

A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal

Matilde Trevisani · Arjuna Tuzzi

Published online: 8 July 2014
© Springer Science+Business Media Dordrecht 2014

Abstract The words that occur in papers published by the journals of an old and prestigious scientific society like the American Statistical Association portray the most relevant research interests of a discipline and the recurrence of words over time show fashions, forgotten topics and new emerging subjects, that is, the history of a discipline at a glance. In this study a set of keywords occurred in the titles of papers published in the period 1888–2012 by the Journal of the American Statistical Association and its predecessors are examined over time in order to retrieve those which appeared in the past and which are today the research fields covered by Statistics, from the viewpoints of both methods and application domains. The existence of a latent temporal pattern in keywords' occurrences is explored by means of (lexical) correspondence analysis and clusters of keywords portraying similar temporal patterns are identified by functional (textual) data analysis and model-based curve clustering. The analyses reveal a definite time dimension in topics and show that much of the History of Statistics may be gleaned by simply reading the titles of papers through an explorative correspondence analysis. However, the functional approach and model-based curve clustering turn out to be better in tracing and comparing the individual temporal evolution of keywords, despite some computational and theoretical limitations.

Keywords Chronological corpora · Correspondence analysis · Functional data analysis · Model-based curve clustering · Textual data

M. Trevisani
Department of Economics, Business, Mathematics and Statistics (DEAMS),
University of Trieste, Piazzale Europa, 1, 34127 Trieste, Italy

A. Tuzzi (✉)
Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA),
University of Padova, Sociology buildings, Via M. Cesarotti 10/12, 35123 Padua, Italy
e-mail: arjuna.tuzzi@unipd.it

1 Introduction

The *American Statistical Association* (ASA) represents the world's largest community of statisticians and is also one of the oldest and most prestigious statistical societies. As stated in the Association's first constitution, in Boston in the year 1839, its purpose was to collect, preserve and publish statistical information in the various departments of human knowledge. Originally called the *American Statistical Society*, the organization changed its name to ASA during its first annual meeting in Boston in 1840. Today, ASA sustains the development, application and dissemination of statistical science through meetings, publications, membership services, education, accreditation and advocacy.

The *Journal of the American Statistical Association* (JASA) has long been considered the world's premier review in its field. Established in 1888, formerly known as *Publications of the American Statistical Association* (PASA 1888–1912) and *Quarterly Publications of the American Statistical Association* (QASA, 1912–1921), JASA (1922–present) is published quarterly and focuses on statistical applications, theory and methods in economic, social, physical, engineering and health sciences. Relevant books contributing to advances in statistics are also reviewed.

In the frame of the analysis of large databases of scientific literature, this study explores the opportunity of learning of the evolution of concepts, methods and applications, i.e., the very History of Statistics, by means of the evolution of keywords in papers published by JASA and its predecessors. A text corpus including the titles of papers published in the period 1888–2012 is examined in order to retrieve those which appeared in the past and which are today the research fields covered by JASA, from the viewpoints of both methods and application domains. The main aims of this study are:

- (a) to make a first overview of the relationship between time and keywords in order to verify the existence of a latent temporal pattern, by means of correspondence analysis (CA);
- (b) to identify keywords showing prototypical temporal patterns, and to cluster keywords portraying similar temporal patterns, by means of functional textual data analysis (FTDA) and model-based curve clustering (MBCC).

The two proposed approaches have deep differences from a conceptual viewpoint, but both are important in that they complement each other in shaping the words' temporal patterns: CA is only an explorative and model-free approach providing a general graphic overview of the structure of the association among years, among keywords, and between years and keywords; FTDA and MBCC represent and compare the individual trajectories of each keyword resorting to a wavelet-based decomposition of the signal-trajectories and a functional (mixed) model for clustering them.

In many applications of the analysis of textual data, corpora include a chronological arrangement (Bolasco 2013) of texts and these timings are crucial to retrieve information about the evolution of a “language”, often in terms of topics, contents, collocations, and other specific linguistics features (at the morphological, lexical or syntactical level). Typical cases emerge in the field of the history of a language, e.g. the analysis of the lexicon of a language based on large corpora of written—mainly literary—texts,¹ and the history of scientific concepts, e.g. the study of specific terminology based on large corpora of scientific texts (e.g. Maggioni et al. 2009). However, the idea of shaping the history of words (Trevisani and Tuzzi 2012, 2013a, b) is unusual in language studies: for example, since the

¹ Examples of historical lexicography for the Italian language are Battaglia's *Grande dizionario della lingua italiana*; the *Vocabolario degli Accademici della Crusca*; Tommaseo and Bellini's *Dizionario della lingua italiana*.

pioneering studies conducted by [Migliorini \(1960\)](#), the main concern in the study of the history of the Italian language has always been dating the birth of individual words and—possibly—their main semantic changes, but little attention has been paid to the fortunes (or death) of words. From a theoretical viewpoint, the provisional results of this study aim at fostering a discussion on the concept of the “quality of life” of words within the scientific community.

2 Corpus and data

2.1 Sources

Within the perspective of analysis of textual data and, more in general, in all cases of data collection based on text harvesting, the construction phases, first of the corpora and then of textual data, are essential moments: choices made before statistical analysis are crucial to guarantee the quality of the available data, if reliable results are to be achieved and also, obviously, to avoid the well-known “garbage in, garbage out” effect.

Our corpus derives from three very popular archives: (1) Thomson Reuters/ISI Web of Knowledge, (2) ASA’s web pages, (3) JSTOR. All references to all papers published by PASA, QASA and JASA, i.e., all information regarding citations (author(s), title, year, volume, issue, number of pages, etc.) and, where possible, the abstracts, were downloaded. Pooling of the three sources was necessary in order to go back to 1888 and fill in some gaps we had observed during text harvesting. A total of 12,557 items was collected for a period of 125 years, from the very first, Volume No. 1, Issue No. 1 (PASA 1888) to the latest, No. 107, Issue No. 500 (JASA 2012).

A corpus is a set of homogeneous texts collected in order to answer a specific research question. To assess the homogeneity of texts in a corpus, several criteria come into play, and the topic, author(s), style and language are all examined ([Muller and Brunet 1988](#); [Popescu et al. 2009](#); [Popescu 2009](#)). The articles cover very many different topics, which are the focus of this research. Many authors were involved in all of them, each with their own writing style, although an analysis by author in this context is not applicable and was not one of the study aims. As regards language ([Cosieriu 1973, 1988](#); [Berruto 1987](#)), again, this variable is not applicable in our case, because all the texts are written in English and diatopic variations (due to different geographical settings) cannot be a study aim, as we do not know whether the texts were originally designed and written in English. However, we have standardized some orthographic variations, e.g., American and British spelling. Variations such as diaphasic (variations reflecting differences in communicative situations), diamesic (variations due to differences in the means of communication) and diastratic (variations in social groups) have been neutralized, as all texts refer to scientific literature and their authors are scholars of the same discipline or from similar research fields. Diachronic variations are the most important trait for this study based on chronological developments, especially with reference to topics, but merely² at the lexical level.

² Although they are obvious in a linguistic environment, it should be noted here that diachronic variations appreciable over such a limited period of time with respect to the timings required for a significant evolution of a language concern only the lexical level, which is also the most superficial part of a language. In the case of the analysis of corpora accounting for several centuries, we should have tackled further specific issues concerning the retrieval of textual data, such as normalization, tokenization and lemmatization aimed at correctly processing orthographic and semantic variants in the course of time.

