TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
ISSN 2029-4913 / eISSN 2029-4921

2018 Volume 24(2): 406–428
doi:10.3846/20294913.2016.1213193

WEB DATA MINING FOR MONITORING BUSINESS EXPORT ORIENTATION

Desamparados BLAZQUEZ, Josep DOMENECH

Department of Economics and Social Sciences, Universitat Politècnica de València,
Camí de Vera s/n, 46022 Valencia, Spain

Received 12 September 2014; accepted 26 January 2016

Abstract. The World Wide Web (WWW) has become the largest repository of information in the world, providing a data stream that grows at the same time as the scope of the Internet does in society. As with most Information and Communication Technologies (ICTs), its digital nature makes it easy for computer programs to analyze it and discover information. This is why it is being increasingly explored as a source of new indicators of technology, economics and development. Web-based indicators can be made available on a real-time basis, unlike delayed official data releases. In this paper, we examine the viability of monitoring firm export orientation from automatically retrieved web variables. Our focus on exports is consistent with the role of internationalization in economic development. To evaluate our approach, we first checked to what extent web variables are capable of predicting firm export orientation. Once these new variables are validated, their automated retrieval is assessed by comparing the predictive performance of two nowcast models: one considering the manually retrieved web variables, the other considering the automatically retrieved ones. Our results evidence that i) web-based variables are good predictors for firm export orientation, and ii) the process of extracting and analyzing such variables can be entirely automated with no significant loss of performance. This way, it is possible to nowcast not only the export orientation of a firm, but also of an economic sector or of a region.

Keywords: automatic indicators, Big Data, corporate websites, export, monitoring, nowcasting, web data mining.

JEL classification: C8, C63, F17, L60.

Introduction

New information is being published daily on the WWW, which has become the largest public source of real-time information in the world. This increased amount of online information is known as “Big Data”, which are transforming the economy and society. This data revolution is called to change, in the near future, the landscape of economic policy and

Corresponding author Desamparados Blazquez
E-mail: *mdeblzso@ade.upv.es*

research (Einav, Levin 2013; Varian 2014) as it is the main driver of the process of social change in the ICTs era. As the scope of web technology in society grows, the data stream increases, which makes people even thirstier for information. This kind of loop process ends up with lots more information posted and updated on the net (Edelman 2012; Einav, Levin 2013). In fact, the WWW has changed the way people and companies interact and communicate. For these reasons, web technology doubtlessly opens up the possibility of improving economic and social policy and research, and has the potential to become the reference for real-time information.

The digital nature of the WWW makes it easy for computer programs to explore and analyze its contents, which enables automatic knowledge discovery and lowers the cost of the information retrieval process (Edelman 2012). Such automatic information extraction opens up the possibility to build up-to-date indicators, which can be useful for a wide range of purposes. This is especially interesting for computing real-time economic indicators without waiting for official data, which are usually released after a long delay. In the particular context of the increasing economic globalization, a topic of much interest about the economy is the engagement in international commerce.

Establishing in foreign markets contributes to the long-term development of firms and economies (Miskinis, Reinbold 2010; Zeng *et al.* 2012). Within the existing alternatives to establish in foreign markets, export is considered the easiest and fastest one. In addition, it represents an attractive and manageable opportunity for firms independently of their size (Nassimbeni 2001; Majocchi *et al.* 2005). Export-oriented companies contribute to increase the competitiveness of an economy, since they become more proactive and adaptable to turbulent environments. For these reasons, exports figure prominently in the minds of policymakers (Girma *et al.* 2004).

To properly design and control export promotion policies, an accurate monitoring system should be implemented. However, current monitoring systems entail some concerns, such as the cost of producing the indicators, over-aggregation of data and the lag between implementing a specific policy and its effect on overseas sales (Wholey, Hatry 1992; Spence 2003). Policy monitoring can be enriched by obtaining firm-level data in real time, which would turn it into a continuous process with a higher level of granularity. This would allow to immediately collect changes in the microeconomic situation to improve their identification and understanding for researchers and policymakers.

In addition to the chance offered by technology to apply Big Data analysis on real-time data, the WWW has the ability to remove a number of geographic constraints and to facilitate instant communication worldwide, thus empowering exports (Dholakia, Kshetri 2004; Vivekanandan, Rajendran 2006). Therefore, corporate websites could reflect the export orientation of firms in different ways, a reflection which would gradually grow as Internet penetration deepens. Confirming this tendency, a previous work demonstrated that adoption of web technology and some web features are good predictors of firms' export orientation (Blazquez, Domenech 2014). Unfortunately, this proposal to obtain an indicator of export orientation with web-based variables relies on a manual retrieval, which renders them inappropriate for designing a real-time monitoring system.

Given the importance of exports to the evolution of an economy, the availability of a new source of prompt information about firm export orientation becomes especially useful. This paper focuses on developing a new monitoring method which relies on automatically obtaining an indicator for the export orientation of firms by analyzing their corporate websites. This way, we can design a model for nowcasting not only the export orientation of a firm, but also of an economic sector or of a region. Nowcasting models exploit the early availability of variables correlated with the target one to obtain an "early estimate" before the official figure becomes available (Choi, Varian 2009; Bánbura *et al.* 2013). These real-time estimates can help policymakers to make informed decisions earlier.

To evaluate our proposal, we build a regression model in a first step with manual web-based variables and compare its predictive performance to a baseline model with firm economic variables. In a second step, after validating the web variables, we check the usefulness of their automatic version by comparing the predictive performance of the manual model to that of an automatic one. Hence this paper has two objectives: to examine the ability of some web-based variables to infer firms' export orientation; and to validate their automatic retrieval so that a nowcast model, which constitutes a real-time monitoring system, can be implemented.

The remainder of the paper is organized as follows. Section 1 reviews some related research on the automatic extraction of web features and nowcasting, linking web activities to the economy and on the website features that are expected to be related to firms' export orientation. Section 2 describes the data used to carry out the performance analysis and shows the results for the baseline model and the manual web-based model. Section 3 explains the construction and validation of the automatic variables, and analyzes the prediction performance of the proposed automatic model. The last section draws some concluding remarks and provides directions for future work.

1. Theoretical background

This section provides background on linking web activities to firms' characteristics and on the automatic extraction of web features. First, we review the related literature; second, we focus on some website features that could provide valuable information on the export orientation of firms; finally, we review some firms' economic characteristics which have been usually related to export behavior.

1.1. Web data mining for science and economic indicators

The digital nature of the WWW makes it easy for computer programs to explore and analyze its contents, thus enabling Big Data techniques and automatic knowledge discovery. In this context, the automatic extraction of web indicators for economic purposes is an incipient research topic, although similar methods have been formerly applied to other purposes such as obtaining indicators for scientific production.

The first approach to systematically use the web as a source of information is the webometrics. These indicators rely on analyzing web page links to compute similar measures to some widespread bibliometrics indicators. The first related work attempted to equate hypertext links with publication cites to generate similar indicators to impact factors (Ingwersen 1998; Smith 1999). The main drawback of this approach is that the large heterogeneity found in the web hinders the reliability of such indicators (Smith 1999; Vaughan, Hysen 2002). However, this heterogeneity was not a limitation when the scientific production of universities or nations was analyzed (Wilkinson *et al.* 2003; Scharnhorst, Wouters 2006; Heimeriks *et al.* 2008). More recent research has successfully focused on economic topics, such as obtaining indicators for the financial situation of banks (Vaughan, Romero-Frias 2010).

