Web-scraping as a source for producing e-commerce indicators: findings from a pilot in Brazil

Thiago Meireles (NIC.br)
Marcelo Trindade Pitta (NIC.br)
Pedro Luis do Nascimento Silva (ENCE)

Introduction

The Brazilian Network Information Centre (NIC.br) is a non-profit civil entity created in 2005 to implement the decisions and projects designed by the Brazilian Internet Steering Committee. Within NIC.br, the Regional Center for Studies on the Development of the Information Society (Cetic.br) is the branch that produces statistics on Information and Communication Technologies (ICT) for policymaking. As part of its portfolio, Cetic.br conducts the ICT Enterprises annual survey of companies with 10+ employees operating in Brazil. This survey aims to measure the presence and use of ICT in these companies, covering topics such as infrastructure, appropriation, and use of new technologies by the private sector, as well as perceptions regarding their potential benefits to their activities.

One of the main indicators produced regularly by the ICT Enterprises survey refers to e-commerce infrastructure and practices adopted by the companies obtained by self-declared responses to the so-called E-module. Considering that e-commerce implies the use of web sites, some years ago a proposal was made to investigate the possibility of collecting this information directly from the web, thus reducing the response burden for companies in the sample. If successful, this approach would help reduce survey costs and perhaps improve response rates. Furthermore, this could enable production of more timely and disaggregated statistics.

We have been experimenting with this idea since 2017 and found that it is possible to automate data collection through web-scraping to classify a company's website in one of two categories: with e-commerce and without e-commerce. However, for Brazil, estimates obtained using web scrapped data have so far failed to match e-commerce prevalence rates when compared to those obtained from traditional survey-based statistics relying on self-declared indication that the company does offer e-commerce facilities.

An alternative approach we have been exploring involves combining the data collected via questionnaires from the ICT Enterprises survey (designed data) with web scrapped data for the sampled companies (found data). This is the basis for our ongoing pilot for replacing part of the ICT Enterprise survey's E-module.

This paper describes the findings of this pilot, discussing the differences between the web universe and the enterprise sampling frame, the challenges faced when collecting data from the web for a specified probability sample of companies, and the model developed to predict a site's e-commerce availability from the web-scrapped data.

History of the project

The idea behind this pilot project – a continuation of a project that begun in 2017 – is to evaluate the possibility of collecting data through web-scraping sites on the Internet and using these to provide e-commerce prevalence indicators that agree with those that are currently produced by the ICT Enterprises survey. If successful, this approach would enable reducing the size of the actual survey questionnaire and possibly enable increasing the data availability both in terms of frequency and disaggregation. The increase of the sample would be made by sampling from the Brazilian official enterprise register and by linking the sampled companies to the Brazilian domains database (*.br*), thus enabling the web-scraping of the sites of the sampled companies and the estimation of the target indicators. This linkage study was conducted between 2021 and 2022. The results are discussed in the final section of this article.

Based on findings of previous studies [4], the ICT Enterprises survey changed the questions on e-commerce. Before 2019 the set of questions asked the way that sales were made through the web, and the specific question about selling through websites did not explicitly separate social networks and marketplaces. Some analysis of web-scraped data in 2017 revealed that the respondents answered that the company sold through own websites even when this was actually done via a third-party marketplace or social network web page. Beginning in 2019 the survey questionnaire started explicitly separating marketplaces and social networks as options for the respondent to enable more precise identification of companies having e-commerce of their own.

The results we present in this study are based on data from the ICT Enterprises survey from 2021. Considering the changes made to the questionnaire since 2019, we had hope that, at this time, the web-scraping of websites would enable correct predictions for the answers to the question: Did this company sell products or services in the last 12 months through the company's own website?

Details of the ICT Enterprises survey methodology and data collection can be found at: https://cetic.br/en/pesquisa/empresas/.

Methodology

Database, target population and web-scraping

The target population of the ICT Enterprises survey comprises all Brazilian private companies with 10+ employees with activities shown in Table 1.

