DATA CENTER SUSTAINABILITY

Best practices and future scenarios

An innovative methodology for analysing sustainability of data centers via websites

Marco Bettiol[1], Shira Fano[2], Gianluca Toschi[2]

[1]University of Padova & DSEA, [2]Fondazione Nord Est

Padova, December 16th, 2022

Background

- Due to the **increase of digitalization** and demand of data related services, **data centers** are becoming a fundamental element of the supply-chain of digital technologies.

- This rapid development of data center has enormous impact on the **environment**.

- Given the growing international attention on climate change, it's essential to investigate the **links** between transition to a **data center-driven economy** and **sustainability**.

Lack of data

- Official statistics are currently not compiled on data center **energy use** at national or global levels, but several authors tried to estimate its **environmental impacts** in terms of carbon footprint.

- Belkhir and Elmeligi (2018) calculate an impact around 2,5 % to 3% of global greenhouse gas emissions (GHGE) at the global level in 2018 with a projection of to reach 3 to 3,5% in 2020.

- **Electricity consumption** of Google data centers, is now bigger than the electricity consumption of the State of Maine in US (Bergen, 2022).

Research questions

1. How are Data centers taking into account **sustainability issues**? Can we differentiate between soft and strong sustainability engagement (i.e. certifications)?

2. Which practices are data centers implementing to **mitigate** their **carbon footprint**?

3. Focusing on Europe, does **geographical location** matter in terms of sustainability initiatives?

We perform an **exploratory analysis** to map and understand the *status quo* related to sustainability initiatives in the data center management in Europe

Methodology

- We use web scraping tools to build a **new database** with the text of websites and analyzing the text using an **innovative methodology**, based on text analysis and machine learning techniques

- We analyzed website content using a **bag of word approach** and **TF-IDF** (term frequency–inverse document frequency) text classification method (Jones, 1972).

- This methodology allowed us to create an **index** of the presence and degree of discussions about **sustainability** on each website.

Advantages

- It is in the company's best interest to provide an **accurate representation** of itself via websites
- Information from websites is relatively inexpensive to obtain (Gök et al., 2015), publicly available, and up to date
- Collecting data from the web is also **nonintrusive** (Arora et al., 2016), which is increasingly relevant in a period when response rates to research questionnaires are in constant decline.

The Internet can now be considered a source of data in combination with traditional tools.

Data

- Our original dataset consists in the text of 342 websites of Europen data centers with at least an English website

- We obtained the list of data centers from different sources/databases:
  1. https://www.datacentermap.com
  2. https://www.impresaitalia.info;
  3. list of companies that are part of the European Data Centre Association (EUDCA);
  4. https://cispe.cloud/members (CISPE's members101);
  5. https://sciencebasedtargets.org/companies-taking-action

- Data was obtained using Qiba (Quantitas Intelligent Business Analyzer), a web crawling and scraping tool.
- Text was cleaned and pre-processed to create a corpus to be analyzed (remove symbols, remove punctuation, transform into lower case…)

Data

- Websites describe data center characteristics and services offered
- But they also talk about sustainability, we see *sustainability, green, cooling*
- *Sustainability ≈5000*
- *Data services ≈50.000*

BoW: Bag of words approach

- We analyse the topic of sustainability using a Bag of Word approach
- The idea is that we can estimate the **relevance** of a topic (sustainability) in a document (website) by the **frequency** of a given set (bag) of pre-selected words. (Zhang et al., 2010)

- Given the complexity of the concept of sustainability, we identify and analyze **five subtopics**: ***metrics, green factors, footprints, circular economy and certifications***.

- In particular, given the word frequency in each BOW we compute the TF-IDF indicator (term frequency – inverse document frequency) allowing us to **rank company** websites with respect to **sustainability** bags.

Identifying sustainability sub-topics

- To identify sustainability sub-topics we implement a hybrid approach apply a two-step procedure
- In the first step we use a qualitative lexicon-based approach and identify keywords linked to sustainability for the five baskets: metrics, green factors, footprints, circular economy and certifications.
- In the second step we used a word embedding methodology to enrich the initial baskets. In detail, we used *word2vec*, one of the most popular word embeddings techniques.

Advantages:

1. enables the retrieval of terms initially missed by the researcher
2. allows to search for terms in the specific language of the corpus analyzed.

