Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Le moissonnage du Web comme source de données de rechange pour prévoir les indicateurs du commerce électronique

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

Résumé

Mots-clés :

1. Introduction

Le Centre d'information du réseau brésilien (*Núcleo de Informação e Coordenação* – NIC.br) est une entité civile à but non lucratif fondée en 2005 pour concrétiser les décisions et les projets du Comité directeur de l'Internet au Brésil (*Comitê Gestor da Internet no Brasil* – CGI.br). Dans le cadre de son programme, le NIC.br effectue des enquêtes annuelles auprès des entreprises de 10 employés et plus qui font affaire au Brésil. L'enquête TIC auprès des entreprises vise à mesurer la présence des technologies de l'information et des communications (TIC) dans les entreprises brésiliennes comptant 10 employés et plus, et couvre des sujets tels que l'infrastructure, l'utilisation et l'appropriation des nouvelles technologies dans le secteur privé, de même que la perception des avantages potentiels de ces technologies pour les activités de ces entreprises.

Certaines des questions de l'enquête TIC auprès des entreprises touchent à l'infrastructure et aux pratiques de commerce électronique adoptées par les entreprises (le « module E »). Tous les indicateurs de commerce électronique ayant été estimés à la lumière des résultats de l'enquête à ce jour sont fondés sur les réponses des répondants à ce module. Au tableau 2 se trouve une liste des variables indicatrices qui présentent un intérêt dans cette étude.

Compte tenu de la présence et de l'importance accrues du commerce électronique, des coûts croissants des enquêtes et de l'émergence d'approches fondées sur les mégadonnées pour compléter, voire remplacer, les sources d'enquêtes traditionnelles, l'équipe de l'enquête a décidé d'envisager une approche différente en vue de l'estimation potentielle de certains indicateurs du commerce électronique, en se basant sur l'observation directe des sites Web des entreprises. Pour ce faire, des techniques de moissonnage du Web devraient être employées pour recueillir certaines données directement sur le site Web de chaque entreprise sans avoir à mener des interviews d'enquête auprès de représentants de l'entreprise.

Le présent document décrit les résultats d'une expérience de moissonnage du Web visant à obtenir des renseignements pouvant servir à l'estimation de plusieurs indicateurs du commerce électronique, à partir d'un échantillon d'entreprises brésiliennes qui avaient 10 employés ou plus en 2017. La section 2 du document décrit l'échantillon, ainsi que l'exercice de collecte de données par moissonnage du Web. La section 3 décrit un exercice de modélisation et l'estimation correspondante effectuée en combinant les données d'enquêtes et les données du moissonnage du Web afin d'estimer plusieurs indicateurs du commerce électronique. La section 4 conclut le document par une évaluation des résultats de l'expérience et par des suggestions de projets liés à ce sujet.

2. Méthodologie

[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

La population cible de l'enquête TIC auprès des entreprises est constituée de toutes les entreprises brésiliennes qui comptent 10 employés ou plus selon le Registre central des entreprises (Cadastro Central de Empresas – CEMPRE) de l'Instituto Brasileiro de Geografia e Estatística (IBGE), qui exercent des activités d'intérêt (voir tableau 1) et dont le type d'entreprise était « entreprise privée ». Les entreprises publiques sont exclues de la population cible en vue d'assurer la comparabilité internationale et parce que le NIC.br effectue déjà une enquête du « gouvernement en ligne », qui vise les entreprises publiques de même que toutes les autres entités gouvernementales.

Pour l'année 2017, la population cible comptait 529 861 entreprises, dont un échantillon de 49 246 a été sélectionné par échantillonnage binomial inverse et aléatoire simple stratifié. L'enquête a obtenu 7 062 interviews complètes (effectuées par ITAO) auprès d'entreprises de l'échantillon qui avaient été considérées à l'origine pour l'exercice de moissonnage du Web.

Les 11 variables prises en considération pour définir les indicateurs de commerce électronique d'intérêt sont indiquées au tableau 2-1. Ces variables ont été obtenues auprès des entreprises ayant répondu à l'enquête TIC auprès des entreprises. Les proportions estimées de la population d'entreprises ayant chacune de ces pratiques de commerce électronique ont été établies et publiées à la lumière des données d'enquête. Ces proportions de la population seraient les « cibles pour l'inférence » dans le cadre de l'exercice de moissonnage du Web.