The ICT Enterprises survey uses a stratified simple random sampling design. The stratification was designed to enable estimation with controlled precision for some domains:

- Brazilian macro-regions (North, Northeast, Southeast, South and Centre-West).
- Eight activity groups (mentioned above).
- Four size bands in terms of the number of employed workers: 10 to 19 workers, 20 to 49 workers, 50 to 249 workers, and 250+ workers.

The stratification comprises up to 160 strata. The total sample size and its allocation to the strata would enable providing estimates with controlled precision separately for domains defined by macro-region, by activity group, and by size band, but not by cross-classifications of these. As in every survey, ICT Enterprises experiences non-response, when some of the selected companies are not reached or refuse to participate. The weighting approach used to compensate for the observed non-response is post-stratification of basic weights for the design strata.

The frame used for sample selection contained 509,049 enterprises. The selected and responding samples contained 83,182 and 4,064 enterprises respectively. Response rates varied substantially across the strata.

Table 1 – Sections of CNAE 2.0[*] covered by ICT Enterprises Survey

Section Code	Section Description
C	Manufacturing
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
L	Real estate activities
M	Professional, scientific and technical activities
N	Administrative and support service activities
R	Arts, entertainment and recreation
S	Other service activities

[*] CNAE 2.0 is the version 2.0 of the Brazilian National Classification of Economic Activities, defined by IBGE to implement the ISIC Rev. 4, i.e., revision 4 of the International Standard Industrial Classification of all economic activities.

The target population for this project includes only companies that have a website. This website's URL address is obtained during the survey's telephone interview and plays the "respondent" role to the web-scraping procedure. The goal is to accurately predict the answer to the current survey question designed to estimate e-commerce prevalence:

> "Indicator I1 - Did this company sell products or services in the last 12 months through the company's own website? (Yes) (No)."

Among the 4,064 responding companies in 2021, 2,688 companies declared having a website, and out of those, 1,030 declared that they sold through that website. This provided estimates that 58% of companies had a website and 29% sold products or services through that website.

To estimate the proportion of companies that sold on the Internet by the alternative approach, we web-scraped data for the first page of every website that had its address informed via the survey's questionnaire. At this stage 719 websites had to be collected by the project team whenever the URL provided did not reach a valid site. The statistics of data collection for websites are shown in Table 2.

Even after the search done manually by team members using web searching engines, some 317 websites for sampled companies were not located. To ensure that the websites found represent the universe of sites and companies with sites estimated by the survey, a second non-response adjustment was made to obtain the final weights for the companies for which a website was found. The approach used for this adjustment was the same as that previously used to calibrate the basic sampling weights: post-stratification.

Table 1 - Situations encountered during the web-scraping with the corresponding frequencies.

Situation of websites	Frequency	%
Selected	2,688	100
Not found / not scraped	317	12%
Web-scraped	2,371	88%

Source: ICT Enterprises 2021 plus web-scraping data.

The first page of the sites was scraped collecting:

- Texts; and
- Texts that were linked (tagged) to other pages.

The resulting group of texts was further processed to enable analysis. A cleaning operation was performed to keep the radical part of the words found. Prepositions, punctuation, plurals, numbers, and special characters were removed. In the end we had two datasets: one with the radicals of words found in the cleaned texts and one with the links found. Tables 3a and 3b present the most frequent words and links found.

Table 3a – Top 10 words found by the web-scraping with the corresponding frequencies.

Word	Frequency	Number of websites where found	Translation
tod	2,898	1,248	all
servic	2,437	823	service, services
empres	2,290	871	company
produt	2,252	763	product, products
client	2,104	851	client, clients
atend	1,942	878	customer services
melhor	1,691	871	best
qualidad	1,647	789	quality
contat	1,481	828	contact
jur	1,471	56	interest rate

Source: ICT Enterprises 2021 plus web-scraping data.

Table 3b – Top 10 links found by the web-scraping with the corresponding frequencies.

