

normalizzazioni doppie i gruppi generati appaiono bilanciati, con un'egua 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 d₁, 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 d₂, 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: ClusteringIndices*. 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.