OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

WELCOME.

OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

ENCE Escola Nacional de Ciências Estatísticas

unesco Centro sob os auspícios da UNESCO

cetic.br Centro Regional de Estudos para o Desenvolvimento da Sociedade da Informação

nic.br Núcleo de Informação e Coordenação do Ponto BR

cgi.br Comitê Gestor da Internet no Brasil

CPS 52
Finance and business statistics VI

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

Marcelo Trindade Pitta
Thiago Meireles
Regional Center for Studies on the Development of the Information Society, Brazilian Network Information Center
Pedro Silva
National School of Statistical Sciences

Tuesday 18 July 4:00PM - 5:25PM

Introduction

- Objective: Estimate selected e-commerce indicators through web scraping of enterprises' websites
- For now, the selected indicators are produced through a bianual traditional ICT Enterprise survey
- The use of web scraping would provide a larger sample, enabling more disaggregated and timely estimates
- The idea: to model the traditional survey responses as functions of the content of web scraped pages

Introduction

BUILDING A PREDICTION MODEL

Web scraping process
Survey Data
Dictionary

PHASE 1

- website1.com.br
- website3.com.br
- website4.com.br
- . . .
- website*n*.com.br

DICTIONARY

Prediction Model

PRODUCING E-COMMERCE INDICATORS

Sampling (Domain Name System Frame) (DNS)
Web data collection
Statistical Model

PHASE 2

DNS DATABASE

Domain Name Database

Prediction Model

Web Scraping Process

Data Analysis & Report

Methodology: target population, survey sample

- The target population of the ICT Enterprises survey comprises all Brazilian private companies with 10+ employees for selected industries

- The sample for the 2021 ICT Enterprises Survey contained 2,688 enterprises that have a website out of 4,064 respondents
  - Of those, 1,030 enterprises declared selling through website
  - Estimates of the survey: 58% have websites and 29% sell through own websites

Methodology: web scraping and modeling

- The web scraping procedure aims to scrap only the home page of respondents' websites
  - The scraping managed to retain texts and links in these pages
  - After the collection of websites, cleaning of the texts (and links) was performed – excluding stopwords, punctuation, stemming, etc.

- Modeling:
  - Logistic regression (sample design considered)
  - Selection of candidate predictors by Information Value (IV) (*SmartEDA R Cran package)*

Some results

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 websites not scraped/found were treated as non response and the weights adjusted through post-stratification

Some results

Table 2a - Top 10 words found by the web-scraping with the corresponding frequencies

Word	Frequency	Number of websites	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.

Some results

- Only 58% of overall enterprises were correctly classified (cutoff threshold found through iterative search – best if diagonal classification is balanced)

- Models were re-fitted separately for Wholesale and Retail trade, Repair of motor vehicles and motorcycles industry ➔ overall discrimination improved to around 24%, and correct classification rose to 72% of seller's websites and 62% of the non-sellers

Discussion and final remarks

- Based on these results web scraping still can not be used to estimate the traditional survey sample indicator

- Diversity of industry sectors seems to impact the ability of the model to correctly predict the survey responses

- Enterprises may trade only with other enterprises (B2B) activity – imposing different challenges when it comes to the web-scaping (login needed before we can access the page for web scraping)

Discussion and final remarks

- For Wholesale and Retail trade, results were better, as business-to-customer (B2C) relations is the norm for most companies, but still we did not get sufficient predictive power.
- 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 needed to ascertain whether this is an issue or not.

OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

THANK YOU.

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