SIS 2017
Statistics and Data Science: new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.
(Proceedings e report ; 114)

http://digital.casalini.it/9788864535210

ISBN 978-88-6453-521-0 (online)

Peer Review Process
All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board
A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

SOCIETÀ ITALIANA DI STATISTICA
Sede: Salita de' Crescenzi 26 - 00186 Roma
Tel +39-06-6869845 - Fax +39-06-68806742
email: sis@caspur.it web:http://www.sis-statistica.it

La Società Italiana di Statistica (SIS), fondata nel 1939, è una società scientifica eretta ad Ente morale ed inclusa tra gli Enti di particolare rilevanza scientifica. La SIS promuove lo sviluppo delle scienze statistiche e la loro applicazione in campo economico, sociale, sanitario, demografico, produttivo ed in molti altri settori di ricerca.

Organi della società:

Presidente:
- Prof.ssa Monica Pratesi, Università di Pisa

Segretario Generale:
- Prof.ssa Filomena Racioppi, Sapienza Università di Roma

Tesoriere:
- Prof.ssa Maria Felice Arezzo, Sapienza Università di Roma

Consiglieri:
- Prof. Giuseppe Arbia, Università Cattolica del Sacro Cuore
- Prof.ssa Maria Maddalena Barbieri, Università Roma Tre
- Prof.ssa Francesca Bassi, Università di Padova
- Prof. Eugenio Brentari, Università di Brescia
- Dott. Stefano Falorsi, ISTAT
- Prof. Alessio Pollice, Università di Bari
- Prof.ssa Rosanna Verde, Seconda Università di Napoli
- Prof. Daniele Vignoli, Università di Firenze

Collegio dei Revisori dei Conti:
- Prof. Francesco Campobasso, Prof. Michele Gallo, Prof. Francesco Sanna, Prof. Umberto Salinas (supplente)

SIS2017 Committees

Scientific Program Committee:
Rosanna Verde (chair), Università della Campania "Luigi Vanvitelli"
Maria Felice Arezzo, Sapienza Università di Roma
Antonino Mazzeo, Università di Napoli Federico II
Emanuele Baldacci, Eurostat
Pierpaolo Brutti, Sapienza Università di Roma
Marcello Chiodi, Università di Palermo
Corrado Crocetta, Università di Foggia
Giovanni De Luca, Università di Napoli Parthenope
Viviana Egidi, Sapienza Università di Roma
Giulio Ghellini, Università degli Studi di Siena
Ippoliti Luigi, Università di Chieti-Pescara "G. D'Annunzio"
Matteo Mazziotta, ISTAT
Lucia Paci, Università Cattolica del Sacro Cuore
Alessandra Petrucci, Università degli Studi di Firenze
Filomena Racioppi, Sapienza Università di Roma
Laura M. Sangalli, Politecnico di Milano
Bruno Scarpa, Università degli Studi di Padova
Cinzia Viroli, Università di Bologna

Local Organizing Committee:
Alessandra Petrucci (chair), Università degli Studi di Firenze
Gianni Betti, Università degli Studi di Siena
Fabrizio Cipollini, Università degli Studi di Firenze
Emanuela Dreassi, Università degli Studi di Firenze
Caterina Giusti, Università di Pisa
Leonardo Grilli, Università degli Studi di Firenze
Alessandra Mattei, Università degli Studi di Firenze
Elena Pirani, Università degli Studi di Firenze
Emilia Rocco, Università degli Studi di Firenze
Maria Cecilia Verri, Università degli Studi di Firenze

Supported by:
Università degli Studi di Firenze
Università di Pisa
Università degli Studi di Siena
ISTAT
Regione Toscana
Comune di Firenze
BITBANG srl

Index

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Preface

The 2017 SIS Conference aims to highlight the crucial role of the Statistics in Data Science. In this new domain of "meaning" extracted from the data, the increasing amount of produced and available data in databases, nowadays, has brought new challenges. That involves different fields of statistics, machine learning, information and computer science, optimization, pattern recognition. These afford together a considerable contribute in the analysis of "Big data", open data, relational and complex data, structured and no-structured. The interest is to collect the contributes which provide from the different domains of Statistics, in the high dimensional data quality validation, sampling extraction, dimensional reduction, pattern selection, data modelling, testing hypotheses and confirming conclusions drawn from the data. In the mention that statistics is the "grammar of data science", statistics has become a basic skill in data science: it gives right meaning to the data. Still, it isn't replaced by newer techniques from machine learning and other disciplines but it complements them. The Conference is also addressed to the new challenges of the new generations: the native digital generations, who are called to develop professional skills as "data analyst", one of the more request professionality of the 21st Century, crossing the rigid disciplinary domains of competence. In this perspective, all the traditional statistical topics are admitted with an extension to the related machine learning and computer science ones. The present volume includes the short papers of the contributions that will be presented in the 4 invited speaker sessions; in the 19 specialized sessions; in the 11 solicited sessions; in the 6 foreign societies sessions and in the 17 contributed sessions as well as, in the panel session.

