Firms innovation capabilities and corporate websites: evidence on Italian SMEs

Carlo Bottai[14] Lisa Crosato[24] Josep Domenech[34]
Marco Guerzoni[1 4] Caterina Liberati[14]

[1]University of Milano-Bicocca, Italy
[2]Ca' Foscari University of Venice, Italy
[3]Universitat Politècnica de València, Spain
[4]WebSight Observatory, University of Milano-Bicocca

WIN 2025 February 4-5, 2025

Conventional data in SMEs innovation studies

- **CIS** — Small businesses are only present on a rotating sample basis. Micro enterprises are not surveyed at all.
- **Balance Sheets** — for SMEs, many R&D expenses occurr informally, rather reported under general costs
- **Patents** — SMEs often don't have the capacity to patent. Different patenting propensity between and within sectors. Existence of non-patentable technological knowledge

All of these data sources suffer from a sensible delay between release and reference period

Corporate website data in SMEs Innovation studies

Enterprises use their publicly-viewable websites as a virtual window (Domènech et al., 2012)

Content Analysis:
more companies than suggested by conventional data sources reported undertaking R&D activities on their websites(UK: Gök at al., 2015)

web-based innovation indicators to detect products innovation can be developed (Germany: Kinnie et al. 2019, 2021)

results of the Community Innovation Survey for SMEs can be reproduced (Netherlands: Daas and van der Doef, 2020).

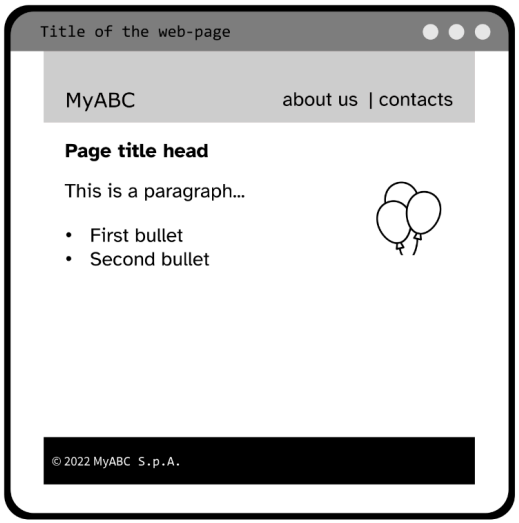
business innovation can be mapped (Flanders: Crijns et al., 2023)

Our Proposal: websites' HTML code to identify innovative SMEs

HTML describes the structure, interactivity and appearance of a web page. The ***tags*** describe website functionality:

HTML tags

- `<title>`
- `<a>`
- `<footer>`
- `<h1>`
- `<img>`
- `<li>`



the HTML code used to create a website reflects the interaction of a company's needs and skills with those of the programmer (Brinck, 2001)

the outcome of this interaction can reveal unobservable characteristics related to high levels of skills and creativity that may be indicative of an overall degree of innovativeness

Research design

need for a conventional label of 'innovative' SMEs to validate results and for websites to be scraped

1. BvD AIDA (inno label + websites' URL)
2. WayBack Machine for websites' homepages

- algorithm for correct attribution of a website to each SME:
  - the websites of 42,238 Italian manufacturing SMEs active in 2016 scraped from the WayBack Machine
  - **178 'innovative SMEs'** were identified according to the definition by the Italian ministry of Econ. Development (Italian Startup Act of 2013)
  - a group of **680 'non-innovative SMEs'** , similar to the innovative ones by geographical area, industry and size was built
  - Innovative and Non-innovative firms were organized in 100 matched samples

1. aggregate statistics measuring website size;
2. natural grouping of tags emerging from the data;
3. differences between innovative vs non-innovative firms with respect to the usage of the tags

Innovative SMEs websites are bigger (visual)

.