Another noteworthy approach to obtain economic indicators from web data is using reports generated by Google Trends (GT). This tool provides up-to-date reports on the volume of web search queries with some specific text. These data can be used to nowcast some economic variables because some specific text querying (e.g., "apply for unemployment benefits") might correlate with some particular aspect of economy (e.g., unemployment). Since they were first introduced as an economic indicator by Choi and Varian (2009), nowcasting models with GT data have been applied to a number of situations, such as proposing indicators for investors' attention (Da *et al.* 2011), tourist arrivals (Bangwayo-Skeete, Skeete 2015), business performance (Vaughan 2014), transaction volumes on the stock market (Preis *et al.* 2010; Moat *et al.* 2014), and well-being (Askitas, Zimmermann 2015). Although GT can supply useful hints on the economic activity at an aggregate level, its ability for characterizing individual firms is limited because it only provides data about what users demand.

Individual firm strategies can be better observed on their corporate websites. In this context, Libaers *et al.* (2010) constructed a taxonomy of technology commercialization models by counting the appearance of some keywords on firms' websites. This analysis was conducted by automating Google queries with each potentially related keyword. The keyword analysis method for tracking firms' strategies has also been used by Youtie *et al.* (2012) and by Arora *et al.* (2013) in the emerging technologies context.

Beyond the keyword analysis, the first attempt to combine different website features to perform a completely automatic analysis of corporate websites to retrieve economic indicators was introduced by Domenech *et al.* (2012). This research work presented an architecture for a web data mining system that manages the download and analysis of corporate websites. The proposed system was applied to find web-based indicators for the size of companies. In this paper, we extend this system to deal with website features related to the export orientation of firms. Section 3 provides more details on the system implementation.

1.2. Export-related indicators built from website features

Web technologies and online platforms have made it possible for companies, independently of their size, to enter new markets and to increase their export sales thanks to the removal of geographical constraints and the instant communication all over the world. In fact, the WWW can at once remove some organizational and resource constraints which exporting presumably entails (Vivekanandan, Rajendran 2006; Sinkovics *et al.* 2013).

At an aggregate level, a number of recent studies revealed that the Internet stimulates trade. For instance, it has been checked that expanding Internet use improves information availability, reduces trade-related costs (informational and transactional, among others) and boosts exports. Moreover, it has been found that an increase in the number of Internet users reduces asymmetric information, increases the business competition level and cuts fixed trade costs, thus contributing to export growth, as verified in the food and manufacturing industries (Clarke, Wallsten 2006; Bojnec, Fertö 2009, 2010).

Focusing on the WWW, the work by Freund and Weinhold (2004) revealed that growth in the number of websites in a country explains its export growth in the following year, since the Internet reduces market-specific fixed costs of trade. In addition, the WWW is useful for increasing firm's visibility and potential customers, and to also improve operational efficiency. This is due to its capacity to make communications and transactions easier and less expensive, which means important efficiency gains (Dholakia, Kshetri 2004; Kažemikaitiene, Bilevičiene 2008; Berthon *et al.* 2012).

Corporate websites have been used in previous works to infer firms' economic characteristics. In line with this, Overbeeke and Snizek (2005) reviewed company websites to find indicators of corporate culture, while Meroño-Cerdan and Soto-Acosta (2007) related web content to firm performance. Similarly, Llopis *et al.* (2010) used corporate website contents to analyze firm strategies. Firm export orientation can also be found on website adoption, as described by Blazquez and Domenech (2014). Therefore at an individual firm's level, a number of website features could be linked to company international strategies.

For all these reasons, we review how different website features could provide valuable information on the export orientation of firms. The objective is to verify whether these features differ between the corporate websites of exporters and non exporters, thus enabling to build a web-based predictive model. To do so, we classified web features in two different groups according to their nature: the "Web presence" group and the "Content-based" group.

Web presence variables

The first group of variables is related to how and when firms implement a corporate website. It includes two variables, namely the domain name age and top-level domain code.

Experienced firms are usually more likely to export as they have had time to increase their knowledge and accumulate useful resources for internationalization (Majocchi *et al.* 2005; Fernández, Nieto 2006). Firms with more experience on the Internet could follow this same pattern towards export.

The domain name is the main identifier of a company on the Internet. The date on which a domain name is registered suggests the approximate date when a company started to go online (Scaglione *et al.* 2009), despite the temporal gap between a domain name being registered and a website being implemented (Murphy *et al.* 2007). Hence, the domain name age is related to the firm's experience in the Internet. As older firms usually own older domains, having an older domain could be indicative of a greater propensity to export.

The top-level domain (TLD), as part of the firm's Internet name, is either an ISO country code (e.g., .es for Spain) or a generic code (e.g., .com). According to Murphy and Scharl (2007), using a country code or a generic one reflects local or global interests, respectively. In addition, it is an important decision in the company's e-branding strategy (Ibeh *et al.* 2005). Thus, its election could be related to the firm's strategic orientation.

Current exporters or companies which intend to start exporting in the near future would prefer to choose any generic domain code to establish its presence on the Internet, as they have a more international profile. Therefore, a generic top-level domain could be positively related to the firm's export orientation.

Content-based variables

This group of variables refers to the contents and functions available in corporate websites. It includes two variables, namely the foreign language version and presence of export-related keywords.

Offering websites in more than one language is usually related to greater marketing effectiveness (Lee, Morrison 2010). Moreover, deploying multilingual websites helps firms to succeed in reaching their target markets and to better deal with clients and suppliers as the cultural language barrier disappears and users feel more confident. In fact offering a multilingual website helps firms gain a competitive advantage in the global market, and enables them to reach a larger number of potential customers (Samiee 2008; Escobar-Rodríguez, Carvajal-Trujillo 2013). Therefore, a website being available in more than one language could be related to the foreign target markets of companies.

Across all languages, English seems the most natural option for exporting firms in non English speaking countries as it is the most widely used language in international businesses.

The WWW is being used as a marketing media by firms. Through their websites, firms can provide information about the markets and countries where they operate and can describe their products and services without limitations. This way, they can easily reach more potential customers throughout the world (Dholakia, Kshetri 2004; Vivekanandan, Rajendran 2006; Berthon *et al.* 2012).

Motiwalla *et al.* (2005) suggest that websites allow companies to gain marketing efficiencies. One fact that this relies on is that website information origination costs are lower than for printed catalogues (Bennett 1997). These characteristics make websites appealing for companies so that they can include as much information about themselves as they consider necessary. In this way, business strategies can emerge on the WWW and they can be monitored by the presence of key terms, as demonstrated by recent research (Youtie *et al.* 2012; Arora *et al.* 2015).

For these reasons, if a firm is selling abroad or intends to reach new markets, it is likely that information about these matters is provided in its corporate website. These activities can be tracked by detecting the presence of certain keywords on websites. Consequently, presence of trade-related keywords on a corporate website could be positively related to the firm's export orientation.

1.3. Structural variables related to export orientation

To assess the prediction performance of the web-based variables, a baseline predictive model was built using the firms' structural variables which have been traditionally related to their export propensity. These included the size, labor productivity and age of the firm.

