One Method, Many Systems

A rank-size approach to socio-economic modelling
IE 2025

Valerio Ficcadenti
London South Bank University
Business School
London, UK

March 8, 2026

Overview

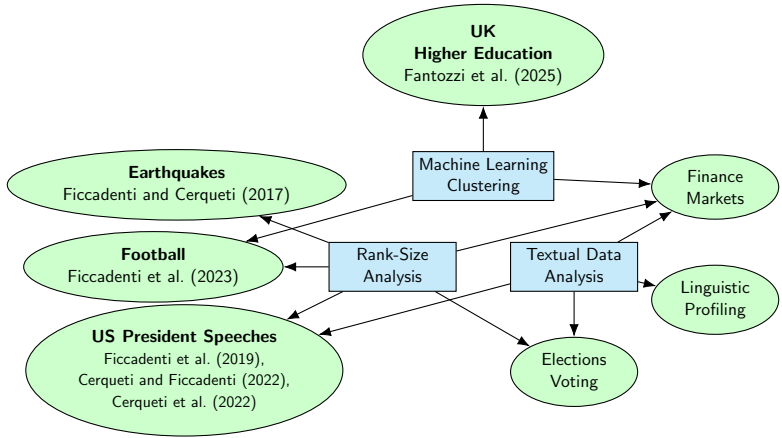
1. Introduction of myself
2. Rank-Size Laws
3. Clustering and Rank-Size
4. Textual Analysis
5. References

About me...

- **Associate Professor in Business Research Methods** @ London South Bank University, UK
- Financial Pricing Specialist @ Deloitte, UK
- Post-Doc in Financial Modeling @ University Polytechnic of Marche, IT
- PhD in Quantitative Methods for Economic Policy @ University of Macerata, IT

I love politics...

Research as a Network, Not a List



Each node = a paper/topic;
Edges = shared methods (rank laws, textual data analysis, Clustering).

Methodological Kernel: Rank-Size Laws

Definition

Rank-size laws describe how elements in a dataset—when ranked in decreasing order—follow a regularity described by a set of mathematical functions.

Historical Origins:

- Pareto (1896): Wealth distribution follows a rank-size regularity!
- Auerbach (1913): Applied rank-size to city populations.
- Zipf (1945, 1949): Observed that word frequencies scale inversely with rank—gave the law its name.
- **Modern Forms**: Zipf–Mandelbrot Mandelbrot (1953, 1965), Lavalette Lavalette (1996), the Discrete Generalised Beta Distribution, and the Universal Law Ausloos and Cerqueti (2016) allow for a more flexible fitting.

Methodological Kernel: Rank-Size Laws

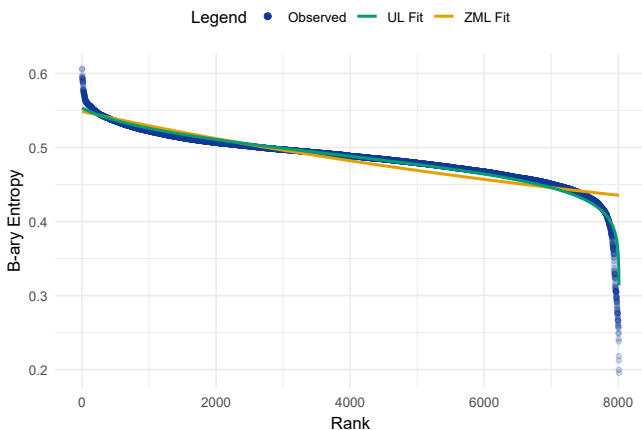
The Zipf–Mandelbrot Law (ZML):

$$z = \frac{\alpha}{(\beta + r)^{\gamma}}, \qquad (1)$$

α, β, and γ to be estimated.
The Universal Law (UL):

$$z = k_5 \cdot \frac{(N(r + \phi))^{-\gamma}}{(N + 1 - r + \psi)^{-\xi}}, \qquad (2)$$

k_5, ϕ, ψ, γ, and ξ to be estimated, and N is the sample size.



