Analisi di dati testuali cronologici in corpora diacronici: effetti della normalizzazione sul curve clustering

Matilde Trevisani[1], Arjuna Tuzzi[2]

[1] DEAMS, Università di Trieste, Trieste – Italy

[2] Dip. FISPPA, Università di Padova, Padova – Italy

Abstract

In bag-of-words approaches textual data are organized in words×texts contingency tables. Diachronic corpora include texts which have a chronological order and produce words×time-points contingency tables, *i.e.* the frequencies of each word in the text (or in the set of texts) that refers to each time-point. The temporal evolution of word frequencies is crucial to highlight the distinctive features of time spans as well as to cluster words portraying a similar temporal pattern. However, to take into account the fluctuating size of available texts for each time-point, the strong asymmetry of word frequencies and the general problem of data sparsity, a transformation of data is necessary. This study aims at examining how different data transformations affect curve clustering in terms of number and composition of word groups. A functional data approach that envisages a smoothing procedure (B-splines) combined with a distance-based curve clustering has been adopted. Examples are taken from the corpus of titles of scientific papers published by the *Journal of the American Statistical Association* (and its predecessors) in the time-span 1888-2012 and consist in the analysis of the life-cycle of 900 keywords through the timeline of 107 volumes.

Riassunto

Negli approcci di tipo *bag-of-words* i dati testuali sono organizzati in tabelle di contingenza parole×testi. I corpora diacronici sono formati da testi in ordine cronologico e producono tabelle di contingenza parole×tempi, cioè frequenze di ogni parola nel testo (o nell'insieme di testi) che fa riferimento a ciascun periodo di tempo. L'evoluzione temporale delle frequenze di una parola è indispensabile per evidenziare le peculiarità dei diversi periodi e anche per raggruppare parole che mostrano un andamento simile. Tuttavia, per tenere in considerazione la diversa dimensione dei testi a disposizione per ciascun periodo di tempo, la forte asimmetria presente nella frequenza delle parole e la generale sparsità dei dati, non si può prescindere dal sottoporre i dati ad una trasformazione preliminare. Lo scopo di questo studio è esaminare come diverse trasformazioni agiscono sui risultati del *curve clustering* in termini di quantità e composizione dei gruppi di parole. Si è scelto di usare un approccio per dati funzionali che combina una procedura di lisciamento (B-splines) delle traiettorie con una di raggruppamento di tipo *distance-based*,. L'esempio riguarda un corpus di titoli di articoli scientifici pubblicati dal *Journal of the American Statistical Association* (e predecessori) nel periodo 1888-2012 e consiste nell'analisi del ciclo di vita di 900 parole chiave attraverso la linea temporale dei 107 volumi.

Key words: diachronic corpora; chronological textual data; data transformation; normalization; curve clustering; splines; functional data analysis

1. Introduzione

Un corpus diacronico è una collezione di testi corredati da informazioni sulla loro collocazione temporale. In molti casi questi testi sono raggruppati in intervalli temporali (subcorpora) con l'obiettivo di studiare comparativamente l'evoluzione di uno o più fenomeni. In quest'ottica, se il corpus riesce a rappresentare la produzione di un ben definito genere testuale in un ben definito periodo di tempo, ha senso osservare la storia del lessico attraverso l'evoluzione temporale delle occorrenze delle parole. L'idea di studiare la «qualità

della vita» delle parole è innovativa perché, a partire dai pionieri della storia della lingua italiana (Migliorini, 1960), si è sempre studiata la loro prima apparizione (datazione), ma poco è stato fatto per seguirne il destino, cioè per capire se nel corso del tempo una parola è diventata più frequente, ha avuto un momento di gloria e poi si è rarefatta, è scomparsa, ecc.

Negli approcci di tipo *bag-of-words*, un corpus diacronico viene comunemente rappresentato con una tabella di contingenza parole×tempi, dove ogni cella riporta le occorrenze di una parola nel subcorpus che fa riferimento a un periodo di tempo. Ogni riga della tabella è una serie temporale[1] di dati discreti che descrive in termini di frequenza l'evoluzione di una parola e costituisce una base di partenza ideale per studiare le peculiarità lessicali del corpus nei diversi periodi di tempo, riconoscere parole che mostrano andamenti temporali prototipici e raggruppare parole che mostrano andamenti temporali simili.

Analogamente ad altri ambiti, questi «dati grezzi» di partenza sono in realtà il frutto di una prima elaborazione basata su conteggi (Michel et al., 2011) e, siccome il numero di occorrenze di una parola non è indipendente dalla dimensione dei testi, una prima riflessione riguarda l'opportunità di trasformare le frequenze assolute in relative (per esempio, dividendole per la dimensione del subcorpus e trasformandole, così, in tassi). Ma, naturalmente, sussiste anche l'effetto della frequenza totale di una parola nell'intero corpus che, alla luce dell'accentuata variabilità, dovrebbe essere eliminato per rendere confrontabile l'andamento temporale di parole con popolarità molto diversa (come avviene, per esempio, nel calcolo della distanza del chi-quadrato). In generale, la forte asimmetria e sparsità[2] presente nella frequenza delle parole complicano ulteriormente le operazioni e la scelta. Un altro problema che sorge quando si confrontano andamenti temporali è quello dell'allineamento: due parole con una traiettoria della stessa forma ma traslata un po' più a destra o un po' più a sinistra devono essere considerate simili? Siccome la nostra risposta è negativa (i movimenti in periodi diversi non sono assimilabili) il problema di allineamento non viene considerato in questo studio.

