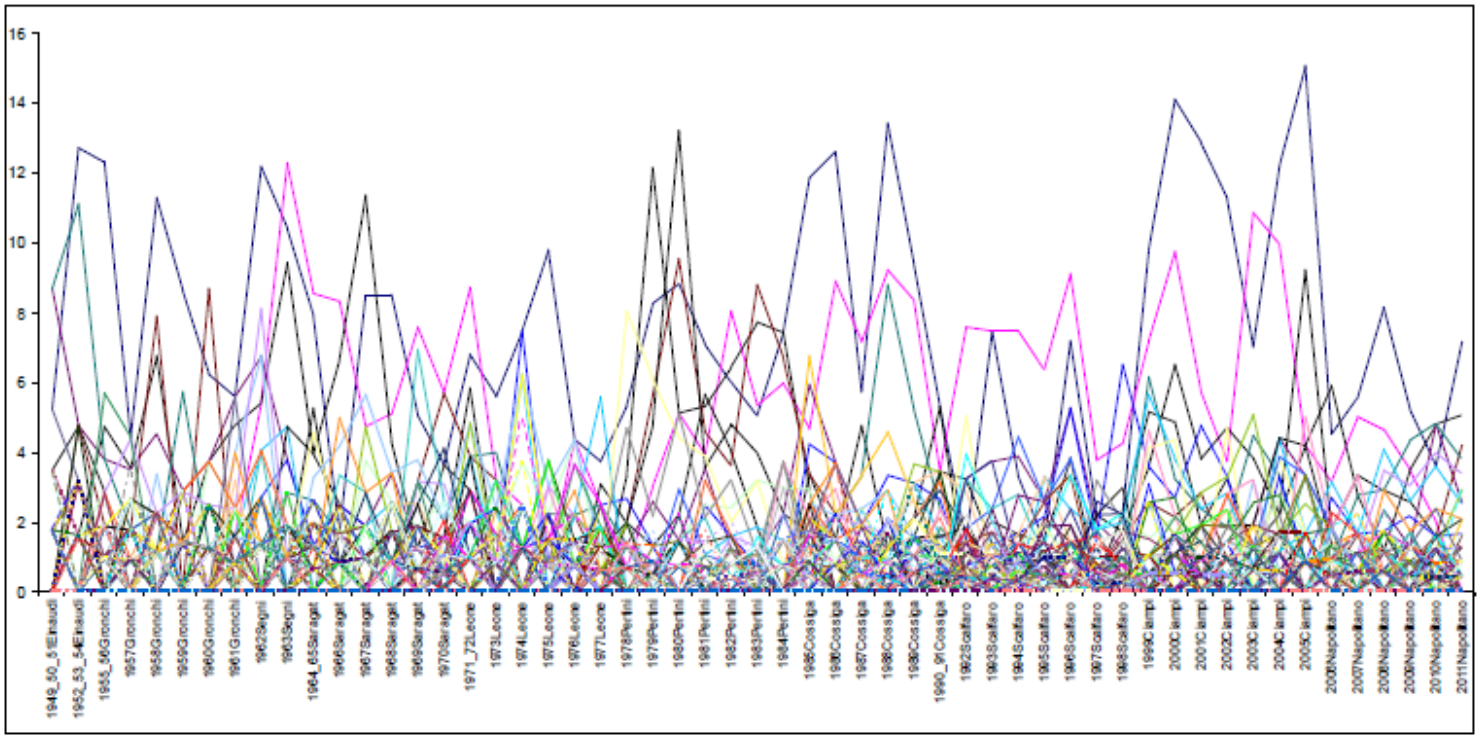


Corpus of end-of-year addresses of Presidents of Italian Republic
Piotrowski-Altmann Law (Altmann 1983)

]

In the "real world"

Though, in the real world, we need uncover regular forms and trends from a chaotic tangle of irregular curves.



Case-study: history of Statistics

We explored the opportunity of **reading** the temporal evolution of concepts, methods and applications in Statistics, i.e., **the history** of the discipline, through the temporal evolution of occurrences of keywords from articles published by an historical and outstanding journal.

Researchers studied the evolution of ideas to detect what were in the past and what are today the research themes of Statistics from both a methodological and application point of view.

Basic assumption is that the shape of trajectories drawn by the frequencies of keywords reflects the importance of themes in scientific discourse.

Case-study: history of Statistics

Research phases

a) identify and select a set of **keywords** in the texts

b) obtain a first representation of the relationship between keywords and time in order to verify the **existence of a temporal pattern**

c) identify keywords showing prototypical trends and cluster keywords showing **similar time trends** to reconstruct a history of the field of interest

JASA Corpus

The *American Statistical Association* (ASA) publications are a fundamental source to investigate the temporal evolution of Statistics and that is why we chose the articles published on the association's official magazine.

1. Publications of the American Statistical Association (PASA, 1888-1912)

2. Quarterly publications of the American Statistical Association (QASA, 1912-1921)

3. Journal of the American Statistical Association (JASA, 1922-current)

In the first study we considered:

- the **titles** of the articles (1888-2012)
  - then updated (1888-2016)
- the **abstracts** of the articles (1946-2016)

Corpus of titles: construction of lexical table

1. Text collection (**text harvesting**)

download of all data and references (title, abstract, volume, number, issue) from archives (ASA, ISI, JSTOR)

- 12,557 titles / articles
- 107 volumes / years (one per year with some exceptions)

2. Parsing words (**tokenization**)

A *word* is a sequence of letters isolated by means of separators (spaces and punctuation marks). A corpus contains a finite set of different words (i.e. *word-types*). A *word-token* is a particular occurrence of a word-type in the corpus

In this phase all the titles that do not represent real scientific articles (e.g. *List of publications*, *News*) or do not include content words (e.g. *Comment*, *Rejoinder*) were discarded. After **normalization** (upper/lower case, acronyms, etc.) and **part-of-speech** (POS, word tagging by grammatical categories)

- 10,077 titles
- 87,060 word-tokens
- 7,746 word-types

Construction of lexical table

3. **Stemming**

to overcome some limitations of word-types (e.g., tenses of verbs, plural forms of nouns), graphic forms have been replaced by stems through the most recent version of Porter's algorithm (e.g. *model*, *models*, *modeling*, and *modeling* forms are replaced by the same stem *model*)

- 4,834 stem-types

4. Identifying **stem- segments**

words have different meanings if considered alongside the adjacent words, then n-stem-grams (i.e., sequences of stems such as *model select*, *addit model*, *hierarch model*, *log linear model*, *dynam model*) occurring at least twice in the corpus and composed of a minimum of two to a maximum of six consecutive stems were recognized

In this phase all the "non-relevant" stem sequences were discarded (e.g. grammatical sequences such as *such as the*) by means of Morrone's IS index and potential relevant segments were identified.

Construction of lexical table

5. **Keyword** labeling (tagging)

to extract only stems and stem sequences relevant to the study of history of Statistics, the corpus vocabulary was compared with a list of 12,700 entries obtained from six **glossaries** of the discipline

```
  i. ISI - International Statistical Institute
 ii. OECD - Organization for Economic Cooperation and Development
iii. Statistics.com - Institute for Statistics Education
 iv. StatSoft Inc.
  v. University of California, Berkeley
 vi. University of Glasgow
```

6. **Thresholding**

all keywords with frequency were selected [arbitrary choice, but it means that on average one keyword was present at least once every 10 years in the titles]

- 900 keywords
- 107 time points (volumes/years) [1888-2012] for titles
  - 111 time points (volumes/years) [1888-2016] after update
- 4,915 keywords and 71 time points for abstracts [1946-2016]

Another example: Philosophy corpora

Corpus consisting of **titles** from journals:

1. Journal of Philosophy (1904-2015)
2. Mind (1897-2016)
3. Philosophical Review (1892-2016)
4. The Monist (1890-2015)

rivista	paese	periodo	volumi		fascicoli	titoli	occorrenze
JPPSM JoP	USA	1904-1920 1921-2015	1-112	112	2.285	10.876	66.439
PhilRev	USA	1892-2016	1-125	125	616	9.465	54.636
Mind	UK	1876-1891 1892-2016	1-16 1-125	141	500	8.638	48.866
Monist	UK	1890-1936 1962-2015	1-98	98	384	3.472	21.710
Corpus						32.451	191.651

Corpus "**full text**" of Journal of Philosophy (1946-1975)

- 30 years
- 1,570 articles
- 7 milions of occurrences

... the all corpora

Social psychology

1. European Journal of Social Psychology
   - titles (1971-2016)
   - abstracts (1992-2016)
2. Journal of Personality and Social Psych.
   - titles (1965-2016)
   - abstracts (1965-2016)

Sociology

1. American Journal of Sociology
   - titles (1895-2016)
   - abstracts (1921-2016)
2. Journal for the Scientific study of Religion
   - titles (1961-2015)
   - abstracts (1967-2015)

Statitics

1. Journal of the American Statistical Ass.
   - titles (1888-2016)
   - abstracts (1888-2016)

Philosophy

1. Journal of Philosophy
   - titles (1904-2015)
   - full text (1946-1975)
2. Mind (1897-2016) (titles)
3. Philosophical Review (1892-2016) (titles)
4. The Monist (1890-2015) (titles)

Linguistics

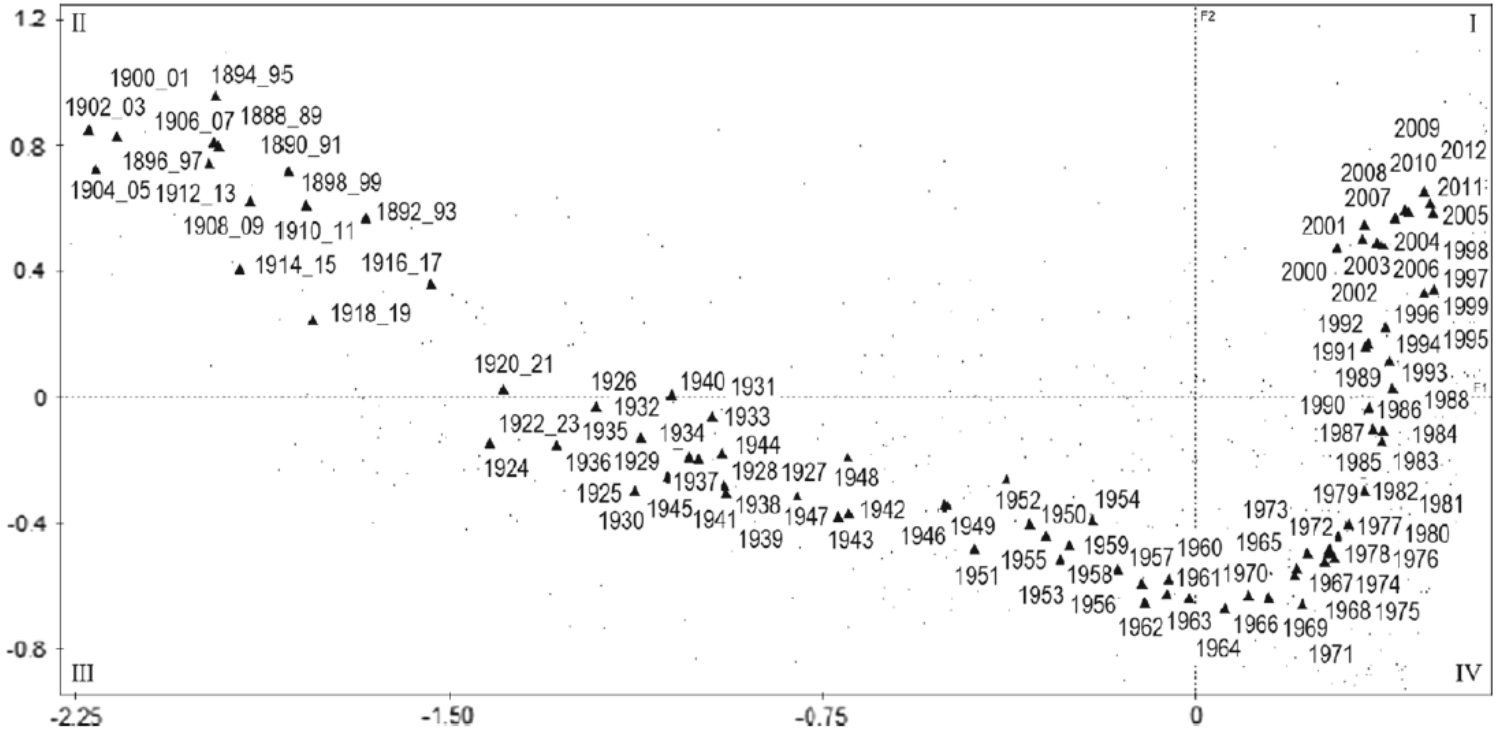
1. Archivio Glottologico Italiano (1873-2016)
2. Italia dialettale (1924-2016)
3. Lingua nostra (1939-2016)

Correspondence Analysis

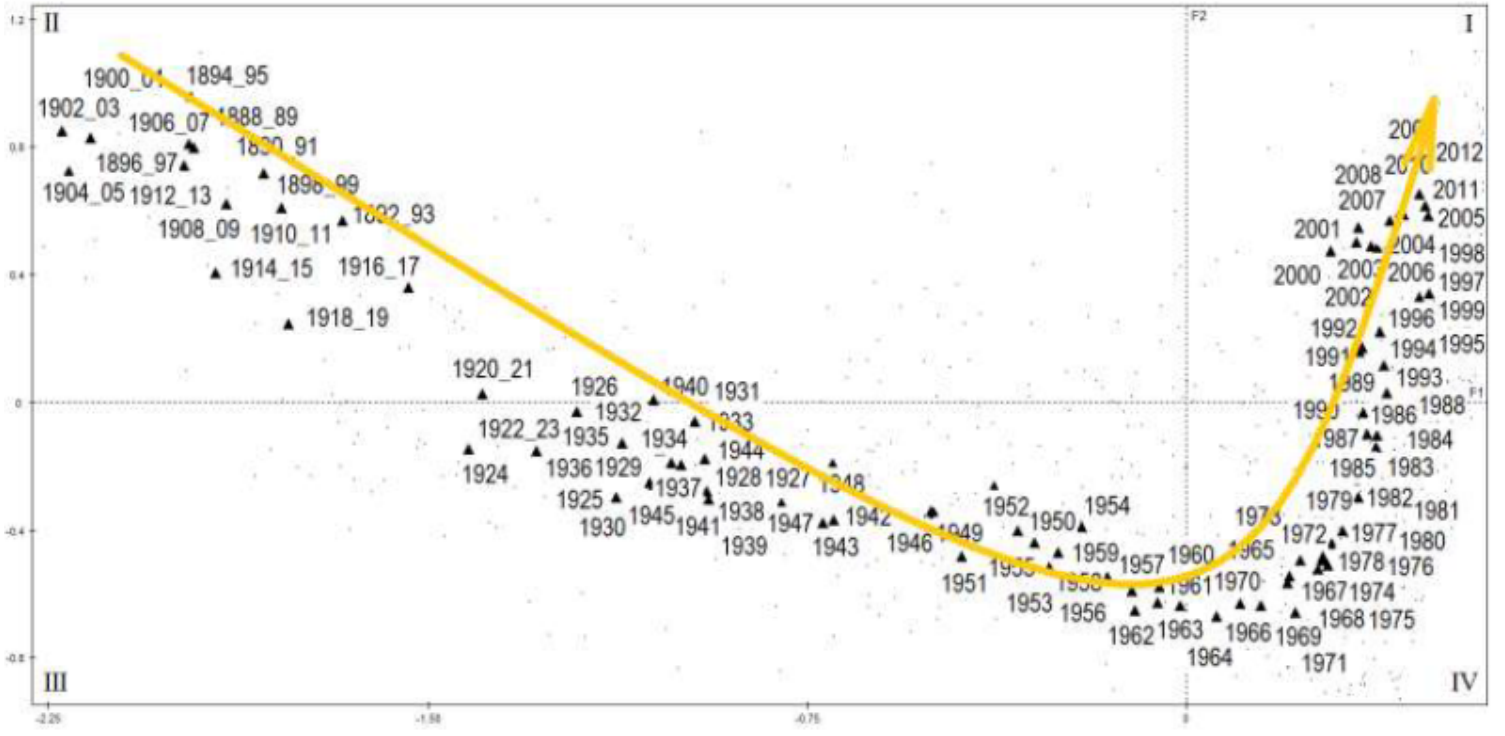
Correspondence analysis (CA) was used to graphically represent the relationships between keywords and time points (years) plus *bi*relationships between keywords and time points.

