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

Parts I & II – Network communities and abstract clustering

Carlo Drago[1] Gentian Hoxhalli[2]

[1]University Niccolò Cusano, Rome, Italy
[2]Luarasi University & Armed Forces Academy, Tirana, Albania

Italian Stata User Meeting – Milano, 2026

Motivation

- Empirical economics increasingly relies on large text corpora: publications, reports, policy documents, press releases.
- Bibliometric analysis maps the **thematic structure** of a research field in a transparent, reproducible way.
- Existing platforms (VOSviewer, Bibliometrix, CiteSpace) are useful but sit *outside* the empirical workflow of Stata users.
- Question: *can we run a full bibliometric pipeline entirely inside a Stata-centered environment?*
- Yes – by combining Stata with its integrated Python engine (`python:` block, Stata 16+).

Goal of Part I

Introduce bibliometric analysis as a methodology for Economics, describe it in theoretical and operational terms, and present a first application on the OpenAlex corpus retrieved through the query `primary_topic_id = t10438` (Energy, Environment, Economic Growth).

Motivation

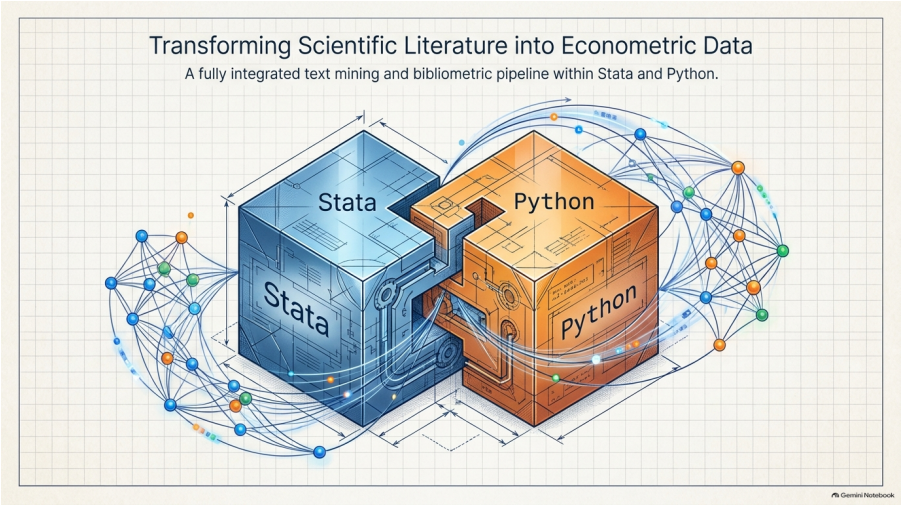

Figure: Integrated Text Mining for Empirical Economics

Motivation

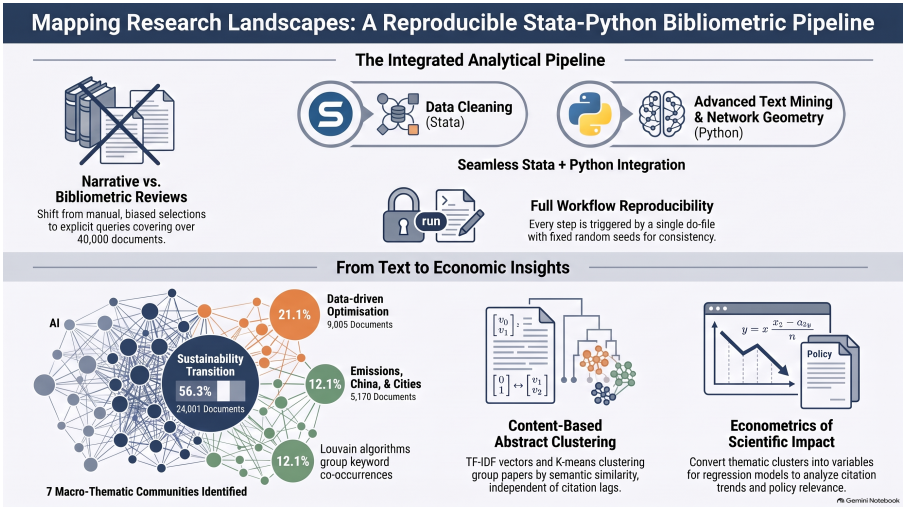

Figure: Mapping Research Landscape

Roadmap of Part I

1. Bibliometric analysis and its role in Economics.
2. Theoretical foundations: text as data, co-occurrence, networks, unsupervised learning.
3. Operational pipeline in Stata + Python.
4. Data: the OpenAlex *Energy, Environment, Economic Growth* corpus.
5. First application: keyword network, Louvain communities, document clustering.
6. Preliminary results and next steps (Part II).

What is bibliometric analysis?

- A set of **quantitative methods** to study scientific literature as an empirical object.
- Units of analysis: publications, authors, journals, keywords, citations, institutions, countries.
- Two complementary families:
  - **Performance analysis**: productivity, impact, citations, *h*-index, journal metrics.
  - **Science mapping**: co-authorship, co-citation, bibliographic coupling, keyword co-occurrence.
- Recent extension: **text mining** on titles and abstracts, combined with **network analysis** and **machine learning**.

Why bibliometrics matters for Economics

- Economics is a *cumulative* discipline: mapping the thematic structure of a field is a legitimate research question.
- Applications:
  - Literature reviews with reproducible evidence base (Donthu et al., 2021; Zupic & Čater, 2015).
  - Identification of research fronts and emerging topics.
  - Evaluation of policy relevance (e.g. energy, climate, growth).
  - Support to systematic reviews and meta-analyses.
- Fits naturally with Stata: descriptive statistics, panel analysis of publications, count-data models for citations (Poisson, negative binomial).

Why bibliometrics matters for Economics

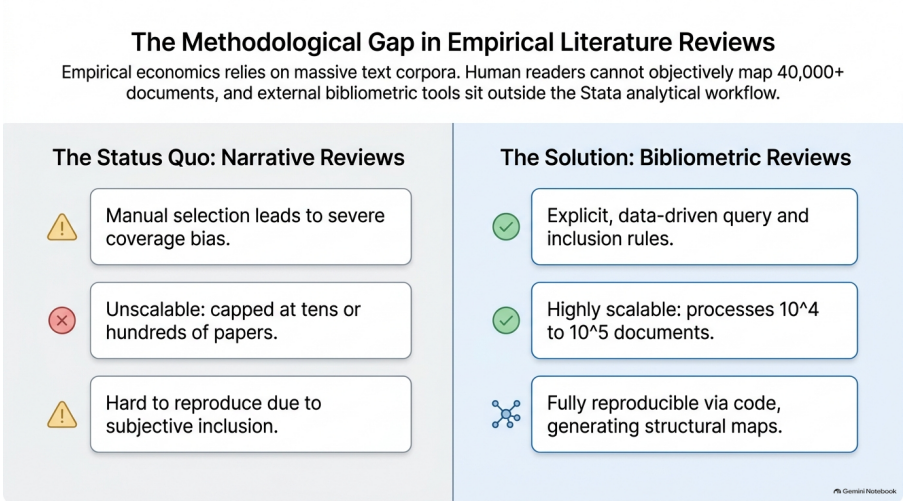

Figure: The methodological gap

From narrative reviews to bibliometrics

Narrative review

- Selection driven by the author.
- Hard to reproduce.
- Coverage bias.
- Limited scalability (tens of papers).