Firm size

Firm size has been usually related to firm enrollment and performance in international activities. Its effect can differ depending on the industry and other variables considered for prediction, as shown in the literature. On the one hand, some authors emphasize that firm size positively impacts export behavior, as stated by the stage theory of internationalization. Larger firms have more resources, so they are better equipped to deal with the internationalization challenge (Majocchi *et al.* 2005; Fernández, Nieto 2006). On the other hand, a number of studies have revealed that firm size is not a restriction in export performance. In fact, it is argued that firm size influences the firm's decision to enter international markets only when it remains under some specific level (Bonaccorsi 1992; Pla-Barber, Alegre 2007).

Firm labor productivity

The literature shows that exporters are generally more productive than non exporters. It is argued that this could be due to two alternative effects: the "learning-by-exporting" effect and the "self-selection" effect. The first establishes that the higher productivity of exporters comes from the international experience and knowledge that they acquire from their presence in international markets. The latter states that the most productive firms decide to enroll in exporting activities because they are better positioned to succeed and to recover the sunk costs associated with entering foreign markets. As pointed out in the literature, both effects may co-exist (Bernard, Jensen 1995; Girma *et al.* 2004; Andersson *et al.* 2008). Among others, a frequently employed measure for firm performance is labor productivity.

Firm age

The firm's age, which is taken as a proxy to its experience, has been usually considered in the literature as being related to export orientation. However, results between different studies diverge. Some authors have found a positive relationship between the firm's age and its export behavior in both propensity and intensity terms. This can be explained because they have had more time to increase their knowledge, resources and capabilities, which are useful business tools to face the internationalization challenge (Majocchi *et al.* 2005; Fernández, Nieto 2006).

Other studies have concluded that the firm's age is not that related to export behavior or that it has a negative effect, which is in line with the "born-global" phenomenon. This maintains that there are firms which have expanded into foreign markets since they were set up and did not need much time or lots of resources because of the role of innovation and ICTs (Baldauf *et al.* 2000; Andersson *et al.* 2004). The differences between studies could be due to two co-existing effects: the greater solidity and experience of older firms, which imply better conditions for exporting and, at the same time, the more receptive and flexible nature of younger firms, which can make it easy to adapt to the current quick changes in trends and markets.

2. Using web-based variables to infer firm export orientation

2.1. The sample

The sample for this study included 350 manufacturing companies (NACE Rev. 2 codes 10-33) with corporate website established at the Region of Valencia, in east Spain. According to INE (2012), the rate of industrial companies with website in this region is 75.9%, similar to the rest of Spain (75%). The sample was retrieved through a simple random sampling design from the SABI[1] database. As the list of corporate websites provided by SABI was incomplete, the missing website URLs were obtained by querying a search engine with the company's name or its VAT number, given that Spanish regulations make firms include this information in their websites. From each website, the following web-based variables were manually retrieved and coded at the end of 2012:

- Domain name's age (DOM_AGE_i): Continuous variable measured as the number of years since the corporate website domain name was registered. It was computed from the information available in the Internet *whois* service.
- Top-level domain (TLD_i): Dichotomous variable that takes a value of 1 if the TLD of the corporate website was generic.
- English version (EN_i): Dichotomous variable that takes a value of 1 if the corporate website had a functional English version available.
- Export-related keywords ($KEYWORDS_i$): Dichotomous variable with a value of 1 if the website contained any term associated with exportation. A word list[2] containing key terms potentially connected to export orientation was prepared and searched for by querying Google with each term on each website using the advanced search tool.

In order to validate our proposal, this set of web-based variables was supplemented with some economic characteristics of the firms. This information was collected from the companies' financial statements, available in SABI, and the records of exporters of the Spanish Institute for Foreign Trade (ICEX) and the Spanish High Council of Chambers of Commerce. After downloading the websites, we had to wait more than one year to access to the economic information from year 2012, as it is made available with delay. Once our proposal is validated, we will be able to provide frequent estimations about firm export orientation without relying on official sources of data. The following variables were included:

- Size of the firm ($SIZE_i$): Continuous variable measured by the logarithm of the number of employees in the firm.
- Firm's labor productivity (LP_i): Continuous variable measured as the value added per employee.
- Age of the firm (AGE_i): Continuous variable measured as the number of years since the firm was established.
- Firm's industry ($INDUSTRY_i$): Vector of binary variables for two-digit NACE Rev. 2 codes used to control for specific industry effects. It included 14 variables, of which

[1] SABI: Sistema de Análisis de Balances Ibéricos. It is published by Bureau van Dijck. It includes information about 5,000 active manufacturing firms with website in the Region of Valencia.

[2] The terms included in the word list (mostly Spanish) were: Continental; continente; continentes; export; exporta; exportación; exportaciones; exportamos; exportando; exporter; extranjero; globalización; internacional; internacionales; internacionalización; mundial; países. These keywords were selected from our experience after visiting many corporate websites.

13 corresponded to different industry categories with at least 10 companies in the sample. The remaining one gathered all those firms in sectors with fewer than 10 companies in the sample. Ensuring that each variable controls for 10 companies or more allowed us to avoid overfitting.

- Export orientation ($EXPORT_i$): Dichotomous variable that takes a value of 1 if the firm was enrolled in exporting activities. It is the dependent variable in the prediction models.

2.2. Data analysis

First, some descriptive statistics were obtained, as Table 1 shows. Firms with export activities accounted for 48.29% of the sample. It can be observed that the majority of the companies owned a generic domain. It is also remarkable that the mean firm age (20.24 years) was much larger than the mean domain name age (8.60 years). This means that firms have little experience in the Internet and that its adoption is a relatively recent practice, which predictably will continue to expand. In addition, the absence of high correlations (>0.7) among the variables means that there was no high risk of information redundancy and multicollinearity when estimating the prediction models.

Table 2 reflects the sector distribution of the firms in the sample. The metal products, textiles and furniture industries, which are highly representative of the Valencian manufacturing sector (Molina-Morales *et al.* 2011), predominate the sample.

In order to test whether the variables behaved differently depending on the firm's export orientation, statistical techniques of group differences were employed. Normality and homogeneity of variance were checked both graphically and numerically for the continuous variables. As none of the variables fulfilled both assumptions, the nonparametric U Mann-Whitney test was employed. For the case of the binary variables, the Pearson's Chi-squared test was employed (Anderson *et al.* 2014). The results of these analyses are reported in Table 3.

Table 1. Descriptive statistics and correlation matrix[3]

Variable	Mean	SD	1	2	3	4	5	6	7
1. $EXPORT_i$	0.48	0.50							
2. DOM_AGE_i	8.60	4.15	0.36***						
3. TLD_i	0.73	0.44	0.01	0.12**					
4. EN_i	0.39	0.49	0.56***	0.42***	0.06				
5. $KEYWORDS_i$	0.37	0.48	0.38***	0.30***	0.01	0.33***			
6. $SIZE_i^{\dagger}$	20.47	48.15	0.44***	0.40***	-0.15***	0.34***	0.29***		
7. LP_i	35.42	24.67	0.22***	0.18***	-0.10	0.15***	0.09***	0.20***	
8. AGE_i	20.24	10.82	0.33***	0.29***	-0.06	0.19***	0.19***	0.28***	0.17***

Notes: ***($p < 0.01$); **($p < 0.05$). †: The mean for variable $SIZE_i$ is expressed in levels instead of logarithms, as this is more informative on the behavior of the variable. However, all analyses were performed using the variable in logarithms.

[3] Procedures employed: Pearson's r coefficient for pairs of continuous variables; Point-biserial coefficient for pairs of a continuous and a binary variable; and Phi coefficient for pairs of binary variables (Cohen *et al.* 2002).

