CARMA 2016

1st International Conference on Advanced Research Methods and Analytics

July 6 – 7, 2016 · Valencia, Spain

Congress UPV
Proceedings of the 1st international Conference on Advanced Research Methods and Analytics, CARMA2016.

The contents of this publication have been evaluated by the Scientific Committee which it relates and the procedure set out http://www.carmaconf.org/

Scientific Editors

Josep Domenech
Alicia Mas-Tur
Norat Roig-Tierno
Maria Rosalia Vicente

www.lalibreria.upv.es / Ref.: 6287_01_01_01

ISBN: 978-84-9048-462-3 (print version)

Print on-demand

DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4355

Editorial Universitat Politècnica de València http://ocs.editorial.upv.es/index.php/CARMA/CARMA2016

First International Conference on Advanced Research Methods and Analytics

Preface

This volume contains the selected papers of the First International Conference on Advanced Research Methods and Analytics (CARMA 2016), which was held in Valencia, Spain, during July 6th and 7th of 2016. Research methods in economics and business are evolving with the increasing availability of comprehensive sources of data. As these methods are becoming more interdisciplinary, CARMA 2016 provided researchers and practitioners with a forum to exchange ideas and advances on how emerging research methods are applied to different fields of social sciences as well as to discuss current and future challenges.

The scientific program was divided into three tracks. Track 1 focused on **Web and Big Data in Social Sciences** and included 23 papers dealing with topics such as Big Data in official statistics, Internet econometrics, geospatial and mobile phone data, and public opinion mining. All received papers were peer-reviewed by two or three members of the scientific committee, leaded by Dr. Josep Domenech and Dr. María Rosalía Vicente. The track also featured a special session on "Big data and nowcasting macroeconomic indicators" organized by Eurostat.

Track 2 and Track 3 dealt with **Qualitative and Comparative Methods** and **Advanced Regression Methods**. These sessions, comprising 10 papers, stand out on social innovation, technology transfer, entrepreneurial activities or business research. As with Track 1, two or three members of the scientific committee led by Dr. Alicia Mas-Tur and Dr. Norat Roig-Tierno performed peer-reviews of every manuscript.

CARMA 2016 also featured two keynote speakers that overviewed important and current topics. The opening keynote speech was delivered by Dr. Antonino Virgillito, head of the "Business Intelligence, Mobile and Big Data Architecture" unit at the Italian National Statistical Institute (Istat). His talk highlighted the challenges, experiences and future steps of Big Data in official statistics. The closing keynote speech was given by Prof. Sascha Kraus, Full Professor and Chairholder in Strategic Management and Entrepreneurship at the University of Liechtenstein. His talk dealt with the usage of fsQCA in innovation and entrepreneurship research.

The conference was hosted by the Faculty of Business Administration and Management of the Universitat Politècnica de València, which has been recently ranked as the best technical university in Spain by the Academic Ranking of World Universities (ARWU)

2015. Valencia is a city of culture and heritage. It is the third largest city in Spain and its location by the Mediterranean Sea provides their citizens and visitors with a privileged weather.

The organizing committee would like to thank all who made the first edition of CARMA a great success. Specifically, thanks are indebted to the invited speakers, authors, scientific committee members, reviewers, session chairs, presenters, sponsors, supporters and all the attendees. Our final words of gratitude must go to the Faculty of Business Administration and Management of the Universitat Politècnica de València for supporting CARMA 2016.

Josep Domenech
Alicia Mas-Tur
Norat Roig-Tierno
Maria Rosalia Vicente

Conference Committee

Conference Co-Chairs

Josep Domenech, Universitat Politècnica de València
Alicia Mas-Tur, Universitat de València
Norat Roig-Tierno, ESIC Business & Marketing School

Conference Secretary

Desamparados Blazquez, Universitat Politècnica de València

Sponsors

BigML
DevStat

Supporters

Universitat Politècnica de València
Facultad de Administración y Dirección de Empresas
Departamento de Economía y Ciencias Sociales

Scientific committee

Track 1: Web & Big Data in Social Sciences

Track chairs: Josep Domenech, Universitat Politècnica de València
María Rosalía Vicente, Universidad de Oviedo

Isidro Aguillo, Consejo Superior de Investigaciones Científicas (CSIC)
Concha Artola, Banco de España
Nikolaos Askitas, Institute for the Study of Labor (IZA)
Giulio Barcaroli, Italian National Institute of Statistics
Marc Bogdanowicz, Institute for Prospective Technological Studies (JRC-IPTS)
Pablo de Pedraza, Universidad de Salamanca / University of Amsterdam
Giuditta de Prato, Institute for Prospective Technological Studies (JRC-IPTS)
Benjamin Edelman, Harvard Business School
Daniel Gayo-Avello, Universidad de Oviedo
José A. Gil, Universitat Politècnica de València
Felix Krupar, Alexander von Humboldt Institute for Internet and Society
Rocío Martínez Torres, Universidad de Sevilla
Suzy Moat, University of Warwick
Esteban Moro, Universidad Autónoma de Madrid / Universidad Carlos III

Maurizio Naldi, Università di Roma Tor Vergata
Michaela Nardo, Joint Research Centre (JRC)
Bülent Özel, Universitat Jaume I
Ana Pont, Universitat Politècnica de València
Tobias Preis, University of Warwick
Ravichandra Rao, Indian Statistical Institute
Fernando Reis, Eurostat
Andrea Scharnhorst, Royal Netherlands Academy of Arts and Sciences
Vincenzo Spiezia, OECD
Pål Sundsøy, Telenor
Sergio L. Toral Marín, Universidad de Sevilla
Michela Vecchi, Middlesex University

Track 2: Qualitative and Comparative Methods
Track chair: Norat Roig-Tierno, ESIC Business & Marketing School

Adrian Dusa, University of Bucharest
Ian Jenson, University of Tasmania
Kati Kasper-Brauer, Freiberg University of Technology
Alexander Leischnig, Bamberg University
Jordi Paniagua, Universidad Católica de Valencia
Domingo Riberio-Soriano, Universitat de València
Carsten Q. Schneider, Central European University

Track 3: Advanced Regression Methods
Track chair: Alicia Mas-Tur, Universitat de València

Hervé Abdi, The University of Texas at Dallas
José Antonio Belso, Universidad Miguel Hernández
Jasmina Berbegal, International University of Catalunya
Ricarda B. Bouncken, University of Bayreuth
Gabriel Cepeda, Universidad de Sevilla
José Manuel Guaita, Valencian International University
Alenka Janko Spreizer, University of Primorska
José M. Merigó, University of Chile
Marcin W. Staniewski, University of Finance and Management in Warsaw

Local Organization

Andrea Burgos Mascarell
Sergi Doménech de Soria
Alejandro Gil Ródenas
Andrea Rey-Martí

Web & Big Data in Social Sciences

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4243

Some guidance for the use of Big Data in macroeconomic nowcasting

Mazzi, Gian Luigi[a1]
[a]Eurostat, European Commission, Luxembourg.

Abstract

This paper develops an operational step by step approach aiming to facilitate the use of Big Data in nowcasting exercises. Each step includes a description of the problem and a set of recommendations addressing the most relevant available solution. The approach includes nine steps starting from the theoretical availability of Big Data until the publication of new nowcasting including also Big Data. In designing this operational step by step approach, the preliminary results of an ongoing Eurostat project on Big Data and macroeconomic nowcasting have been used as a starting point. Further elaboration has been carried out in order to make the operational step by step approach more concrete and prescriptive. Its aim is to provide a concrete help for experts involved in the construction of nowcasting especially in the judgment about the usefulness of the presence of Big Data in their models. It also provides guidance related to the dissemination of new nowcasting based also on Big Data.

Keywords: *Big Data; Nowcasting.*

[1] The information and views set out in this paper are those of the author and do not necessarily reflect the official opinion of the European Commission.

1. Introduction

The availability of Big Data is opening new challenging ways of producing statistics. Big Data can be particularly relevant to increase the timeliness of macroeconomic indicators by means of new types of nowcasting. At present, Big Data should be viewed realistically more as a complement of traditional information to produce nowcasting instead of an alternative. The presence of Big Data can substantially change the traditional ways of building up nowcasting. This paper, also based on the preliminary results of an ongoing project on Big Data and macroeconomic nowcasting, is proposing a step by step approach for the utilization of Big Data in a nowcasting exercise.

2. A step by step approach for the use of Big Data in a nowcasting exercise

2.1. The step by step approach

In this section we are proposing some guidance, expressed in a form of a step by step approach, for using Big Data when building up macroeconomic nowcasting. Each step will be accompanied by a detailed description as well as one or more recommendations, suggested to the compilers of macroeconomic nowcasting. Table 1 summarises the various steps together with their aim while, in the following subsection, each step will be further detailed.

Table 1.

Steps	**Title**	**Aim**
Step 0	Big Data usefulness within a nowcasting exercise	Checking for the existence of adequate Big Data sources
Step 1	Big Data search	Identification of the appropriate Big Data
Step 2	Availability and quality	Verification of Big Data availability and its quality
Step 3	Accounting for Big Data specific features	Move from unstructured Big Data to a structured dataset
Step 4	Big Data pre-treatment	Removing deterministic and periodic undesirable effects
Step 5	Presence of bias	Checking Big Data for the presence of bias

Step 6	Big Data modelling	Identifying the best modelling strategy
Step 7	Results evaluation of Big Data based nowcasting	Checking for the effective contribution of Big Data
Step 8	Implementing Big Data based nowcasting	Timing and scheduling for the new nowcasting

2.2. Big Data usefulness within a nowcasting exercise

2.2.1. Description

This first step should investigate the potential usefulness of Big Data for a specific indicator of interest, such as GDP growth, inflation or unemployment rate. Big Data sources should be considered for their ability of improving the overall quality of existing nowcasting or of producing timelier estimates. The theoretical soundness of the relationships between existing Big Data sources and the target variables should also be investigated.

2.2.2. Recommendations

- Suggest the use of Big Data only when there are well founded expectations of their usefulness either for fixing problems in existing nowcasting or to improve the timeliness.
- Do not consider Big Data sources with doubtful or even spurious correlations with the target variable.

2.3. Big Data search

2.3.1. Description

Once Big Data passes the "need check" in the previous step, the next action of the Big Data based nowcasting exercise is a careful search for the specific Big Data to be collected. There are many potential providers such as social networks, traditional business systems, the Internet of things, etc. It is very difficult to give general guidelines on a preferred data source because the choice is heavily dependent on the target indicator of the nowcasting exercise.

2.3.2. Recommendations

- Searching in the wider possible set of Big Data having clearly in mind the specificities and the characteristics of the target variable as well as what we want to nowcast.
- Checking for the adherence of available Big Data to what the target variable is really measuring.

2.4. Availability and quality

2.4.1. Description

Having identified the preferred source of Big Data, the second step requires assessing the availability and quality of the data. A relevant issue is whether direct data collection is needed, which can be very costly, or a provider makes the data available. In case a provider is available, its reliability and cost should be assessed, together with the availability of meta data, the likelihood that continuity of data provision is guaranteed, and the possibility of customization (e.g., make the data available at higher frequency, with a particular disaggregation, for a longer sample, etc.). All these aspects are particularly relevant in the context of applications in official statistical offices. As the specific goal is nowcasting, it should be also carefully checked that the temporal dimension of the Big Data is long and homogeneous enough to allow for proper model estimation and evaluation of the resulting nowcasts.

2.4.2. Recommendations

- Privileging data providers which are able to give sufficient guarantee of the continuity of the data process and of the availability of a good and regularly updated metadata associated to the Big Data
- Privileging Big Data sources which ensure sufficient time coverage to properly building up a nowcasting exercise.

2.5. Accounting for Big Data specific features

2.5.1. Description

The third step analyzes specific features of the collected Big Data. A first issue concerns the amount of the required storage space and the associated need of specific hardware and software for storing and handling the Big Data. The second issue is the type of the Big Data, as it is often unstructured and may require a transformation into cross-sectional or time series observations.

2.5.2. Recommendations

- Creating a Big Data specific IT environment where the original data are collected and stored with associated routines to automatically convert them into structured, either cross-sectional or time-series datasets.
- Ensure the availability of an exhaustive documentation of the Big Data conversion process.

2.6. Big Data pre-treatment

2.6.1. Description

Even when already available in numerical format or after their transformation into numerical form as in the previous step, pre-treatment of the Big Data is often needed to remove deterministic patterns such as outliers and calendar effects and deal with data irregularities, like missing observations. Furthermore, seasonal and non-seasonal short-term movements (i.e. infra-monthly ones) should be removed accordingly to the characteristic of the target variable. Since not all seasonal and calendar adjustment methods can be applied when data are available at high frequency, appropriate adjustment techniques need to be identified when the data are available at high frequency. The size of the datasets suggests resorting to robust and computationally simple univariate approaches.

2.6.2. Recommendations

- Whenever possible, all the data treatment described in this step should be done within a unique framework in order to avoid inconsistencies between different parts of the process.
- The filtering of Big Data should be consistent to the one used for the target variables: for example if the target variable is not seasonally adjusted, there is no reason to remove the seasonal component from Big Data and vice-versa.

2.7. Presence of bias

2.7.1. Description

This step requires assessing the presence of a possible bias in the answers provided by the Big Data, due to the so-called “digital divide” or the tendency of individuals and businesses not to report truthfully their experiences, assessments and opinions. Another relevant and partially related problem, particularly relevant for nowcasting, is the possible instability of the relationship with the target variable. This is a common problem also with standard indicators and traditional nowcasting exercises. Both issues can be however tackled at the modelling and evaluation stages.

2.7.2. Recommendations

- If a bias in the Big Data answers is observed, provided that it has been reasonably stable in the last few years, a bias correction can be included in the nowcasting strategy.
- If a bias in the Big Data answers is very unstable, then the Big Data should be considered not reliable enough to be used in a nowcasting exercise.
- In order to deal with a possible instability of the relationships between the Big Data and the target variables, nowcasting models should be re-specified on a regular basis (e.g. yearly) and occasionally in presence of unexpected events.

2.8. Big Data modelling

2.8.1. Description

This step requires the identification of the most appropriate econometric technique when building up a nowcasting exercise with Big Data. It is important to be systematic about the correspondence between the nature and the size of the selected Big Data and the method that is used. There are a number of dimensions along which we wish to differentiate.

In the first one we address the choice between the use of methods suited for large but not huge datasets, and therefore applied to summaries of the Big Data (such as Google Trends, commonly used in nowcasting applications), or of techniques specifically designed for Big Data. For example, nowcasting with large datasets can be based on factor models, large BVARs, or shrinkage regressions.

Huge datasets can be handled by sparse principal components, linear models combined with heuristic optimization, or a variety of machine learning methods, such as LASSO and LARS regression which, though, are generally developed assuming i.i.d. variables. It is difficult to provide an a priori ranking of all these techniques and there are few empirical comparisons and even fewer in a nowcasting context, so that it may be appropriate to apply and compare a few of them for nowcasting the specific indicator of interest. In absence of a multifrequency problem, those techniques can work for variable selection or data reduction as well as for the estimation of the nowcasting of the target variable.

In the second dimension we address the problem of the frequency of the available data. If this frequency is mixed, then specific techniques for mixed frequency data become relevant after having selected the variables or having reduced the dimension of the variables space accordingly with the techniques discussed above. Among mixed frequency models, UMIDAS stands out but also Bridge models can deserve a certain attention. UMIDAS provides a very flexible framework of analysis and can be adapted to work together with most if not all Big Data methods be they machine learning or econometric.

2.8.2. Recommendations

- In absence of any a priori information on the relative performance of various techniques, as many methods as possible should be evaluated and compared in a nowcasting context in order to select the best performing one.
- Alternative modelling strategies should be compared also by looking at the balance between their complexity in computational terms and their empirical performance.
- In case of mixed frequency data, linear methods such as UMIDAS and, as a second best, Bridge, should be privileged.
- Forecast combination and model averaging techniques, also when the mixed frequency aspect is present, can be used as an alternative to a large-scale comparison among competing techniques.

2.9. Results evaluation of Big Data based nowcasting

2.9.1. Description

The final step consists of a critical and comprehensive assessment of the contribution of Big Data for nowcasting the indicator of interest. This should be carried out within a real-time or a pseudo-real time exercise. In order to avoid, or at least reduce the extent of, data and model snooping, a cross-validation approach should be followed, whereby various models and indicators, with and without Big Data, are estimated over a first sample and they are selected and/or pooled according to their performance, but then the performance of the preferred approaches is re-evaluated over a second sample.

This procedure provides a reliable assessment of the gains in terms of enhanced nowcasting performance from the use of Big Data. For some critics about the usefulness of Big Data see Hartford (2014) and Lazer et al. (2014).

2.9.2. Recommendations

- Conducting an in-depth real-time or pseudo real-time simulation of competing models in order to evaluate their relative performance in nowcasting the variable of interest.
- Models including Big Data should be preferred when they significantly lead to an improvement of the reliability and accuracy of the nowcasting at the same point in time.
- Models including Big Data should also be preferred when they allow for timelier nowcasting without any significant loss in terms of reliability and accuracy.

2.10. Implementing Big Data based nowcasting

2.10.1. Description

In case the in-depth comparative analysis carried out in the previous steps suggests that the use of Big Data can improve the nowcasting for a given variable of interest, they can be then implemented. At this stage, the institution in charge of producing nowcasting should take several relevant decisions related to the number of the nowcasting to be implemented and their scheduling. For example, it is possible to decide to publish just one nowcasting (e.g. at the very end of the reference period or at the very beginning of the following one), to produce two nowcastings (e.g. on in the middle of the reference period and one at the very end), or to produce a sequence of nowcasts scheduled at weekly or even daily frequency. Such decisions should take into account, among other, the trade-off between timeliness and reliability, the user needs as well as some more institutional considerations.

2.10.2. Recommendations

- Implementing and publishing the most reliable nowcasts available either at the end of the reference period or at the beginning of the following one.
- Moving towards a daily or weekly update on nowcasting already during the reference period, only after detailed pros and cons analysis and a consultation of the most relevant stakeholders.
- The new Big Data based nowcasting should be accompanied by clear metadata and widely available reference and methodological papers.

3. Conclusions

In this paper, we have proposed a new operational step by step approach for using Big Data in a nowcasting exercise. It aims to facilitate the activity of experts involved in the construction of nowcasting by providing a set of recommendations associated to various operational steps.

References

Hartford, T. (2014, April). Big data: Are we making a big mistake? Financial Times. Retrieved from http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#ixzz2xcdlP1zZ.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science, 143, 1203-1205.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4226

Nowcasting with Google Trends, the more is not always the better

Combes, Stéphanie[a] and Bortoli, Clément[b]
[a]Department of statistical methods, INSEE, France, [b]Department of conjoncture, INSEE, France.

Abstract

National accounts and macroeconomic indicators are usually published with a consequent delay. However, for decision makers, it is crucial to have the most up-to-date information about the current national economic situation. This motivates the recourse to statistical modeling to "predict the present", which is referred to as "nowcasting". Mostly, models incorporate variables from qualitative business tendency surveys available within a month, but forecasters have been looking for alternative sources of data over the last few years. Among them, searches carried out by users on research engines on the Internet – especially Google Trends – have been considered in several economic studies. Most of these exhibit an improvement of the forecasts when including one Google Trends series in an autoregressive model. But one may expect that the quantity and diversity of searches convey far more useful and hidden information. To test this hypothesis, we confronted different modeling techniques, traditionally used in the context of many variables compared to the number of observations, to forecast two French macroeconomic variables. Despite the automatic selection of many Google Trends, it appears that forecasts' accuracy is not significantly improved with these approaches.

Keywords: *nowcasting; Google Trends; macroeconomics; high dimension; machine learning; time series.*

1. Introduction

Official statistics are often published within irreducible delays[1]. But for decision makers, it is crucial to have access to the most up-to-date information about the current national economic situation. This is why developing some efficient forecasting tools and identifying the most relevant data sources are a serious issue in macroeconomics. Real time forecasting (or *nowcasting*) of macroeconomic indicators usually implies to incorporate variables from qualitative business surveys or sometimes financial variables. Over the last few years, forecasters have also been looking into data from the Internet and at trending searches made by Google users in particular.

In 2006, Google launched *Google Trends*, a tool that provides data series free of charge which reflect the interest of Internet users in a query or a set of semantically linked search terms. If this application has been popularized by advertising the most popular searches of the moment, it has also become a well-known source of data for economic studies. Characterized by their high frequency compared to official indicators and their short delay of publication (a week), they have been investigated by numerous economists over the last few years. Indeed, the global evolution of queries made by users about particular products or subjects via the search engine is likely to reflect the potential volume of sales of these products or the predominance of the subject for individuals at the time. These data could therefore be considered as indicators of consumer purchase intention or concerns (for example queries about unemployment benefit may give a hint of the evolution of the unemployment rate). Plus, the soaring penetration rate of equipment of households in computers and Internet connection makes them a credible source of information on individuals (less likely on companies).

The most famous use case of prediction with Google Trends is the Google Flu application developed by Google to forecast the spread of the flu epidemic in real time, based on user queries, in 2008. First launched in the United States, the tool was extended the following year to a dozen European countries, including France. In 2009, the group published an analysis of the benefits of using these series to forecast socio-economic indicators (Choi and Varian, 2009). According to this study, which used American data, forecasting automobile purchases, retail sales and purchases of dwellings could be improved by introducing this type of series into simple models using the dynamics of the series of interest (autoregressive model).

[1] In France, the main quantitative data available on household expenditure for example is the monthly household consumption expenditure on goods, published within one month and its equivalent in services is published within two months. Finally, an initial estimate of quarterly spending on all goods and services is published in the middle of the following quarter.

Askitas and Zimmermann (2009) used the frequency of use of certain search terms to forecast the unemployment rate in Germany; Kulkarni et al. (2009) suggested a link between the frequency of several search terms and housing prices in the United States; Vosen and Schmidt (2013) also used this type of series to forecast household expenditure in the United States.

Using Google Trends data in a variety of fields and incorporating them into more complex econometric models were also tested subsequently. It is along those lines that our study contemplates to contribute. Indeed Google Trends supplies a large pool of series that may convey useful yet hidden information. Automatic variables selection or extraction methods seem to match perfectly this situation where a lot of potential regressors are available, the number of observations is limited, and the expert may not want to constrain the specification of the model too much. In this study, we confronted several approaches used for forecasting in high dimension: variables selection techniques well-known in macroeconomics, variables extraction methods which aim at summarizing a large set of data in a smaller one, averaging methods to take into account the modeling uncertainty, and, eventually, non-parametric methods borrowed from machine learning, which appear to provide accurate predictions in numerous and various fields.

The next section describes more precisely our data and the treatments that were operated on them. Then, we remind our reader quickly with the concepts behind the different techniques we used, and eventually, we present and discuss the main results.

2. Data

The main attraction of the Google Trends data for the economic outlook lies in the fact that they can be mobilized quickly and at a higher frequency than most traditional economic series. Indeed, data related to one given week are published at the end of the very week. Data can also be filtered by geographic origin: we could therefore restrict our study to searches carried out in France. Available data are pretreated which means that raw series corresponding to the real frequency of use of a search term are not made public. Applied treatments are not very well documented but series are supposedly corrected accordingly to a trend resulting from an increase in popularity of the search engine itself. They are normalized too so that their maximum always equals 100, which means that they might be revised between one extraction at a certain date and another one later on and that direct comparison between two distinct series is not possible.

Google provides categories grouping queries by topics. More of one hundred of them are available organized in a three levels hierarchy. Normalization of categories differs from keyword's one: the frequency of the category in the first week of 2004 is used as a reference, the following points in the series are expressed as deviations from this level. Since the

meaning of a search term can evolve over time, it seems preferable to work on categories or concepts rather than on specific terms. Plus the strategy of choice of keywords would be very subject to subjectivity in addition to consequent manual task. For example, the "Sports" category aggregates all search terms linked with the field of sport. French Google users have shown an increased interest in this topic in the summers of even years (figure 1). Indeed searches related to sport showed a marked increase during the football World Cup 2006, 2010, 2014, the European football championships and the Olympic Games in the summers of 2004, 2008 and 2012. Purchases of televisions usually increase significantly at times of major sports events, so using the "Sports" category seems to be a natural choice to measure the degree of interest that a sports event can generate among French consumers.

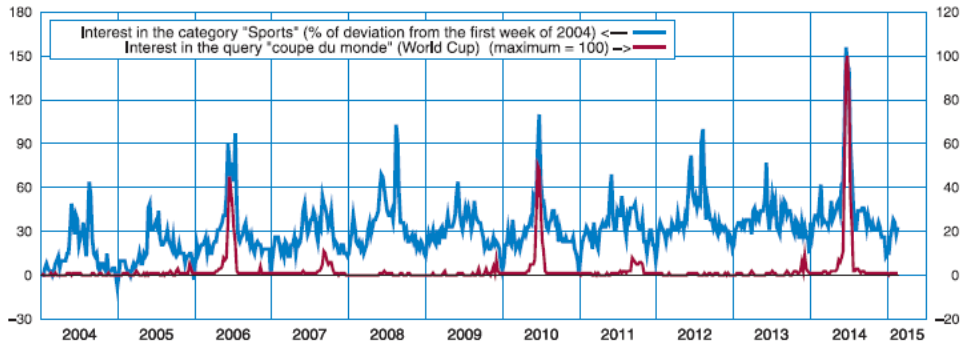

Figure 1. Examples of Google Trends chronicles for two chosen keywords. Source: Google Trends (2015)

In the context of our study, we selected a pool of 50 categories which may be correlated with the macroeconomic situation in one way or the other. The Google Trends categories were first transformed into a monthly format, weeks overlapping a month were distributed accordingly to the number of days in each month[2]. Series were then seasonnally adjusted; their monthly growth rates were computed to produce the explanatory variables, as well as their first time lag (i.e. the value of this growth rate in the previous month).

For this study, two targets have been considered: the household consumption in goods and the manufacturing production index. Representing more than half of GDP, household consumption is the largest item in final domestic demand, its estimate gives therefore a good outline of the whole activity. The first available data is the monthly household consumption expenditure on goods, published within one month. The publication of the manufacturing production index takes more time (two months), but its variation explains most of the quarterly GDP's evolution, it is then crucial to be able to produce accurate advanced estimates. In order to forecast these indices in real time or before they are published, usual

[2] This aspect makes it difficult to use techniques mixing data with different frequencies such as MIDAS (Mixed-data sampling), the advantage of the higher frequency was then not exploited here.

models incorporate variables from qualitative business tendency surveys available within a month. It seems reasonable to believe that the volume of queries made by users about particular products via the search engine could also reflect the potential volume of sales of these products, and, to a lesser extent, about production.

3. Methods

In this study, we confronted several approaches usually used for forecasting in high dimension: variables selection techniques well-known in time series, model averaging, and some machine learning techniques involving regression trees.

In dimension reduction problem, we can adopt two standard approaches: either we suspect that some variables are more important than the others, then the emphasis is put on identifying them; either we believe that some latent – unobserved – variables explain most of the comovements of the series considered altogether. In the first case, it is common to use iterative algorithms which add (respectively remove) a certain number of variables from an initial empty (respectively full) linear model, on the basis of a significance criterion. The main risk is to select a model which is far from the best possible model or to overfit, *i.e.* to adjust perfectly on observations used for estimation, which is generally associated with bad performance in forecasting new ones. More efficient approaches in terms of optimization have been developed such as penalized regressions - like LASSO (Tibshirani, 1996) or Elastic Net (Efron et al. 2004) - incorporating a penalty term in the objective function to favor parsimonious solutions. The main idea is to trade-off between the quality of the adjustment and some metric calculated on coefficients which prevents overfitting. In case the hypothesis of sparsity is challenged, variable extraction methods like principal component regression or partial least squares may reveal to be more efficient since they summarize the large set of data in a smaller set supposed to approximate the latent yet important variables.

When an estimator is the result of a model search amongst a collection of models in which multiple estimators are computed, one can potentially obtain an even better predictor by averaging these estimators for some selected models. In this respect, bayesian model averaging approach (Raftery et al., 1997) – BMA – combines multiple bayesian regressions weighted with their likelihood given the data (adjusted for complexity in a BIC criterion fashion) and a prior distribution on models (choosing here a binomial distribution with a probability lower than 0.5 for each variable to be included in order to search among parsimonious models).

Eventually, given the momentum of machine learning techniques in the context of a growing interest for “Big Data”, and their aknowledged performances in multiple and various fields, we also tried regression trees aggregation techniques like bagging and random forests.

Bagging (Breiman, 1996) consists in aggregating regression trees built on bootstrap samples (block bootstrap samples were used here to account for autocorrelation of time series). Any modeling technique can actually be bagged. Random forests (Breiman, 2001) introduce more randomness by sampling a set of regressors from the initial set of variables at each separation step of each trees. It entails more diversity in the aggregated trees since trees in bagging are usually very close since some variables get systematically selected. Boosting (Schapire et al., 1998) is quite different, it is an additive adaptive procedure which takes into account the biggest forecasting errors at one iteration when calibrating at the next iteration. This is done by actualizing some observations' weights.

To evaluate and compare the different approaches, we computed Root Mean Square Error (RMSE) in a pseudo real time fashion. After we fixed a first window - from 2004 to mid 2011; we proceeded for each month from end 2011 to end 2015 as follows: we extended the window by one month, estimated and calibrated the model, produced the one step ahead forecast and computed the forecast errors. The series of forecasts errors were eventually used to compute RMSE for each variable and methods.

4. Results

For households consumption in goods, a simple autoregressive model with one lag gets a RMSE of 0.56, we can see in Table 1 that this is hard to beat. For manufacturing production, given the publication delay, it makes no sense to compare to an AR(1) model, therefore we used the historical mean as benchmark (with a RMSE about 0.99). Again, the improvements are very limited. But it would be fairer to compare with models founded on business tendency surveys.