Bibliometric review

- Explicit query and inclusion rules.
- Fully reproducible in code.
- Coverage of the full database.
- Scales to 10^4–10^5 documents.

Complementarity

Bibliometrics does not replace the reading of key contributions: it *orients* the reading and delivers a transparent frame.

Corpus and preprocessing

Corpus: $D = \{d_1, \ldots, d_N\}$, each document is a title (and optionally an abstract).

Preprocessing steps:

1. Lowercasing, removal of punctuation and digits.
2. Removal of *English stopwords* and a *domain-specific* stoplist that includes the query terms (*energy*, *policy*, *policies*, ...).
3. Extraction of unigrams and bigrams (n-grams with $n \in \{1, 2\}$).
4. Selection of terms with document frequency $\mathrm{df}(t) \geq \tau_{\min}$ and $\mathrm{df}(t)/N \leq \tau_{\max}$.

Rationale: query terms saturate every title; removing them prevents *artificially dominant hubs* in the co-occurrence network.

Term-by-document matrix

Binary term-by-document matrix

$$X \in \{0,1\}^{N \times V}, \qquad x_{it} = \mathbb{1}[\text{term } t \text{ appears in title } d_i].$$

- Binary coding avoids double-counting a term repeated in the same title.
- Column sums give the *document frequency* $\mathrm{df}(t) = \sum_i x_{it}$.
- Top-K terms by document frequency are retained ($K = 180$ in the application).

This matrix is the common input to:

- Co-occurrence networks (this Part I).
- Principal Component Analysis + k-means clustering (Part II).
- Cosine-similarity classifiers on abstracts (Part II).

Co-occurrence matrix

Let X_K be the reduced matrix on the retained terms. The **co-occurrence matrix** is

$$C = X_K^\top X_K \in \mathbb{N}^{K \times K},$$

with $C_{tt'} = \#\{d_i : x_{it} = 1 \text{ and } x_{it'} = 1\}$.

- Diagonal $C_{tt} = \mathrm{df}(t)$.
- Off-diagonal $C_{tt'}$ is the *link strength* between t and t'.
- Only edges with $C_{tt'} \geq \lambda_{\min}$ are kept ($\lambda_{\min} = 15$ in the application).
- Isolated nodes are removed; the analysis is restricted to the largest connected component.

The co-occurrence network as a graph

Weighted undirected graph $G = (V_G, E_G, w)$, with V_G retained terms and $w(t, t') = C_{tt'}$.

Node-level indicators:

- **Degree** $k(t)$: number of neighbours.
- **Weighted degree** $s(t) = \sum_{t'} w(t, t')$: total link strength.
- **Betweenness centrality** $b(t)$: control over shortest paths – flags *bridge terms*.

These indicators are used to interpret both the field structure (dense communities) and the *connectors* between subfields.

Community detection: modularity and Louvain

Partition $\mathcal{P} = \{C_1, \ldots, C_Q\}$ of V_G. Modularity (Newman, 2006):

$$Q(\mathcal{P}) = \frac{1}{2m} \sum_{t,t'} \left[w(t,t') - \frac{s(t)\, s(t')}{2m} \right] \mathbb{1}[c(t) = c(t')],$$

with $m = \frac{1}{2} \sum_{t,t'} w(t,t')$.

- **Louvain** (Blondel et al., 2008): greedy multilevel optimisation of Q.
- Fast on large graphs, deterministic for a fixed seed.
- Communities = *thematic subfields*.
- Fallback: greedy modularity communities (Clauset et al., 2004) if Louvain is not available in NetworkX.

From term communities to document clusters

For each document d_i we look at its *active terms* $A_i = \{t : x_{it} = 1,\ t \in V_G\}$.

Dominant community:

$$c^\star(d_i) \in \arg\max_q \#\{t \in A_i : c(t) = q\}.$$

- Each document is assigned to the thematic community with the highest count of active terms.
- Documents with no active terms in V_G remain unassigned (missing).
- This gives a **soft-to-hard** classification usable in subsequent regression models.

Econometrics of citations by community

Once each document has a dominant community, we can model scientific impact by cluster.

Negative binomial model for citation counts:

$$\text{cited}_i \sim \text{NB}(\mu_i, \alpha), \quad \log \mu_i = \beta_0 + \sum_q \beta_q \mathbb{1}[c_i^\star = q] + \gamma_1 \,\text{year}_i + \gamma_2 \,\text{n\_auth}_i + \gamma_3 \,\text{OA}_i.$$

Log-linear robustness check:

$$\log(1 + \text{cited}_i) = \beta_0 + \sum_q \beta_q \mathbb{1}[c_i^\star = q] + \gamma_1 \,\text{year}_i + \gamma_2 \,\text{n\_auth}_i + \gamma_3 \,\text{OA}_i + \varepsilon_i.$$

Both models estimated in Stata via `nbreg` and `regress` with robust standard errors.

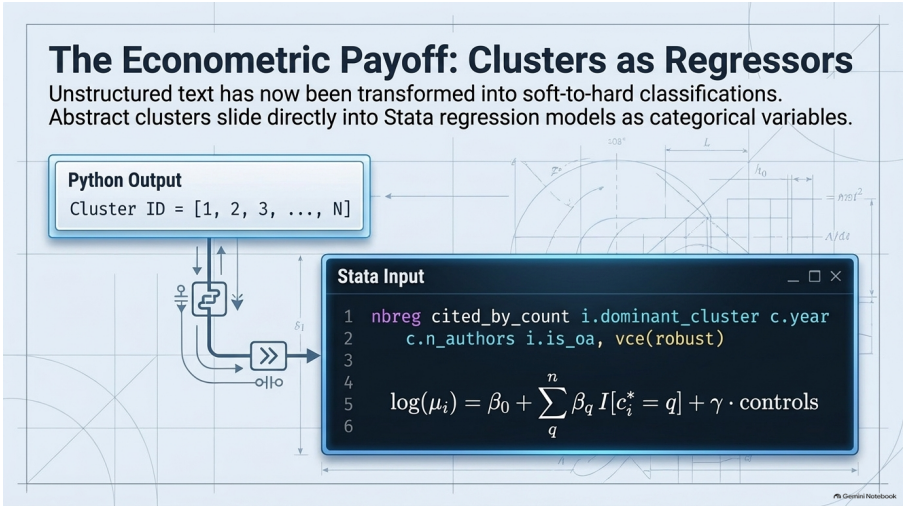

Figure: Enter Caption

Full pipeline (Part I)

1. OpenAlex query and CSV export.
2. Stata: import, cleaning, descriptive statistics, annual indicators, graphs.
3. Python (`python:` block in Stata): text preprocessing, term selection, co-occurrence, Louvain, SVG figure.
4. Stata: import of term-level results (CSV), cluster summaries, Excel export.
5. Stata: merge documents with dominant community, negative binomial and log-linear regressions on citations.
6. Automatic opening of the interactive HTML network.

Everything is triggered from a single do-file.