Table 2. Sector distribution of the firms in the sample

NACE Rev. 2 Codes	N	%
10. Food products	26	7.43
13. Textiles	29	8.29
15. Leather and related products	18	5.14
16. Wood and products of wood and cork, except furniture; articles of straw and plaiting materials	18	5.14
18. Printing and reproduction of recorder media	26	7.43
20. Chemicals and chemical products	21	6.00
22. Rubber and plastic products	27	7.71
25. Fabricated metal products, except machinery and equipment	52	14.86
27. Electrical equipment	12	3.43
28. Machinery and equipment n.e.c	24	6.86
31. Furniture	28	8.00
32. Other manufacturing (jewelry, games and toys, etc.)	13	3.71
33. Repair and installation of machinery and equipment	11	3.14
Various	45	12.86
Total	350	100

Note: "Various" includes the firms under those NACE manufacturing codes with fewer than 10 firms.

Table 3. Results of the comparison made between exporters and non exporters

Variable	Mean $EXPORT_i = 1$	Mean $EXPORT_i = 0$	U Mann-Whitney (Sig.)	Chi-squared (Sig.)
DOM_AGE_i	10.096	7.200	0.000	–
TLD_i	0.740	0.729	–	0.922
EN_i	0.675	0.127	–	0.000
$KEYWORDS_i$	0.562	0.193	–	0.000
$SIZE_i^*$	32.030	9.508	0.000	–
LP_i	40.912	30.200	0.000	–
AGE_i	23.924	16.807	0.000	–

Note: * The mean for variable $SIZE_i$ is expressed in levels instead of logarithms, as this is more informative on the behavior of the variable. However, all analyses were performed using the variable in logarithms.

Within the domain name age, exporters owned significantly older domains than non exporters on average (10.1 years *vs.* 7.2 years). This suggests that a relationship between Internet experience and export behavior exists as exporters started the implementation of corporate websites earlier than non exporters. Furthermore, older firms have the possibility of owning older domains, thus the firm's experience, domain name age and enrollment in exporting activities are connected.

For TLD, no statistically significant differences between exporters and non exporters were found. This finding, though contrary to what was expected, is actually reasonable.

First, although a generic domain is related to e-business, it does not necessarily imply an international profile. Second, the legal and bureaucratic obstacles when registering Spanish domains, which were in force until 2005, probably made them less appealing than generic domains. This could have favored adopting the latter among the majority of firms.

Regarding an English website version, its availability was statistically higher for exporters (67.5%) than for non exporters (12.7%), which indicates that the relation between exports and the most widely used language in international trade is reflected on corporate websites. Presence of keywords on exporters' websites was higher than on the non exporters' ones (56.2% *vs.* 19.3%), and the difference was statistically significant. However, the positive percentage for non exporters websites suggests that some words considered in the analysis may not be appropriate for distinguishing between both groups of firms. An analysis on the separate effect of each export-related keyword was conducted when evaluating the automatic extraction of web features (see Section 3).

For the firm's structural variables, exporters showed higher values for the three variables under study (size, labor productivity and age of firm), and the differences were statistically significant in all cases. Therefore, they can be safely included in a baseline model to check the effectiveness of the web-based predictions.

Overall, the univariate analysis exhibited that exporters have earlier implemented corporate websites on which the availability of an English version and the presence of export-related keywords are also more frequent than for the websites of non exporters. These results bring up the potential of the information extracted from corporate websites for monitoring firms' export behavior. Regarding the structural variables, exporters seem larger, more experienced and more productive than non exporters.

2.3. The predictive models

This section describes the predictive model based on the variables retrieved from corporate websites, and compares its prediction performance against the baseline model based on firms' structural variables. To do this, two logistic regression models were built after identifying which characteristics varied across exporters and non exporters. The estimations of both models were compared to determine the validity of our proposal.

About the statistical methods, logistic regression was applied because it is the most appropriate when a dependent variable is binary, as is the case in this study. The selected variables were those that varied with an admissible level of significance (p < 0.05) between both groups of firms and did not correlate highly (Nassimbeni 2001). According to these criteria, the web-based model was defined as follows:

$$Prob(EXPORT_i = 1) = \frac{e^{Z_i}}{1+e^{Z_i}};$$

$$Z_i = \beta_0 + \beta_1 \cdot DOM\_AGE_i + \beta_2 \cdot EN_i + \beta_3 \cdot KEYWORDS_i + \gamma \cdot INDUSTRY_i, \qquad (1)$$

where β_0 is a constant and the coefficients β_1, β_2, β_3 and γ indicate the relative influence of each feature on the prediction of the category of the dependent variable. Table 4 shows the estimation results, including the estimated regression coefficients and the Standard Error

(SE), p-value and Odds Ratio for these estimations. The Odds Ratio (OR) is a measure of association between the presence of a particular characteristic and the presence of exports, that is, our dependent variable. Thus, an OR greater than 1 indicates that the probability of being an exporter increases with a given independent variable, an OR lower than 1 indicates that this probability decreases, while OR equals 1 when there is no association between the independent and the dependent variable. For binary variables, it can be expressed as follows:

$$e^{\beta Z} = \frac{\dfrac{Prob(Y=1)}{(1-Prob(Y=1))}(Z=1)}{\dfrac{Prob(Y=1)}{(1-Prob(Y=1))}(Z=0)}\,. \tag{2}$$

Results show that the domain name's age effect is positive and statistically significant, thus increasing the probability of exporting. An English version being available on the website is the feature that most contributes to inferring the export orientation, being associated with a high OR. Similarly, presence of export-related keywords is also connected to the export orientation since it significantly raises the probability of exporting. The model performs relatively well, as pointed out by the pseudo-R^2 (0.534), the high prediction accuracy (81.4%) and the Hosmer-Lemeshow test which, in this case, indicates that the model is adequate to explain the data. Table 5 shows the model prediction performance by comparing the firm's actual export orientation to the predictions made by this model.

Table 4. Prediction of export orientation with manually retrieved WWW variables

Variables	β	SE	p-value	OR
DOM_AGE_i	0.068	0.039	0.083	1.070
EN_i	2.186	0.333	0.000	8.901
$KEYWORDS_i$	1.203	0.311	0.000	3.329
(Constant)	–1.717	0.489	0.000	0.180
Pseudo-R^2	0.534			
Hosmer-Lemeshow	0.112			
Prediction accuracy	81.4%			

Notes: The null hypothesis of the Hosmer and Lemeshow test is that the model is fit. The industry dummies have been included in the model specification.

Table 5. Comparison of the model predicting business export orientation from manually retrieved WWW features to the actual export orientation of the firm

Export orientation	*MANUAL* = 0	*MANUAL* = 1
EXPORT = 0	44.6%	11.4%
EXPORT = 1	7.1%	36.9%

The results of this model were compared with the prediction performance of the baseline model, which included the firms' structural variables and was made up as follows:

$$Prob(EXPORT_i = 1) = \frac{e^{W_i}}{1+e^{W_i}};$$

$$W_i = \beta_0 + \beta_1 \cdot SIZE_i + \beta_2 \cdot LP_i + \beta_3 \cdot AGE_i + \gamma \cdot INDUSTRY_i, \tag{3}$$

where β_0 is a constant and the coefficients β_1, β_2, β_3 and γ indicate the relative influence of each feature on the prediction of the category of the dependent variable. The estimations for this model are reported in Table 6. The effect of the three considered variables is positive and statistically significant, thus contributing to the probability of being an exporter. This model also performs relatively well, with a pseudo-R^2 of 0.468, a prediction accuracy of 77.7% and a good data fit according to the Hosmer-Lemeshow test results. When comparing both models, it can be stated that the web-based variables contain as much information about a firm's export orientation as the firm's size, age and labor productivity. Table 7 summarizes the model prediction performance by comparing the actual export orientation of the firm to the predictions made by this model.