Rank-Size Laws: Earthquakes' economic cost

From: Ficcadenti and Cerqueti (2017)

Calibrated Parameter	Value
$\hat{\alpha}$	9.48
$\hat{\beta}$	68.80
$\hat{\gamma}$	0.14
R^2	0.98

Table: Calibrated parameters of best-fit on Eq. 1 on earthquakes from April 16, 2005 to March 31, 2017 ($N = 13239$, $M \geq 2.5$), in Italy.

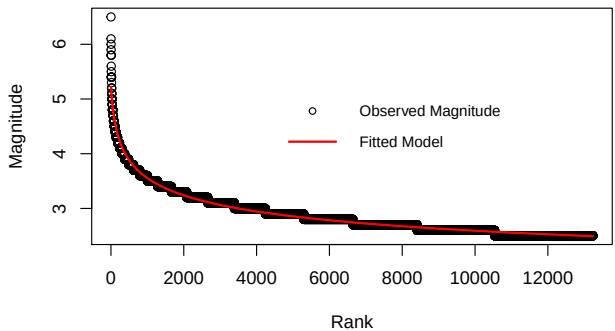


Figure: All the earthquakes registered from 16/04/2005 to 31/03/2017 with magnitudes not smaller than 2.5 with ZML fit, from Eq. 1.

Rank-Size Laws: US President Speeches

From: Ficcadenti et al. (2019)

	R^2	Std err.
Max	1,00	5,68
Min	0,91	0,26
m	0,99	1,00
μ	0,98	1,25
σ	0,01	0,87

Table: Stat. summary of R^2s and non-linear regression std err. for each fit with Eq. (1). They represent the models goodness of fit calibrated over each US President speech's words frequencies.

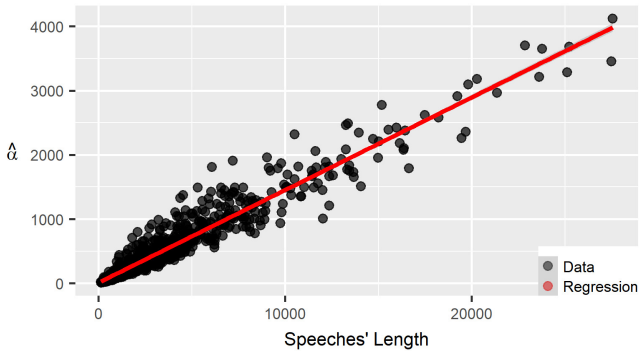


Figure: Relationship between Eq. 1's $\hat{\alpha}$ and speeches' length in term of total number of words used per corpus

Rank-Size Laws: COVID-19

From: Cerqueti and Ficcadenti (2022).

$$z = a + b \cdot r + c \cdot r^2 + d \cdot r^3, \quad (3)$$

where z is new death per million and r represents the rank.

Parameter	Italy	United Kingdom
$\hat{a}$	13.09504	15.66925
$\hat{b}$	-0.04930	-0.12922
$\hat{c}$	-0.00000	0.00038
$\hat{d}$	1.0×10^{-7}	-3.8×10^{-7}
R^2	0.99	0.99
RMSE	0.48	0.41

Table: Calibrated parameters and goodness-of-fit metrics (R^2 and RMSE) for the fitted model in Italy and the United Kingdom.