Table 1. Best performances (in RMSE/ out RMSE) obtained for different methods.

Variable	Stepwise	E. Net	BMA	Bagged E Net	Bagging	RF
Households consumption	0.48/0.52	0.53/0.53	0.49/0.52	0.58/0.53	0.61/0.56	0.54/0.53
Manufacturing production	0.88/0.99	0.90/0.99	0.88/0.97	0.91/0.98	0.82/0.92	0.64/0.98

Source: Bortoli & Combes (2016).

In each case, it seems not possible to expect more than 5-10% improvement. For households consumption forecasts, there is no clear winner. Selected Google Trends by stepwise selection or Elastic Net are multiple and diverse but the most contributing variable remains the lagged target. For manufacturing production, best performances are obtained for bagging of regression trees, other approaches don’t really compete with the historical mean.

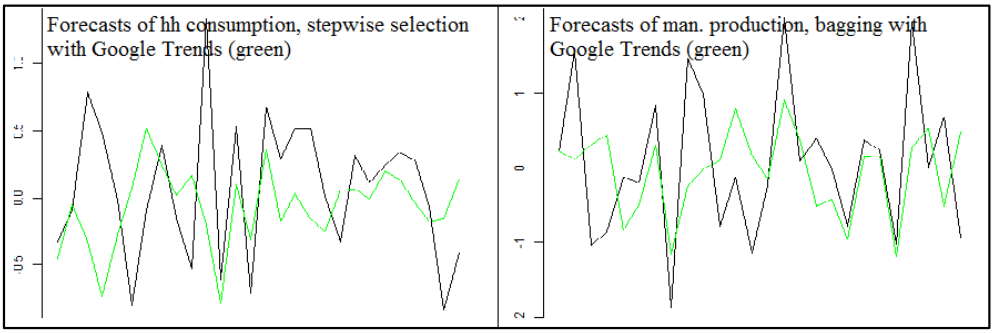

Figure 2. Examples of forecasts for households consumption and manufacturing prod.. Source: Bortoli & Combes

5. Discussion

Several reasons may explain these mitigated results. First, Google Trends series are very short, the comparison and evaluation protocol may suffer from the small number of observations and conclusions discussed here must be considered with caution. Plus, targets used in this study are not exempted of flaws. Household consumption in goods, for example, exhibits almost the properties of a white noise. Whatever the model or inputs, this series will remain hard to forecast. On the other hand, results for the purchase of certain goods (especially clothing and household durables) are more positive and some Google Trends categories seem to be probable explanatory variables (Bortoli and Combes, 2015). But expecting to get better forecasts for aggregated variables with a wide range of series thanks to automatic methods without any human intervention may be too naive. This may well be one limit to the attractiveness of "Big Data".

As far as Google Trends is concerned, it is worth noting than if they were to be used in the context of recurrent and official forecasts, we would not be in a position to judge the way these categories are built. Indeed their composition is unknown and we couldn't guarantee that the volume of searches is always enough to make statistical assumptions. Their composition may also change over time, especially when a popular new query appears at a given date. In addition, the series provided are the result of random sampling and can therefore differ from one data extraction to another. Repetitive forecasts founded on different extractions of the data will impose to re-estimate models systematically. The lack of transparency about treatments processed or sampling is one of the serious weaknesses of this tool, even if it proves to be more effective in another application than ours. From the official statistics' point of view, the sustainability use of the tool is also questionable. Indeed, Google Trends application is, by design, dependent on the technological developments in the search engine itself, continuously adapted to meet the needs of its users: the performance of the search engine and the underlying algorithms may evolve and lead to a change in the way in

which users use it (Lazer et al., 2014). Plus, since it was first created, the tool and the range of series available have changed substantially. Likewise, the free of charge access is based on the current marketing strategy of the company. Finally the behavior of users are continuously evolving, yet the quality of the data depends greatly on the individuals' habits of looking for information through the research engine. The growing share of smartphone applications could eventually lead to a reduction in the part played by search engines: the ability of trending searches to capture their behavior may decrease.

References

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2), 107–120.

Bortoli, C., & Combes, S. (2015). Apports de Google Trends pour prévoir la conjoncture française : des pistes limitées. *Note de Conjoncture, INSEE.*

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and Regression Trees. *Wadsworth.*

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, 5.

Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. *Technical report, Google*.

Claeskens G. (2012). Focused estimation and model averaging for high-dimensional data, an overview. *Statistica neerlandica*, 66(3), 272–287.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics,* 32(2), 407–499.

Kulkarni R., Haynes, K., Stough, R., & Paelinck, J. (2009). Forecasting housing prices with Google econometrics. *Research Paper, George Mason University School of Public Policy,* 457(10), 1012–1014.

Kunsch, H. R. (1989). The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics,* 17(3), 1217-1241.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science,* 343, 1203–1205.

Raftery, A., Madigan, D., & Hoeting, J. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association,* 437, 179– 191.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58, 267–288.

Vosen, S. & Schmidt, T. (2011). Forecasting Private Consumption: Survey-Based Indicators vs. Google Trends. *Journal of Forecasting*, 30(6),565–578.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*,26,1651-86

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.4245

Weighting machine learning solutions by economic and institutional context for decision making

Alvarez-Jareño, Jose A.[a] and Pavía, Jose M.[b]
[a] Departament of Economics, Universitat Jaume I, Spain.
[b] Department of Applied Economics, Universitat de València, Spain.

Abstract

It is quite common that machine learning approaches reach high accuracy forecast rates in imbalanced datasets. However, the results in the category with few instances are usually low. This paper seeks to improve the results obtained applying different techniques (such as bagging, boosting or random forests) with the inclusion of cost matrices. We propose applying the actual costs incurred by the company for misclassification of instances as a cost matrix. This approach, along with an economic analysis of the different solutions, makes it possible to incorporate a business perspective in the decision making process. The approach is tested on a publicly available dataset. In our example, the best ratings are obtained by combining the cost matrix with random forests. However, our analysis shows that the best technical solution is not always the best economical solution available. A company cannot always implement the optimal solution, but has to adopt a solution constrained by its social, institutional and economic context. Once an economic analysis is carried out, it seems the final decision of the company will depend on its economic situation and its institutional policy.

Keywords: *imbalanced datasets, random forest, cost matrix, economic analysis, uplift modeling.*

The authors wish to thank M. Hodkinson for translating into the English the text of paper and the support of the Spanish Ministry of Economics and Competitiveness through grant CSO2013-43054-R.

1. Introduction

This paper takes as reference Moro et al. (2014) and proposes a set of methods of statistical learning (logistic regression [LG], decision trees [DT], neural networks [NN] and support vector machine [SVM]) to analyse and subsequently predict the response of clients of a bank to a telephone marketing campaign.

As suggested by Radcliffe and Surry (2011), three types of models, which have continued evolving over time, can be identified:

- Entry models, which aim to profile those clients that are consumers of the product. These models answer the question "Who is buying?".
- Purchasing models, which aim to profile customers that have recently bought a product. As well as responding to the question "Who is buying?", the model also looks at "When does the purchase take place?".
- Response models, whose objective is to profile customers that have bought the product, ostensibly in response to a marketing campaing. These types of models look for answers to "Who is influenced by the marketing campaign in question?".

The model used by Moro et al. (2014) is a purchasing model and as such uses a unique set of data without a control group. A control group would be necessary to be able to estimate an uplift model, as shown by Guelman et al. (2015).

The data used are available on the webpage of UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.

2. The problem

The objective of the modelling is to forecast a dichotomous response variable, with a reliable degree of probability, by way of 20 predictors. The purpose of the forecast is to increase the success rate among the people contacted, avoiding contacting those clients who have a lower or null probability of opening a term deposit.

To be able to analyse the predictive capacity of the models the initial set of data was divided into two subsets, where the instances that belong to each set were randomly selected. The first subset (of learning or training) makes up 80% of the data and will be used to train the models.

Table 1 shows the predictive capacity of conventional processes analysed with default options. All the methods classify correctly in 90% of the instances, with the differences between them being minimal. The data set, however, is clearly imbalanced. The percentage

of instances with "No" responses is 88.73%, so a naïve classifier would obtain similar percentages of overall success.

Table 1. Predictive capacity of the models.

	Logistic Regression	Decision Tree	Neural Networks	Support Vector Machines
Correctly Classified	90.75%	90.67%	90.17%	89.9%
Incorrectly Classified	9.25%	9.33%	9.83%	10.1%
True Positive (TP) Rate	0.973 (no)	0.957 (no)	0.966 (no)	0.979 (no)
	0.421 (yes)	0.533 (yes)	0.427 (yes)	0.303 (yes)
False Positive (FP) Rate	0.579 (no)	0.467 (no)	0.573 (no)	0.697 (no)
	0.027 (yes)	0.043 (yes)	0.0034 (yes)	0.021 (yes)

Source: Own elaboration from Weka ouputs.

The TP rate of the methods in the target category (yes) is, however, very low. On average, only 42.1% of the clients who contracted the product are properly targeted. The objective of these models should be to obtain a higher rate of true positives in the target category, even though the false positive (FP) may be high in the complementary category or the ROC area may be lower. From this perspective, the decision tree would be preferable to the other models.

Regarding the measures of goodness of fit, there is not consensus. Whilst decision trees obtain the best measure of the Kappa statistic, logistic regression reaches a higher value in the ROC area. The lowest average absolute error is obtained using support vector machine.

Table 2. Model assessments.

	Logistic Regression	Decision Tree	Neural Networks	Support Vector Machines
Kappa statistic	0.4715	0.5242	0.4555	0.3696
ROC Area	0.9330	0.8940	0.8750	0.6410
Mean Absolute Error	0.1256	0.1197	0.1031	0.1010

Source: Own elaboration from Weka ouputs.

3. Methodologies to improve the results

The approaches analysed so far have not taken into account the fact that there is an imbalance in the data. This can be seen in the poor results obtained in predicting the uptake by customers of the product. The main objective for constructing these models is to identify

with greater accuracy those clients who would open a term deposit. In other words, which bank clients could be more easily encouraged to contract this product.

The imbalanced data sets need to introduce techniques which allow correct identification of the minority response, the solution to this problem being found in two distinct levels: at data level and algorithm level.

At the data level, the proposed solutions include different ways of model ensemble and resampling. Its aim is to ensure the predictions generated are robust. The different ensemble approaches are outlined by Rokach (2009). One of the first model assembly systems was bagging, proposed by Breiman (1996) and Buhlmann and Yu (2002) and implemented in R by Spanish researchers, Alfaro et al. (2013). Among the resampling techniques, we can found boosting, in particular, the algorithm AdaBoost M1, introduced by Freund and Schapire (1997) and extensively assessed in many studies and analyses, most notably by Eibl and Pfeiffer (2002) and Meir and Rätsch (2003). Random forests is another resampling method developed by Breiman (2001) as a variant of the bagging methodology using decision trees.

At the algorithm level, solutions include adjustments to the costs of various classes in order to counteract the imbalanced class: cost matrix. The aim is to adjust the probabilistic estimate of decision tree leaves (when working with decision trees), to adjust the decision threshold and to base learning on recognition rather than discrimination.

4. Results obtained with the proposed methodologies

After applying the different techniques described above, the following results were obtained:

- At data level, the different ensemble techniques (bagging, boosting and random forest) improve the robustness of the results, but without significantly improving the predictive capacity of the models.
- At algorithm level, the application of the cost matrix enables better identification of true positives. However, false positives are also increased, reducing the global accuracy of the model.

Tables 3 and 4 show the results obtained for the models in Table 1 with the application of the cost matrix. Table 4 also presents the evaluation of the model for the methodology of random forests with cost matrix. This approach is the one showing the greatest predictive capacity in the target category.

While the results of the quality of adjustment have worsened overall (see Table 4), the predictive capacity of all the models has improved for the target category (yes), even reaching levels of around 97% for Logistic Regression (see Table 3).

Table 3. Predictive capacity of the models with cost matrix.

	Logistic Regression	**Decision Tree**	**Neural Networks**	**Support Vector Machines**
Correctly Classified	76.57%	80.79%	88.02%	79.49%
Incorrectly Classified	23.23%	19.21%	11.98%	20.51%
True Positive (TP) Rate	0.738 (no)	0.794 (no)	0.903 (no)	0.772% (no)
	0.969 (yes)	0.912 (yes)	0.709 (yes)	0.965 (yes)
False Positive (FP) Rate	0.031 (no)	0.088 (no)	0.291 (no)	0.035 (no)
	0.262 (yes)	0.206 (yes)	0.097 (yes)	0.228 (yes)

Source: Own elaboration from Weka ouputs.

These results can be further improved by using random forests [RF] with cost matrix, achieving TP rates of 0.771 in the "no" category and 0.972 in the "yes" category.

Table 4. Assessments of models with cost matrix.

	Logistic Regression	**Decision Tree**	**Neural Networks**	**Support Vector Machines**	**Random Forest**
Kappa statistic	0.3875	0.435	0.5169	0.4296	0.4314
ROC Area	0.854	0.853	0.806	0.869	0.872
Mean Absolute Error	0.2343	0.192	0.1198	0.205	0.2051

Source: Own elaboration from Weka ouputs.

5. Economic evaluation of the results

With so many results, and with often very little difference between them, it is difficult to decide which model to use to select customers worth contacting. To simplify this decision making process, the economic criteria of income and costs can be used.

Assuming that the experiment was real, and under each prediction approach, the contacts would be reduced to those customers likely to say "yes" according to the model; and no contact would be made with customers identified by the model as unlikely to buy the financial product. Therefore, the cost of the campaign would be the sum of the second column of the confusion matrix multiplied by the cost of each contact (5 units), and the

income would be obtained by multiplying the number of clients that enter into a contract (amongst those who were contacted) and the average income of the bank for each contract (100 units). The difference would be the financial gain. The results are summarised in Table 5.

Table 5. Costs, income and ratios (over costs) of the campaign using different models.

Selection model	Costs	Income	Gain	Inc/Cost	Gain/Cost	Variation
No Selection	41,190	97,900	56,710	2.38	1.38	-
LG	3,035	41,200	38,165	13.58	12.58	-18,545
DT	4,165	52,200	48,035	12.53	11.53	-8,675
NN	3,335	41,800	38,465	12.53	11.53	-18,245
SVM	2,235	29,700	27,465	13.29	12.29	-29,245
RF	3,510	46,400	42,890	13.22	12.22	-13,820
LG-wCM	14,245	94,900	80,655	6.66	5.66	23,945
DT-wCM	11,945	89,300	77,355	7.48	6.48	20,645
NN-wCM	6,980	69,400	62,420	9.94	8.94	5,710
SVM-wCM	13,000	94,500	81,500	7.27	6.27	24,790
RF-wCM	13,075	95,200	82,125	7.28	6.28	25,415

Source: Own elaboration.

If no selection were made, and all clients were called, as is the norm in many real campaings, the cost would be 41,190 units, the income 97,900 and, consequently, the gain would be 56,170 units.

If a selection of people to contact had been made using the traditional models as proposed by Moro et al. (2014), the costs of the marketing campaign would have been significantly reduced, to a tenth of the original costs. However, there would have been a substantial reduction in the income of the financial institution and, consequently, in the gains. The highest income and costs had been achieved with the decision tree model, which shows the smallest losses compared to the situation where no selection was made.

The results (relative to a general campaign) change to a positive sign when the cost matrix is applied, achieving greater gains in all cases. In this scenario, the marketing campaign costs had been higher than under the proposals of Moro et al. (2014), ranging from 6,980 units for model NN to 14,245 units for model LG, but with income increasing substantially. With respect to a universal campaign, the costs are reduced by up to a quarter and the outcome is a greater gain.

All financial institutions may not have the same target or the same restrictions. Some might want to minimise the costs of the campaign, others maximise the income or the gain. The strategy of the bank will determine which method is the best to use for the marketing

campaign. In any case, from an economic perspective, using any of the models for the marketing campaign is still more efficient that not using one at all.

From the point of view of the strength of investment, we can compute the indices of income over costs or gains over costs to observe that the strategy of contacting all potential clients yields the lowest return on investment (1.38 units) with a cost almost even 13 times higher. Models without cost matrix would require an average investment of 3,200 monetary units, while models with cost matrix would need an average of 11,500 units. In other words, the necessary investment would be 3.6 times higher.

Although the task set for the models is the same in all cases, the best solution will depend on a series of financial variables and on the institutional policy adopted by the company. In the situation of the example studied, where capturing passive assets (by way of long term deposits) was one of the strategic objectives of Portuguese banks to improve the cash balances and to pass the European Central Bank stress tests, an increase in passive assets was favourable above a cost/gain analysis (Moro et al. 2014).

The selection of the model to be used for identifying the target audience of a marketing campaign will depend on the strategic objective set by the company, and could be very different for each situation. The best scientific solution is not always the best economic solution for a company, and having different options ensures the most appropriate strategic solution is selected.

6. Conclusions

The results obtained from the application of new models indicate that the inclusion of a cost matrix in the imbalanced sets significantly improves the classification of true positives to the detriment of true negatives. The other techniques used (boosting, bagging and random forest without cost matrix, and with or without using cross-validation, whose results have not been included due to lack of space), do not show any substantial improvement in the results obtained by Moro et al. (2014). Although it is true that they add robustness to the results, this does not always lead to an improvement and when there is, it is usually marginal.

From a scientific point of view, the best results are obtained combining the cost matrix with random forests. However, since the data is of an economic nature, the results should be approached from an economical-financial perspective. When entering costs and income as variables for decision making, we observe a variety of strategies to be taken by companies according to their economic situation and institutional policy.

The different technical solutions lead to different economic consequences, which in many cases need to fit in with the individual bank's circumstances. Faced with a set of economic restrictions not all the available technical solutions are viable and one should be chosen which either maximises or minimises a target within the capabilities of the company.

References

Alfaro, E., Gamez, M. & Garcia, N. (2013) adabag: An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software*, 54(2), 1-35.

Breiman, L. (1996). Bagging predictors. *Machine Learning*. 24(2), 123-140.

Breiman, L. (2001) Random Forests. *Machine Learning*. 45(1), 5-32.

Buhlmann, P., & Yu, B. (2002) Analyzing bagging. *Annals of Statistics*, 30, 927-961.

Eibl, G., & Pfeiffer, K. P. (2002). How to make AdaBoost. M1 work for weak base classifiers by changing only one line of the code. In *Machine Learning: ECML 2002* (pp. 72-83). Berlin-Heidelberg: Springer.

Freund, Y & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift random forests. *Cybernetics and Systems*, Vol. 46 (3-4), pp. 230-248.

Meir, R., & Rätsch, G. (2003). An Introduction to Boosting and Leveraging. In *Advanced Lectures on Machine Learning*, *LNAI 2600* (pp. 118-183). Springer.

Moro, S., Cortez, P., & Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, 62, 22-31.

Radcliffe, N., & Surry, P. (2011). Real-World Uplift Modelling with Significance-Based Uplift Trees. *Stochastic Solutions Limited*. http://stochasticsolutions.com/pdf/sig-based-up-trees.pdf

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, 53(12), 4046-4072.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3137

Quantifying and comparing web news portals' article salience using the *VoxPopuli* tool

Bonacci, Duje[a, b]; Jelinić, Antonija[b]; Jurišić, Jelena[a] and Alujević-Vesnić, Lucija[a]
[a]Communicology Department, School of Croatian Studies, University of Zagreb, Croatia
[b]VoxPopuli project

Abstract

VoxPopuli *tool enables quantification of absolute and relative salience of news articles published on daily news web portals. Obtained numerical values for the two types of salience enable direct comparison of audience impact of different news articles in specified time period. Absolute salience of a news article in a specified time period is determined as the total number of distinct readers who commented on the story in that period. Hence, articles that appear on web portals with larger audiences will in general be (absolutely) more salient as there are more potential commentators to comment on them. On the other hand, relative salience of a particular article during a particular time period is calculated as the quotient of a number of distinct readers who comented on that particular story and the number of all readers who in the same period commented on any news story published on the same news portal. As such relative salience will always be a number between 0 and 1, irrespective of the popularity of particular news portal, the (relative) salience of news stories on different news portals can be compared.*

Keywords: *VoxPopuli; news article salience; agenda setting theory; daily news web portals; readers' comments analysis.*

1. Introduction

1.1. Concept of issue salience and its measurement

As elaborated by Wlezien (2005), the concept of issue salience emerged in political sciences referring to the ''importance'' an individuals placed on certain issues. It indicates the perceived importance and/or prominence that a person attaches to particular (most often political) issue. Hence it can be said that salience of an issue determines the "ranking" of particular item on that individual's "private agenda". The higher the salience of the item, the higher up the agenda it is positioned.

When taken over a certain population, the cummulative "private agenda" turns into "population agenda". Hence, the capability to determine the salience of particular issue opens up the possibility of measuring – in quantitative terms - the structure of "public agenda" of the specified "public".

In practical research terms, salience is usually measured using a survey, typically by asking respondents to indicate 'the most important problem' facing the nation (Wlezien (2005)).

New media platforms, such as daily news web portals, enable new approach to measurement of issue salience. Namely, as interaction of the audience with the news stories published on such portals can be precisely quantitatively and systematically monitored, and as that interaction is necessarily related to the subjective importance the portal visitors attach to the issue elaborated by the respective news story, the quantitative measure of salience of issues raised by various news stories can be constructed/calculated and compared. As such web portals' audience can be taken to represent to some extent some broader social group (e.g. inhabitants of a particular town or city for a small local web portal, or citizens of particular country for a broader national web portal), this opens up the possibility of systematic real-time monitoring of issue salience

1.2. Readers' comments as an indicator of the news article salience

Salience of a particular news article published on some daily news web portal can be quantified using several distinct parameters. First that comes to mind is the number of (page)views the story attracts, as this parameter indicates how many portal visitors opened the story in their browser. However, as will be argued below, the validity of this parameter as an indicator of news story salience is questionable. The second potential indicator – and we argue much more valid – is the number of distinct visitors who engaged in a commenting of a particular news story.

Comments are better indicator than pageviews in several respect. First, not all daily news web portals have publicly available number of pageviews, whereas all the readers' comments (albeit only for those portals that enable commenting – but in fact nowdays great majority of them do) are necessarily publicly available and visible. Certailny, from the publishers' perspective, one of the point of the readers comments is actually to attract more visitors pageviews and so to expand the opportunity for online advertising.

Second, pageviews can be much more easily artificially "boosted" by the interested parties using automated software scripts that send page requests for a particular news stories on a particular portal. The comments, however, cannot be so easily manipulated for two reasons. First, it is not such an easy taks to generate a large number of false readers' profiles, and without these it is impossible to post comments. Namely, great majority of commenting platforms (e.g. Facebook comments plugin - https://developers.facebook.com/docs/plugins/comments, Disqus - https://publishers.disqus.com/) require some form of user authentification in order to be able to post the comments. Second, as comments are very contextual and intertextual pieces of text, it is impossible to generate huge number of diverse random comments which will appear as posted by the real person in a particular comment discussion. Hence, artificial/fake commentators and all of their respective comments can easily be filtered out from the analysis.

Third, the number of pageviews can be significantly influenced by the editorial interventions such as physical positioning of the news story on the web portal frontpage and/or vesting it with the exaggerated or attractive but misleading headline. On the other hand, it is safe to assume that readers will make an effort to comment on an article they opened not based on its primary appearance (reflected in afore mentioned editorially controlled parameter), but on the content of the news story contained. In other words, whereas readers of the daily news web portals will open many articles published on the news portal (hence increasing respective articles' pageview number), they will comment only on those articles whose content they find substantially important, as judged by their own subjective personal standards – i.e. the ones that really are for some reason salient to them.

In this light, we argue that the number that best quantifies the salience of a particular news article published on some web portal is actually not the number of comments the article attracts, but the number of distinct commentators that engage in commenting the article. This number indicates how many readers found the issues behind the news article personally significant to such an extent that they had an urge to actively state their opinon regarding that issue. What is more, from the qualitative content analysis of the comments themselves, the deeper reasons and explanations for such readers attitutes can be glanced. The number of comments then additionally indicate how controversial the issue is, because

the more controversial the issue, the more discussion develops among the involved commentators, and the more comments are hence generated.

2. VoxPopuli tool

2.1. Data harvesting engine properties

VoxPopuli is a software system/tool that enables automatic and systemtic monitoring and comparison of salience of issues published on daily news web portals. It achieves this by analyzing in real time the number of visitors who commented any of the articles published at these portals in a specified period by monitoring .

Engine currently harvests data from 43 Croatian local, regional and national daily news web portals, but is fully universal and can be adapted to harvest data from any daily news portal in the world.

Harvesting of news articles is set to 10-15 minutes intervals whereas each article is checked for new comments at interval between 8 minutes and several hours, depending on the current intensity of commenting activity for particular article, which is automatically determined by the engine. All these parameters can be modified to greater or smaller values if required.

The system is build as an server Java™ application which stores the data in Oracle™ MySQL database. System also has a web front-end (http://voxpopuli.hr) built in PHP which enables users to search for the news stories published on monitored portals in various periods.

2.2. Data quantities and data interpretation

VoxPopuli is a big data tool. Table 1. presents the typical quantity of data harvested by the *VoxPopuli* harvesting engine during a day, week and month.

Table 1. Typical daily, weekly and monthly quantities of data harvested by the VoxPopuli engine.

	Day (18.2.2016)	Week (8.-14.2.2016)	Month (january 2016)
Articles published	1.757	9.768	45.745
Articles commented	1.222	4.460	23.211
Total comments	24.578	117.793	699.251
Distinct commenters	6.166	15.974	42.453

The number of articles published by the monitored portals can be taken as an estimate of the number of (public) issues raised by these portals in specified period. Certainly, as "big news stories" always get covered in multiple articles, the number of issues is actually smaller than the number of articles. But this number is certainly still much greater that the one that can be administered through any survey questionnaire. As can be seen from data in Table 1., daily number of issues raised is counted in hundreds (~1700 articles for the day analised).

The fact that the number of published articles is smaller than the number of articles commented (roughly half of all the published articles attract any comments) indicates that not all news stories are considered equally significant (i.e. salient) by the readers-commentators. What is more, the data presented in Figure 1. show that, even among the articles that did attract some comments, the overwhelming majority - about 70% - of them attracted less than 10 commentators, whereas only about 2% of them attracted more than 100 commentators. This clearly demonstrates that even though daily news web portals raise a huge number of issues through the articles they publish, the readers' interest in these issues is actually very focused.

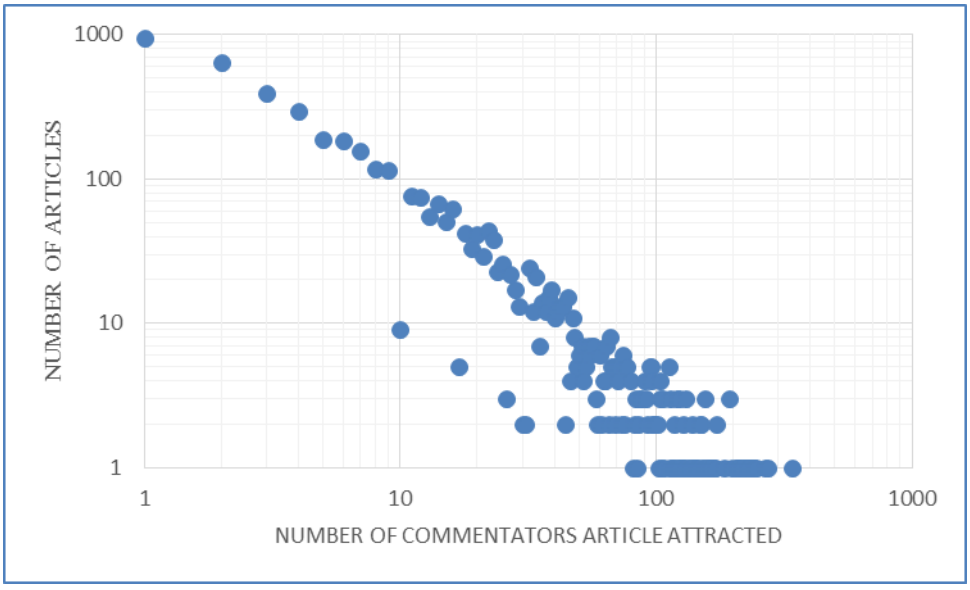

Figure 1. Distribution of commented articles with respect to the number of commentators who engaged in commenting them. Data correspond to the one week period between 8.-14. February 2016..

2.3. VoxPopuli *analysis*

Finally, Figure 2. presents the "*VoxPopuli* analysis" - comparative overview of all articles on all the daily news web portals currently monitored by the *VoxPopuli* system that

attracted some readers' comments, with respect to 3 distinct parameters of each article: absolute salience, relative salience and controversiality. Each bubble on the chart represents a single commented article.

Absolute salience of a news article in a specified time period is calculated as the total number of distinct readers who commented on the story in that period and is represented by the size of the bubble – the greater the bubble, the more commentators engaged in commenting the article. Articles that appear on web portals with larger audiences will in general be absolutely more salient as there are more potential commentators to comment on them and their respective bubbles will always stand out.

Further, **relative salience** of particular article during particular time period is calculated as the quotient of a number of distinct readers who comented on that particular article and the number of all readers who in the same period commented on any news article published on the same news portal. As such relative salience will always be a nulber between 0 and 1 (i.e. 0% and 100%), irrespective of the popularity of particular news portal, the relative salience of news articles on different news portals can be compared. This parameter is represented by the position of the article's circle on the vertical axis.

Finally, **controversiality** of the article is calculated as the percentage of commentators who posted more than one comment to the respective article. This parameter is represented on the horizontal axis of the graph. Namely, for any article analysed, great majority of commentators leave just one single comment. Multiple comments from a single author usually stem from the discussion in which that commentator engaged with other commentators. The more commentators engaged in discussion usually indicates that they "hotly debate" the issue presented by the article.

