Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Web Scraping as an alternative data source to predict E-Commerce indicators

Marcelo Trindade Pitta, José Márcio Martins Júnior, João Victor Pacheco Dias, and Pedro Luis do Nascimento Silva[1]

Abstract

Key Words:

1. 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 (CGI.br). As part of its portfolio, NIC.br conducts annual surveys of the companies with 10+ employees operating in Brazil. The ICT Enterprises survey on the use of Information and communication technologies in Brazilian enterprises (ICT Enterprises) aims to measure presence of information and communication technologies (ICT) in the companies with 10+ employees, covering topics such as infrastructure, use and appropriation by the private sector of new technologies, as well as perception of potential benefits of these to their activities.

Some of the questions asked in the ICT Enterprises survey refer to e-commerce infrastructure and practices adopted by the companies (the so-called E module). All e-commerce indicators estimated from the survey to date are based on self-declared responses to this module. See Table 2 for a list of the indicator variables of interest in this study.

Given the increased presence and importance of e-commerce, coupled with increasing survey costs and the emergence of 'big data' approaches to complement or even replace traditional survey sources, the survey team decided to consider an alternative approach for potential estimation of some e-commerce indicators based on direct observation of companies' websites. This should be accomplished by using web-scraping techniques to collect some data directly from companies' websites, without having to rely on survey interviewing of company representatives.

This paper describes the outcome of an experiment of web scraping to obtain information that might be used to estimate several e-commerce indicators using a sample of Brazilian companies with 10+ employees in 2017. Section 2 describes the sample and the web-scraping data collection exercise. Section 3 describes a modelling exercise and corresponding estimation carried out using the combination of survey and web-scraped data, aiming to estimate several e-commerce indicators. Section 4 concludes the paper with an assessment of the outcomes of the experiment, and some suggestions for future work on the topic.

2. Methodology

The target population of the ICT Enterprises survey comprises all Brazilian companies with 10+ employees according to the Central Business Register (Cadastro Central de Empresas - CEMPRE) from the Instituto Brasileiro de Geografia e Estatística (IBGE), classified in activities of interest (see Table 1) and for which the type of business was "private enterprise". The exclusion of public enterprises aims to ensure international comparability and considers that NIC.br carries out an e-government survey, which targets this group of enterprises together with all other government entities.

[1]Marcelo Trindade Pitta, Brazilian Network Information Centre, Brazil; José Márcio Martins Júnior, Brazilian Network Information Centre, Brazil; João Victor Pacheco Dias, Brazilian Network Information Centre, Brazil; Pedro Luis do Nascimento Silva, ENCE

For the year 2017, the target population contained 529.861 companies, from which a sample of 49,246 was selected using stratified simple random inverse sampling. The survey obtained 7,062 complete interviews (carried out using CATI) with companies in the sample, which were initially considered for the web-scraping exercise.

The 11 variables considered for defining the e-commerce indicators of interest are listed in Table 2-1. These variables were collected from the ICT Enterprises responding companies. Estimates of the population proportions of companies having each of these e-commerce practices were published using the survey data. These population proportions would be the 'targets for inference' for the web-scrapping exercise.

Table 2-1
Sections of CNAE 2.0(*) covered by TIC Empresas

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 activites
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 study aimed to assess the possibility of estimating the population proportions of the various e-commerce indicators using the web-scraped data, instead of data collected from the survey respondents. For this purpose, the web-scraped data were combined with those of the traditional ICT Enterprises survey, and statistical models were fitted to assess whether the web-scraped data could successfully predict the company level variables, as well as enable precise estimation of the corresponding population proportions.

The sample for the ICT Enterprises survey is a stratified simple random inverse sample. Stratification of the companies was used to enable estimation with controlled precision for some target domains of interest. The stratification was done in two steps. First, companies were stratified according to the five Brazilian macro-regions (North, Northeast, Southeast, South and Centre-West) cross-classified by eight activity groups (C, F, G, H, I, J, L+M+N, R+S). This step gave rise to 40 strata. Within each of the strata formed in step 1, companies were further stratified in step 2 by size into four size bands: 10 to 19 workers, 20 to 49 workers, 50 to 249 workers, and 250+ workers. If any stratum was empty (i.e. had no companies), the size stratum was collapsed with the size band immediately smaller, preserving the stratification by macro-region and group of activity.

Hence, the stratification adopted should enable providing estimates by macro-region, by activity group, and by size band, separately. The small sample sizes used in some of the cross-classified strata, however, do not permit the regular production of estimates for all the 160 or so domains resulting from the cross-classification of the three stratification variables.