2.2 Word-types

Words are sequences of letters isolated by means of separators, i.e., blanks and punctuation marks. In order to list the words contained in a corpus, two concepts must be distinguished: word-tokens and word-types. A corpus contains a limited set of different word-types and a word-token is the particular occurrence of a word-type, e.g., word-type *statistics* has many tokens in our corpus, and there are also many word-types which only appear once, e.g. *hemocytometer*. The frequency of a word-type is the number of corresponding word-tokens in the whole corpus or in a subset of texts (*subcorpus*). The number N of word-tokens is the size of the corpus in terms of occurrences; the number V of word-types is the size of the corpus in terms of different words; and the V/N type-token ratio provides a rough measure of lexical richness.

Some of the items retrieved from archives are not articles (e.g., *List of publications, News*); some of them do not include content words in the title (e.g., *Comment, Rejoinder*) and, since many of them are works of the past, they often do not have abstracts.³ Taking into account only the texts of titles including content words, our corpus is composed of 10,077 titles (out of 12,557), 87,060 word-tokens and 7,746 word-types, i.e., this is a small corpus of relatively limited lexical richness, as the type-token ratio is 8.9 % and the number of occurrences of word-types is on average 11.

One of the first steps of the data construction involved the problem of identifying the keywords on which to concentrate analysis. An over-simplified initial solution to reveal terms concerning statistical methods and application domains would be to consider the most frequent words, excluding grammatical ones (articles, conjunctions, prepositions, pronouns, modal verbs). Clearly, this is too naive a method, since it does not take into account redundancies, compounds and ambiguities, and we therefore resorted to content words and sequences of words.

2.3 Stem-types and stem-segments

In a text the nature of some lexical choices is contingent, e.g., verb tenses and the plural forms of nouns. To overcome some of the limitations of analyses based on simple word-types, we replaced words with stems (e.g., word-types: *model, models, modeling* and *modelling* are replaced with the same stem *model*) by means of the version 2012 of Porter's stemming algorithm⁴ (Porter 1980; Sparck Jones and Willet 1997). The procedure yielded a new vocabulary comprising 4,834 stem-types (out of 7,746 word-types).

Texts are sequences of words which have different meanings when they are considered in their context of use and together with adjacent words. To increase the amount of information conveyed by each stem, we identified all stem-segments (or *n-stem-grams*), i.e., sequences of stems occurring in our corpus at least twice and composed of a minimum of two and a maximum of six consecutive stems (e.g., *model select, addit model, hierarch model, log linear model, dynam model*).

³ 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. This is another indication of the standardization of scientific literature as it appears in the structure of published works.

⁴ The Porter's algorithm is a popular stemmer in information retrieval applications. The original stemmer software appeared for the first time in 1979 and was written in BCPL, a once famous but now defunct language. Many versions appeared after 1980, and we chose version 2012 (September 7), available on-line.

In the list of stem-segments sorted by decreasing frequencies,⁵ high-frequency values are mostly populated by segments composed of articles, conjunctions and prepositions (grammatical segments such as *in which*, *and of*, *such as the*, etc.). In order to select content segments, the list was ranked according to Morrone's *IS* indexes (Morrone 1996; Tuzzi 2012):

$$IS = \sum_{i=1}^q \frac{f_{segm}}{f_i} \times P \quad \text{and} \quad IS_{rel} = \frac{IS}{q^2} \quad (1)$$

where q is the number of stems composing the segment, f_{segm} is the frequency of the segment, and f_i is the frequency of the i -th stem of the segment. Multiplicative factor P is the number of non-grammatical stems in the segment. The IS_{rel} index is the standardized version of IS and may vary from 0 to 1, where 1 means that the segment is composed of non-grammatical stems occurring only in that specific sequence.

In this phase of data preparation, it should also be noted that, as longer sequences of stems may contain shorter sequences of stems (e.g., *linear model*, in *log linear model*), text segments were identified and isolated in the corpus starting with the longest ones and the frequencies were calculated afterwards. We also discarded grammatical stems and stem-segments (i.e., articles, conjunctions, prepositions, pronouns, auxiliary and modal verbs), as previous studies show that, although they are good markers of writing style, content words are suitable for harvesting topics (Stamatatos 2009; Tuzzi 2010).

2.4 Tagging and selecting keywords

In order to identify the most important items of the vocabulary, we tagged relevant statistical keywords (stem-types and stem-segments) matching our vocabulary with lists of items of six statistics glossaries available on-line (over 12,700 unique entries):

- (1) ISI—International Statistical Institute (<http://isi.cbs.nl/glossary/index.htm>);
- (2) OECD—Organisation for Economic Cooperation and Development (<http://stats.oecd.org/glossary/index.htm>);
- (3) Statistics.com—Institute for Statistics Education (www.statistics.com/glossary/);
- (4) StatSoft Inc. (www.statsoft.com/textbook/statistics-glossary/);
- (5) University of California, Berkeley (<http://statistics.berkeley.edu/stark/SticiGui/Text/gloss.htm>);
- (6) University of Glasgow (www.stats.gla.ac.uk/steps/glossary/).

Lastly, we selected all keywords with frequencies equal to or higher than 10 in the period 1888–2012 as we need sufficient information to trace over time the temporal pattern of frequencies of individual keywords. The final matrix included 900 keywords and 107 time-points (volumes). The number of volumes is slightly less than the number of years (125) since at the very beginning (1888–1921) the volumes of the ASA's journal were biennial.

3 Methods

From this point onward, we worked with the logic of the “bag of words” approach and used counts, i.e., the frequencies of each keyword over time-points. These discrete data may be

⁵ Procedures for identification and selection of n -word-grams (multiwords, compounds, collocations and, more generally, sequences of words repeated several times in a text) are available in Taltac2. Developed by a research team from the University of Rome “La Sapienza”, Taltac2 is an Italian software program which targets statistical and linguistic resources for computer assisted statistical analysis of textual data.

viewed as a contingency table of keywords \times time-points (CA) and its rows as discrete observations of continuous objects represented by functional relationships (functional data analysis). We expected to identify and automatically isolate trajectories that are easily interpreted, i.e. words whose presence has grown in time and whose increasing trend is continuing; words that have become less frequent and have disappeared; words that were very successful; words that have been constantly present in time. In contrast, many trajectories show multiple peaks that are difficult to interpret in terms of their chronological evolution, although they may hide interesting patterns of regularity.

3.1 Correspondence analysis

Correspondence analysis of the lexical contingency table keywords \times time-points is a common, well-known, and classical tool in the literature on textual data analysis. Here, it was used from an exploratory perspective in order to position years and keywords (stem-types and stem-segments) on a map to verify the existence of a chronological dimension (Lebart et al. 1998) and to represent relationships among keywords, among time-points and between keywords and time-points. CA (Greenacre 1984, 2007; Murtagh 2005) transforms the frequencies of keywords over time into coordinates on a multidimensional Cartesian system. CA displays articles in a low-dimensional space by transforming an appropriate distance (χ^2) into a specific Euclidean distance and, then, mapping distances into Cartesian planes.

3.2 Model-based curve clustering

The temporal evolution of a keyword is expressed by the sequence of its occurrences over a set of time-points. These discrete data may be imagined as the discrete observation of a curve, i.e., a functional observation. Time-series, or rather functional data thus generated, pose some problems: irregular (bumpy, abrupt peak-and-valley) curves, high inter-individual (inter-word) variability, and high dimensionality.