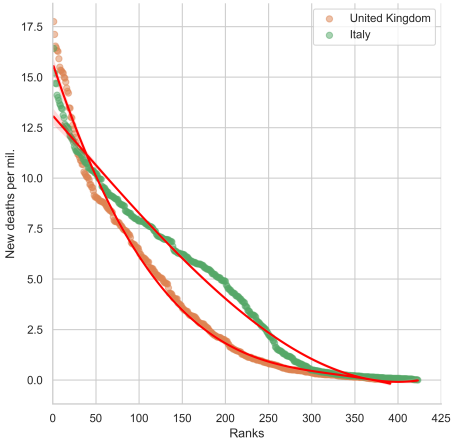
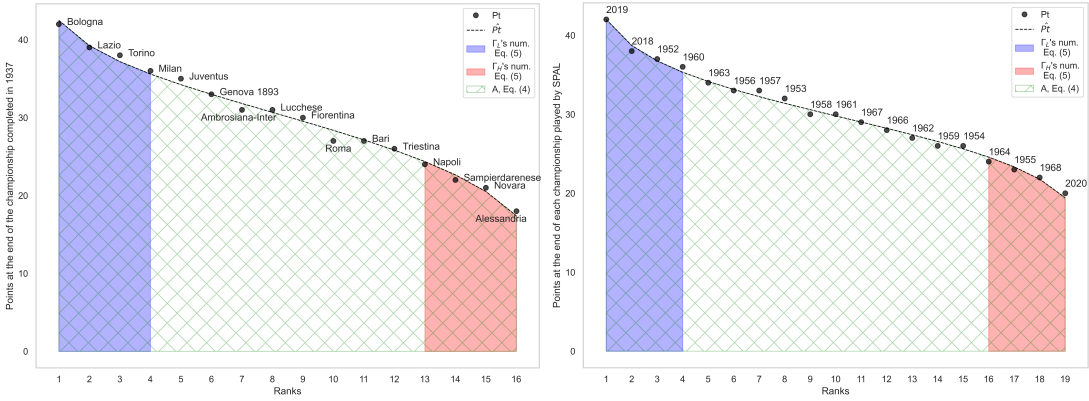


Figure: UK and Italy new deaths per mil. & best fits (red lines) via Eq. 3.

Rank-Size Laws: Soccer

From: Ficcadenti et al. (2023)



(a) Analysis *by year* for the championship finished in 1937.

(b) Analysis *by team* for championships in which SPAL has played.

Figure: DGBD: $z = \frac{\alpha(R+1-r)^{\beta}}{r^{\gamma}}$, where α, β, and γ are non-negative parameters to be calibrated, and $R = \max(r)$ over the considered data sample.

Rank Laws: Meta-Observation

- Ranks capture hierarchy, competition, institutional inertia in a certain setting.
- Same laws fit different phenomena, e.g., earthquakes' magnitude, words' frequencies, scored points, each of these rank under a different framework of rules.
- Deviations from law signal structural change, especially when observed in the same framework for the same phenomenon.
- In the same phenomenon, rank-size law parameters contain information on the phenomenon when the fit is run on repetitions under the same framework of rules.

Rank Laws: Meta-Observation - Soccer

Discrete Generalised Beta Distribution:

$$z = \frac{\alpha(R+1-r)^{\beta}}{r^{\gamma}} \quad (4)$$

where α, β, and γ are parameters to be calibrated, and $R = \max(r)$.

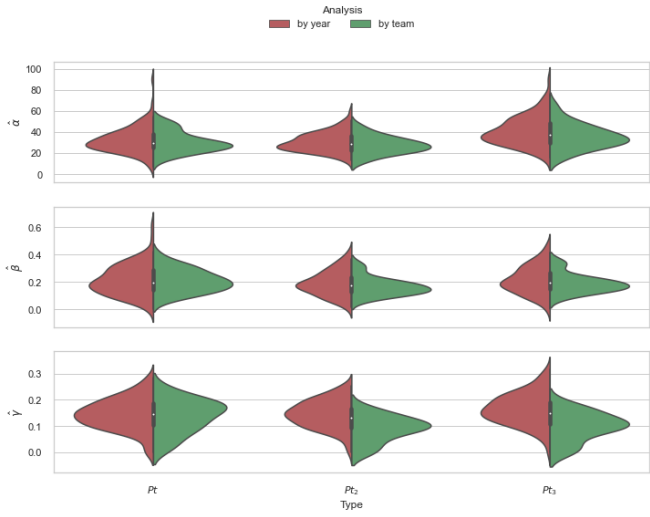


Figure The figure presents the Eq. (4) estimated parameters' probability density (smoothed by a kernel density estimator) when one performs the best fits "by year" and "by team'.' Figure from Ficcadenti et al. (2023).