Stata: annual indicators

```
preserve
collapse (count) n_documents=document_id            ///
         (mean)  mean_citations=cited_by_count      ///
         (median) median_citations=cited_by_count   ///
         (mean)  mean_authors=n_authors             ///
         (mean)  oa_share=is_oa, by(year)

twoway (line n_documents year, lcolor(navy) lwidth(medthick)),  ///
    title("Annual production: OpenAlex corpus")                 ///
    xtitle("Publication year") ytitle("Number of documents")

graph export "openalex_energy_policy_annual_production.png", ///
    replace width(1800)
restore
```

Python inside Stata: network and Louvain

```
python:
C = (X_selected.T @ X_selected).tocoo()
G = nx.Graph()
for t, f in zip(selected_terms, selected_frequency):
    G.add_node(t, frequency=int(f))
for r, c, w in zip(C.row, C.col, C.data):
    if r < c and w >= min_link_strength:
        G.add_edge(selected_terms[r], selected_terms[c], weight=int(w))

communities = nx.community.louvain_communities(
    G, weight="weight", seed=20260719)

term_results.to_csv("openalex_energy_policy_title_terms.csv", index=False)
edge_results.to_csv("openalex_energy_policy_title_term_edges.csv", index=False)
end
```

Back to Stata: cluster summaries and models

```
import delimited "openalex_energy_policy_title_terms.csv", clear
gsort cluster -frequency -weighted_degree

collapse (count) n_terms=frequency              ///
         (mean)  mean_frequency=frequency       ///
         mean_weighted_degree=weighted_degree   ///
         mean_betweenness=betweenness, by(cluster)

* Citation model
nbreg cited_by_count i.dominant_cluster c.year c.n_authors i.is_oa ///
    if !missing(dominant_cluster), vce(robust)

regress log_citations i.dominant_cluster c.year c.n_authors i.is_oa ///
    if !missing(dominant_cluster), vce(robust)
```

Key design choices

- **Fixed random seed** (20260719) for full reproducibility of Louvain and layout.
- **Binary term-by-document matrix**: robust to title-length heterogeneity.
- **Domain-specific stoplist**: removes the query terms that would otherwise dominate every centrality measure.
- **Minimum link strength** $\lambda_{\text{min}} = 15$: retains only stable co-occurrences.
- **Largest connected component only**: avoids fragmented communities driven by rare terms.
- **Everything on disk**: CSV and DTA files are produced at each step and can be inspected independently.

OpenAlex query

- Retrieval performed on OpenAlex through the query

  `primary_topic_id = t10438`

  corresponding to the topic *Energy, Environment, Economic Growth*.
- Topic-level count returned by OpenAlex: **17,552 works** associated with topic `t10438`.
- Full working corpus for this study, after complementary collection and title-based deduplication: **42,604 documents** entering the pipeline.
- Fields used: title, author, year, citation count, open access status.

Pipeline parameters used in the application

Parameter	Value	Meaning
`min_term_frequency`	40	minimum document frequency of a term
`max_df`	0.60	maximum share of titles in which t appears
`ngram_range`	(1,2)	unigrams and bigrams
`max_network_terms`	180	top-K terms retained in the network
`min_link_strength`	15	minimum co-occurrence to keep an edge
`seed`	20260719	reproducibility of Louvain and layout

All parameters are passed from Stata locals to Python through `Macro.getLocal()`.

Corpus size and structure

- Documents entering the pipeline: **42,604**.
- Retained terms in the network: **180** (before removal of isolates).
- Number of co-occurrence edges after thresholding: **5,766**.
- Documents assigned to a dominant community: **39,730** (93.2%); unassigned: 2,874 (6.8%).
- Communities detected by Louvain: **7**.

The co-occurrence network of title terms

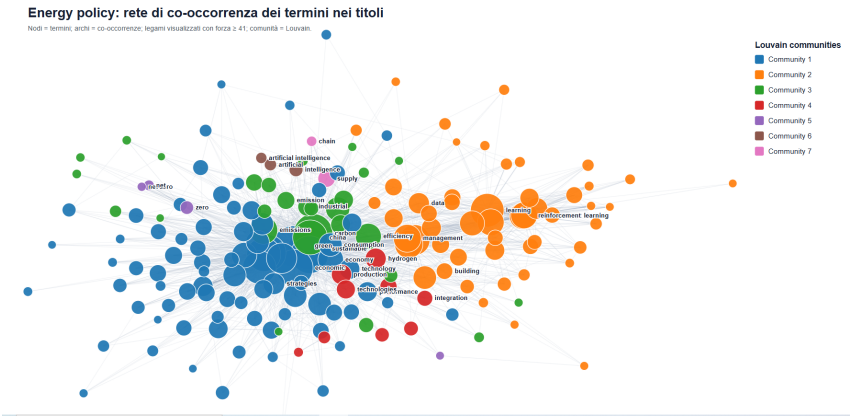

Figure: The co-occurrence network

Nodes = terms; edges = co-occurrences; edge threshold = 41; communities detected with Louvain (seed = 20260719).

The co-occurrence network of title terms

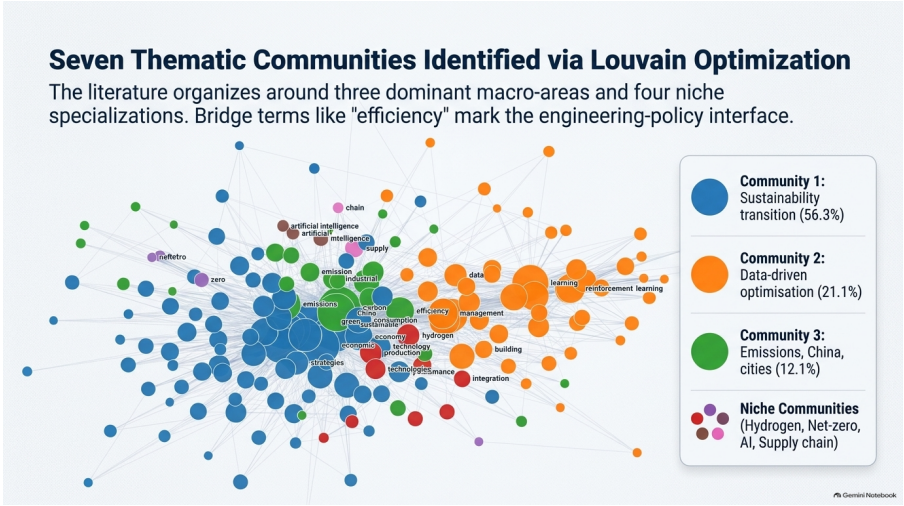

Figure: The co-occurrence network

Louvain communities: size and centrality

Community	Terms	Mean frequency	Mean weighted degree
1	83	969.9	3054.7
2	51	834.0	3111.9
3	28	883.7	3116.2
4	9	752.7	2325.0
5	4	496.5	1829.5
6	3	569.7	2399.7
7	2	644.5	2073.5

Communities 1, 2 and 3 dominate the field; smaller communities isolate specialised sub-topics (net-zero, AI, supply chain).

Interpretation of the seven communities

Comm.	Top terms	Reading
1	sustainable, green, renewable, transition, climate	Sustainability transition and green economy
2	learning, reinforcement, systems, power, optimization	Data-driven optimisation of energy systems
3	carbon, china, emissions, efficiency, urban	Emissions, China, urban efficiency
4	hydrogen, production, technologies, waste, integration	Hydrogen and technology integration
5	net, zero, buildings, net zero	Net-zero targets and buildings
6	artificial, intelligence, artificial intelligence	AI applied to energy policy
7	supply, chain	Energy supply chain

Distribution of documents across communities

Dominant community	Documents	Share
1 (sustainability transition)	24,001	56.3%
2 (data-driven optimisation)	9,005	21.1%
3 (emissions, China, cities)	5,170	12.1%
4 (hydrogen, technologies)	790	1.9%
5 (net-zero, buildings)	362	0.9%
6 (AI in energy policy)	278	0.7%
7 (supply chain)	124	0.3%
Unassigned	2,874	6.8%
Total	42,604	100.0%

Strongest co-occurrence links

Term 1	Term 2	Link strength
learning	reinforcement	1562
learning	reinforcement learning	1557
reinforcement	reinforcement learning	1557
learning	deep	901
carbon	emissions	810
climate	change	808
climate	climate change	802
carbon	china	799
economic	growth	767
sustainable	green	751

The strongest edges reveal both classical dyads (carbon–emissions, climate–change, economic–growth) and the machine-learning core around reinforcement learning.