La trasformazione dei dati da adottare dipende dagli obiettivi dell'analisi ed è fondamentale per la corretta interpretazione dei risultati. In questo studio si parte dall'ipotesi di voler riconoscere gruppi di parole con andamenti temporali simili e si esamina come diverse trasformazioni dei dati di partenza agiscono sui risultati in termini di quantità e composizione dei gruppi di parole (*clusters*) individuati. Si è scelto di usare un approccio per dati funzionali che combina una procedura di lisciamento (*B-splines*) delle traiettorie con una di raggruppamento di tipo *distance-based*. In questa prospettiva, le osservazioni discrete nel tempo della frequenza sono interpretate come una realizzazione di una funzione continua che raffigura la storia di vita di una parola. La forma di questa traiettoria viene utilizzata per confrontare e raggruppare le parole che mostrano andamenti simili.

L'esempio di applicazione riguarda un corpus di titoli di articoli scientifici pubblicati dal *Journal of the American Statistical Association* (e predecessori) nel periodo 1888-2012 e consiste nell'analisi del ciclo di vita di 900 parole chiave della statistica attraverso la linea temporale scandita dai 107 volumi della rivista (Trevisani e Tuzzi, 2015).

[1] In questo lavoro, le serie temporali non vengono considerate perché hanno obiettivi diversi dallo studio della forma delle traiettorie: mirano, infatti, a studiare la correlazione delle osservazioni nel tempo e a costruire modelli di previsione.

[2] Una matrice è sparsa quando è in larga parte costituita da zeri.

2. Il corpus

Le pubblicazioni dell'*American Statistical Association* sono una fonte molto autorevole per lo studio dell'evoluzione temporale di concetti, metodi e applicazioni di ambito statistico e, grazie alla loro lunga storia, ci permettono di risalire alla fine dell'Ottocento attraverso tre diverse riviste: *Publications of the American Statistical Association* (PASA, 1888-1912), *Quarterly Publications of the American Statistical Association* (QASA,1912-1921) e *Journal of the American Statistical Association* (JASA, 1922-corrente).

	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

Figura 1. Estratto della tabella di contingenza.

La tabella di contingenza (Figura 1), che costituisce la base delle analisi di questo lavoro, è il frutto di un lungo processo di costruzione del dato, che può essere sintetizzato in sei passi:

1. sono stati raccolti i titoli di tutti gli articoli pubblicati dalle tre riviste per un totale di 12.557 titoli, in 107 volumi, che coprono un arco di 125 anni (1988-2012, i primi volumi sono biennali);
2. dal corpus sono stati eliminati i titoli che non fanno riferimento a veri e propri articoli (*List of publications*, *News*) e non contengono parole chiave (*Comment*, *Rejoinder*) pervenendo, così, a un totale di 10.077 titoli e a un corpus che include 87.060 occorrenze e 7.746 forme;
3. le forme sono state trasformate in "radici" mediante la versione 2012 della procedura di *stemming* dell'algoritmo di Porter (1980) ottenendo un vocabolario di 4.834 *stem* (es. le forme *model*, *models*, *modeling* e *modelling* sono state sostituite dalla stessa radice *model*);
4. sono stati riconosciuti mediante la procedura presente nel software Taltac tutte le sequenze (*n-stem-grams* come *model select*, *addit model*, *hierarch model*, *log linear model*, *dynam model*) presenti nel corpus almeno due volte e costituite da un minimo di due e un massimo di 6 *stem* consecutivi;
5. per ricavare le parole chiave della statistica, il vocabolario è stato confrontato con una lista di 12.700 termini ottenuta dall'unione di sei noti glossari (*International Statistical Institute, Organisation for Economic Cooperation and Development*, *Institute for Statistics Education*, *StatSoft* Inc., *University of California* – Berkeley, *University of Glasgow*);

6. i titoli sono stati raggruppati per volume (da 1 a 107) e sono state scelte le 900 parole chiave con frequenza almeno pari a 10.

Per studiare la relazione tra parole chiave, tra volumi e tra parole chiave e volumi, è stata usata in uno studio precedente l'analisi delle corrispondenze ed è emerso un chiaro andamento cronologico (Trevisani e Tuzzi, 2015). Sebbene risulti molto utile per ottenere una rappresentazione grafica efficace, l'analisi delle corrispondenze non è, tuttavia, in grado di far emergere il contributo degli andamenti delle singole parole. Spostando la prospettiva di indagine nell'ambito dell'analisi di dati funzionali, i metodi di *curve clustering* con rappresentazione delle traiettorie mediante funzioni base (*splines*, *wavelets*) si rivelano più efficaci per tracciare e comparare l'evoluzione temporale delle singole parole (Trevisani e Tuzzi, 2012; 2015).

3. Metodo

In una prospettiva di analisi per dati funzionali (FDA), le osservazioni discrete y_{ij} della frequenza di una parola chiave i (1…900) nel volume j (1…107) sono considerate come una realizzazione di una funzione continua $x_i(t)$, sufficientemente liscia (*smooth*) o regolare, che ne rappresenta l'evoluzione temporale:

$$y_{ij} = x_i(t_j) + "rumore"$$

Il "rumore" (*noise*) tiene conto del fatto che le osservazioni sono disturbate, ossia affette da un termine di errore ε_{ij} di media nulla e, nel modello standard, varianza costante σ^2 o, più opportunamente, fatta dipendere dal tempo Σ_ε (matrice di dispersione per il vettore degli errori riferiti agli istanti t_j).

3.1. Lisciamento con B-splines

Per rappresentare i dati come funzioni regolari, un approccio comunemente usato è quello di esprimere $x_i(t)$ come una combinazione lineare finita di funzioni a valori reali ϕ_k, dette funzioni base,

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \; (c_{ik} \in \Re)$$

per K sufficientemente grande (Ramsay e Silverman, 2005).