The first plan shows a clear-cut temporal trend.

CA first factorial plane: years



CA first factorial plane: years



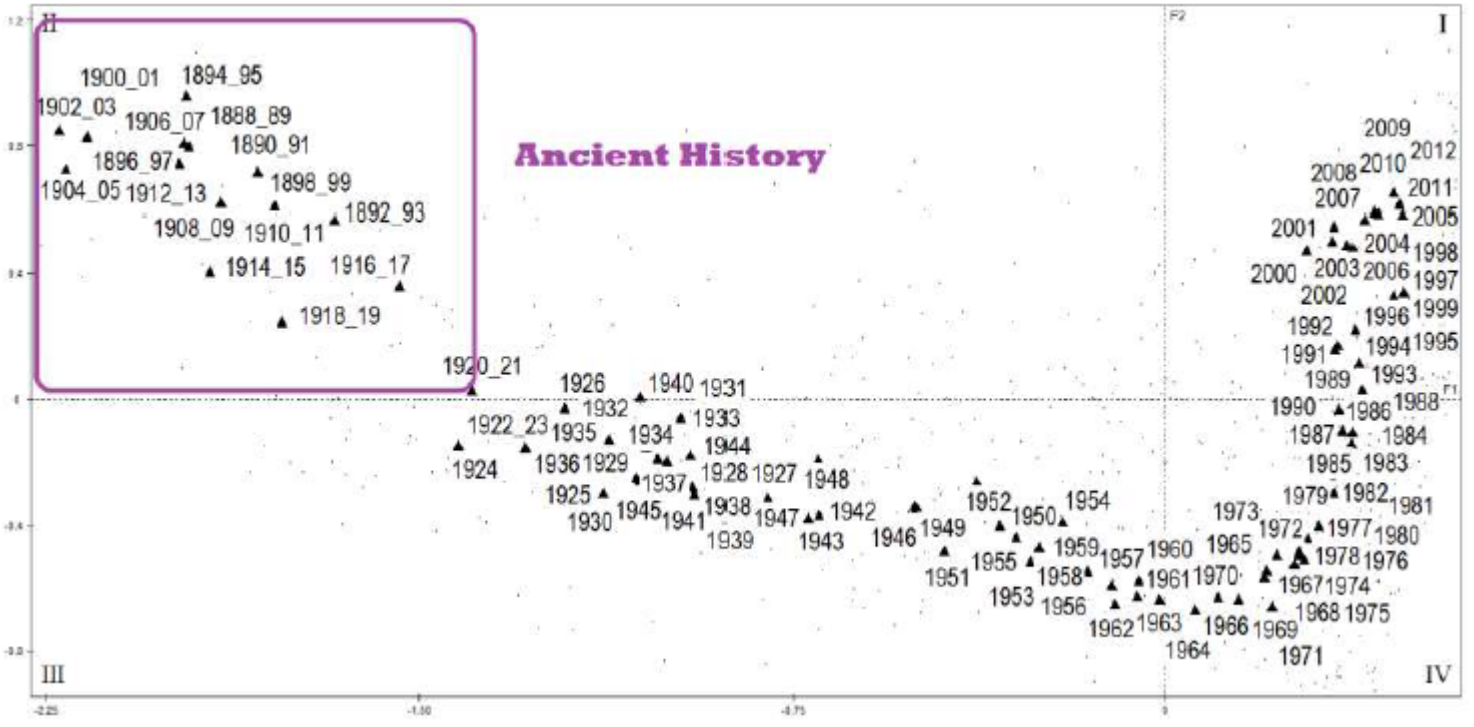
CA Results

CA shows that in titles of articles published on JASA (and predecessors) there is a progressive reduction in variability over time.

Over the years the language of Statistics has become more technical and specialized but also less varied (also in the choice of research topics).

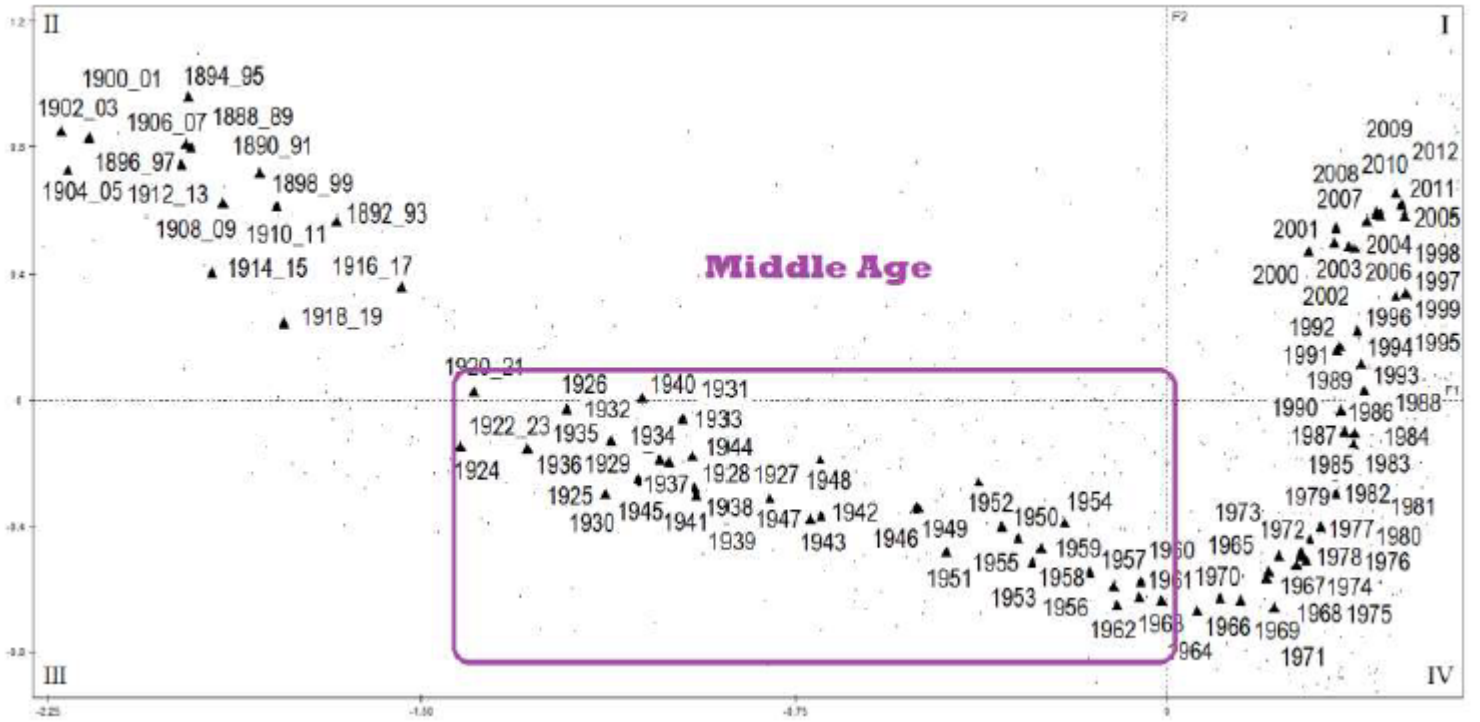
CA clearly highlighted several eras in the history of Statistics.