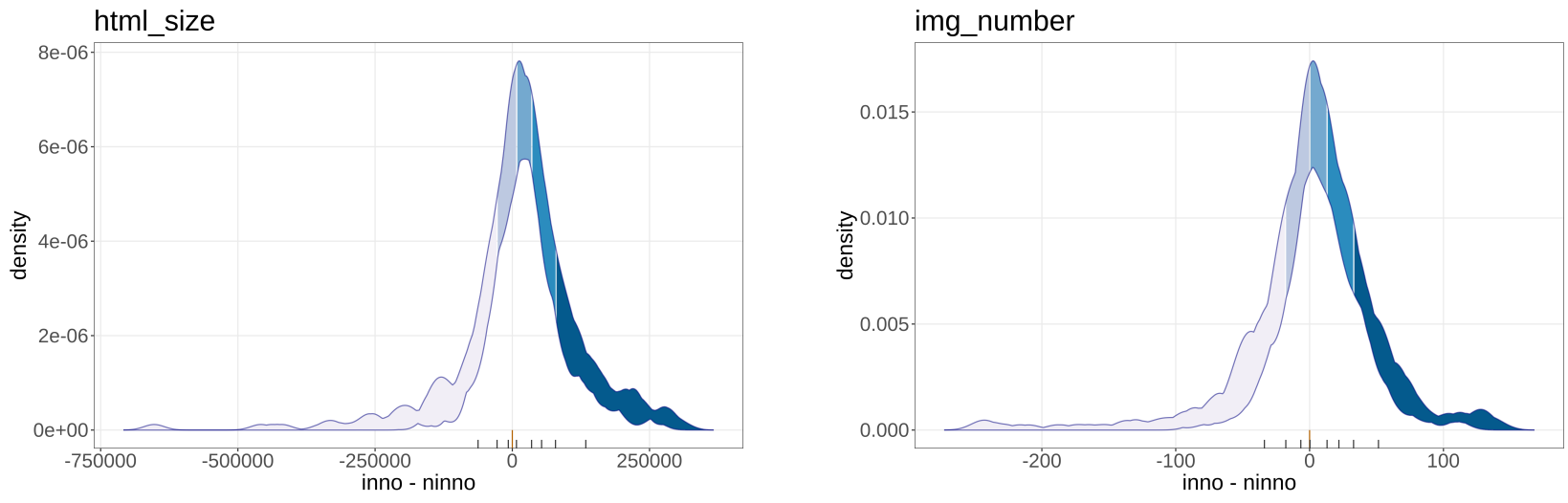


Figure 1: Density distributions of the difference in aggregate statistics between innovative and non-innovative SMEs (fan on 100 samples)

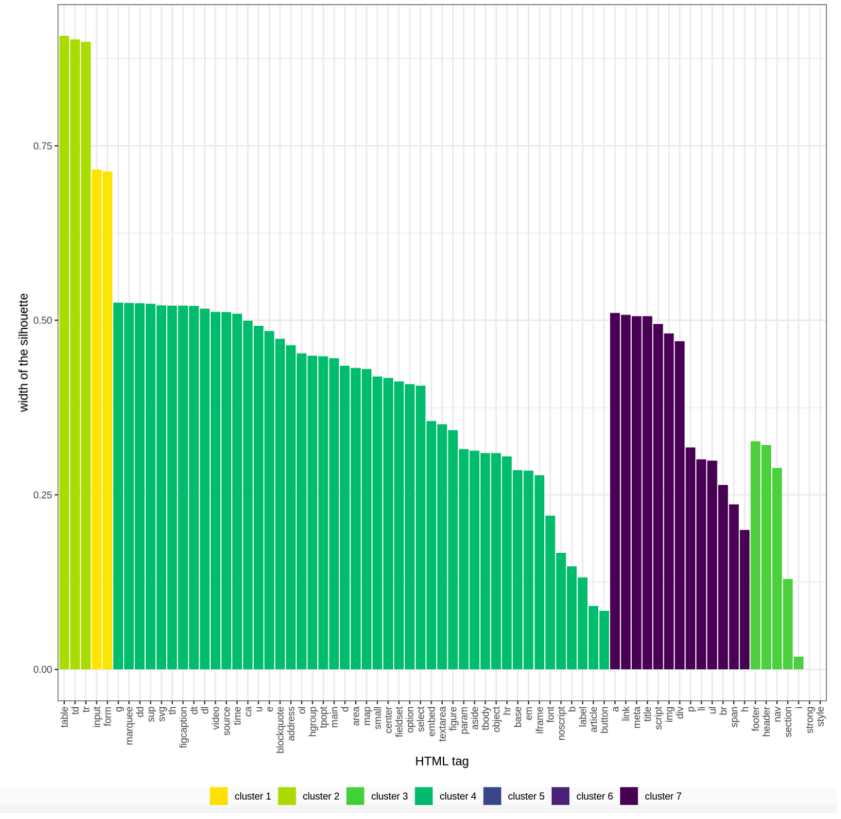
Innovative SMEs websites are bigger (tests)

Variable	observed	Quantile test				Wilcoxon	paired
	differences	10%	20%	30%	40%	signed test	t-test
html_size	+	0.002	0.000	0.000	0.000	0.000	0.000
gztext_size	+	0.009	0.000	0.000	0.000	0.000	0.000
text_size	+	0.010	0.000	0.000	0.000	0.000	0.002
img_number	+	0.039	0.006	0.006	0.002	0.006	0.023
href_number	+	0.000	0.000	0.000	0.000	0.000	0.005
linkhref_number	+	0.000	0.000	0.000	0.000	0.000	0.000

Table 1: Quantile test (D method), Wilcoxon signed-rank test and t-test on the differences of web-based aggregates between innovative SMEs and paired firms. Median *p*-values on one hundred sets of matched samples.