In chronological corpora, data are typically sparse over time, so that many cells of the contingency table have small counts or are empty. The latter are due to a large number of word-types with a relatively low number of corresponding word-tokens—an intrinsic feature of textual data commonly known as the “large p , small n ” problem (Lebart et al. 1984; Johnstone and Titterton 2009)—and to the size of time-point subcorpora. In terms of number of documents and their size in word-tokens, the richness of information and regularity of the corresponding signals may vary greatly over time.

In order to partially reduce the effect of the dimensions of the subcorpora collected for each time-point (available titles including content words for each volume), we replaced absolute frequencies with relative frequencies (the frequency of a word on a time-point is divided by the total number of tokens in titles referring to the same time-point).⁶ Figure 1 shows the temporal trajectories of some of the most (left) and the least (right) frequent keywords in titles. A bumpy course is evident although it is somewhat attenuated for the keywords selected among the most popular (left).

Here, we focus on methods for MBCC when the above problems occur. Although curve clustering has long been studied with splines as the basis for signal decomposition (James and Sugar 2003; Reithinger et al. 2008), they are not appropriate when dealing with high-dimensional data or for modeling irregular functions such as sharp peak-and-valley curves.

⁶ The dependence of the linguistic measures on text length is well known in the field of Quantitative Linguistics (Altmann 1978; Köhler and Galle 1993; Wimmer and Altmann 1999; Popescu 2009). The relative frequency attenuates but does not prevent the effect of dimensions.

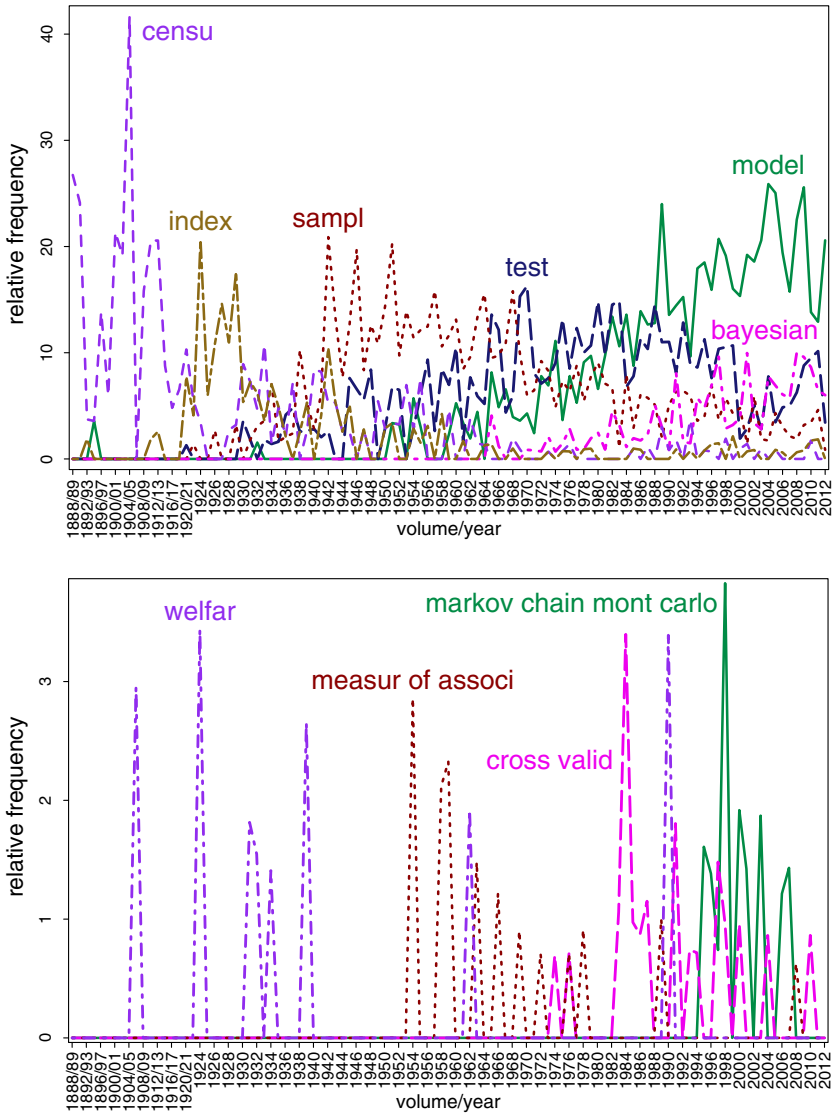


Fig. 1 Temporal trajectories for some of the most- (*top*) and least-frequent (*bottom*) keywords within the corpus of titles, 1888–2012. y-axis represents the keyword relative frequency ($\times 1000$ tokens) for each volume

Wavelet-based decompositions can accommodate a wider range of functional shapes and are more flexible than splines (Morris and Carroll 2006; Morris et al. 2008; Giacomfi et al. 2013), besides being more computationally efficient. Moreover, word-specific fluctuations, often a major source of variability, can be taken into account by considering random word effects in addition to the main fixed component in the functional model framework. We therefore applied the class of wavelet-based functional clustering mixed models (FCMM), originally developed by Giacomfi et al. (2013) in the context of microarray-type data clustering problems, to the setting of chronological corpora. In the wavelet domain, the model refers to a linear

mixed-effects model which can be used to create a model-based clustering algorithm. For estimation, we applied the EM-algorithm for maximum likelihood estimation provided with the *curvclust* (Giacofci et al. 2012) R package (R Core Team 2013), which implements the method proposed by the above authors. A brief description of the FCMM class, here adjusted to chronological corpora, follows.

In our modeling approach, the temporal evolution of each word i ($i = 1, \dots, n$) is represented by a curve $y_i(t)$, observed on M equally spaced time-points ($t = t_1, \dots, t_M$) with $M = 2^J$ for some integer J . In the unsupervised setting, L unknown clusters (L a priori being unknown) grouping words with similar patterns are presumed to exist. We then assume that, for a word i in cluster l ($l = 1, \dots, L$), curve $y_i(t)$ is generated by a functional fixed effect $\mu_l(t)$ characterizing cluster l , a random functional effect $U_i(t)$ accounting for word-specific deviation from the cluster mean curve, and a random measurement error term $E_i(t)$. We thus assume the linear functional mixed model:

$$y_i(t) = \mu_l(t) + U_i(t) + E_i(t)$$

$$E_i(t) = N(0, \sigma_E^2 \delta_{tt'}) \quad U_i(t) = N(0, K_l(t, t')), \tag{2}$$

where $E_i(t)$ is a zero mean Gaussian process with covariance kernel $\text{cov}(E_i(t), E_i(t')) = \sigma_E^2 \delta_{tt'}$, in which $\delta_{tt'}$ is the Kronecker product, and $U_i(t)$ is modeled as a centered Gaussian process with kernel $K_l(t, t') = \text{cov}(U_i(t), U_i(t'))$ independent of $E_i(t)$.