I sistemi di basi più noti in letteratura sono quelli costituiti da funzioni di tipo monomiale, Fourier e spline, oltre che di tipo wavelet, esponenziale e di potenza, anche se meno usati.

In questo studio abbiamo scelto le B-splines per la loro flessibilità e capacità di rappresentare funzioni complesse per dati non ricorrenti. Esse formano un particolare sistema di basi tramite cui costruire *splines*, cioè funzioni costituite da polinomi che si saldano tra loro in modo "liscio" dopo che l'intervallo di definizione è stato suddiviso in più sotto-intervalli fissando una sequenza di "nodi" (*knots*).

L'approccio scelto per lisciare i dati funzionali è quello noto come regolarizzazione (*regularization*) o "penalizzazione dell'irregolarità" (*roughness penalty*) attraverso cui x_i è stimata come funzione che minimizza la somma dei quadrati penalizzata

$$PENSSE_\lambda(x|\boldsymbol{y}_i) = [\boldsymbol{y}_i - x_i(\boldsymbol{t})]'W[\boldsymbol{y}_i - x_i(\boldsymbol{t})] + \lambda \cdot PEN(x_i)$$

dove $\boldsymbol{y}_i$ e $x_i(\boldsymbol{t})$ sono, rispettivamente, i vettori del dato funzionale i e della funzione sottesa, riferiti al vettore degli istanti $\boldsymbol{t}$, W è il reciproco di Σ_ε, $PEN(x_i)$ una misura dell'irregolarità

della funzione. Il bilanciamento tra bontà di adattamento ai dati, misurata dalla somma dei quadrati degli errori, e irregolarità della funzione è realizzato dal parametro di lisciamento λ.

Il criterio scelto rende esplicita l'idea di attuare un *trade-off* tra *bias* e varianza della stima, che qui in particolare si attua attraverso la funzione di perdita nota come *mean squared error*.

Nel presente lavoro i nodi sono stati fatti coincidere con i punti temporali di osservazione, mentre per la scelta della penalità si è considerato, come d'uso, il quadrato della derivata integrato sull'intervallo di osservazione. L'ordine della derivata è stato fatto variare per provare diversi gradi di lisciamento.

Per la scelta del numero *K* di funzioni base, definito, nell'approccio di stima adottato, dal parametro di lisciamento λ, abbiamo usato il metodo noto come *generalized cross validation* (GCV): un'estensione del metodo *cross validation* che ne ovvia la tendenza a "sotto-lisciare" i dati (Ramsay et al., 2009).

3.2. Distance-based curve clustering

L'andamento temporale delle occorrenze nel corpus viene considerato come una *proxy* del ciclo di vita delle parole chiave. L'obiettivo è riconoscere la forma delle traiettorie per poter costruire gruppi di parole che condividono lo stesso destino.

In generale, i metodi di raggruppamento possono essere suddivisi in metodi *soft,* o *model-based,* e *hard,* o *distance-based*. In un contesto FDA, i metodi *soft* assumono che i dati funzionali siano realizzazioni di un processo mistura, dove i pesi della mistura corrispondono alle probabilità dei gruppi; l'appartenenza ad un gruppo non è fissata ma segue una distribuzione multinomiale, per cui un modello di *soft-clustering* equivale a una mistura di densità. I metodi *hard* creano una partizione dei dati funzionali in gruppi, generalmente disgiunti, secondo una misura di similarità tra traiettorie.

Mentre in un precedente lavoro abbiamo seguito un approccio *model-based* (Trevisani e Tuzzi, 2015), nel presente studio abbiamo adottato una tecnica di *hard-clustering* che mira a mettere a punto una procedura esplorativa e automatizzata per l'analisi del corpus, che possa essere ragionevolmente usufruibile anche da parte di ricercatori di altre discipline impegnati nello studio di corpora diacronici con struttura simile[3]. In particolare, abbiamo utilizzato un algoritmo di partizionamento *K-means*, provando vari tipi di distanza (euclidea, di Manhattan, dissimilarità basata sulla correlazione, indice di dissimilarità adattivo – sia con distanza euclidea che con *Dynamic Time Warping*), e vagliato diversi criteri di qualità del clustering per la selezione del numero di cluster.

4. Trasformazioni

Obiettivo del presente lavoro è studiare come diverse trasformazioni dei dati contenuti nella tabella di contingenza influiscano sui risultati del *curve clustering*.

Vari tipi di trasformazione o normalizzazione della tabella parole×tempi sono pensabili: per colonna, al fine di omogeneizzare la dimensione dei subcorpora variabile nel tempo; per riga, sia al fine di omogeneizzare le traiettorie delle parole chiave depurandole della loro frequenza totale o popolarità nel corpus, sia al fine di risolvere il problema della sparsità (eccesso di "zeri" nelle traiettorie) e, infine, doppia, cioè sia per colonna che per riga.

[3] Rif: Progetto di Ateneo "Tracing the History of Words. A Portrait of a Discipline Through Analyses of Keyword Counts in Large Corpora of Scientific Literature" finanziato dall'Università di Padova.

La normalizzazione per colonna – operazione che ci pare preliminare a ogni altro successivo trattamento – può essere effettuata, per esempio, dividendo le frequenze:
(c_1) per il totale di colonna (il numero di occorrenze totali del volume della rivista);
(c_2) per il totale di parole chiave presenti nel subcorpus (parole chiave contenute nel volume);
(c_3) per il totale di testi contenuti nel subcorpus (numero di titoli del volume).