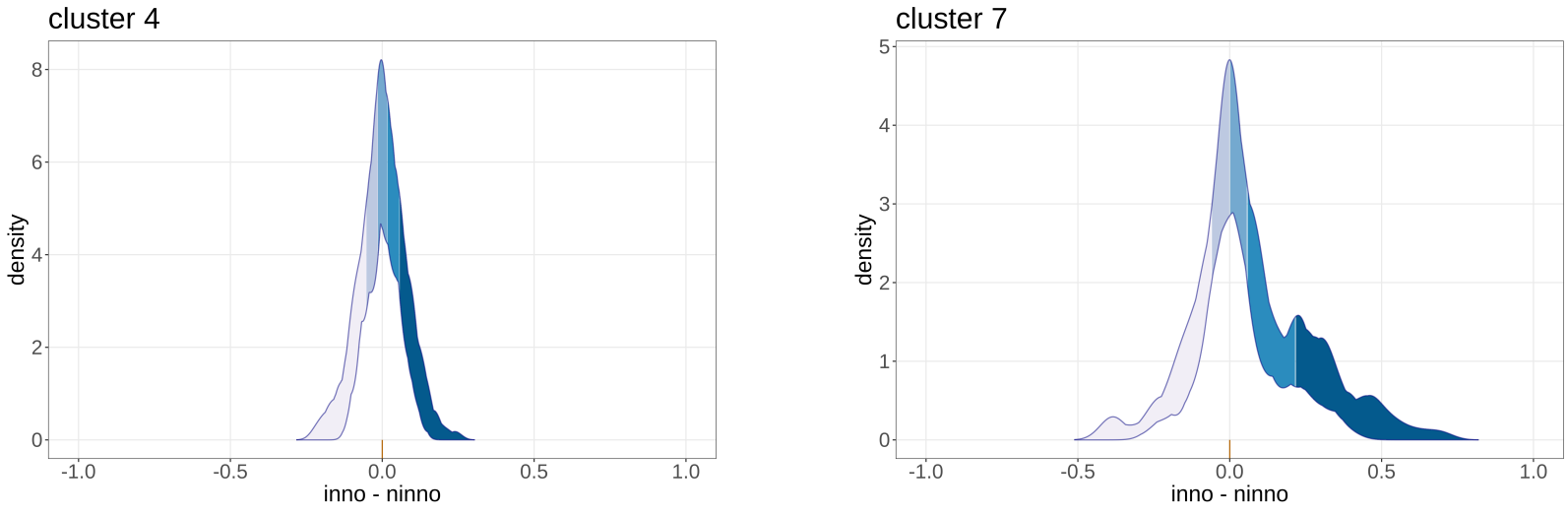
Tags grouped in coding ways

To detect natural grouping of the tags, we apply a hierarchical cluster analysis on our HTML tags based on their pairwise similarity (presence in the same webpage)
We choose 7 clusters, based on four evaluation criteria. (in the figure Silhouette - Rousseeuw, 1987)



- C1 : 2 tags (user inputs)
- C2 : 3 tags (table building)
- C3 : 5 tags (HTML5)
- C4 : 46 tags (seldomly used)
- C5: 1 tag `<strong>`
- C6: 1 tag `<style>`
- C7 : 13 tags (of common use)

Differences in coding ways (visual)



(a) (b)

Figure 2: Density plot of the difference in the adherence to clusters between innovative and non-innovative firms (fan over 100 samples).

Differences in coding ways (tests)

Cluster	observed differences	Quantile test 20%	Quantile test 40%	Wilc. signed	paired t-test
cluster 1	none	0.412	0.379	0.391	0.399
cluster 2	-	0.000	0.000	0.001	0.001
cluster 3	+	0.008	0.003	0.003	0.003
cluster 4	none	0.663	0.524	0.603	0.663
cluster 7	+	0.000	0.009	0.003	0.001

Table 2: Quantile test, Wilcoxon Signed-Rank test and t-test on the difference of adherence to a given cluster (Median p-values over 100 matched samples). Clusters 5 and 6 are removed since composed by one tag only.

Difference in the use of single HTML tags (tests)

HTML tag	Cluster	observed differences	Quantile test 20%	Quantile test 40%	Wilc. signed	paired t-test
<table>	C2	-	0.007	0.002	0.005	0.175
<td>	C2	-	0.012	0.002	0.006	0.184
<tr>	C2	-	0.009	0.002	0.005	0.158
<footer>	C3	+	0.000	0.000	0.001	0.000
<header>	C3	+	0.011	0.014	0.011	0.051
<i>	C3	+	0.002	0.002	0.002	0.022
<nav>	C3	+	0.022	0.028	0.019	0.015
<section>	C3	+	0.008	0.015	0.009	0.004
<a>	C7	+	0.006	0.000	0.002	0.008
	C7	+	0.002	0.000	0.000	0.001
<h>	C7	+	0.004	0.000	0.002	0.009
<img>	C7	+	0.036	0.048	0.041	0.082
	C7	+	0.029	0.000	0.004	0.037
<link>	C7	+	0.000	0.000	0.000	0.000
<meta>	C7	+	0.006	0.004	0.002	0.007
<p>	C7	+	0.008	0.004	0.005	0.065
<script>	C7	+	0.002	0.000	0.000	0.001
	C7	+	0.002	0.000	0.001	0.013
<title>	C7	none	0.051	0.059	0.054	0.171
	C7	+	0.004	0.000	0.002	0.011

Table 3: Quantile test, Wilcoxon Signed-Rank test and t-test on the differences of tags usage (Median p-values over 100 matched samples). C2, C3 and C7

Conclusions

- Contributions:
  - innovative SMEs websites are bigger, richer, more up-to-date and more complex
  - HTML tags naturally group into coding ways, three of which discriminate between innovative and non-innovative firms
  - the same was found also for single HTML tags within clusters
- Advantages:
  - free and real-time
  - HTML-indicators are more stable than text-indicators
  - Cross-Countries comparisons are more simple
- Limitations:
  - The sample of innovative SMEs is small
  - We did not use text/images of the corporate websites
  - firms' age?
- Future research
  - Build an innovativeness index (probability to be innovative given the combination of HTML tags that you use), to be used also at the local level
  - Estimate a supervised classifier to predict innovative SMEs with conventional/unconventional data