Table 6. Prediction of export orientation with the firm's structural variables

Variables	β	SE	p-value	OR
$SIZE_i$	0.847	0.178	0.000	2.333
LP_i	0.016	0.006	0.016	1.016
AGE_i	0.058	0.015	0.000	1.060
(Constant)	–3.556	0.604	0.000	0.029
Pseudo-R^2	0.468			
Hosmer-Lemeshow	0.658			
Prediction accuracy	77.7%			

Notes: The null hypothesis of the Hosmer and Lemeshow test is that the model is fit. The industry dummies have been included in the model specification.

Table 7. Comparison of the model predicting business export orientation from structural variables to the actual export orientation of the firm

Export orientation	*BASELINE* = 0	*BASELINE* = 1
EXPORT = 0	40.3%	11.6%
EXPORT = 1	11.0%	37.1%

3. Automating the retrieval of web-based variables

This section describes the method which was followed to obtain the export orientation indicator built from automatic web-based variables, as well as the evaluation of their performance. To do so, we first describe the implementation of a web data mining tool to automatically obtain information from corporate websites, and second, the statistical techniques applied to construct the automatic web-based variables. Finally, we describe the replication of the manual web-based model with the automatic web-based variables, which was done to check their predictive power.

3.1. Architecture of the web data mining system for analyzing corporate websites

To automatically extract and analyze the contents from the corporate websites, we extended the web mining model presented in Domenech *et al.* (2012) with specific analysis modules. Figure 1 shows the architecture of this system, which consists of three main modules: the *Capture Module*, the *Analysis Module* and the *Production Module*.

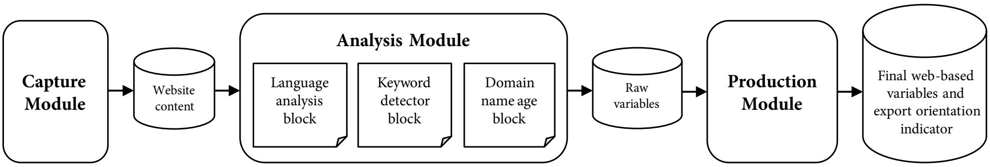

Fig. 1. Model for a web data mining system to retrieve the web-based variables

The *Capture Module* basically acts as a crawler that parses and downloads all the website contents from the corporate sites provided as input. It has been implemented as a modified version of HTTrack (Roche 2014), which is a robot that recursively parses and downloads the links found in the initial URI.

The *Analysis Module* examines the contents downloaded by the *Capture Module* to produce some raw variables that potentially relate to the firm's economic variable under study; i.e., the export orientation in this case. This module is composed of several independent blocks, each one computing related variables.

The *language analysis block* detects the language in which every HTML file on the site is written. Its output is the number of resources in each considered language. The *keyword detector block* departs from a list of keywords and counts the number of occurrences of each keyword in the text of the website. It provides counting not only for strict matching (i.e., exact coincidence), but also for wide matching, that is, derived words are also considered a coincidence. The *domain name age block* makes a request to a *whois* server to find the date on which the provided domain name was registered.

Finally, the *Production Module* takes as input all the raw variables generated by the *Analysis Module* to compute the web-based variables for detecting, in this study case, the export orientation of firms. For this purpose, statistical methods to estimate the probability of exporting given the raw variables were used. More details on these methods are provided below.

3.2. Construction and validation of automatic web-based variables

The web-based model described in Section 2.3 relied on two website features that were manually retrieved (EN_i and $KEYWORDS_i$). This section describes the supervised learning methods applied to estimate the manually retrieved variables from the raw variables generated by the *Analysis Module* of the system. These methods, which are particularly useful with big data, bring up much more realistic prediction performance measures (in terms of obtaining good out-of-sample predictions) than other measures generally used in economics (Varian 2014).

English version

The detection of the foreign language version of the website from the related raw variables (number of HTML documents in each language) relied on the ratio of documents in the foreign language (English) to the number of documents in the local language (Spanish). The rationale behind this is that the English version can be functional, although not all the website's sections are translated.

One of the limitations of our *Capture Module* is its ability to detect duplicate content. This makes that the number of apparently different documents grows uncontrollably with some dynamic websites. This happens, for instance, when two (or more) different terms in an HTML form lead to the same page. To alleviate this problem, a saturation parameter was included. It was defined as the maximum number of files to be considered in each language so that the number of documents saturates at this level.

Both the language ratio and saturation threshold parameters were tuned by a 10-fold cross-validation method. This method assesses in which way the results of a particular statistical analysis would generalize to an independent data set. Basically, this method involves splitting the data sample into a number of complementary subsets, then performing the analysis on one subset (referred to as the *training set*) and validating it on the other subsets (referred to as the *test set*). In this case, 10 partitions were made so that 10 rounds of cross-validation were performed (to reduce variability in the test error estimation). With this method, we were able to choose the values for both the parameters that led to the lowest test error, thus limiting the problem of overfitting.

The results of this method are shown in Figure 2. The saturation threshold varied from 1 to 40 documents, while the language ratio ranged from 0.1 to 2.0. However, for the sake of clarity, the figure shows only a few of these values. The results indicate that the optimal value for the saturation threshold is 11 HTML files, while the optimal value for the language ratio is 0.4. These parameter values were used to compute the estimated *EN* variable.

When compared to the manually retrieved variable, the estimated one (EN_i^A) provides an overall prediction accuracy of 84.3%, as Table 8 shows. A detailed analysis of the classi-

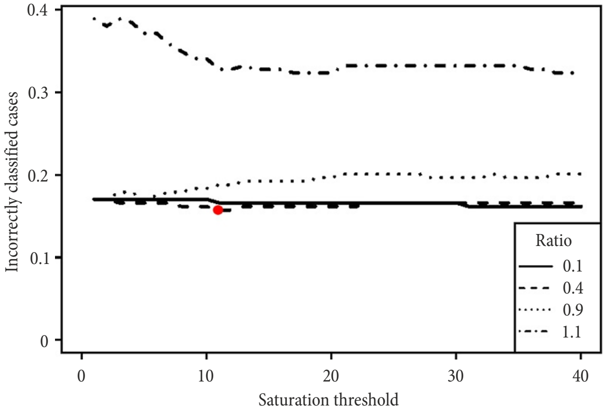

Fig. 2. Cross-validation test error for a range of parameter values of the automatic English version indicator

fication errors reveals that false positives occur when some error messages are found in the HTML text (generated by the web server). False negatives are found to be usually caused by crawling errors (i.e., not all the pages are downloaded). In captured information terms, the moderate correlation found between both variables (0.676) shows that the estimated variable contains similar information to the manually retrieved one.