La normalizzazione per riga può essere realizzata, per esempio:
(r_1) dividendo le frequenze per il totale di riga (frequenza della parola nell'intero corpus);
(r_2) dividendo le frequenze per la massima frequenza presente
(r_3) calcolando lo *z-score* di ogni riga.

Le normalizzazioni che in questo studio vengono confrontate sono:

- per colonna, c_2: $y_{ij} = n_{ij}/N_j$, dove N_j è il totale di parole chiave nel subcorpus j;
- doppia, d_1 (χ^2): $y_{ij} = (n_{ij}/n_{i.})/\sqrt{N_j}$, dove $n_{i.}$ è il totale di riga;
- doppia, d_2: $y_{ij} = (n_{ij}/M_i)/N_j$, dove M_i è il massimo di riga.

5. Risultati

Una volta trasformati i dati, per ciascuno dei tre casi di normalizzazione, si è ottenuto il lisciamento ottimale tramite *B-splines* facendo variare sia la funzione di penalizzazione $PEN_d(x_i)$ (ordine della derivata d = 0, 1, 2 e m-2, con m ordine della spline) sia, in un opportuno range di valori, l'ordine della *spline* e il parametro di lisciamento. Sulla base del criterio GCV è stato scelto di lisciare i dati trasformati secondo c_2 con *B-splines* di ordine m = 5 e gradi di libertà gdl = 7,7 ($\lambda = 10^3$) avendo imposto una penalizzazione di tipo PEN_2, mentre sia per i dati trasformati secondo d_1 che secondo d_2 il lisciamento ottimale è risultato fissando m = 3 e gdl = 7,4 ($\lambda = 10^{1,75}$) dopo aver imposto una penalizzazione di tipo PEN_1.

Le curve ottenute nei tre casi sono state partizionate con il metodo *K-means* usando vari tipi di distanza (di cui qui illustriamo solo i risultati con la più nota distanza euclidea); inoltre facendo variare il numero di cluster da 2 a 26 e, per ognuno di questi, eseguendo 20 replicazioni. Al fine di valutare quale numero di cluster emerge come migliore nel cogliere la struttura di raggruppamento sottesa ai dati, sono stati esaminati in maniera congiunta i valori di una numerosa serie di criteri di qualità di una partizione (una cinquantina tra cui gli indici di Calinski-Harabasz, Davies-Bouldin, Dunn, Ray-Turi, silhouette, BIC, AIC, etc., prendendo spunto dalle liste contenute in Genolini et al., 2015, e Desgraupes, 2015). Anticipiamo che in tutte e tre le trasformazioni le partizioni a due o tre cluster sono risultate in generale come le migliori; esito che riflette la sostanziale bipartizione dell'arco storico considerato in due periodi: uno antecedente e uno successivo alla nascita della Statistica come disciplina autonoma e affermata, collocabile intorno agli anni '60. Anche le partizioni con un numero di cluster prossimo all'estremo del *range* considerato (25, 26) sono state frequentemente prescelte dai criteri, risultato che, da un lato, riflette la mancanza di una struttura definita e parsimoniosa di raggruppamento ma, dall'altro, è anche un esito prevedibile per la massiccia presenza di criteri di qualità che sottendono ai dati una distribuzione normale multivariata (che è l'ipotesi standard del *model-based clustering* ma è assunta anche da diversi criteri di qualità e dallo stesso algoritmo *K-means*). Esaminando quali partizioni sono state più spesso valutate migliori, o anche come seconda/terza/quarta scelta, possono essere giudicate pressoché *ex-aequo*:

- nel caso della trasformazione c_2, le soluzioni con 5/9/22/6 gruppi,
- nel caso della trasformazione d_1, quelle con 6/4/19/12 gruppi,
- nel caso della trasformazione d_2, quelle con 5/6/20/12 gruppi (Figura 2).