CA first factorial plane: IV quadrant



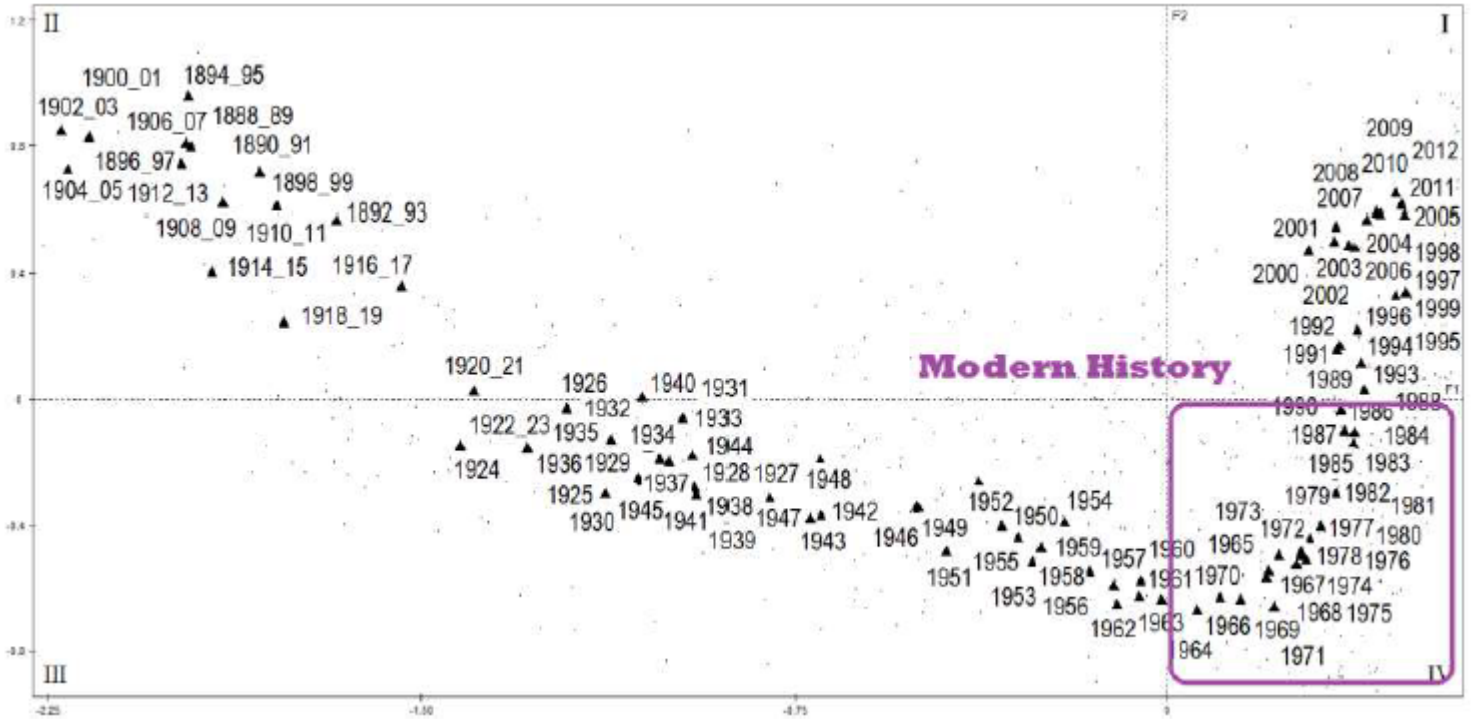
JASA corpus
CA first plane: time points

CA first factorial plane: III quadrant



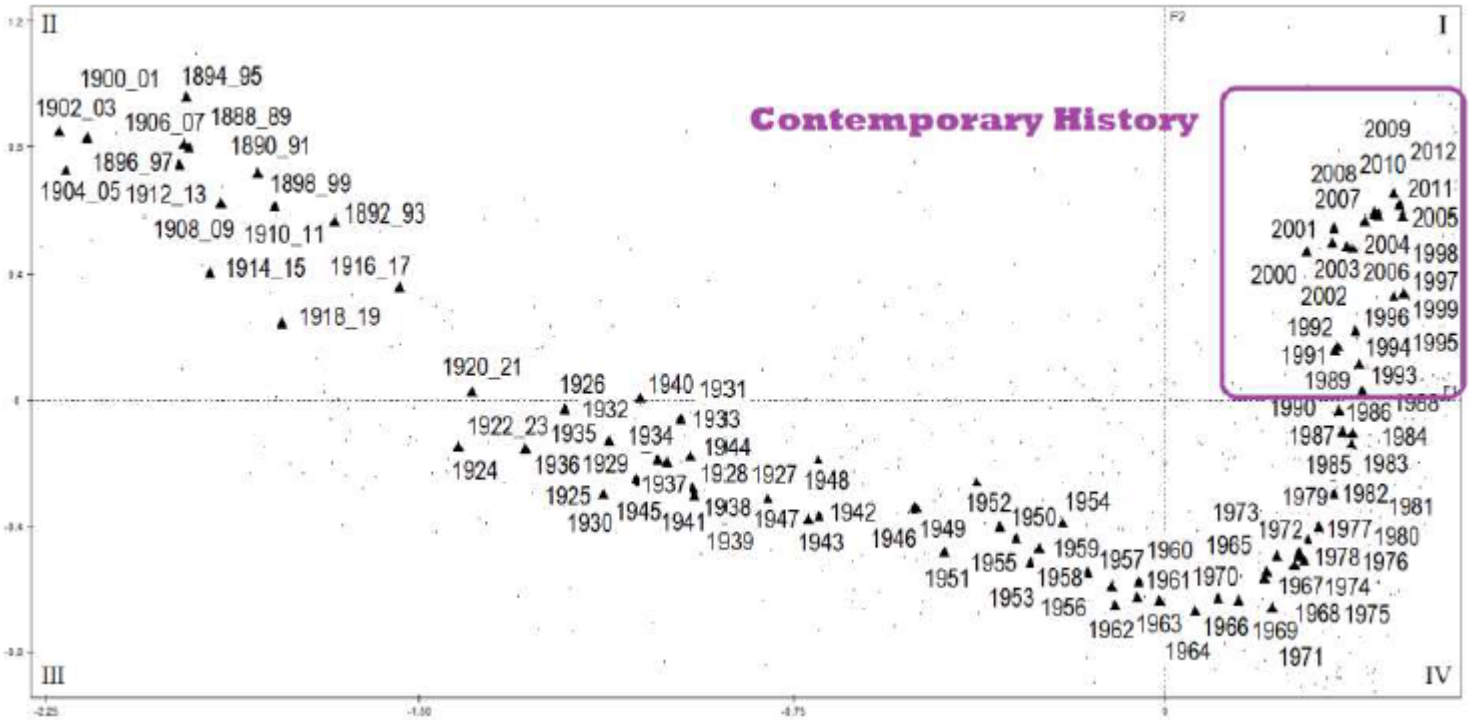
JASA corpus
CA first plane: time points

CA first factorial plane: II quadrant



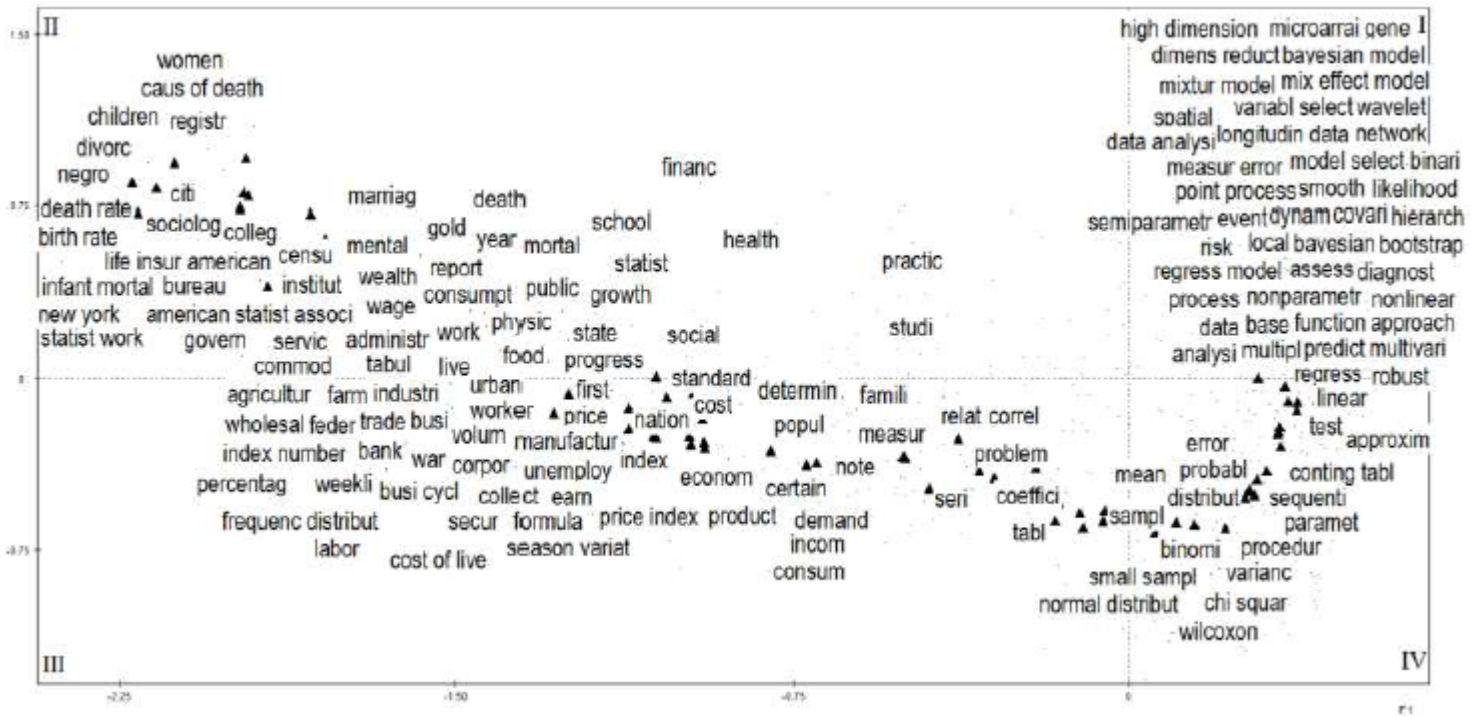
JASA corpus
CA first plane: time points

CA first factorial plane: I quadrant



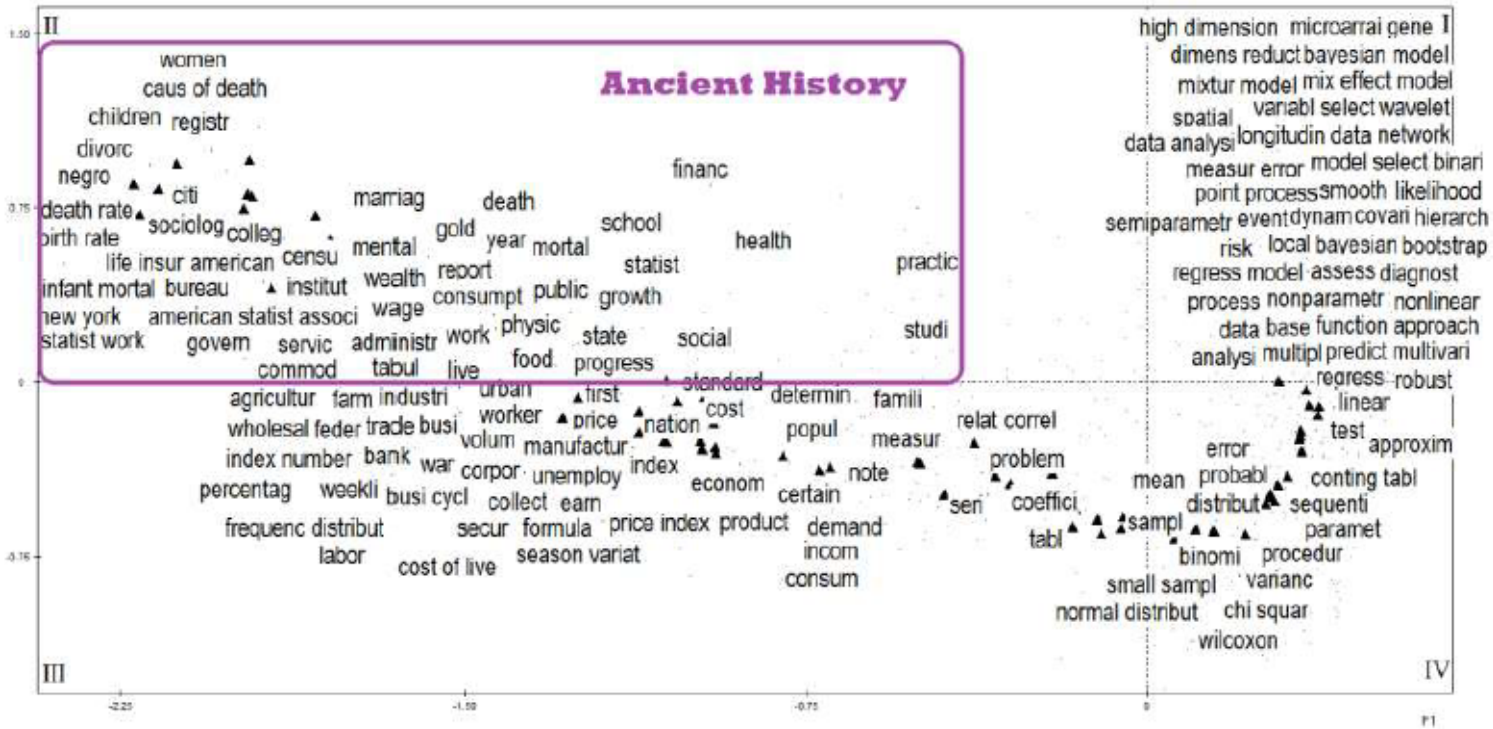
JASA corpus
CA first plane: time points

CA first factorial plane: words and years



JASA corpus
CA first plane: time points (triangles) and words

CA first factorial plane: IV quadrant



JASA corpus
CA first plane: time points (triangles) and words

Ancient history of statistics (1888-1920)

In this period Statistics is mainly Social statistics and Demography.

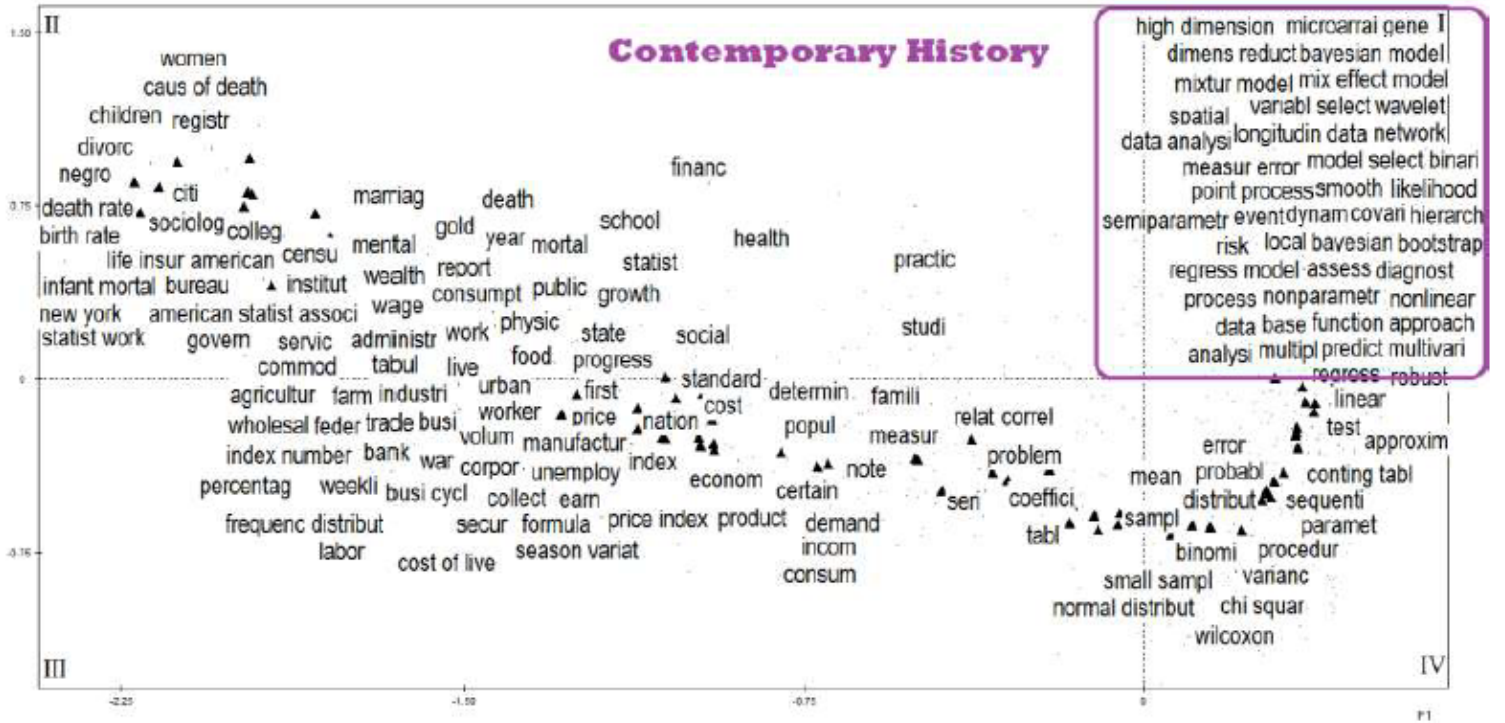
Main topics:

- living conditions of population (*children*, *negro*, *women*, *cities*, *census*, *living*),
- survival (*vital statistics*, *birth rate*, *death rate*, *cause of death*, *life insurance*, *infant mortality*),
- *health*, *food*, schooling (*school*, *college*),
- economic status (*wealth*, *wages*, *labor statistics*, *consumption*, *progress*, *finance*) amd
- social characteristics (*marriage*, *sociology*).

Statisticians of this first generation show their ability to "think big" through scientific articles dealing with the great problems of humanity.

... an era of dreamers?

CA first factorial plane: I quadrant



JASA corpus
CA first plane: time points (triangles) and words

Contemporary history (1990–2012)

New approaches:

algorithms (*bootstrap*), *smoothing* (*wavelets*), *functional* data analysis, *risk* analysis, neural *networks*, *robust* methods, *mixed effect models*, *mixture model*, *hierarchical* models, *additive*, *latent*, *point process*, *longitudinal data*, *binary* data, *dynamic covariance*),

new estimation methods: *bootstrap*, *empirical likelihood*, *model selection* (*variable selection*),

interesting contrast between schools:

frequentist versus *Bayesian*, *parametric* versus *non-parametric*, linear versus *non-linear* and early attempts to find a compromise (*semiparametric*).

The new millennium brings new challenges in

epidemiological studies, medicine, biology, environmental and health field (*surveillance*), spatial statistics, genome (*microarray*, *gene expression*)

and new technical-computational problems:

dimensionality and complexity of information to be processed (*high dimensional*, *dimension reduction*, *mixing*).

Contemporary history (1990–2012)

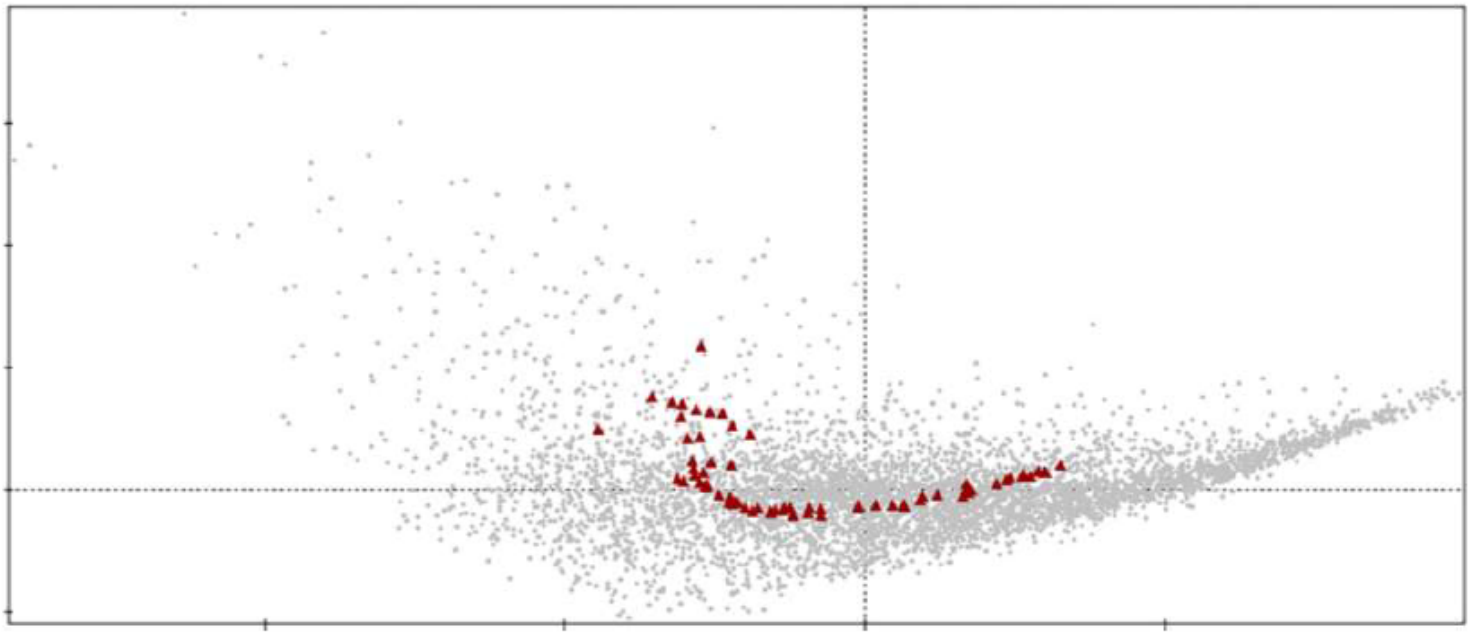
The new generation of statisticians is hyperspecialized and scientific articles dealing with the great problems of humanity have disappeared …

… an era of specialists?