References

- BlÃ¡zquez, DomÃ¨nech, Gil, and Pont (2019). Monitoring e-commerce adoption from online data. Knowledge and Information Systems
- Crijns, A., Vanhullebusch, V., Reusens, M., Reusens, M., & Baesens, B. (2023). Topic modelling applied on innovation studies of Flemish companies. Journal of Business Analytics, 1-12.
- Oslo Manual (1992 and subsequent versions). OECD proposed guidelines for collecting and interpreting technological innovation data.
- Daas and van der Doef (2020) Detecting innovative companies via their website. Statistical Journal of the IAOS
- GÃ¶k, Waterworth, and Shapira (2015). Use of web mining in studying innovation. Scientometrics
- Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. Journal of classification, 3, 5-48.
- Kinne and Lenz (2021) Predicting innovative firms using web mining and deep learning. PLOS ONE
- 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.
- Sokal, R.R. and Michener, C.D. (1958): A Statistical Method for Evaluating Systematic Relationships. The University of Kansas Scientific Bulletin, 38, 1409-1438.
- Wilcox, R. R., and Erceg-Hurn, D. M. (2012): Comparing two dependent groups via quantiles. Journal of Applied Statistics, 39(12), 2655-2664.

Innovators: definitions from Official Statistics

EU Official Statistical Offices use the following definitions:

Innovators with realized innovations: Enterprises that realized and successfully implemented technological innovation in the period under review. Technological innovation consists of product and/or process innovation.

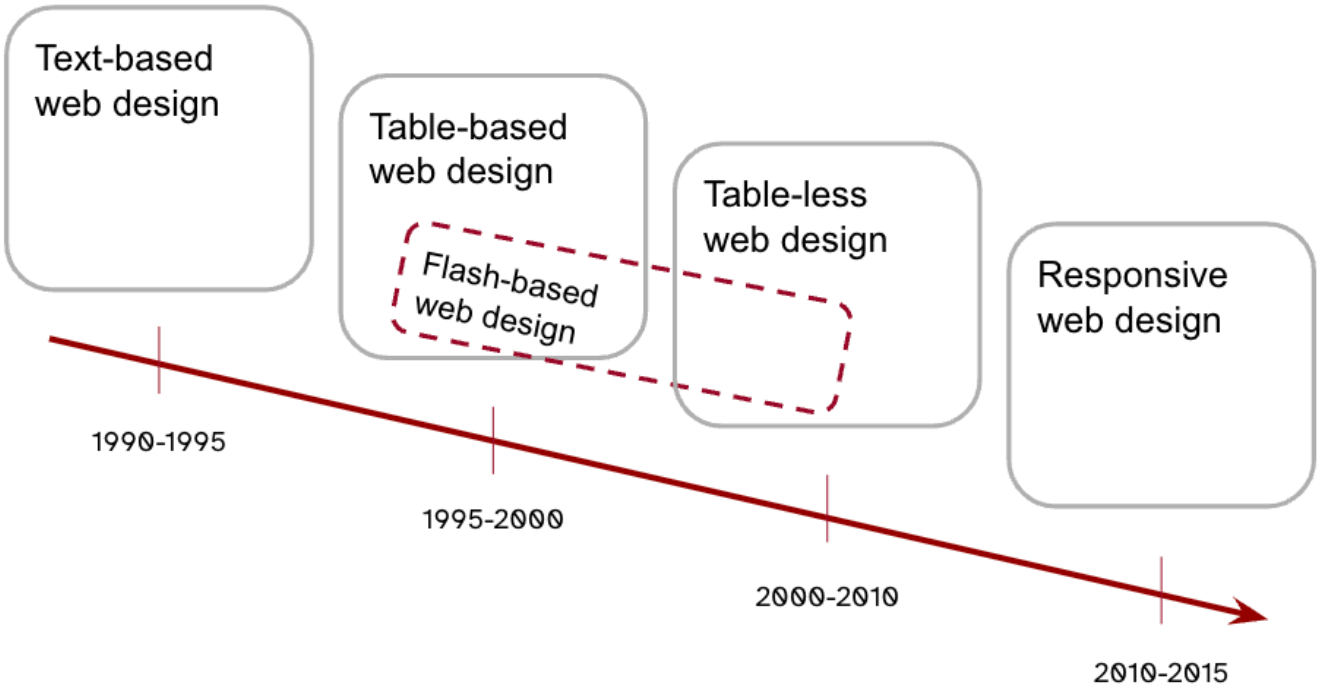
- **Technological innovators**
  Enterprises with product and/or process innovation.
- **Product innovators**
  Enterprises that conducted innovation projects that resulted in the implementation of new or significantly improved goods or services.
- **Process innovators**
  Enterprises that conducted innovation projects which resulted in the implementation of new or significantly improved production processes, distribution methods, or support activities for goods or services.

Innovative SMEs

The Italian Startup Act implied the creation of specific sections in the Italian companies register for classifying innovative startups and innovative SMEs.

- Firms must not distribute profits and must develop, produce, and commercialize innovative goods or services of high technological value
- Firms must fulfill at least one of the following conditions:
  - They must allocate at least 15% of expenses to R&D
  - Employ PhD students or Master's degree holders comprising at least one third or two thirds of the workforce, respectively
  - Have deposited, or have in license, a registered patent or a legally registered computer program.