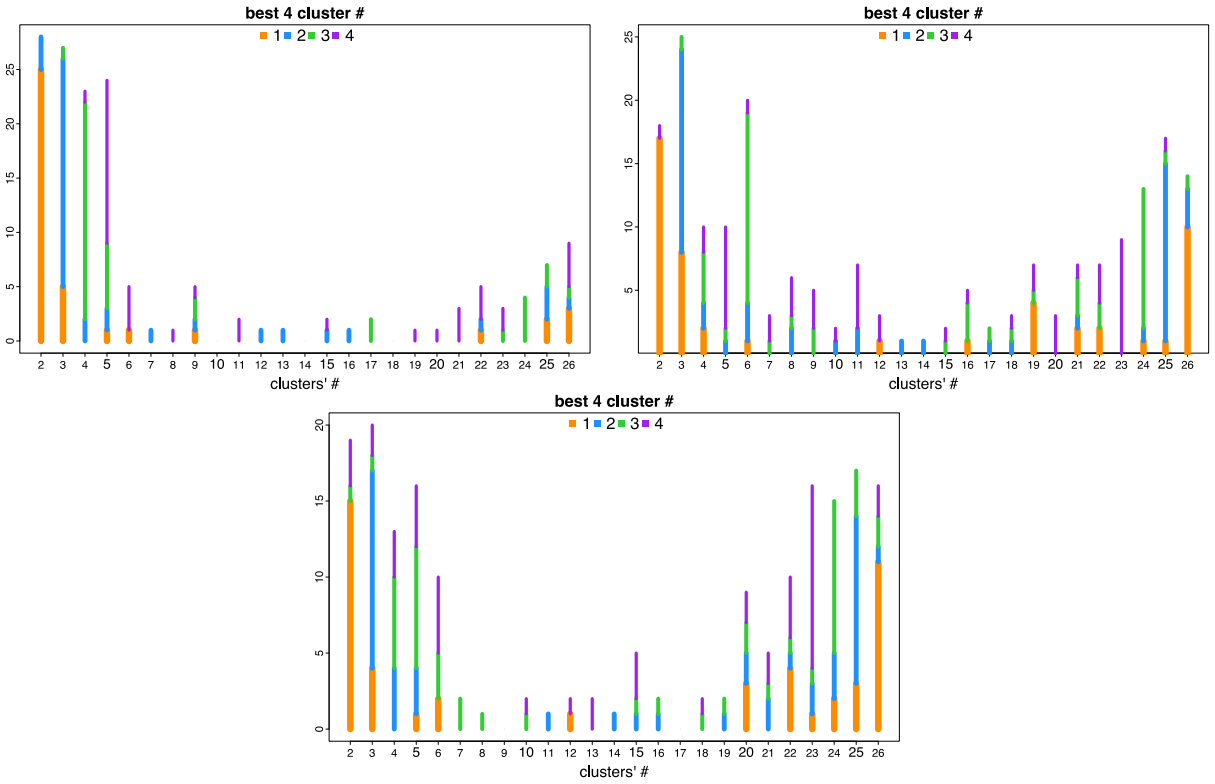

Figura 2. Selezione del numero di cluster nei tre casi di normalizzazione: c_2 (in alto a sinistra), d_1 (in alto a destra) e d_2 (in basso).

Per confrontare alcuni aspetti dell'effetto delle tre diverse trasformazioni sui risultati del clustering, presentiamo la partizione risultata tra le migliori con un numero contenuto di cluster nei tre casi esaminati: a cinque per c_2 (Figura 3), a sei per d_1 (Figura 4), a cinque per d_2 (Figura 5).

Nel caso dei dati normalizzati solo per colonna, la "popolarità" della parola assume un ruolo dominante: i cluster sono principalmente determinati dalle curve di livello alto (parole di elevata frequenza) e la maggior parte delle parole a bassa frequenza viene ammassata in uno o più gruppi "amorfi". Nell'esempio del clustering a cinque gruppi sui dati normalizzati secondo c_2 (Figura 3), solo tre cluster – che raccolgono circa il 10% del totale delle parole – appaiono di un certo interesse, mentre i due rimanenti sono rappresentati da un singleton (*statist*, la parola più frequente in assoluto che forma il gruppo E) e da un indistinto agglomerato di parole a bassa frequenza (il massivo gruppo A). Al contrario, una normalizzazione doppia consente un partizionamento più bilanciato, dove sia la forma che il livello delle curve giocano, in generale, un ruolo alla pari. Nel risultato del clustering a sei gruppi sui dati normalizzati secondo d_1 (Figura 4) e in quello a cinque secondo d_2 (Figura 5), alcuni *patterns* già emersi per la trasformazione c_2 vengono confermati e meglio strutturati con la costruzione di gruppi più numerosi: il lungo arco temporale è chiaramente diviso da cluster di parole che condividono cicli di vita simili. Ad esempio, il gruppo C per c_2, che rappresenta il già citato periodo antecedente alla nascita della statistica come disciplina autonoma, si articola temporalmente nei gruppi F (demografia, studi sulla popolazione e statistica pubblica) e C (statistica economica e sociale) per d_1 e si sostanzia nel gruppo B per d_2; il gruppo B per c_2 si suddivide temporalmente nei gruppi A (la statistica nasce e si afferma come disciplina autonoma con un proprio vocabolario), D (il "periodo d'oro" della statistica classica, anni '60-'80), B (la statistica contemporanea) ed E (la statistica moderna) per d_1.