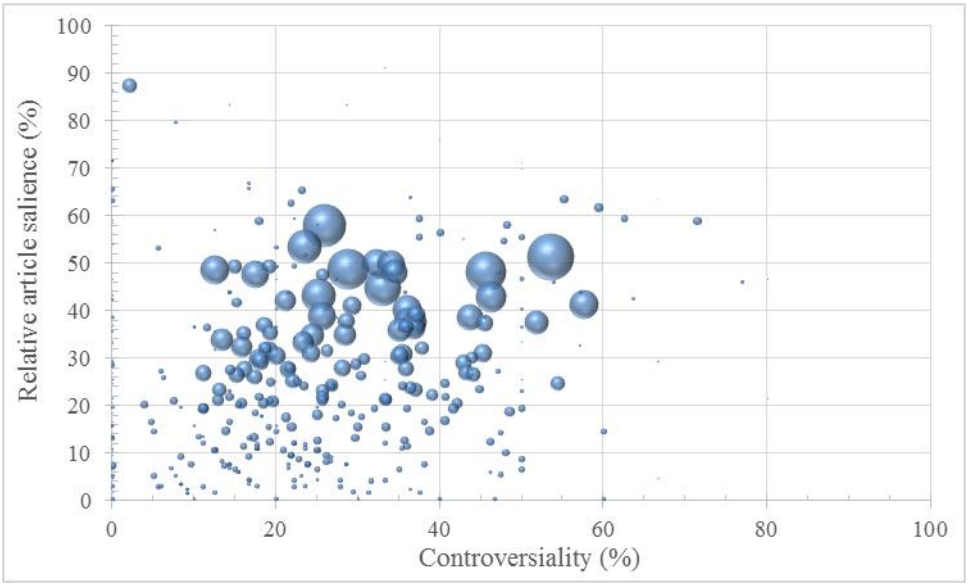

Figure 2. VoxPopuli *analysis graph. Data correspond to the one week period between 8.-14. February 2016.*

Using such chart, which can be (and in fact in current version of the VoxPopuli tool is) generated in real time, enables us to simultaneously monitor the salience of different news articles on various news portals and hence to quickly identify the "issue of the hour/day/week/etc." among all the issues raised by the monitored portals.

References

Wlezien, C. (2005). On the salience of political issues: The problem with 'most important problem'. Electoral Studies, 24, 555-579.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3642

Identification of Influencers in eWord-of-Mouth communities using their Online Participation Features

Olmedilla, M.[a]; Arenas-Marquez, F. J.[b]; Martinez-Torres, M. R.[a] and Toral, S. L.[c]
[a]Departamento de Administración de Empresas y Comercialización e Investigación de Mercados (Marketing), Universidad de Sevilla, Spain, [b]Departmento de Economía Financiera y Dirección de Operaciones, Universidad de Sevilla, Spain [c]Departmento de Ingeniería Electrónica, Universidad de Sevilla, Spain.

Abstract

The identification of influencers in any type of online social network is of paramount importance, as they can significantly affect consumers' purchasing decisions. This paper proposes the utilization of a self-designed web scraper to extract meaningful information for the identification of influencers and the analysis of how this new set of variables can be used to predict them. The experimental results from the Ciao UK website will be used to illustrate the proposed approach and to provide new insights in the identification of influencers. Obtained results show the importance of the trust network, but considering the intensity and the quality of both trustors and trustees.

Keywords: *e-word of mouth; influencers; Social Network Analysis; virtual communities.*

1. Introduction

The emergence of user-generated content has facilitated the interactions among users, so they can easily share opinions and exchange experiences. In this regard, the electronic interactions are complementing traditional word of mouth (WOM). The importance of WOM is widely accepted in traditional marketing research (Lee et al., 2008) and it is usually considered to be a very effective marketing tool with major repercussions on consumer behavior. However, it has evolved to a more impersonal but more pervasive form of WOM, the so-called electronic worth-of-mouth (eWOM), which is based on technology information advances and the growing access to the Internet (Law et al., 2014). eWOM is also providing an alternative and effective marketing channel to firms, which does not require huge investments in advertising (Ku et al., 2012). As a result, the identification of possible influencers is of great interest to business given the importance and impact that their reviews can cause on other consumers' purchase intentions. Marketing information can be propagated faster and promoted better via recommendations by influencers to their followers and peers (Cheung & Thadani, 2012). Previous approaches for the identification of influencers have been mainly focused on the idea of trust (Kim & Tran, 2013) and the degree of expertise in a specific domain (Ku et al., 2012). However, modern computational techniques can collect much more information about the social networking practices of users within these communities. For instance, the reputation of users can be measured using the ratings that their reviews receive from the rest of the community. Popularity is another feature of users that can also be measured using several metrics such as the number of comments or the number of readings received.

In this paper, we propose using a combination of reputation and popularity. Collected information can also enrich the dependent variables of the study. Typically, the trust network was included in previous studies by considering the size of the trust network. However, more recent works propose studying the trust network as a 2-hop network, considering also the quality of trustors (Kuk et al., 2012). Moreover, the social networking practices of users also include the possibility of scoring other posted reviews and trusting other users. This information is also publicly available in many eWOM websites. Finally, the domain in which users post their reviews can also be collected, considering different domain levels. All these new variables will be considered

The remainder of this paper is organized as follows. Section 2 details the related work about eWOM, the collection of information and the identification of influencers. The proposed methodology for collecting information and the definition of collected variables are detailed in section 3. Section 4 presents the empirical work and reports the evaluation results. The last section concludes the paper by summarizing the most important features of the proposed approach and by suggesting future research directions.

2. Related work

eWOM websites provide tools for consumers to discuss products and learn from other customer how to better use them (King et al., 2014). Among others, the online reviews usually include aspects such as a main text with the comments about the product, a general rating and the scoring of certain attributes and key phrases related to the product's perceived weaknesses and strengths. Additionally, some consumer-opinion websites include mechanisms that report reviewers' reputation (e.g. ratings received from other consumers) and allow members to add other members to a trust network (Ku et al., 2012).

Influencers are usually early adopters in markets, have multiple interests and are trusted by other consumers in a wide social network (Kiss & Bichler 2008). One major challenge of eWOM research consists in determining the characteristics that are more suitable for identifying influencers. Reviewer's exposure in the eWOM community (usually measured by how many times a user posts reviews on the website) is an important magnitude in previous studies. Hu et al. (2008) state that consumers pay more attention to reviewers with high exposure and their reviews are more likely to change consumers' uncertainties and transaction costs for buying a product. Lu et al. (2010) indicate that the number of reviews contributed by focal members positively correlates with the helpfulness of their reviews. Meanwhile, Huang et al. (2010) state that when a user has a great expertise in a field, she or he often writes more reviews on that specific field.

A number of papers suggest that reviewers' degree of expertise positively relates to their reputation and is likely manifested in their review behaviour. From this point of view, probably a high-level reviewer is a very active contributor in a certain product category or domain (Ku et al., 2012; Martínez-Torres & Diaz-Fernandez 2013). Arenas-Marquez et al. (2014) conclude that influencers usually review a wider range of products (i.e. products of different brands, technical features or benefits), which reflects their greater expertise with regard to a certain domain. Hung & Yeh (2014) state that influencers often post useful and knowledgeable contents. Therefore, these authors propose a text mining-based approach to evaluate features of quality of information and to identify influencers.

Finally, other existing works are based mainly on social network analysis. These papers study the topological features of the network formed by registered users within the consumer platform to identify influencers. Luarn et al. (2014) examine the influence of Facebook user's networks on the dissemination of information. They conclude that users with high network degree (more connections) and high clustered connections (frequency of information dissemination) have a greater influence on the dissemination process.

3. Methodology

After developing a set of tools for crawling the eWOM website, the gathered information must be transformed into a structured data format. The aim is obtaining a set of metrics representing the social networking practices of users from the collected information, which are going to be used for modelling and analysis in subsequent stages. The website includes some structured statistical information such as the number of reviews written, the review rating values or the number of reads received that are directly obtained from the programmed crawler. The user trust relationships (number of users trusted and number of users trusted-by) were obtained from the circles of trust information. Table 1 lists the variables considered in this study.

Table 1. Metrics describing the social networking practices of users.

Variable	Description
CritCap	∑ rating scores given per user
Tint	Size of the trust network
AvTInt	∑ network size of trustors / Size of the trust network
Tint-by	∑ users trusted-by
AvTInt-by	∑ network size of trustors / ∑ users trusted-by
ExpertCat	∑ categories of posted reviews per user
ExpertSubCat	∑ subcategories of posted reviews per user
MaxExpertCat	Maximum number of reviews in one category

- Critical capacity (CritCap). This variable is obtained as the sum of the rating scores given by each user.
- Trust intensity (Tint): Number of members who trust a given user (the size of his circle of trust).
- Average trust intensity of trustors (AvTInt): It is the average trustworthiness of all the trustors of a given user (i.e., average trust intensity of members who trust this user).
- Trust-by intensity (Tint-by): Number of users trusted by a given member (i.e., this user is included in other circles of trust)
- Average trust intensity of trustees: (AvTInt-by) It is the average trustworthiness of all the users trusted-by a given user.
- Level of expertise per category (ExpertCat): Total number of distinct categories in which a reviewer has written.
- Level of expertise per subcategory (ExpertSubCat): Total number of distinct subcategories in which a reviewer has written.
- Maximum level of expertise (MaxExpertCat): Maximum number of reviews written by a reviewer in a particular category.

4. Results

A crawler that follows the hyperlink structure of the users' webpages at Ciao has been developed using Scrapy with Python. As a result, the whole website Ciao.co.uk was crawled gathering information from about 45 thousand registered users within Ciao UK.

Although the number of registered users at Ciao UK is about 45 thousands, only a fraction, 12886 users, has posted at least one review. This is the typical participation inequality exhibited in many virtual communities (Martinez-Torres, 2013). However, the number of users posting only one review is still quite high. Therefore, in this study, we have filtered the original data and we have only considered those users posting more than one review. The number of users accomplishing this condition is 3158.

The condition of being an influencer can be defined in terms of reputation and popularity, following the studies by Kuk et al. (2012) and Arenas-Marquez et al. (2014). The reputation was measured by the average value of the received rating scores and the popularity by the average value of comments received. In this study we have considered two different thresholds given by the percentiles 90 and 95. More specifically, two definitions of influencers will be considered, as given by equations (1) and (2):

$$Infl90 = Reputation90 \ \& \ Popularity90 \quad (1)$$

$$Infl95 = Reputation95 \ \& \ Popularity95 \quad (2)$$

$Infl_{90}$ considers as influencers those users located in the percentile 90 of both reputation and popularity, while $Infl_{95}$ considers the percentile 95. Note that in both cases is a dichotomous variable, which takes the value 1 when the double condition is accomplished and 0 otherwise. The number of obtained influencers is 190 in the case of percentile 90 and 76 in the case of percentile 95. As we have a dichotomous variable, a binary logistic regression is appropriate to determine the variables that characterize the behaviour of influencers. However, obtained results show that influencers only represent a small fraction of community users. That means that the dependent variable contains a high number of zeros (which is the value for non influencers) and a low number of ones (which is the value for influencers). This kind of problems, where the dependent variable contains a disproportionally high number of zeros, are known as zero inflated problems, and they can lead to biased/inconsistent parameter estimates, inflated standard errors and invalid inferences (Lee et al., 2006). A possible alternative consists in considering generalized linear modelling with Poisson distribution. However, generalized linear modelling with Poisson distribution has problems with overdispersion (Hinde & Demetrio, 1998). The model with negative binomial distribution is an alternative way to fix over-dispersion problem in Poisson distribution (Hinde & Demetrio, 1998), as the variance and mean are

not assumed to be equal. This is the chosen regression model for this study. Obtained results for the two definitions of influencers are detailed in Table 2.

Table 2. Negative binomial regression results for the influenc-ers measured with the percentile 90 and 95.

	Dependent variable:	
	$Infl_{90}$	$Infl_{95}$
CritCap	-0.001***	-0.000
	(0.000)	(0.000)
Tint	0.01***	0.03***
	(0.002)	(0.002)
AvTInt	0.1***	0.1***
	(0.004)	(0.01)
Tint-by	0.03***	0.04***
	(0.004)	(0.005)
AvTInt-by	0.01***	0.01***
	(0.002)	(0.004)
ExpertCat	0.1***	0.1***
	(0.01)	(0.02)
ExpertSubCat	-0.01**	-0.03***
	(0.003)	(0.01)
MaxExpertCat	0.004***	-0.002
	(0.001)	(0.003)
Constant	-4.9***	-6.0***
	(0.1)	(0.2)
Observations	3,158	3,158
Log Likelihood	-434.1	-214.0
Akaike Inf. Crit.	886.1	446.1
Bayesian information criteria (BIC)	940.6	500.6
Precision	0.415	0.621
Recall	0.513	0.237
Overall	0.971	0.978
McFadden R^2	0.553	0.520
Note:	**p<0.1; **p<0.05; ***p<0.01*	

Results from Table 2 show that the critical capacity is not an important feature of influencers. Only in the case of $Influ_{90}$ definition there is a significant negative relationship, but with a very low coefficient. The negative relationship is the one expected, as the

influencers are supposed to have a good knowledge about the reviews they are scoring. More important is the influence of the circles of trust. The same than previous studies (Ku et al., 2012; Liu et al., 2015), both the trust intensity and the average trust intensity of trustors have a significant and positive influence over the condition of being an influencer. However, and as a novel contribution of this paper, the trust-by intensity and the average trust intensity of trustees also show this significant positive relationship, although at a lower level. Therefore, it is not only important the size of the circle of trust, but also the quality of this circle of trust, which means that people trusting an influencer also have a high circle of trust.

Table 2 shows a positive and significant relationship with the number of categories where the reviewer posts his or her reviews, but a negative relationship with the number of subcategories. According to previous studies, influencers exhibit a high level of expertise, which means they should focus on specific categories. Therefore, a negative relationship with the number of categories (ExpertCat), subcategories (ExpertSubCat) and the maximum number of reviews (MaxExpertCat) is expected. However, Table 2 only shows a negative relationship with the number of subcategories. This result can be explained by the way categories and subcategories are defined at Ciao. Ciao establishes 28 categories and the subcategories are then defined by reviewers. That means the main categories have a wide scope, so it is easy that a reviewer posts reviews belonging to several main categories.

4. Conclusions

This paper proposes a methodology for collecting user-generated content within eWOM websites in order to extend the number of variables usually considered for the identification of influencers. The data collection is based on the design of a self-programmed crawler to access the meaningful information related to the social networking practices at eWOM. Obtained results show the importance of the trust network, but considering the intensity and the quality of both trustors and trustees. We have also confirmed the low relevance of the critical capacity and the specialization of influencers but considering the level of subcategories rather than the level of the main categories.

References

Arenas-Márquez, F. J., Martínez-Torres, M. R., Toral, S. L. 2014. Electronic word-of-mouth communities from the perspective of social network analysis, Technology Analysis & Strategic Management, 26 (8): 927-942.

Cheung, C.M.K., Thadani, D.R. 2012, The impact of electronic word-of-mouth communication: A literature analysis and integrative model, Decision Support Systems, 54(1):461–470.

Hinde, J., and Demetrio, C. 1998, Overdispersion: Models and Estimation, Comput. Statistics and Data Analysis, 27:151-170.

Hu, N., Liu, L. and Zhang, J.J. 2008. Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. Information Technology and Management 9(3):201−214.

Huang, S., Shen, D., Feng, W., Baudin, C., Zhang, Y. 2010. Promote product reviews of high quality on e-commerce site, Pacific Asia Journal of the Assoc. for Information Systems 2(3):51-71.

Hung, C. Yeh, P.W. 2014. Identification of opinion leaders using text mining technique in virtual community. 1st Symp. on Information Management and Big Data, Cusco (Peru), 1318:8-13.

Kim, Y. S., & Tran, V. L. 2013. Assessing the ripple effects of online opinion leaders with trust and distrust metrics. Expert Systems with Applications, 40(9): 3500-3511.

King, R.A., Racherla, P., Bush, V.D. 2014. What we know and don't know about online word-of-mouth: a review and synthesis of the literature. Journal of Interactive Marketing 28:167-183.

Kiss, C. and Bichler, M. 2008. Identification of influencers - Measuring influence in customer networks. Decision Support Systems 46:233–253.

Ku, Y.C., Wei, C.P. and Hsiao, H.W. 2012, To whom should I listen? Finding reputable reviewers in opinion-sharing communities, Decision Support Systems, 53: 534–542.

Law, R., Buhalis, D. and Cobanoglu, C. 2014, Progress on information and communication technologies in hospitality and tourism, International Journal of Contemporary Hospitality Management, 26(5): 727-750.

Lee, A.H., Wang, K., Scott, J.A., Yau, K.K., McLachlan, G.J. 2006. Multi-level zero-inflated poisson regression modelling of correlated count data with excess zeros, Statistical Methods in Medical Research, 15:47–61.

Liu, S., Jiang, C., Lin, Z., Ding, Y., Duan, R., & Xu, Z. 2015. Identifying effective influencers based on trust for electronic word-of-mouth marketing: A domain-aware approach. Information Sciences, 306: 34-52.

Lu, Y., Tsaparas, P., Ntoulas, A. Polanyi, L. 2010. Exploiting social context for review quality prediction. International Conference on World Wide Web.

Luarn, P., Yang, J,C., Chiu, Y.P. 2014. The network effect on information dissemination on social network sites. Computers in Human Behavior 37:1–8.

Martínez-Torres, M. R. 2013. Application of evolutionary computation techniques for the identification of innovators in open innovation communities. Expert Systems with Applications, 40(7): 2503-2510.

Martínez-Torres, M.R., Díaz-Fernandez, C. 2013. Current Issues and Research Trends on Open Source Software Communities. Technology Analysis & Strategic Management 26(1):55-68.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.2991

Big Data Matching Using the Identity Correlation Approach

McCormack, Kevin and Smyth, Mary
Methodology Division, Central Statistics Office, Cork, Ireland

Abstract

The Identity Correlation Approach (ICA) is a statistical technique developed for matching big data where a unique identifier does not exist. This technique was developed to match the Irish Census 2011 dataset to Central Government Administrative Datasets in order to attach a unique identifier to each individual person in the Census dataset (McCormack & Smyth, 2015[1]). The unique identifier attached is the PPS No. (Personal Public Service No.[2]). By attaching the PPS No. to the Census dataset, each individual can be linked to datasets held centrally by Public Sector Organisations. This expands the range of variables for statistical analysis at individual level. Statistical techniques developed here were undertaken for a major European Structure of Earnings Survey (SES) compiled by the CSO using administrative data only, and thus eliminating the need for an expensive business survey to be conducted (NES, 2007[3,4,5]). A description of how the Identity Correlation Approach was developed is given in this paper. Data matching results and conclusions are presented here in relation to the Structure of Earnings Survey (SES)[6] results for 2011.

Keywords: *Identity Correlation Approach, Big Data matching, Unique identifier*

1. Identity Correlation Approach

The Identity Correlation Approach (ICA) was developed as part of a big data matching Project known as the SESADP (Structure of Earnings Survey Administrative Data Project) carried out by the Central Statistics office, Ireland (McCormack & Smyth, 2015[1]). The aim of the SESADP was to produce data to meet the EU SES 2014 Regulation from administrative data sources, from 2011 on an annual basis going forward. This eliminated the need for an expensive business survey to be conducted each year (NES, 2007[3,4,5]).

The ICA involves combining a number of individual variables for each person until a unique identifier is arrived at (McCormack,K 2015[7]). An example of this is combining the individual characteristics of each person in the Irish Census Dataset. Beginning with the variable for *date of birth*, then combine it with the variable *gender*, then adding variable for *county,& marital status, etc.* until a unique identifier is arrived at for each person. This is illustrated in Figure 1 below.

2.1 Theoretical Application

There was 1.6m employees in Ireland in 2011. Of these, an average of 65,000 persons were born in the same year (years 1946 to 1995), as illustrated in Figure 1.

Figure 1: Identity Correlation Approach: Simple Model - Combining Variables

Operation	Variable	No. of Records
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
	No. Persons with same DoB	178
Divide by:	Gender	2
	No. Persons with same DoB and gender	89
Divide by:	No. Counties	26
	No. Persons with same DoB, Gender, County	3
Divide by:	Marital Status (married & other)	2
	No. Persons with same DoB, Gender, County, marital status	1

The 65,000 persons born in the same year are divided by 365 days in the year to give approximately 178 persons with the same date of birth. The 178 persons with the same date of birth (DoB) can be divided by 2 for gender, to give 89 persons with the same DoB and gender. Dividing by the no. of counties a person lives in (89 divided by 26) results in 3 persons with the same DoB, gender & county. The 3 persons can be further subdivided by marital status resulting in 1 person with the same DoB, gender, county and marital status. Other variables used to further breakdown the data are industrial sector (NACE[5] code), no. of dependent kids, etc. A unique combination of variables for each person allows a person to be uniquely identified. This method is termed the Identity Correlation Approach (ICA).

2.2 Complexity and Variables Added

Complexity added to the simple model for the Identity Correlation Approach is shown in Figure 2. Populations are not evenly distributed, thus we allow for the fact that up to a third of the working population may be based in Co. Dublin. As a result, duplicates increase 10 fold for the No. of persons with the same DoB, gender & county. Similarly, the employee population is not evenly distributed in the various NACE sectors (industrial sector). Other variables added include no. of dependent children, which allows further breakdowns. In this way the Identity Correlation Approach arrives at a unique identity for each individual by combining a number of personal characteristics for the person.

Figure 2: Identity Correlation Approach: Complexity & Variables Added

Operation	Variable	No. of Records
	Approx. No. of births each year	65,000
Divide by:	No. days in the year	365
	No. Persons with same DoB	178
Divide by:	Gender	2
	No. Persons with same DoB and gender	89
Divide by:	No.Counties (allowing for approx. one third employees living in Dublin)	3
	No. Persons with same DoB, Gender, County	30
Divide by:	NACE industrial code (15) - allow for one fifth employees in same NACE Sector	5
	No. Persons with same DoB, Gender, County, NACE	6
Divide by:	Marital Status (married & other)	2
	No. Persons with same DoB, Gender, County, NACE, marital status	3
Divide by:	No. of dependent kids (3 groups)	3
	No. Persons with same DoB, Gender, County, NACE, marital status, no. dependent kids	1

2. Practical Application

2.1 Census 2011 data

The identity Correlation approach was applied to the Irish Census Data 2011 as described above. This allowed for a Unique Identifier (UI) to be applied to each individual by combining their personal characteristics (i.e. DoB, gender, county residence, etc.). The unique identifier is called the matching variable (matchvar) which is used to link each person's record to other datasets (see Fig. 3).

Figure 3: Applying Identity Correlation Approach to Create Unique Identifier (Matchvar)

Date of Birth	Gender	County	NACE	Marital Status	No.Kids	Matchvar (all variables)
15031949	M	CORK	42	M	0	15031949MCORK42M 0
11021945	F	LIMERICK	31	S	1	11021945FLIMERICK31S1
21111954	M	DUBLIN	25	D	2	21111954MDUBLIN25D2
19051964	M	CARLOW	55	O	2	19051964MCARLOW55O2
22091966	M	GALWAY	82	M	3	22091966MGALWAY82M3
24031971	F	CAVAN	84	M	0	24031971FCAVAN84M0

2.2 Public Sector Administrative Datasets (ADS)

A single master Administrative Dataset (ADS) was created by linking a number of Public Sector Administrative Datasets (Revenue Commissioners Tax data, Social Security Administrative Datasets and the CSO's Administrative Datasets (e.g. Central Business Register (CBR), Earnings and Labour Force Survey)). These datasets were combined using the PPS No. for each individual and the CBR Enterprise No. for Establishment Surveys. The identity Correlation approach was applied to the master Administrative Dataset (ADS) also, allowing for a Unique Identifier (UI) to be applied to each individual by combining their personal characteristics (i.e. DoB, gender, county residence, etc.). This Unique Identifier known as the match variable (matchvar) was then used to link to the UI (matchvar) in Census.

2.3 Linking Census to ADS

Variables common to both the Census dataset and the master Administrative Data Source (ADS) were identified (e.g. DoB, gender, etc.). These common variables were joined to each other to create a Unique Identifier on each dataset using the Identity Correlation Approach. By linking the two datasets using the Unique Identifier, a PPS No. could be applied to each individual person in the Census. This is shown in Figure 3. Once the PPS No. was assigned to the Census dataset, it enabled Census data to be linked to any Public Sector Administrative Dataset.

3. Dataset Matching

3.1 Census dataset

A total of 1.6 million employee records were extracted from the 4.6 million Census Records. Approximately 200,000 records had a unique Business No. identifier attached (CBR No.). Another 500,000 records had a CBR No. attached using the Employer's Business name on the Census. The first matching variable (Matchvar1) created for Census used the following variables combined: CBR No., Dob, gender, county, NACE 2, marital status, No. of kids. A second matching variable was created (Matchvar2) excluding NACE2 (see Figure 4). Up to ten matching variables (Matchvar1 – Matchvar10) were created. Each matching variable is similar to the previous one with one change to the composition variables for each subsequent matching variable created. Figure 4 illustrates the construction of each subsequent matching variable.

Figure 4: Matching Variables

Date of Birth	Gender	County	NACE	Ent No.	Marital Status	No.Kids	Match Var 1	Match Var 2	Match Var 3
15031949	M	CORK	42	EN12345678	M	0	15031949MCORK42EN12345678M0	15031949MCORK42EN12345678M	15031949MCORK42EN12345678
11021945	F	LIMERICK	31	EN52345679	S	1	11021945FLIMERICK31EN523456791	11021945FLIMERICK31EN52345679S	11021945FLIMERICK31EN52345679
21111954	M	DUBLIN	25	EN52795680	O	2	21111954MDUBLIN25EN527956802	21111954MDUBLIN25EN52795680O	21111954MDUBLIN25EN52795680
19051964	M	CARLOW	55	EN32795681	D	2	19051964MCARLOW55EN327956812	19051964MCARLOW55EN32795681D	19051964MCARLOW55EN32795681
22091966	M	GALWAY	82	EN22795682	M	3	22091966MGALWAY82EN227956823	22091966MGALWAY82EN22795682M	22091966MGALWAY82EN22795682
24031971	F	CAVAN	84	EN52795683	M	0	24031971FCAVAN84EN527956830	24031971FCAVAN84EN52795683M	24031971FCAVAN84EN52795683
28021977	F	DUBLIN	71	EN84355684	S	1	28021977FDUBLIN71EN843556841	28021977FDUBLIN71EN84355684S	28021977FDUBLIN71EN84355684
30061990	F	KERRY	35	EN73795687	M	1	30061990FKERRY35EN737956871	30061990FKERRY35EN73795687M	30061990FKERRY35EN73795687

3.2 ADS (Public Sector Administrative Datasets)

The records in the master Public Sector Administrative Dataset (ADS) also contained all the above variables used in Census to create the matching variables (Matchvar1 – Matchvar10). Therefore the matching variables created in the ADS were used to match to the same variable in the Census.

3.3 Incremental Matching Process

To match both datasets (Census to ADS) the matching variables (Matchvar1 – Matchvar10) were used.

First Matching (using *Matchvar1*)

In practice, duplicates will occur when the Unique Identifier (*Matchvar1*) is created. This can be due to errors in coding. Therefore in the dataset linking process, each dataset (Census and the ADS) is edited to extract records where the Matchvar1 is unique. If there are duplicate occurrences of *Matchvar1*, then these records are deleted from the matching process. The two datasets (Census and the ADS) were matched using *Matchvar1*. This yielded a match of 307,300 Census records to the ADS dataset.

Second Matching (using *Matchvar2*)

In the second stage of matching, the remaining unmatched records in each dataset (Census and the ADS) are edited to extract records where the variable *Matchvar2* is unique. Only unique occurrences of Matchvar2 are used to match both datasets. This yielded a match of 75,400 Census records to the ADS dataset.

Third and consecutive matches (Matchvar3 to Matchvar11)

The matching process continued in this incremental process using Matchvar3 up to Matchvar10. This yielded a total of 797,100 Census records to the ADS dataset.

Figure 5: Incremental Matching Process

Unique Identifier (UI)	Variables constituting UI	No. of Records Linked
Matchvar1	(CBRno|DoB|sex|NACE2|PP|County|MS)	307,393
Matchvar2	(CBRno|DoB|sex|NACE2|PP|County)	75,410
Matchvar3	(CBRno|DoB|sex|NACE2|PP|MS)	29,854
Matchvar4	(CBRno|DoB|sex|NACE2|PP)	14,885
Matchvar5	(DoB|sex|NACE2|PP|County|MS)	47,963
Matchvar6	(DoB|sex|NACE2|PP|County)	16,898
Matchvar7	(DoB|sex|NACE2|PP|MS)	25,996
Matchvar8	(DoB|sex|NACE2|PP)	15,170
Matchvar9	(DoB|sex|NACE2|County|MS)	204,099
Matchvar10	(DoB|sex|NACE2|County)	59,102
	Total no. of records linked	**796,770**

3.4 False Positives

False positives can occur in the matching process if a variable is incorrect on one of the datasets. For example is the county variable has not been updated on the Social Welfare dataset then the county will be different on the person's record on Census. Similarly, if the NACE code is incorrect on either dataset, then it will not match a person to their correct record.

4. Mathematical Representation of Identity Correlation Approach (ICA)

Creating a Unique Identifier (UI) for each record using the Identity Correlation Approach (ICA) will result in a perfect match of records across two datasets, if there is a sufficient overlap of variables on both datasets. The probability of matching records across two datasets can be calculated by a formula known as the Matching Rate of Unique Identifier (MRUI). This is illustrated in Table 1.