The intrinsic infinite dimension of functions complicates the statistical analysis of functional data. Hence, once defined in the functional domain, a classic approach is to convert the original problem into a finite-dimensional one by means of a functional basis representation of the model. Following Morris and Carroll (2006) and Giacofci et al. (2013), we used a wavelet-based representation of the model (2) and the Discrete Wavelet Transform (DWT) to examine continuous functions on the set of M sampled points. Wavelet representation is based on a father wavelet (or scaling) ϕ and a mother wavelet (or simply wavelet) ψ . Curve $y_i(t)$ has the following decomposition:

$$y_i(t) = \sum_{k=0}^{2^{j_0}-1} c_{i,j_0k}^* \phi_{j_0k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j-1} d_{i,jk}^* \psi_{jk}(t)$$

and for the mixed-effects model (2) the DWT coefficients $\mathbf{c}_i = (c_{i,j_0k})_{k=0 \dots 2^{j_0}-1}$ and $\mathbf{d}_i = (d_{i,jk})_{j=j_0 \dots J-1}^{k=0 \dots 2^j-1}$ are:

$$\mathbf{c}_i = \boldsymbol{\alpha}_l + \mathbf{v}_i + \boldsymbol{\epsilon}_{c_i} \quad \mathbf{d}_i = \boldsymbol{\beta}_l + \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_{d_i},$$

where: $\boldsymbol{\alpha}_l = (\alpha_{l,j_0k})_{k=0 \dots 2^{j_0}-1}$ and $\boldsymbol{\beta}_l = (\beta_{l,jk})_{j=j_0 \dots J-1}^{k=0 \dots 2^j-1}$ represent the scaling and wavelet coefficients of fixed effects $\boldsymbol{\mu}_l(\mathbf{t}) = (\mu_l(t_1), \dots, \mu_l(t_M))^T$; $\mathbf{v}_i = (v_{i,j_0k})_{k=0 \dots 2^{j_0}-1}$ and $\boldsymbol{\theta}_i = (\theta_{i,jk})_{j=j_0 \dots J-1}^{k=0 \dots 2^j-1}$ the scaling and wavelet random coefficients of random effects $\mathbf{U}_i(\mathbf{t}) = (U_i(t_1), \dots, U_i(t_M))^T$; lastly, $\boldsymbol{\epsilon}_{c_i}$ and $\boldsymbol{\epsilon}_{d_i}$ the errors on scaling and wavelet coefficients, distributed as $N(\mathbf{0}_M, \sigma_E^2 \mathbf{I}_M)$. The pair $(\mathbf{v}_i^T, \boldsymbol{\theta}_i^T)^T$ is supposed to follow a $N(\mathbf{0}_M, \text{Diag}(G_v, G_\theta))$, with G_v and G_θ the (diagonal) covariance matrices of v_i and θ_i , and to be independent from $(\boldsymbol{\epsilon}_{c_i}^T, \boldsymbol{\epsilon}_{d_i}^T)^T$. Several ways of modeling the variance of random effects (constant, group-specific or scale-location varying) were also considered.

Lastly, model selection was guided by two criteria from the framework of penalized likelihoods, i.e., the well-known Bayesian Information Criterion (BIC) and the Integrated Classification Likelihood criterion (ICL). The latter is based on the integrated likelihood of the complete data (Biernacki et al. 2000) and includes prediction of random effects (Giacofci

et al. 2013), and thus is considered an alternative when mixed models are assessed in a model-based clustering perspective.

3.3 Model selection problems in clustering histories of a word's "quality of life"

Within the rationale of standard model selection, an implicit assumption is that a "true model" exists and that it is one of the models in competition. This assumption commonly underlies the use of standard penalized likelihood criteria such as BIC or, in the case in question, ICL. However, besides being unrealistic, this assumption may lead to choose overly complex and useless models from a practical point of view. This problem is crucial in model-based clustering settings, in which the main aim is to find interpretable and possibly parsimonious group partitions. Ultimately, the specific modeling purpose—which is not considered in the routine use of standard criteria—should be taken into account in model selection to find practically relevant answers to scientific questions.

In this work, defining the modeling aim was decisive in choosing the type of variance for random effects. Once both selection criteria clearly showed that a mixed structure was needed, the choice still lay between a constant versus a group-specific variance set-up. Automatic use of standard criteria indicated a more complex model as the best one, i.e., a mixed model with group-specific variance for random effects and, given this choice, a relatively higher number of clusters. However, the resulting clustering was not quite meaningful in answering our main questions. On the contrary, the best model within the constant variance class provided better clustering. In short, using group-specific variance typically involves a greater number of uninteresting clusters, i.e., groups of curves with no apparent similarity moving on average along a flat mean trend, as well as more singletons, or outlier curves. In addition, interesting clusters, i.e., groups showing on average a non-amorphous temporal pattern, are made up solely of curves which are clearly synchronized. In contrast, partitions provided by using constant variance consist of all but one meaningful clusters, showing a rather lower degree of synchronization, and driven by the most frequent words of the corpus. The rest of the curves left out by significant grouping, generally corresponding to low or very low-frequency words, plotted in a single amorphous cluster.

We believe that these results were the direct consequence of two model features: the specification of normal distribution for curves, determining partitions primarily driven by high-level curves, corresponding to high-frequency words, combined with the type of variance chosen for random effects. If the variance varies with group, the model is sufficiently flexible to produce as many amorphous clusters as are needed; if it is constant, the model can form just one "garbage" cluster containing all the remaining unpaired curves. In addition, although high-level curves are also dominant in determining clusters with group-specific variance, their power is somehow reduced by lower-frequency words, again thanks to the higher flexibility of the more complex model.

Our aim in this study was to construct a "general history" of Statistics, specifically emphasizing the "grand themes" which have always characterized the scientific literature on the subject, and not the many, albeit interesting, "micro-histories". This matches the fact that the most "popular" words guided the formation of clusters. If our intention had been, instead, to obtain groups of words with a time trend similar in form and independent of their frequency or "popularity", we would have normalized the lexical contingency table by row right from the beginning. Our modeling aim therefore led us to choose constant variance: all groups of words with temporal trends of some significance, "driven" by the most popular topics, are highlighted, although the degree of correlation is admittedly less strong and many of

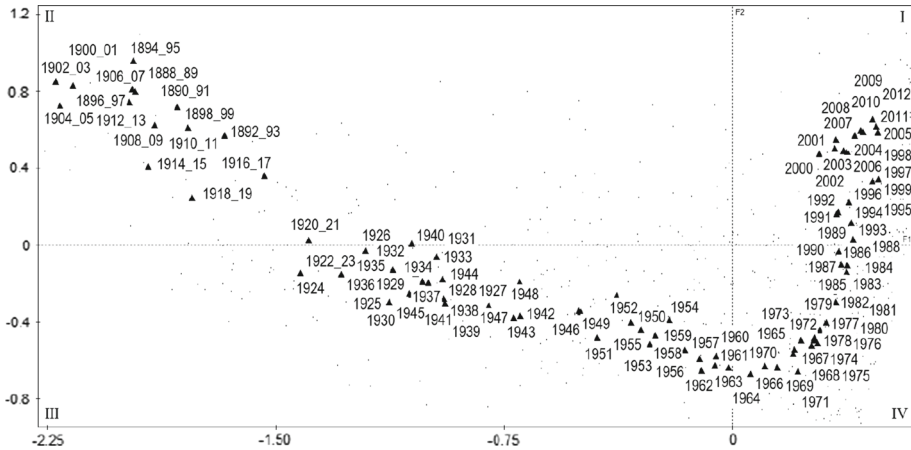


Fig. 2 First factorial plan of correspondence analysis. Projection of volume-points

the “little histories” are sacrificed. However, in the future, we will compare more flexible choices, to reveal aspects of the corpus which, here, inevitably remain in shadow.