Corpus of abstracts

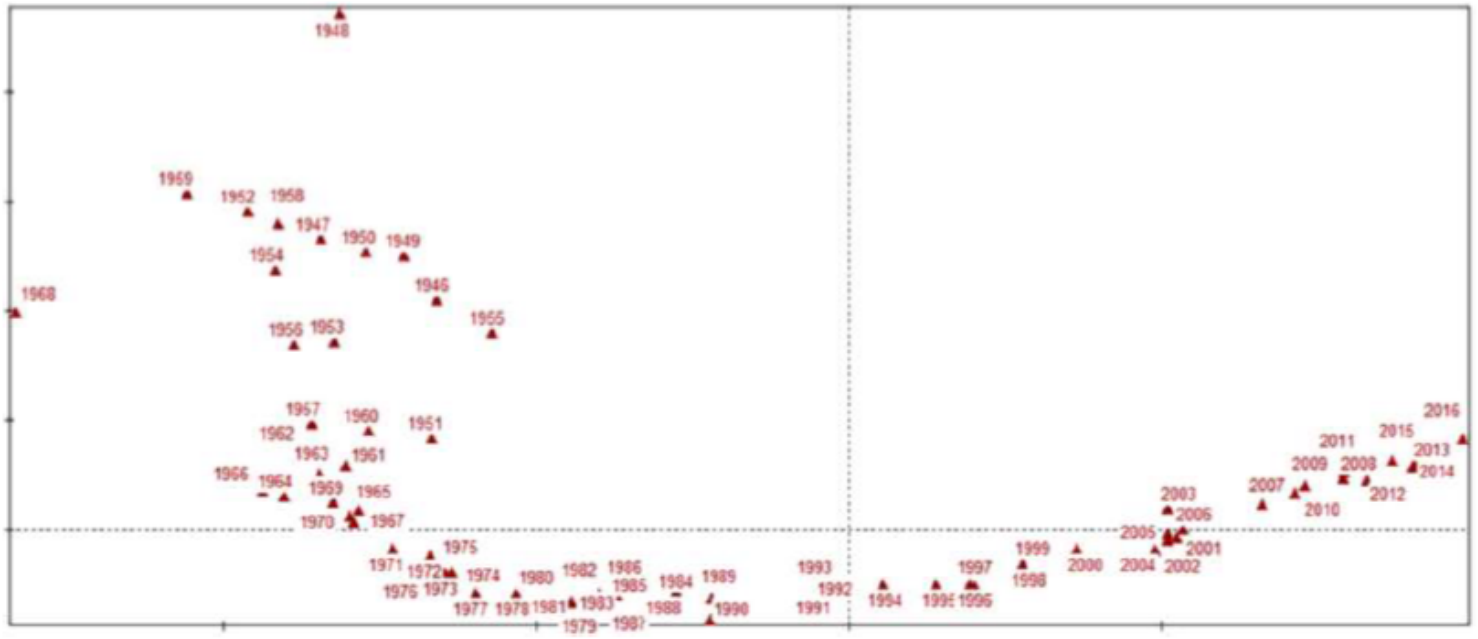
Similar results were obtained from the analysis of **abstracts** [1946-2016] (4,915 keywords × 71 years).

CA highlights a progressive specialization, hence standardization, of the language.

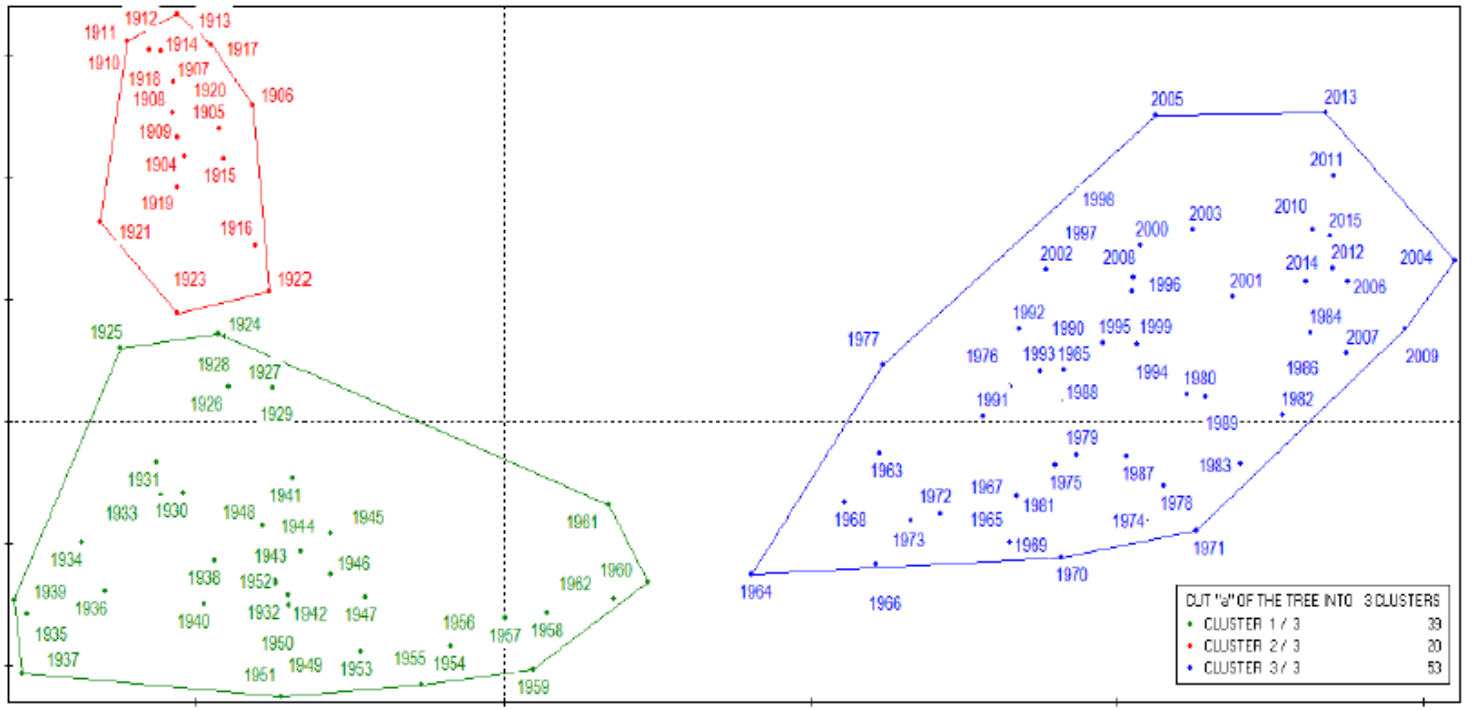


Corpus of abstracts: CA

and there appears to be a chronological trend

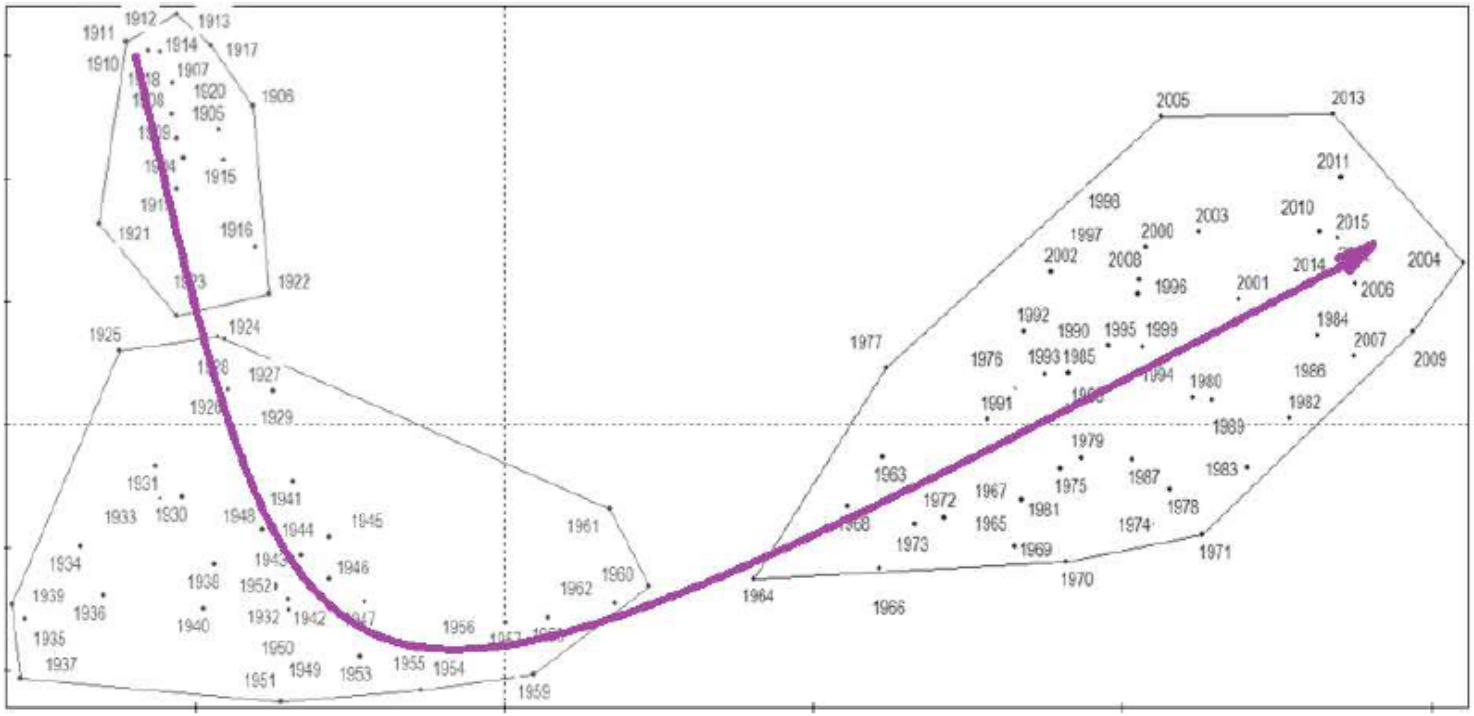


Journal of Philosophy



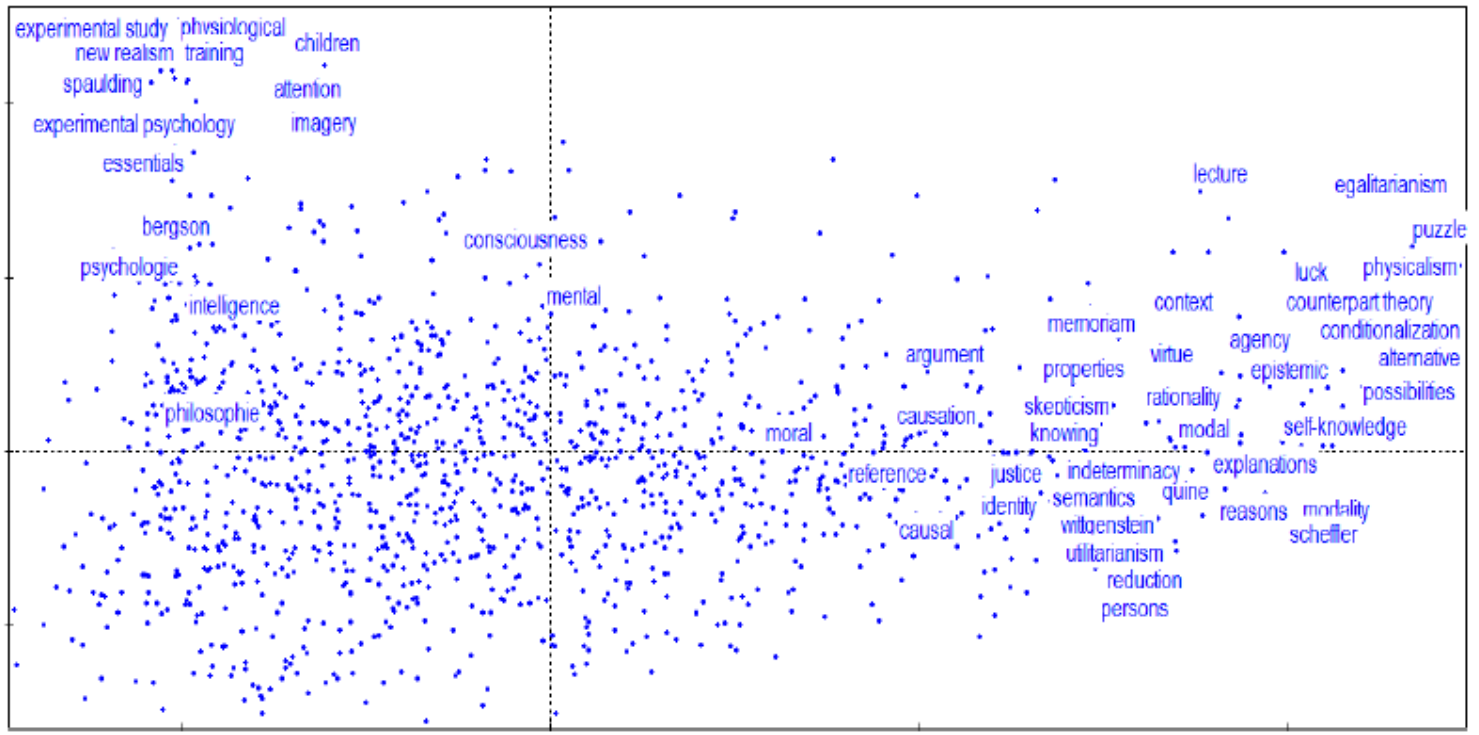
JOP: titles
Distribution of years on first factorial plane
(1388 words and syntagmas, threshold = 5)

Journal of Philosophy



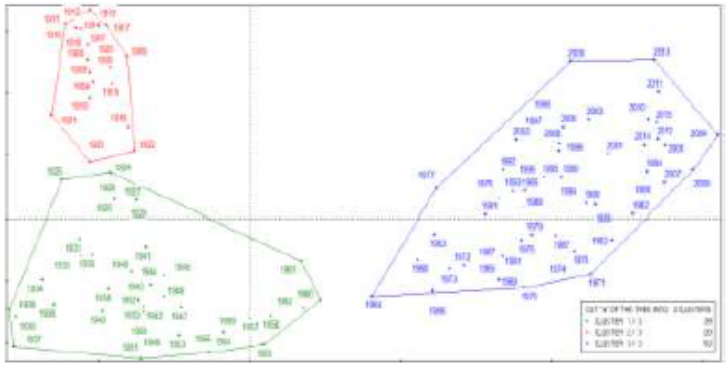
JOP: titles
Distribution of years on first factorial plane
(1388 words and syntagmas, threshold = 5)