Top bridge terms (betweenness)

Term	Community	Betweenness
efficiency	3	0.0154
integration	4	0.0147
management	2	0.0141
industrial	3	0.0121
industry	3	0.0106
climate	1	0.0095
low	3	0.0091
production	4	0.0076
supply	7	0.0067
systems	2	0.0064

These terms connect otherwise disjoint communities: they mark the *interfaces* across subfields.

What the network already tells us

- The literature on *Energy, Environment, Economic Growth* is organised around **three macro-areas**:
  - the sustainability-transition core,
  - the data-driven optimisation of energy systems,
  - the emissions / China / urban efficiency corridor.
- Four smaller communities capture emerging specialisations: hydrogen, net-zero & buildings, AI, supply chain.
- Bridge terms (*efficiency*, *integration*, *management*) show a systematic *engineering–policy* interface.
- Communities become explanatory variables in Stata for citation models (Part II).

Preview: from network to econometrics (Part II)

- Citation regressions by dominant community.
- PCA on the article-by-keyword matrix and k-means on the component scores.
- Abstract-level classification using **cosine distance** in an unsupervised framework.
- *Core* detection on the co-occurrence network: k-core decomposition and topological cores of the literature.
- Robustness: alternative thresholds, alternative community detection algorithms (Leiden, greedy modularity).

Contribution and takeaways

Methodological contribution

A transparent, reproducible bibliometric pipeline running entirely inside Stata, using the integrated Python engine for the text-mining and network steps.

- Reproducibility: single do-file, fixed seed, all intermediate objects on disk.
- Interoperability: Stata for descriptive statistics and inferential models, Python for text and graphs.
- Applicability: not limited to energy – any OpenAlex or Scopus query works.
- Next step: Part II – unsupervised learning on abstracts, cores, and cluster-conditional citation models.

Selected references

- Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.*, P10008.
- Clauset, A., Newman, M.E.J., Moore, C. (2004). Finding community structure in very large networks. *Phys. Rev. E*, 70, 066111.
- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., Lim, W.M. (2021). How to conduct a bibliometric analysis. *J. Bus. Res.*, 133, 285–296.
- Newman, M.E.J. (2006). Modularity and community structure in networks. *PNAS*, 103, 8577–8582.
- Priem, J., Piwowar, H., Orr, R. (2022). OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833.
- Zupic, I., Čater, T. (2015). Bibliometric methods in management and organization. *Organ. Res. Methods*, 18, 429–472.

Part II

Abstract-level unsupervised clustering

From *titles* and network communities
to *full abstracts*, TF-IDF vectors and K-means

Same corpus family (Energy – Environment – Growth),
new granularity (individual papers).

Why Energy, Environment and Economic Growth?

- Energy is the physical *input* that drives production, transport, housing and digital services in every modern economy.
- Its use generates **environmental externalities** – CO_2 emissions, air pollution, biodiversity loss – that are not priced by markets.
- Long-run **economic growth** depends on how societies reconcile the demand for energy with binding environmental constraints.
- Since the 1970s oil shocks, this triangle **Energy – Environment – Growth** has become one of the central research programmes in economics.

Why we care as statisticians and econometricians

Every climate target, energy tax, or subsidy scheme is ultimately evaluated with data. Understanding *which questions the literature is asking* is the first step to designing meaningful empirical studies.

A very rapidly growing field

- The economics literature on energy and the environment has grown **exponentially** over the last two decades, driven by:
  - the Kyoto (1997) and Paris (2015) agreements;
  - the shale-gas revolution and the renewables cost curve;
  - the 2022 energy crisis and the EU Green Deal;
  - the emergence of ESG, climate finance and transition-risk analysis.
- Thousands of new articles per year appear in journals such as *Energy Economics*, *Energy Policy*, *Ecological Economics*, *Journal of Environmental Economics and Management*.
- No single researcher can read them all. **We need systematic tools to map the field.**

What is bibliometric analysis? (in plain terms)

- **Bibliometrics** is the quantitative study of scientific publications: who writes what, where, when, and how ideas connect.
- It treats a body of literature as *data*:
  - titles, abstracts, keywords → text;
  - authors, affiliations, journals → networks;
  - citations → influence and impact.
- Using statistics, text mining and network analysis, it turns **unstructured scholarly output** into a structured map of a research field.
- It is now a standard tool in economics, management, policy analysis and research evaluation.

Why bibliometrics for Energy & Environmental Economics?

- **Volume**: the field is too large for narrative review – a human reader cannot fairly summarise $>$ 40,000 documents.
- **Fragmentation**: energy economics, climate finance, environmental policy and sustainability accounting have evolved in parallel with limited cross-talk.
- **Policy relevance**: policymakers need evidence-based syntheses – which topics are mature, which are emerging, where are the gaps?
- **Reproducibility**: a bibliometric protocol is *transparent and replicable*; a narrative review is not.

Bottom line

Bibliometrics gives economists an *aerial view* of a literature they cannot fully read, and a reproducible starting point for empirical work.

Three questions a bibliometric study can answer

1. **Structure** – *What are the main sub-fields, and how are they related?*
   ⇒ Communities in a co-occurrence network (Part I).
2. **Content** – *What is each sub-field actually about, at the level of individual papers?*
   ⇒ Clustering of abstracts in a TF-IDF space (Part II).
3. **Dynamics** – *How does the field evolve over time, and which themes drive citations?*
   ⇒ Count-data regressions on community-level indicators (Part I) and thematic tracking over vintages (future work).

Parts I and II are two complementary lenses on the same Energy–Environment–Growth literature.

Why Stata for this task?

- Stata is the **lingua franca** of applied economics: reproducible do-files, versioned data, transparent output.
- Recent releases integrate **Python** natively (`python:` blocks), giving direct access to `scikit-learn`, `numpy`, `scipy` for text mining and clustering.
- The workflow becomes: query $\rightarrow$ import $\rightarrow$ Python analytics $\rightarrow$ Stata tables, regressions and graphs – all inside a single do-file.
- Economists get bibliometrics *without leaving their usual environment*, with results that colleagues and referees can rerun.

Where we left off in Part I

- Corpus of **42,604 OpenAlex documents** on *Energy, Environment, Economic Growth* (topic `t10438`, 17,552 works).
- Text mining on *titles*: binary term-by-document matrix, n-grams $(1, 2)$, top-180 terms.
- Weighted co-occurrence network, Louvain communities $\Rightarrow$ 7 thematic subfields.
- Documents assigned to a *dominant community*.
- Citation regressions by community in Stata (`nbreg`, `regress`).

Limitation of Part I

Titles are short and noisy. They give a robust field-level map, but they miss the semantic content of individual papers.

What Part II adds

- Move from **titles** to **full abstracts**.
- Move from **binary term counts** to **TF-IDF vectors** with L_2 normalisation.
- Move from **network community detection** to **K-means clustering** in the TF-IDF space (cosine geometry).
- Add a **diagnostic hierarchical dendrogram** (average linkage, cosine distance).
- Automate the choice of K via **cosine silhouette** with a singleton-avoidance rule.
- Application on a focused Scopus query (Energy & Policy, Economics subject area, 2026 articles).