4 Results

4.1 Correspondence analysis

The first important aspect revealed by the CA based on titles⁷ was that, in the course of history, the language of Statistics became considerably specialized and increasingly standardized as the discipline⁸ became established and accredited as autonomous. The clear-cut spread of Figs. 2 and 3 shows that, before the 1960s, the wider the range of subjects, the richer the lexicon, whereas topics have become more circumscribed and the lexicon more technical in recent years. Lexical richness is reduced in the name of standardization. One important phenomenon which influenced this learning process of a “special language” (Cortelazzo 1990) was JASA’s specialization which, with the establishment of other journals in other disciplines (demography, economics, biometrics, etc.), has increasingly become less generalistic and more focused in its contents, methods and applications (although today it is considered less specialized than others in the statistic sector).

The timings which we obtained with observations of statistical methods and fields of interest obviously does not coincide with the initial discovery of these methods and first interest in these fields, but reflects the moment when they became established in the scientific community and spread to the literature (something which of necessity arrives late).

The titles of articles published in ASA journals (Fig. 2, 1888–2012) fall into a very clear-cut time dimension and four reference periods (quadrants). The keywords which mainly

⁷ Correspondence analysis of the contingency table 900 keywords \times 107 volumes led to 106 eigenvalues and eigenvectors. The first two axes collect the 18% of explained inertia (11% the first-horizontal axis and 7% the second-vertical axis).

⁸ Following Healy (1978), Bartholomew (1995) and Straf (2003), we never used the term “science” to refer to Statistics.

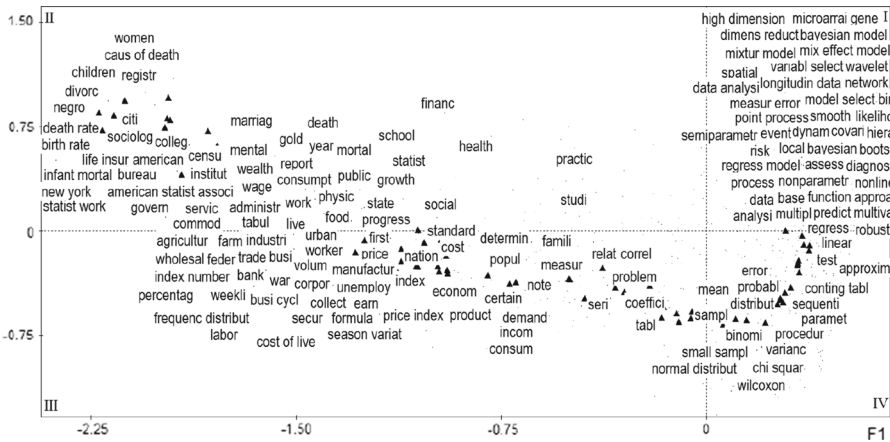


Fig. 3 First factorial plan of correspondence analysis. Projection of volume-points and 20% of key-words with highest contributions

contributed to the factorial solution, falling in the same periods, show which applicative environments and which statistical methods characterized the various periods (Fig. 3).

If, as a metaphor, we apply the timings traditionally used in historical studies and compare the advent of writing with the appearance of documents published in ASA’s scientific journals, we could imagine that Statistics had a period of *Prehistory*, with great names (e.g., political arithmetic with John Graunt and William Petty, German statistics with Corning and Achenwell, probability calculation with Pascal, Fermat, Bayes, Laplace and Gauss, demography with Malthus and Quet  let, etc.). We cannot examine it here, owing to the lack of available documentation. However, continuing with the same metaphor, the *Ancient History* of Statistics begins in 1888 with the publication of newly conceived scientific journals and the birth of a new civilization of statisticians: the ASA. Like the great civilizations of the world, which developed along great rivers, the statisticians of these first generations showed their capacity for “thinking big” by writing scientific articles which dealt with humanity’s great problems. In this period (Fig. 3, quadrant II, top left), Statistics was prevalently *social* statistics and demography, studying the *living* conditions of the population (*children, negro, women, cities, census*), survival (*vital statistics, birth rate, death rate, cause of death, life insurance, infant mortality*), *health, food*, education (*school, college*), economic (*wealth, wages, labour statistics, consumption, progress, finance*) and social status (*marriage, sociology*). Statistics was still mainly an instrument of *government* and ancillary to other subjects (*bureau, administration, public, service*), but it was already showing signs of developing (Ogburn 1940; Stigler 1988).

The two-year period 1920–1921 (Fig. 3, quadrant III, lower left) represents a passage to a new era, the *Middle Ages* (1920–1960). In the early part of these Middle Ages (our *Early Middle Ages*), researchers concentrated on the economy and the Depression, and Statistics returned to “political arithmetic”, producing documentation on state resources (*manufactured goods, farms, cost of living, price index, earnings, costs*). After the “Dark Ages” between the two world wars, the state’s administrative, economic and juridical organization gave way to post-war reconstruction and a phase of concentrating scientific interests on tools for data analysis. Between 1945 and the early 1960s (our *Late Middle Ages*), Statistics was still very much anchored to economic statistics, but a new reason for unity appeared in the

first rudimentary mathematical instruments (*measure, series, table, correlation, coefficient*) (Ogburn 1940). A true sense of *Humanism* was about to open its doors to the *Renaissance* of Statistics.

The beginning of *Modern History* (1960–1990) can be collocated in the early 1960s (Fig. 3, quadrant IV, lower right): after having served mainly demography and economics, statistics now affirmed its status as an autonomous discipline and gained its own lexicon and methods. The rudiments of modern statistics burst upon the world with great significance, like the Discovery of the Americas, marking the end of the Middle Ages and the beginning of the Modern Age. The new Far West was what we now call the essentials of the discipline, like *probability, sampling, distributions (Normal, Chi-square)* and *tests (Wilcoxon)*. Development proceeded with a rich supply of instruments and methods, now called “classic” (*regression, sampling*), until the late 1980s, when technological revolutions stimulated new modes of statistical analysis which were unthinkable before the arrival of modern computers (first algorithms and simulation, later *high dimensional* statistics and increasingly refined modeling).

These innovations led to a new era which we can call *Contemporary History* (1990–2012). Between 1990 and 2012 (Fig. 3, quadrant I, top right), *algorithms* advanced (*bootstrap*) and *smoothing (wavelets)*, *functional data analysis, risk analysis, neural networks* and robust methods became highly fashionable, like many other classic topics in statistical methodology (*multiple, mixed effect models, mixture model, hierarchical models, additive, latent, point process, longitudinal data, binary data, dynamic covariance*). This period was characterized by interesting dualities in *data analysis*, such as frequentist versus *Bayesian, parametric* versus *non-parametric*, linear versus *non-linear* and first attempts at compromise (*semi-parametric*). The new millennium heralded new topics and new challenges about epidemiology, medicine, biology, environment, surveillance, *spatial statistics* and the impact—partly cultural—of knowledge of the genome is apparent (*microarray, gene expression*). Estimation methods like *bootstrap* and empirical *likelihood* emerged as largely popular and the problem of *model selection (variable selection)* is a topic that remains central.

What clearly stands out in the last few decades is a truly new paradigm of complexity, which pervades sciences. On one hand, it raises its head in Statistics with true problems such as the *dimensionality* and complexity of the information to be processed (*high dimensional, dimension reduction*) and, on the other, with the need for *mixing, hybridization* and interdisciplinarity (Scheaffer 2002; Sen 2002; Straf 2003; Brown and Kass 2009).