Table 1: Matching Rate for Unique Identifier

$$N * \frac{1}{V1_{ui}} * \frac{1}{V2_{ui}} * \frac{1}{V3_{ui}} * \frac{1}{V4_{ui}} * \ldots\ldots \frac{1}{Vx_{ui}} = MRUI$$

Where:

N = Population , V = Variable, x = no. of variables

ui = Uniqueness Factor. ui = no. of classes where variable is distributed evenly across all classes

MRUI The Matching Rate for Unique Identifier (MRUI) is the ability to identify a unique record in a dataset, given the combination of variables used to deduce the record. Mathematically it is assumed that variables are discreet (non-dependent).

5. Results & Conclusion

Results for matching Census records to Administrative Datasets are given in Figure 6, classified by NACE industrial sector. The lower rate of matching in some sectors (e.g. Construction) can be attributed to records not being updated for certain variables. If the theoretical MRUI value indicates a perfect match, but this is not reflected in practice, then there are issues with coding or with records not being updated.

Figure 6: Employee Population Coverage 2011 - Census & Administrative Datasets Matched

Nace Economic Sector	No. Employees Total
B-E Industry	55
F Construction	26
G Wholesale and retail trade	44
H Transporation and Storage	51
I Accommodation and Food Services	31
J Information and communication	52
K-L Financial, insurance, etc.	61
M Professional, scientific & technical	39
N Administrative and support services	26
O Public administration & defence	69
P Education	63
Q Health & social work	46
R-S Arts, entertainment, other services	41
Total	47

5.1 Impact of ICA Data Matching

The ICA (Identity Correlation Approach) enabled 50% of the employee population in Ireland (750,000 of 1.5 million employees) to be matched to the Census 2011 dataset, as part of the SESADP. This enabled the Census dataset to provide variables for the SES (e.g. education level and occupation). Outputs from the SESADP produced the SES data for 2011 to 2014 and avoided having to do an expensive business Survey. IT and statistical infrastructure are now in place to produce the SES on an annual basis going forward, reducing costs from €1.6million annually to €0.1million. A Cost/Benefit Analysis of the SESADP is given in Figure 8.

5.2 Data Quality

An analysis of the data was undertaken to determine if the ICA matched the Census correctly to the other administrative data sources (ADS). There was a 90% correlation with the individual's *employer name* on Census with the *employer name* on the Business Register. The 10% with a different business name were eliminated as false positives. In Figure 9 a distribution of employees by age group and NACE in the Census dataset shows a very good comparison with the SESADP (similar results were obtained for education, occupation & gender comparisons).

Figure 8: Cost/Benefit Analysis of the SESADP

	Business Survey former NES	**SESADP Project**	**SESADP (Annual)**
Survey Type	Annual Survey	Data for 4 years	Annual basis - going forward
Reference period	Years 2002 to 2009	2011 to 2014	2015 onwards
Cost	€ 1.6 million p.a.	€0.4m	€0.1m p.a.
Timeliness	T+ 18	2 years to develop	T+ 10
Data edits	Data Edits	No edits - Revenue data	No edits - Revenue data
Sample size	65k	800k	800k+
Coverage of employee population	4%	50%	50%+
Burden	70,000 employees	None	None
Burden	5,000 enterprises	None	None
Staff Nos.	15 FTEs	4	2
Savings	-	€6.0m	€1.5m p.a.

Figure 9: Employees Nos. in SESADP compared to Census dataset by Nace Sector and Age Group 2011

NACE Economic Sector	**% difference in employee nos.**					
	Age Group in years					
	15-24	**25-29**	**30-39**	**40-49**	**50-59**	**60 and over**
	%	%	%	%	%	%
B-E Industry	-1	-1	2	1	0	-1
F Construction	-1	-1	3	0	-1	-1
G Wholesale & Retail Trade; Repair of Motor Vehicles and Motorcycles	0	-1	1	1	0	-1
H Transportation and Storage	-1	-1	0	2	1	-1
I Accomodation and Food Service Activities	-1	-1	2	1	0	-1
J Information and Communication	-1	-2	3	0	1	-1
K-L Financial, Insurance and Real Estate	-1	0	3	0	0	-1
M Professional, Scientific and Technical Activities	0	1	3	-1	-2	-1
N Administrative and Support Service Activities	0	1	4	0	-2	-2
O Public Administration and Defence; Compulsory Social Security	-1	-1	2	1	-1	-1
P Education	-2	-2	2	2	0	0
Q Human Health and Social Activities	0	-1	1	1	-1	-1
R-S Arts, entertainment, recreation and other service activities	0	1	1	1	-1	-2

References

McCormack,K (2015). Constructing structural earnings statistics from administrative datasets: Structure of earnings survey – Administrative data project. *The Statistics Newsletter – OECD***,** 62, 3-5[8].

McCormack,K. & Smyth,M. (2015). Constructing structural earnings statistics from administrative datasets. *New Techniques and Technologies for Statistics (NTTS) 2015. Collaboration in Research and Methodology for Official Statistics*[1].

McCormack,K. & Smyth,M. (2015). Specific Analysis of the Public/Private Sector Pay Differential for National Employment Survey 2009 & 2010 Data. *Research Paper. Central Statistics Office, Ireland*[3]

NACE is the "statistical classification of economic activities in the European Community" [9]

National Employment Survey 2007 (2009), *Central Statistics Office, Ireland*[6].

National Employment Survey 2008 and 2009 (2011), *Central Statistics Office, Ireland*[5].

National Employment Survey 2009 and 2010 Supplementary Analysis (2012), *Central Statistics Office, Ireland*[4].

The European Union Structure of Earnings Survey (SES) in accordance with Council Regulation n° 530/1999 is conducted in the 28 Member States of the European Union as well as candidate countries and countries of the European Free Trade Association (EFTA) [7].

The Personal Public Service Number (PPSN) is a unique reference number that allows individuals access social welfare benefits, public services and information in Ireland. State agencies that use PPSNs to identify individuals include the Department of Social Protection, the Revenue Commissioners and the Health Service Executive (HSE). [2]

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3133

Online CASE CPI

Radzikowski, Bartosz[a] and Śmietanka, Adam[a]
[a]CASE - Center for Social and Economic Research, Poland

Abstract

Online CASE CPI is an example of using Big data in public statistics. In principle, it is a consumer price index based entirely on online prices: a combination of Central Statistical Office of Poland's methodology and online data sets. An innovative method of data collection – data scrapping – allowed us to substantially reduce a time delay between data collection and a publication of results. A short, nine-month period of data collection has not given rise to make important conclusions, hence the aims of this paper are: to discuss a general framework of measuring consumer inflation online, to present preliminary results for Poland and to highlight the strengths and weaknesses of this approach. Finally, we believe that online consumer price indices have a complementary nature to conventional inflation measurement, but it might be a serious alternative, having in mind a huge growth potential of e-commerce in coming years.

Keywords: *Big data in public statistics; scrapping data; inflation measurement; e-commerce; online vs offline consumption*

1. Introduction

All actors in the economy are interested in accurate and timely information on changes in price levels. A rate of inflation influences households' decision whether to consume or save, enables economists to predict an economy's position in the business cycle, and determines the level of central banks' interest rates.

National statistical offices employ pollsters to visit thousands of retail outlets, restaurants and service units to collect price data across the country. Since the process is time-consuming, final information about prices, e.g. the consumer price index (CPI), is announced with some delay. Moreover, it is hard to establish a precise day for data collection because it lasts a whole month. For instance, Central Statistical Office of Poland (GUS) publishes its CPI index two weeks after the end of the month, meaning the data for a certain month is only available, in the best case, two weeks after the fact. Indeed, having in mind the fact that the data collection lasts almost the whole month, the real delay (between a day of collection and announcement of the results) might be longer than two weeks.

The research embodied in this paper has been performed in order to reduce the time delay and to measure prices volatility online. We developed a methodology to collect the data on prices in real time, along the lines of the innovative project elaborated by Massachusetts Institute of Technology's academics in *The Billion Price Project*. Integrating a model of household consumption used in the CPI index published by GUS, we have created an *online CASE CPI* for Poland that is calculated entirely on the price data from the Internet. What is worth mentioning, our CPI is not a substitute for the GUS's CPI and other official measures of inflation, but rather represents a faster and more frequent estimation of a similar nature. From this standpoint alone, the project is, thus, a unique and innovative study using *Big data* for statistical purposes in Poland.

2. Online vs. offline prices

Researchers, who blazed a trail in terms of analyzing prices via the Internet was Cavallo & Rigobon (2011). They collected individual-product prices in 36 supermarkets across 22 countries and 5 continents using a scraping software. In effect, they received 5-million-observation data sets to test for a price stickiness. Since that time, at least 7 studies have already been performed on these data sets, and some findings are quoted in this paper.

Although Internet shopping is becoming increasingly popular, online traders are still not necessarily representative of the typical consumer. Similarly, prices on the Internet may differ from those in brick-and-mortar retailers. If prices online and offline behave differently in long term, then index based on online prices alone cannot be extrapolated to the whole economy. However, studies comparing prices of the same goods in traditional

stores and online retailers found that those prices are either identical or there is a stable deviation between them. Cavallo *et al.* (2014). Basically, we can distinguish three types of relation between online and offline prices:

a) **Permanent shift**. Online prices may have a tendency to outrun prices in the real world. The shift may be around 2-3 months, like in France.
b) **Cohesion**. Online and offline prices behave in a similar manner, having in mind short-term deviations, like in the USA.
c) **Different strength.** Online prices may react stronger or weaker than those in the real economy, like in Columbia.

Moreover, a relation between prices may also differs due to a reference period. Even though online and offline prices follow similar trend on an annual basis in the USA, monthly indices vary significantly in some points.[1]

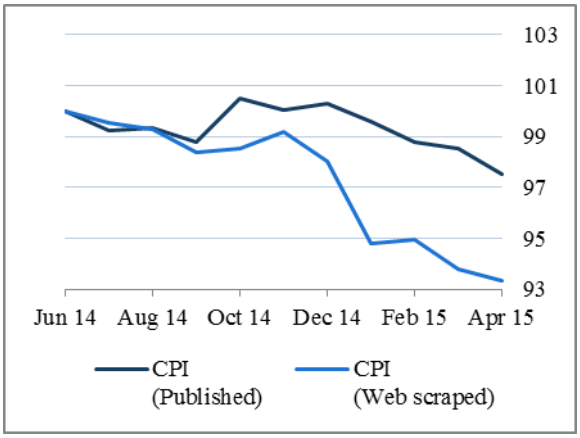

Figure 1. Price of food and drinks in the UK (June 2014 = 100), source: Office for National Statistics in the UK.

A proof of different behaviors between online and offline prices came from the UK. The UK's statistical agency compared a conventional consumer price index with an online-based version on data from supermarket websites. During the nine-month period between June 2014 and April 2015, the official CPI for 35 items of food and drink fell 2.5%, while the equivalent built on online prices from three large supermarkets fell by 6.7% (Figure 1) (Giles, 2015).

[1] Figures and data for each country could be found in Cavallo A. (2015)

Short term disparities between online and offline prices result from number of reasons. First, pricing strategies of companies. Since online customers and their behaviors vary in comparison to offline buyers, companies adjust their offer to maximize profits. Second, a number of online transactions is still far beyond those in the real economy. In 2014, an estimated share of online goods in total retail of goods was around 5.9% globally, and 6.4% in Europe. Although this share has increased more than doubled since 2010, it is still a small fraction of the global consumption (Ecommerce Foundation, 2015). Third, online market is more competitive. Due to price-comparison websites, an asymmetry of information about the prices on the market is reduced. In effects, retailers' markups are modest and the prices are more flexible than in bricks-and-mortar stores. Fourth, *menu costs* are extremely low, actually they refer to price data update, which also contributes to a lower price rigidity.

3. GUS's CPI vs. online CASE CPI

A composition of the online CASE CPI's basket of goods and services follows GUS's methodology; thus, it based on 12 main aggregates announced by GUS in February each year (the basket composition for 2016 is presented in the Table 1).

However, those main aggregates do not provide enough precision to take into account the consumption patterns of a typical person. Therefore, each of the main aggregates is divided into smaller categories (93 in total), according to latest available the Household Budget Survey. For instance: according to the Household Budget Survey in 2014, consumption of cereal accounts for 0,59% of consumption of Food and non-alcoholic beverage, which, as an aggregate, accounts for 24,04% of total expenditures, according to the 2016 revised inflation basket (Table 1). A list of representatives, which form each category, come from RAMON Eurostat's Metadata, according to COICOP classification.

Due to missing categories (about 15% of goods and services do not have prices available online) the shares are accordingly scaled. The share of missing categories is split proportionally between other categories inside the same aggregate (the same is done if one or more aggregates is missing – its share is divided accordingly between the remaining ones). For example, services in recreational and cultural category are characterized by high heterogeneity, variability of the offer over time, and dispersion of pricing data. Those factors might influence reliability of the data and as a result that category is omitted. The share of recreational and cultural services in expenditures is proportionally divided between other categories in Recreation and culture aggregate. The idea of scaling missing categories is in line with Cavallo (2012).

Table 1. Official weights of main aggregates of goods and services in inflation basket in 2016

Category	Share
Food and non-alcoholic beverages	24.04%
Alcoholic beverages. Tobacco	6.56%
Clothing and footwear	5.47%
Housing. water. electricity. gas and other fuels	21.04%
Furnishings. household equipment and routine maintenance of the house	4.99%
Health	5.45%
Transport	8.72%
Communication	5.27%
Recreation and culture	6.63%
Education	1.01%
Restaurants and hotels	5.04%
Miscellaneous goods and services	5.78%
Total	100.00%

3.1. Sources of data

To ensure that online CASE CPI is not influenced by one retailer and its pricing strategy, data on prices are collected mostly from the websites which compare prices from online shops. This allows us to easily track prices of certain goods and services from a number of outlets. What is more, those sellers do not have to be predefined - if a new retailer enters the market, price-comparison websites will automatically include its offer in the search results. By using price-comparison websites, online CASE CPI takes into account a dynamically changing market environment. For representatives which are not listed on those websites, we use dedicated websites, like commodity exchange for fresh products or industry portal for petroleum prices. In case of services and utilities data are collected either from private websites that contain listings of prices or public websites publishing current tariffs (for example websites of municipal transportation authorities). In effect, we tracks

prices on about 50 different pages across the Internet, however through the use of price-comparison websites, data from more than 3000 outlets is taken into account. Some of the data sources are updated in real time, some in regular intervals (at least once a month), others only when price of certain commodity has changed (for example prices of electricity are in effect until new tariff is announced).

To gather information on prices, Central Statistical Office of Poland sends out more than 200 agents to collect data from about 35,000 points of sales throughout the country each month. Data for online CASE CPI are scrapped using sophisticated internet robots. They gather approximately 60 000 observations each week, which means 240 000 observations monthly. The main advantage is the speed of collection because the whole process takes a few hours. As a result, we are able to calculate and announce the index almost in the real-time, even CASE's CPI, like the GUS's CPI, tracks approximately 1400 representatives. Main statistics for both indices are presented in Table 2 below.

Another advantage of the *scrapped data* is that prices of newly introduced products and services can be collected from the first day of their online appearance without need for any administrative adjustments in the basket. There is no need to decide whether new products should be included in the index - they automatically are. Moreover, GUS tracks prices for chosen representatives until they stop being offered. Then they look for the closest substitute in order to remain continuity. Official CPI measures control and "divide" price between attributes of certain products - if such product is discontinued (for example when a new technology is introduced) it can be compared to a new product with slightly different attributes[2].

Table 2. Comparison of CPI conducted by GUS and CASE

	CASE CPI	**GUS CPI**
Number of representatives	~1400	~1400
Number of observations	~240 000	~260 00
Frequency of data collection	Weekly	1-3 times a month
Time of availability	Real-time	Up to 4 weeks delay

[2] For more information please refer to: IWGPS: Consumer Price Index Manual: Theory and Practice (2004), chapter 7

We decided to track prices for chosen representatives only in a few categories, including processed food and apparel. These goods or services are characterized by relatively large heterogeneity in terms of price range and attributes and thus may disrupt the final result (for example a price reduction of an upscale dress does not mean that cost of living of a typical person fell to any degree). In general, online CASE CPI does not adjust goods and services for quality changes, as all items are treated independently. Because of use of a number of representatives for each product type, it could simply omit discontinued products.

4. Preliminary results

Online CASE CPI has been calculated since August 2015. The project is ongoing and we expect significant conclusions after 2 years of observation. At this point, we present a graph based on 9-month observation period. Green bar on Figure 2 presents an absolute value (in percentage point) of a difference between online CASE CPI and GUS's CPI in a given month. A negative value means online CASE CPI noted stronger deflation or weaker inflation. On the other hand, a positive value means GUS CPI index noted weaker deflation or stronger inflation. As we can see the differences between both indices did not exceed 0.3 p.p. apart from October 2015, in which we noted a significant decreases in online price in the three categories: Food and non-alcoholic beverages, Clothing and footwear, and Restaurants and hotels.

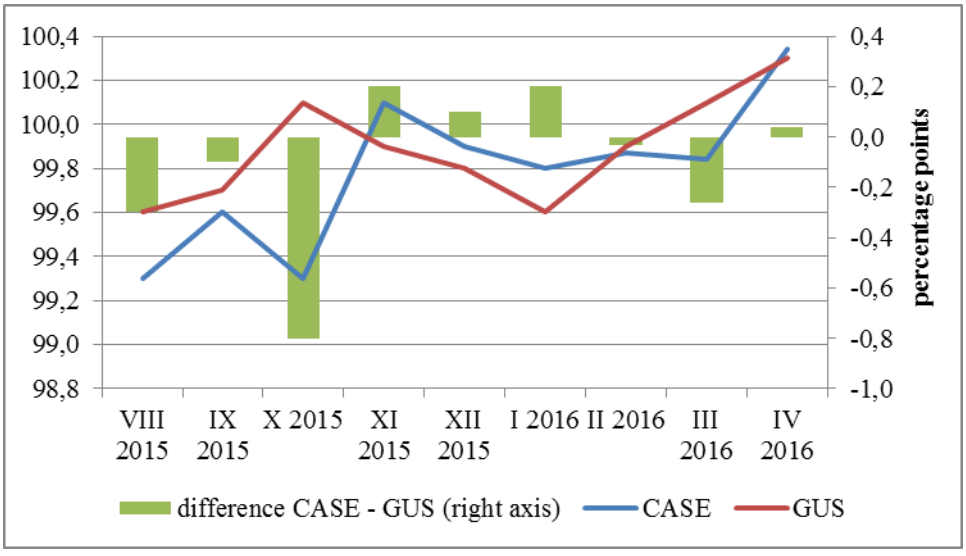

Figure 2. Comparison between GUS's and CASE's indices since August 2015

Moreover, since November 2015 we have published online CASE CPI weekly[3]. For this purposes, we compute an average price level for the last 4 week and compare it with the same average for the 4 week before. In other words, we analyze 4-week moving average each week. GUS estimates CPI on a monthly basis, so a direct comparison is impossible but we can expect some relations that would outrun the changes in prices earlier. Figure 3 presents GUS's and CASE's weekly indices.

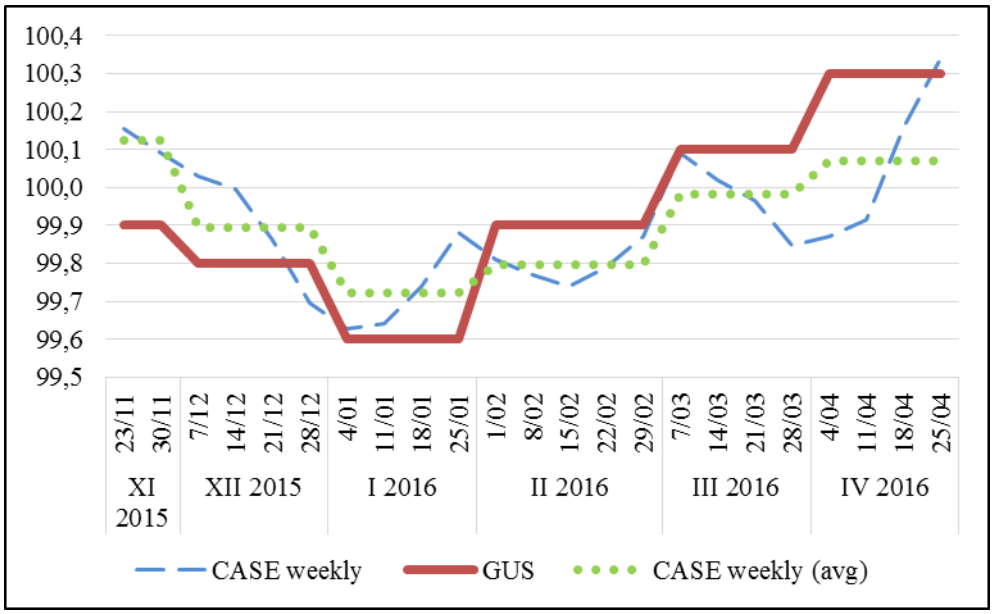

Figure 3. Comparison between GUS's monthly and CASE's weekly indices since November 2015

A blue dashed line shows online CASE CPI counted each week. After a decrease in the end of 2015, the online prices fluctuated rather regularly and local maximums were reflecting an upward trend. A green dotted line is an average of weekly estimates obtained for each month. Both lines are presented in reference to GUS's CPI (red line).

[3] http://www.case-research.eu/

5. Findings and further research

Summarizing, due to a short observation period it is too early to conclude on a correlation over time between both indices. Testing Cavallo's types of relationships: Permanent shift, Cohesion or Different strength is also premature at this phase of the project.

At present, we can highlight the main findings that appeared during our research. First, in terms of time and money spend on data collection, online based indices are more economical. In effect, online CPIs can be measured and published more frequently, with a small time delay. Second, an availability of certain goods and services in the Internet. According to Cavallo (2012), about 60% of the basket that makes-up the CPI is available online. The proportion is different for each country and depends on the distribution of CPI basket and how well-developed the online market is. CASE's CPI includes about 87% of goods and services of a typical household basket (77 out of 93 weighted categories are covered). Third, compared to the traditional CPI measure, more goods than services are available online, or at least they have had a price in the Internet so far. For instance, services like hairdressing, construction services, or a dentist's appointment have a better representation out of the Internet - it mean an overrepresentation of goods in online indices. Fourth, since profiles of online and offline customers are not the same, both online sellers and buyers behave differently in comparison to offline counterparts. Therefore, to analyze online consumption exclusively, the basket of goods and services should stem from virtual markets instead of the official weights from the offline surveys. In effect, online prices are measured according to habits of conventional consumers, instead of using a structure of consumption occurring in the Internet. Such surveys have not been performed in Poland yet.

As previously explained, a consumer price index based on online prices may precede classical measures of consumer prices. This allows one to predict the level of officially announced measures, as well as, to predict the possible reactions of financial and public institutions which use inflation measures in their decision making process. In this sense, the online CASE CPI could be used as an inflation expectations measure. The micro scale of collected data may also allow for investigating different properties of price adjustments in Poland such as price stickiness, frequency, and scope of price changes. Furthermore, we are able to disaggregate our index up to 93 subcategories, which allows us to make analyses of price trends in certain product groups and branches of economy.

Online CASE CPI will be constantly analyzed and developed in order to cover goods and services which will be entering the virtual markets. As an example of so-called *Big Data* this exercise will also allow for the permanent storage of big datasets relating to prices in Poland. Furthermore, alongside e-commerce development around the world, we can expect increasingly importance of inflation statistics based on the online data.

References

Cavallo, A., & R. Rigobon (2010). The Distribution of the Size of Price Changes NBER Working Paper, w16760. Link

Cavallo, A. (2012). The Billion Prices Project: Building Economic Indicators From Online Data, MIT Sloan, Geneva, May 31st, 2012. Link

Cavallo, A., Cruces G., & Perez-Truglia R. (2014). Inflation Expectations, Learning, and Supermarket Prices: Evidence from Field Experiments, NBER Working Paper 20576, November 2014. Link

Cavallo A. (2015). The Billion Prices Project and PriceStats. AEI Conference: The federal statistical system in a Big Data world, March 2015. Link

Ecommerce Foundation (2015). Global B2C E-commerce Report 2015, Ecommerce Foundation. Link

Giles, C. (2015). Supermarket prices fall faster than official inflation measure, Financial Times 2015. Link

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.2974

A Multivariate Approach to Facebook Data for Marketing Communication

Arrigo, Elisa[a]; Liberati, Caterina[a] and Mariani, Paolo[a]
[a] Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Italy.

Abstract
The aim of this paper is to propose a method to explore and synthesize social media data in order to aid businesses to make their communication decisions. The research was conducted at the end of 2014 on 5607 Italian Facebook subjects interested in drugs and health. In this study, we refer to the pharmaceutical market that is characterized by strict legal constraints, which prevent any promotional activities (such as advertising) of companies on prescription drugs. Thus, pharmaceutical businesses tend to promote their corporate brand instead of a single product brand. In such context, social media offer the opportunity to gather customers' information about their attitudes and preferences, helpful to address marketing activities. Through a multivariate statistical approach on Facebook data, we have highlighted the associations existing between TV channels and users' profiles. Therefore, depending on the value proposition to promote, every business could choose, first, the target group to reach and, then, the nearest suitable channel where to develop the corporate brand communication.

Keywords: *Social Media; Business Analytics; Marketing Communication; Facebook; Binary Correspondence Analysis; Pharmaceutical Industry.*

Elisa Arrigo (E.A.), Caterina Liberati (C.L.) and Paolo Mariani (P.M.) share the final responsibility for this paper, however E.A. wrote 1, 2, 5 sections; C.L., 1, 4, 5 sections and P.M wrote 1, 3, 5 sections.

1. Introduction

According to the 49th Censis Report, the percentage of Italian population who accessed a media at least once in 2015 was equal to 96.7% for television, 83.9% for radio, and 52.9% for newspapers. Likewise, Internet users continue to increase, by reaching a penetration rate of 70.9% of the Italian population and a percentage equal to 50.3% pertains to Italians who accessed Facebook, the most famous social network, at least once during the last year. In the light of such evidence, businesses should try to integrate successfully both traditional and social media in order to realize the best marketing communication strategy.

In this study we refer to the pharmaceutical market, that is characterized by strict legal constraints which prevent any promotional activities (such as advertising) of companies on prescription drugs. Thus, pharmaceutical businesses promote their corporate brand instead of the product brand. In such context, social media offer the opportunity to gather customers' information about their attitudes and preferences helpful to address marketing activities. The aim of this study is to propose a method to explore and synthesize social media data in order to aid businesses to make their communication decisions; and more precisely, to orientate their media choices into a marketing communication plan.

2. Theoretical Background

The 21st century has seen a big change in the media landscape due to the introduction of social media and the consequent increase in the variety of channels that businesses can employ to reach their customers. In fact, the digital environment has enlarged the set of communication media that businesses had used for 50-100 years: television, radio, newspapers, magazines and outdoor. Actually, traditional media are not disappeared from firms' media choices and, on the contrary, some of them are experiencing a new boom such as television and radio through the launch of satellite and other digital formats.

Social media have emerged as a new powerful marketing tool useful to achieve business purposes by providing businesses with new ways to communicate, collaborate and share content with customers (Lee, 2014). They encompass a wide range of tools and technologies and can be defined as "a group of internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of User Generated Content" (Kaplan & Haenlein, 2010, p. 61). Although few accepted classifications exist to distinguish them, there are several social media formats and platforms such as blogs, social networks, virtual social worlds, collaborative projects, content communities, virtual game worlds, etc. In particular, social networks are applications in which users create personal profiles accessible to others in the exchange of personal content and communication. Facebook has over one billion registered accounts

worldwide and among these many refer to companies that have created their own Facebook page where customers can sign up to become fans. Social media represent for businesses both an efficient channel to display ads, commercial and institutional communications and to collect data and information on customers' lifestyles, needs and problems encountered with existing products. Social media have empowered customers to publish opinions and spread new content online that is argued to be particularly valuable in business strategies for both marketing communication and intelligence purposes (Lee, 2014; Kietzmann *et al*., 2011). From a marketing communication perspective, the relevance of digital marketplace lies exactly in the interaction between consumers and the online community and in the immediate, interactive and low-cost communications. With a shift towards a multimedia environment, where traditional and digital media are available, the nature of marketing communication model has dramatically changed and businesses have lost the full control on their marketing communications. In the past, marketing communication was created by businesses and pushed towards the customers, which have a passive role of recipient; businesses could control the content, timing and frequency of the communication flow. Instead, nowadays, customers interact online with each other by taking part in conversations about brands and products and freely expressing their opinions and, in doing so, they alter the original marketing communication flow and become themselves a source of communication towards the online community (Fill, 2009). From a marketing intelligence perspective, market research analysts have recognized social media as an excellent base for tracking the behavior of customers. Generally, these latter create a virtual identity and, being aware of the fact that they are unidentifiable, say what they really think and produce information that is considered quite truthful. By continuously scanning social media, firms can acquire users' data at any moment, which greatly improves the availability of information about customer experiences (Bose, 2008) and allows monitoring the customer evolution over a specific period. In fact, digital technologies and access to social media only via login subscription enable businesses to collect enormous quantities of customer data with which to build customer web analytics (Zeng *et al*., 2010). Although social media appear to offer significant opportunities to businesses in terms of customer knowledge acquisition, since the most of social data is free, available in large quantities and in real time, the amount of these data is overwhelming and, consequently, the search of the desired information can be very complex and expensive to find. Then, in this study, we propose a method to reduce the complexity of analysis of social media data in order to orientate the media decisions of businesses within their marketing communication strategy.

3. The Data

The research was conducted at the end of 2014 on 5607 Italian Facebook subjects interested in drugs and health[1]. We collected all the possible interactions among people and brands, products and services (i.e. shares, likes, tweets, pins, posts, etc.). Of course, such huge amount of information could not be handled and processed with standard computing engineering. Therefore, the raw data were stored on a cloud platform with 20 servers active on Amazon Web Services (AWS) infrastructure. More than 5 Terabyte (distributed) database were gathered and updated daily via Hadoop2[2]. In our case, the synthesis process stored information into three main tables: the *Behavioral* Table, that contained each user by Facebook page interactions, the *User Demographic* Table that collected unstructured data about users profiles, the *Pages Demographic* Table that stored unstructured data about Facebook pages (Fig. 1). In each table, records were extracted with queries based on users' keys and behavior.