Table 2-2
Variables to be predicted via web scraping

Variable	Indicator Variable Description
Y_1	The company's website provides a catalog of products and services
Y_2	The company's website provides a price list
Y_3	The company's website provides a system for ordering, reserving or a shopping cart
Y_4	The company's website provides on-line payment for completing purchases
Y_5	The company's website provides post-sales suport or customer services
Y_6	The company's website provides institutional information about the company, such as contact and
Y_7	The company's website offers customization or personalization of products or services
Y_8	The company sells products or services via internet via e-mail
Y_9	The company sells products or services via internet via company's website
Y_{10}	The company sells products or services via collective buying sites
Y_{11}	The company sells products or services via internet via social networks

Source: ICT Enterprises 2017 plus *web-scraping* data.

Within each stratum, companies were sampled using simple random inverse sampling (see Vasconcellos et al 2005). This method is like simple random sampling without replacement (SRS) but has a key difference: in SRS, a fixed sample size n is pre-specified, but the effective sample size $m \leq n$ is random (due to non-response and other fieldwork related reasons). In simple random inverse sampling, the number of companies sampled (n) until the target effective sample size (m) is achieved is random, but m is fixed. For example, when the target effective sample size in a stratum is 30, then as many units are sampled sequentially at random from this stratum and approached for the survey as is needed to ensure that 30 complete interviews are obtained in the stratum. Further details about the survey's methodology are available from NIC.br (2018).

Out of the 7,062 companies that responded to the 2017 ICT Enterprises survey, 4,786 indicated having a website. Therefore, the survey estimates that close to 286,000 companies had websites at the time of the survey taking.

For all companies that declared having a website in the ICT Enterprises survey, the corresponding primary website addresses at domain level where collected. Such websites were then accessed using a web-scraping computer program written to obtain the information available on the corresponding main pages. The web-scraping program was written in Java, and no ready-made app was used. The web-scraping collection process followed the three steps indicated below:

- The main page for each website was accessed, with all words in the HTML code for this page being stored, and identification of all the words which had associated links to other pages;
- Text processing was carried out to clean the words database by removal of words such as prepositions, stopping words, and identification of word radicals;
- Manual fine-tuning of the words database to identify word radicals that were not automatically recognised.

During the data collection process, it was impossible to collect data from many pages. Table 2-3 provides a list of situations encountered with the corresponding frequencies.

Table 2-3
List of situations encountered in the web scraping with the corresponding frequencies

Situation	Frequency
Selected websites	4,786
Websites not found	2,026
Websites found	2,760
Websites found and scraped	2,256
Website found and not scraped (various reasons)	504

Source: ICT Enterprises 2017 plus *web-scraping* data.

The 25 most frequent word radicals appearing on the scraped websites were: atend, ativ, client, cont, contat, desenvolv, equip, experienc, marc, merc, oferec, process, produt, profiss, projet, receb, reserv, seguranc, serv, sistem, soluc, tecn, tecnolog, trabalh, vend. The set of word radicals for which there were links which were selected for analysis included: aces, atend, brasil, client, conhec, contat, desenvolv, empr, entr, equip, event, facebook, instituc, notic, parc, poli, port, produt, projet, reserv, serv, soc, som, trabalh, vend. Whenever a company's website contained one of these word radicals, a corresponding value of one was recorded for the corresponding word radical indicator; otherwise, a value of zero was recorded.

3. Model fitting and analysis

Logistic regression models were fitted having each of the e-commerce indicator variables listed in Table 2-2 as the response, and the set of indicators for all the word radicals listed in Section 2 considered as potential predictors. In addition, the three stratification variables (macro-region, activity group and size band) available from the company sampling frame were also included as potential predictors. The goal of such model fitting exercises would be to enable prediction of the various e-commerce indicator variables at company level from the information gathered via the web scrapping of company websites. Such predicted values of the e-commerce indicators could then be used to obtain estimates of the population level proportions of the various e-commerce practices.

The logistic regression models were fitted taking account of the survey design used to obtain the data. As previously mentioned, out of the 4,786 companies indicating that they had a website in the 2017 survey, the web-scraping exercise managed to obtain the required information from 2,256 (47%) websites considered. The non-response arising from the web-scraping exercise was compensated for by multiplying the survey weights of the 2,256 companies considered in the modelling by the corresponding stratum-level inverse response rate.

The model considered for each of the e-commerce indicator variables is given by:

$$\pi(\mathbf{X}_{ij}) = \Pr\left(Y_{ij}=1 \middle| \mathbf{X}_{ij}\right) = \frac{exp\left(\alpha_i+\boldsymbol{\beta}_i\mathbf{X}_{ij}\right)}{1+exp\left(\alpha_i+\boldsymbol{\beta}_i\mathbf{X}_{ij}\right)}$$

where
Y_{ij} is the response given by company j to the ith e-commerce indicator, taking value one if the company had the corresponding infrastructure or practice, and zero otherwise;

$\mathbf{X}_{ij}$ is the vector of predictor variables for company j selected for predicting the ith e-commerce indicator, which includes indicators for the selected word radicals and stratification variables; and

α_i and $\boldsymbol{\beta}_i$ are regression parameters to be estimated for the ith e-commerce indicator, after selection of the relevant predictors.