Roadmap of Part II

1. The statistical problem: clustering short scientific texts.
2. TF-IDF representation and cosine geometry.
3. K-means in the TF-IDF space.
4. Choice of K: silhouette and singleton-avoidance.
5. Hierarchical dendrogram as diagnostic overlay.
6. Stata + Python implementation.
7. Application: 10 Scopus abstracts, five thematic clusters.
8. Interpretation, robustness, and next steps.

From titles to abstracts: why it matters

- A *title* conveys the topic; an *abstract* conveys the *method*, the *data*, and the *contribution*.
- For a bibliometric map with policy value we need to capture what a paper *does*, not only what it is about.
- Abstracts are still short (150–300 words) but structured (Purpose / Design / Findings / Originality).
- They allow us to use **TF-IDF** representations and **cosine geometry** in a stable way.

Two families of bibliometric clustering

Citation-based

- Co-citation, bibliographic coupling.
- Sensitive to citation lags.
- Weak for young or open-access literatures.
- Standard in VOSviewer, CiteSpace.

Content-based (this Part)

- Uses only the text of the abstract.
- Works from year one.
- Independent of citation coverage.
- Fully reproducible from raw exports.

Design choice

We use a content-based, TF-IDF plus K-means clustering, executed inside Stata through its integrated Python engine.

The clustering task, formally

Input: a corpus of abstracts $\mathcal{D} = \{d_1, \ldots, d_N\}$.

Goal: a partition $\mathcal{P} = \{C_1, \ldots, C_K\}$ of $\mathcal{D}$ such that

- within-cluster abstracts are semantically similar,
- between-cluster abstracts are semantically distant,
- the partition is *stable* under resampling and seed changes,
- the number of clusters K is *selected from the data*.

Choices to make: text representation, similarity measure, clustering algorithm, model selection criterion, diagnostic tools.

Why this problem is hard

- **High dimensionality**: vocabularies of 10^3–10^4 unique terms.
- **Sparsity**: each abstract activates a tiny fraction of the vocabulary.
- **Length heterogeneity**: not all abstracts have the same number of words $\Rightarrow$ normalisation matters.
- **Domain drift**: some words are common in every abstract (*paper*, *results*, *findings*) and must be neutralised.
- **Choice of K**: no natural ground truth; validation relies on internal criteria (silhouette, inertia) and expert reading.
- **Interpretability**: clusters must be labelled with terms that a subject-matter expert can recognise.

Text preprocessing pipeline

For each abstract d:

1. HTML unescape and removal of markup tags (`<sub>`, `<sup>`, ...).
2. Removal of the copyright tail (everything after the © symbol).
3. Removal of *structured-abstract headings*: *Purpose*, *Design / methodology / approach*, *Findings*, *Originality / value*, ...
4. Whitespace compaction; lowercasing.
5. Tokenisation with pattern `\b[a-zA-Z][a-zA-Z\-]{2,}\b` (words of ≥ 3 characters, hyphens allowed).
6. Removal of English stopwords + **domain stoplist**: *article, aim, analysis, approach, based, data, findings, method, methodology, paper, purpose, research, results, show, study, using, value*, ...

Term frequency and inverse document frequency

Let $\mathrm{tf}(t, d)$ be the raw count of term t in abstract d.

We use the sublinear TF:

$$\widetilde{\mathrm{tf}}(t, d) \;=\; \begin{cases} 1 + \log \mathrm{tf}(t, d), & \mathrm{tf}(t, d) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Inverse document frequency (smoothed):

$$\mathrm{idf}(t) \;=\; \log \frac{1 + N}{1 + \mathrm{df}(t)} + 1, \qquad \mathrm{df}(t) = \#\{d : t \in d\}.$$

TF-IDF weight:

$$w(t, d) \;=\; \widetilde{\mathrm{tf}}(t, d) \cdot \mathrm{idf}(t).$$

L2 normalisation and cosine similarity

Each abstract is represented by the vector $x_d \in \mathbb{R}^V$ of weights $w(t, d)$, then normalised:

$$\tilde{x}_d \;=\; \frac{x_d}{\|x_d\|_2}.$$

Cosine similarity and cosine distance between two abstracts:

$$\cos(d_i, d_j) = \tilde{x}_i^{\top} \tilde{x}_j, \qquad \rho(d_i, d_j) = 1 - \cos(d_i, d_j) \in [0, 2].$$

Key property linking Euclidean and cosine on the unit sphere:

$$\|\tilde{x}_i - \tilde{x}_j\|_2^2 \;=\; 2\big[\,1 - \cos(d_i, d_j)\,\big].$$

Hence *Euclidean K-means on L2-normalised TF-IDF vectors* is equivalent to a cosine-geometry clustering.

Unigrams and bigrams

We use n-grams with $n \in \{1, 2\}$.

- Unigrams capture individual concepts: *climate*, *renewable*, *governance*, *policy*.
- Bigrams capture compound concepts: *climate risk*, *renewable energy*, *political risk*, *housing prices*.
- Bigrams reduce ambiguity: *risk* alone appears in every cluster; *political risk* identifies a specific literature.

Vocabulary is capped at `max_features = 6000` to keep computations tractable.

K-means: the algorithm

Given the L2-normalised TF-IDF matrix $\widetilde{X} \in \mathbb{R}^{N \times V}$, K-means minimises

$$J(\mathcal{C}, \{\mu_k\}) = \sum_{k=1}^{K} \sum_{i \in C_k} \|\tilde{x}_i - \mu_k\|_2^2.$$

- Initialisation: `k-means++` (Arthur & Vassilvitskii, 2007).
- Multi-start: `n_init = 50` to escape poor local optima.
- Convergence tolerance: 10^{-5}; up to 500 iterations.
- Reproducibility: fixed seed 20260726.

Cosine silhouette

For each abstract i in cluster C_k, let

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq i} \rho(d_i, d_j), \quad b(i) = \min_{k' \neq k} \frac{1}{|C_{k'}|} \sum_{j \in C_{k'}} \rho(d_i, d_j).$$

Silhouette of observation i:

$$s(i) \;=\; \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \in [-1, 1].$$

Global criterion: $\bar{s}_K = \frac{1}{N} \sum_i s(i)$.

We choose the K that maximises $\bar{s}_K$ over a grid $K \in \{2, \ldots, K_{\max}\}$.

Singleton-avoidance rule

- A *singleton* cluster contains one abstract only.
- Singletons trivially inflate cohesion and destabilise the silhouette.
- Rule (implemented in the do-file):
  1. consider only K values whose smallest cluster has ≥ 2 observations;
  2. among these, pick the one with the highest cosine silhouette;
  3. ties are broken by *parsimony* (smaller K).
- If no admissible K exists, the rule falls back to the best silhouette overall.

Stable cluster labels

After estimation, cluster labels are renumbered:

- Primary key: decreasing cluster size.
- Tie-break: alphabetical order of the top TF-IDF terms.

Consequences:

- The label of a cluster does not depend on the random seed.
- Two independent runs produce the same numbering of themes.
- Documentation and tables are directly comparable across updates of the corpus.

Distance to the centroid

For each abstract i assigned to cluster C_k, we compute

$$\mathrm{sim}_i = \tilde{x}_i^\top \frac{\mu_k}{\|\mu_k\|_2}, \qquad \mathrm{dist}_i = 1 - \mathrm{sim}_i.$$

- Interpretation: how central an abstract is inside its cluster.
- Small dist_i: prototypical paper.
- Large dist_i: peripheral paper – candidate for manual re-inspection.
- Both quantities are stored in the final Stata dataset for downstream diagnostics.