Bags

Metrics	Green Factors	Footprint	Circular Economy	Certifications
carbon usage effectiveness	clean energy	carbon	circular	certified energy efficient datacenter award
data center infrastructure efficiency	environmental policy	carbon emissions	circular economy	green globes building certification
renewable energy factor	environmental report	carbon footprint	disposal	materials analytical services certified green
cue - Carbon Usage Effectiveness	geothermal	carbon free	dispose	building energy innovators council
dcie - Data Center Infrastructure Efficiency	geothermal energy	carbon neutral	e-waste management	building energy innovators council
electronics disposal efficiency	hydro	carbon reduction	life cycle assessment	ceeda
efficiency metrics	hydroelectric	climate neutral	lifecycles	certification
electronics disposal efficiency	recycle heat	co2 emission	recycle	certified
energy consumption	renewable energy	decarbonisation	recycled	certified recycling company
energy efficiency	renewables	decarbonizing	repair	climate neutral data center
energy reuse effectiveness	solar	dioxide	repairing	climate neutral data center pact
energy use	solar energy	emission	repairs	fossil free data
energy used	water conservation	emissions	reuse	gbi
energy reuse effectiveness	wind	environmental footprints	reused	green building initiative
green energy coefficient	wind energy	environmental impacts	waste	green certification
green energy coefficient		footprint		iso 14001
metric		greenhouse		iso 14040
power usage effectiveness		greenhouse gas emissions		iso 50001
power usage effectiveness		low carbon		iso 14001
ref- Renewable Energy Factor				iso 50001
sustainability metrics				leadership in energy and environmental design
water consumption				leadership in energy and environmental design
water usage				mas certified green (Materials Analytical Services) Certified Green® certifies low VOC (volatile organic compound) emitting products and materials
water usage effectiveness				sustainability certification
wue - Water Usage Effectiveness				

TF-IDF Term Frequency-Inverse Document Frequency

- We used a TF-IDF indicator to rank websites according to the strength of their communication about sustainability.

- Higher the frequency of a certain term, the more important it is to the company.

- TF-IDF score for a term is higher when it appears *i)* frequently and *ii)* only in a small number of websites (rare)

$$TFIDF(t,d,D) = tf(t,d)*idf(t,D) = tf(t,d)*log(N/n_t)$$

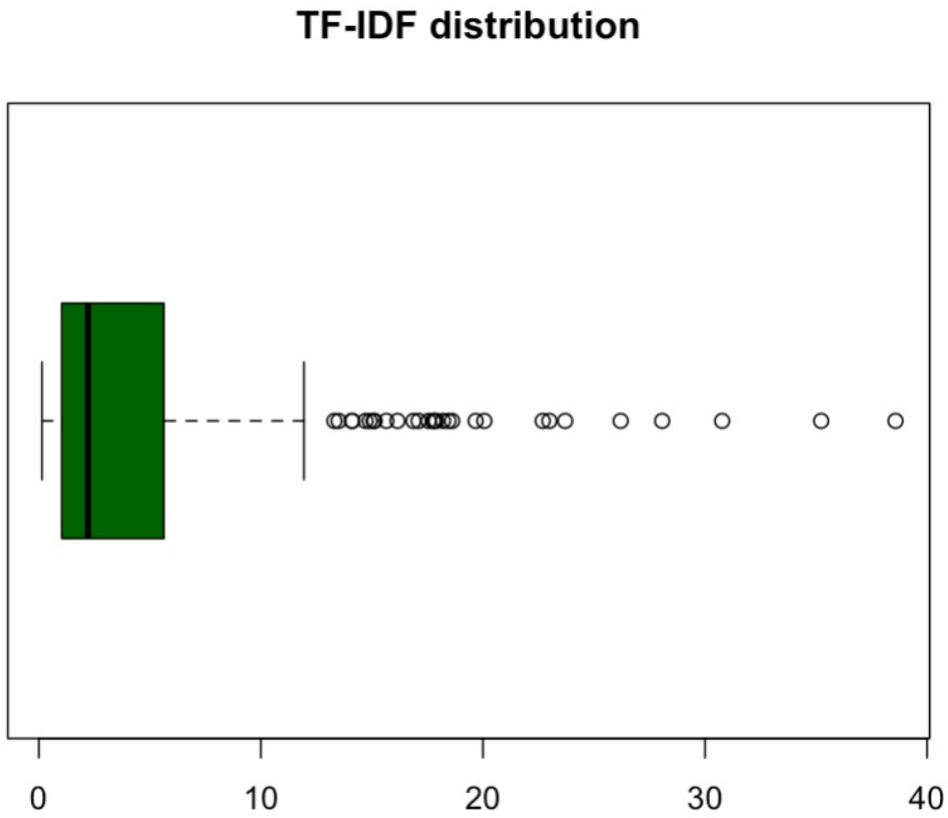
N number of documents,
n_t number of documents containing a given word

- In detail, we implement the version by Paik (2013) that takes into account:
    1. different website size
    2. different basket length