Journal of Philosophy

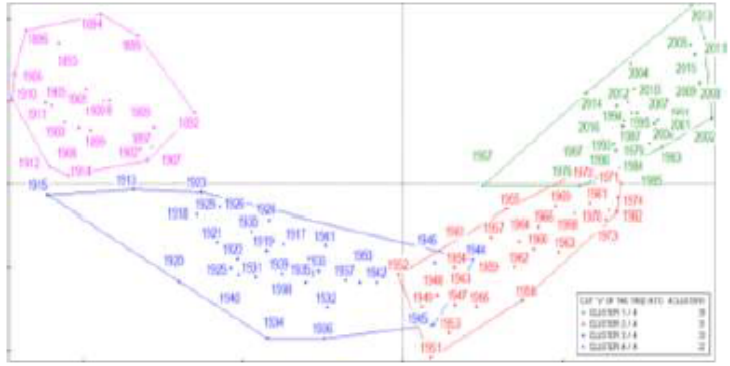


JOP: titles
Distribution of 5% of most relevant words on first factorial plane

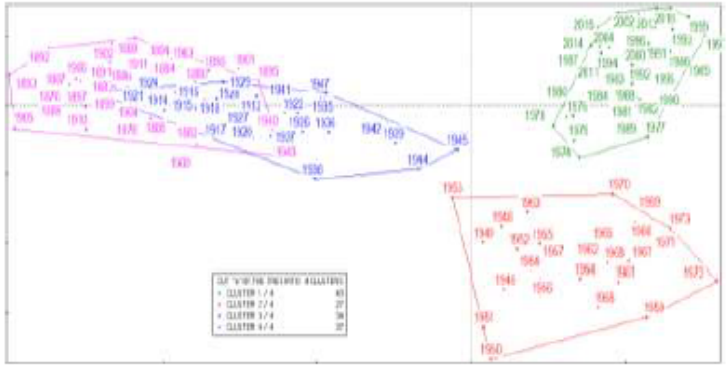
Philosophy: all the 4 corpora



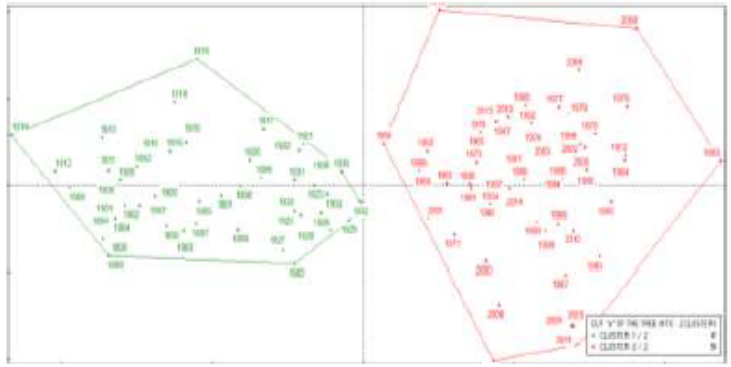
Journal of Philosophy



Philosophical Review



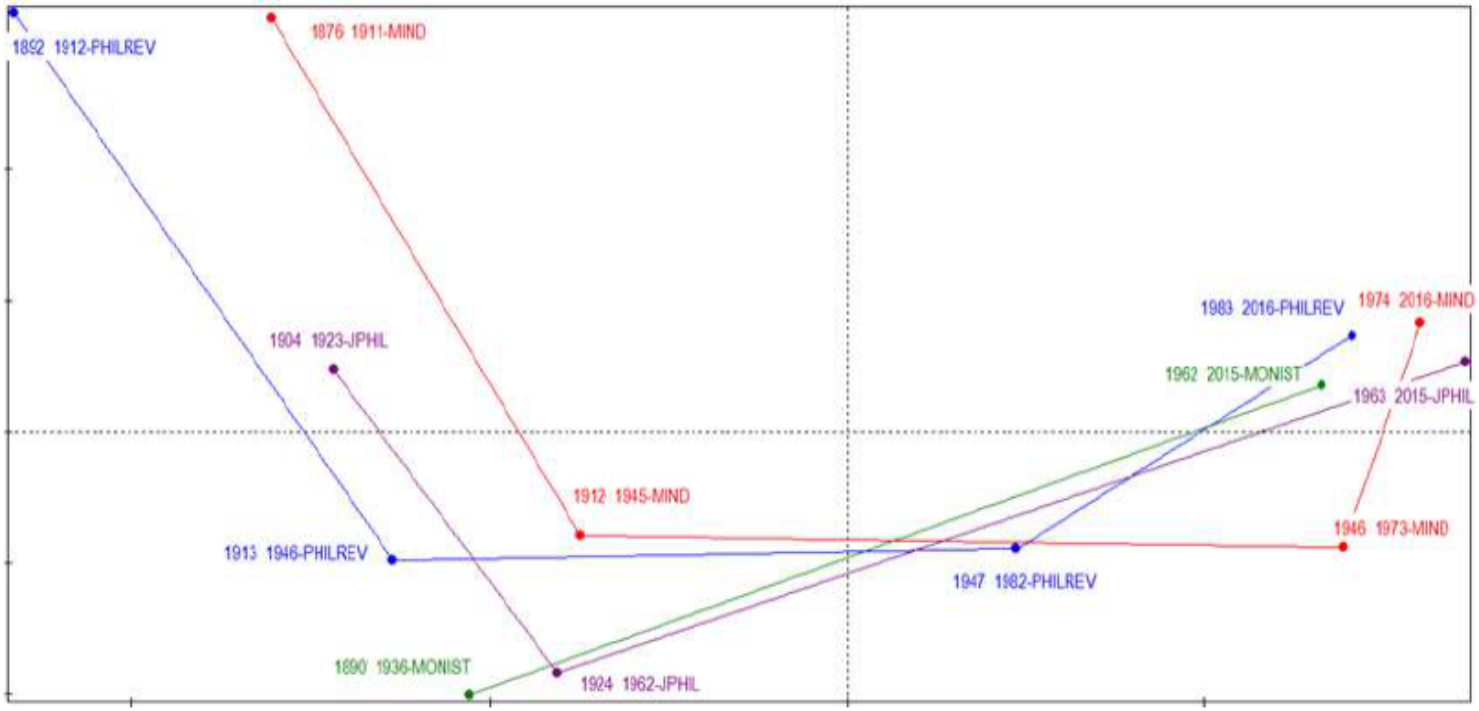
Mind



Monist

Philosophy corpora: four journals (titles)
Clusters on first factorial plane
(1388 words and syntagmas, threshold = 5)

Philosophy: all the 4 corpora



Philosophy corpora: four journals (titles)
Time trends through clusters on first factorial plane

Rivista	gruppi	periodi di tempo (classi)			
Journal of Philosophy	3	1904-1923	1924-1962	1963-2015	
Philosophical Review	4	1892-1912	1913-1946	1947-1982	1983-2016
Mind	4	1876-1911	1912-1945	1946-1973	1974-2016
Monist	2	1890-1936	1962-2015		

Limits of CA

CA is useful to

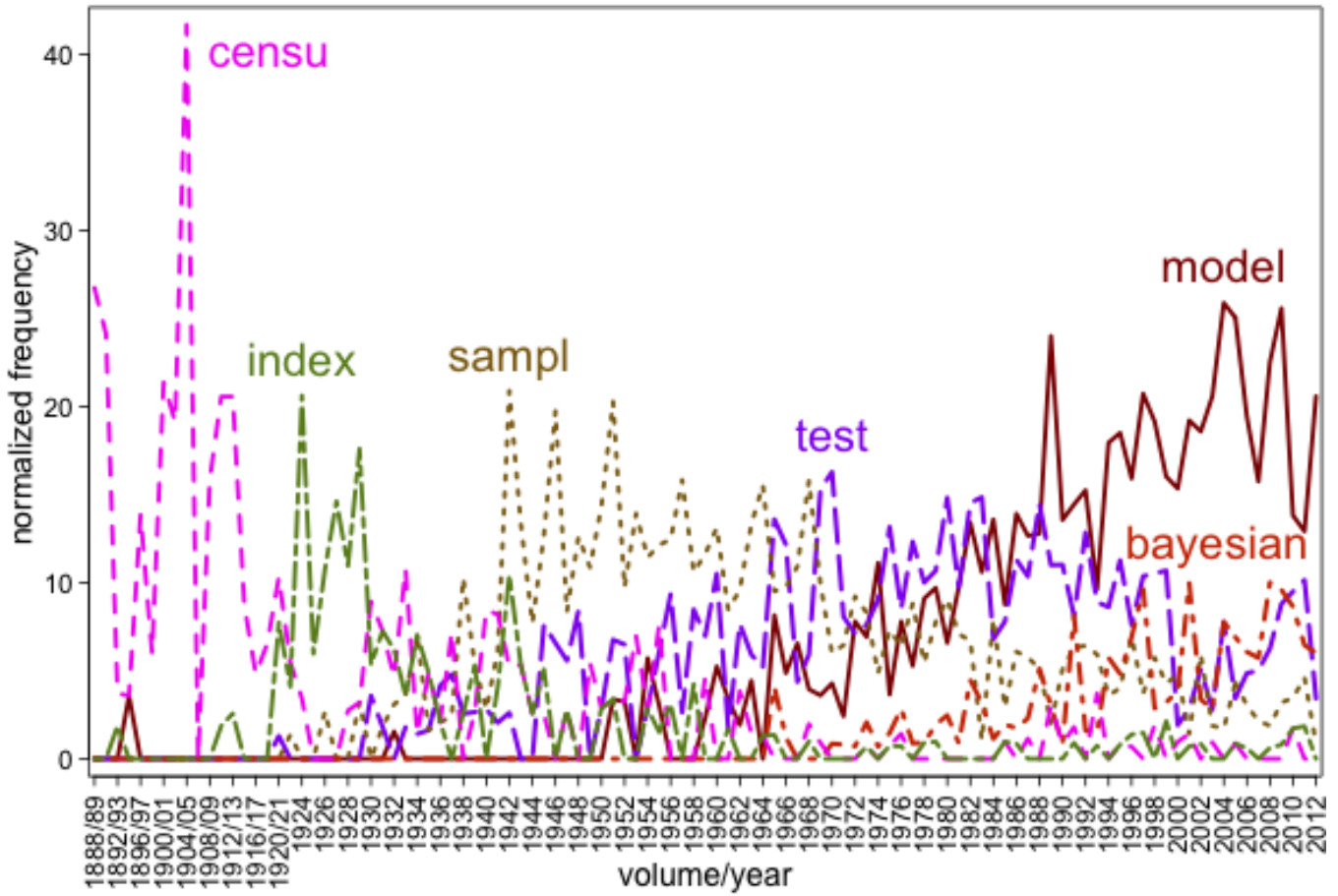
- obtain an effective graphical representation offering the possibility to see at "a glance" main trends or clusters
- reveal the existence of an apparent time dimension

CA is not useful to:

- study individual temporal patterns of keywords (life cycle)
- identify prototypical patterns (increasing, decreasing, constant, meteor-like trends)
- group keywords showing similar individual patterns over time

Words' life cycle

In our work, the temporal course of a word occurrence is viewed as a proxy of a word diffusion and vitality, i.e. of a word life-cycle.



FDA approach

We adopt a functional data analysis (**FDA**) approach under which the observations (occurrences) through time are viewed as a realization of an underlying continuous function representing the temporal development of a word.

- Data: Discretely recorded data

- FDA: is a snapshot of the function at time

where , an underlying function that generates ( ), is *smooth*.

Basis function approach for representation

For representing FD as smooth functions one method is the basis function approach where

- Basis *f*: is represented by a finite-dimensional linear combination

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

The most common basis systems are *monomial*, *Fourier* (convenient when functional data are cyclical), *B-spline* basis functions.

There are other useful basis systems to model functional data, for example, *wavelets*, *exponential* and *power bases*.

Curve clustering

After *smoothing* / *filtering* FD, estimated curves can be grouped by means of a

- model-based
- distance-based

curve clustering (CC).

Example with wavelets

A wavelet-based filtering was successful in recognizing the typical bumpy peak-and-valley trend of word trajectories (due to data sparsity).

Moreover, is more adapted to high dimensional data than spline decomposition thanks to computational efficiency.

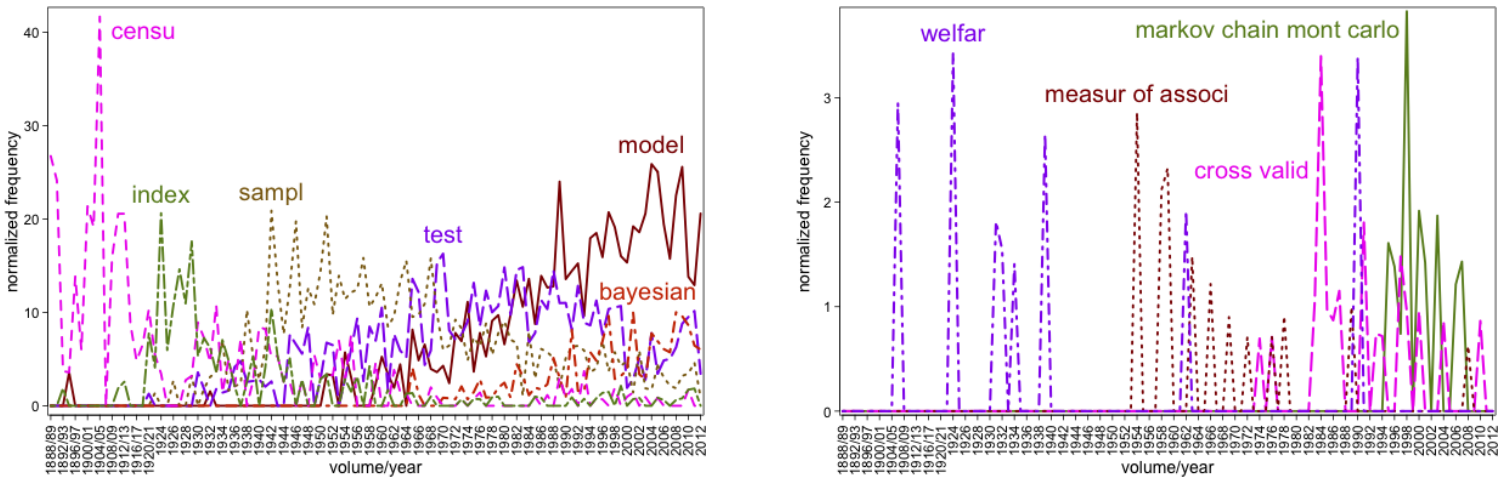
Within a model-based curve clustering approach, we applied a **wavelet-based functional clustering mixed model** (FCMM), proper to account for:

- irregular curves,
- high dimensional CC,
- high inter-(word) variability

Typical characteristic of a chronological corpus

If we consider the words x documents table by row, a typical feature of a word trajectory over time is a sharp peak-and-valley trend, mainly due to the sparsity affecting frequency data of the most of the words.

many cells of the contingency table have small counts or are empty



Temporal trajectories for some of the most-frequent (left) and least-frequent (right) keywords within the corpus of titles, 1988–2012.

Typical characteristic of a chronological corpus

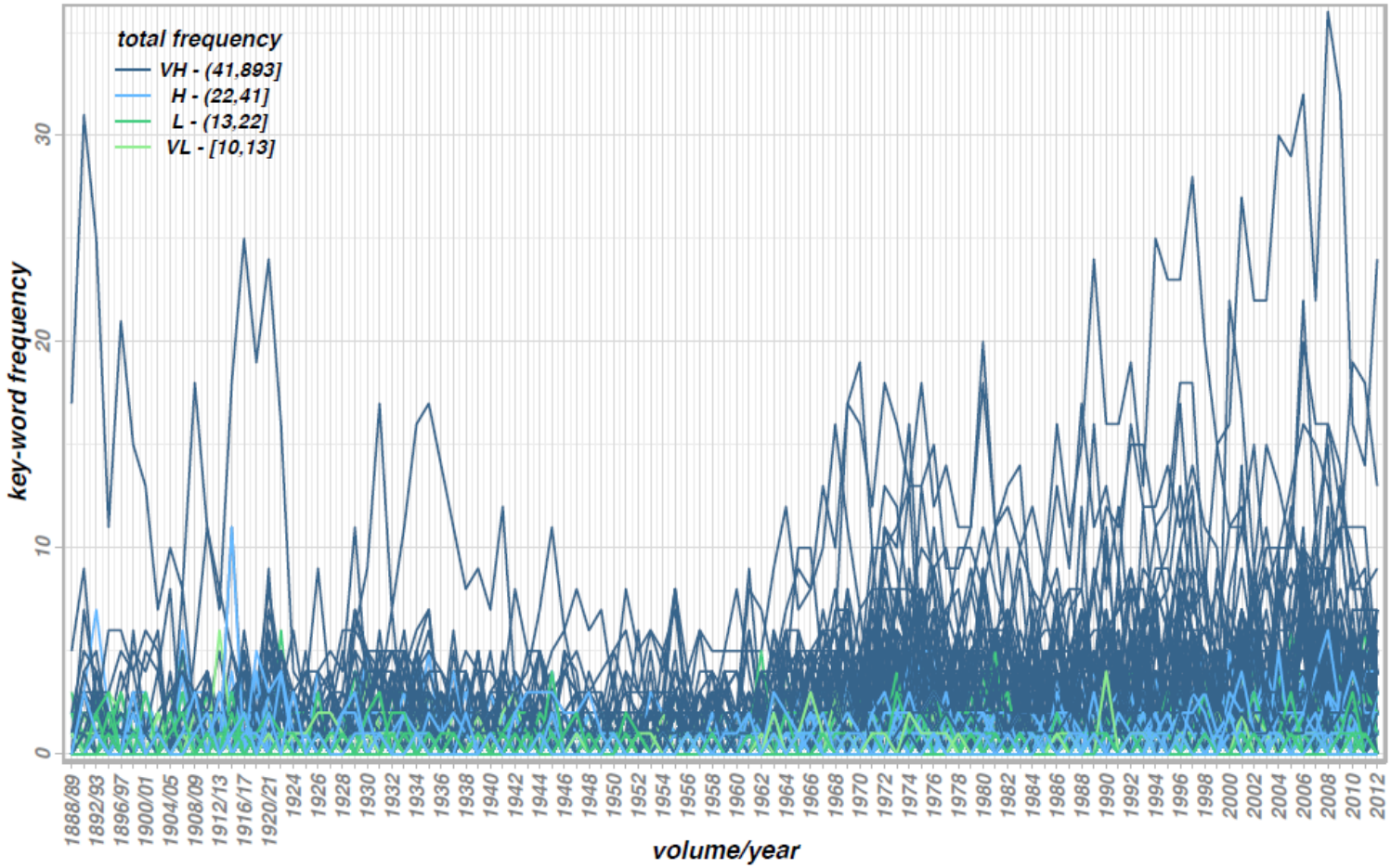
On the other hand, if we look at data by column they appear strongly asymmetrical, in particular for the marked disparity of frequency classes between the most popular words and the rest of the others.