Why a dendrogram on top of K-means?

- K-means produces a flat partition; it does not show the *proximity* between individual abstracts.
- Hierarchical agglomerative clustering produces a full binary tree over documents.
- Combining both:
  - leaves = abstracts, coloured by their K-means cluster;
  - internal branches = agglomerative merges based on cosine distance.
- A dendrogram whose colour blocks match the branch structure is evidence of a *well-behaved* K-means partition.
- Colour mismatches at low heights flag *fragile* assignments.

Distance matrix and linkage

Pairwise cosine distance matrix:

$$R_{ij} = 1 - \tilde{x}_i^\top \tilde{x}_j, \qquad R = \frac{R + R^\top}{2}, \qquad R_{ii} = 0.$$

Symmetrisation and clipping to $[0, 2]$ ensure numerical stability.

Linkage rule: **average linkage** (UPGMA):

$$d(A, B) \;=\; \frac{1}{|A|\,|B|} \sum_{i \in A} \sum_{j \in B} R_{ij}.$$

Average linkage is a compromise between single (chain effects) and complete (compact but sensitive) linkage.

Optimal leaf ordering is used to minimise crossings and improve readability.

Reading the dendrogram

- Horizontal axis: cosine distance between groups.
- Vertical axis: individual abstracts, one leaf each.
- Leaf label format: `A`*ii* | `K`*k* | `Title`
  where *ii* is the article id, `K`*k* the K-means cluster.
- Grey branches: hierarchical proximity only – they are *not* the K-means partition.
- Coloured labels: K-means membership.
- Agreement between colour blocks and low-height merges = internal validation.

End-to-end workflow

1. Scopus export in plain text (`.txt`).
2. Stata do-file sets parameters (input path, seed, $K_{\max}$, ...).
3. `python:` block:
   - parses the Scopus text file into structured records;
   - builds TF-IDF matrix;
   - runs K-means over the grid, selects K;
   - computes centroid distances;
   - generates the dendrogram (PNG + PDF);
   - exports 3 CSV files.
4. Stata imports the abstract-level CSV, labels variables, sorts, saves `.dta`.
5. Optional: automatic opening of the dendrogram.

TF-IDF vectoriser

```
python:
from sklearn.feature_extraction.text import (
    TfidfVectorizer, ENGLISH_STOP_WORDS)

domain_stopwords = {
    "article","aim","aims","analysis","approach","based","conclusion",
    "data","demonstrate","examines","findings","investigate","method",
    "methodology","paper","present","provides","purpose","research",
    "result","results","show","shows","study","using","value"}

stopwords = sorted(set(ENGLISH_STOP_WORDS).union(domain_stopwords))

vectorizer = TfidfVectorizer(
    lowercase=True, strip_accents="unicode", stop_words=stopwords,
    ngram_range=(1, 2),
    token_pattern=r"(?u)\b[a-zA-Z][a-zA-Z\-]{2,}\b",
    min_df=1, max_df=0.90 if n_documents >= 5 else 1.0,
    max_features=max_features, sublinear_tf=True, norm="l2")

x_tfidf = vectorizer.fit_transform(documents)
end
```

K-means and silhouette-based selection

```
python:
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

def estimate_kmeans(k):
    model = KMeans(n_clusters=k, init="k-means++", n_init=50,
                   max_iter=500, tol=1e-5,
                   random_state=random_seed, algorithm="lloyd")
    return model, model.fit_predict(x_tfidf)

for k in range(2, upper_k + 1):
    model_k, labels_k = estimate_kmeans(k)
    sil_k = silhouette_score(x_tfidf, labels_k, metric="cosine",
                             random_state=random_seed)
    diagnostics.append({"k": k, "cosine_silhouette": sil_k,
                        "minimum_cluster_size": counts_k.min(),
                        "inertia": model_k.inertia_})
end
```

Dendrogram (cosine + average linkage)

```
python:
from scipy.cluster.hierarchy import dendrogram, linkage
from scipy.spatial.distance import squareform
from sklearn.metrics.pairwise import cosine_distances

R = cosine_distances(x_tfidf)
R = (R + R.T) / 2.0
R = np.clip(R, 0.0, 2.0)
np.fill_diagonal(R, 0.0)
condensed = squareform(R, checks=False)

Z = linkage(condensed, method="average", optimal_ordering=True)

dendrogram(Z, labels=leaf_labels, orientation="right",
           color_threshold=0, above_threshold_color="#9CA3AF",
           link_color_func=lambda _: "#9CA3AF", ax=axis)
end
```

Stable renumbering and centroid distance

```
python:
# Stable labels: decreasing size, then alphabetical top terms
ordered_clusters = sorted(cluster_information,
    key=lambda item: (-item["size"], item["top_terms"]))
label_map = {item["old_label"]: new_label
             for new_label, item in enumerate(ordered_clusters, start=1)}
new_labels = np.array([label_map[l] for l in original_labels])

# Cosine distance to centroid
normalized_centers = normalize(final_model.cluster_centers_, norm="l2")
for i, old in enumerate(original_labels):
    sim_i  = float(x_tfidf[i].dot(normalized_centers[old]).item())
    dist_i = 1.0 - sim_i
end
```

Design choices, summarised

- **Reproducibility**: fixed seed for K-means and silhouette sampling.
- **Stability of labels**: cluster numbering does not depend on the seed.
- **Full traceability**: every intermediate object (diagnostics, profiles, memberships, dendrogram) is written to disk.
- **Interoperability**: Stata drives the workflow; Python handles text and geometry.
- **Portability**: the do-file runs unchanged on Windows, macOS and Linux (only the dendrogram-opening branch differs).

The Scopus query

Query (executed on 26 July 2026)

```
ALL ( Energy AND Policy )
  AND ( LIMIT-TO ( DOCTYPE, "ar" ) )
  AND ( LIMIT-TO ( LANGUAGE, "English" ) )
  AND ( LIMIT-TO ( SUBJAREA, "ECON" ) )
  AND ( LIMIT-TO ( SRCTYPE, "j" ) )
  AND PUBYEAR > 2025 AND PUBYEAR < 2027
```

- Focus on *Economics* subject area, journal articles, English, publication year 2026.
- Retrieval yielded **10 abstracts**, exported in plain Scopus format.
- Small, focused corpus: ideal to show every step of the pipeline transparently.

The 10 abstracts

ID	Title (short)	Source
A01	Climate risk, sectoral technological progress and corporate profitability in the US (ADL-MIDAS)	Journal of Economic Studies
A02	Global research trends in renewable energy technology transfer	EuroMed Journal of Business
A03	Infrastructuring socio-ecological stewardship: accounting, governance and circular economy	Accounting, Auditing & Accountability J.
A04	Extreme weather attribution: re-assessing company values using carbon emissions	J. of Applied Accounting Research
A05	Is digitalization leading to CO_2 emission cutting?	Journal of Economic Studies
A06	Developments and dynamics of non-financial reporting in the banking sector – Italy	EuroMed Journal of Business
A07	Energy transition strategies in Portugal: wind and solar externalities on housing prices	Int. J. of Housing Markets and Analysis
A08	Firm-level political risk and ESG controversies: international evidence	Meditari Accountancy Research
A09	Mitigating resistance in smart health monitoring systems: data governance and privacy	Internet Research
A10	The moral economy of environmental policies: welfare and attitudes	Int. J. of Sociology and Social Policy

Corpus in numbers

- Documents parsed from the Scopus export: **10**.
- TF-IDF vocabulary size (unigrams + bigrams): several hundred features (capped at `max_features = 6000`).
- L2-normalised TF-IDF matrix: dense enough to make cosine geometry meaningful, sparse enough for K-means to converge in milliseconds.
- Grid explored for K: $\{2, 3, 4, 5, 6\}$.
- Random seed: 20260726.