Table 8. Prediction performance of the automatic English version indicator

English version	*AUTO* = 0	*AUTO* = 1
MANUAL = 0	50.7%	8.3%
MANUAL = 1	7.4%	33.6%

Presence of export-related keywords

The automatic variable for export-related keywords ($KEYWORDS_i^A$) was built from the raw features that included the number of occurrences that apply strict and wide matching algorithms to each word in the list of terms. This list consisted of the same keywords related to business exports as the ones used in the manual model. Though these raw variables are numeric variables that doubtlessly include valuable information, they were transformed into binary in order to replicate the manual variable and thus checking the validity of their automatic extraction, as it is one of our objectives. Since the number of features was large, the *Least Absolute Shrinkage and Selection Operator* (LASSO) method was employed to find a more parsimonious model. The LASSO, which is derived from the *Elastic Net Regression* method, is a statistical method for variable selection which includes a penalty term (shrinkage parameter) and works by producing regressions where some coefficients are set at zero. Hence, problems such as multicollinearity are limited feasibly. The shrinkage parameter (λ) required by this method was tuned by a 10-fold cross-validation procedure, whose results are presented in Figure 3. This procedure resulted in the selection of a logistic regression model with 15 features to be used to estimate the presence of the export-related keywords. That is, 15 binary variables that take a value of 1 when a match with a given word in the resulting list occurs.

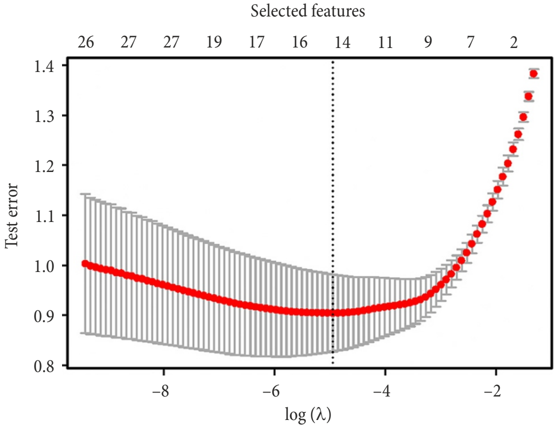

Fig. 3. Cross-validation test error (with 5% confidence intervals) for a range of λ parameter values of the LASSO Method for computing the automatic *KEYWORDS* indicator

The prediction performance of the estimated variable (*KEYWORDS*$_i^A$) is summarized in Table 9. As observed, the proposed method works relatively well since the prediction accuracy is 85.1%. Most misclassifications come from false negatives, which are found on 13.1% of corporate websites. A more detailed analysis reveals that most of these false negatives are due to the incomplete crawling of the site. In captured information terms, the strong correlation found between both variables (0.714) indicates that the estimated variable contains similar information to the manually retrieved one.

Table 9. Prediction performance of the automatic *KEYWORDS* indicator

KEYWORDS	*AUTO* = 0	*AUTO* = 1
MANUAL = 0	52.8%	1.7%
MANUAL = 1	13.1%	32.3%

3.3. Predicting firm export orientation from automatic web-based variables

The automatic export-related variables *EN*$_i^A$ and *KEYWORDS*$_i^A$ computed in the previous section are now employed to finally estimate the business export orientation. To do so, the manually retrieved variables used in the manual web-based model are replaced with the automatically retrieved ones.

The results of the estimation of this model are reported in Table 10. This model attempts to capture most of the prediction accuracy achieved with the manual web-based model. The estimation results of this new model give a prediction accuracy of 78.2%, which is slightly below the 81.4% resulting from the manual web-based model. Taking into account that the model based in manually retrieved web variables acts as an upper bound of the prediction performance of the automatic model, this result means that 96% of the prediction power of the manual model has been successfully reproduced. For each variable effect, in this case the domain name's age and presence of export-related keywords are not statistically significant. The English version variable remains statistically significant, and is associated with an OR of 6.6. This model also performs relatively well, as pointed out by the pseudo-R^2 (0.481), the high prediction accuracy mentioned above and the Hosmer-Lemeshow test, which indicates that the model correctly fits the data.

Table 10. Prediction of export orientation with automatically retrieved WWW variables

Variables	β	SE	p-value	OR
DOM_AGE$_i$	0.050	0.046	0.274	1.052
EN$_i^A$	1.888	0.379	0.000	6.604
KEYWORDS$_i^A$	0.541	0.379	0.152	1.718
(Constant)	–1.721	0.652	0.008	0.179
Pseudo-R^2	0.481			
Hosmer-Lemeshow	0.732			
Prediction accuracy	78.2%			

Notes: The null hypothesis of the Hosmer and Lemeshow test is that the model is fit. The industry dummies have been included in the model specification.

Table 11 shows the model performance by comparing the firm's actual export orientation to the predictions made by this model. A detailed comparison of the results of the manual and automatic web-based models (see Tables 5 and 11) evidences that, as expected, the main difference lies in the false negative rate (7.1% *vs.* 10.0%), which derives from limitations in website crawling. Overall, the comparison shows that the automatic variables are good predictors of firms' export orientation, despite losing some performance if compared to the manually retrieved ones.

Table 11. Comparison of the model predicting business export orientation from automatically retrieved WWW features to the actual export orientation of the firm

Export orientation	*AUTO* = 0	*AUTO* = 1
EXPORT = 0	38.0%	11.8%
EXPORT = 1	10.0%	40.2%

Conclusions

The online data stream increases on a daily basis as people and companies adopt and use the Internet and web technologies. Corporate websites, which are being widely adopted by any kind of firm, reflect the intentions and activities of companies. Following the Big Data paradigm, they can be used as a source of information to produce real-time indicators of the evolution of some economic variables. This is particularly important given that the availability of fresh and frequent data about the economy gives governments more time to react and correct imbalances. As use of web technologies and their economic and social importance are fully expanding, more granular and updated information is available and also demanded at the same time.

This paper has explored the use of Big Data analysis on corporate websites for nowcasting firms' export orientation by automatically producing a web-based indicator. This objective has been accomplished in two steps: first, by finding the corporate website features related to the firms' export orientation; second, by implementing and validating the automatic extraction of these features through a web data mining system.

Our results show that the selected website features contain as much information about the export orientation of companies as the main firm's structural variables (size, age and labor productivity). In contrast to the classic variables obtained from official sources, which are usually made available with long delays, these web features can be retrieved and analyzed in real time. Moreover, our system for automatically analyzing corporate websites achieved 96% of the prediction accuracy of the model with manually retrieved web features, thus validating a new inexpensive and timely source of information about individual firm's export orientation.

From the academic point of view, these web-based variables can complement firms' data from other sources to understand the role played by corporate websites in the internationalization strategy. The results of this study also have implications for policymakers, particularly for the evaluation of export promotion policies. By demonstrating that there are website features from which export indicators can be built, a new way for timely and inexpensive monitoring opens. As their retrieval has been automated, the continuous monitoring of export orientation is now possible. This would allow policymakers to detect how fast companies are reacting to some export promotion policies or what the trend in trade openness is, among others. Furthermore, as website contents are usually related to the designed corporate strategy, it is expected that the decision to export is reflected earlier on the website than in foreign sales, thus anticipating future exports.

There are some limitations of the study that are worth mentioning. First, caution should be taken when generalizing the implications beyond the scope of this study. The results come from only a sample of firms from the Region of Valencia, in east Spain, so they may be specific to this setting, particularly those variables related to language. Further studies using samples from other regions and countries should be carried out. Second, only cross-sectional data are analyzed. A longitudinal analysis would help determine how fast changes in export behavior translate into website changes.

Given the system's ability to retrieve a large number of website features in a short period of time, and the advantages and possibilities offered by web technology, in future works we will explore the relation between other website features and exports, and with other business activities.

Acknowledgements

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness under grant TIN2013-43913-R, and the Spanish Ministry of Education under grant FPU14/02386.