4.2 Model-based clustering

From model-based clustering of keywords included in the corpus of titles, a functional clustering mixed model with 13 clusters (Fig. 4, top left: mean curves of all groups) and constant variance for functional random effects performs best according to the viewpoints of statistics (in particular, the BIC criterion) and subject-matter. The algorithm was run several times (and, moreover, for different EM tolerance thresholds ε) and a good option for the number of clusters was found to range from 12 to 20 according to the BIC (the ICL curve has an elbow point at 16 from where it is practically steady). After checking that the partitioning of words into groups and, more importantly, that the core of clusters remained essentially constant across this range of numbers, we finally chose the optimal number 13 (and run).

Apart from one cluster with a rather flat, undifferentiated trend (Fig. 4, Group 13, the “garbage” cluster containing all the curves left out by significant grouping, consisting of low or very low-frequency words) and one outlier (Group 1, corresponding, interestingly enough, to the most frequent word *statistics*, which was largely used in the formative period of

Statistics, either meaning “summary” or associated with various areas of application, although it declined after the 1960s, the period we make coincide with the beginning of *Modern History* when Statistics had become fully established as a autonomous discipline), other clusters

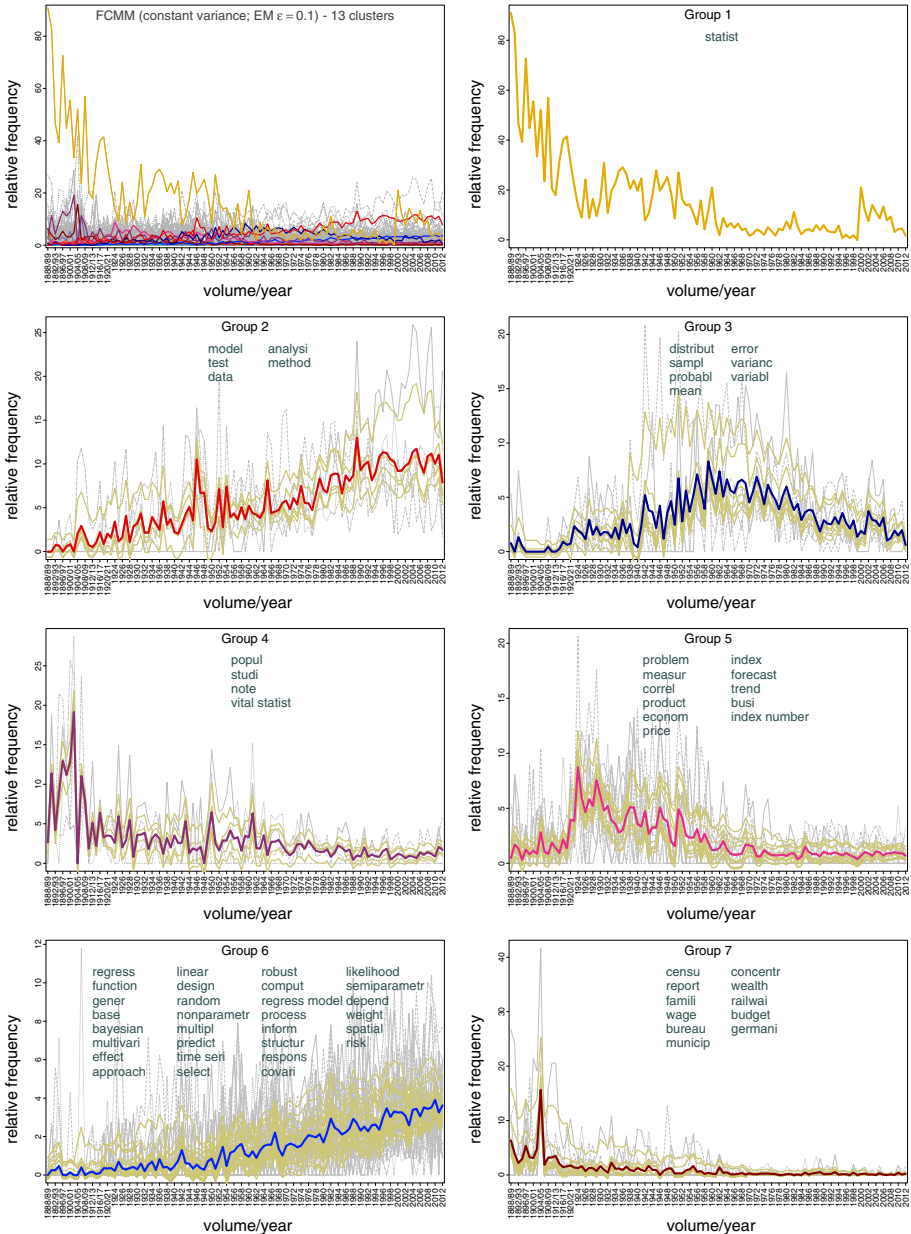


Fig. 4 The best model with the overall groups (top-left) and the single 13 clusters ordered according to the overall incidence of the frequency (y-axis scale) of cluster words. Group-specific functional fixed effect (bold darker line) and functional random effects (bold lighter lines) are superimposed to the individual word curves (thin lines) (continues to the next page)

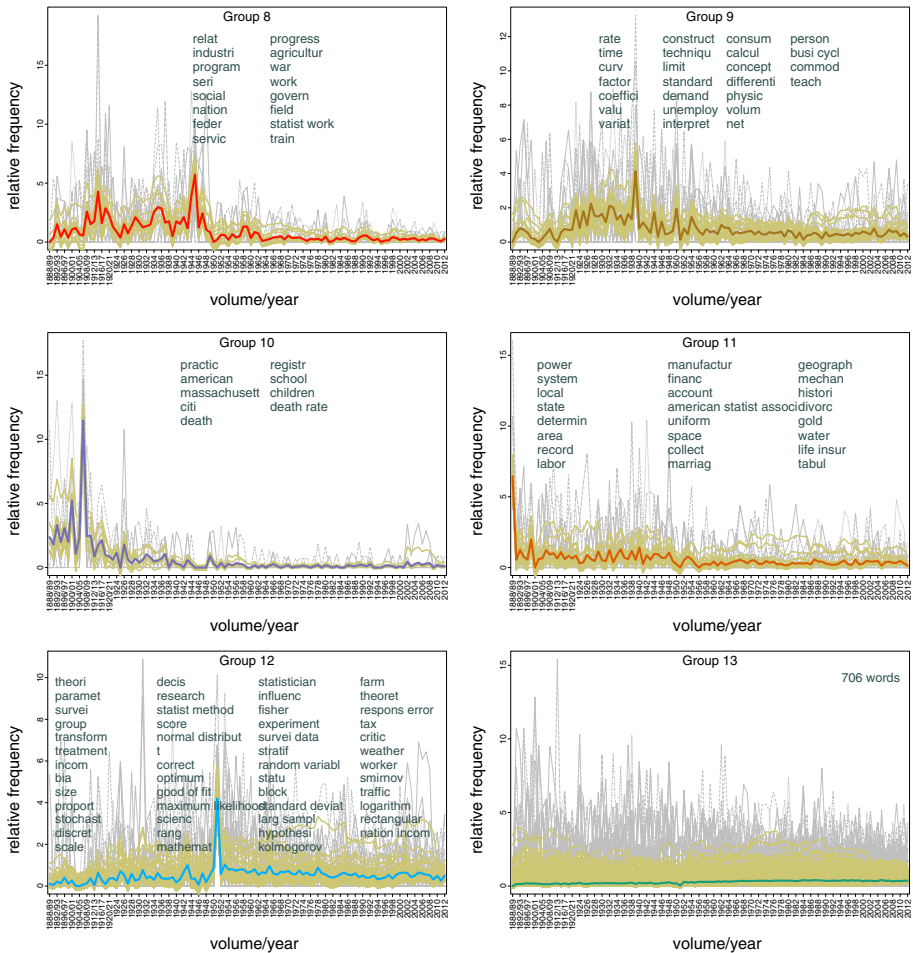


Fig. 4 continued

displayed interesting chronological backgrounds. Note that groups are numbered according to their ranking in the overall incidence of the frequency of “group words” (see y-axis scale in Fig. 4). When interpreting results, we follow this order of “popularity”, eventually grouping distantly ranked clusters by theme.