Word	Frequency	Number of websites where found	Translation
contat	2,453	1,269	contact
ver	1,650	282	see
saib	1,622	451	know, to know
conosc	1,507	727	talk to us
hom	1,345	769	home
servic	1,341	574	service, services
produt	1,305	443	product, products
empres	1,147	488	company
compr	1,146	174	buy, to buy
sobr	983	517	about

Source: ICT Enterprises 2021 plus web-scraping data.

Those two databases were merged to the survey database for analysis and modeling. The modeling approach aimed to predict the answers to the target survey question (selling through own website) using the web-scrapped data.

Modeling strategy and Results

To estimate the probability of a website being classified as defined by indicator I1 we adopted a two-step approach:

- considering the number of words/links we found on all the companies websites, we used a statistic to identify candidate variables (indicators to the words or links) to consider as predictors of the answer to the survey question (indicator I1); and
- we then fitted a Logistic Regression model having the answer to the survey question as response and the candidate variables selected in step one as predictors, aiming to evaluate the predictive power of the information gathered via web-scraping.

The statistic used to find candidate predictors was the Information Value (IV). The statistic was calculated using the R package *SmartEDA*.[1] The IV statistic is calculated as:

$$IV = ln\left[\frac{t_{11}/t_{12}}{t_{21}/t_{22}}\right]\times[(p_0-p_1)]$$

Where:

t_{ij} equals the frequency of crosstabulation of Indicator I1 and a candidate predictor (words/links) for row i and column j
p_0 equals the proportion of 0s for indicator I1
p_1 equals the proportion of 1s for indicator I1

Considering this statistic, 9 variables were selected as candidate predictors. They are shown in Table 4 with their respective IVs.

[1] The package and references in> https://cran.r-project.org/web/packages/SmartEDA/index.html.

Table 4: Word radicals used in modeling (links and text)

Word	Frequency	Number of websites where found	Translation	IV
jur	1,471	56	interest rate	0.31
compr	1,146	174	buy, to buy	0.19
privac	694	475	privacy	0.20
descont	314	75	discount, discounts	0.18
cont	228	119	contact, contacts	0.29
carrinh	221	100	checkout	0.32
ped	204	109	product list	0.26
troc	113	88	sale exchange	0.22
devoluco	46	43	sale return	0.21

Source: ICT Enterprises 2021 plus web-scraping data, IV statistic.

The model considered is given by:

$$\Pr(Y_k{=}1|\mathbf{X}_k) \; = \frac{exp(\alpha{+}\boldsymbol{\beta}\mathbf{X}_k)}{1 + exp(\alpha{+}\boldsymbol{\beta}\mathbf{X}_k)}$$

Where:

Y_k is the response given by company k to I1, taking value one if the company sold through its own website, and zero otherwise.

$\mathbf{X}_k$ is the vector of selected predictor variables for company k plus two summation variables – the number of words and the number of links found in each website by the web-scraping.

α and $\boldsymbol{\beta}$ are regression parameters to be estimated for the I1 e-commerce indicator.

Models were fitted using the R survey package (see Lumley, 2010). To evaluate the predictive power, different cutoffs for the predicted probabilities were tested by construction of crosstabulations. For the entire database, consisting of different industries and company sizes, the model did not provide powerful discrimination – 58% overall companies where correctly classified considering a cutoff of 0.339.

Looking in more detail at the websites and companies' answers, we modeled only the Wholesale and retail trade, repair of motor vehicles and motorcycles industry. For this specific activity the overall discrimination improved to around 24%, and corrected classifying 72% of seller's websites and 62% of the non-sellers (cutoff of 0.31575).