Rank Laws: Meta-Observation - COVID19

Country	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$	R2	RSME	Inflection Point
Italy	13.09504	-0.04930	-0.00000	1.000000e-07	0.99	0.48	2.78
Japan	0.56195	-0.00512	0.00002	-2.000000e-08	1.00	0.01	288.22
Norway	0.95166	-0.01289	0.00006	-8.000000e-08	0.97	0.04	243.20
South Korea	0.19842	-0.00168	0.00001	-1.000000e-08	0.98	0.01	269.90

Table: Estimated parameters and clusters per each country. The last column represents the rank at which the best fit of Eq. (3) presents a change in concavity.

Rank Laws: Meta-Observation - US President Speeches

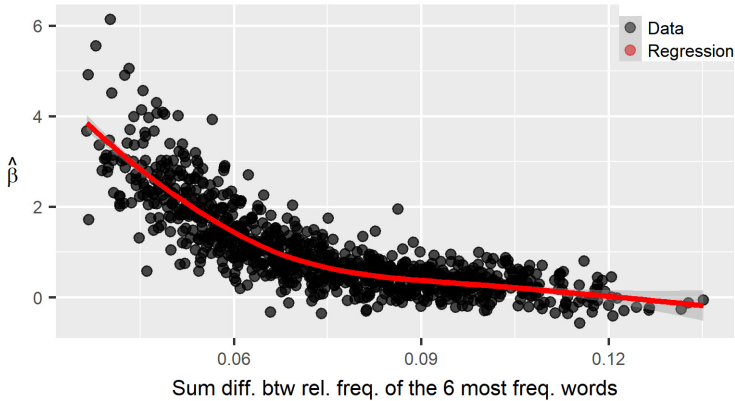


Figure: Eq. (1) $'s$ $\hat{\beta}$s from US President speeches against the summed differences in relative frequencies of the top 6 repeated words within each speech. Figure from Ficcadenti et al. (2019)

Reviewer 3 asks:

"So what?"

Rank Laws: Machine Learning? k-means clustering

Let $\boldsymbol{\theta}_n = (\theta_{n,1}, \ldots, \theta_{n,P})$ be the P-dimensional parameter vector estimated from a rank-size model at instance n (e.g., a specific election, year, or entity).
We construct the parameter matrix:

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,P} \\ \vdots & \ddots & \vdots \\ \theta_{N,1} & \cdots & \theta_{N,P} \end{bmatrix}$$

After standardizing each column to mean zero and unit variance, we define squared Euclidean dissimilarities:

$$D_{n,k}^2 = \sum_{p=1}^{P} \left(\tilde{\theta}_{n,p} - \tilde{\theta}_{k,p} \right)^2$$

Clustering seeks the partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ minimizing:

$$J = \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} D_{n,k}^2$$

The optimal number of clusters K has to be selected for balancing fit and complexity.

Clustering Countries by Rank-Size of COVID-19 Deaths

- **Cluster 0**: Countries with flat peaks, and gentle decay. Includes: Australia, Japan, South Korea, Canada, Finland.
- **Cluster 1**: Countries with **sharp peaks** and **high curvature**. Includes: Slovenia, Hungary, Croatia, Belgium, Poland.
- **Cluster 2**: Countries with **prolonged high mortality** and sustained waves. Includes: Italy, Czechia, US, Montenegro, France.

Countries grouped by structural similarity in their rank-size parameters.

