Bibliometric Analysis in Stata
Text Mining Network Community Detection and Unsupervised Learning

Carlo Drago — University Niccolò Cusano, Italy

Gentian Hoxhalli — Luarasi University and Armed Forces Academy, Albania

Abstract

This study presents a Stata-centered workflow integrating text mining, unsupervised learning, and network analysis to map the thematic structure of energy-policy research. Building on established approaches to bibliometric science mapping (Zupic & Čater, 2015; Aria & Cuccurullo, 2017; Donthu et al., 2021), it combines Stata with its integrated Python environment to analyse a large OpenAlex corpus of article titles and a Scopus dataset of abstracts (Priem et al., 2022; Baas et al., 2020).

The workflow links three complementary analytical perspectives. First, transparent linguistic preprocessing, domain-specific filtering, and frequency-based selection produce an interpretable keyword dictionary and an article-by-keyword matrix. Principal component analysis summarizes the principal dimensions of variation in this matrix, and k-means clustering of the retained component scores groups articles with similar thematic profiles (Jolliffe & Cadima, 2016; Lloyd, 1982). Second, a separate analysis of abstracts uses cosine dissimilarity to identify groups of documents with similar textual content (Salton & Buckley, 1988). Third, keyword co-occurrence networks represent relationships among terms, and community detection identifies interconnected thematic cores (Callon et al., 1983; Fortunato & Hric, 2016). This distinguishes clusters of documents from communities of keywords, allowing complementary interpretations of the literature's conceptual organization.

The network perspective connects with previous bibliometric applications to energy resilience and telemedicine (Gatto & Drago, 2020; Drago, Gatto, & Ruggeri, 2023), and with recent developments in keyword-ensemble community detection applied to remote work and sustainable land planning (Drago & Errichiello, 2026; Morelli et al., 2026). Related conference-based research on AI and knowledge management further situates this approach within an evolving methodological programme (Drago, Giglio Moro, & Leogrande, 2026b).

Stata coordinates data preparation, descriptive statistics, visualization, and export, while Python supplies text-processing and machine-learning routines. Outputs include the keyword dictionary, component loadings and scores, article-level cluster memberships, and cluster-level bibliometric profiles. The contribution is an integrated and interpretable workflow for bibliometric exploration within a Stata-centered research environment. Database coverage, missing text, preprocessing decisions, and clustering choices require explicit reporting and sensitivity assessment when interpreting the resulting maps (Grimmer & Stewart, 2013; Culbert et al., 2025).

Keywords: bibliometrics; Stata; Python; text mining; principal component analysis; k-means; cosine dissimilarity; keyword co-occurrence; community detection; energy policy.

Bibliography

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies, 1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235. https://doi.org/10.1177/053901883022002003

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402. https://doi.org/10.1002/asi.21525

Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics, 130*(4), 2475–2492. https://doi.org/10.1007/s11192-025-05293-3

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Drago, C. (2018). MCA-based community detection. In F. Mola, C. Conversano, & M. Vichi (Eds.), *Classification, (big) data analysis and statistical learning* (pp. 59–66). Springer. https://doi.org/10.1007/978-3-319-55708-3_7

Drago, C. (2024a). A network-based community detection ensemble bibliometric analysis of knowledge assets and competitive advantage. In *Proceedings IFKAD 2024: Translating knowledge into innovation dynamics* (pp. 15–29). Institute of Knowledge Asset Management. Official proceedings record

Drago, C. (2024b). Ensemble-based community detection bibliometric analysis of sustainability trends in digital economics. In *Book of abstracts of the 8th Business Systems Laboratory International Symposium* (pp. 45–48). Business Systems Laboratory. Official book of abstracts

Drago, C. (2024c, May 9). *Text mining in economics and health economics using Stata* [Conference presentation and slides]. 18th Italian Stata Conference, Florence, Italy. Official conference archive

Drago, C. (2025). Technology-driven economic inequality: A ensemble keyword community detection approach of the literature. In *Book of abstracts of the 9th Business Systems Laboratory International Symposium* (pp. 138–140). Business Systems Laboratory. Official book of abstracts

Drago, C., & Aliberti, L. A. (2019). Interlocking directorship networks and gender: A bibliometric analysis. In P. Paoloni & R. Lombardi (Eds.), *Advances in gender and cultural research in business and economics* (pp. 115–136). Springer. https://doi.org/10.1007/978-3-030-00335-7_8