References

Anderson, D. R.; Sweeney, D. J.; Williams, T. A.; Camm, J. D.; Cochran, J. J. 2014. *Statistics for Business & Economics, 12th Edition.* Cengage Learning.

Andersson, S.; Gabrielsson, J.; Wictor, I. 2004. International activities in small firms: examining factors influencing the internationalization and export growth of small firms, *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration* 21(1): 22–34. https://doi.org/10.1111/j.1936-4490.2004.tb00320.x

Andersson, M.; Lööf, H.; Johansson, S. 2008. Productivity and international trade: firm level evidence from a small open economy, *Review of World Economics* 144(4): 774–801. https://doi.org/10.1007/s10290-008-0169-5

Arora, S. K.; Li, Y.; Youtie, J.; Shapira, P. 2015. Using the wayback machine to mine websites in the social sciences: a methodological resource, *Journal of the Association for Information Science and Technology* 67(8): 1904–1915. https://doi.org/10.1002/asi.23503

Arora, S. K.; Youtie, J.; Shapira, P.; Gao, L.; Ma, T. T. 2013. Entry strategies in an emerging technology: a pilot web-based study of graphene firms, *Scientometrics* 95(3): 1189–1207. https://doi.org/10.1007/s11192-013-0950-7

Askitas, N.; Zimmermann, K. F. 2015. Health and well-being in the great recession, *International Journal of Manpower* 36(1): 26–47. https://doi.org/10.1108/IJM-12-2014-0260

Baldauf, A.; Cravens, D. W.; Wagner, U. 2000. Examining determinants of export performance in small open economies, *Journal of World Business* 35(1): 61–79. https://doi.org/10.1016/S1090-9516(99)00034-6

Bánbura, M.; Giannone, D.; Modugno, M.; Reichlin, L. 2013. *Now-casting and the real-time data flow*. European Central Bank Working Paper Series, Vol. 1564.

Bangwayo-Skeete, P. F.; Skeete, R. W. 2015. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach, *Tourism Management* 46: 454–464. https://doi.org/10.1016/j.tourman.2014.07.014

Bennett, R. 1997. Export marketing and the Internet: experiences of Website use and perceptions of export barriers among UK businesses, *International Marketing Review* 14(5): 324–344. https://doi.org/10.1108/02651339710184307

Bernard, A. B.; Jensen, B. J. 1995. Exporters, jobs, and wages in U.S. manufacturing: 1976–1987, *Brookings Papers on Economic Activity: Microeconomics* 1995: 67–119. https://doi.org/10.2307/2534772

Berthon, P. R.; Pitt, L. F.; Plangger, K.; Shapiro, D. 2012. Marketing meets Web 2.0, social media, and creative consumers: implications for international marketing strategy, *Business Horizons* 55(3): 261–271. https://doi.org/10.1016/j.bushor.2012.01.007

Blazquez, D.; Domenech, J. 2014. Inferring export orientation from corporate websites, *Applied Economics Letters* 21(7): 509–512. https://doi.org/10.1080/13504851.2013.872752

Bojnec, Š.; Fertö, I. 2009. Impact of the Internet on manufacturing trade, *Journal of Computer Information Systems* 50(1): 124–132.

Bojnec, Š.; Fertö, I. 2010. Internet and international food industry trade, *Industrial Management and Data Systems* 110(5): 744–761. https://doi.org/10.1108/02635571011044768

Bonaccorsi, A. 1992. On the relationship between firm size and export intensity, *Journal of International Business Studies* 23(4): 605–635. https://doi.org/10.1057/palgrave.jibs.8490280

Choi, H.; Varian, H. R. 2009. *Predicting the present with Google Trends* [online], [cited 20 May 2014]. Available from Internet: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

Clarke, G. R. G.; Wallsten, S. J. 2006. Has the Internet increased trade? Developed and developing country evidence, *Economic Inquiry* 44(3): 465–484. https://doi.org/10.1093/ei/cbj026

Cohen, J.; Cohen, P.; West, S. G.; Aiken, L. S. 2002. *Applied multiple regression/correlation analysis for the behavioral sciences.* 3rd Edition. Routledge.

Da, Z.; Engelberg, J.; Gao, P. 2011. In search of attention, *Journal of Finance* 66(5): 1461–1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

Dholakia, R. R.; Kshetri, N. 2004. Factors impacting the adoption of the Internet among SMEs, *Small Business Economics* 23(4): 311–322. https://doi.org/10.1023/B:SBEJ.0000032036.90353.1f

Domenech, J.; de la Ossa, B.; Pont, A.; Gil, J. A.; Martinez, M.; Rubio, A. 2012. An intelligent system for retrieving economic information from corporate websites, in *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE*, 4–7 December 2012, Macau, China, 573–578. https://doi.org/10.1109/WI-IAT.2012.92

Edelman, B. 2012. Using Internet data for economic research, *Journal of Economic Perspectives* 26(2): 189–206. https://doi.org/10.1257/jep.26.2.189

Einav, L.; Levin, J. D. 2013. The data revolution and economic analysis, *Innovation Policy and the Economy* 14(1): 1–24. https://doi.org/10.1086/674019

Escobar-Rodríguez, T.; Carvajal-Trujillo, E. 2013. An evaluation of Spanish hotel websites: informational vs. relational strategies, *International Journal of Hospitality Management* 33: 228–239. https://doi.org/10.1016/j.ijhm.2012.08.008

Fernández, Z.; Nieto, M. J. 2006. Impact of ownership on the international involvement of SMEs, *Journal of International Business Studies* 37(3): 340–351. https://doi.org/10.1057/palgrave.jibs.8400196

Freund, C. L.; Weinhold, D. 2004. The effect of the Internet on international trade, *Journal of International Economics* 62(1): 171–189. https://doi.org/10.1016/S0022-1996(03)00059-X

Girma, S.; Greenaway, D.; Kneller, R. 2004. Does exporting increase productivity? A microeconometric analysis of matched firms, *Review of International Economics* 12(5): 855– 866. https://doi.org/10.1111/j.1467-9396.2004.00486.x

Heimeriks, G.; van den Besselaar, P.; Frenken, K. 2008. Digital disciplinary differences: an analysis of computer-mediated science and "Mode 2" knowledge production, *Research Policy* 37(9): 1602–1615. https://doi.org/10.1016/j.respol.2008.05.012

Ibeh, K. I. N.; Luo Y.; Dinnie, K. 2005. E-branding strategies of internet companies: some preliminary insights from the UK, *Journal of Brand Management* 12(5): 355–373. https://doi.org/10.1057/palgrave.bm.2540231

Ingwersen, P. 1998. The calculation of web impact factors, *Journal of Documentation* 54(2): 236–243. https://doi.org/10.1108/EUM0000000007167

Instituto Nacional de Estadística (INE) 2012. *Encuesta sobre el uso de TIC y comercio electrónico en las empresas* [online], [cited 16 March 2015]. Available from Internet: http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t09/e02&file=inebase.