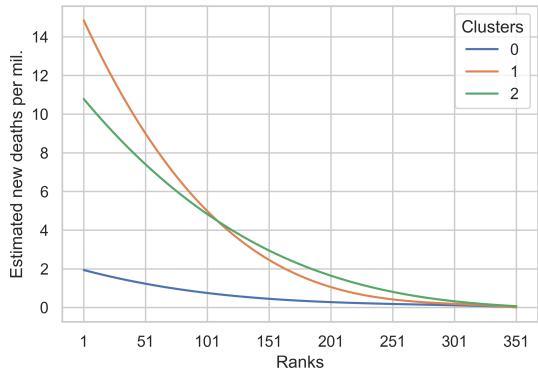


Figure: The 3 curves are obtained by plugging into Eq. (3), $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\hat{d}$ corresponding to the centroid of the clusters $\{0,1,2\}$.

Clustering Soccer Teams by Rank-Size Parameters

- **Cluster 0** — Teams with more uniform performance and lower variability *(e.g., Verona, Cagliari, Chievo, Lecce)*
- **Cluster 1** — Historically successful teams with high scores and broader rank gaps *(e.g., Juventus, Inter, Roma, Napoli)*

Clusters reflect differences in rank concentration and stability across decades of competition.

Type	Cluster	$\bar{\hat{\alpha}}$	$\bar{\hat{\beta}}$	$\bar{\hat{\gamma}}$
Pt	0	23.78	0.257	0.096
	1	38.68	0.159	0.187
Pt_2	0	21.46	0.212	0.065
	1	33.17	0.132	0.113
Pt_3	0	25.92	0.247	0.055
	1	42.29	0.157	0.128

Table: Parameters are derived from DGBD-fitted rank-size curves across three point systems.

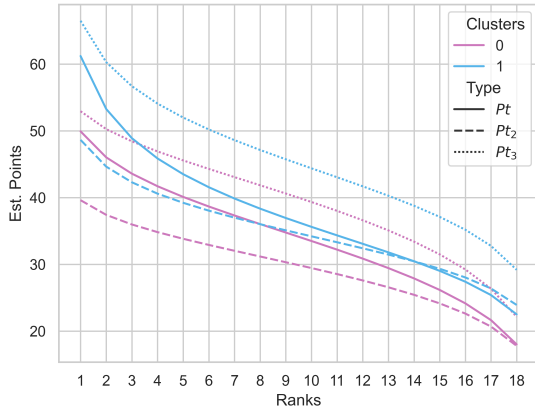


Figure: Cluster profiles based on average fitted curves for teams with $\geq$ 16 Serie A seasons.

Methodological Kernel: Textual Data Analysis

Definition of Text Data Analysis

Textual data analysis is the quantitative study of written/transcribed language aimed at uncovering structural, statistical, or semantic patterns from word distributions, frequency rankings, entropy-based measures, words embeddings and more.

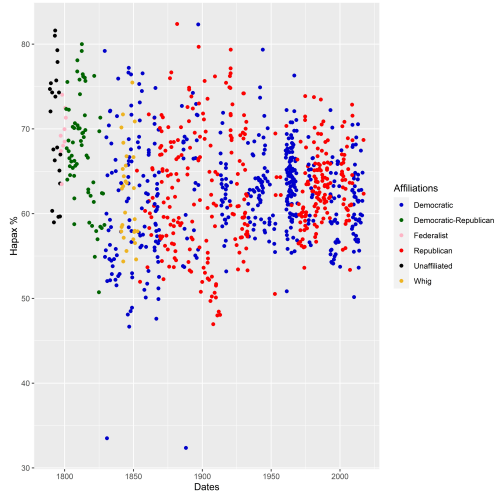
- Treats text as a signal: sequences of tokens carrying statistical regularities.
- Applies rank-size models (e.g. Zipf-Mandelbrot) to analyse word frequency decay in corpus.
- Identifies rare yet meaningful words (e.g., hapax legomena) to detect stylistic or strategic shifts.
- Incorporates entropy and inequality measures to assess complexity and concentration in discourse.