Tableau 2-1
Sections du CNAE 2.0(*) couvertes par l'enquête TIC auprès des entreprises (*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 activities
R	Arts, entertainment and recreation
S	Other service activities

(*) Le CNAE 2.0 est la version 2.0 de la Classification nationale des activités économiques au Brésil, établie par l'IBGE pour appliquer le CITI, révision 4, c'est-à-dire la révision 4 de la Classification internationale type, par industrie (CITI) de toutes les activités économiques.

Cette étude visait à évaluer la possibilité d'estimer les proportions de la population pour les divers indicateurs du commerce électronique en utilisant des données moissonnées sur le Web plutôt que des données obtenues auprès des répondants à l'enquête. À cette fin, les données moissonnées sur le Web ont été combinées aux données de l'enquête TIC auprès des entreprises traditionnelles, et les modèles statistiques ont été ajustés de manière à déterminer si les données moissonnées sur le Web permettaient de prédire de manière fiable les variables au niveau de l'entreprise et d'estimer avec précision les proportions correspondantes de la population.

L'échantillon de l'enquête TIC auprès des entreprises est sélectionné par échantillonnage binomial inverse et aléatoire simple stratifié. La stratification des entreprises a été utilisée pour permettre une estimation avec une précision contrôlée dans certains domaines d'intérêt cibles. La stratification a été effectuée en deux étapes. Tout d'abord, les entreprises ont été stratifiées selon les cinq macro-régions du Brésil (Nord, Nord-est, Sud-est, Sud et Centre-ouest), avec classification croisée selon huit groupes d'activité (C, F, G, H, I, J, L+M+N, R+S). Cette étape a permis de créer 40 strates. À l'étape 2, les entreprises de chacune des strates formées à l'étape 1 ont été de nouveau stratifiées en quatre catégories selon leur taille : de 10 à 19 travailleurs, de 20 à 49 travailleurs, de 50 à 249 travailleurs, et 250 travailleurs et plus. Si une strate était vide (c'est-à-dire qu'elle ne comptait aucune entreprise), la strate de la taille était combinée à la catégorie de taille située juste en dessous afin de préserver la stratification selon la macro-région et le groupe d'activité.

La stratification adoptée devrait donc permettre la production d'estimations selon la macro-région, selon le groupe d'activité et selon la catégorie de taille, séparément. Cependant, comme certaines des strates à classification croisée ont un échantillon très restreint, il n'est pas possible de produire régulièrement des estimations pour tous les plus ou moins 160 domaines créés par la classification croisée des trois variables de stratification.

Tableau 2-2
Variables à prédire au moyen du moissonnage du Web

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 : l'enquête TIC auprès des entreprises de 2017, plus les données obtenues par *moissonnage du Web.*

À chaque strate, les entreprises ont été sélectionnées par échantillonnage binomial inverse et aléatoire simple (voir Vasconcellos et al 2005). Cette méthode est similaire à l'échantillonnage aléatoire simple sans remise (EAS), avec une différence clé : alors que dans l'EAS, une taille d'échantillon fixe *n* est spécifiée au départ, la méthode employée ici utilise un échantillon dont la taille réelle (*m*) est aléatoire et est égale ou inférieure à *n* (en raison de la non-réponse et autres raisons liées au travail sur le terrain). Selon l'échantillonnage binomial inverse et aléatoire simple, on échantillonne un nombre aléatoire d'entreprises (*n*) jusqu'à atteindre la taille réelle de l'échantillon (*m*), mais *m* est un nombre fixe. Par exemple, si la taille réelle cible de l'échantillon d'une strate est de 30, on échantillonne aléatoirement et de manière séquentielle autant d'unités de cette strate que nécessaire et on communique avec ces unités en vue d'obtenir 30 interviews complètes pour cette strate. De plus amples détails sur la méthodologie de l'enquête sont disponibles auprès du NIC.br (2018).

Sur les 7 062 entreprises ayant répondu à l'enquête TIC auprès des entreprises de 2017, 4 786 ont indiqué avoir un site Web. On estime qu'environ 286 000 entreprises avaient donc un site Web au moment de l'enquête.