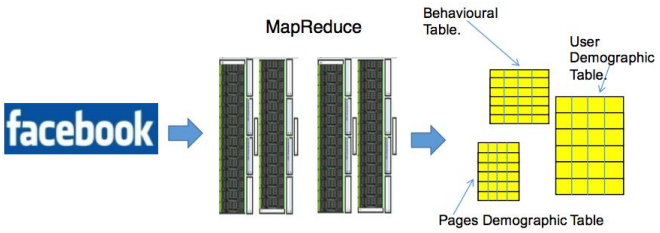

Figure 1. MapReduce Process Flow

Finally, the built matrix had size 5607 rows (Facebook users) and 140 dummy columns (pages visited/liked). In order to reduce the dimensionality of the data, we employed the subjects' classification of Kosinski *et al.* (2013), which distinguish users into 19 alternative psychographic profiles: Pet Lovers, Outdoor Enthusiast, Techies, Car Lovers, Book Lovers, Social Activist, Gamers; Movie Lovers, Politically Active, Sport Lovers, Fashion Lovers, Music Lovers, Travel Lovers, Public Figures Followers, Food Lovers, Home Decorators, Beauty and Wellness Aware, Business People and House-keepers[3].

[1] The research was jointly conducted with Cubeyou, which is a company that delivers customer insights based on Social Media data. We monitored only websites of top pharmaceutical companies and Italian Public health istitutions.

[2] Hadoop is an open-source software designed to handle extremely high volumes of data in any structure. It has two components: 1) the Hadoop Distributed File System (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between and 2) the MapReduce paradigm for managing applications on multiple distributed servers and to perform parallel computations.

[3] The Kosinski *et al.* (2013) approach transformed the unstructured data in meaningful customer information as personality traits, user location, hobbies and interests. After coding process, all the information is synthesized via a singular vector decomposition. The classification of the istances into a specific psychografic group is obatained by means of a logistic regression.

4. Empirical Analysis

Facebook dataset is generally considered a very complex example of unstructured information. Its volume and variety of contents makes impossible to deal with it without considering any manipulation for reducing the size of the data. Before considering any technique or model, the question that should be faced with is: "How do we elaborate this data set?" We run a pre-processing step to squeese the dimensions of the original matrix: we summed users and correspondent likes in order to obtain a contingency table of 19 psychographics profiles and 140 topics (i.e. Celebrites, TV shows, TV channels, Magazines, On-line sources). Due to the fact we focused our attention on the choice of TV channels for addressing a media campaign, we further reduced the number of the columns of the matrix to 20. In order to uncover and visualize the associations between the levels of a two-way contingency table we employed a Correspondance Analysis (Benzècri, 1973; Grenacre, 1984). The technique provides a geometric representation of the rows (psychographic-profiles) and columns (TV channels) as points in a low-dimensional space, according to the chi-square metric. In our case the hypothesis of independence between rows and columns is rejected (χ^2=613.679, p-value=0.000) and the first two factors, retained for our analysis, generate a principal factor (f_1, f_2) plane that explains 68.50% of the total inertia. For sake of brevity, in Tables 1-2 we show only the main statistics relative to those TV channels and profiles that crucially weight in characterizing the two axis. According to the contributions[4] displayed in Table 1, the first factor contrasts *Action type TV* (positive pole) vs *Generalist TV* (negative pole).

[4] Contribution (or absolute contribution) measures the proportion of variance provided by each element (row/column) in explaining a principal axis.

Table 1. Main statistics on selected columns.

Column	Mass	Coordinates		Contributions		Quality of representation	
		Axis 1	**Axis2**	**Axis 1**	**Axis 2**	**Axis 1**	**Axis2**
Real Time	0.204	0.130	0.446	0.028	0.346	0.078	0.877
Sky Sport	0.074	0.653	-0.501	0.259	0.159	0.552	0.310
DMAX Italia	0.068	0.386	-0.203	0.082	0.024	0.659	0.175
Rai.tv	0.068	-0.450	-0.266	0.112	0.041	0.540	0.180
Sky TG24	0.051	-0.302	-0.457	0.038	0.092	0.246	0.537
ALICE TV	0.029	-0.501	0.620	0.060	0.096	0.278	0.406
Laeffe	0.030	-0.490	-0.042	0.058	0.000	0.557	0.004

Table 2. Main statistics on selected rows.

Row	Mass	Coordinates		Contribution		Quality of representation	
		Axis 1	**Axis2**	**Axis 1**	**Axis 2**	**Axis 1**	**Axis2**
Techies	0.050	-0.451	-0.310	0.084	0.041	0.539	0.243
Gamers	0.016	0.879	-0.807	0.104	0.092	0.456	0.367
Politically Active	0.043	-0.740	-0.500	0.194	0.093	0.588	0.256
Sport Lovers	0.069	0.637	-0.437	0.229	0.113	0.631	0.284
Fashion Lovers	0.050	0.146	0.543	0.009	0.125	0.059	0.773
Beauty and Wellness Aware	0.043	-0.036	0.650	0.000	0.154	0.003	0.813
Housekeepers	0.034	-0.125	0.810	0.004	0.190	0.018	0.731

In particular, the variance of f_1 is explained for 34.10% by Sky Sport (0.259) and DMAX Italia (0.082), whereas Rai.tv (0.11.2) and Laeffe (0.052) weight 16.40%. Also the quality of representation[5] of the points that provides additional richness to the interpretation of the relationships in the contingency table, confirm such picture (Tab. 1). The psychographic profiles associated to the factor are coherent with the description provided: Sport Lovers and Gamers show positive coordinates and relevant contributions with the principal axis 1 (Tab.2), instead Politically Active and Techies are located and associated with the opposite side of the same factor. The second axis discloses differences between *Entertaining TV* (positive pole) and *Newcast* (negative pole). It is reasonably clear from the inspection of the contributions that Real Time (0.346), ALICE TV (0.096) on one side, and Sky TG24 (0.092) on the other side, are the relevant elements for interpreting the factor f_2 (Tab.1). The reading of the axis is also upholded by the quality of representation relative to those channels that are always greater than 40%. It is interesting to highlight a coherent association of those results with the psychograpic profiles Fashion Lovers, Beauty and Wellness Aware and Housekeepers. A complete representation of the associations is visualized in Figure 2.

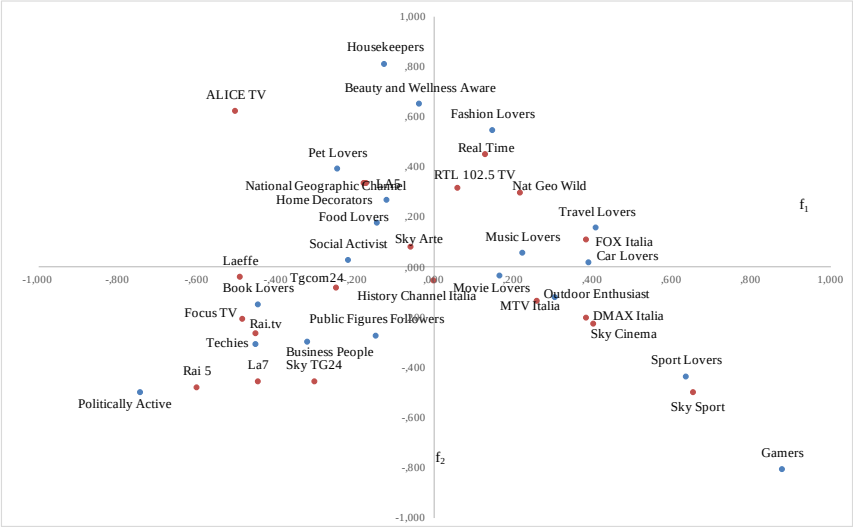

Figure 2. Normalized 2-dimensional plot of psychographic profiles and TV channels

[5] The quality of representation (or relative contribution) measures the proportion variance provided by each principal axis in explaining a single point.

The interpretation of the map is straightforward: the proximities among channels and profiles indicate high similarities (associations). For example, in the first quadrant of the map (where f_1 and f_2 are both positive), we find RTL 102.5 TV, Real Time and Fox Italia characterized by an emotional hedonistic broadcast style. Such result is a useful suggestion for a business that has to select a suitable channel for reaching target audience as Fashion Lovers, Travel Lovers, Music Lovers, Car Lovers.

5. Conclusion

As before illustrated, the four quadrants allow to highlight the associations between channels and users' profiles. Depending on the product, every pharmaceutical business can choose the target group and consequently the nearest suitable channel where to deliver the corporate brand communication. Together with the discussed provided opportunities, it is important to point out limitations of Facebook data: differently from survey data, that are a collection of conscious customers answers, pages without likes are not necessary a users' aware choice. These situations could occur both for lack of users' visits or for a reasoned decision of them. In our study it is likely that the users were exposed to all the TV channels (due to their high popularity). In the light of such considerations, the hope is that the usage of social media, always growing, will be done through a rigorous research design that illustrates gains (and limitations) of the results. Having more data does not necessarily mean having more information, since the knowledge extraction process is not only an automatic computational syntesis. The future research could focus on analyzing other media as magazines or celebrities and other statistical exploration.

References

Benzècri, J. (1973). *Analyse des Donnèes*, Dunod, Paris.

Bose, R. (2008). Competitive intelligence process and tools for intelligence analysis. *Industrial Management & Data Systems*, 108(4), 510–528.

Censis (2015). 49 Rapporto sulla Situazione Sociale del Paese. Milano: FrancoAngeli.

Fill, C. (2009). *Marketing communications. Interactivity, communities and content*. Fifth Ed., London: Prentice Hall.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, New York.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68.

Kietzmann, J. H., Hermkens, K., & McCarthy, I. P. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3), 241-251.

Kosinski, M., Stillwell D. & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *Proceedings of the National Academy of Sciences*, 110, 5802-5805.

Lee, I. (Ed.) (2014). *Integrating Social Media into Business Practice, Applications, Management, and Models*. IGI Global, Hershey, PA: Business Science Reference, USA.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *Intelligent Systems*, IEEE, 25(6), 13-16.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3597

Know your customers from Twitter contacts: automatic discrimination of peer contacts from news sources

Munar, Antoni[a]; Chiner, Esteban[a]
[a]GFT Group, Av Barón de Carcer 48, 46001, València, Spain.

Abstract

Know your customer is a core element of any customer relationship management system for mass service organizations. The emergence of social networking services has provided a radically new dimension, creating a more personalized, deeper, ubiquitous and almost real time relation with customers. At the same time, some of the more widespread social network platforms seem to be evolving not only as social networks between individuals but also as mass information distribution media. When knowing your customer through social networking services, it may be of interest to disambiguate which part of the customer context in the network relates to his peers from other sources. In this paper we present an algorithmic approach to disambiguate one aspect of such relation, as expressed in the nature of the contacts established in the social network: with peers or with organizations, news media or influencers. We focus in the case of Twitter where a simple supervised linear regression can provide a ranking score, effectively discriminating and ordering by closeness peer and other types of contacts (mass media or influencers). Such discrimination can serve as a preliminary step for deeper analysis or privacy protection of customer interaction and is suitable for implementation in automated Big Data systems.

Keywords: *Big Data; Social Mining; Twitter; Social Networks; Know Your Customer; KMeans clustering; Linear Regression.*

1. Introduction

Since the advent of mass production, mass services and mass marketing in the second half of the twentieth century, customer relationship management (CRM) systems have become a core component of the modern company in the race to improve customer satisfaction and retention in an increased competitive environment (Injazz 2003). Such technology applications have required an integration of front office systems (sales, marketing, customer service) and back office systems (financial, operations, etc..) with a set of well defined "customer touch points" (e-mail, direct mail, media marketing, etc..) carrying mostly a transactional (customer – company) or unidirectional (company – mass media) character.

The full emergence of social networking services in the early 2000s (Boyd 2008) supporting a series of globally wide spread social networks has added a radical new dimension. Information about customer preferences, customer-company customer-product and company-competitors is nowadays hidden inside of vast amounts of unstructured social data as blogs, posts in social networks, reviews in collaborative sites or opinions instantaneously expressed in micro-blogging sites, to name a few (Oberhofer 2015). Companies have been increasingly turning to the most widespread social networks (Facebook, Twitter, etc…) to engage with customers trying to integrate the anonymous social network information into their processes and operations (Heller 2011), to gain a 360º view of their customers, in what is known as "know your customer" (Bielski 2001). This integration has consisted mainly in engaging in conversations of different type (customer complains, brand promotion and company news), reputation alerts and sentiment analysis and community analysis of opinion leaders and influencers (Oberhofer 2015, Zhang 2015).

However, it has been stated that some of these Social Network Services (*e.g* the microblogging platform Twitter) act more as information networks that social networks (Myers 2014), being consistenly used by customers increasingly as a source of news of events outside the realm of direct acquitances and relatives (Mitchell 2015). These news can be originated and propagated either from peers or directly from media or companies with Social Network engagement.

In this paper we present a machine learning procedure allowing in a Social Network like Twitter, discriminate and rank by a perceived degree of "closeness" which contacts of given customers in the social network are most probably news sources (media or influencers) rather than peers from their most inmediate circle. Such discrimination can provide a first disambiguation of the community context of the interacting customers ("external" news media from peer circle) in a scalable and prompt way, avoiding biases in subsequent analysis. Also, such automated procedure can prove itself very appropriate as a first step of automated analysis of customer social network communities, where the computational complexity and amount of information needed for the analysis of the

community graph (relations, sentiment analysis, natural language processing, theme detection) are greatly complicated by the presence of news media or extremely popular contacts, due to the great activity and number of subsequent contacts.

This paper is organaized as follows. Section 2 describes some relevant aspects of the type of data used from the Twitter public application program interface (API). Section 3 describes the developed KMeans algorithm together with the linear regression model. Section 4 presents the obtained results. Section 5 discusses the validation of both models. Finally, the conclusions summarizes the obtained results.

2. Twitter Data

The public open Twitter API (REST APIs 2016) provides information about any Twitter account that has decided to make public the account. The Twitter Social Network Service can be considered as a directed graph (Myers 2014) where each account represents a node. The graph is directed because the relation (edges) between two nodes are asymmetrical. In Twitter terminology, Node B is said to be a friend of node A if A is subscribed to receive all the posts of B. On the contraire, node B is a follower of A, if B is subscribe to receive all the posts from A. For the account A the degree of the outgoing edges is called the number of "friends". The degree of incoming edges is called the number of "followers". For the purposes of this paper, for a given account only the number of followers and friends is retrieved. For model training and evaluation purposes, the twitter id (an arbritary combination of letters and numbers that can be used to access the timeline of posts of the account through the public Twitter web page) is also retrieved.

3. Model

The topological structure of the graph formed by the Twitter accounts follows a power law (Myers 2014). This means that some of the accounts have an exponential big number of followers (the celebrities or news media) while a very long tail of "normal" users will have a reduced number of both followers and friends. 95% of Twitter accounts have less than few hundreds of either friends or followers, with usually a bigger number of friends –*i.e.* accounts that they are following- than followers. For the 50% quantile, the number of friends is 39 while the number of followers is more than half: 26 (Myers 2014). The reduced number of contacts of normal people (compared with influencers) fits with the hypothesis that current people has limited resources to absorb and emit information, roughly of the same order, with a higher tendency to absorb than emit. Another factor is the reciprocity. If I follow you, and you are my real friend, most probable you will follow me as well. On the contraire, the accounts of celebrities, mass media and influences show a different structure. Usually an extremely number of followers, in occasions with comparable number of friends, if they choose to reprocicate or not. Another type is bots, spammers or marketers, with high number of friends, but very reduced number of

followers. Despite the filter implement in Twitter (can not be a friend of more than 2000 accounts if more than 2200 accounts do not follow you) they may play a role. Because of its straightforward interpretation and availability the number of followers and friends, together with the reprocitiy –considered here as a categorical value- will be used as variables.

The discrimination of the contacts of a given account can modeled as a binary classification model: peer or mass media (or celebrity). For such classification, we study two different methods: KMeans clustering, which is an unsupervised method and linear regression, that is a supervised method requiring training. The value of the regressand variable of the linear model can also serve as an effective ranking score to order the contacts by perceived "closeness". Both methods can be easily implemented and parallelized in Big Data Systems. Unsupervised methods have the advantage that do not require training and are based only in some very general heuristic principles. Supervised models require training but can potentially lead to more precise results.

3.1 Unsupervised model: KMeans

KMeans (Kanungo 2002) is one of the more populars algorithms due to its simplicity and straightforward interpretation. KMeans tries to find clusters with centroids such that the mean square distance from each data point to its nearest centroid is minimized. As attributes we consider the difference of number of friends between the contacts of the prospect and the number of contacts of the prospect itself $d_{fr} = n_{fr}^{c} - n_{fr}^{p}$ together with the difference beteween the number of followers of the contact n_{fl}^{c}, and of the prospect $d_{fl} = n_{fl}^{c} - n_{fl}^{p}$. The prospect itself will have then $d_{fr} = 0$ and $d_{fl} = 0$. One meta-parameter of the KMeans algorithm is the *a-priori* number of clusters, that must be set ahead based on heuristic considerations. In our case, the "elbow method" which selects the number of clusters with the minimum total variance, is used (Kanungo 2002).

3.2 Linear regression

Linear regression for binary classification suffers from several shortcomings (Agresti 2011), namely heterocedascity of the residuals (wich causes difficulties interpreting the confidence intervals), unrealistic constant regression factors and the fact that the predicted regressand variable can be either bigger than 1 or negative, which is at odds with the interpretation of the regressand as a probability. However, they are sometimes used when what we want is simply to build a rank which values indicate some tendency and we are interested mainly in the extreme cases (contacts representing news media or celebrities are expected to be quite different from normal accounts due to the power law).

In our model the rank variable y that can take values of 0 (peer contacts) or 1 (news media or celebrity) is modeled as a linear combination of: the categorical variable *reciprocity* with takes the value 1 when contact and prospect are followers and friends of each other (0 otherwise) and an followers-friends term composed of the product of the difference of number of friends between the contacts and the prospect $d_{fr} = n_{fr}^{c} - n_{fr}^{p}$ multiplied by the difference of the number of followers between the contacts and the prospect $d_{fl} = n_{fl}^{c} - n_{fl}^{p}$. This interaction term between variables tries to capture the reinforced effect of similar number of friends and followers.

$$y = C + B \cdot rec + A \cdot d_{fl} \cdot d_{fr}$$

4. Model Fit

4.1 KMeans Clustering

25 individuals with Twitter accounts were selected and asked to classify if they contacts were people that they personally know or celebrities or sources of news. Mass media or firm accounts are usually easy to identify because of the explicit name and content of the posts. Another 10 accounts were selected for model validation purposes.

Figure 1 shows the result of clustering for the contacts of a single individuals. The within cluster variance plot shows that the optimal number of clusters is three. The cluster containing the prospect (cluster No 1) is composed of contacts with almost the same number of friends and followers, and its clearly separated from two other clusters, where its components have a very high number of friends or followers. Therefore, by choosing this clusters one could discriminate between peer contacts and news or other contacts. The high variance in the number of followers inside the cluster of the prospect is due to the fact that the cluster comprises not only peer contacts.

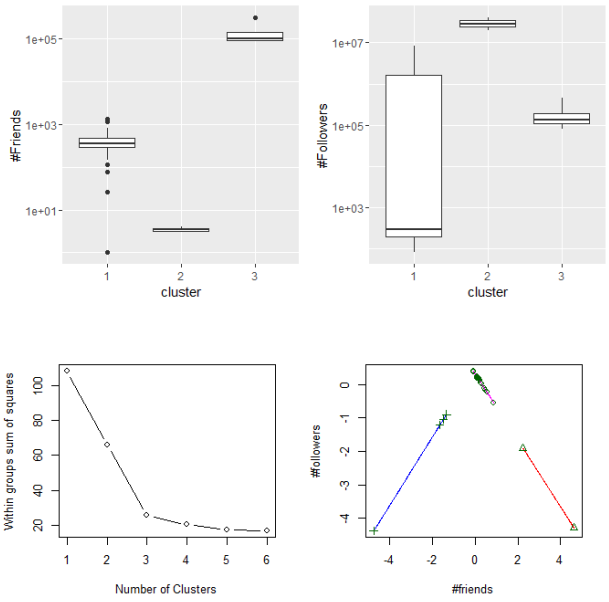

Figure 1. Top left: box plot of the number of firiends for the three different clusters. Top Right: box plot of the number of followers for the three different clusters. Bottom left: sum of within clusters variance as a function of the number of clusters. Botton right: cluster found the Kmeans algorithm. Circles correspond to cluster No 1, crosses to cluster No 2 and triangles to cluster No 3.

4.2 Linear Regression Model

Table 1 shows the results of the fit to the linear regression model with a set of contacts of different individuals. A set of 25 individuals with several tenths of contacts each where asked to indicate whether the contacts were peers or contacts from which they obtain information out of their inner circle. All the parameters with the exception of the intercept

Parameter	Estimate	Std. Error	t-score	p-value
Intercept C	0.044	0.039	1.124	0.266
Reciprocity B	0.84	0.05	16.777	<2e-16
Followers-Friends A	0.15	0.024	6.147	1.13e-07

Table 1. Estimated values for the linear regression model.

are statistical significant. As it could be expected, the reciprocity is a very strong factor in discriminating between peers, accounting for the most part of the discrimination. The followers-friends plays a residual role usefull to discriminate in the case when, *e.g.* a celebrity or organization systematically reciprocicates contacts. Figure 2 shows fit residuals

that as expected for a linear regression classification present non-gaussian tails. As in the case of the clustering, contacts classified as peers show reduced and roughly equal number of friends and followers with small variance. Other regression classifiers like logistic regression do not yield statistical significant values for the coefficients due to the huge range of values that the variable followers-friends can take for one of the classes.

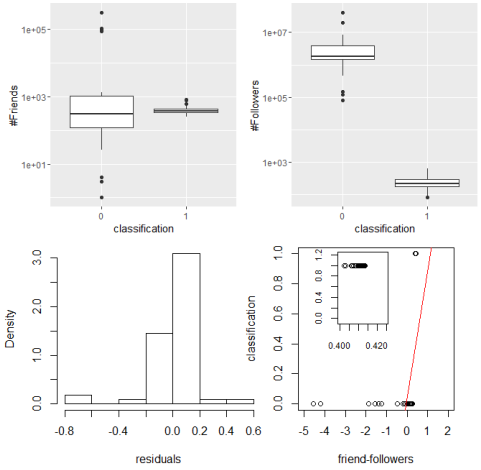

Figure 2. Top left: box plot of the number of friends for peer contacts (classification 0) and others (classification 1). Top right: box plot for the number of friends. Bottom right: fit residuals. Bottom left: classification versus friend-followers, points are data, red line is the regression line. The inset expands the range for some data points.

5. Model validation

To validate the model, 25 individuals of similar characteristics of those used in the training where asked to annotate their contacts as peers or information sources. The reduced sample is due to the high cost to obtain such sample because different individuals are personally requested to evaluate the results. Two kinds of validations can be performed. First, building the corresponding confusion matrix (Table 3).

Method	Precision	Recall	F-Factor
Clustering	0.94	0.75	0.83
Linear Regression	0.95	0.81	0.87
Reciprocity cut	0.96	0.80	0.90

The results show that precision and recall are quite high. Linear regression performance is quite similar to a simple cut based in the reciprocity of the contacts, that is used to asses the

real discrimination power of the method. Both methods yield slightly better resuls than the clustering method. However, the linear regression method offers the possibility to rank the obtained results, so results can be ordered by "closeness" quite in the same ways as in other information retrieval systems (Lopresti 1998). When users were confronted with such ordering, usually expressed that the firsts results in the ranking were the most closest to their inner circle, while the results with the lowest ranks usually were newspapers or companies, with a perceived utility in such ranking.

6. Conclusions

For mass service oriented organizations interaction with their customers in the social networks services like Twitter, Facebook, etc… is becoming a core component of their customer relationship management. Big Data technologies allow for the massive analysis of the data generated during customer interaction. In this paper we have investigated an scalable automatic algorithm for one aspect of such interaction. Namely, to automatically disambiguate which part of the customer contacts in a social network relate to his peer circle and which part to other information sources used by the customer. Despite the reduce sample used in the evaluation, results show that a linear regression model based in robust observable variables like the number of contacts of each of the contacts of the customer itself can provide a ranking score automatically discriminating which contacts are peers and which ones other sources. The performance is similar to more straightforward rule based methods, but the linear regression offers a ranking that can be interpreted as a perceived "closenesses" similar to other information retrieval methods. Furthermore, the regression model can be further enriched with futher information. The obtained results show that this model could potentially be implemented at large scale yielding significant results.

Acknowledgements

The authors wish to thanks the colleagues from the GFT Data Practice Ignacio Sales, Sergio Gomez and Angel Rey for their enlighting discussions and support.

References

Agresti, A. & Kateri A.. (2011) *Categorical data analysis*. Springer Berlin Heidelberg.

Bielski, L. (2001). Giving your customer a face. *American Bankers Association. ABA Banking Journal*, 93.4, 49-53.

Boyd, D. & Ellison N. (2008). Social Network Sites: Definition, History and Scholarship. *Journal of Computer-Mediated Communication*, Vol 13, 210-230.

Heller C. & Parasnis G. (2011) From social media to social customer relationship management", *Strategy & Leadership*, Vol 39, No 5, 30-77

Injazz J. C., & Popovich, K. (2003). Understanding customer relationship management (CRM). *Business Process Management Journal*, Vol 9, No 5, 672-688.

Kanungo, T., Mount D. *et al*. (2002) "An efficient k-means clustering algorithm: Analysis and implementation." *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.7*, 881-892.

Lopresti, D., & Zhou. J. (1998) Document analysis and the world wide web. *Series in Machine Perception and Artificial Intelligence* 29 479-502.

Mitchell A. & Page D. (2015) The evolving role of news on Twitter and Facebook. *Pew Research Center.*

Myers S., Sharma A. *et al.* (2014) Information network or social network?: the structure of the twitter follow graph. *Proceedings of the 23rd International Conference on World Wide Web*. ACM New York, 493 – 498

Oberhofer M., Hechler E. *et al* (2015) *Beyond Big Data. Using Social MDM to Drive Deep Customer Insight*. Pearson plc publishing as IBM Press.

REST APIs, Twitter Developers (2016) https://dev.twitter.com/overview/api/users

Russell, M. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. O'Reilly Media, Inc.

Zaharia, M. , Chowdhury M. *et al*. (2012) "Fast and interactive analytics over Hadoop data with Spark." USENIX; login 37.4: 45-51.

Zhang L., Zhao, J.*et al* (2015). Who creates trends in online Social Media: the crowd or opinion leaders. *Journal of Computer-Mediated Communication*, Vol 21, 1 1-16.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3110

Effects of colored contrast of mobile websites on behavioral intentions

Pelet, Jean-Eric[a] and Taieb, Basma[b]
[a]Department of Marketing, KMCMS.net, ISC Business School, University of Nantes, France, [b]Department of Marketing, University of Cergy Pontoise, France

Abstract

This study examines the effects of the mobile-phone website colored contrasts and the affective states of the consumer (emotions and moods) and trust respectively on intention to revisit, buy on and recommend the mobile website. For this purpose, a factorial plan 2x2 was developed and a mobile website, with two different alternatives, was designed especially for the experiment: positive contrast (yellow text on green background) and negative contrast (green text on yellow background). The research was conducted on French consumers. 312 valid responses were collected through online and personal survey questionnaires. Data was analysed using the method of structural equations. The results show the significant effects of mobile website's color contrast on behavioral intentions. Perceived dominance and trust towards the website have positive effects on behavioral intentions, whereas mood has non-significant effects on behavioral intentions. Managerial implications are discussed.

Keywords: *m-commerce, color, contrast ratio, dominance, trust, behavioral intention.*

1. Introduction

Reading through the screen of a smartphone in the street can sometimes prove to be a nightmare, especially when the sun shines, and when legibility conditions are not very good. Brands that don't pay attention to the ease of reading the content of their m-commerce (mobile commerce) website may lose consumers' intentions to buy and revisit as well as their intentions to recommend the website. To that extent, the contrast ratio created by the foreground (text) and background colors of the website need to be carefully designed. As already mentioned by Pelet (2014) regarding e-commerce, color contrast of the website enhances the memorization of information and the intention to purchase (Pelet & Papadopoulou, 2010, 2012). Especially in the context of the Web design, color serves as a support to the layout of the information. Hence, it works as a contrasting base. Reading time is shorter when there is good contrast. Also, reading is more enjoyable due to the correlation with improved legibility. Some studies show that a positive contrast obtained with a light text over a dark background provokes online visitors to abandon a website and puts users in a negative mood (Pelet, 2010). A positive contrast inflicts a visual strain accompanied by eye fatigue, contrary to a negative contrast (a dark text over a light background). This is why we choose to distinguish the background and foreground colors in the experiment we will present hereafter, by using both negative and positive contrasts to conduct our measurements. The positive contrast being yellow on green, and the negative one being green on yellow, following Hill & Scharff's (1997) findings.

This paper presents a literature review on the importance of the colored contrast of mobile website interfaces, placing importance on emotions and dominance in particular, the mood and trust of users regarding the website. The methodology section then presents the website built for the experiment, followed by results aiming at highlighting the behavioral intentions derived from positive and negative colored contrasts exhibitions. We conclude the paper with a discussion, some limits and future research pathways, affording some managerial implications.