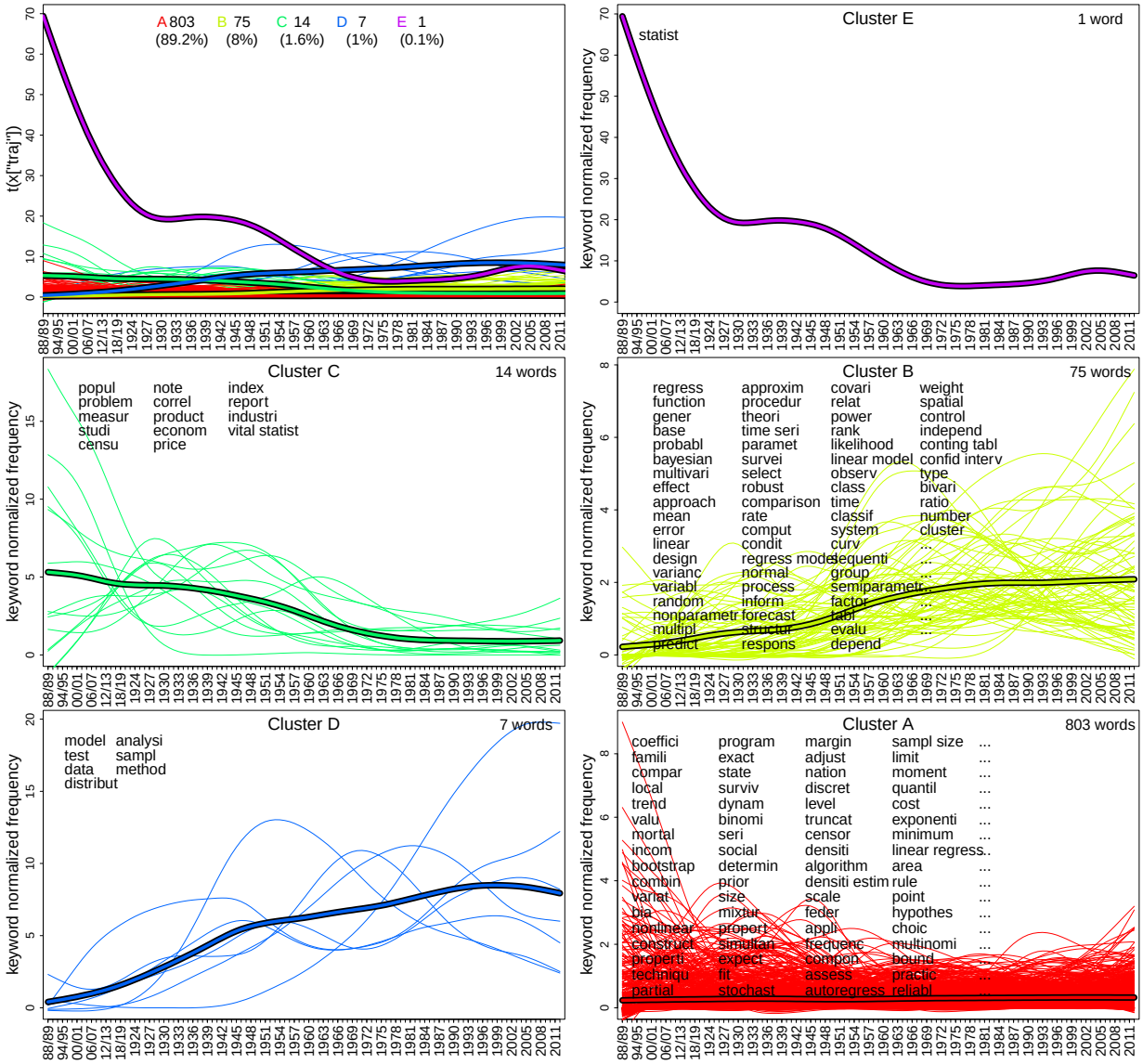

Figura 3. Clustering sui dati normalizzati secondo c_2: tutti i cinque gruppi (in alto) e i cinque cluster individuali

Infine, il gruppo D per c_2 contiene le parole "base" della statistica, che si distribuiscono in cluster differenti nelle partizioni generate dai dati con doppia normalizzazione. Il clustering con d_2, pur presentando forti analogie con il raggruppamento appena esaminato per d_1, appare cadenzare l'intervallo temporale in periodi meno netti e creare gruppi trasversali. Una possibile interpretazione è che, mentre il clustering con d_1 è principalmente determinato dalla assenza/presenza delle parole lungo il tempo (in pratica, da quali sono i momenti di vitalità delle parole e quali, invece, i periodi di assenza), con d_2 sono soprattutto le peculiarità della vita delle parole, cioè la pura forma delle traiettorie, a comporre i gruppi.