Pour toutes les entreprises qui avaient un site Web au moment de l'enquête TIC auprès des entreprises, l'adresse Web correspondante au niveau du domaine a été obtenue. Ces sites Web ont alors été soumis à un programme de moissonnage du Web conçu pour obtenir les renseignements disponibles sur les pages principales correspondantes. Le programme de moissonnage du Web a été codé avec Java; aucune appli toute faite n'a été utilisée. Le processus de collecte par moissonnage du Web a suivi les trois étapes suivantes :

- La page principale de chaque site Web a été ouverte, tous les mots du code HTML de cette page ont été stockés, et tous les mots qui avaient des liens vers d'autres pages ont été détectés.
- Le texte a été traité et la base de données des mots a été nettoyée de manière à supprimer les mots tels que les prépositions et mots vides et à identifier le radical des mots.
- Une manipulation manuelle de la base de données des mots afin d'identifier les radicaux qui n'ont pas été reconnus automatiquement.

Des données contenues dans un bon nombre de pages n'ont pu être recueillies au cours du processus de collecte de données. Le tableau 2-3 donne une liste des situations qui se sont présentées au cours du processus et indique la fréquence de chacune de ces situations.

Tableau 2-3
Liste de situations qui se sont présentées au cours du processus de moissonnage du Web et leur fréquence respective

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

Source : l'enquête TIC auprès des entreprises de 2017, plus les données obtenues par *moissonnage du Web.*
Les 25 radicaux les plus fréquents obtenus des sites Web moissonnés étaient les suivants : 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. Les radicaux pour lesquels il y avait des liens ayant été sélectionnés aux fins d'analyse comprenaient les suivants : 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. Si le site Web d'une entreprise comprenait l'un de ces radicaux, une valeur de 1 était enregistrée pour l'indicateur du radical correspondant; dans le cas contraire, une valeur de 0 était enregistrée.

3. Ajustement et analyse des modèles

Les modèles de régression logistique ont été ajustés de manière à avoir pour réponse chaque variable indicatrice de commerce électronique indiquée au tableau 2-2, et de manière à avoir comme prédicteurs potentiels l'ensemble d'indicateurs de tous les radicaux mentionnés à la section 2. En outre, les trois variables de stratification (macro-région, groupe d'activité et catégorie de taille) du cadre d'échantillonnage des entreprises ont également été incluses comme prédicteurs potentiels. Ces exercices d'ajustement du modèle visent à permettre la prédiction des diverses variables indicatrices de commerce électronique au niveau de l'entreprise à la lumière des renseignements obtenus par moissonnage des sites Web des entreprises. Ces valeurs prédites des indicateurs du commerce électronique seront alors utilisées pour estimer des proportions des diverses pratiques de commerce électronique au niveau de la population.

Les modèles de régression logistique ont été ajustés en tenant compte du plan d'enquête utilisé pour obtenir les données. Comme on l'a mentionné plus tôt, sur 4 786 entreprises qui ont indiqué avoir un site Web dans l'enquête de 2017, l'exercice de moissonnage du Web a permis d'obtenir les renseignements requis pour 2 256 (47 %) des sites Web pris en considération. On a compensé pour la non-réponse de l'exercice de moissonnage du Web en multipliant le poids de sondage des 2 256 entreprises prises en compte dans la modélisation par le taux de réponse inverse correspondant, au niveau de la strate.

Le modèle pris en considération pour chacune des variables indicatrices de commerce électronique est obtenu ainsi :

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

où

Y_{ij} est la réponse donnée par l'entreprise j pour le $i^{ème}$ indicateur de commerce électronique, qui prend la valeur d'un (1) si l'entreprise avait l'infrastructure ou la pratique correspondante, ou de zéro (0) dans le cas contraire;

$\mathbf{X}_{ij}$ est le vecteur des variables prédictives pour l'entreprise j sélectionnées pour prédire le $i^{ème}$ indicateur de commerce électronique, qui comprend des indicateurs pour les variables de stratification et les radicaux sélectionnés;

α_i et $\boldsymbol{\beta}_i$ sont des paramètres de régression à estimer pour le $i^{ème}$ indicateur de commerce électronique, après la sélection des prédicteurs pertinents.

Les modèles ont été ajustés à l'aide du progiciel R pour les enquêtes (voir Lumley, 2010). Une procédure pas à pas a été appliquée pour sélectionner les variables prédictives pour chaque réponse cible, en utilisant une option pour optimiser le point de coupure en vue de l'estimation des proportions de la population.