Models were fitted using the R survey package (see Lumley, 2010). A stepwise procedure was applied for predictor variable selection for each target response, using an option for optimising the cut-off point for estimating the population proportions.

The fitted models were assessed using a global goodness of fit test statistic proposed by Archer, Lemeshow & Hosmer (2007). Table 3-1 presents the values of the goodness of fit statistics for each of the 11 logistic regression models fitted with corresponding p-values. Small p-values signal cases for which the model fit was not good. The results indicate that none of the 11 models fitted well to the data, indicating that the predictive power of the available covariates is not substantial. Nevertheless, the analysis proceeded for all the e-commerce indicators, as described below.

Table 3-1
Goodness of fit statistics for each of the fitted logistic regression models

e-commerce Indicator	F statistic	p-value
Y_1	8765.801	< 2.22e-16
Y_2	2521.543	< 2.22e-16
Y_3	4864.996	< 2.22e-16
Y_4	5962.148	< 2.22e-16
Y_5	7531.174	< 2.22e-16
Y_6	1349.309	< 2.22e-16
Y_7	11833.24	< 2.22e-16
Y_8	6447.785	< 2.22e-16
Y_9	5320.713	< 2.22e-16
Y_{10}	1094.335	< 2.22e-16
Y_{11}	4322.619	< 2.22e-16

Source: ICT Enterprises 2017 plus *web-scraping* data.

Table 3-2 provides information about the estimated confusion matrices. Table 3-3 provides estimates for the population proportions (in %) using the predicted company level e-commerce indicators resulting from the fitted logistic regression models.

Table 3-2
Proportions of correct predictions using the fitted models

	Observed and predicted		
e-commerce indicator	No	Yes	% of correct predictions
Y_1	68%	64%	65%
Y_2	86%	70%	83%
Y_3	85%	68%	82%
Y_4	86%	64%	83%
Y_5	79%	60%	71%
Y_6	78%	79%	79%
Y_7	77%	61%	72%
Y_8	73%	64%	71%
Y_9	80%	69%	77%
Y_{10}	94%	78%	93%
Y_{11}	88%	66%	85%

Source: ICT Enterprises 2017 plus *web-scraping* data.

Table 3-3
Estimates for the population proportions (%) of e-commerce indicators by method

e-commerce indicators	Estimates (%)		
	ICT Enterprises	Companies with scraped websites	Fitted models
The company's website provides a catalog of products and services	74.1%	75.3%	56.4%
The company's website provides a price list	23.3%	22.1%	25.7%
The company's website provides a system for ordering, reserving or a shopping cart	21.0%	18.7%	24.1%
The company's website provides on-line payment for completing purchases	17.6%	16.6%	22.1%
The company's website provides post-sales suport or customer services	42.6%	41.5%	37.2%
The company's website provides institutional information about the company, such as contact and address	96.4%	97.2%	77.8%
The company's website offers customization or personalization of products or services	31.9%	33.2%	35.5%
The company sells products or services via internet via e-mail	21.5%	23.4%	35.5%
The company sells products or services via internet via company's website	20.1%	21.9%	30.9%
The company sells products or services via collective buying sites	7.7%	6.9%	10.5%
The company sells products or services via internet via social networks	14.1%	14.8%	19.9%

Source: ICT Enterprises 2017 plus *web-scraping* data.

4. Conclusions

Analysis of Table 3-3 reveals that the estimates obtained using the predicted e-commerce indicators at company level are not close to those obtained considering the observed e-commerce indicators from the survey questionnaire. This results from insufficient model predictive power, as hinted from the results in Tables 3-1 and 3-2. Considering a sample the same size of that in the current ICT Enterprises, and the proposed web-scraping approach, it would not be recommended to replace collection of the survey module E by web-scrapping and model-based prediction of the e-commerce indicators at company level, for subsequent use in estimating the population level e-commerce proportions. Reasons for the rather large differences between the estimates using the predicted company level e-commerce indicators and the published survey estimates might include the large non-response observed in the web-scraping exercise, as well as the low predictive power of the fitted models. In many cases, the company's websites were simply not found. NIC.br may have access to a database of registered domains for Brazilian companies, which may improve the hit rate for websites selected for web scraping. This is one area for potential improvement of the alternative approach. The other would be to consider alternative models for predicting the company level e-commerce indicators, such as neural networks. A third direction for future work will be to consider using price-comparing websites which cover some activities, and where information about websites for companies practicing e-commerce may be obtained indirectly to support web-scraping.

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*.

Vasconcellos De, M. T. L., P. Silva, P. do Nascimento, and C. L. Szwarcwald (2005), "Sampling design for the World Health Survey in Brazil", *Cadernos de Saúde Pública*, 21, pp. S89–S99.