Starting from the highest-ranking cluster after the singleton *statistics*, Group 2 shows the ongoing development of the most frequent keywords in statistical vocabulary (*model, test, data, analysis, method*): indeed, if we look closer, the frequency of the first-ranked words is markedly rising until the late '90s when they reach a plateau. Group 3 includes the basic elements of statistics (*distribution, sample, probability, mean, error, variance, variable*), showing that it became established as a separate discipline in the second half of the 20th century and that the *Golden Age* of its classical formulation occurred during the period 1950–1980. Groups 4, 7, 10 and 11 include key-words related to demography (*population, vital statistics, registration, death rate, life insurance*) and social economic statistics (*families, wages, cities, schools, children, marriages, divorces, wealth, railways, labor, manufacture,*

finance, accounts, geography, gold, water), mainly addressed in the oldest articles at the turn of the century till the first decade, as well as to the gradual consolidation of statisticians' work and the public role of statistics, being of significant impact until the 1950s (*census, report, bureau, municipalities, budget, Massachusetts, state, records, collection, practical, uniformity, mechanical, tabulation*). Group 8 again stresses the service of statistics to public government and social-economic planning "in problems of peace and war" throughout the first half of the 20th century, with two peaks during the years of the world wars (*industrial, program, series, social, nation, federal, service, progress, agriculture, war, work, government, statistical work, train*). Groups 5 and 9 highlight the dominance of economic statistics in the first half of the 20th century after World War I, especially before the first forebodings of World War II and almost disappearing after the 1960s (*measure, correlation, production, economic, price, index, forecast, trend, business, index number, rate, time, curve, coefficient, variation, demand, unemployment, consumer, differential, volume, net, business cycle, commodity*). Group 6 shows, since the 1950s, the emergence of modern statistics, with the expansion and refinement of approaches to estimation, modeling techniques and mainstream areas of application (*Bayesian, non-parametric, likelihood, semi-parametric, selection, robust, computation, regression, functional, generalized, multivariate, multiple, process, covariance, dependence, weighted, time series, spatial, risk*).

5 Discussion and conclusions

The approach to textual data and text mining techniques differs from that of functional data analysis, although both approaches are useful, and the integration of various perspectives always poses interesting challenges.

By means of CA the titles of papers show a clear-cut chronological pattern and four different eras in the History of Statistics emerge from the first factorial plan: from the origins to the twenties and World War I; from the twenties to World War II and, then, to early sixties; from the sixties to early eighties; from the late eighties to nowadays. Moreover, titles show a progressive reduction in variability along time, i.e. the scientific language has become more technical and more specialized in recent decades.

Initial results from FTDA and MBCC have begun to bring out prototypical temporal patterns and cluster keywords with similar patterns. However, there are various modeling aspects proposed here which we intend to examine in greater depth in the future. Apart from comparisons between various types of variance for random effects, already discussed, the topics on which we are working are: (1) the choice of basis systems alternative to wavelets for smoothing of functional data; (2) the opportunity of further reducing data by functional principal component analysis; (3) the use of distributions alternative to Normal for frequency curves; (4) in-depth study of the problem of choosing the number of groups in the context of model-based clustering, both by comparing consolidated penalized likelihood criteria and by their integration with the modeling purpose assumed as a principle, wrapping the two rationales into an automatic selection procedure.

As regards the choice of alternative basis systems, we are already analysing data processed by B-splines. The latter generally produce greater smoothing, revealing clearer and more easily interpreted time trends, but with the obvious sacrifice of sometimes much of the information contained in the data. As regards the choice of alternative distributions, we are concentrating on the class of skew-normal distributions. They seem to be more suitable for asymmetric data, allowing better and more efficient clustering and a potentially better balance between high- and low-frequency curves in determining cluster formation. This last

aspect is linked to the more general discourse of defining primary aims. Emphasizing once again that the “popularity” of words remains crucial in describing the general history of a corpus, the search for methods which can best highlight the “shape” of temporal patterns—the synchronization, or correlation, between words, independently of their popularity—remains one of the main aims of the various in-depth analyses mentioned above.

A simple exploratory analysis reveals a definite time dimension in JASA articles and much of the History of Statistics may be gleaned by simply reading their titles. CA clearly reveals timings, although on one hand it is difficult to interpret the abundance of keywords and, on the other, to understand—beyond the chronological pattern—how single concepts evolved and which keywords shared the same destiny in the course of the History of Statistics. In brief, CA is useful for an overall view of the corpus, but MBCC turns out to be better in tracing and comparing the individual trajectories of keywords.

Although at present our analysis should be considered purely explorative, the joint application of CA and MBCC proved promising and one of our research aims is to adopt the same perspective for analyzing abstracts of recent articles in order to examine in depth the evolution of the discipline during a period of full lexical maturity and substantial development of methods.

The history recounted by these data is composed of fashions, forgotten topics and new emerging subjects and these results are a first step towards a new concept of “quality of life” of words: we have a toolbox to trace the “life cycle” of words in a chronological corpus and to recognize clusters of words sharing the same “luck” or destiny. In other words we are ready to introduce in the scientific debate the importance of discriminating the first occurrence and the “settlement” of a given word in chronological corpora. Unfortunately, clear patterns could not be observed in our application as each word possesses its own irregular series of frequency values (cfr. Fig. 1). Moreover, since every model in quantitative linguistics should be checked for important properties such as reliability, interpretability, and simplicity (Altmann 1978, 1988), we have to highlight the weaknesses of our explorative approach and the lack of a robust linguistic theory to justify and interpret such a complex model.

Acknowledgments The present study is included in the activities conducted within the frame of the GIAT—Interdisciplinary Group on Text Analysis (www.giat.org).