Les modèles ajustés ont été évalués à l'aide d'une statistique de test de la qualité globale de l'ajustement proposée par Archer, Lemeshow et Hosmer (2007). Le tableau 3-1 présente les valeurs de la statistique sur la qualité de l'ajustement pour chacun des 11 modèles de régression logistique ajustés, avec leur valeur *p* correspondante. Une petite valeur *p* signifie une mauvaise qualité de l'ajustement. Les résultats indiquent qu'aucun des 11 modèles utilisés n'avait un bon ajustement par rapport aux données, ce qui signifie que la puissance prédictive des covariables disponibles n'est pas substantielle. Néanmoins, l'analyse a été effectuée pour tous les indicateurs du commerce électronique, comme on le décrit ci-dessous.

Tableau 3-1
Statistiques sur la qualité de l'ajustement pour chacun des modèles de régression logistique ajustés

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 : l'enquête TIC auprès des entreprises de 2017, plus les données obtenues par *moissonnage du Web.*

Le tableau 3-2 présente des renseignements sur les grilles de correction estimées. Le tableau 3-3 offre des estimations de la proportion de la population (en %), à l'aide des indicateurs du commerce électronique prédits au niveau de l'entreprise à partir des modèles de régression ajustés.

Tableau 3-2
Proportion de prédictions correctes obtenues à l'aide des modèles ajustés

Indicateur du commerce électronique	Observé et prédit		% des prédictions qui se sont avérées correctes
	Non	Oui	
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 : l'enquête TIC auprès des entreprises de 2017, plus les données obtenues par *moissonnage du Web.*

Tableau 3-3
Estimations de la proportion de la population (en %) pour les indicateurs du commerce électronique, selon la méthode

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 : l'enquête TIC auprès des entreprises de 2017, plus les données obtenues par *moissonnage du Web.*

4. Conclusions

L'analyse du tableau 3-3 révèle que les estimations obtenues à l'aide des indicateurs du commerce électronique prédits au niveau de l'entreprise ne sont pas proches des estimations basées sur les indicateurs du commerce électronique observés à la lumière du questionnaire d'enquête. Cela est attribuable à une puissance prédictive insuffisante, comme on peut le voir dans les résultats des tableaux 3-1 et 3-2. Avec un échantillon de la même taille que celui utilisé par l'enquête TIC auprès des entreprises actuelle et avec l'approche de moissonnage du Web proposée, il ne serait pas recommandé de remplacer la collecte du module E de l'enquête par le moissonnage du Web et la prédiction des indicateurs du commerce électronique au niveau de l'entreprise basée sur un modèle aux fins d'estimation des proportions du commerce électronique au niveau de la population.

Ces différences assez importantes entre les estimations basées sur les indicateurs du commerce électronique prédits au niveau de l'entreprise et les estimations de l'enquête ayant été publiées pourraient s'expliquer en partie par le taux élevé de non-réponse observé dans l'exercice de moissonnage du Web et par la faible puissance prédictive des modèles ajustés. Dans de nombreux cas, nous avons tout simplement été incapables de trouver le site Web de l'entreprise. NIC.br a peut-être accès à une base de données des domaines enregistrés par des entreprises brésiliennes, ce qui pourrait accroître la capacité à trouver les sites Web de ces entreprises en vue du moissonnage du Web. Il s'agit d'une manière parmi d'autres dont l'approche alternative pourrait être améliorée. Une autre manière consisterait à envisager d'autres modèles pour prédire les indicateurs du commerce électronique au niveau de l'entreprise, par exemple des réseaux neuronaux. Une troisième direction possible pour les travaux futurs serait le recours à des sites Web de comparaison des prix qui couvrent certaines activités, et dont on peut obtenir des renseignements indirects sur les sites Web des entreprises qui s'adonnent au commerce électronique afin d'appuyer le moissonnage du Web.

Bibliographie

Archer, K. J., S. Lemeshow, et 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* vol. 51, nº 9, p. 4450–4464, 2007.

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, et C. L. Szwarcwald (2005), « Sampling design for the World Health Survey in Brazil », *Cadernos de Saúde Pública* vol. 21, nº S, p. S89–S99, 2005.