Rosanna Verde
President of the Scientific Programme Committee

Alessandra Petrucci
President of the Local Organizing Committee

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Knowledge mapping by a functional data analysis of scientific articles databases

Mappare la conoscenza attraverso un'analisi di dati funzionali di basi di dati di articoli scientifici

Matilde Trevisani and Arjuna Tuzzi

Abstract Scientometrics studies in quantitative fashion the evolution of science focusing on the analysis of publications. One of its objectives is the development of information systems that can help to explore the enormous amount of scientific articles unceasingly published. Our study proposes an information system to reconstruct a dynamical knowledge mapping from a functional data analysis perspective. The source database is a diachronic corpus which originates a words×time-points contingency table displaying the frequencies of each keyword in the set of texts grouped by time-points in the observed time span. The information system consists of an information retrieval procedure for keywords' selection and a two-stage functional clustering to reconstruct the historical evolution of the knowledge field under investigation.

Abstract *La scientometria studia con un approccio quantitativo l'evoluzione della scienza attraverso l'analisi delle pubblicazioni. Uno degli obiettivi é lo sviluppo di sistemi di informazione di ausilio nell'esplorare l'enorme mole di articoli scientifici pubblicati incessantemente. Il nostro studio propone un sistema di informazione atto a ricostruire una mappatura dinamica della conoscenza secondo una prospettiva di analisi di dati funzionali. Il database di partenza é un corpus diacronico che dá origine a una tabella di contingenza parole×punti temporali contenente le frequenze di ogni parola chiave nell'insieme dei testi raggruppati per punti temporali lungo l'arco di tempo osservato. Il sistema informativo é costituito da una procedura di recupero delle informazioni per la selezione delle parole chiave e un clustering funzionale a due stadi per ricostruire l'evoluzione storica del campo di conoscenza in esame.*

Matilde Trevisani
Department of Economics, Business, Mathematics and Statistics, University of Trieste, Via Tigor 22, 34124 Trieste (Italy), e-mail: matilde.trevisani@deams.units.it

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

Alessandra Petrucci, Rosanna Verde (edited by), *SIS 2017. Statistics and Data Science: new challenges, new generations. 28-30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*
ISBN (online) 978-88-6453-521-0 (online), CC BY 4.0, 2017 Firenze University Press

Key words: scientometrics, diachronic corpus, functional data analysis, cluster validation

1 Introduction

Scientometrics studies in quantitative fashion the evolution of science focusing on the analysis of publications. One of its major objectives is the development of information systems that can help to explore the enormous amount of scientific articles unceasingly published. The two main methods for automatically designing lexical maps are *citation-based analysis* and *co-word analysis*. Co-citation analysis maps the literature under consideration from the interaction of document citations whereas co-word analysis deals directly with the interaction of key terms shared by documents. Dynamical science mapping is another challenge that aims at describing dynamical patterns in science evolution.

In our study a dynamical knowledge mapping is reconstructed from a functional data analysis (FDA) perspective. The source database is a diachronic corpus which is a collection of texts including information on the time period to which they relate. In *bag-of-words* approaches, a diachronic corpus originates a words$\times$time-points contingency table displaying the frequencies of each keyword in the set of texts grouped by time-points in the observed time span. Diachronic corpora represent the ideal ground for studying the history of linguistic phenomena, e.g., when a corpus is able to reflect the relevant features of a text genre in a well-defined time period, the temporal evolution of word occurrences mirrors the historical development of the corresponding concepts [3].

This study proposes an information system consisting of (1) an information retrieval procedure for keywords' selection and (2) a functional clustering two-stage approach to identify words showing prototypical temporal patterns and cluster words portraying similar temporal patterns.