Results

Etichette di riga	Rank	2 Metrics	3 Green Factors	4 Footprint	5 Circular Economy	6 Certifications	Totale
Nlyte Software	1	13,88	2,08	5,30	4,51	12,81	38,58
DigiPlex	2	6,71	10,43	9,39	5,31	3,38	35,23
CONAPTO	3	2,17	10,59	7,28	4,35	6,38	30,77
Lamda Hellix	4	11,87	3,56	6,22	1,36	5,06	28,07
Echelon Data Centres	5	3,62	5,68	10,58	6,32		26,20
Lefdal Mine Data Center	6	5,49	5,67	8,33	2,54	1,67	23,70
Green Mountain Data Center	7	4,27	5,98	5,81	5,63	1,30	22,99
Dassault Systemes (DELMIAworks)	8	1,30	3,60	8,80	8,38	0,61	22,69
Atos	9	1,06	4,26	7,65	6,21	0,88	20,06
Submer	10	11,20	0,29	5,81	2,15	0,22	19,67
Tele2 AB	11	1,96	2,18	8,19	6,14	0,14	18,61
Telia Company AB	12	0,49	2,91	9,15	5,81	0,10	18,44
Google	13	1,77	3,74	6,66	4,33	1,70	18,20
IBM	14		6,50	5,99	4,80	0,60	17,89
Turk Telecom	15	2,86	4,43	5,80	2,20	2,51	17,81
EnerKey	16	3,28	1,27	5,32	5,53	2,35	17,75
Vodafone Group Plc	17	1,30	1,92	5,66	7,38	1,28	17,55
Ark Data Centres	18	3,97	2,06	6,44	1,40	3,22	17,10
Kao Data	19	2,25	3,93	6,81		3,89	16,87
Capacity Media	20	2,37	1,48	8,07	1,97	2,26	16,15
Verne Global	21		11,15	3,99		0,51	15,65
HPE	22	1,59	2,37	3,74	6,93	0,50	15,12
Workspace Technology	23	4,31	1,37	6,62	1,57	1,21	15,07
Telenor Group	24	1,03	4,80	7,52	1,56		14,90
Orange	25	1,31	2,01	5,13	6,15	0,12	14,71
EcoDataCenter	26	3,91	3,39	4,63	1,02	1,18	14,12
NDC-GARBE	27	5,34	3,30	3,75	1,71		14,09
Cisco Systems	28	0,87	2,28	5,29	4,39	0,67	13,50

TF-IDF distribution



TFIDF distribution quartiles				
Min	1st quartile	Median	3rd quartile	Max
0,14	1,1	2,2	5,6	38,6

Sustainability and Data Center Location

	Northern Europe		
	1st Qu.	Median	3rd Qu.
Metrics	1,10	1,90	3,30
Green factors	0,20	0,90	2,18
Footprint	0,60	1,80	5,10
Circular economy	0,53	1,00	1,78
Certifications	0,50	0,80	2,20

	Western Europe		
	1st Qu.	Median	3rd Qu.
Metrics	0,60	1,80	2,88
Green factors	0,20	0,25	1,43
Footprint	0,60	1,90	2,70
Circular economy	0,58	0,90	1,50
Certifications	0,50	0,70	1,23

	Eastern Europe		
	1st Qu.	Median	3rd Qu.
Metrics	0,60	1,20	2,00
Green factors	0,20	0,20	0,20
Footprint	0,40	0,60	1,63
Circular economy	0,60	0,70	1,30
Certifications	0,30	0,40	0,70

	Southern Europe		
	1st Qu.	Median	3rd Qu.
Metrics	0,50	0,95	2,03
Green factors	0,20	0,50	1,35
Footprint	0,50	1,30	2,65
Circular economy	0,50	0,80	1,60
Certifications	0,38	0,65	1,48

- Descriptive statistics of the TF-IDF confirm differences across regions, although not very large in terms of magnitude
- Data centers in Northern EU highly consider sustainability issues
- Southern countries are catching up

Correlation between bags

	Dependent variable:				
	Metrics			Certifications	
	(1)	(2)	(3)	(4)	(5)
Green factors	0.337***				0.229**
	(0.118)				(0.102)
Footprints		0.295***			−0.091
		(0.063)			(0.083)
Circular eco			0.308***		−0.027
			(0.097)		(0.097)
Metrics				0.453***	0.452***
				(0.050)	(0.068)
Constant	1.152***	0.766***	1.068***	0.477***	0.470
	(0.311)	(0.246)	(0.222)	(0.116)	(0.298)
Observations	102	130	138	157	69
R^2	0.075	0.146	0.069	0.346	0.474
Adjusted R^2	0.066	0.140	0.062	0.342	0.441
Note:				*p<0.1; **p<0.05; ***p<0.01	

- There is a positive correlation between sustainability principles (green factors, footprints and circular eco) and specific topics in the metrics basket: PUE power usage effectiveness, WUE water usage effectiveness, green energy coefficient
- Companies mentioning sustainability metrics tend to have sustainability certifications.

Conclusions

- Large digital companies (Facebook, Google, IBM, Microsoft, ecc.) are not the only companies involved in greening the data center.
- We discover that data center providers located in North Europe are on average more than double keen to adopt environmental sustainability initiatives (culture/favourable climate)
- There is a coherence between what is said (online) and what is done (certifications)

Thank you for the attention!