The frequency spectrum shows the typical frequency 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.

A corpus, no matter how large, is still a sample far from saturating the entire population, or, otherwise said, it falls in the so called *large number rare events* (**LNRE**) zone where the chance of encountering a new word (indicated by the presence of hapax legomena) is never null.

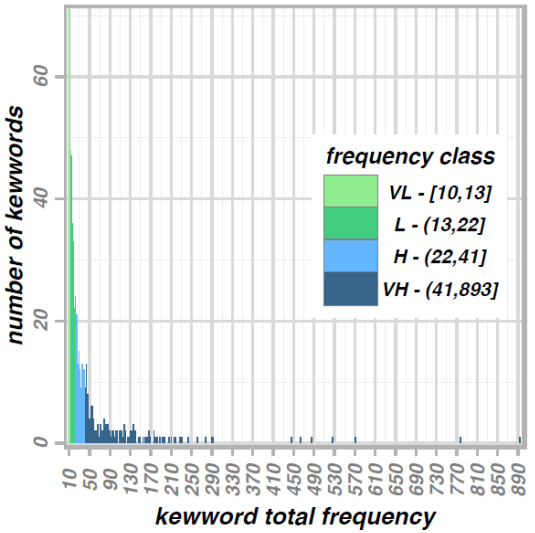
A direct consequence is the sparsity mentioned before.

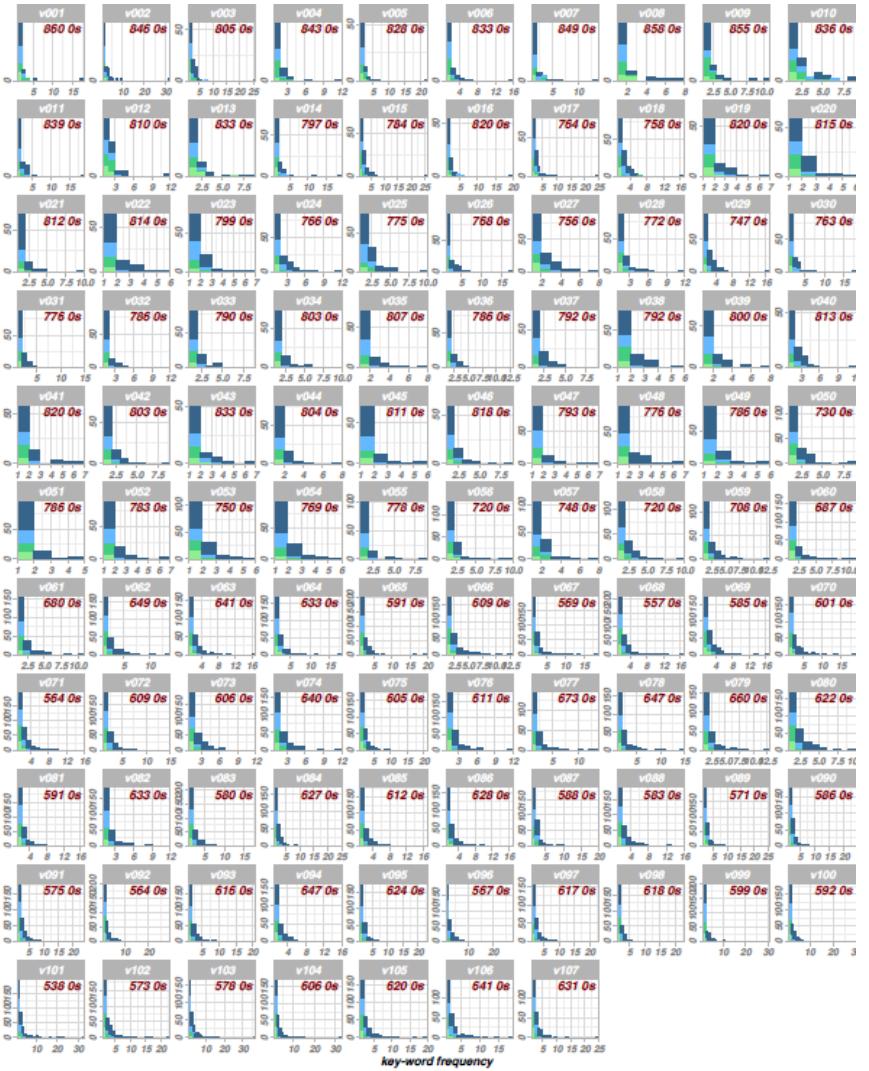
Word trajectories



Keyword trajectories: non-normalized data

Frequency spectrum

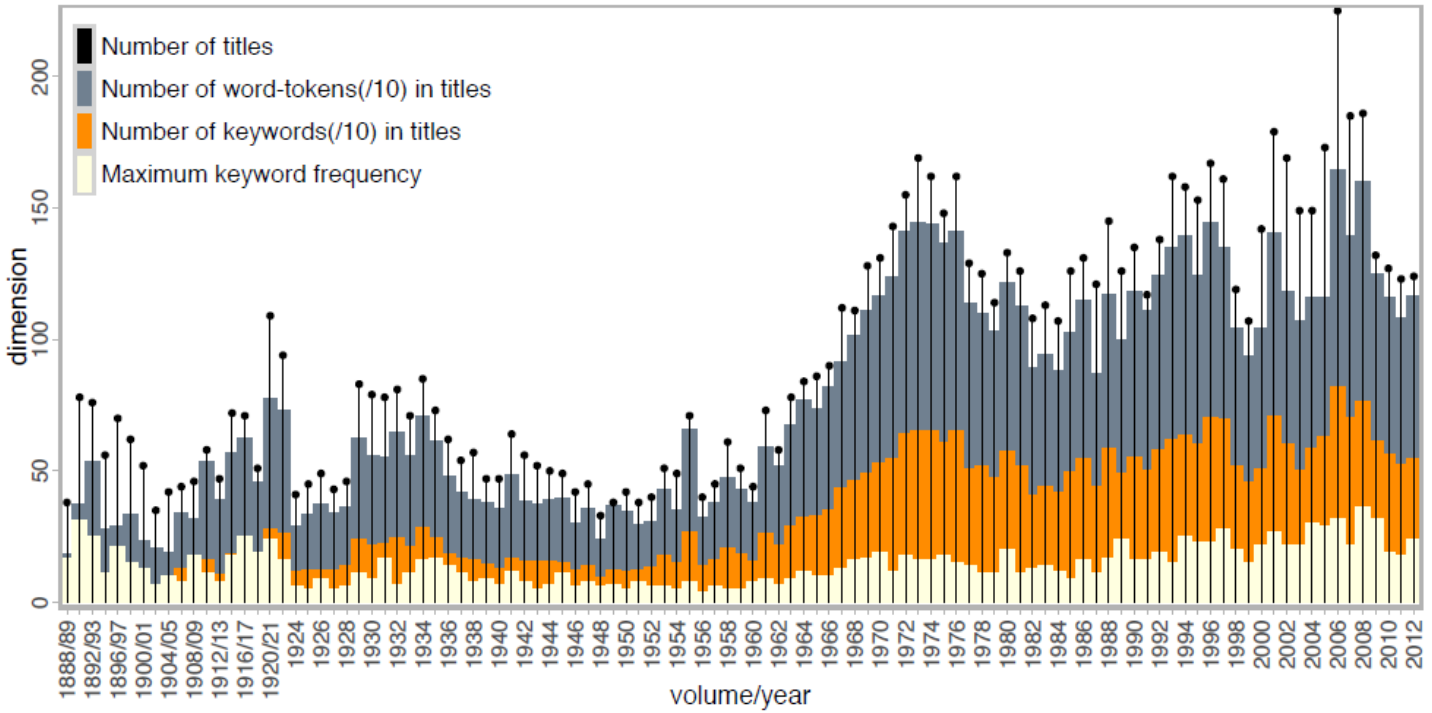




Frequency spectrum of the entire corpus (left) and divided by volume/year.

Varying signal across time

Lastly, the size of time-point subcorpora (number of documents and their size in word-tokens) may vary greatly over time. (Again, this reinforces the data sparsity.)



Subcorpora dimension: for each volume, number of titles/articles, total number of tokens in titles/10, total number of key-word tokens (column sum) in titles/10, maximum key-word frequency.

Question

What does represent the frequency spectrum of the entire corpus in the correspondence analysis (CA) perspective?

And the subcorpora dimension across time?

Data transformation in light of the clustering objs

In studying these functions data transformation (**normalization**) is essential, especially in light of the clustering objectives.

Normalization plan

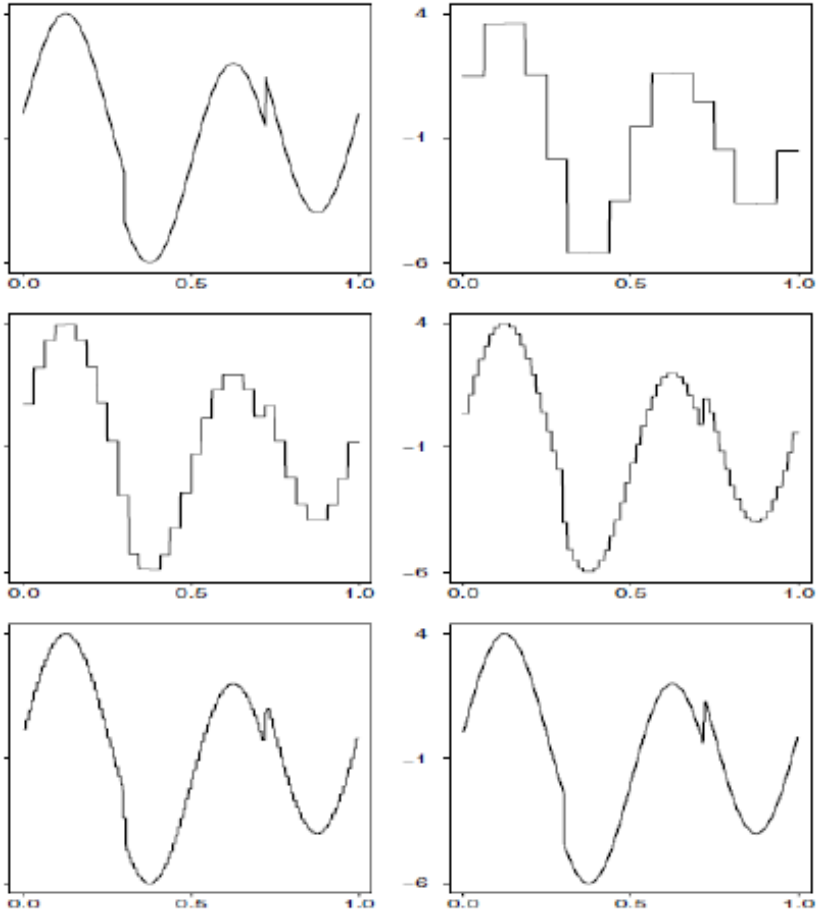
	normalized by column / normalized by row	(corpus logic) subcorpus #titles	(corpus logic) subcorpus #tokens	("table" logic) column sum ($\sqrt{\cdot}$)	("table" logic) column max. freq.	(LNRE) dynamic density	
Strong asymmetry	row sum	d	d	d_1 (χ^2)	d	d_{1b}	r_1
	z-score by row	d	d_2	d	d	d	r_2
	maximum 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	

in assessing two curves as similar, we consider

- height (**word popularity**)
- timing (**synchrony**)
- height and timing jointly (**double normalization**)

Wavelet-based filtering

Filtering step (in which FD are represented as smooth functions by a basis-expansion method) based on wavelets.



FCMM (in brief)

- Let be the individual curve of word , with from to , observed at equally spaced time points , with for some integer , and
- suppose that individuals are spread among unknown clusters of prior size , with from to .

Then, a FCMM takes the form

where

- is the **functional fixed effect** that characterizes cluster
- is a **random functional effect** introduced for handling inter-individual random deviation from the cluster average curve and modelled as centered Gaussian process independent from
- , a random measurement error modelled as a zero-mean Gaussian process.

FCMM (in brief)

Using a discrete wavelet expansion, the equation above defined in the functional domain can be written in matrix notation as

where is the -dimensional vector of *wavelet coefficients* and is the orthogonal matrix that contains the discretized scaling and wavelet functions.

The DWT of corresponds to the multiplication of by

which resumes to a linear mixed-effect model in the wavelet coefficients domain such that

Furthermore, this class of FCMMs allows random effect variance to vary over wavelet positions and/or groups.

Model selection

For estimation we resort to the EM-algorithm for maximum likelihood estimation provided by the R package curvclust `curvclust` which is dedicated to model-based curve clustering and was originally thought for microarray-type data.

Two criteria for model selection have been used:

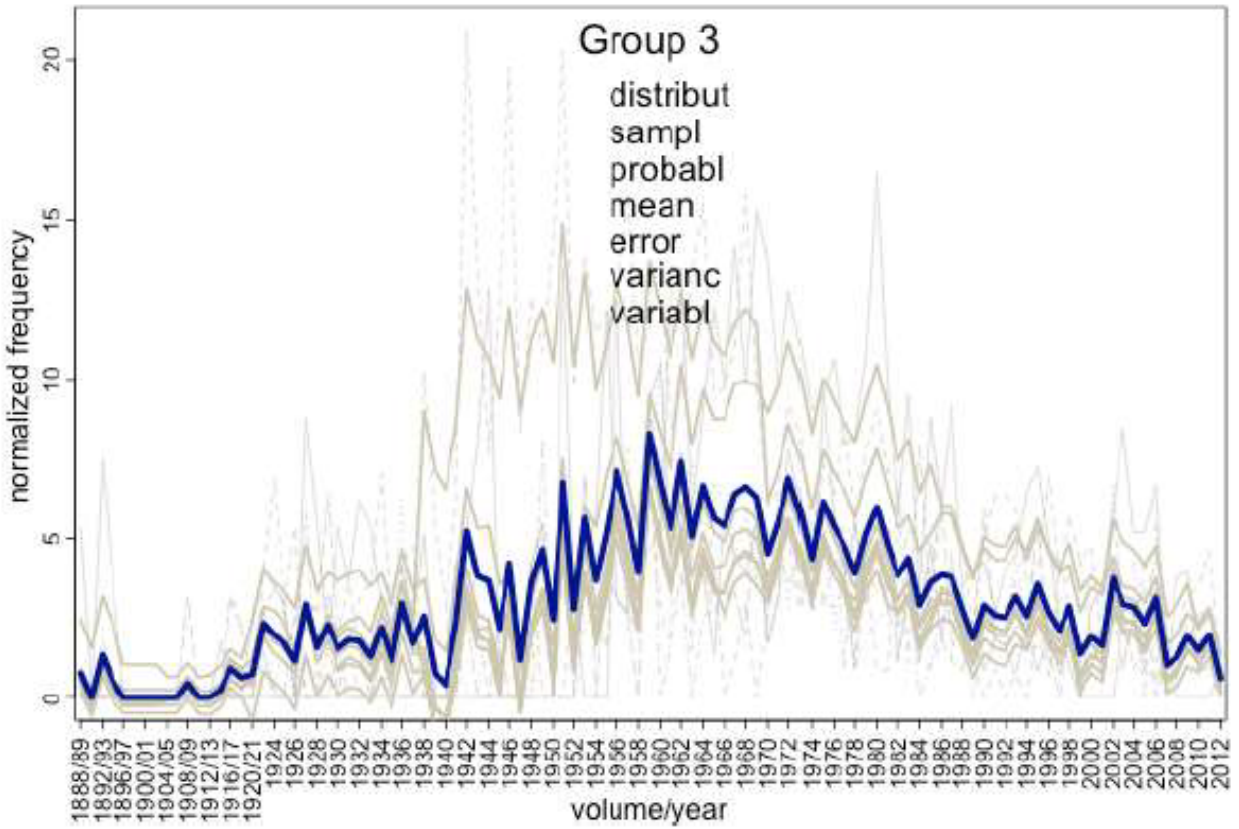
- the Bayesian Information Criterion (BIC)
- Integrated Classification Likelihood criterion (ICL).