2. Literature review

2.1. Controlling the Contrast Ratio of Mobile Website's Interfaces

Initial feelings of the users are so crucial because during the first 50 milliseconds they decide whether to continue browsing the website or not (Lindgaard *et al.*, 2006). The aim of designers is therefore to provide interactive systems that are not only useful and usable, but are also capable of transmitting positive emotions through their aesthetic characteristics.

Practitioners are advised to emphasize strong contrast between the foreground and background, especially for text messages. An important factor is appearance, especially proper choice of font and color contrast, which creates reliability and enhances e-service

quality or perception if the website is user friendly (Lowry *et al.*, 2014). Thus, we can hypothesize:

H1. The color contrast of the mobile website positively influences behavioral intentions

2.2. Emotions and Mood's Effects on Behavioral Intentions

When consumers buy on the Web, they fulfill certain tasks that can elicit responses in both cognition and affect. These responses can determine the intention of the consumer to return to the site (Koufaris, 2002). In the context of electronic commerce, more specifically, the perceived control or the feeling of dominance, has been studied as an experiential factor that can influence the attitudes and behavioral intentions of users (Mazaheri *et al.*, 2011). Hence, we assume the following hypothesis:

H2. The perceived dominance when visiting the mobile website positively influences behavioral intentions

Previous research has indicated that people in different mood states tends to process information differently (e.g. Kim *et al.*, 2015). Indeed, people in a positive mood avoid spending cognitive effort whereas people in a negative mood process information more centrally. A good mood may contribute to consumers' intention to return to a website at which they have made previous purchases (Wu *et al.*, 2008). Therefore, we can hypothesize:

H3. The mood of the consumer when visiting the mobile website has an influence on behavioral intentions

2.3. Online Trust Induces Better Behavioral Intentions

The lack of trust is one of the reasons most frequently cited by consumers who do not buy on the Internet (Mukherjee & Nath, 2007). Previous research has shown that trust towards the website can have a significant and positive influence on behavioral intentions of consumers in terms of repeat visits, purchase intention, and positive word-of-mouth (Mukherjee & Nath, 2007). We assume the following hypothesis:

H4. Trust towards mobile website positively influences behavioral intentions

Coming from these hypotheses, we propose the following conceptual model:

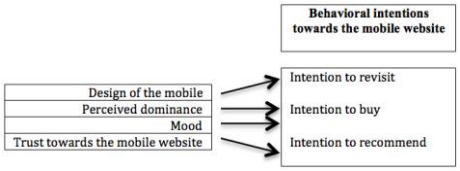

Figure 1. Conceptual model

3. Methodology

3.1. Conditions of the Survey, Stimuli and Data Collection

Respondents answered the questionnaire in the real-life, outdoor conditions of using a smartphone. For this reason, respondents were tested for color blindness using the Ishihara test[1] (Pelet, 2010). This guaranteed the validity of our sample's responses, by keeping people with perfect color vision. In order to control environmental variables – ambient lighting[2], humidity[3] and temperature[4] were also measured during each respondent's visit. These measurements enabled us to neutralize variables and were reported on each questionnaire of each respondent.

Two versions of a mobile website specializing in the sale of music CDs, have been specially designed for this experimental study: a site with a positive contrast (yellow text on a green background) and another with a negative contrast (green text on a yellow background). The experimental site contains 30 CDs available in 10 categories (3 CDs / category). For each CD, participants could see the CD cover, the album title, the artist name and other details (music category, status (new or used), price, time delivery...).

312 participants visited the site's two versions: 160 participants for the site with positive contrast and 152 participants for the site with negative contrast. Each participant had to view the details of at least two CDs of their choice and add them to their cart without making real purchases. After viewing at least two CDs, a link appeared asking them to respond to a questionnaire.

The questionnaire includes measures related to site design (5 items borrowed from Bressolles, 2006), perceived dominance (5 items of Mazaheri *et al.*, 2011), mood (5 items borrowed from Mayer & Gaschke, 1988) trust towards the website (7 items of Chung & Shin, 2009), intention to revisit (3 items of Mukherjee & Nath, 2007), purchase intent (4 items of Limayem & Rowe, 2006) and intention to recommend (3 items of Goyette*et al.*, 2010).

3.2. Validation of Measurement Scales

The measurement scales have a good internal consistency (Cronbach α and ρ of Jöreskog are above 0.7). The variance extracted (AVE) by items is greater than 0.5. The variance

[1] The Ishihara test is the most common color-blindness test used today (Deeb & Motulsky, 2011). It consists of a number of plates, 24 or 38, the Ishihara plates, each containing a circle of dots which appear in random colors and sizes. Within the circle there are dots forming a number which should be clearly visible to viewers with normal color vision and hard to see or invisible to viewers with defective color vision. To pass the test participants should recognize the number in every plate.

[2] an interval between 300 lux and 500 lux was fixed

[3] a maximum of 65% humidity was retained

[4] data was collected when the temperature ranged from 19 to 29°C

extracted by each construct is greater than the squared correlations among constructs (Fornell & Larker, 1981), thus discriminant validity was supported. The indices for the measurement model also indicate a good fit (RMSEA = .08, RMR= .03, $\chi^2/df = 3.27$, $p < .01$).

3.3. Hypothesis Testing

The hypotheses were tested using the method of structural equations (AMOS). Their validity has been verified for mobile websites with positive and negative contrast. The results are presented in Table 1 below.

Table 1. Test of the hypotheses

	Site with negative contrast (Green text on a yellow background)				Site with positive contrast (Yellow text on a green background)			
	Estim	S.E.	C.R.		Estim	S.E.	C.R.	
Revis<--- Desig	.10	.04	2.73**	H1 partially supported	.23	.03	7.00**	H1 supported
Buy<--- Desig	.02	.04	0.36		.20	.04	5.30**	
Recom<--- Desig	.16	.04	4.40**		.17	.03	5.21**	
Revis<--- Dom	.33	.05	6.35**	H2 supported	.17	.05	3.79**	H2 partially supported
Buy<--- Dom	.22	.05	4.12**		.24	.05	4.57**	
Recom<--- Dom	.11	.05	2.49*		-.03	.04	-.71	
Revis<--- Mood	-.19	.11	-1.79	H3 non validated	.05	.08	.64	H3 non validated
Buy<--- Mood	-.19	.12	-1.58		.10	.09	1.04	
Recom<--- Mood	-.06	.10	-0.57		.09	.09	1.06	
Revis<--- Trust	1.07	.08	12.86**	H4 supported	1.19	.09	12.95**	H4 supported
Buy<--- Trust	1.15	.09	12.19**		.98	.09	10.44**	
Recom<--- Trust	1.03	.08	13.32**		1.2	.09	13.76**	

Note: ** $p < .01$; * $p < .05$

Results show that the website's positive contrast has a significant impact on purchase, revisit and recommendation intentions. Trust towards the website has very significant effects on behavioral intentions regardless of the positive or negative contrast. The perceived dominance has an impact only on purchase and revisit intentions when the

contrast is positive. No significant effect was found for the mood regardless of contrast. Unlike the site with positive contrast, the website's negative contrast does not influence purchase intent. Finally, the dominance has an impact on the intention to recommend when the contrast is negative.

4. Discussion, Limits and Future ways of Research

Our results show that the color contrast of the m-commerce website, in particular a positive contrast, strongly influences the consumer's behavioral intentions. It should also be noted that the consumer would be more inclined to buy from the mobile website with a positive contrast (in our case, a yellow text on a green background). On a mobile website, in the outdoor conditions of the experimentation, respondents have better behavioral intentions when the contrast ratio is a bright text over a dark background, which is the opposite of most actual mainstream websites, which use dark texts (such as a black ones) over bright backgrounds (such as white ones).

In conditions where the sun shines and where it is often quite difficult to read on the smartphone screen, users seem to prefer positive contrasts. Interestingly, this contrast is exists in any smartphone as an "accessibility" option, in the "View" folder of the "Settings" section when pressing the "Negative Colors" button. However, it is not very easy to set up, especially when the lighting conditions are difficult, such as when the sun shines. An initial managerial implication of this research could therefore be to make this service available when the image sensor of the smartphone detects the sunshine and switches it on when it reaches a certain level. This would enhance the legibility of the screen and preserve users' eyes at the same time. A second managerial implication that comes from this finding is that it could save the smartphone battery somewhat. It uses more energy to show a bright screen than a dark one. This is why our smartphone turns off relatively quickly when we don't touch its screen: in order to preserve the battery life, the more it shows a black screen, the less energy it uses. A website that is predominantly dark colors rather than bright colors may help to save the battery life and therefore preserve the environment: it becomes an "eco-friendly" strategy for the brands which will work in this way and make users "Socially Conscious Consumers" as depicted by Brooker (1976). The second result is in line with the preliminary one, showing that the perceived dominance when visiting the m-commerce website positively influences behavioral intentions: a user who can easily read the screen may feel more confortable and by extension have a greater feeling of dominance with the m-commerce website's interface. The user knows exactly what he can do with the website he is viewing, because everything seems clear to him. Such a feeling reinforces the importance of ergonomics where the content must be easy-to-use and agreable to read and understand. We finally found that trust in the mobile website positively influences behavioral intentions. These results support previous work in e-commerce (e.g. Pelet & Papadopoulou, 2012).

Our study contributes to a better understanding of consumer behavior using mobile websites, nevertheless, some limitations may be noted. Consideration of other variables such as feelings of privacy and perceptions of waiting time could provide more details on the purchasing behavior of people with sight problems. While maintaining constant levels of brightness and saturation, which are two components of color (along with hue) to study the overall effect of color on behavioral intentions without distinguishing the effect of text color from background color is one limitation of this research. Future studies are advised to consider others colors and to examine separately the effect of background color, and text color, as well as the interaction between the two. Moreover, there's still no agreement on whether human reactions to color are innate or are learned and can be conditioned by each individual's social experience (Crozier, 1999). Cultural aspects must also be taken into consideration in as far as we perceive color differently according to our origins.

Whereas smartphones continue to offer more and more graphic possibilities thanks to their ever growing power, enabling them to show more than 16 billion colors, our eyes enable us to disinguish only 8 billion (Chrisment *et al.*, 1994). Among the latter, it is easy to create contrast ratios enabling users suffering from color-blindness, as well as other sight problems, using free and easy-to-use tools (for a list of them, please visit http://www.scoop.it/t/color). As already described by Pelet & Papadopoulou (2012), by taking into account the W3C and Web Accessibility Initiative (WAI) guidelines[5], the use of color becomes more professional and the choices web designers make are more informed in terms of usability, as well as in terms of human-computer interaction in general.

References

Bressolles, G. (2006). La qualité de service électronique: NetQu@lProposition d'une échelle de mesureappliquéeauxsitesmarchandseteffetsmodérateurs. Recherche et Applications en Marketing, 21(3), 19-45.

Brooker, G. (1976).The self-actualizing socially conscious consumer. Journal of Consumer Research, 3(2), 107-112.

Chrisment, A., Durchon, P., Lanthony, P., & Tavernier, I. (1994). Communiquer par la couleur - Mesurer, Reproduire, Observer, Vivre la couleur. Paris, 3C Conseil.

Crozier, W.R. (1999). The meanings of colour: preferences among hues. Pigment and Resin Technology, 28(1), 6-14.

Deeb, S.S. & Motulsky, A.G. (2015). Red-Green Color Vision Defects. retrieved from the Internet at http://www.ncbi.nlm.nih.gov/books/NBK1301/October 12, 2015.

[5]The Web Access Initiative (WAI) works with the W3C (World Wide Web Consortium) working groups to address and improve accessibility within specifications and W3C technologies.

Fornell, C., & Larcker, D.F. (1981). Evaluatingstructuralequationmodels with unobservablevariables and measurementerror. Journal of Marketing Research, 18(1), 39-50.

Goyette, I., Ricard, L., Bergeron, J., & Marticotte, F. (2010). e-WOM Scale: word of mouth measurement scale for e-services context. Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration, 27(1), 5-23.

Hill, A., & Scharff, L.V. (1997). Readability of websites with various foreground/background color combinations, font types and word styles.In Proceedings of 11th National Conference in Undergraduate Research, 2, 742-746.

Kim, H., Choi, Y., & Lee, Y. (2015). Web atmospheric qualities in luxury fashion brand websites. Journal of Fashion Marketing and Management, 19(4), 384-401.

Koufaris, M. (2002). Applying the technology acceptance model and flow theory to online consumer behavior. Infomation System Research, 13(2), 205-224.

Limayem, M., & Rowe, F. (2006). Comparaison des facteursinfluençant les intentions d'achat à partir du Web à Hong Kong et en France: influence sociale, risques et aversion pour la perte de contact, Revue Française du Marketing, 209(4/5), 25-48.

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention Web designers: You have 50 milliseconds to make a good first impression. Behavioral Information Technology 25(2), 115-126.

Lowry, P.J., Wilson, D.W., & Haig, W.L. (2014). A picture is worth a thousand words: Source Credibility Theory applied to logo and website design for heightened credibility and consumer trust. International Journal of Human-Computer Interaction,30(1),63-93.

Mayer, J.D., & Gaschke, N.Y. (1988). The experience and meta-experience of mood, Journal of personality and social psychology, 55(1), 102-111.

Mazaheri, E., Richard, M.O., & Laroche, M. (2011). Online consumer behavior: Comparing Canadian and Chinese website visitors. Journal of Business Research, 64(9), 958-965.

Mukherjee, A., & Nath P. (2007). Role of electronic trust in online retailing: A re-examination of the commitment-trust theory. European Journal of Marketing, 41(9/10), 1173-1202.

Pelet, J.-É., (2014). Investigating the Importance of Website Color Contrast in E-Commerce: Website Color Contrast in E-Commerce. In Khosrow-Pour, M. (2015). Encyclopedia of Information Science and Technology, Third Edition (10 Volumes). Hershey, PA: IGI Global.

Pelet, J.-É., & Papadopoulou, P. (2012). The effect of colors of e-commerce websites on consumer mood, memorization and buying intention. European Journal in Information Systems, 21(4), 438-467.

Pelet, J.-É. (2010). Effets de la couleur des sites web marchands sur la mémorisation et sur l'intention d'achat. Systèmes d'Information et Management, 15(1), 97-131.

Wu, C.S., Cheng, F.F., & Yen, D.C. (2008). The atmospheric factors of online storefront environment design: an empirical experiment in Taiwan. Information & Management, 45(7), 493-498.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3643

The impact of Internet on the artist reputation

Castelló, Daniel; De la Poza, Elena; Guadalajara, Natividad
Center of Economic Engineering, Faculty of Business and Management, Universitat Politècnica de València, Spain.

Abstract

This work analyzes the relationship between the digital information provided by the search engine Google on the Internet and the volume of sales or revenues in the art market in order to quantify the variable artist's reputation. In particular, the usefulness of digital media information is analyzed to interpret past sales in the art market or, conversely, to estimate future sales in the short term. Finally, we study the ability of digital information to explain the volume of activity (number of lots sold) or what is the same the degree of liquidity of the secondary market of artworks.

Keywords: *Art market; Internet; artist's reputation; lots, revenues.*

1. Introduction

In the art market all agents (artists, galleries, auction houses, dealers…) involved make use of the double use of Internet as a channel of information and communication, and to support their activity.

Internet has supported and dynamic the secondary art market, promoting its growth through online auctions. Major international auction houses such as Christie's and Sotheby's expose their collections online before each session, or even carry out some of these sessions online (Adam, 2002), and their purchases are collected in large digital databases, providing transparency to the art market (Beaux Arts Magazine, 2007).

In fact, recent studies such as Dass et *al*. (2011) and Schlaegel (2015) analyze changes in auction prices of works of art in the field of online auctions; those researches identify which variables influence in order to improve artworks pricing models and generate bidding strategies for the buyer of works of art in the secondary market.

But not only the secondary market for art has been affected by the Internet, the artists use it to inspire and develop their creative abilities (Hansen, 2006), as well as to promote and sell their work from their own blogs, from social networking (MySpace, Facebook, LinkedIn) (Wikberg, 2014) or from websites like eBay or Craiglist (Grant, 2008; Canals-Cerdá, 2014). This continuous and sustained development has meant that these sites acquire the role of a source of information for analyzing market prices of artworks (Deng et al, 2014).

Studies on valuation of works of art, both in the academic field as in traditional media (newspapers, magazines, etc.) are numerous. In most of these, the value of the artworks is estimated by building mathematical models focused on their physical and artistic characteristics, the place of sale of the work, the date of sale, the nationality of the artist and the artists ‘number of exhibitions. On contrast, the artist's reputation which is difficult to quantify (Ursprung and Wiermann, 2008) has not been included in these analysis, but clearly influences the value of works of art and, therefore, in the art market. Authors like Anderson (2004) analyzes the reputation of the museum through variables results or "outcomes" resulting from the activity of the museum, such as the number of results that appear in the Google search engine, the number of visits to the museum, or the economic impact of the museum measured through the turnover generated by tickets to the museum.

Caballer and De-la-Poza (2005), Guadalajara and De-la-Poza (2007) and De-la-Poza, *et al*., (2009) included in their valuation models the artist's reputation, measured by the media digital information, and studying the influence of turnover and average prices of the auctioned works of each artist in those media.

In this work we analyze the relationship between the digital information provided by *Google* and the volume of sales or revenues in the art market in order to quantify the variable artist's reputation.

In particular, the usefulness of digital media information is analyzed to interpret past sales in the art market or, conversely, to estimate future sales in the short term. Finally, we analyze the ability of digital information to explain the volume of activity (number of lots sold) or what is the same the degree of liquidity of the secondary market of artworks.

2. Methods

2.1. Sources of information

Search engine Google was selected as primary source of digital information. Its selection was based on the volume of queries conducted throughout 2006 by Internet users worldwide. In 2015 Google continues ranking the list of the most used search engine with a market share of 75.2% followed by Yahoo with a market share of 10.4%.

Also, the database Artprice was employed as a source of information on the art market. Artprice provides information about the sale prices of works traded in the international auction houses since 1997. The choice of ArtPrice compared to other similar databases is due to the extensive information provided on each work and artist as well as the rapid updating of information and, most importantly, by the possibility of online research.

2.2. Database construction

The collection of data was carried out in several stages:

Firstly, we classify the top-100 artists in terms of the turnover reached by their works of art sold at international auction houses in the period from 1997 to 2015. It is a sufficiently long period of time to make a selection of artists with greater presence in the market and comprises the whole economic cycle, (economic expansion and recession).

However, the top 100 ranking varies from year to year changing the composition of the sample, with the exception of Picasso who always was ranked first throughout the period. Thus, looking for maintaining the sample of artists constant, we selected those artists who ranked the top 100 for at least 9 years of our period of study but also had at least 10 works of art sold each semester. These criteria were applied with the dual purpose of choosing figures of great relevance in the market but with a continuous presence in it.

Then, the number of results obtained in Google for each artist was quantified (January 2007; December 2015 and again in January 2016). The search was performed entering the name and surname of the artist, date of birth and death in quotes. This search system has

given rise to the cybermetric variables, "GoogleC" followed by the reference year. The aim of this type of search is to ensure that we only take into account those results that are directly related to the artist.

3. Methodology

In accordance with the study by De la Poza, *et al*., (2009), we want to test the following hypotheses:

1. The results (number of results) for each artist in Google in the year (n + 1) are related to the turnover reached by the artist in the art market art at year n.

2. The volume of sales in the art market in the year n is related to short-term results in Google in the year n + 1.

3. Finally, it is tested if the number of lots sold on the art market is related to the results on Google.

The methodology used has been the simple regression analysis. The purpose of this analysis is to investigate and explain the linear relationship established between a dependent variable (in this case the turnover in the art market), and an independent variable (number of results in Google); adjusted-R^2 coefficient measures the goodness of fit of the proposed models.

4. Results

The sample of artists is collected in Table 1, determining their position in the ranking in 2006 and 2015. All artists remained in the top 100 in 2015 with the exception of Edgar Degas, Kees Van Dongen and Maurice de Vlaminck.

Table 1. Sample of artists and their ranking position in 2006 and 2015.

Artist	2006	2015
Picasso, Pablo	1	2
Warhol, Andy	2	1
Monet, Claude	14	7
Kooning, Willem de	4	16
Chagall, Marc	6	23
Basquiat, Jean-Michel	37	11
Léger, Fernand	27	35
Lichtenstein, Roy	10	14
Miró, Joan	22	25
Matisse, Henri	9	28
Renoir, Auguste	11	36
Dongen, Kees van	23	154
Giagometti, Alberto	28	8
Modigliani, Amedeo	5	18
Degas, Edgar	31	101
Pissarro, Camille	24	49
Calder, Alexander	32	27
Fontana, Lucio	17	26
Cézanne, Paul	18	74
Richter, Gerhard	21	4
Twombly, Cy	120	12
Vlaminck, Maurice de	41	261

The regression models were computed, (Table 2). The table shows the regression models to value the turnover of the artists selected for each year (2006, 2007, 2014 and 2015) throughout the collected cybermetric variable GoogleC at different years.

Table 2. Regression Models

Function	a	p-value	b	p-value	Adjusted R^2	Durbin-Watson
F2006= f(GoogleC2007)	30,766,933.72	0	7.76	0	0.70	1.95
F2007= f(GoogleC2007)	56,381,249.98	0.001	6.08	0.007	0.27	1.6
F2015= f(GoogleC2016)	6,411,786.09	0.917	450.11	0.027	0.18	1.08
F2015= f(GoogleC2015)	16,584,411.14	0.778	431.23	0.031	0.17	1.1
F2014= f(GoogleC2015)	10,210,560.81	0.831	313.67	0.051	0.14	1.45

By the models obtained it is remarkable to highlight the loss of explanation of the cyber metric variable over time. Thus, the variable "GoogleC2007" explained 70% variability of the turnover of the group of selected artists in 2006, while the turnover of the same group of artists in 2015 was explained by 18 % by the "GoogleC" variable in 2016.

Also, it is important to highlight the importance that acquires the cyber metric variable "GoogleC2015" and "GoogleC2016" in the art market turnover. Thus, each result in the search engine "GoogleC2016" increases the 2015 turnover by 450 Euros. In fact, the coefficient of the variable "GoogleC2007" was 6.08, meaning that each additional result in the search engine Google increased turnover 6.08 Euros in 2006, while in 2015 the variable "GoogleC" contributed to the marginal increase in turnover in 2015 431.23 Euros.

One of the reasons that explain the increase in the value of the coefficient "GoogleC" in recent years (2015 and 2016), is the smallest number of results recorded by that search engine. The reason lies in the advances implemented by Google in capturing information available on the Internet. This has meant that for every artist the number of results has reduced considerably provided that both the name is included, as the years of birth and death, ("GoogleC"). This results in the mathematical expression that relates terms (turnover and number of citations), the slope coefficient and also value of the variable "GoogleC (term b) take high values.

Finally table 2 shows the results for testing the third hypotheses about the capacity of the cyber metric information to measure the number of lots sold (L) in the art market:

Table 3. Regression Models

Function	a	p-value	b	p-value	Corrected R^2	Durbin-Watson
L(2007)= f(GoogleC2007)	155,82	0.009	0.0000401	.000	0.54	2.27
L(2015)= f(GoogleC2015)	-355,68	0.235	0.003	0.006	0.29	1.28

As table 3 shows the explanatory power of the cyber metric information to measure the liquidity of the market drops over time.

4. Conclusions

This study presents a useful methodology for the assessment and quantification of the artist's reputation. In previous work (De la Poza, *et al*., 2009) was shown that there was a relationship between the information provided by search engines on the Internet and the volume of revenue in the art market, both in past and future in the short term.

As presented, the paradigm today is different; the results show how the explanatory power of the cyber metric variables is less at present than in the past. Thus, the explanatory power of the model valuation has dropped from values close to 70%, 30% and 46% for the cases studied in the years 2006 and 2007 to models whose explanatory power is below 20% and even very close to zero.

The same result was found when explaining the market liquidity, measured by the number of lots sold by the artists and the same cyber metric variables. One of the causes of the decline of explanatory power of the turnover by the cyber metric variables is the variation in the number of results. This is the case of Google, which from 2007 onwards has made continuous improvements to its search algorithm, and updating Caffeine in 2009, apart from reducing the search speed considerably increased refining searches, offering better results and better indexed temporarily.

In 2006 and 2007, the relationship between sales in the art market and cyber metric information in the Google search engine is greater when it comes to explaining past sales that future, suggesting that this information digital is able to collect a greater extent what happened in the market in the short term to estimate future sales.

In summary, these results open a path to new research to assess the artist through its impact on cyber metrics, and use the means of digital information to analyze the art market in long term.

References

Artprice.com: les raisons d'un succès. (2007). In *Beaux arts magazine*(276), 68-69.

Adam, G. (2002). "The opportunity to find bargains?". *Art newspaper, 13*(121), 33.

Anderson, M. L. (2004). Metrics of Succes in Art Museums. *The Getty Leadership Institute. United States*.

Caballer, V., & De la Poza, E. (2005). Modelos econométricos para la valoración de obras pictóricas. *VIII Congreso Internacional Cultura Europea. Universidad de Navarra*, 1-10.

Canals Cerdá, J. (2012). The value of a good reputation online: an application to art auctions. *Journal of cultural economics, 36*, 67-85.

Dass, M., Jank, W., & Shmueli, G. (2011). Maximizing bidder surplus in simultaneous online art auctions via dynamic forecasting. *International Journal of Forecasting, 27*, 1259-1270.

De la Poza, E., & Guadalajara, N. (2007). The Influence of the Net-Metric and Biblio-Metric Variables on the Top Artists. *Estudios de Economía Aplicada, 25*(1), 5-22.

De la Poza, E., Guadalajara, N., & Moya, I. (2009). El rol de los medios de información digitales en los precios en el mercado del arte. *El profesional de la información, 18*(4), 382-388.

Deng, S., Mitsubuchi, T., & Sakurai, A. (2014). Stock price shange rate prediction by utilizing social network activities. *The Scientific World Journal*, 14.

Grant, D. (2008). The internet art market. *American artist, 72*, 76.

Guadalajara, N., & De la Poza, E. (2007, abril). The influence of the netmetric and bibliometric variables on the top artists of the international art market. *Estudios de economía aplicada*, *25*(1), 5-22.

Hansen, G. (2006, May). I'm feeling lucky: using Google to break a creative slump. *Craft reports*, *32*, 36.

Schlaegel, C. (2015). Understanding individuals' initial and continued use of online auction marketplaces. A meta-analysis. *Management Research Review*, *38*(8), 855-907.

Ursprung, H. W., & Wiermann, C. (2008). Reputation, Price, and Death: An Empirical Analysis of Art Price Formation. *CESifo Working Paper*(2237).

Wikberg, E., & Strannegard, L. (2014). Selling by Numbers: The quantification and marketization of the Swedish Art World for Contemporary Art. *Organizational Aesthetics*, *3*(1), 19-41.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3140

Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community

Rietsch, Constant[a]; Beaudry, Catherine[a] and Héroux-Vaillancourt, Mikaël[a]
[a]Department of Mathematics and Industrial Engineering, Ecole Polytechnique de Montréal, Canada,

Abstract

In this exploratory study, we explore a methodology using a web mining technique to source data in order to analyse innovation and commercialisation processes in Canadian nanotechnology firms. 79 websites have been extracted and analysed based on keywords related to 4 core concepts (R&D, intellectual property, collaboration and external financing) especially important for the commercialisation of nanotechnology. To validate our methodology, we compare our web mining results with those from a classic questionnaire-based survey. Our results show a correlation between the indicators from the two methods of r=0.306 (p-value=0.007) for R&D, of r=0.368 (p-value=0.002) for IP, of r=0.222 (p-value of 0.071) for Collaboration and of r=0.222 (p-value=0.067) for external financing. We conclude that some of the data extracted by our web mining technique can be used as proxy for specific variables obtained from more classical methods.

Keywords: *Web-mining, Innovation, Commercialisation, Nanotechnology.*

1. Introduction

Data is often hard to come by, and firms are increasingly solicited to answer surveys and participate in interviews. In this paper, we explore a methodology using a web mining technique to source data and analyse innovation and commercialisation processes in Canadian nanotechnology firms and help to overcome surveys issues.

Public websites are generally freely available and provide relevant information about a firm's products, services, business models, R&D activities and so on. All this information can be mined by researchers to study innovation and technology management. The question is whether this information is reliable and whether there is enough to give a good portrait of a firm characteristics - can the content of a commercial website be used to identify various innovation characteristics of a company? And if so, can we validate this methodology with concrete evidence?

Nanotechnology-related firms are especially interesting because of their broad set of applications and business sectors. As enabling technology vectors of the 21st century (Siegrist et al., 2007), the vast majority of nanotechnology-related companies have a website that is regularly updated. Regularly updated websites have the advantage of displaying more accurate data than what can be found in governmental databases (Gök, Waterworth, and Shapira 2014). In this study we analysed and compared the commercialisation of nanotechnology in Canada using two different techniques.

The remainder of the article is organised as follows: Section 2 presents the theoretical framework around web mining and our hypotheses about nanotechnology innovation and commercialisation; Section 3 describes the data and survey-based methodology; Section 4 presents and analyses the results; and finally Section 5 presents our conclusion.

2. Theory and hypotheses

The use of Internet data has the advantage of not being in direct contact with the subjects of the study and would ensure a distance between them and the study. Thus, the subject is not led to adapt his behaviour to the study, as can be the case with questionnaires and interviews. These types of unobtrusive measures are suitable for research inquiring for real actions but are restricted by the access of such a given population (Webb et al. 1966). Usually, this type of study is less expensive compared to intrusive studies such as questionnaires and interviews, which require researchers to perform extensive data collection (Lee 2000).

Nowadays, more innovation studies tend to rely on online questionnaires that companies must complete themselves inducing the multiple bias related with this technique. According to Sauermann (2013), numerous studies about innovation that had based their data

collection process on these Internet surveys typically received low response rates (between 10 and 25%) affecting the results of analysis by non-response bias. These online questionnaires are often complex and time-consuming for business managers, which explains why such a low response rate can be found.