The procedure has been and is being applied to corpora of scientific papers published by leading journals of several disciplines, namely, statistics, social psychology, sociology and philosophy. This work connects to the project *Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature* (University of Padova, CPDA145940, 2015-2017), involving an interdisciplinary research group whose aim is to construct chronological corpora, and, hence, to investigate whether a discipline history can be traced from analyzing the keywords' temporal pattern. Several analyses are performed to reconstruct a dynamical evolution: correspondence analysis, topic latent Dirichlet allocation, similarity analysis (using co-occurrences), and—the object of the present work—curve clustering.

2 Material: the corpus

Our databases are collections of articles published by a selection of premier journals of the disciplines of interest over a long time period. Text under consideration consist of titles and/or abstracts and/or full texts of the scientific papers. Time is typically discretized by years according to the cadence of volume publication.

As an example, consider one of the corpora analyzed for exploring the historical evolution of Statistics. The database is the collection of papers published by the *Journal of the American Statistical Association* (JASA, 1922-) and its predecessors, *Publications of the ASA* (1888-1912) and *Quarterly Publications of the ASA* (1912-1921). Taking into account only the texts of titles including content words and disregarding items that not refer to research papers (e.g., *List of publications*, *News*, *Comment*, *Rejoinder*), the corpus includes $10,077$ titles of articles published in the period 1888-2012 (125 years, from Volume No. 1, Issue No. 1 to Volume No. 107, Issue No. 500, since at the very beginning the volumes were biennial). The corpus is composed of $87,060$ word-tokens and $7,746$ word-types. To solve the problem of identifying a set of keywords that prove relevant for the study of the history of Statistics, we adopt a stepwise procedure:

1. to overcome some of the limitations of analyses based on simple word-types, we replace words with stems by means of the popular Porter's stemming algorithm;
2. to take into account compounds, multi-words and sequences of words which have different meanings when they are considered in their context of use and together with adjacent words, we identify n-stem-grams;
3. to identify the most relevant statistical keywords, we match the vocabulary with popular statistics glossaries available on-line: ISI-International Statistical Institute; OECD-Organisation for Economic Cooperation and Development; Statistics.com-Institute for Statistics Education; StatSoft Inc.; University of California, Berkeley; University of Glasgow.
4. to reduce low frequency keywords we select keywords with frequencies ≥ 10.

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

3 Method: a functional clustering two-stage approach

From a FDA perspective, discrete observations $\mathbf{y}_i = \{y_{ij}\}$ of the frequency of a keyword $i(= 1,\ldots,N)$ in the volumes $j = 1,\ldots,T$ are viewed as a realization of an underlying continuous function $x_i(t)$. As $\mathbf{y}_i$ is a noisy observation of $x_i(t)$, an adequate model is $\mathbf{y}_i = x_i(\mathbf{t}) + \boldsymbol{\varepsilon}_i$, where $\mathbf{t} = \{t_j\}$ is the finite set of time-points and $\boldsymbol{\varepsilon}_i = \{\varepsilon_{ij}\}$ is a zero mean vector with dispersion matrix Σ_ε.

For representing functional data (FD) as smooth functions one method is the basis function approach where $x_i(t)$ is represented by a finite-dimensional linear combination $x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$, $c_{ik} \in \Re$, for sufficiently large K, of real-valued func-

tions ϕ_k called basis functions. In this study we consider B-splines as they consist in a very flexible basis for non-periodic FD. As regards the positioning of breakpoints, a direct and reasonable choice is placing knots at each time-point t_j.

We adopt the *roughness penalty* approach for estimation under which the estimate $\hat{x}_i$ is that function minimizing the penalized residual sum of squares, $\text{PENSSE} = \text{SSE} + \lambda \cdot \text{PEN}_r$, where $\text{SSE} = \{\mathbf{y}_i - x_i(\mathbf{t})\}^T W \{\mathbf{y}_i - x_i(\mathbf{t})\}$ ($W = \Sigma_\varepsilon^{-1}$) is the residual sum of squares, $\text{PEN}_r = \int [D^r x_i(s)]^2 ds$ is the penalty term and λ is a smoothing parameter.

A standard practice for choosing λ is to use the generalized cross validation, $\text{GCV}(\lambda) = T/(T - df(\lambda))^2 \text{SSE}(\hat{x}_i)$, which provides a convenient approximation to leave-one-out CV. $df(\lambda)$ is the effective degrees of freedom, which is monotone decreasing in λ with maximum equal to K when $\lambda = 0$.