References

- Altmann, G.: Zur Verwendung der Quotienten in der Textanalyse. In: Altmann, G. (ed.) *Glottometrika 1*, pp. 91–106. Brockmeyer, Bochum (1978)
- Altmann, G.: Linguistische Meßverfahren. In: Ammon, U., Dittmar, N., Mattheier, K. (eds.) *Sociolinguistics. Soziolinguistik*, pp. 1026–1039. Walter de Gruyter, Berlin (1988)
- Bartholomew, D.J.: What is statistics? *J. R. Stat. Soc. Ser. A Stat. Soc.* **158**(1), 1–20 (1995)
- Berruto, G.: *Sociolinguistica dell'italiano contemporaneo*. Studi Superiori NIS. Carocci, Roma (1987)
- Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intel.* **22**(7), 719–725 (2000). doi:[10.1109/34.865189](https://doi.org/10.1109/34.865189)
- Bolasco, S.: *L'analisi automatica dei testi. Fare ricerca con il text mining*. Carocci, Roma (2013)
- Brown, E.N., Kass, R.E.: What is statistics. *Am. Stat.* **63**(2), 105–110 (2009)
- Cortelazzo, M.: *Lingue speciali, La dimensione verticale*. Unipress, Padova (1990)
- Coseriu, E.: *Lezioni di linguistica generale*. Boringhieri, Torino (1973)
- Coseriu, E.: *Einführung in die Allgemeine Sprachwissenschaft*. UTB für Wissenschaft / UTB für Wissenschaft. Francke, Tübingen (1988)
- Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F.: *Curvclust: curve clustering* (2012). <http://CRAN.R-project.org/package=curvclust>. R package version 0.0.1

- Giacofci, M., Lambert-Lacroix, S., Marot, G., Picard, F.: Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* **69**(1), 31–40 (2013). doi:[10.1111/j.1541-0420.2012.01828.x](https://doi.org/10.1111/j.1541-0420.2012.01828.x)
- Greenacre, M.: *Theory and Applications of Correspondence Analysis*. Academic Press, Michigan (1984)
- Greenacre, M.: *Correspondence Analysis in Practice*. Chapman & Hall/CRC, London (2007)
- Healy, M.J.R.: Is statistics a science? *J. R. Stat. Soc. Ser. A Gen.* **141**(3), 385–393 (1978)
- James, G.M., Sugar, C.A.: Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.* **98**, 397–408 (2003)
- Johnstone, I.M., Titterton, D.M.: Statistical challenges of high-dimensional data. *Phil. Trans. R. Soc. A* **367**(1906), 4237–4253 (2009). doi:[10.1098/rsta.2009.0159](https://doi.org/10.1098/rsta.2009.0159)
- Köhler, R., Galle, M.: Dynamic aspects of text characteristics. In: Hřebíček, L., Altmann, G. (eds.) *Quantitative Text Analysis*, pp. 46–53. Wissenschaftlicher, Trier (1993)
- Lebart, L., Morineau, A., Warwick, K.M.: *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Applied Probability and Statistics. Wiley, Chichester (1984)
- Lebart, L., Salem, A., Berry, L.: *Exploring textual data*. Kluwer Academic Publishers, Dordrecht (1998)
- Maggioni, M.A., et al.: Mapping the evolution of “clusters”: a meta-analysis. Working Papers 2009.74, Fondazione Eni Enrico Mattei (2009)
- Migliorini, B.: *Storia della lingua italiana*. Sansoni, Firenze (1960)
- Morris, J.S., Brown, P.J., Herrick, R.C., Coombes, K.R.: Bayesian analysis of mass spectrometry proteomics data using wavelet based functional mixed models. *Biometrics* **64**(2), 479–489 (2008). doi:[10.1111/j.1541-0420.2007.00895.x](https://doi.org/10.1111/j.1541-0420.2007.00895.x)
- Morris, J.S., Carroll, R.J.: Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**(2), 179–199 (2006). doi:[10.1111/j.1467-9868.2006.00539.x](https://doi.org/10.1111/j.1467-9868.2006.00539.x)
- Morrone, A.: Temi generali e temi specifici dei programmi di governo attraverso le sequenze di discorso. In: Villone, M., Zuliani, A. (eds.) *L'attività dei governi della Repubblica italiana (1948–1994)*, pp. 351–369. Il Mulino, Bologna (1996)
- Muller, C., Brunet, E.: La statistique résout-elle les problèmes d'attribution? *Strum. Crit.* **3**(3), 367–387 (1988)
- Murtagh, F.: *Correspondence analysis and data coding with Java and R*. Chapman & Hall/CRC Computer Science & Data Analysis, London (2005)
- Ogburn, W.F.: Statistical trends. *J. Am. Stat. Assoc.* **35**(209b), 252–260 (1940). doi:[10.1080/01621459.1940.10500563](https://doi.org/10.1080/01621459.1940.10500563)
- Popescu, I.: *Word Frequency Studies*. Mouton De Gruyter, Berlin (2009)
- Popescu, I., Mačutek, J., Altmann, G.: Aspects of Word Frequencies. *Studies in Quantitative Linguistics*. RAM, Ludenscheid (2009)
- Porter, M.F.: An algorithm for suffix stripping. *Program* **14**(3), 130–137 (1980)
- R Core Team: *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2013). <http://www.R-project.org/>
- Reithinger, F., Jank, W., Tutz, G., Shmueli, G.: Modelling price paths in on-line auctions: smoothing sparse and unevenly sampled curves by using semiparametric mixed models. *J. R. Stat. Soc. Ser. C* **57**(2), 127–148 (2008)
- Scheaffer, R.L.: Statistical bridges. *J. Am. Stat. Assoc.* **97**(457), 1–7 (2002). doi:[10.1198/016214502753479176](https://doi.org/10.1198/016214502753479176)
- Sen, P.: Shifting goals and mounting challenges for statistical methodology. *J. Mod. Appl. Stat. Methods* **1**(1), 2–12 (2002)
- Sparck Jones, K., Willet, P.: *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco (1997)
- Stamatatos, E.: A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* **60**(3), 538–556 (2009). doi:[10.1002/asi.v60:3](https://doi.org/10.1002/asi.v60:3)
- Stigler, S.M.: A look backward on the occasion of the centenary of JASA. *J. Am. Stat. Assoc.* **83**(403), 583–587 (1988). doi:[10.1080/01621459.1988.10478637](https://doi.org/10.1080/01621459.1988.10478637)
- Straf, M.L.: Statistics: the next generation. *J. Am. Stat. Assoc.* **98**(461), 1–6 (2003). doi:[10.1198/016214503388619030](https://doi.org/10.1198/016214503388619030)
- Trevisani, M., Tuzzi, A.: Chronological analysis of textual data and curve clustering: preliminary results based on wavelets. In: Società Italiana di Statistica (ed.) *Proceedings of the XLVI Scientific Meeting*, pp. 1–4. Cleup, Padova (2012)
- Trevisani, M., Tuzzi, A.: Shaping the history of words. In: Obradovic, I., Kelih, E., Köhler, R. (eds.) *Methods and Applications of Quantitative Linguistics: Selected Papers of the VIIIth International Conference on Quantitative Linguistics*, pp. 84–95. Academic Mind, Belgrad (2013a)
- Trevisani, M., Tuzzi, A.: Through the JASA's looking-glass, and what we found there. In: *Proceedings of the 28th International Workshop on Statistical Modelling*, vol. 1, pp. 417–422. Istituto Poligrafico Europeo, Palermo (2013b)

- Tuzzi, A.: What to put in the bag? Comparing and contrasting procedures for text clustering. *Ital. J. Appl. Stat. Stat. Appl.* **22**(1), 77–94 (2010)
- Tuzzi, A.: Reinhard Köhler's scientific production: words, numbers and pictures. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.) *Synergetic Linguistics Text and Language as Dynamic Systems*, pp. 223–242. Praesens, Vienna (2012)
- Wimmer, G., Altmann, G.: Review article: on vocabulary richness. *J. Quant. Ling.* **6**(1), 1–9 (1999)