Web design history (tentative)



Data quality assessment

We searched in the websites correspondance with the following info:

firm-URL $\rightarrow$

- identification number (*codice fiscale*)
- business address (street name, number, and postal code)
- telephone number

	Predicted False	Predicted True
Actual True	0.060	0.940
Actual False	0.930	0.070

Table 4: Confusion matrix from data-quality assessment procedure

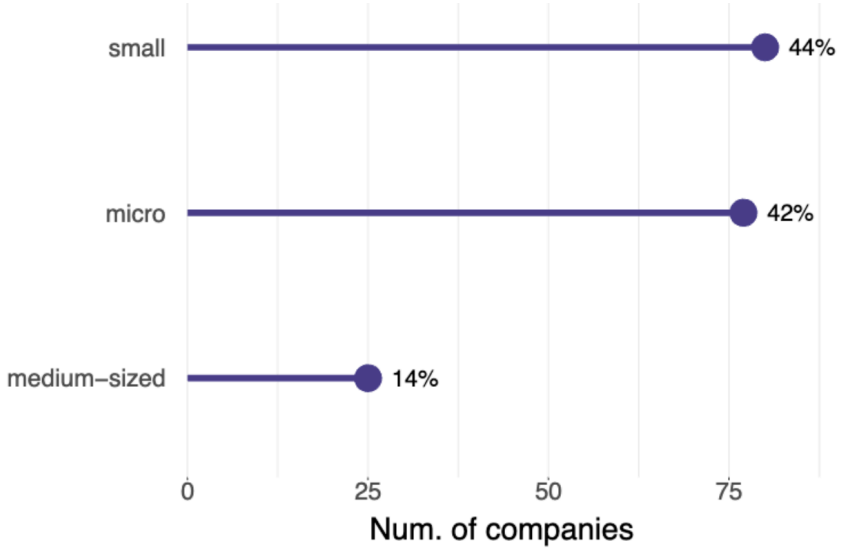
HTML indicators

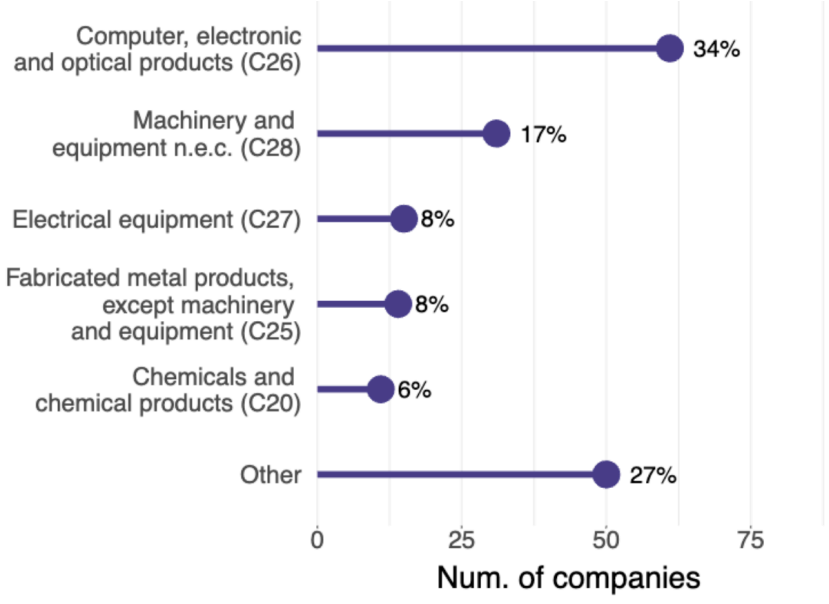
```
<!DOCTYPE html>
<html>
  <head>
    <title>A page containing some text</title>
    <meta name="author" content="John Doe">
    <meta name="keywords" content="lorem;ipsum">
    <link href="main.css" rel="stylesheet"></link>
    <script src="script.js"></script>
  </head>
  <body>
    <header><img src="nameplate.png"></header>
    <div>
      <h1>Lorem ipsum dolor sit amet</h1>
      <p>Ut enim ad minim veniam,
        quis nostrum exercitationem</p>
      <p>Duis aute irure dolor in
        <a href="https://www.dolor_sit_am.et">
          reprehenderit</a></p>
    </div>
    <footer style="text-align:center">
      &copy; 2022 Author &ndash; VAT Num. X123A.
    </footer>
  </body>
<html>
```

id	variable	value
X123A	html_size	516
X123A	text_size	175
X123A	gztext_size	138
X123A	img_number	1
X123A	href_number	1
X123A	linkhref_number	1

id	tag	count
X123A	head	1
X123A	body	1
X123A	title	1
X123A	meta	2
X123A	link	1
X123A	script	1
X123A	header	1
X123A	div	1
X123A	footer	1
X123A	img	1
X123A	h	1
X123A	p	2
X123A	a	1