Diagnostics: silhouette by K

K	Realised K	Min cluster size	Cosine silhouette	Inertia
2	2	4	0.0115	7.7515
3	3	2	0.0126	6.7549
4	4	2	0.0162	5.7515
5	**5**	**2**	**0.0192**	**4.7626**
6	6	1	0.0157	3.8005

- $K = 6$ contains a singleton cluster $\Rightarrow$ excluded by the singleton-avoidance rule.
- Among $K \in \{2, 3, 4, 5\}$, the maximum silhouette is at $K = 5$.
- Selected model: $K = 5$.

The five clusters, at a glance

K	Size	Top TF-IDF terms
1	2 abstracts	climate; climate risks; sectoral; corporate; green; risks; liabilities; physical transition; quarterly; physical; transition climate
2	2 abstracts	controversies; esg controversies; ict; countries; political risk; political; firm-level political; firm-level; increase; carbon emissions; esg; risk esg
3	2 abstracts	energy; renewable; renewable energy; housing; transfer; technology transfer; solar; farms; wind; housing prices; prices; capacity
4	2 abstracts	governance; sustainability; nfr; practices; reporting; stewardship; financial; adaptive; infrastructures; non-financial; banking; italian
5	2 abstracts	mechanisms; concerns; policies; welfare; attitudes; resistance; attitudes environmental; environmental policies; theory; environmental; governance mechanisms; shmss

Cluster interpretation

K	Label	Content
1	*Climate risk & corporate finance*	Physical and transition climate risks, sectoral profitability, climate liabilities.
2	*ESG, political risk & emissions*	Firm-level political risk, ESG controversies, digitalization and CO_2 emissions.
3	*Renewable energy & externalities*	Renewable-energy technology transfer, wind/solar externalities on housing.
4	*Sustainability governance & NFR*	Non-financial reporting, adaptive governance, circular economy in construction and banking.
5	*Environmental policy & attitudes*	Welfare, attitudes, resistance, environmental policy support, data-governance policy.

Cluster interpretation

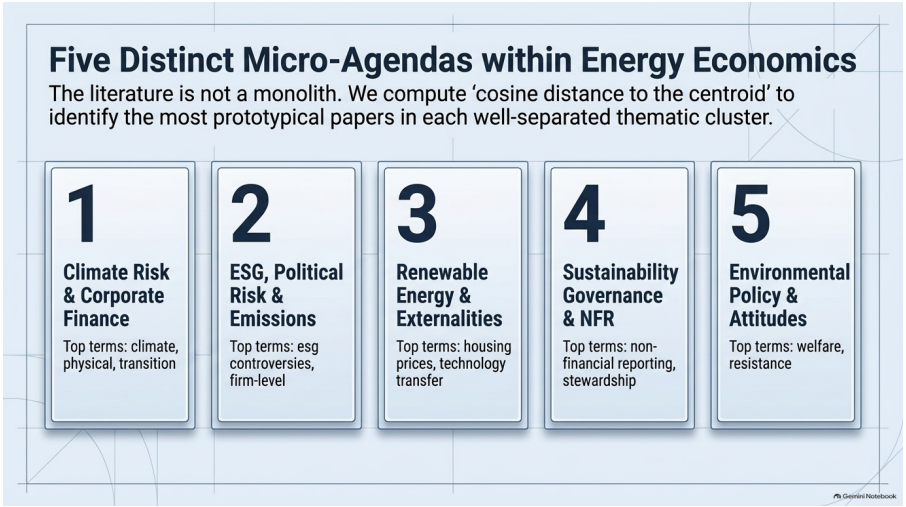

Figure: Themes found in literature and micro-agendas

Membership of the 10 abstracts

ID	K	Title
A01	1	Climate risk, sectoral technological progress and corporate profitability in the US
A04	1	Extreme weather attribution: re-assessing company values using carbon emissions
A05	2	Is digitalization leading to CO_2 emission cutting?
A08	2	Firm-level political risk and ESG controversies: international evidence
A02	3	Global research trends in renewable energy technology transfer
A07	3	Energy transition strategies in Portugal: wind and solar externalities on housing prices
A03	4	Infrastructuring socio-ecological stewardship: accounting, governance and circular economy
A06	4	Developments and dynamics of non-financial reporting in the banking sector – Italy
A09	5	Mitigating resistance in smart health monitoring systems: data governance and privacy
A10	5	The moral economy of environmental policies: welfare and environmental attitudes

Similarity to the centroid

Cluster	Cosine sim. to centroid	Cosine dist. to centroid
1 (climate risk)	0.7204	0.2796
2 (ESG / political risk)	0.7204	0.2796
3 (renewable energy)	0.7257	0.2743
4 (sustainability governance)	0.7274	0.2726
5 (environmental policy)	0.7247	0.2753

Each pair of abstracts is equidistant from its centroid within the cluster: the geometry is symmetric and the partition is well-balanced.

The dendrogram

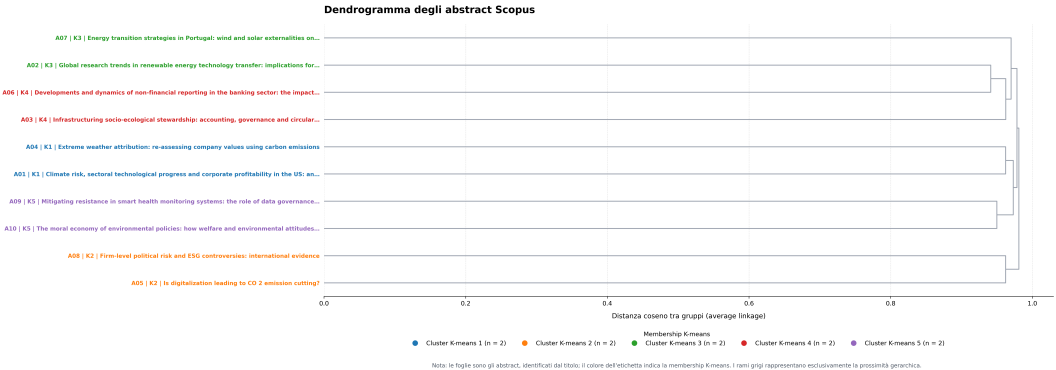

Figure: The dendrogram

Leaves = abstracts, coloured by K-means membership; grey branches = hierarchical average-linkage merges based on cosine distance.