Discussion

The goal of quantifying the proportion of enterprises that trade through Internet was not achieved by just web-scraping the first pages of websites. As this pilot has shown, for the universe of the ICT Enterprises survey, considering all industries in the target population, the predictive power of the logistic model was low. On one hand, the diverse characteristics of the industry sectors seems to impact the ability of the model to correctly predict the survey responses. For some industries the trade is a business-to-business (B2B) activity – imposing different challenges when it comes to the web-scaping (for example, customers would need to login before they can access the page where trade is done). On the other hand, for trade and wholesale, the results were better, given that for companies in these sectors, a business-to-customer (B2C) approach is the norm, but still we did not get sufficient predictive power.

Internal discussions raised the possibility of yet some new approaches:

- web-scraping the second level of the pages, the collection of texts and links for the pages linked to the first page. This possibility would generate a lot more text and links – increasing substantially the time and cost for processing and analyzing the data – as information, the web-scraping of the first page got more than 30,000 texts and 15,000 links.
- increase the ICT Enterprises sample for designated industries (B2C) using a questionnaire design to collect more data about the website of the enterprise and its local units separately, collecting more detailed information about the website and its features (a questionnaire rich in information about the website only). With this more detailed information we could understand better the presence of an enterprise on the Internet from the point of view of the human respondent in comparison to the website content (web-scraping).

The first option, to scrap the second level pages of companies' websites, is still in discussion, but the results of this and past pilot studies do not point in the direction that this would result in better outcomes. The option of conducting a bigger study to understand the connections between the human respondent about the website and the web-scraped information is on the table.

As mentioned in the section "History of the project", as this study was meant to enable the production of more timely and disaggregated e-commerce indicators, a study about linking a traditional sample based on the frame of Brazilian enterprises with the websites was conducted in 2021. In this study a probability sample of wholesale and trade enterprises was selected from the Brazilian frame of enterprises (a frame maintained by IBGE). With the information of enterprises' ID, we looked for domain registrations in the Brazilian *.br* register. The success of this linkage would enable the selection of larger samples and the scraping+modeling (if successful) estimation of e-commerce indicator. As the result of this project we found that:

- For a large part of the companies, there are no *.br* registered domains under their ID.
- For some companies the domain is "rented" from another company or even built and administered by some third party company (the domain is linked to a different enterprise ID).
- For some companies the domain is registered under the personal ID of an associate or an associate relative.
- For many companies we were not able to determine if a domain existed or not.

The project concluded that the *.br* domain register covers a different target population than the Brazilian companies frame, and there is no easy way to link these two frames. The only feasible alternative is to ask the enterprises themselves what are their websites, which returns to a process of traditional data collection. On the other hand, if the web-scraping+modeling methodology worked, it would be possible to produce e-commerce indicators (I1) for the universe of *.br* population, based on a probabilistic sample done directly in this frame. This information is not the same produced by the ICT Enterprises survey, but would provide a measure of e-commerce in Brazil.

While our efforts have so far not been successful in providing a feasible alternative to the current method of estimating the prevalence of e-commerce by using web-scrapping, it has highlighted potential difficulties with the current approach of computing such estimates based on the survey data alone. It is possible that the way many respond to the survey does not match the target concept that we are trying to capture via the single survey question. More work would be needed to ascertain whether this is an issue or not.

References

Archer, K. J., S. Lemeshow, and D. W. Hosmer (2007), "Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design", Computational Statistics & Data Analysis, 51(9), pp. 4450–4464.

Lumley, T. (2010), Complex Surveys: A Guide to Analysis Using R, Hoboken: John Wiley & Sons.

NIC.br. (2017), ICT Enterprises Survey on the Use of Information and Communication Technologies in Brazilian Enterprises.

José Márcio Martins Júnior, Marcelo Trindade Pitta, Joao Victor Pacheco Dias and Pedro Luis do Nascimento Silva, Brazilian Network Information Center and National School of Statistical Science from IBGE, Brazil (2018), "Web-scraping as an alternative data source to predict E-Commerce indicators", Proceedings of Statistics Canada Symposium.

Alec Zhixiao Lin, Loan Depot, Using Information Value, Information Gain and Gain Ratio for Detecting Two-way Interaction Effect. SAS Global Forum Proceedings (2018). https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2528-2018.pdf