The concept of the exploration of Internet data can be explained by the way in which we retrieve information about companies via their websites to convert them into analytical data. The vast majority of companies working in high technological fields such as the ones using nanotechnology keep their website updated in order to inform potential customers and investors about the current activities of the company. Of course, the information is made available online by the companies themselves, which indicates the possibility of a strong self-reporting bias. However, this source of information access would be suitable for the study of emerging technologies such as nanotechnology (Gök et al. 2014). Furthermore, Youtie et al. (2012) note that small businesses tend to have smaller websites which would facilitate the handling of data. However, it is clear that companies do not disclose all strategic and business data on their websites as it is already the case with other available data sources such as scientific publications or patents. A successful web mining analysis would have several advantages over questionnaires, scientific publications and patents. To start with, the population covered by a study using a search of the Web (web mining) is very wide (Herrouz, Khentout, and Djoudi, 2013) in an area where questionnaire studies find few returns, particularly in the field of new technologies. Contrary to government data, the frequency of updates is high, even daily, in most cases (Gök et al. 2014). Thus the information contained in websites is perfectly suited to many possible types of studies in the field of new technologies. The main disadvantage is the difficulty to organise and interpret data, with each site having different information and being organised differently.

In this study, we focus on parameters influencing innovation and commercialisation of the nanotechnology of Canadian firms. Based on Lee et al (2013), 4 important factors are considered to especially influence the commercialisation of nanotechnology: R&D, intellectual property, collaboration and external financing. Innovation and R&D efforts are likely to influence positively the firms' commercialisation and financial performance as mentioned in many studies (Geroski et al. 1993; Klette and Griliches, 2000). For nanotechnology firms, R&D efforts are likely to give them a technological superiority on the market. Intellectual property, especially patents, are the research outputs giving the company a competitive advantage over the competition by providing the exclusive product research for commercialisation. Technology patenting implies a return on investment by marketing the technology, reselling the patent or selling licenses. Moreover, patent statistics are also often use as a proxy for innovative activities (Pavitt 1985). Collaboration is essential for the development and the deployment of emerging technologies. McNeil et al. (2007) show that collaboration with universities or government institutes allows young companies to access especially expensive tools. Furthermore, Kim et al. (2008) stress the

impact of university research and scientists in the industry by providing specialized manpower, patents and innovation. Finally, most nanotechnology projects are still in their early stages, meaning they need private or public funding to attain the commercialisation phase. Most SMEs require public funding or venture capital investment to support nanotechnology commercialisation helping them to bridge the valley of death (Kalil, 2005; McNeil et al., 2007).

R&D, intellectual property, collaboration and external financing have all synonyms and other related terms that a company can use to refer to it. When we visit a company's website, we are directed to read what the company wants us to read. Companies use words that can give insight into what they actually do. We suggest that the more a company uses terms related to a certain factor, the more they are likely to perform activities related to that specific factor. Thus, from the 4 factors we mentioned earlier, we suggest the 4 following propositions:

Proposition 1: The more words related to R&D are used on a firm's website, the more a firm would be likely to perform R&D activities.

Proposition 2: The more words related to intellectual property are used on a firm's website, the more a firm would be likely to perform intellectual property related activities.

Proposition 3: The more words related to collaboration are used on a firm's website, the more a firm would be likely to perform collaborations.

Proposition 4: The more words related to external financing are used on a firm's website, the more a firm would be likely to perform external financing activities.

Each proposition will be tested with the help of the results of a classic questionnaire-based survey using the methodology explained in the following section.

3. Methodology

3.1. Data collection and sample methodology

We started by conducting a classic questionnaire-based survey of which the core is based on the Oslo Manual (OECD and Eurostat, 2005) and explored the following themes: innovation, commercialisation, collaboration and intellectual property. A sample of the questionnaire can be found in Annex I.

Firms that either use or develop nanotechnology are not labelled nor searchable in any obvious way. We used a list of 583 firms from AGY consulting, a Canadian firm specialized in emerging technologies such as nanotechnology, clean technology and biotechnology. We asked the companies whether they were performing nanotechnology

activities or using nano-enabled products or processes. When the companies were eligible to the study, we listed them with their associated NAICS code. We used a total of 23 NAICS codes representing 67% of all the cumulated frequencies with which we bought lists of over 3000 companies. We thus contacted 2971 high technological Canadian firms. 973 firms did not respond, 1439 were not eligible to the survey, 380 refuses to participate and a total of 222 were eligible. The first 13 fully-completed questionnaire served to test and validate the questionnaire in order to mitigate any self-reporting and fatigue bias. We did remove 6 questions in order reduce the time of completion and reduce potential fatigue bias. A total of 89 respondents finally accepted to participate to our study allowing us to reach a response rate of 40%. Since the population is unknown, we are in a presence of a non-probabilistic convenience sample for which it is possible the methodology induced a selection bias. Of course, we assume that the respondents were honest and answered the survey with goodwill.

Our sample represent a wide range of Canadian nanotechnology firms. Moreover, 74 % of the firms are nanotechnology intensive, which means that at least 80% of their revenues come from nanotechnology-related innovations. The different application domains in nanotechnology are wide with 54% for advanced materials, 21% for biotechnology and medicine, 24.4% for electronics, 23.30% for equipment and devices, 13.3% for photonics and 33.3% for other. More than 50% of respondents are small businesses and 83.5 % are SMEs with an average of $94 M revenues and $31M without the 3 biggest firms. Finally, 85% of the firms came from Quebec and Ontario and 12 % are from British Columbia and Alberta.

In order to test several types of bias such as self-reporting bias, non-respondent bias and non-selection bias, we gathered 79 eligible enterprises that did not participate to the study into a control sample. To do so, we needed an external source of data to validate our main sample. Industry Canada provides a database of companies in different sectors. The database is comprised of data provided by the companies themselves on a voluntary basis. While Industry Canada does not guarantee the accuracy or the reliability of the content, we assumed the companies that willingly updated information in an official public database will input accurate information and thus mitigate the self-reporting bias from this source. We used the data available from Industry Canada where we found the number of employees for 37 firms and revenues for 30 firms from our main sample and the number of employees for 29 firms and revenues for 26 firms from our control sample. We compared our main sample and our control sample with these two metrics with a Mann-Whitney U test and we did not find a significant difference between both samples for both metrics (p-value=0.115, p-value=0.166) which leads us to assume we are not likely to face a non-respondent bias.

We then compared these two metrics between the data obtained via our questionnaire-based survey and the data from Industry Canada in order to verify if any important self-reporting

bias can be found. For every firm for which we had both data from our Questionnaire and from Industry Canada, we tested each data pair with a Wilcoxon Signed Ranks Test. We did not find any significant differences between our questionnaire results and the data from Industry Canada (p-value=0.058, p-value=0.714), which leads us to assume the self-reporting bias issued from the questionnaire is not different from the one we can find in an official public database.

3.2. Web mining methodology

Next we selected these 89 enterprises, and used a web scraper, Nutch, to extract and store the text from their website. Due to technical limitations such as the structure of the websites, only 79 of these firms (88%) provided enough information to be included in our study. We then used a content mining technique to perform a word frequency analysis with the text present on the websites. More specifically, in the 79 websites, we looked for innovation and commercialisation core factors : R&D, intellectual property, collaboration and external financing. For each factor, we listed all the relevant keywords that appear in company web pages. Factors, keywords and the web mining construct are described in Annex II. R&D and collaboration keywords were selected from the literature while intellectual property and external financing are issued from our own research. The Government of Canada offers many public programs and funding opportunities to companies for the development of nanotechnology projects. The website of Industry Canada identifies funds and programs offered to Canadian nanotechnology firms that we have used for our research.

Clustering using keyword frequency analysis with a text mining software enabled us to get the occurrences of each keyword for each factor. We transformed these clusters of occurrences into 4 continuous variables. Because the 79 companies are different in structure and size and therefore, present different amounts of information in their websites, we standardized each variable by dividing all occurrences by the total number of words appearing on their website and multiplied the resulting value by 1000. For each continuous variable, we obtained the Kurtosis and Skewness measures in order to determine whether our variables were following a normal distribution. All 4 variables did not follow a normal distribution so we transformed them by applying a natural logarithm (LN) or an inverse function (INV). In the case of External financing, we did not reach normality and thus, we treated this variable with a non-parametric test.

Since we selected only the companies that answered the survey, there is a possible selection bias. We ran our web mining technique on our control sample and generated the same variables. We then used a Student's t-test to test the difference of means for the following variables LN_WEB_MINING_RD (p-value=0.13), INV_WEB_MINING_IP (p-value=0.083) and LN_WEB_MINING_COLLAB (p-value=0.144) and conclude that the

difference is not significant between the two sample for these three variables. We tested the means of the variable WEB_MINING_EXTERN_FINAN (p-value=0.008) with a Mann-Whitney U test and found it was significant, so we cannot conclude for that variable that the means of the two sample are the same. Therefore, a selection bias is present for WEB_MINING_EXTERN_FINAN and will be included in our limits of research.

3.3. Questionnaire-based survey data methodology

In order to validate our 4 continuous variables from our web mining results, we identified all the relevant questions from the questionnaire-based survey and transformed them into different types of variables. The questions used can be found in Annex 1. We transformed every continuous variables from the survey that did not follow a normal distribution by applying a natural logarithm (LN) or an inverse function (INV). Since several 7-point Likert scale questions described the concept of R&D, we used Principal Component Analysis (PCA) with a Varimax rotation to reduce the number of variables and combine them into relevant dimensions corresponding to specific factors of the concept examined. Two factors were created but both the K-M-O and Cronbach alpha did not reach an acceptable level which would satisfy the validity and the reliability of the construct. In addition, these combined variables do not correlate with each other which hints towards using a formative construct. We thus proceeded to treating each item individually.

At the end, we generated a total of 9 variables corresponding to R&D, 1 variable related to collaboration, 2 variables corresponding to external financing and finally, 2 variables measuring intellectual property. The details of the Questionnaire-based survey construct can be see in Annex III.

3.4. Web mining validation with questionnaire-based survey methodology

Each pair of variables related to the same concept from the two methods (Web mining and survey) was examined via a Pearson correlation analysis when the subjects were following a normal distribution or a Spearman correlation when they were not following a normal distribution, to assess whether the variables stemming from the Web mining analysis can be used as a proxy for similar concepts measured by a survey. The details concerning our construct comparing a Web Mining technique and a Questionnaire-based survey can be see in Annex IV.

4. Results

The results of this paper aim to validate the utilisation of a web-mining-based methodology using firms' websites as a data source to analyse the extent of commercialisation and innovation, which can be used to better understand innovation practices. Comparing the variables constructed from the web mining and from the survey, we find a correlation of

0.306 (p-value of 0.007) between R&D measures and whether a firm is likely to provide R&D services to third parties. Additionally, we find a correlation of 0.306 (p-value of 0.010) when we associate the R&D concepts on websites with whether a firm has a high percentage of employees allocated to R&D tasks. Moreover, we find a correlation of 0.284 (p-value of 0.013) when we associate the R&D concepts on websites with whether a firm is likely to contract R&D service from external providers. Finally, we find a non-significant correlation of 0.197 (p-value of 0.100) when we associate the R&D concepts on websites with whether a firm has a long R&D process or not. It is important to note that the variable LN_NUMBER_RD which is related to the number of R&D projects did not correlate at all with our web mining variable with r =0.002 (p-value=0.985).

Terms related to intellectual property strongly correlate with the variables from the survey with a correlation of 0.368 (p-value of 0.002) regarding the use of intellectual property mechanisms and with a correlation of 0.351 (p-value of 0.033) regarding the activities related to patenting. Web mining methods therefore appear to be able to capture the importance of the use of IP mechanisms.

Collaboration terms from the Web sites are partially correlated with r=0.222 (p-value of 0.071) with the firms that confirmed collaborating from our questionnaire but the result is not significant at 5%.

External financing terms (from the web-based analysis) are also partially correlated with the extent of the use of external funds for commercialisation purposes (r=0.222 – p-value of 0.067) but the result is not significant at 5% regarding their importance for funding R&D activities.

To conclude, our latest results confirm the data extracted by our web mining technique can be used as a proxy at least for some of the variables coming from classical methods. If the collaboration and financing concept did not have a significant correlation, intellectual property and most of R&D web mining variables seem to be, according to our findings, good proxies for innovation studies.

5. Discussion and conclusion

Websites are a gold mine of informations. Researchers in innovation and technology management are now investigating if they can datamine enterprises' websites in order to get valuable data to their research. Nowadays, researchers rely on questionnaire-based survey to get most of the data. These questionnaires are costly, time consuming and a source of multiple bias. We thus explore a technique using data mining to determine if whether or not we can use data from websites as proxy for certain information that would have required a questionnaire-based survey to be obtained. We tested 4 factors that are determinant for the success of nanotechnology commercialisation: R&D, intellectual property, collaboration

and external financing. While results seems conclusive for intellectual property factors and some indicators of R&D, results did not show significant correlation with neither collaboration and external financing factors. Therefore, our proposition 3 and 4 are not yet validated.

For the specific case of R&D, we can observe that the web mining indicator seems to reflect the promotion needs in terms of R&D. Our web mining R&D indicator did correlate the most when firms are more likely to provide R&D services to third parties. This might be explained by the fact that the company use its website to promote their offer of R&D service. Also, our web mining R&D indicator did correlate really high with firm has a high percentage of employees allocated to R&D tasks. This can be explained by the willingness of a firm to attract new talents in R&D through their websites. Finally, our web mining R&D indicator did correlate significantly when firms are more likely to contract R&D service from external providers. However, one of the most important R&D indicator, the number of R&D projects, did not correlate at all with our web mining variable which may seems counter intuitive. Thus, we can hardly use our web indicator as a proxy since 3 independent indicators correlates with it and strong indicator of R&D activities are ignore. Therefore, our proposition 1 is partially true. A better definition of R&D activities would be required in order to use a R&D web mining proxy.

Our intellectual property web mining correlates with both the use of intellectual property mechanisms and the activities related to patenting. In that sense, our second proposition seems to be true i.e. intellectual property activities seems it can be explained by an IP web mining proxy.

More data would allow our research to be more robust, especially when it comes to verifying the concept of collaboration and external financing, normally addressed with classical methods, can be appropriately measured on web sites. For instance, we were not able to crawl data from all the companies from our survey due to technical limitations and only 79 out of 89 companies were used in this paper. Another limitation of our methodology is that we did not take into account the context of our keywords, possibly leading to multiple false positives. For instance, the mention of the word ‘collaboration’ on a website does not necessarily means that the company does collaboration with second parties at all. Qualitative data analysis of the websites’ content could be used to reduce the risk of false positives and to gather more accurate data. Moreover, our data are limited to textual content, while website also display, images, sounds and videos which are difficult to take into account in our study. Of course, websites, questionnaire-based survey and the official public database we used are all subject to self-reporting bias and it is part of our limitations.

Websites can be updated from time to time and the results can change accordingly depending on what companies want to display publicly. Thus, it is important to note that a

punctual web mine crawl might not be sufficient to capture all relevant information and results are subject to change with updated websites. Thus, longitudinal study would be required to better assess the validity of our methodology over time.

In the very near future, Partial Least Square (PLS) regression will be tested to determine if it is possible to create reliable and valid reflective indexes from the factors found by the PCA. In addition, we are currently investigating the use of a Multitrait-multimethod matrix (MTMM) to verify the validity and reliability of our constructs and to determine whether our methodology can be used as a valid approach to provide data for future innovation and technology management studies. Future studies will allow to better understand whether these web mining indicators capture all the information required to understand the proposed factors and can be used as a proxy for questionnaire-based survey questions or if these variables propose additional information that was not captured before by traditional means.

References

Geroski, Paul, Machin, Steve & Van Reenen, John. (1993). "The profitability of innovating firms." *The RAND Journal of Economics*, 198-211.

Gök, Abdullah, Alec, Waterworth, & Philip Shapira. (2014). "Use of Web Mining in Studying Innovation." Scientometrics, September, 1–19. doi:10.1007/s11192-014-1434-0.

Herrouz, Abdelhakim, Chabane Khentout, & Mahieddine Djoudi. (2013). "Overview of Web Content Mining Tools." *arXiv Preprint arXiv:1307.1024*. http://arxiv.org/abs/1307.1024.

Kalil, Thomas A. (2005). "Nanotechnology and the valley of death." 265.

Kim, Jinyoung, Sangjoon John Lee, & Gerald Marschke. (2008). "Impact of University Scientists on Innovations in Nanotechnology."

Klette, Tor Jakob, Jarle Møen, & Zvi Griliches (2000). "Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies." *Research Policy* 29, no. 4, 471-495.

Lee, C. J., Lee, S., Jhon, M. S., & Shin, J. (2013). Factors influencing nanotechnology commercialization: an empirical analysis of nanotechnology firms in South Korea. Journal of nanoparticle research, 15(2), 1-17.

OECD and Eurostat. (2005). Oslo Manual. The Measurement of Scientific and Technological Activities. OECD Publishing. http://www.oecd-ilibrary.org/science-and-technology/oslo-manual_9789264013100-en.

Pavitt, K. (1985). "Patent Statistics as Indicators of Innovative Activities: Possibilities and Problems." *Scientometrics* 7 (1-2): 77–99. doi:10.1007/BF02020142.

Ramdani, Anas, Vue de, Du diplôme de maîtrise l'obtention, and others. (2014). "revue systématique de la littérature sur les mesures de la collaboration inter-organisationnelle dans un contexte d'innovation." http://publications.polymtl.ca/1624/1/2014_AnasRamdani.pdf.

Sauermann, Henry, & Michael Roach. (2013). "Increasing Web Survey Response Rates in Innovation Research: An Experimental Study of Static and Dynamic Contact Design Features." Research Policy 42 (1): 273–86. doi:10.1016/j.respol.2012.05.003.

Siegrist, Michael, Carmen Keller, Hans Kastenholz, Silvia Frey, & Arnim Wiek. (2007). "Laypeople's and experts' perception of nanotechnology hazards."*Risk Analysis* 27, no. 1, 59-69.

Webb, E. J., Campbell, D. T., & Schwartz, R. D. (1966). Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally.

Youtie, Jan, Diana Hicks, Philip Shapira, & Travis Horsley. (2012). "Pathways from Discovery to Commercialisation: Using Web Sources to Track Small and Medium-Sized Enterprise Strategies in Emerging Nanotechnologies." Technology Analysis & Strategic Management 24 (10): 981–95. doi:10.1080/09537325.2012.724163.

Annex I - Questions from the questionnaire-based survey

R&D

1- How many nanotechnology-related and/or advanced material products in development do you actually have in each of the following phases?

1- Applied Research, 2- Product Scoping and Business Case Building, 3- Development, Testing and Validation, 4- Commercialisation

2- How important to your plant's innovation activities are each of the following sources of knowledge and innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Internal R&D in your firm
- Commercial laboratories / R&D firms / Technical Consultants

3- Please indicate the level of importance of each of the following innovation activities to your plant during the period 2010 to 2014 (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Contracting of external R&D service providers
- Providing R&D services to third parties

4- How long did it take to develop your most significant and recent (MSR) nanotechnology-related product innovation?

5- How important were each of the following organisations as collaborators in the development and commercialization of your MSR product innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Private research laboratories / Research and Development firms

6- How important were the following reasons in deciding to collaborate for the development and the commercialisation of your MSR product innovation? (1-Not important, 2-Very Low, 3-Low, 5-High, 6-Very high, 7-Essential)

- Accessing research and development

7- What proportion of Canadian employees from your firm are assigned primarily in R&D (%)?

Collaboration

1- Did your firm participate in alliances or collaborative agreements with other organisations to develop or commercialise your MSR product innovation? Y/N

2- How important were each of the following organisations as collaborators in the development and commercialisation of your MSR product innovation? (1-Not important, 2-Very low, 3-Low, 5-High, 6-Very high, 7-Essential).

- Universities or higher education institutions, College centres for technology transfer (CCTT) and CEGEPs, university technology transfer offices

External financing

1- Please indicate the proportion (%) of the total amount of financing provided by each of the following sources for the development and commercialisation of your MSR product innovation.

Y: 1- Internal funds of your firm or establishment, 2- Government subsidies / tax credits / academics grants, 3- Debt capital (such as bank loans), 4- Venture capital (public/private), 5- Collaboration agreements, 6- Programs from organisations such as nanoQuebec (now PRIMA-Quebec), nanoOntario, nanoAlberta, etc.., 7-Other

X: 1- Development of innovation, 2- Commercialisation of innovation

Intellectual property

1- Which of the following mechanisms are used by your firm to protect the intellectual property rights (IPR) for your MSR product innovation?

- Patents
- Trademarks
- Confidentiality agreements
- Trade secrets
- First mover advantage
- Other

2- How many patents does your firm own? Please note that the same patent filed in different countries is considered as only one patent.

Y: 1- Patent applications, 2- Existing patents, 3- Patents assigned / (sold) to others

X: 1- All patents, 2- Nanotechnology-related and advanced materials patents

Annex II - Web Mining construct

Topic	Factors	Keywords	Indicators	Variables
Innovation and commercialisation of nanotechnology (Lee et al, 2013)	R&D	research and development, r&d, laboratories, researcher, scientist, product development, technology development, development phase, technical development, development program, development process, development project, development cent, development facility, technological development, development effort, development cycle, development research, research & development, development activity, fundamental research, basic research (Gök et al. 2014)	Number of keywords frequencies per webpage	LN_WEB_MINING_RD (Continuous, normal)
	Intellectual property	Patent, intellectual property, trade secret, industrial design	Number of keywords frequencies per webpage	INV_WEB_MINING_IP (Continuous, normal)
	Collaboration	affiliation, collaboration, cooperation, partners, partnership (Ramdani et al. 2014)	Number of keywords frequencies per webpage	LN_WEB_MINING_COLLAB (Continuous, normal)
	External financing	atlantic canada opportunities agency, business development bank of canada, sustainable development technology, venture capital, atlantic innovation fund, nrc-irap, fednor, Industrial research assistance program, grants, private investment	Number of keywords frequencies per webpage	WEB_MINING_EXTERN_FINAN (Continuous, not normal)

Annex III - Questionnaire-based survey for web mining validation construct

Concepts	Indicators	Variables
R&D	• Number of R&D projects in nanotechnology • Level of importance of internal R&D as a source of knowledge • Level of importance of Commercial laboratories / R&D firms / Technical Consultants as a source of knowledge • Level of importance of contracting of external R&D service providers • Level of importance of providing R&D services to third parties • Time of R&D • Level of importance of Private research laboratories / Research and Development firms as collaborators for the development and the commercialisation • Level of importance of accessing research and development from collaborators for the development and the commercialisation • Proportion of Canadian employees assigned primarily in R&D (%)	• LN_NUMBER_RD (Continuous, normal) • D_INTENSITY_INTERN_INFO_RD (Dummy) • INTENSITY_EXTERN_INFO_RD (Continuous, normal) • INTENSITY_CONTRACTING_RD (Continuous, normal) • INTENSITY_PROVIDING_RD (Continuous, normal) • LN_TIME_RD (Continuous, normal) • D_INTENSITE_COLLAB_RD (Dummy) • D_INTENSITE_COLLAB_REASON_RD (Dummy) • PROP_RD (Continuous, normal)
Intellectual property	• Number of IP mechanisms used • Number of patents	• SUM_IP (Continuous, normal) • LN_NUMB_PATENT (Continuous, normal)
Collaboration	• Use of collaboration for the latest innovation	• D_COLLAB (Dummy)
External financing	• Proportion of external financing for R&D (%) • Proportion of external financing for commercialisation (%)	• RD_EXTERN_FINAN (Continuous, normal) • COMM_EXTERN_FINAN (Continuous, normal) • TOTAL_EXTERN_FINAN (Continuous, normal)

Annex IV - Questionnaire-based survey for web mining validation construct

Concepts	Web Mining Variables	Questionnaire Variables	Correlation type
R&D	LN_WEB_MINING_RD (Continuous, normal)	• LN_NUMBER_RD (Continue, normal) • D_INTENSITY_INTERN_INFO_RD (Dummy) • INTENSITY_EXTERN_INFO_RD (Continue, normal) • INTENSITY_CONTRACTING_RD (Continue, normal) • INTENSITY_PROVIDING_RD (Continue, normal) • LN_TIME_RD (Continue, normal) • PROP_RD (Continue, normal) • D_INTENSITE_COLLAB_RD (Dummy) • D_INTENSITE_COLLAB_REASON_RD (Dummy)	Pearson
Intellectual property	INV_WEB_MINING_IP (Continuous, normal)	• SUM_IP (Continuous, normal) • LN_NUMB_PATENT (Continuous, normal)	Pearson
Collaboration	LN_WEB_MINING_COLLAB (Continuous, normal)	• D_COLLAB (Dummy)	Pearson
External financing	WEB_MINING_EXTERN_FINAN (Continuous, not normal)	• RD_EXTERN_FINAN (Continuous, normal) • COMM_EXTERN_FINAN (Continuous, normal) • TOTAL_EXTERN_FINAN (Continuous, normal)	Spearman

ETLAnow: A Model for Forecasting with Big Data
Forecasting Unemployment with Google Searches in the EU

Tuhkuri, Joonas
M.I.T, Massachusetts Institute of Technology, and ETLA

Abstract

In this paper we document the ETLAnow project. ETLAnow is a model for forecasting with big data. At the moment, it predicts the unemployment rate in the EU-28 countries using Google search data. The model is publicly available at the ETLAnow's website, http://www.etlanow.eu.

The forecast model is based on the idea that volumes of Google searches could be associated with the current and future level of an economic index. And these data are available earlier than official statistics.

The motivation for our approach is that big data could help produce more accurate economic forecasts. Those forecasts would inform better policy and decisions, and help real people—especially during an economic crisis.

Keywords: *Big Data, Google, Internet, Nowcasting, Forecasting, Unemployment*

Macroeconomic Nowcasting Using Google Probabilities

Koop, Gary[a] and Onorante, Luca[b]
[a] University of Strathclyde, United Kingdom. [b]European Central Bank.

Abstract

Many recent papers have investigated whether data from internet search engines such as Google can help improve nowcasts or short-term forecasts of macroeconomic variables. These papers construct variables based on Google searches and use them as explanatory variables in regression models. We add to this literature by nowcasting using dynamic model selection (DMS) methods which allow for model switching between time-varying parameter regression models. This is potentially useful in an environment of coefficient instability and over-parameterization such as can arise when forecasting with Google variables. We extend the DMS methodology by allowing for the model switching to be controlled by the Google variables through what we call Google model probabilities. That is, instead of using Google variables as regressors, we allow them to determine which nowcasting model should be used at each point in time. In an empirical exercise involving nine major monthly US macroeconomic variables, we find DMS methods to provide large improvements in nowcasting. Our use of Google model probabilities within DMS often performs better than conventional DMS.

Keywords: *internet search data; nowcasting; dynamic model averaging, state space model.*

This working paper should not be reported as representing the views of the ECB. The views expressed are those of the authors and do not necessarily reflect those of the ECB. [a]This research was supported by the ESRC under grant RES-062-23-2646. Gary Koop is a Fellow at the Rimini Centre for Economic Analysis. [b]Currently on leave to the Central Bank of Ireland

Showcasting the use of big data for policy purposes

Per Nymand-Andesen
Adviser, Director General Statistics, European Central Bank.

"Progress lies not in enhancing what is, but in advancing towards what will be" (Khalin Gibran).

Abstract

While the availability and accessibility of large data sources is a rich field for statisticians, economists, econometricians and forecasters, it has been relatively unexploited for producing substainable and reliable statistics for policy purposes. It is of particular interest if such large sources could help to detect trends and turning points within the economy, thereby providing supplementary and more timely information compared to the "traditional" toolkit of policy makers. These supplementary statistics may provide further insights contributing to guiding policy actions as well as to assessing the subsequent impact and associated risks of these policy decisions on the financial system and real economy. Big data sources could assist policy makers in obtaining a nearly real-time snapshot of the economy as well as providing early warning indicators. This may be particularly welcome in the light of the shortcomings observed in the run-up to the financial crisis. In particular regulators have been keen to expand the data collection so as to better monitor financial intermediaries, financial risks and vulnerabilities.

The way forward may therefore be to develop and apply a structural approach for piloting the use of big data. While it may be reasonable to expect that "big data" suppliers have different business and data models, it is important to assess their usefulness according to a standardised set of five key criteria following a production cycle; "Input", "Quality", "Production", "Results" and "Assessment" as part of exploring its relevance for producing substainable and reliable statistics for policy making purposes.

Keywords: *Policy making; big data; statistics.*

Forecasting Births Using Google

Billari, Francesco[a]; D'Amuri, Francesco[b] and Marcucci, Juri[b]
[a]Department of Sociology, University of Oxford, United Kingdom, [b]Department of Economics, Statistics and Research, Bank of Italy, Italy

Abstract

Monitoring fertility change is particularly important for policy and planning purposes. New data may help us in this monitoring. We propose a new leading indicator based on Google web-searches. We then test its predictive power using US data. In a deep out-of sample comparison we show that popular time series specifications augmented with web-search-related data improve their forecasting performance at forecast horizons of 6 to 24 months. The superior performance of these augmented models is confirmed by formal tests of equal forecast accuracy. Moreover, our results survive a falsification test and are confirmed also when a forecast horse race is conducted using different out-of-sample tests, and at the state rather than at the federal level. Conditioning on the same information set, the forecast error of our best model for predicting 2009 births is 35% lower than the Census bureau projections. Our findings indicate the potential use of Googe web-searches in monitoring fertility change and in informing fertility forecasts.