We smooth the data by trying different spline orders combined with various roughness penalties and varying the smoothing parameter over an opportune range of values.

We adopted a distance-based approach, in particular the *k*-means algorithm combined with the L_2 metric to measure distance between curves. Besides the L_2 metric other measures of proximity can be considered, such as the L_1 metric, the adaptive dissimilarity index, and the correlation-based dissimilarity [2].

Cluster validation is an essential step in the cluster analysis process. Within the approaches to cluster validation [1], the use of external information is a valuable and ultimately necessary tool. Here, external information consists of an informal assessment of subject matter experts. On the other side, a large number of indexes has been proposed in the literature for a validation based on the clustered data alone. In this study we combine a large number of internal validation indexes without integrating subject matter knowledge, so as to let the data bring out the best rated groupings. Our clustering procedure is thought of as a tool of thorough investigation before submitting the results to experts who possibly will guide towards other analyses.

4 Theory: corpus data transformation

The decision about what data to use is an important part of the clustering process, and often has a fundamental impact on the resulting clusters.

If we consider the keywords×time-points table by row, a typical feature of a word trajectory is a sharp peak-and-valley trend, mainly due to the sparsity affecting frequency data of a corpus. On the other hand, if we look at data by column they appear strongly asymmetrical, in particular for the marked disparity of frequency classes between the most popular words and all of the others. This is a typical feature of word-type frequency distributions aka *large number rare events* property. Lastly, the size of time-point subcorpora may vary greatly over time.

In our research, we envision several transformations which address two different objectives: whether, in assessing two curves as similar, we should consider height (word popularity) and timing (synchrony) jointly, or timing only. In the first case

we just need to normalize data by column, in the other case we need to normalize by row, or better still, since a sort of column-normalization should be regarded as preliminary, to resort to some double normalization.

The normalization step (Table 1) of our procedure provides several transformations by colum (c_1-c_5) and by row (r_1-r_5).

Table 1 Normalization plan

	normalized by column / normalized by row	(corpus logic) subcorpus #titles	#tokens	("table" logic) column sum ($\sqrt{\cdot}$)	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	

Crossing a column- by a row-normalization generates a double normalization. Our comprehensive study examines all the transformations specifically indicated in the table. Here we present a small subset: c_2, d_1, d_3.

5 Results and conclusions

Optimal smoothing for c_2 normalized data is achieved with $m = 5$ and $\lambda = 10^3$ ($df = 7.7$) after setting a PEN_2 roughness penalty, whereas for both d_1 and d_3 normalized data the criterion lead to $m = 3$ and $\lambda = 10^{1.75}$ ($df = 7.375$) under a PEN_1 roughness penalty. Curves are then partitioned by the k-means algorithm on the basis of the Euclidean distance. The algorithm is re-run, for each k from 2 to 26, 20 times from different initial configurations set through the k-means++ seeding method.

A set of 49 quality criteria are then computed in order to identify the best partition/number(s) of clusters. By pooling rankings from all the quality indices, the frequency of being in the top-1 up to the top-4 is calculated for each cluster number k. In general partitions into two/three clusters are the best rated. This reflects the substantial bifurcation of the historical period around the sixties at which the birth of Statistics as an autonomous and established discipline can be placed. Moreover, partitions with a number of clusters close to the maximum of the considered range have also been frequently selected. This result may be a failure due to the standard assumption of data normally distributed. From the foregoing, once discarded the solutions picking the extremes, the most selected cluster numbers are: 5 for c_2, 6 for d_1 and 5 for d_2. To compare some aspects of how the three transformations affect

clustering, we consider the best partition found with the above numbers of clusters (conclusions are below).
Let us now examine some aspects of clustering, in the three cases of normalization, by varying the number of clusters (Table 2): how much groups are balanced; how many groups are singletons; how much groups are heterogeneous in being composed of words of different frequency class or popularity.

Table 2 Balance, presence of singletons and heterogeneity of frequency classes

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

A summary of conclusions follows.

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

References

1. Hennig, C., Meila, M., Murtagh, F., Rocci, R. E.: Handbook of cluster analysis. Chapman & Hall (2016).
2. Montero, P., Vilar, J.: Tsclust: An R package for time series clustering, J. Stat. Softw. **62** (1), 1–43 (2014)
3. Trevisani, M., Tuzzi, A.: A portrait of JASA: the history of Statistics through analysis of keyword counts in an early scientific journal, Quality and Quantity **49** (3), 1287–1304 (2015)