Descriptive Stats of Innovative SMEs





Clustering HTML tags

To detect natural grouping of the tags, we cluster our HTML tags based on their pairwise similarity ('simple matching coefficient' $s_{tt'}$ by Sokal and Michener,1958)

$$s_{tt'} = \frac{\alpha+\delta}{\alpha+\beta+\gamma+\delta}$$

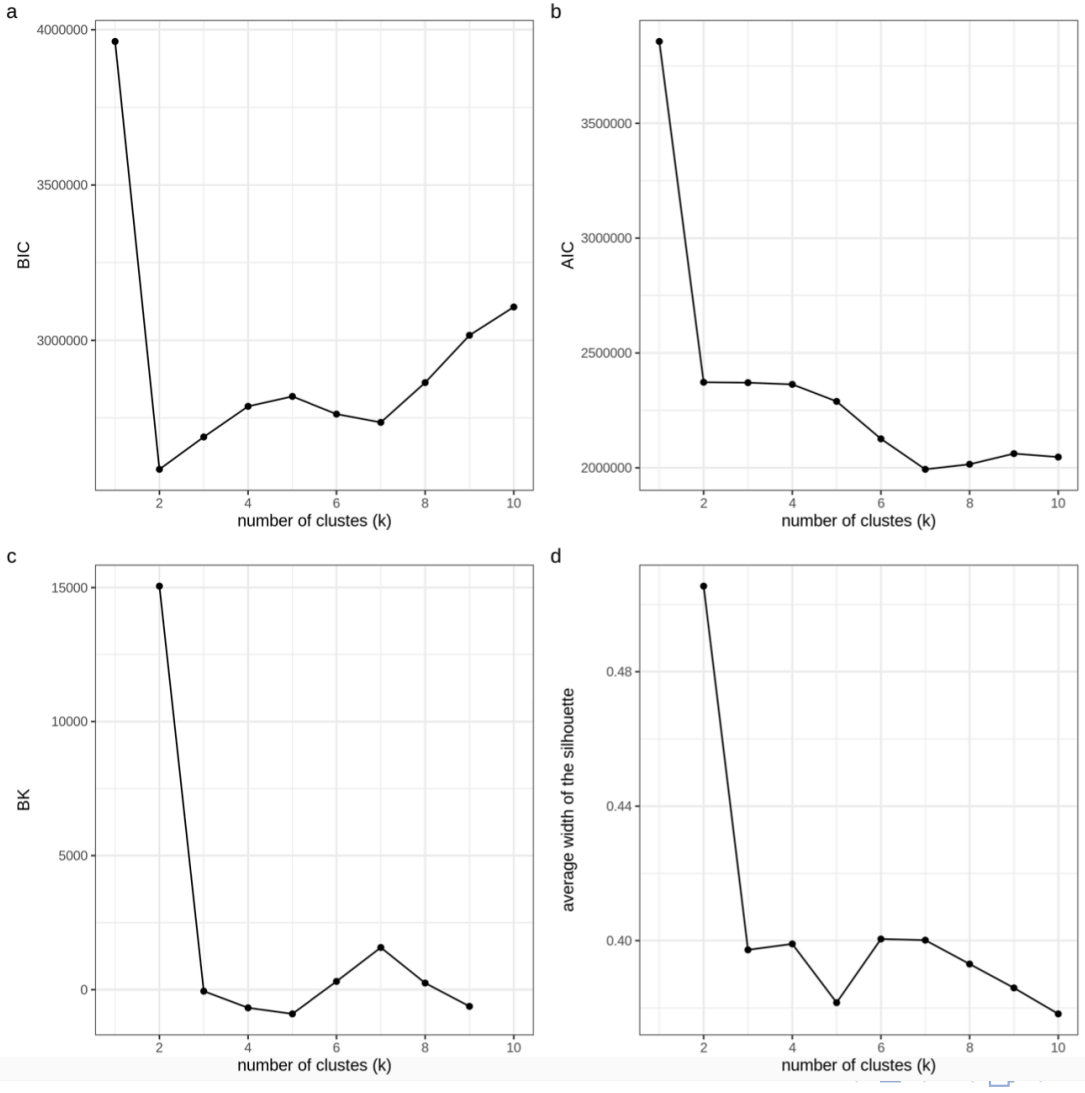
t \ t'	Absent	Present
Absent	α	β
Present	γ	δ

Euclidean Distance matrix between tags $d_{tt'} = \sqrt{1 - s_{tt'}}$ (Gower and Legendre,1986)

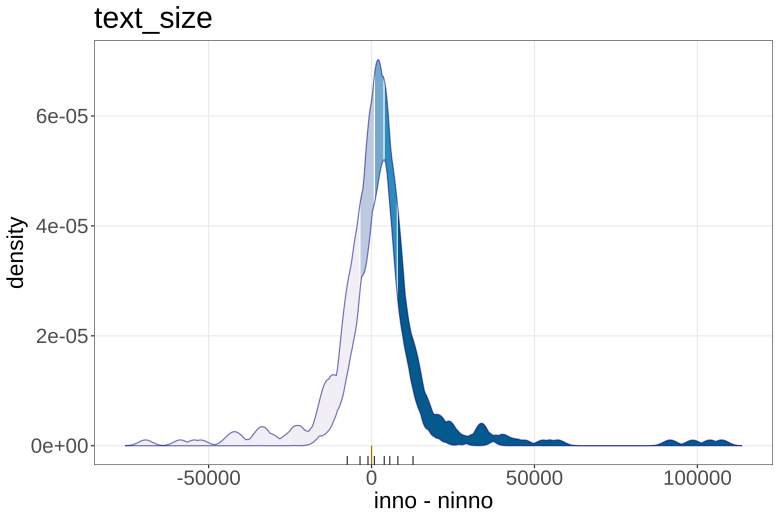
We used hierarchical clustering UPGMA (Sokal and Michener, 1958)