> The selected model was a FCMM with group-specific random effect variance and 13 clusters (normalization).

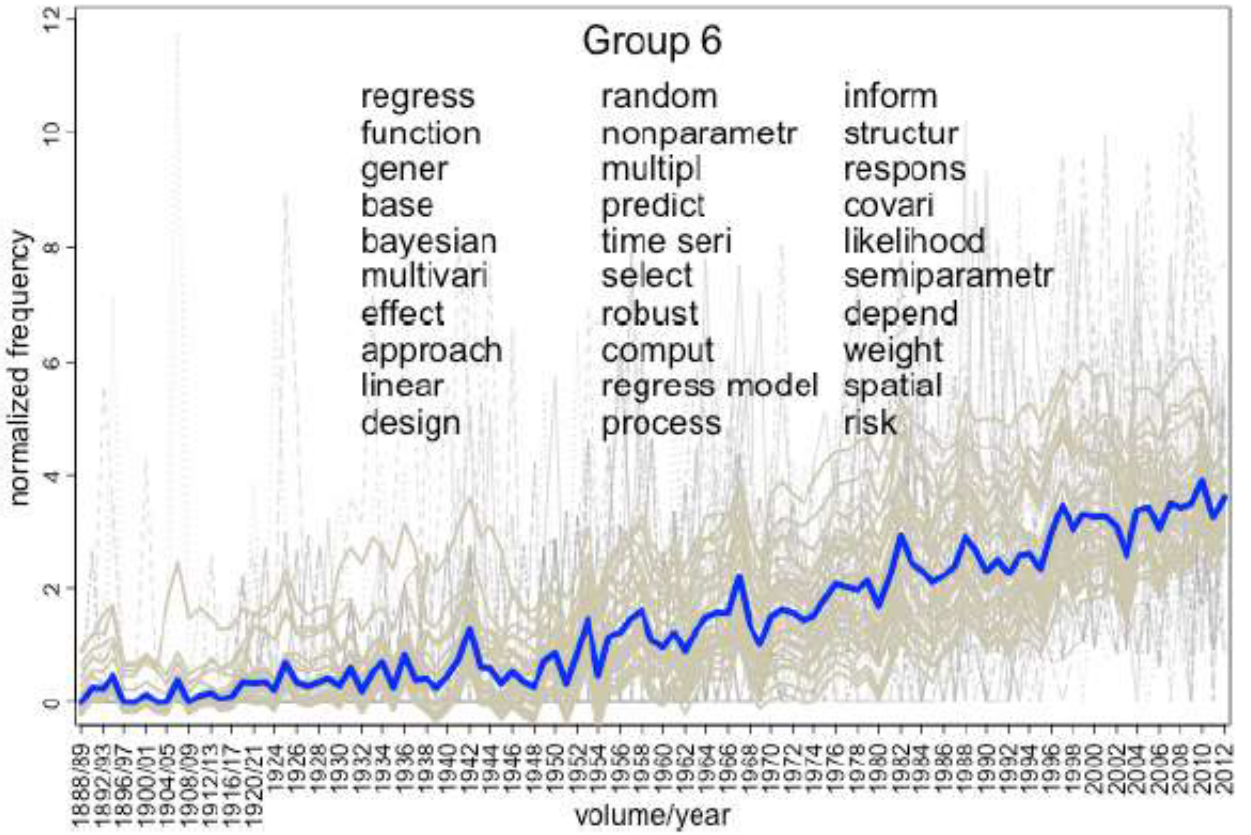
Two examples are the following.

The group-specific functional fixed effect (bold line) and functional random effects (beige lines) are illustrated.

Cluster 3 contains all the basic elements of Statistics (*distribution*, *sample*, *probability*, *mean*, *error*, *variance*, *variable*) and shows that Statistic has established itself as discipline in the second half of the 20th century. This is the golden age cluster of Statistics in its most classic and traditional formulation which dates from the 1950-1980.



Cluster 6 shows instead the birth of modern statistics starting from the 1950s, with the expansion of new approaches to estimation, models and areas of application then have become mainstream (*Bayesian*, *non-parametric*, *likelihood*, *semi-parametric*, *selection*, *robust*, *computation*, *regression*, *functional*, *generalized*, *multivariate*, *multiple*, *process*, *covariance*, *dependence*, *weighted*, *time series*, *spatial*, *risk*).



Example with splines

A B-spline filtering was chosen to recognize continuous shapes, more easily interpretable.

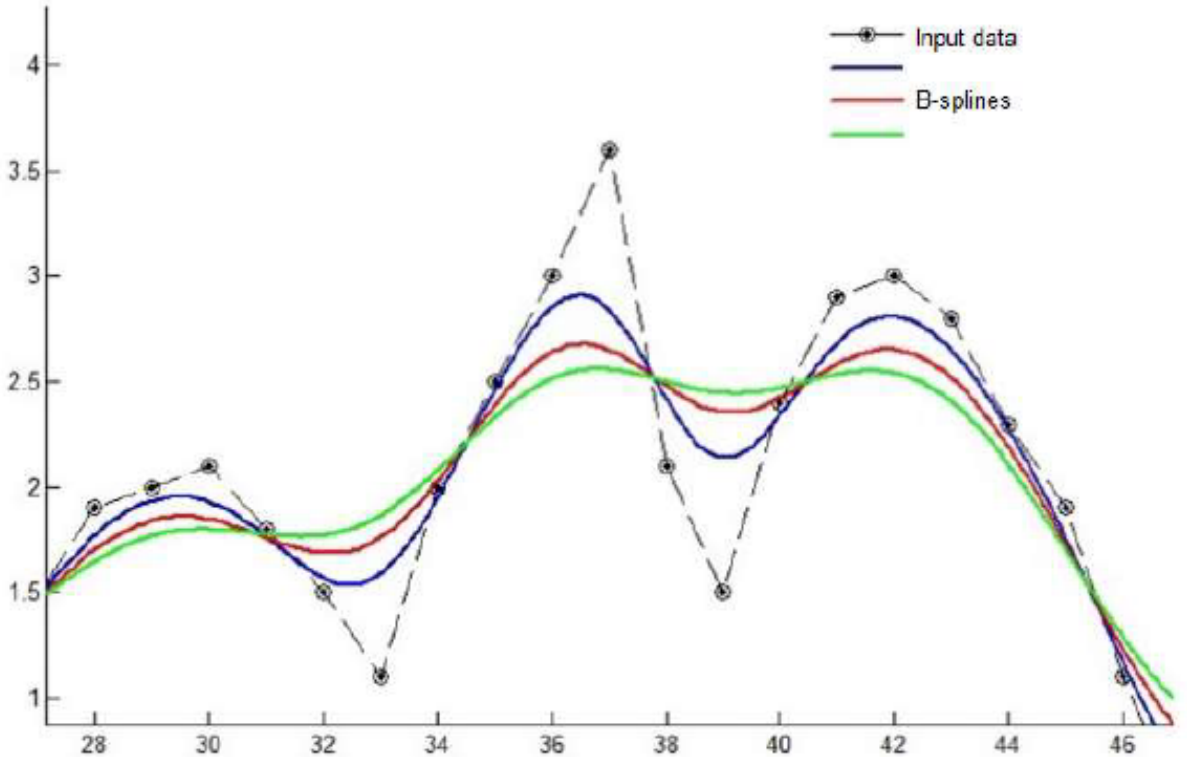
Moreover, we set up a distance-based approach to curve clustering to produce an explorative and mostly automated technique, useful for interdisciplinary research groups (+ an R pkg).

- different types of **data normalization**
- different types of distances
- cluster number selection based on pooling a large basket of quality criteria.

Here we present a -cluster partition with (-like) normalization.

Spline-based filtering

Filtering step (in which FD are represented as smooth functions by a basis-expansion method) based on wavelets.



Spline-based filtering

Knots are placed at each time-point (107).

By the **roughness penalty** approach, we smoothed the data by varying

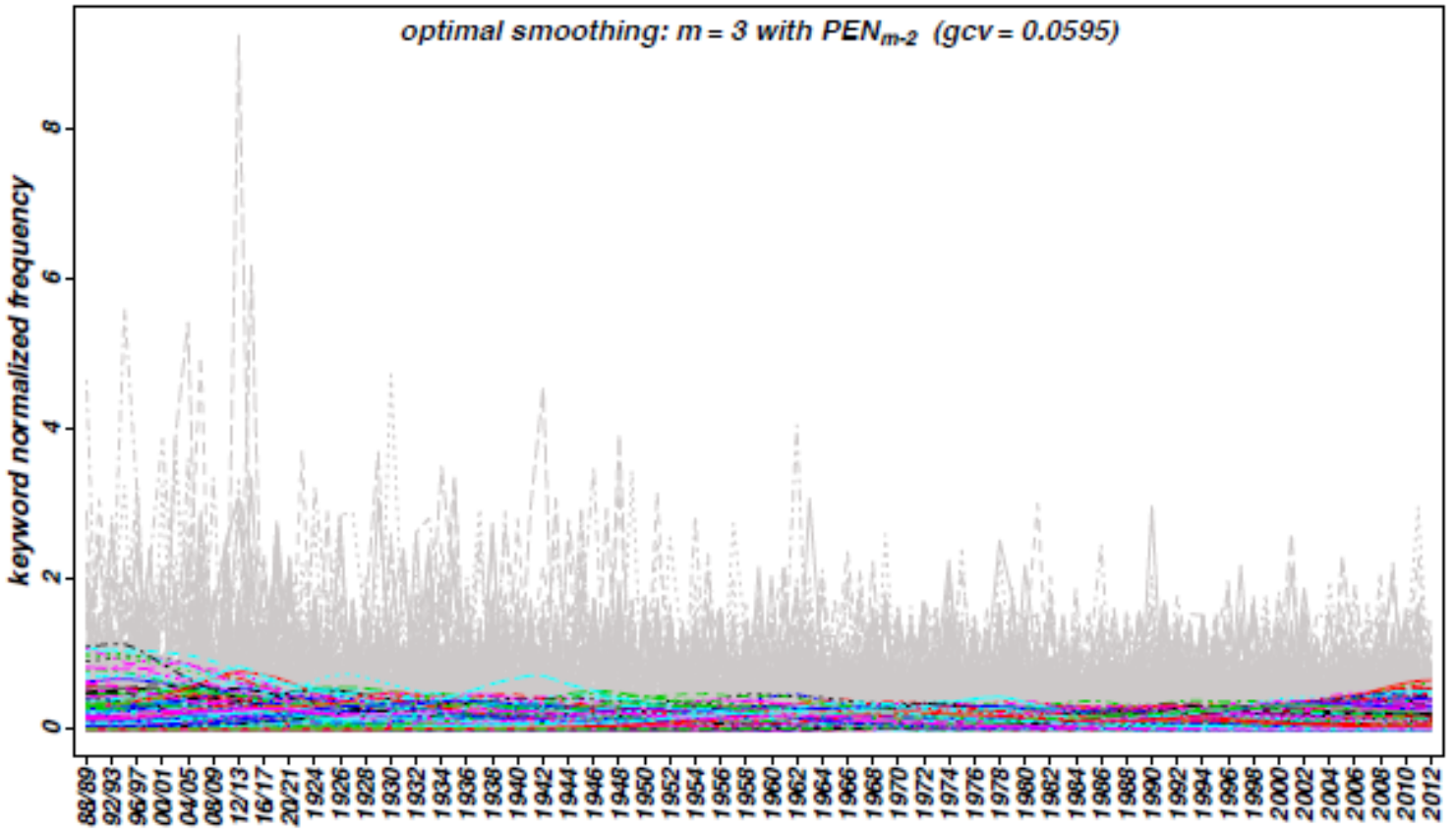
- spline order (from 1 to 8)
- roughness penalty:
  - ;
  - for each ;
  - for each ;
  - for every .
- smoothing parameter over an opportune range of values (from to).

The smoothing selection, according to the *generalized cross validation* (GCV) criterion, led to

(-like) normalization:

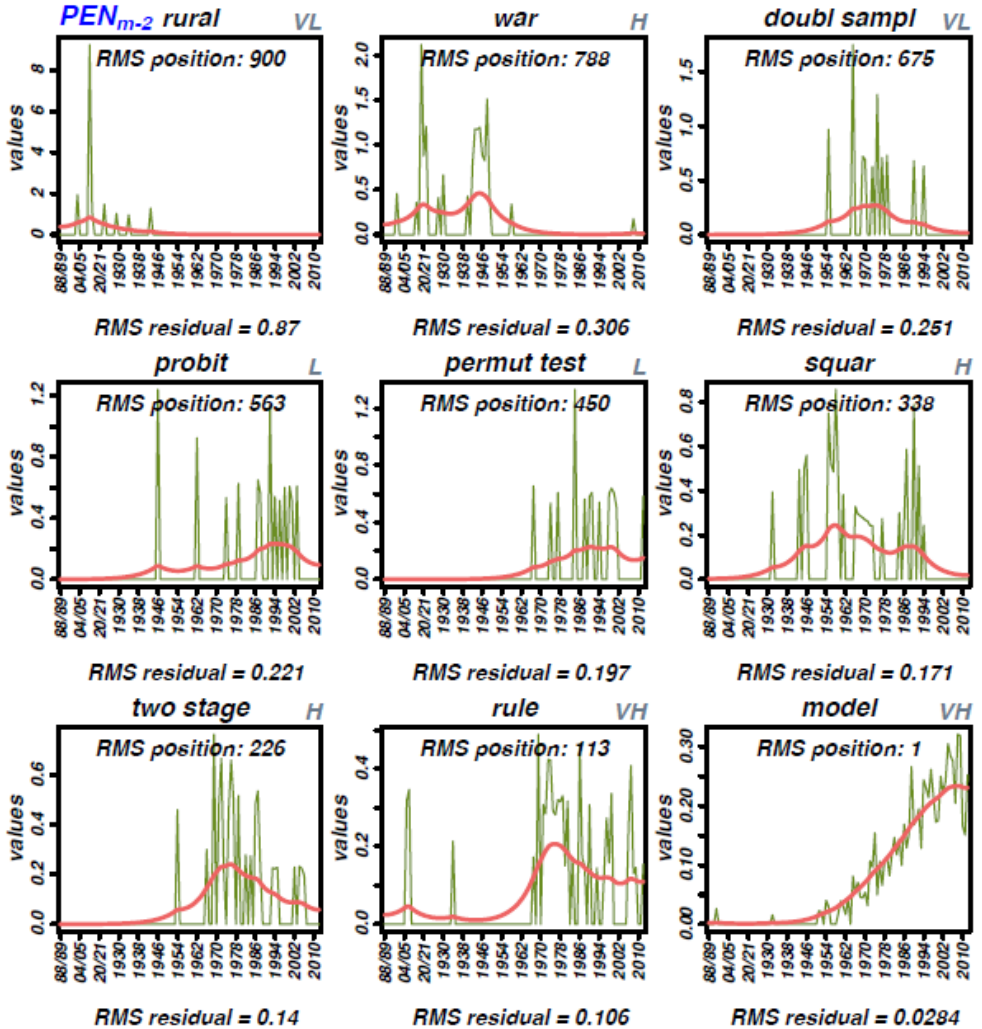
with a roughness penalty and ( )]

Optimal smoothing



Curves from fitting a order smoothing spline, with , to normalized data.