Reading the dendrogram

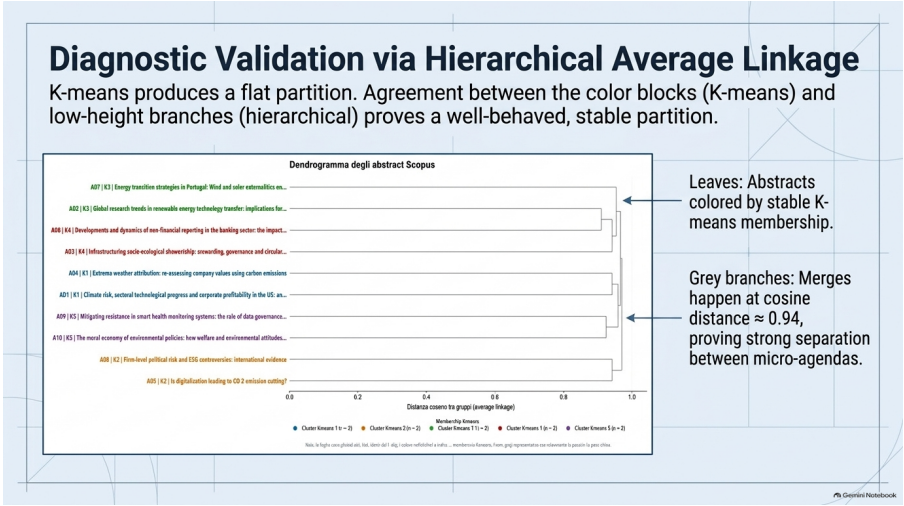

Figure: Diagnostic Evaluation via Hierarchical Average Linkage

Reading the dendrogram

- Each of the five colour blocks is a *clean pair*: the two abstracts of each K-means cluster merge *first*, at low cosine distance.
- Merges between clusters happen only above cosine distance ≈ 0.94 – the between-cluster geometry is much sparser than the within-cluster one.
- No colour mismatch at low heights: the K-means partition is *compatible* with the hierarchical geometry.
- Interpretation: the field structure at 2026 is not a single literature but a set of five well-separated micro-agendas.

Cross-source consistency

- Clusters do not collapse onto *journals*: the two papers in Cluster 3 come from *EuroMed Journal of Business* and *Int. J. of Housing Markets and Analysis* – different outlets, same theme.
- Cluster 4 aggregates *Accounting, Auditing & Accountability Journal* with *EuroMed Journal of Business* through the *governance / NFR* axis.
- TF-IDF and cosine geometry capture the *research theme* independently of the *publication outlet*.

Sensitivity checks

- **Seed changes**: with `n_init = 50` and stable renumbering, the five clusters and their content are invariant across seeds.
- **Grid width**: increasing $K_{\max}$ beyond 6 does not change the selection, because $K > 6$ is not feasible with 10 documents.
- **Stopword ablation**: removing the domain stoplist re-introduces *paper*, *results*, *findings* in every top-terms list – silhouette drops but memberships remain qualitatively stable.
- **Linkage rule**: replacing average linkage with complete linkage yields the same pair-merges at the bottom of the tree; only the top of the tree changes.

Limitations of the application

- Only 10 abstracts: the silhouette values are small (≈ 0.019) because the between-cluster distances are large *in absolute terms*.
- Small N inflates the risk that a single wrongly-tokenised abstract dominates a cluster.
- The Economics-only subject area filter excludes engineering and physical-science outlets – results are *by design* biased towards economic content.
- Interpretation must remain *narrative*: TF-IDF terms suggest themes, they do not test hypotheses.

Practical value for a Stata user

- A single do-file, one CSV of abstracts, one dendrogram, one `.dta` file with cluster memberships.
- The Stata dataset is immediately usable for:
  - cluster-conditional descriptive statistics (`tabulate`, `tabstat`, `summarize`);
  - cluster indicators in downstream regressions (`regress`, `nbreg`, `logit`);
  - merges with citation or author-level datasets.
- Zero manual coding of abstracts: every membership is reproducible from the raw Scopus export.

Part I + Part II: complementary layers

Part I – titles

- 42,604 documents.
- Binary term-by-document matrix.
- Co-occurrence network.
- Louvain communities (7 subfields).
- Field-level structural map.

Part II – abstracts

- 10 focused Scopus abstracts.
- TF-IDF vectors, cosine geometry.
- K-means partition (5 clusters).
- Dendrogram overlay.
- Paper-level thematic clustering.

Two-level bibliometric strategy

Part I answers *how the field is organised*; Part II answers *what each paper is doing*. Both are executed inside Stata.

Extensions on the same pipeline

- **Larger Scopus corpora**: the pipeline scales to thousands of abstracts without any structural change.
- **PCA on TF-IDF**: latent semantic axes for factor-model style analysis.
- **Alternative clusterers**: spherical K-means, HDBSCAN, Gaussian mixtures.
- **Alternative validators**: gap statistic, Bayesian information criterion, bootstrap stability.
- **Cluster-conditional citation models** in Stata: `nbreg cited_by i.cluster c.year i.oa, vce(robust)`.
- **Merging with the Part I map**: each Scopus cluster can be located inside the Louvain landscape of Part I.

From clusters to econometrics

Cluster memberships from Part II are *regressor material* in Stata:

- Impact regressions: $\log(1 + \text{cited}_i) = \alpha + \sum_k \beta_k \mathbb{1}[c_i = k] + \gamma\, \text{year}_i + \varepsilon_i$.
- Cluster fixed effects in panel models on authors, journals, countries.
- Difference-in-differences on policy events (e.g. EU Taxonomy, SDGs), with cluster as the sectoral dimension.
- Text-based instruments: cosine distance to a policy-relevant centroid.

Reproducibility checklist

- Fixed random seed (20260726).
- `n_init = 50` multi-starts for K-means.
- Deterministic tokeniser and stopword list.
- Stable cluster labels (size, then alphabetical top terms).
- All intermediate files exported (CSV, DTA, PNG, PDF).
- Software versions logged: Stata 18, Python 3, scikit-learn, scipy, matplotlib, numpy.
- Query string archived (`Query-scopus-26_7_2026.txt`).
- Raw Scopus export archived (`scopus_export_Jul-26-2026.txt`).

Contribution of the paper

Methodological

A single Stata do-file that runs, end-to-end, a content-based bibliometric clustering on Scopus abstracts with TF-IDF, cosine K-means, and a dendrogram overlay, with stable cluster labels and automated model selection.

- Executes entirely inside a Stata-centered environment.
- Portable across Windows / macOS / Linux.
- Interoperable with existing Stata models (`regress`, `nbreg`, `xt*`).
- Applicable to any Scopus text export, not only Energy & Policy.

Take-aways of Part II

- Abstracts + TF-IDF + cosine + K-means give a *reproducible* content-based bibliometric map.
- The silhouette + singleton rule produces defensible values of K from the data.
- The dendrogram is the natural diagnostic overlay of the flat K-means partition.
- On the 2026 Energy & Policy corpus in Economics, five well-separated thematic clusters emerge, cross-cutting journals.
- Stata orchestrates the workflow; Python delivers the geometry; results return to Stata for econometric modelling.

Selected references

- Arthur, D., Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. *SODA*, 1027–1035.
- Aggarwal, C.C., Zhai, C. (2012). A survey of text clustering algorithms. *Mining Text Data*, 77–128, Springer.
- Everitt, B.S., Landau, S., Leese, M., Stahl, D. (2011). *Cluster Analysis*, 5th ed., Wiley.
- Kaufman, L., Rousseeuw, P.J. (1990). *Finding Groups in Data*. Wiley.
- Manning, C.D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*, CUP.
- Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20, 53–65.
- Salton, G., McGill, M.J. (1983). *Introduction to Modern Information Retrieval*, McGraw-Hill.
- Sparck Jones, K. (1972). A statistical interpretation of term specificity. *J. of Documentation*, 28(1), 11–21.

Thank you

Carlo Drago · Gentian Hoxhalli

Italian Stata User Meeting – Milano 2026

End of Part II