Textual Data Analysis: Hapax in US President Speeches

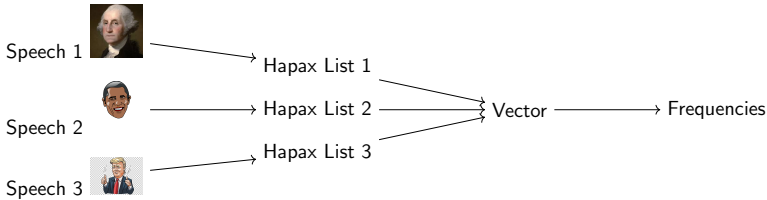
Definition of Hapax Legomenon

A token (word) that appears only once within a corpus.

Figure: Each dot represents a speech, and the colour is the affiliation of the US President speeches, at the time of the speech. The y-axis presents the % of hapax legomena in the speech. x-axis presents time.



Textual Data Analysis: Hapax in US President Speeches



Rank	Word(s)	Frequency
1	sense	250
2	given	247
3	bring, house	240
4	give	239
...	...	...

Table: The most frequent hapax legomena, along with their frequencies.

Textual Data Analysis: Identifying the Core Hapaxes

The **hapax legomena core** is identified using a **Hirsch index**, Hirsch (2005).

$\bar{H}$ Definition

$\bar{H}$ is the maximum number of hapaxes that appear in at least $\bar{H}$ speeches.

Core = 182 most frequent hapaxes (e.g., *sense*, *bring*, *house*). It captures ˜6.6% of total information despite being only 0.58% of the hapax set.

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
Estimation (a)	6.029×10^8	2540	1.896
95% CI (a)	$(5.676 \times 10^8, 6.381 \times 10^8)$	(2525, 2554)	(1.890, 1.902)
Estimation (b)	287.7	5.903	0.084
95% CI (b)	(281.8, 293.6)	(4.288, 7.519)	(0.080, 0.088)
Estimation (c)	4.359×10^8	2668	1.861
95% CI (c)	$(4.083 \times 10^8, 4.634 \times 10^8)$	(2652, 2685)	(1.854, 1.867)

Table: Best-fit parameters of the Zipf-Mandelbrot law (Eq. 1).
(a) All hapaxes; (b) core hapaxes; (c) all hapaxes excluding the core.
95% confidence intervals in parentheses.

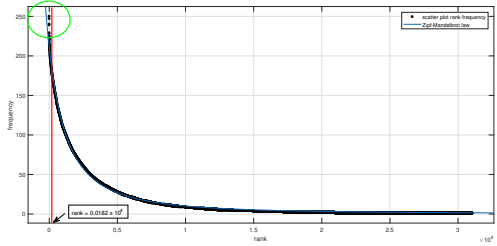


Figure: Best-fit of Eq. 1 with the core highlighted.

Reviewer #1:

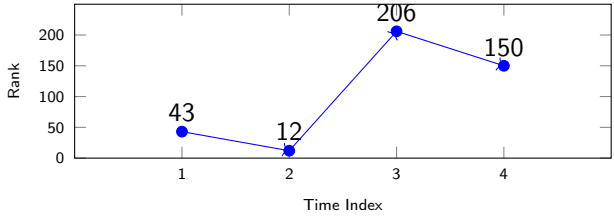
"What is the underpinning mechanism that generates the core of the hapax legomena in the US President speeches corpora?"

Textual Data Analysis: Hapax in US President Speeches

Fellow Citizens of the **Senate** *and the House of Representatives:*
Among *the* **vicissitudes incident** *to* **life**, *no event could have* **filled** *me with* **greater anxieties** *than that of which the* **notification** *was* **transmitted** *by your order, and* **received** *on the* **fourteenth** *day of the present* **month.**

Fellow Citizens of the **43** *and the House of Representatives:*
12 *the* **206 150** *to* **54**, *no event could have* **143** *me with* **41 215** *than that of which the* **207** *was* **154** *by your order, and* **64** *on the* **214** *day of the present* **98.**