Kažemikaitiene, E.; Bilevičiene, T. 2008. Problems of involvement of disabled persons in e. government, *Technological and Economic Development of Economy* 14(2): 184–196. https://doi.org/10.3846/1392-8619.2008.14.184-196

Lee, J. K.; Morrison, A. M. 2010. A comparative study of website performance, *Journal of Hospitality and Tourism Technology* 1(1): 50–67. https://doi.org/10.1108/17579881011023016

Libaers, D.; Hicks, D.; Porter, A. L. 2010. A taxonomy of small firm technology commercialization, *Industrial and Corporate Change* 25(3): 371–405. https://doi.org/10.1093/icc/dtq039

Llopis, J.; Gonzalez, R.; Gasco, J. 2010. Web pages as a tool for a strategic description of the Spanish largest firms, *Information Processing and Management* 46(3): 320–330. https://doi.org/10.1016/j.ipm.2009.06.004

Majocchi, A.; Bacchiocchi, E.; Mayrhofer, U. 2005. Firm size, business experience and export intensity in SMEs: a longitudinal approach to complex relationships, *International Business Review* 14(6): 719–738. https://doi.org/10.1016/j.ibusrev.2005.07.004

Meroño-Cerdan, A. L.; Soto-Acosta, P. 2007. External Web content and its influence on organizational performance, *European Journal of Information Systems* 16(1): 66–80. https://doi.org/10.1057/palgrave.ejis.3000656

Miskinis, A.; Reinbold, B. 2010. Investments of German MNEs into production networks in central European and Baltic states, *Technological and Economic Development of Economy* 16(4): 717–735. https://doi.org/10.3846/tede.2010.44

Moat, H. S.; Curme, C.; Stanley, E. H.; Preis, T. 2014. Anticipating Stock Market Movements with Google and Wikipedia, in D. Matrasulov, H. E. Stanley (Ed.) 2014. *Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale*. Springer, 310 p. https://doi.org/10.1007/978-94-017-8704-8_4

Molina-Morales, X. F.; Martínez-Fernández, T. M.; Torlò, V. J. 2011. The dark side of trust: the benefits, costs and optimal levels of trust for innovation performance, *Long Range Planning* 44(2): 118–133. https://doi.org/10.1016/j.lrp.2011.01.001

Motiwalla, L.; Khan, R. M.; Xu, S. 2005. An intra- and inter-industry analysis of e-business effectiveness, *Information and Management* 42(5): 651–667. https://doi.org/10.1016/j.im.2003.12.001

Murphy, J.; Hashim, N. H.; O'Connor, P. 2007. Take me back: validating the wayback machine, *Journal of Computer-Mediated Communication* 13(1): 60–75. https://doi.org/10.1111/j.1083-6101.2007.00386.x

Murphy, J.; Scharl, A. 2007. An investigation of global versus local online branding, *International Marketing Review* 24(3): 297–312. https://doi.org/10.1108/02651330710755302

Nassimbeni, G. 2001. Technology, innovation capacity, and the export attitude of small manufacturing firms: a logit/tobit model, *Research Policy* 30(2): 245–262. https://doi.org/10.1016/S0048-7333(99)00114-6

Overbeeke, M.; Snizek, W. E. 2005. Websites and corporate culture: a research note, *Business and Society* 44(3): 346–356. https://doi.org/10.1177/0007650305275748

Pla-Barber, J.; Alegre, J. 2007. Analysing the link between export intensity, innovation and firm size in a science-based industry, *International Business Review* 16(3): 275–293. https://doi.org/10.1016/j.ibusrev.2007.02.005

Preis, T.; Reith, D.; Stanley, E. H. 2010. Complex dynamics of our economic life on different scales: insights from search engine query data, *Philosophical Transactions Of The Royal Society A-Mathematical Physical And Engineering Sciences* 368: 5707–5719. https://doi.org/10.1098/rsta.2010.0284

Roche, X. 2014. *HTTrack*. [online], [cited 23 May 2014]. Available from Internet: http://www.httrack.com.

Samiee, S. 2008. Global marketing effectiveness via alliances and electronic commerce in business-to-business markets, *Industrial Marketing Management* 37(1): 3–8. https://doi.org/10.1016/j.indmarman.2007.09.003

Scaglione, M.; Schegg, R.; Murphy, J. 2009. Website adoption and sales performance in Valais' hospitality industry, *Technovation* 29(9): 625–631. https://doi.org/10.1016/j.technovation.2009.05.011

Scharnhorst, A.; Wouters, P. 2006. Web indicators – a new generation of S&T indicators?, *International Journal of Scientometrics, Informetrics and Bibliometrics* 10(1).

Sinkovics, N.; Sinkovics, R. R.; Jean R.-J. "B." 2013. The internet as an alternative path to internationalization?, *International Marketing Review* 30(2): 130–155. https://doi.org/10.1108/02651331311314556

Smith, A. G. 1999. A tale of two web spaces: comparing sites using web impact factors, *Journal of Documentation* 55(5): 577–592.

Spence, M. M. 2003. Evaluating export promotion programmes: U.K. overseas trade missions and export performance, *Small Business Economics* 20(1): 83–103. https://doi.org/10.1023/A:1020200621988

Varian, H. R. 2014. Big data: new tricks for econometrics, *Journal of Economic Perspectives* 28(2): 3–28. https://doi.org/10.1257/jep.28.2.3

Vaughan, L.; Hysen, K. 2002. Relationship between links to journal Web sites and impact factors, *Aslib Proceedings* 54(6): 356–361. https://doi.org/10.1108/00012530210452555

Vaughan, L.; Romero-Frias, E. 2010. Web hyperlink patterns and the financial variables of the global banking industry, *Journal of Information Science* 36(4): 530–541. https://doi.org/10.1177/0165551510373961

Vaughan, L. 2014. Discovering business information from search engine query data, *Online Information Review* 38(4): 562–574. https://doi.org/10.1108/OIR-08-2013-0190

Vivekanandan, K.; Rajendran, R. 2006. Export marketing and the World Wide Web: perceptions of export barriers among tirupur knitwear apparel exporters – an empirical analysis, *Journal of Electronic Commerce Research* 7(1): 27–40.

Wholey, J. S.; Hatry, H. P. 1992. The Case for Performance Monitoring, *Public Administration Review* 52(6): 604–610. https://doi.org/10.2307/977173

Wilkinson, D.; Harries, G.; Thelwall, M.; Price, L. 2003. Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication, *Journal of Information Science* 29(1): 49–56. https://doi.org/10.1177/016555150302900105

Youtie, J.; Hicks, D.; Shapira, P.; Horsley, T. 2012. Pathways from discovery to commercialisation: using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies, *Technology Analysis and Strategic Management* 24(10): 981–995. https://doi.org/10.1080/09537325.2012.724163

Zeng, R.; Zeng, S.; Xie, X.; Tam, C.; Wan, T. 2012. What motivates firms from emerging economies to go internationalization?, *Technological and Economic Development of Economy* 18(2): 280–298. https://doi.org/10.3846/20294913.2012.677588

Desamparados BLAZQUEZ received a BS in Business Administration and Management and a MSc in Data Analytics Engineering from the Universitat Politècnica de València (Spain). She is currently a PhD student and a predoctoral research fellow under the Programme for the Training of University Lecturers (FPU) from the Spanish Ministry of Education. She develops her work at the Department of Economics and Social Sciences of the Universitat Politècnica de València. Her research interests include web economic indicators and internet economics.

Josep DOMENECH received a BS, MSc and PhD in Computer Science and a MSc in Multimedia Applications from the Universitat Politècnica de València (Spain), and a BS and MSc in Business Administration and Economics from the Universitat de València (Spain). Since 2009 he is an Associate Professor at the Department of Economics and Social Sciences of the Universitat Politècnica de València. His research interests include web economic indicators, internet economics and web performance characterization.