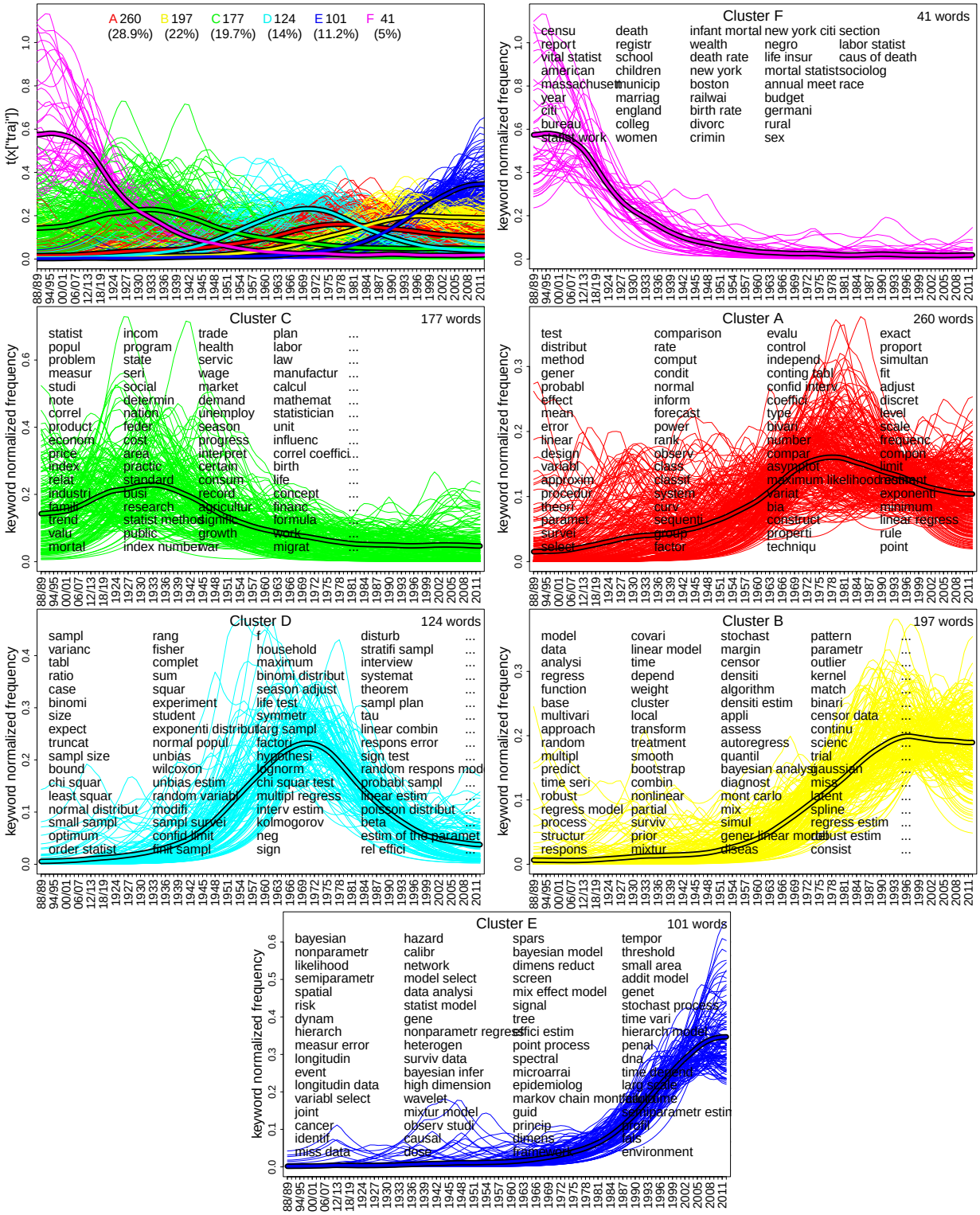

Figura 4. Clustering sui dati normalizzati secondo d_1: tutti i sei gruppi (in alto) e i sei cluster individuali

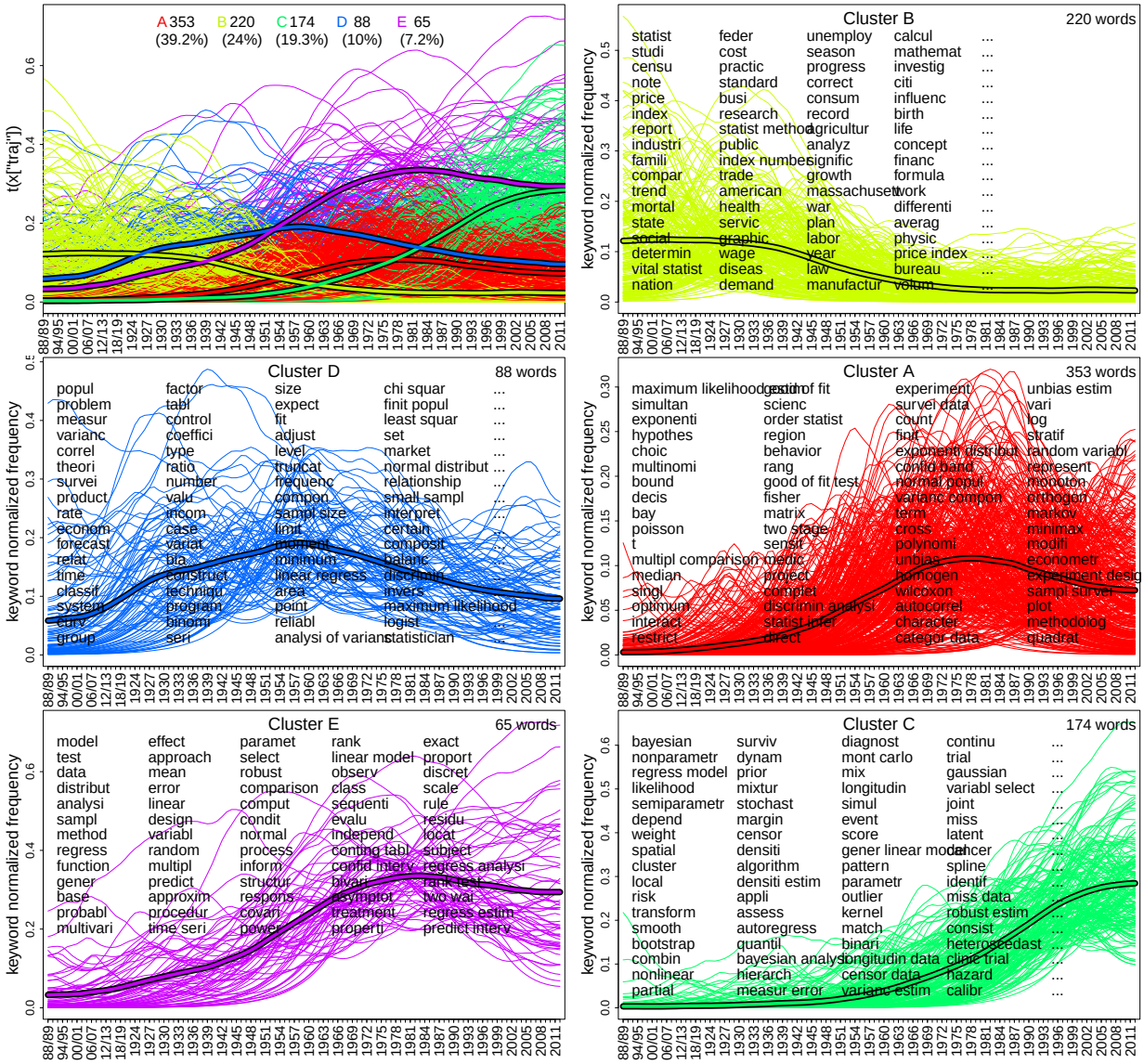

Figura 5. Clustering sui dati normalizzati secondo d_2: tutti i cinque gruppi (in alto) e i cinque cluster individuali

Analizziamo ora alcuni aspetti del clustering, al variare del numero di gruppi, nei tre casi di normalizzazione (Tabella 1):

- quanto i gruppi sono bilanciati (ovvero l'eterogeneità della distribuzione nei gruppi);
- quanti *singleton* sono presenti;
- quanto i gruppi sono eterogenei nella composizione rispetto alla presenza di parole con diversa classe di frequenza o popolarità.