Criteria for setting the number of clusters

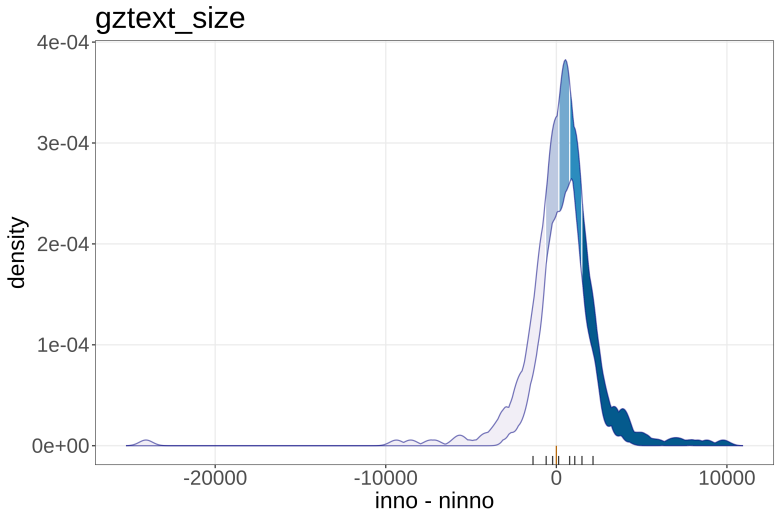
Criteria: BIC, AIC, Best K, Silhouette



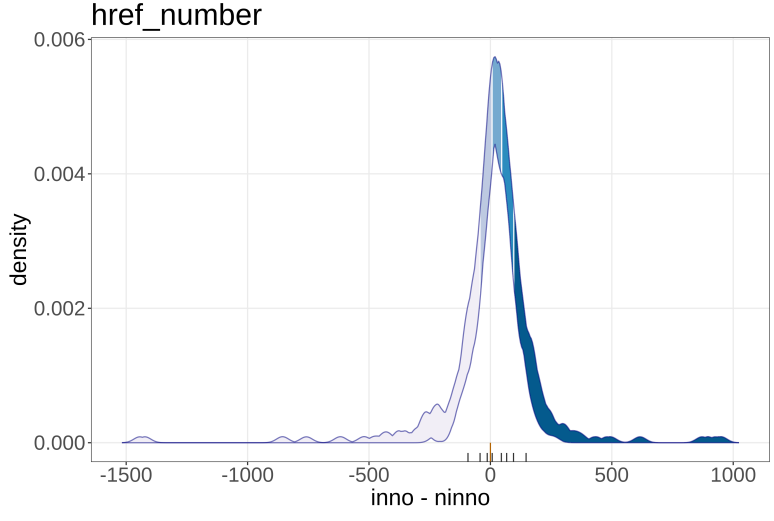
Size difference, remaining variables

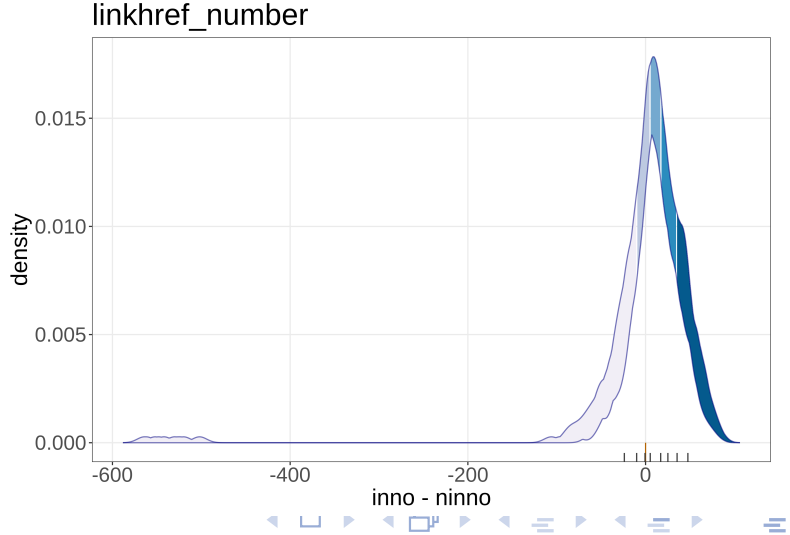


(a)

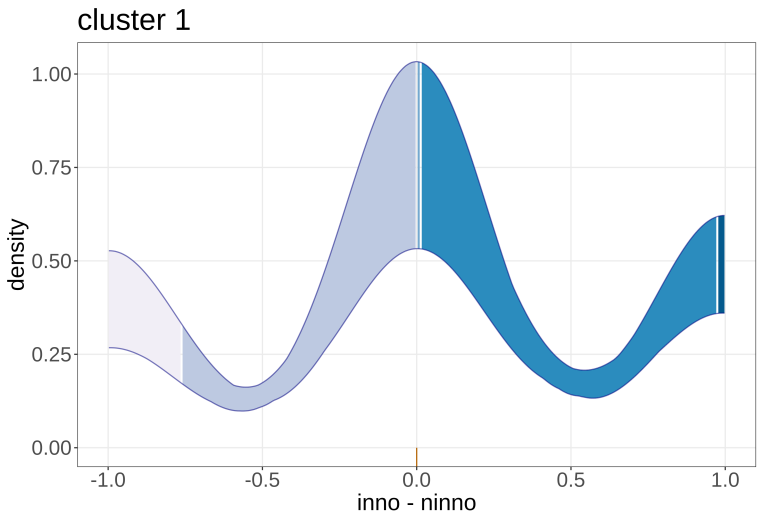


(b)