Optimal smoothing: some words



A selection of fitted curves according to worst RMS residual order.

Distance-based curve clustering

Curves are partitioned by means of

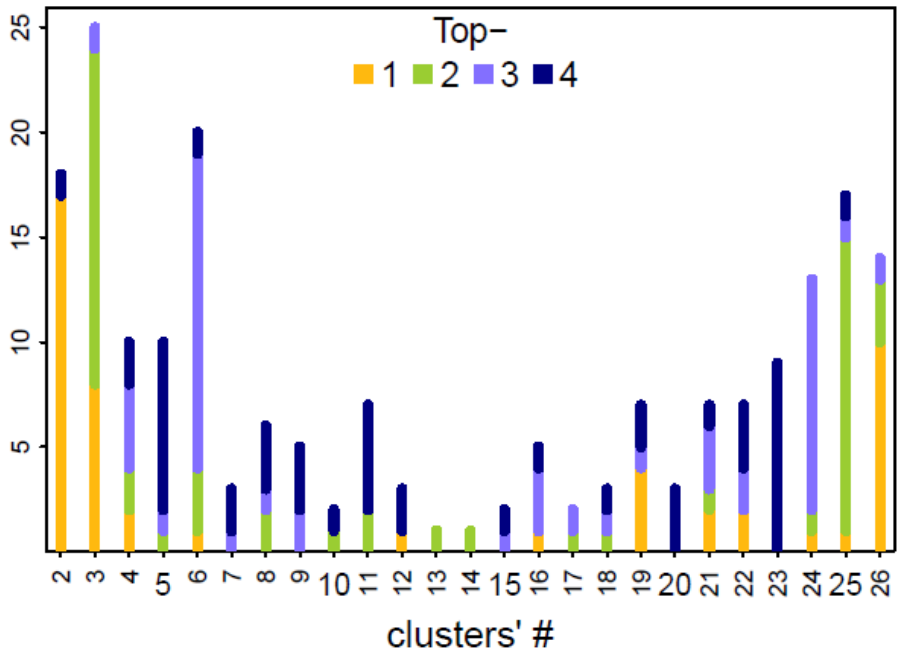
- **K-means** algorithm
- several types of distance between trajectories can be chosen: **euclidean**, Manhattan, correlation-based dissimilarity ( ), adaptive dissimilarity index (with either euclidean distance or Dynamic Time Warping).
- for each cluster number (from 2 to 26), **20 re-runs** from different initial configurations set through the k-means++ seeding method.
- the best partition/number(s) of clusters is/are identified by **pooling** several clustering quality criteria ( ) (the first time without integrating with subject-matter considerations).

After discarding the top-rated solutions (-cluster partition),

(-like) normalization:

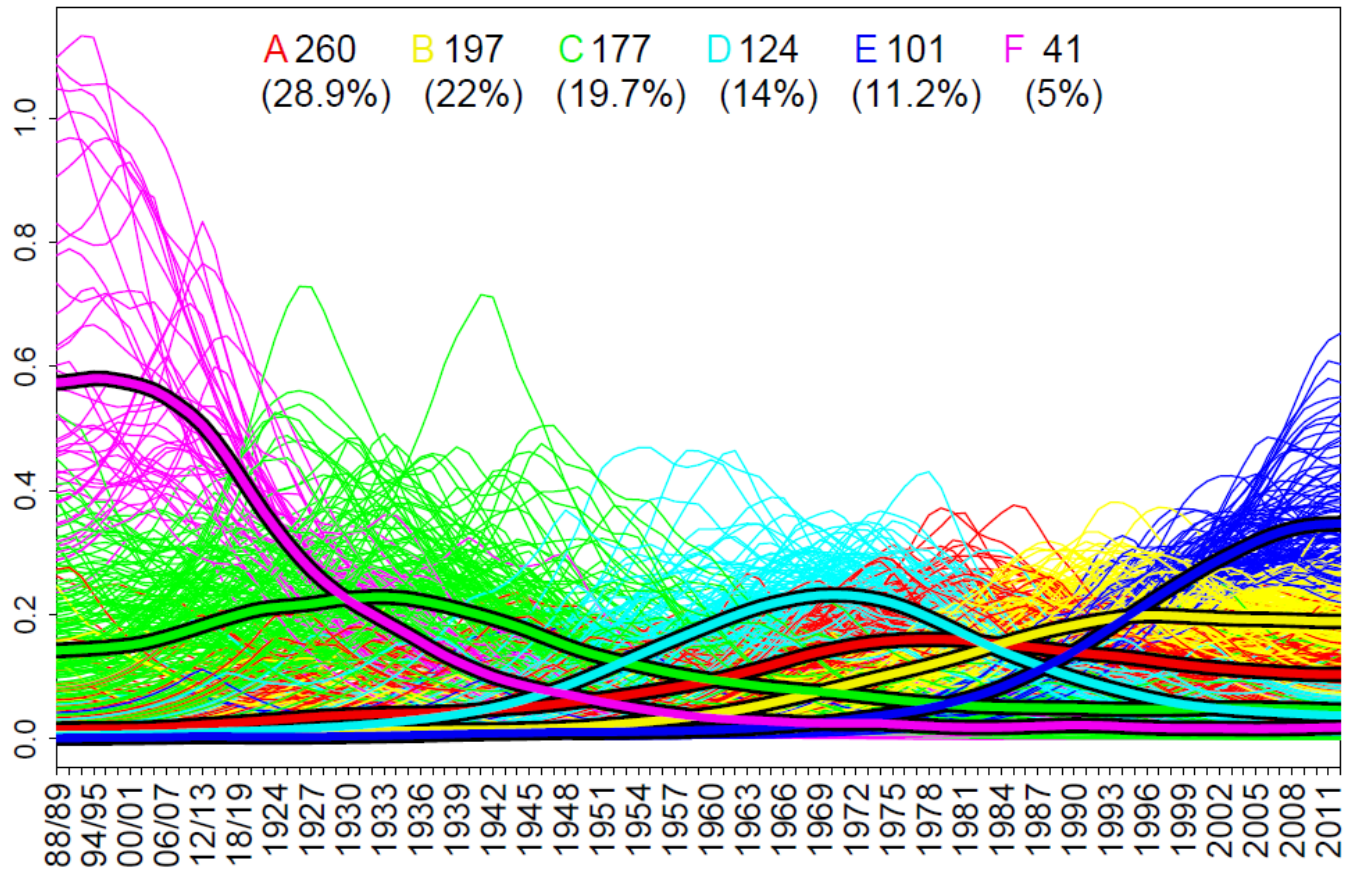
1st: 6 (2nd: , 3rd: *ex-aequo*) groups

Cluster number selection

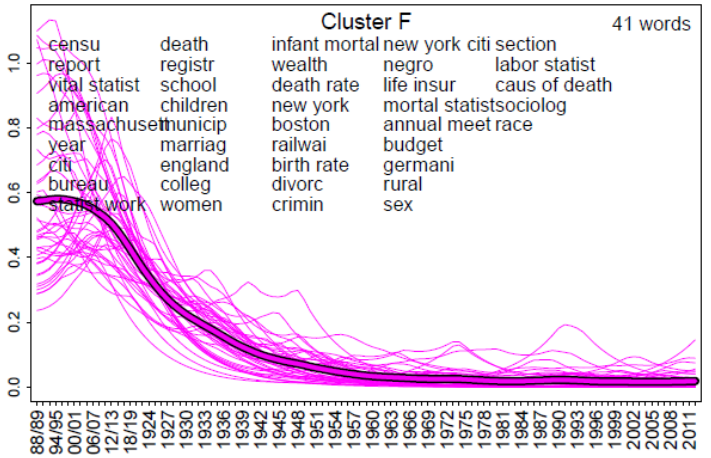


Cluster number selection for normalized data

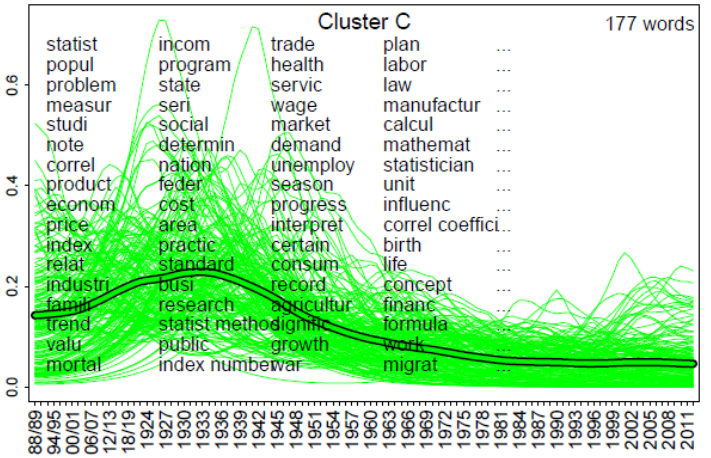
Cluster partition



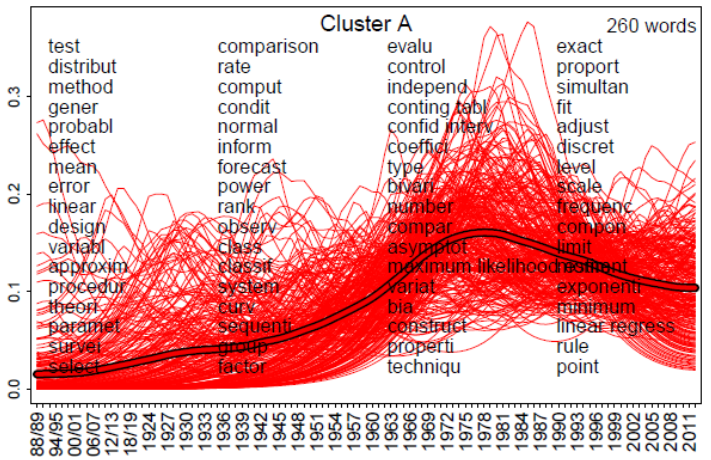
6-cluster partition for normalized data



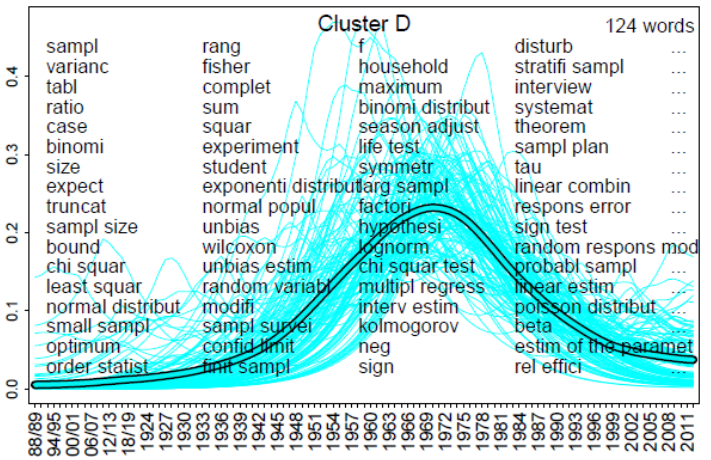
Group F: Demography, Population Studies and Public Statistics



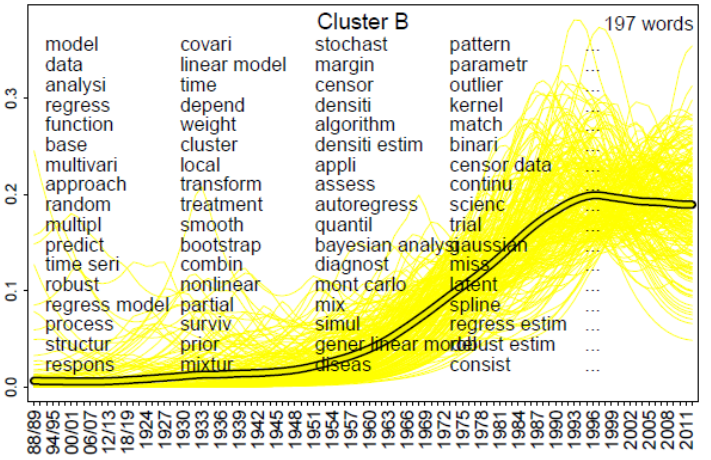
Group C: Social and Economic Statistics



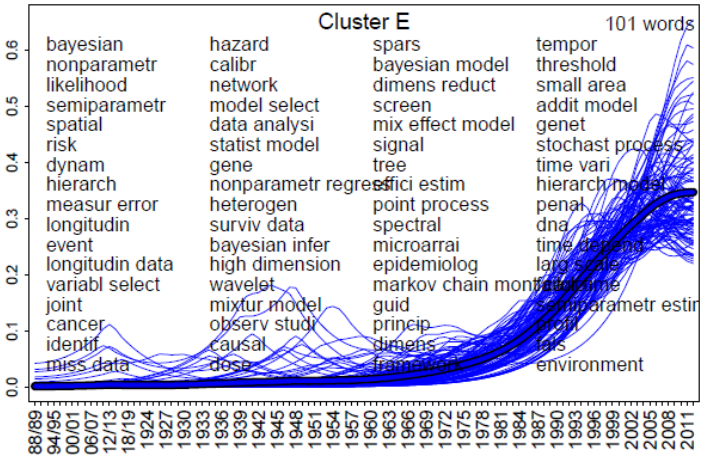
Group A: Statistics become a well-extablished discipline



Group D: Golden Age of Classical Statistics



Group B: Contemporary Statistics



Group E: Modern Statistics

What did we learn?

The idea of studying textual data with a chronological perspective is important and opens the way leads to new challenges (in some cases even computational problems that are not easy to solve).

The effort to develop effective methods for analyzing and representing chronological textual data leads to interesting applications in different fields.

The idea of drawing the history of words and of grouping words that show similar temporal trends are promising.

The foregoing and other methods have been set up for a **project** (PRA 2015-2017):

> Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature."

involving an interdisciplinary research group (linguistics, philosophy, psychology, sociology, statistics) whose aim is to construct corpora of scientific literature, and, hence, to investigate whether a discipline history can be traced from analyzing the temporal pattern of relevant keywords included in papers published by mainstream scientific journals of the discipline.

- Correspondence Analysis
- Reinert's method
- Topic models
- information system consisting of
  1. an information retrieval procedure for keywords selection
  2. a two-stage functional clustering for statistical learning to reconstruct the historical evolution of the knowledge field.

A. Tuzzi (2018, ed), Tracing the Life Cycle of Ideas in the Humanities and Social Sciences, Springer, Cham