Sia il grado di bilanciamento che quello di eterogeneità rispetto alle classi di frequenza sono stati misurati tramite l'indice di Gini normalizzato (considerando la mediana dei valori calcolati per le 20 replicazioni per ogni numero fissato di cluster); il numero di singleton è quello massimo rinvenuto nelle 20 replicazioni. Nel caso della trasformazione c_2, emerge il forte sbilanciamento nella numerosità dei gruppi, di cui un aspetto particolare è la presenza elevata di singleton, e la dominanza di una sola classe di frequenza nella composizione dei gruppi (è ormai noto che si tratta delle parole di frequenza molto elevata). Al contrario, nelle

normalizzazioni doppie i gruppi generati appaiono bilanciati, con un'esigua presenza di singleton, ed eterogenei nella composizione quanto alla classe di frequenza delle parole.

Tabella 1. Bilanciamento, presenza di *singleton* ed eterogeneità delle classi di frequenza nei cluster

	# cluster normalizz.	2	3	4	5	6	7	8	9	10	15	20	25
Bilanciamento	c_2	0,00	0,12	0,26	0,29	0,44	0,49	0,56	0,59	0,63	0,80	0,86	0,91
	d_1	0,72	0,93	0,90	0,90	0,94	0,96	0,95	0,97	0,97	0,98	0,98	0,99
	d_2	0,84	0,88	0,92	0,93	0,93	0,95	0,95	0,96	0,97	0,97	0,98	0,99
Singleton	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_2	0	0	0	0	1	0	0	1	0	3	5	5
Eterogeneità classi frequenza	c_2	1,00	0,50	0,06	0,09	0,02	0,02	0,05	0,09	0,09	0,11	0,05	0,12
	d_1	1,00	1,00	1,00	0,99	0,99	0,99	0,98	0,98	0,97	0,96	0,95	0,94
	d_2	0,90	0,95	0,95	0,93	0,81	0,85	0,80	0,82	0,80	0,80	0,78	0,77

7. Conclusioni

L'analisi di tre esempi di trasformazione delle frequenze assolute del corpus ha messo in evidenza l'influenza che il tipo di normalizzazione adottata può avere sui risultati del *curve clustering*. In sintesi:

- la sola normalizzazione per colonna della matrice parole×tempi mantiene inalterato il diverso livello di popolarità delle parole e provoca una dominanza delle parole con alta frequenza sui risultati della partizione. Lo sbilanciamento rilevante nella numerosità dei gruppi, la presenza consistente di singleton, la scarsa eterogeneità nella composizione dei gruppi di parole con diversa popolarità e, infine, la presenza di uno o più gruppi "amorfi", composti quasi esclusivamente da parole a bassa frequenza, sono alcuni degli effetti più evidenti di questo tipo di trasformazione.
- La normalizzazione doppia, ossia per riga e per colonna, se è vero che produce gruppi in genere ben bilanciati, rari *singleton* e la quasi totale assenza di gruppi "amorfi", fa perdere l'informazione sulla popolarità delle parole.
- La normalizzazione doppia d_1, che riproduce in un certo senso la distanza del chi-quadro, tende a raggruppare parole che hanno un ciclo simile quanto a presenza/assenza, nascita/morte, lungo l'arco temporale considerato, mentre la variante d_2, che più propriamente "normalizza" la frequenza (dividendo per la massima frequenza osservata), nel costruire i gruppi guarda prioritariamente alla forma della curva, ossia a quanto la "popolarità relativa" di una parola sia stata costante nel tempo o abbia avuto oscillazioni (e quali) durante il suo ciclo di vita.

Nel piano di lavoro futuro si prevede lo studio di altri tipi di normalizzazione, in particolare alcune trasformazioni che riescano a superare il problema dell'eccesso di "zeri" (dovuto all'accentuata sparsità dei dati testuali). Inoltre, relativamente al clustering, si intende approfondire il discorso sia sul piano tecnico, ad esempio studiando l'effetto dei diversi tipi di distanza per misurare la similarità tra traiettorie, che su un piano più metodologico, riferendoci in particolare alla linea di pensiero da tenere nella scelta finale del numero di cluster. Infine, parallelamente allo studio del curve clustering di tipo *distance-based* si proseguirà nella rassegna e proposta di approcci *model-based*, dove l'intreccio tra trasformazione preliminare dei dati e assunzione di modelli di probabilità diventa ancora più complesso.

Riferimenti

Desgraupes B. (2015). *clusterCrit: ClusteringIindices*. R package version 1.2.6, URL https://cran.r-project.org/web/packages/clusterCrit/index.html.

Genolini C., Alacoque X., Sentenac M. and Arnaud C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65 (1): 1-34.

Guérin-Pace F., Saint-Julien T. and Lau-Bignon A.W. (2012). The Words of L'Espace géographique: A Lexical Analysis of the Titles and Keywords from 1972 to 2010. *L'Espace géographique* 41 (1): 4-31.

Michel, J.-B., Shen, Y.K., Aiden A.P., Veres A., Gray M.K., Pickett J.P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A. and Aiden E.L (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014): 176-182.

Migliorini B. (1960). *Storia della lingua italiana*. Sansoni, Firenze.

Porter M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130-137.

Ramsay J.O., Hooker G. and Graves S. (2009). *Functional Data Analysis with R and MATLAB*. Springer, New-York.

Ramsay J.O. and Silverman B.W. (2005). *Functional Data Analysis. Second edition.* Springer, New-York.

Salem A. (1991). Les séries textuelles chronologiques. *Histoire & Mesure*, 6 (1-2). Séries temporelles: 149-175.

Trevisani M. and Tuzzi A. (2015). A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal. *Quality and Quantity*, 49:1287-1304.

Trevisani M. and Tuzzi A. (2014). Shaping the history of words. In Obradović I., Kelih E. and Köhler R., editors, *Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO), Belgrade, Serbia, April 16-19, 2012*, Akademska Misao, Belgrado, Serbia: 84-95.