From George Washington's First Inaugural Address, April 30, 1789



Textual Data Analysis: Rank-Size approach

Two problems (P and P')

P *If a President has pronounced a hapax – whose rank in the whole corpus is $i \in$ **Rank** – which is the probability that the next hapax has rank $j \in$ **Rank**?*

P' *Identify a squared probability matrix Π of order R such that the homogeneous Markov chain $X = (X(t) : t \in \mathbb{N})$ with Π as transition matrix has stationary distribution F.*

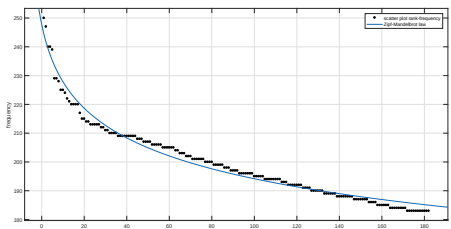


Figure: Best-fit curve of hapax legomena core from Ficcadenti et al. (2020); Cerqueti et al. (2022), using Eq. (1).

Textual Data Analysis: Markov Chain of order 1?

Check the validity of

$$P(X(t+1) = i_{t+1} | X(t) = i_t) = P(X(t+1) = i_{t+1} | X(t) = i_t, X(t-1) = i_{t-1}), \quad (5)$$

for each $t \in \mathbb{N}$ and $i_{t-1}, i_t, i_{t+1} \in$ **Rank**.

1. Simulate series 1000 with observed Transition Matrix (TM) of order 1 and 2.
2. Series simulated from 1^{st}-order TM VS those simulated from the 2^{nd}-order TM.

Estimate the Markov Chain order- P

The pairwise comparison of the simulated series shows no difference, so we have a MC of order one.
Tests run with Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney

Textual Data Analysis: Does it convergence to the rank-size empirical distribution?

Identify a squared probability matrix Π of order R such that the homogeneous Markov chain $X = (X(t) : t \in \mathbb{N})$ with Π as transition matrix leads to a stationary distribution F. Resolved with MCMC procedure $\rightarrow$ Metropolis-Hastings algorithm Brooks et al. (2011).

Verify convernce to F - P'

The series resulting from MCMC converge to the empirical distribution F based on estimated parameters of Eq. 1. Tested with χ^2 and Kolmogorov-Smirnov tests

Additional details in Cerqueti et al. (2022)

Algorithm 1: Rank-based Metropolis-Hastings MCMC

Input: Steps N, initial state x_0, rank-frequency vector $Fr(\cdot)$

for $t \leftarrow 1$ **to** $N - 1$ **do**
  Sample $j \sim$ Uniform(Rank)
  $a \leftarrow \min\left(1, \frac{Fr(j)}{Fr(x_t)}\right)$
  Sample $u \sim$ Uniform$(0, 1)$
  if $u \leq a$ **then**
    $x_{t+1} \leftarrow j$ `// accept new state`
  else
    $x_{t+1} \leftarrow x_t$ `// stay in current state`
  end
end

Textual Data Analysis: US Presidents Economics

Network analysis of US President speeches economic content.

Research Question:

Can economic terminology help uncover patterns of similarity across US Presidents' speeches?

- Focused on economic and financial vocabulary.
- Built a network where nodes are speeches and links reflect similarity in economic content.

Textual Data Analysis: US Presidents Economics - Methods

Step 1: Text Mining

- Created an economic glossary (383 terms from Bishop (2009) + Wikipedia).
- Counted term frequencies in each speech.
- Built a document-term matrix.

Step 2: Network Construction

- Cosine similarity between speeches.
- Filtered network using the Bonferroni-adjusted permutation test.
- Final network: 942 speeches, statistically significant links.

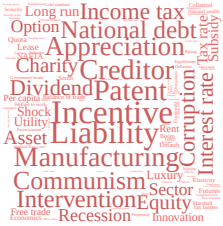
Textual Data Analysis: US Presidents Economics - Findings

- **High clustering coefficient:** presence of thematic clusters.
- **Temporal assortativity:** speeches closer in time use similar economic terms.
- **Weak political assortativity:** party and president have limited impact.