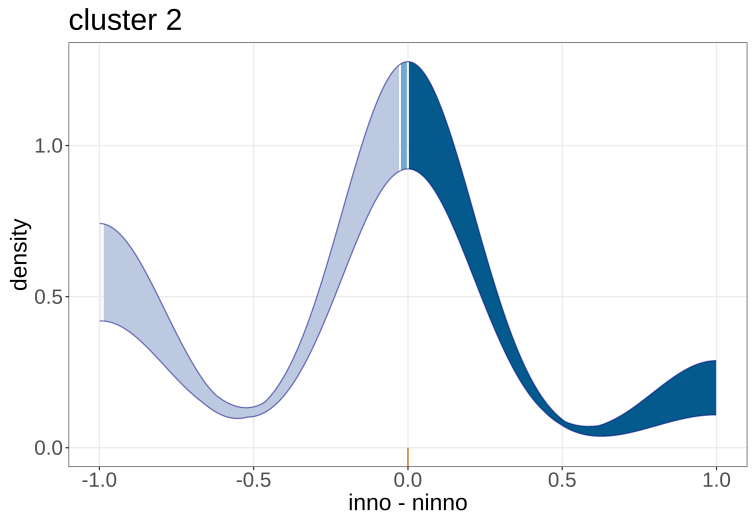




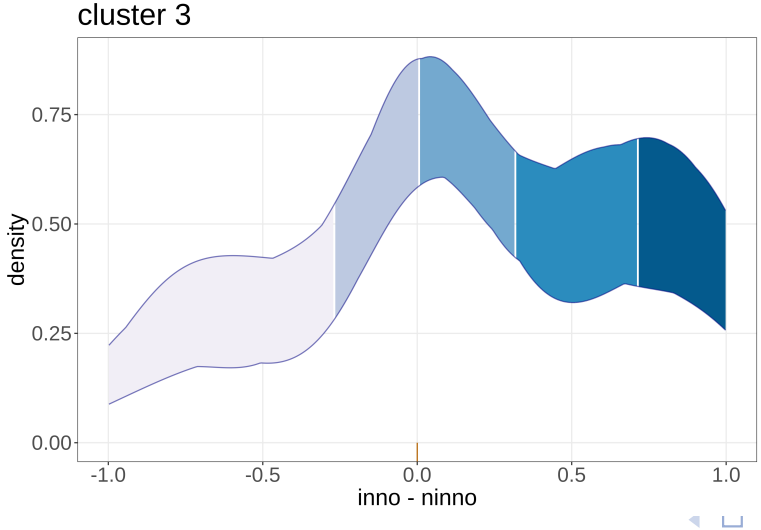
Difference in adherence to clusters, remaining clusters



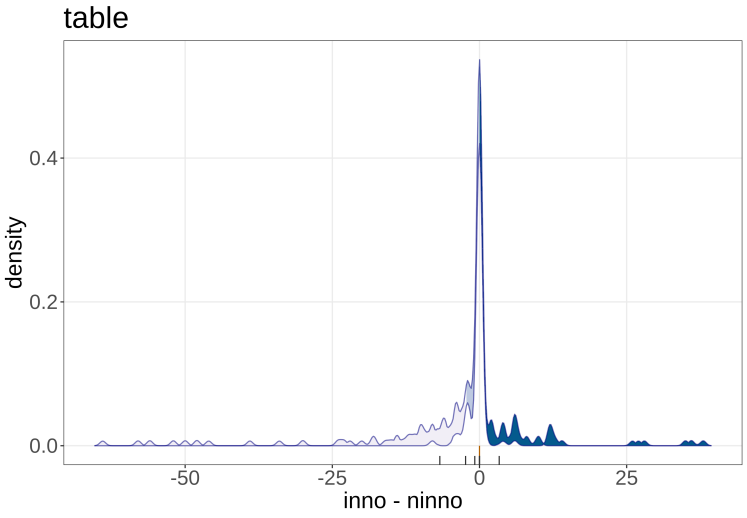
(a)



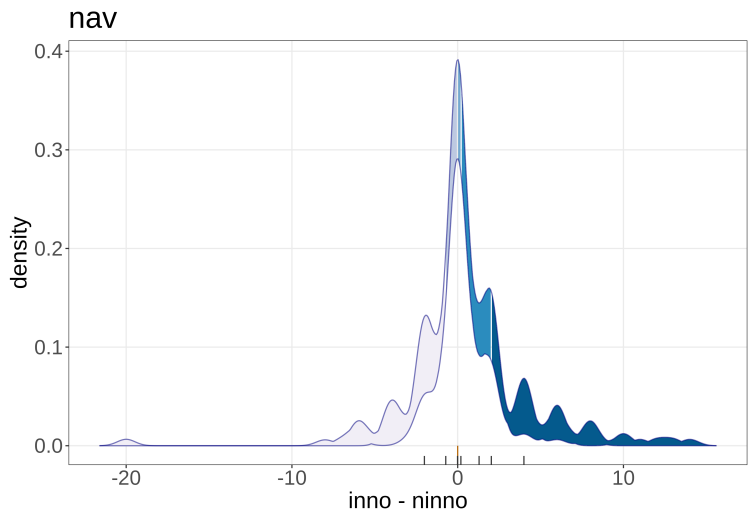
(b)



Difference in the use of single HTML tags (visual)



(a)



(b)