Drago, C., & Errichiello, L. (2026). Remote work during pandemics: a systematic review based on keyword-ensemble community detection. *Quality & Quantity, 60*(5), 14643–14689. https://doi.org/10.1007/s11135-026-02700-6

Drago, C., & Gatto, A. (2023). Comments on “Detecting the research trends and evolution of energy resilience: A bibliometric analysis” by Yu et al. [Commentary]. *Environmental Science and Pollution Research, 30*(19), 57203–57204. https://doi.org/10.1007/s11356-023-25999-2

Drago, C., Gatto, A., & Ruggeri, M. (2023). Telemedicine as technoinnovation to tackle COVID-19: A bibliometric analysis. *Technovation, 120*, 102417. https://doi.org/10.1016/j.technovation.2021.102417

Drago, C., Giglio Moro, F. V., & Leogrande, A. (2025). Bibliometric ensemble-based community detection analysis of AI, knowledge management, and emerging technologies convergence in business paradigms. In G. Bruno, C. Ponsiglione, C. Piccolo, & G. Schiuma (Eds.), *Proceedings IFKAD 2025: Knowledge futures: AI, technology, and the new business paradigm* (pp. 1436–1440). Institute of Knowledge Asset Management. Official proceedings record

Drago, C., Giglio Moro, F. V., & Leogrande, A. (2026a). Bias in AI-driven recruitment: An ensemble-based community detection bibliometric analysis of challenges in hiring and employment. In I. Novo-Corti, X. Picatoste, & M. Valeri (Eds.), *Human resource development for sustainability and social responsibility* (pp. 33–42). Springer. https://doi.org/10.1007/978-3-032-09683-8_3

Drago, C., Giglio Moro, F. V., & Leogrande, A. (2026b). Bibliometric ensemble-based community detection analysis of AI, knowledge management, and emerging technologies convergence in business paradigms. In G. Schiuma & A. Lerro (Eds.), *IFKAD25 knowledge insights: Exploring knowledge futures* (Lecture Notes in Networks and Systems, Vol. 1940, pp. 295–301). Springer. https://doi.org/10.1007/978-3-032-23684-5_33

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports, 659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Gatto, A., & Drago, C. (2020). A taxonomy of energy resilience. *Energy Policy, 136*, 111007. https://doi.org/10.1016/j.enpol.2019.111007

Gatto, A., Drago, C., & Ruggeri, M. (2023). On the frontline—A bibliometric study on sustainability, development, coronaviruses, and COVID-19. *Environmental Science and Pollution Research, 30*(15), 42983–42999. https://doi.org/10.1007/s11356-021-18396-0

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Giglio Moro, F. V., & Drago, C. (2025a). Addressing gender bias in AI business implementation: A latent Dirichlet allocation-based analysis [Conference abstract]. In *Book of abstracts IFKAD 2025: Knowledge futures: AI, technology, and the new business paradigm* (p. 51). IFKAD. Official book of abstracts

Giglio Moro, F. V., & Drago, C. (2025b). Workforce impact of artificial intelligence: A bibliometric analysis. In *Book of abstracts of the 9th Business Systems Laboratory International Symposium* (pp. 148–151). Business Systems Laboratory. Official book of abstracts

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grund, T. (2015). *Social network analysis using Stata* [Conference tutorial]. UK Stata Users Group Meeting. Official conference slides

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference* (pp. 11–15). https://doi.org/10.25080/TCWV9851

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/bf01908075

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks. *Scientific Reports*, *2*, 336. https://doi.org/10.1038/srep00336

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/tit.1982.1056489

Morelli, G., Gurrieri, A. R., Drago, C., Gatto, A., & Magazzino, C. (2026). A biblio-ecological study on sustainable land planning research: Evidence from machine learning methods. *Socio-Economic Planning Sciences*, *105*, 102431. https://doi.org/10.1016/j.seps.2026.102431

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.01833

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

StataCorp. (2025). *Stata 19 programming reference manual* [Software documentation; PyStata integration]. Stata Press. https://www.stata.com/manuals/p.pdf

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z

van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651. https://doi.org/10.1002/asi.21075

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Williams, U., & Williams, S. P. (2014). txttool: Utilities for text analysis in Stata. *The Stata Journal*, *14*(4), 817–829. https://doi.org/10.1177/1536867X1401400407

Zuo, X. (2023). *Comparing with Python: Text analysis in Stata* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.10480

Zupic, I., & Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, *18*(3), 429–472. https://doi.org/10.1177/1094428114562629