Implication: Economic discourse is shaped more by historical context than political identity.

Textual Data Analysis: US Presidents Economics - Findings

- **Rich-club structure**: ∼430 central speeches.
- Core uses *generic terms* (e.g., *trade*, *interest*).
- Periphery includes *event-specific terms* (e.g., *yield*, *inflation*).



Textual Data Analysis: US Presidents Economics - Findings

- **Shared Economic Lexicon:** A stable economic vocabulary bridges centuries of presidential rhetoric.
- **Crisis Periods Drive Similarity:** Talks during economic turmoil show stronger economic focus.
- **Tool Transferability:** Approach generalizes to any large textual corpus with thematic focus.

Conclusion: Economic terms offer a robust axis for comparing political discourse across time.

Future Directions and Open Questions

Open Research Questions

- **Rank-size distribution of topics**:
  Are semantic or thematic topics in a corpus distributed according to a rank-size law?
- **Inverse inference from rank-size parameters**:
  Can we infer structural or contextual features of the originating phenomenon by interpreting the rank-size laws' fitted parameters?
- **Model selection across domains**:
  How can we rigorously determine which rank-size functional form (e.g., ZML, DGBD, UL) is best suited for a specific empirical setting?

References I

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, pages 74–76. Translation and introduction by Antonio Ciccone.

Ausloos, M. and Cerqueti, R. (2016). A Universal Rank-Size Law. *PLOS ONE*, 11(11):1–15.

Bishop, M. (2009). *Essential economics: an A to Z guide*, volume 22. John Wiley & Sons.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Cerqueti, R. and Ficcadenti, V. (2022). Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million. *Chaos, Solitons & Fractals*, 158:111975.

Cerqueti, R., Ficcadenti, V., Dhesi, G., and Ausloos, M. (2022). Markov Chain Monte Carlo for generating ranked textual data. *Information Sciences*, 610:425–439.

Fantozzi, P., Ficcadenti, V., and Naldi, M. (2025). The university research assessment dilemma: a decision support system for the next evaluation campaigns. *Scientometrics*, pages 1–42.

Ficcadenti, V. and Cerqueti, R. (2017). Earthquakes economic costs through rank-size laws. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(8):083401.

References II

Ficcadenti, V., Cerqueti, R., and Ausloos, M. (2019). A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches. *Expert Systems with Applications*, 123:127–142.

Ficcadenti, V., Cerqueti, R., Ausloos, M., and Dhesi, G. (2020). Words ranking and Hirsch index for identifying the core of the hapaxes in political texts. *Journal of Informetrics*, 14(3):101054.

Ficcadenti, V., Cerqueti, R., and Varde'i, C. H. (2023). A rank-size approach to analyse soccer competitions and teams: the case of the Italian football league 'Serie A'. *Annals of Operations Research*, 325(1):85–113.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572.

Lavalette, D. (1996). Facteur d'impact: impartialité ou impuissance. *Report, INSERM U*, 350:91405.

Mandelbrot, B. (1953). An informational theory of the statistical structure of languages. In Jackson, W., editor, *Communication Theory: Papers Read at a Symposium on "Applications of Communication Theory" held at the Institution of Electrical Engineers, London, September 22nd–26th 1952*, pages 486–502, London, UK. Butterworths. Symposium held at the Institution of Electrical Engineers (IEE).

Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In Wolman, B. B. and Nagel, E., editors, *Scientific Psychology*, chapter 29, pages 550–562. Basic Books Publishing Co., Inc., New York.

References III

Pareto, V. (1896). *Cours d'économie politique*, volume 1. Librairie de l'Université.

Zipf, G. K. (1945). The repetition of words, time-perspective, and semantic balance. *The Journal of General Psychology*, 32(1):127–148.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

The End