Keywords: *fertility forecasting, US birth rates, Google econometrics, time series models, Forecast comparison.*

Evaluation of Business-Oriented Performance Metrics in e-Commerce using Web-based Simulation

Mitrevski, Pece[a] and Hristoski, Ilija[b]
[a]Faculty of Information and Communication Technologies – Bitola, "St. Kliment Ohridski" University, Republic of Macedonia, [b]Faculty of Economics – Prilep, "St. Kliment Ohridski" University, Republic of Macedonia.

Abstract

The Web 2.0 paradigm has radically changed the way businesses are run all around the world. Moreover, e-Commerce has overcome in daily shopping activities. For management teams, the assessment, evaluation, and forecasting of online incomes and other business-oriented performance measures have become 'a holy grail', the ultimate question imposing their current and future e-Commerce projects. Within the paper, we describe the development of a Web-based simulation model, suitable for their estimation, taking into account multiple operation profiles and scenarios. Specifically, we put focus on introducing specific classes of e-Customers, as well as the workload characterization of an arbitrary e-Commerce website. On the other hand, we employ and embed the principles of the system thinking approach and the system dynamics into the proposed solution. As a result, a complete simulation model has been developed, available online. The model, which includes numerous adjustable input variables, can be successfully utilized in making 'what-if'-like insights into a plethora of business-oriented performance metrics for an arbitrary e-Commerce website. This project is, also, a great example of the power delivered by InsightMaker®, free-of-charge Web-based software, suitable for a collaborative online development of models following the systems thinking paradigm.

Keywords: *e-Commerce; business-oriented performance metrics; evaluation; Web-based simulation; system dynamics.*

Automatic detection of e-commerce availability from web data

Blazquez, Desamparados[a]; Domenech, Josep[a]; Gil, José A.[b] and Pont, Ana[b]
[a]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain
[b]Department of Computer Engineering, Universitat Politècnica de València, Spain.

Abstract

In the transition to the digital economy, the implementation of e-commerce strategies contributes to foster economic growth and obtain competitive advantages. Indeed, national and supranational statistics offices monitor the adoption of e-commerce solutions by conducting periodic surveys to businesses. However, the information about e-commerce adoption is often available online in each company corporate website, which is usually public and suitable for being automatically retrieved and processed.

In this context, this work proposes and develops an intelligent system for automatically detecting and monitoring e-commerce availability by analyzing data retrieved from corporate websites. This system combines web scraping techniques with some learning methods for Big Data, and has been evaluated with a data set consisting of 426 corporate websites of manufacturing firms based in France and Spain.

Results show that the proposed model reaches a classification precision of about 85% in the test set. A more detailed analysis evidences that websites with e-commerce tend to include some specific keywords and have a private area. Our proposal opens up the opportunity to monitor e-commerce adoption at a large scale, with highly granular information that otherwise would have required every firm to complete a survey.

Keywords: *e-commerce indicator; Big Data; web scraping*

Public Opinion Mining on Sochi-2014 Olympics

Kirilenko, Andrei[a] and Stepchenkova, Svetlana[a]
[a]Department of Tourism, Recreation & Sport Management, University of Florida, USA

Abstract

The requirements of evidence-based policymaking promote interest to real-time monitoring of public's opinions on policy-relevant topics, and social media data mining allows diversification of information portfolio used by public administrators. This study discusses issues in public opinion mining with respect to extraction and analysis of information posted on Twitter about Sochi-2014 Olympic. It focuses on topics discussed on Twitter and sentiment analysis of tweets about the Games. Final database contained 613,333 tweets covering time span from November 1, 2013 until March 31, 2014. Using hash tags the data were classified into the following categories: Anticipation of the Games (9%), Cheering of the teams (31%), News (6%), Events (11%), Sports (18%), and Problems & Politics (15%). Research reveals considerable differences in the outcomes of machine sentiment classifiers: Deeply Moving, Pattern, and SentiStrength. SentiStrength produced the most suitable results in terms of minimization of incorrectly classified tweets. Methodological implications and directions for future research are discussed.

Keywords: *Mega-events; Opinion mining; Sentiment analysis; Sochi Olympics; Social media; Twitter.*

Geocoding of German Administrative Data

vom Berge, Philipp[a] and Wurdack, Anja[a]
[a]Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB), Germany

Abstract

Wherever we choose to move, we often find ourselves living close to people with similar social status, income or ethnicity. While many mechanisms have already been proposed to explain the driving factors of segregation, including the willingness to pay for local amenities, a preference for ethnical homogeneity, or the spatial distribution of jobs there is still missing an unified empirical framework to assess the relative importance of these mechanisms. Improvements in urban research requires data at a very granular spatial resoultion. We use geo-tagged administrative micro data to obtain more insights into urban research and develop an empirical framework to measure urban segregation. Our analysis is based on the Integrated Employment Biographies (IEB) from the Institute of Employment Research (IAB). The data excerpt used here is restricted to all main records effective on June 30th 2009. We identified the primary notification for the observation period and restrict our analysis to this. This subset of data has been linked to geocoded address material from the Federal Agency for Cartography and Geodesy (GAB) by a deterministic linkage. The individual data have been aggregated into grid cells with an edge length of 500 meters. The median daily wage has been calculated within each grid cell as well as the percentage of employees earning below several low-earning thresholds, including 2/3 of the national median gross daily wage, 2/3 of the city-specific median gross daily wage. The data from grid cells were used to produce various city-wide segregation indexes. This allows consistent comparisons of segregation in a larger cross-section of cities for the first time. We visually demonstrate the potential of our approach comparing segregation patterns in the the three largest cities in Germany. We find substantial variation across cities in both the spatial pattern of sorting and the extent of separation between social groups. This variation can be used to analyze social segregation and its relations to the local economical and demographical characteristics as well as to help city governments to reduce inequality.

Keywords: *Geocoding; labour market; segregation; urban planning*

Is content king? Job seekers' engagement with social media employer branding content

Moser, Kilian[a,b]; Tumasjan, Andranik[a] and Welpe, Isabell M.[a]
[a]TUM School of Management, Technical University Munich, Germany
[b]Center for Digital Technlogy & Management, Technical University Munich & University of Munich (LMU), Germany.

Abstract

Increasing digitization and the emergence of social media have radically changed the recruitment landscape adding interactive digital platforms to traditional means of employer communication. Removing barriers of distance and timing, social media enable firms to continue their efforts of promoting their employment brand online. However, social media employer communication and employer brand building remains woefully understudied. Our study addresses this gap by investigating how firms use social media to promote their employer brand. We analyze employer branding communication in a sample of N = 216,828 human resources (HR) related Tweets from N = 166 Fortune 500 companies. Using supervised machine learning we classify the Tweet content according to its informational and inspirational nature, identifying five categories of employer branding social media communication on Twitter.

Keywords: *Recruitment, social media, employer brand equity, Twitter, machine learning, automated text categorization.*

An extended manuscript is available from the authors on request.

Applied Webscraping in Market Research

Herrmann, Markus[a] and Hoyden, Laura[b]
[a]Data Lab, Data & Technology, GfK SE, Nuremberg, Germany
[b]Marketing & Data Sciences, Data & Technology, GfK SE, Nuremberg, Germany

Abstract

Modern Webscraping tools and APIs facilitate the extraction of information from the Internet significantly. We outline, that Webscraping, as a common practice to load, prepare and statistically analyze specific structured or unstructured data from the Internet, has become an essential technique in Marketing and Data Science. Furthermore, we emphasize the importance of Open Data and social media data as a scraping target. While we argue that Webscraping of internet data is an enabler and driver of product innovation in Market Research, it should also be noted that just gathering and integrating more data cannot replace research and modeling expertise; and that focusing on easily available data only, may inevitably lead to wrong conclusions or cause legal issues in commercial environments. As an result, data management concepts have to be applied to ensure accuracy, comparability, findability, re-usability and legality of the scraped data. In this presentation we discuss how data lakes, (meta-)data management and data integration processes help to extract most insight of scraped data.

Keywords: *Market Research, Marketing Science, Data Science, Webscraping, Open Data*

The impact of companies' websites on competitiveness and productivity performance

Domenech, Josep[a]; Rizov, Marian[b]; Vecchi, Michela[c]
[a]Department of Economics and Social Sciences, Universitat Politècnica de València, Spain
[b]Lincoln International Business School, University of Lincoln, United Kingdom
[c]Department of Economics, Middlesex University Business School, United Kingdom

Abstract

We investigate the role of the Internet in inducing market competition and firm productivity performance using a sample of UK and Spanish firms over the 1995-2010 period. For each firm we collect unique information on its online status (website) and the number of years of Internet activity. Our results show that the Internet is associated with reduced market concentration in both countries. Using a semiparametric estimation algorithm we find that firms' Internet presence is positively associated with the level of TFP but not with its rate of growth. This suggests that selection is likely to drive website adoption and that the Internet by itself cannot replace traditional sources of competitive and productivity advantage.

Keywords: *ICT, website, competitiveness, production function estimation, TFP*

Qualitative and Comparative Methods

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3636

Firm survival strategies for entrepreneurs and freelancers in the translation and interpreting sector

Gieure, Clara[a]; Berbegal-Mirabent, Jasmina[b]
[a]Department of Education, Universidad Católica de Valencia, Spain, [b]Department of Economy and Business Organisation, Universitat Internacional de Catalunya, Spain.

Abstract

This study examines a set of factors that contribute to firm and self-employed survival in the Spanish translation and interpreting sector (henceforth, T&I). In the midst of a global downturn firm and self-employed survival is key for the progress of economies and for a better and more stable future. The empirical analysis explores different patterns that lead to firm survival. Using comparative qualitative analysis,we identify seven combinations of causal conditions that explain the outcome. This study contributes towards a better understanding of entrepreneurial translators and interpreters' lifespan. The results indicate that entrepreneurs and freelancers can follow different pathways, all of them, conducing to firm survival. With little literature on the topic of firm survival in the T&I sector, thisstudy aims to fill this gap and make a valuable contribution to the current literature on T&I firms andself-employed survival.

Keywords:*firm survival, translation and interpreting, entrepreneur, freelance, self-employed, qualitative comparative analysis*

1. Introduction

In the current turbulent and uncertain economic environment, it is of utmost importance to better understand how firms align their resources in their attempt to survive and rapidly scale (Audretsch et al., 2016). As firm creation becomes central for the economic and social development of countries and regions, so it does the duration and survival of those firms. In recent years, scholars have increasingly focused their attention on the entry and exit modes of firms (Carreira&Teixeira, 2016), the entrepreneurial activity (Acs et al., 2015)or motivation (Solesvik, 2013). However, little attention has been paid to the strategies entrepreneurs follow once a business has been created. Considering that entrepreneurs and freelancers are today one of the principal agents for a prosperous economy (Kressel& Lento, 2012) and that only about half of all new small businesses survive after 4 years (Cader& Leatherman, 2011) there is a need to shed new light on how firms perform and what strategies contribute to their success. This paper offers an analysis of survival factors of T&I firms and self-employed entrepreneurs in Spainaiming at reducing the high levels of unemployment this sector has traditionally experienced and in doing so, contributing to the literature with factors that guarantee firm sustainability. The remainder of this article has the following structure. We first discuss the theoretical backgrounds and present the propositions posed herein. Then, we present the data and method used to empirically analyzethe propositions. Sectionfour develops a model that positsrelationships between all theoretical constructs and presents the results. Section fivepresents our concluding remarks, implications for business owners and limitations.

2. Theoretical background

Firm survival is considered the period of time the company stays in business, that is, the duration a business performs in the market (Van Praag, 2003). This metric is essential because it gives an idea of the expertise of the firm in the market, and therefore, combined with other variables it is possible to analyze how different factors have shaped its performance. Literature on firm survival varies depending on the sector. The focus here is the T&I service sector, known for its long history and for its prosperous market within European institutions where hundreds of translators and interpreters work in a full time basis to meet all the language requirements of the European Commission and other related institutions and agencies. Gouadec (2007) claims that the duration of a firm will depend on its internationalization, technology, specialization, and ultimately, know-how. The T&I market has abolished distances, having strongly benefited from globalization. On the contrary, it heavily relies on strong investments in translation tools, namely, computer-aided programs, and features numerous oral and written specializations.

Considering the specificities of the T&I market, the following paragraphs summarize the factors that, according to the literature, may influence firm survival.

Business size. For certain industries, firm size explains productivity and growth. Nevertheless, there is a minimum efficiency size, below which firms are generally destined to fail (Ribeiro-Soriano & Urbano, 2010). Some academics even suggest that larger firms are more likely to obtain public aid and subsidies (Görg & Strobl, 2007). This study analyzes small and medium-sized firms andfreelancers, although this is a sector where the freelancers prevail (Gouadec, 2007). Therefore, we pose that:*The size of a T&I firm does not have an influence on the survival of the firm.*

Entrepreneurial education. Numerous experts recognize that education and training can enhance business skills. However, although the traditional managerial skills taught in business schools are essential, they are not sufficient. More attention is neededto the development of entrepreneurial skills, attributes and behaviors, which go beyond purely commercial abilities (Kirby 2004). For the purposes of this study, we define entrepreneurial educationas the entrepreneurial knowledge and skills a person acquires when undertaking training and the specific knowledge acquired to start and run a T&I business. The first type of education aims at training people to undertake and manage a new business, thus learning the basics of venture creation and managerial skills; the latter focuses on gaining particular skills or knowledge applicable to the T&I sector. Thus, we argue that: *If the owner of a T&I business has received entrepreneurial education before starting a business, this training will have a positive influence on the survival of the firm.*

Entrepreneurial background. It is also paramount to consider the effect of having entrepreneurial kith and kin, or even relatives and friends with successful entrepreneurial stories (Kim et al., 2006). In this respect, Gatewood et al. (1995) claimed that entrepreneurs and business owners coming from business family backgrounds enjoy of a greater success than their peers that come from families without such a business background. Mintzberg (2004) also claimed that management skills are gained only through previous experience and stressed the fact that understanding abstract concepts is challenging when there is no understanding of the relationship between those concepts and the real experience. This rationale suggests that any prior experience that takes place before setting up a business could be beneficial for the business. Thus, we posit that: *For the business owner, having an entrepreneurial family background has a positive influence on the survival of the firm.*

Financial investment. Firm creation involves investing in different resources: knowledge-based, financial, organizational, social capital and intellectual property (Frid at al., 2015). When an entrepreneur has a business opportunity but lacks the resources to exploit it, it is difficult to pursue a business venture. Having access to resources constitutes a key factor in the decision to start a new venture, given that it influences the perception of business

viability. From the entrepreneur's standpoint, access to capital and financing is therefore a critical issue, particularly for small and medium firms (Verheul & Thurik, 2001). Accordingly, we suggest that: *Investing in a business has a positive influence on the survival of the firm.*

Media. The Web 2.0 has evolved to become part of almost every facet of our lives and its many tools benefit both customers and business owners. First, it enables customers to express their feelings and personal thoughts on firms and products to a group of people, and thus, firms have direct access to customers' feedback. Second, the Web 2.0 and its social media tools help businesses expandtheir markets. Today, firms are increasingly aware of the potential of using the media to create brand recognition, generate revenue, gain feedback and insights from customers, and to improve customer relationships (Müllern, 2011). In this study, the term "media" denotes the collective communication outlets or tools that firms useto interact online with the customer. Based on these considerations, we pose that: *The use of the media by the business owner has a positive influence on the survival of the firm.*

3. Data and method

3.1. Sample and data

The study considers 46 firms and self-employed entrepreneurs operating in the T&I sector in Spain. Information was obtainedby directly contacting the owners of these firms, and sending them the link to an online questionnaire. This process took from July to August 2014. The first part of the survey consisted of a set of basic questions to better characterize the profile of each firm and the owner/entrepreneur. In the second part, specific questions concerning the areas of interest outlined in the previous section were formulated.Table 1 displays the main characteristics of the sample under analysis.

Table 1. Descriptive statistics of the sample

Variables	Descriptives
Gender	28% Men
	72% Women
Legal form	49% Freelance
	51% Other forms
Company age	37% Between 0-5 years
	35% Between 6-10 years
	15% Over 10 years
Educational background of the founder/freelance	41% Degree in Translation and Interpreting
	12% English, French and Spanish Studies
	12% Master's Degree in Translation

4.5% Degree in Law
30.5% Other

To explain the outcome, five antecedent conditions grouped in three main dimensionswere considered: human capital, investment, and media. The first dimension, human capital, was operationalized by three antecedent conditions: size of the business, potential experience in starting up a business (either by the owner of the business or someone from his/her family), and entrepreneurial education. The second dimension, considered the importance of receiving external funding for stating-up the business. Lastly, the third dimension referred to the use of social media as a tool for generating awareness of the business activity.

3.2. Method

Because we are interested in exploring different patterns that lead to firm survival, in this study we use qualitative comparative analysis (QCA). According to Woodside (2016)this technique overcomes some of the traditional drawbacks of traditional methods as QCA allows identifying which combinations of antecedent conditions leads to a specific outcome by assuming complex causality and focusing on asymmetric relationships. Configurations consist of conditions or factors that can be positive, negative, or absent.

The first step in QCA consists in transforming variables into fuzzy or crisp set terms. Variables with continuous variables can be expressed in fuzzy terms (Ragin, 2008), with values ranging from 0 (full non-membership) to 1 (full membership), and 0.5 denoting the point of maximum ambivalence. For dummy variables, crisp-set is preferred. Table 2 shows how this step, known as calibration, was conducted.

Table 2. Calibration values

Variable definition		**Membership threshold values** [a]		
		Full non-membership (0.05)	Crossover point (0.5)	Full membership (0.95)
Outcomes	Firm survival [b]	0		1
Antecedent conditions	Size	1.1	2.9	4.1
	Entrepreneurial background [b]	0		1
	Entrepreneurial education [b]	0		1
	Financial investment [b]	0		1
	Media [b]	0		1

[a] Observations falling in the percentile-90 represent full set membership. Percentile-10 indicates full non-membership. The crossover point is defined by the median.
[b] Expressed in crisp-set terms.

Before proceeding with the analysis, it is required to check whether any of the antecedent conditions is necessary. Following Shneiderand Wagemann(2010)a condition is

"necessary" when its consistency score exceeds the threshold value of 0.9. The analysis reveals that all values loaded below this cut-off value, except of entrepreneurial education, which consistency score was slightly above 0.9. However, in the interest of including this variable in the analysis we decided to also include it.

In the following step, the truth table is constructed, and afterwards the number of rows is reduced using Boolean algebra. In this process, the Quine–McCluskey algorithm (Quine, 1952) computes the commonalities among the configurations that yield to the outcome and return the minimum a set of combinations of causal conditions that are sufficient to produce the outcome. The reduction is based on the empirical relevance of the solution (coverage) and the extent to which cases sharing similar conditions exhibit the same outcome (consistency). This study uses the fsQCA software program, in its version 2.5.

4. Empirical results and discussion

Table 3 shows the results. Following Ragin's (2008) recommendation, this study reports the intermediate solution. The coding for the solution table follows the approach of Ragin and Fiss (2008), where black circles (●) indicate the presence of a condition, white circles (○) denote its absence, and blank cells represent ambiguous conditions.

Table 3. Sufficient configurations of antecedent conditions for the different outcomes

Configuration	Antecedent conditions					Coverage		Consistency
	Size	Entrepreneurial background	Entrepreneurial education	Financial investment	Media	Raw coverage	Unique coverage	
1	○	●	●			0.4021	0.1700	1.0000
2	●		●		●	0.1774	0.0948	0.9790
3	○	○			●	0.1090	0.1005	1.0000
4	○	●		●	●	0.2521	0.0229	1.0000
5	●	○		●	●	0.0543	0.0133	0.9344
6	●	○	●	●		0.0648	0.0238	0.9444
7	○	○	○	●	○	0.0229	0.0229	1.0000

Solution coverage: 0.7298, Solution consistency: 0.9948
Frequency threshold = 1.0000, Consistency threshold = 0.9149

Seven different combinations of the antecedent conditions explain the outcome, confirming our initial intuition that there is no unique recipe to explain firm survival. On the contrary, results indicate that entrepreneurs/freelancers can follow different pathways, all of them,

conducing to firm survival. This finding is of great interest because the profile of the entrepreneurs in this sector that decide to become self-employed is fairly heterogeneous. In terms of the fit of the model, the consistency of all the configurations is above the acceptable cut-off point of 0.8. Raw coverage values also validate our approach.

Turning to the specific results, we find that the firm size is not a key determinant, meaning that there are different formulas for succeeding in this sector regardless of the size of the firm. As for the effect of the entrepreneurial background of the owner, it seems to play a weak role, but contributes to the outcome when the firm is small. Somehow different is the effect of entrepreneurial education. In three out of seven configurations,it adds value to the firm, signaling that business skills are desirable, particularly when there is a lack of previousentrepreneurial experience or the firm is small.As expected, results indicate that access to financial resources is critical. This effect is accentuatedwhen the owner has no previous entrepreneurial experience. Lastly, the use of media as a way to reach a wider audience and attract customers is also found to positively contribute to the outcome.

Following Ragin's (2008) recommendation, the two causal paths with greater raw coverage (configurations 1 and 4) deserve further attention. Both recipes apply for small firms in which the entrepreneur/owner possesses some kind of entrepreneurial background. Configuration 1 suggests that if under these circumstances the owner has received training on how to start up a business, his/her business activity has chances to be successful in the mid-long term. The absence of entrepreneurial education can be compensated by financial resources (which will provide access to resources) combined with the use of social media to generate firm awareness (configuration 4).

5. Concluding remarks, implications and limitations

Aiming at finding which combinations of antecedent conditions guarantee firm sustainability in terms of survival in the T&I sector, we conducted a qualitative-quantitative study.The contribution of this study to the existing literature is twofold. First, because the survival of firms is necessary for the progress of the economies and knowledge transfer, this research presents a new approach to the study of the determinants of firm survival. Second, the T&I sector reveals a limited number of related research studies, therefore, this gap should be addressed to provide empirical evidence that help future entrepreneurs to successfully exploit the resources and capabilities they possess when deciding entering the sector.These findings have important implications for T&Ibusiness owners, but in particular, for start-ups. In this way, firms should make efficient use of the media, financial resources and undertake entrepreneurial training to achieve success.

Despite covering an existing gap in the literature, this study has several limitations that represent unique opportunities for future studies. First, the survey failed to

differentiatebusiness owners from self-employees, thus, addressing both groups separately would be more informative. Second, because much of the service offered by firms in the T&I sector can be delivered online, it would also be interesting expanding the sample and analyzing the effect of internal/external customers. Third, although the study uses reliable variables, a need remains to question whether some other variables might be best proxies to capture the outcome. Four, future studies should also consider longitudinal analyses.

References

Acs, Z. J., Szerb, L., &Autio, E. (2015). *Global entrepreneurship and development index 2014*. New York, NY: Springer.

Audretsch, D., Guo, X., Hepfer, A., Menendez, H., & Xiao, X. (2016).Ownership, productivity and firm survival in China. *Economia e PoliticaIndustriale*, 1-17.

Cader, H.A., & Leatherman, J.C. (2011).Small business survival and sample selection bias. *Small Business Economics*, *37*(2), 155-165.

Carreira, C. &Teixeira, P. (2016) Entry and exit in severe recessions: lessons from the 2008-2013 Portuguese economic crisis. *Small Business Economics*, *46*, 1-27.

Frid, C.J., Wyman, D.M., & Gartner, W.B. (2015).The Influence of Financial 'Skin in the Game' on New Venture Creation. *Academy of Entrepreneurship Journal*, *21*(2), 1.

Gatewood, E.J., Shaver, K.G., & Gartner, W.B. (1995).A longitudinal study of cognitive factors influencing start-up behaviors and success at venture creation.*Journal of Business Venturing*, *10*(5), 371–391.

Görg, H., &Strobl, E. (2007).The effect of R&D subsidies on private R&D.*Economica*, *74*(294), 215-234.

Gouadec, D. (2007). *Translation as a Profession* (Vol. 73).John Benjamins Publishing.

Kim, P.H., Aldrich, H.E., &Keister, L.A. (2006).Access (not) denied: The impact of financial, human, and cultural capital on entrepreneurial entry in the United States. *Small Business Economics*, *27*(1), 5-22.

Kirby, D.A. (2004). Entrepreneurship education: Can business schools meet the challenge?.*Education + Training*, *46*(8/9), 510 – 519.

Kressel, H., & Lento, T.V. (2012). *Entrepreneurship in the Global Economy: Engine for Economic Growth*.Cambridge University Press.

Mintzberg, H. (2004). *Managers not MBAs.*San Francisco: Berrett-Koehler.

Müllern, T. (2011).*Facebook as a marketing cannel.A study of eCommerce retailers' Facebook page ambitious.*Master Thesis.InternationellaHandelshögskolan, Jönköping, Sweden.

Quine, W.V. (1952). The problem of simplifying truth functions. *The American Mathematical Monthly*, *59*(8), 521–531.

Ragin, C.C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ragin, C.C., &Fiss, P. (2008).Net effects analysis versus configurational analysis: An empirical demonstration.In C.C. Ragin (ed.): *Redesigning Social Inquiry: Fuzzy Sets and Beyond* (pp. 190–212). Chicago, IL: University of Chicago Press.

Ribeiro-Soriano, D., & Urbano, D. (2010). Employee-organization relationship in collective entrepreneurship: An overview. *Journal of Organizational Change Management*, *23*(4), 349-359.

Schneider, C.Q.,&Wagemann, C. (2010).Standards of good practice in qualitative comparative analysis (QCA) and fuzzy-sets. *Comparative Sociology*, 9(3), 397-418.

Solesvik, M. Z. (2013). Entrepreneurial motivations and intentions: investigating the role of education major. *Education+ Training*, *55*(3), 253-271.

Van Praag, C.M. (2003). Business survival and success of young small business owners. *Small Business Economics*, *21*(1), 1-17.

Verheul, I. &Thurik, R. (2001). Start-up capital: Does gender matter?.*Small Business Management*, *16*, 329-345.

Woodside, A.G. (2016). The good practices manifesto: Overcoming bad practices pervasive in current research in business. *Journal of Business Research*, 69(2), 365-381.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3139

Examining technology transfer activities at universities: Does one recipe explain all outcomes?

Berbegal-Mirabent, Jasmina[a]; Guerrero, Adrián[a]
[a]Department of Economy and Business Organisation, Universitat Internacional de Catalunya, Spain.

Abstract

The growing importance knowledge and innovation are acquiring as the basis for economic development and growth has lead universities to expand their traditional functions, spreading their commitment in the contribution to economic and social welfare through their so-called third mission.

Certainly, universities have turned into one of the most important engines for regional development, therefore, they are called to play a paramount role in the provision of new knowledge, which is expected to have a positive impact in the innovation systems of their neighbouring regions. However, these new demands are not followed by a larger availability of resources. On the contrary, universities are struggling to simultaneously carry out teaching and research activities alongside with technology transfer ones.

In this study we aim at scrutinize which are the determinants of technology transfer outcomes at universities. By means of an empirical analysis, we examine which combination of resources lead to higher technology transfer outcomes when these take the form of patents, spin-offs and R&D contracts. Data from the Spanish higher education system for the period 2004-2011 is used. Implications for policy and practice are discussed.

Keywords: *technology transfer; universities; third mission; spin-offs; patents; R&D contracts.*

1. Introduction

Over the last decades, research activities have expanded and new figures have appeared professionalising different tasks related with research activities (Berbegal-Mirabent et al., 2013). In this context, universities are seen as institutions that generate knowledge and transmit it to people. The contemporary university is a combination of teaching, research, entrepreneurial and scholastic interests. Universities do not only provide highly qualified graduates and researchers, but they also offer innovative solutions through technology-transfer mechanisms that foster links with the local industry system. This means that, despite teaching and research are key actions for satisfying society's knowledge demands, it is in the local development where this process articulates. This situation not only requires a high standing university able to educate people with the latest technologies, but also a capable university to translate research results into marketable outcomes.

The growing awareness of how universities can contribute to regional development has lead governments to rethink how to maximize the benefits arising from universities. Collaboration with businesses, local and regional public authorities and other local actors are the traditional practices to reach this purpose, however, benefits arising from these relationships are still far from their true potential, and strongly differ from one university to another (Shattock, 2009).

To cope with these requirements for a major collaboration with firms, universities have enlarged their service portfolios and introduced significant changes in their traditional ways of operating. However, universities face with important constraints in terms of resources, which impede them to successfully engage in different forms of technology transfer activities while maintaining high levels in teaching and research duties. This translates into saying that universities are struggling to survive in a competitive environment where they are constantly asked to simultaneously excel at multiple tasks.

Recognizing this limitation, the original contribution of this study relies in assessing the antecedent conditions of specific technology transfer outcomes, but also in examining a new model, where the desired outcome is a technology transfer output whatever being its form (patent, spin-off or R&D contract). Because resources are scarce and universities might address their objective function differently, not all universities should allocate their efforts in producing the same technology transfer outcomes. The empirical application considers the Spanish higher education system for the period 2004-2011. Using qualitative comparative analysis, we test if there is a unique formula that leads to the desired outcomes.

2. Technology transfer activities

Both academics and policy makers are interested in how universities can develop their third stream function to become more adept at exploiting their knowledge-base and transfer it to the private sector (Lockett & Wright, 2005). Third mission denotes activities primarily designed to support regional engagement and regional economic growth more generally (Shattock, 2009). This mission allows universities to stimulate knowledge creation, knowledge flows and its valorisation and commercialisation into the marketplace. The convergence of three main axes (entrepreneurship, innovation and social commitment) has helped universities respond to the social pressures they face.

Two main ways are primarily envisioned when talking about third mission activities (Berbegal-Mirabent et al., 2013): placing products and services in the marketplace directly (through the creation of spin-offs) or indirectly (by interacting with firms). University spin-offs can be defined in terms of the identification and exploitation of an opportunity carried out by individuals with a particular commitment in starting up a business in the university context. Scientific knowledge can be turned into the starting point of a business idea, and by extension, the birth of new a company. In the second case, possible ways of partnering with firms include cooperation agreements, consulting services, incubator facilities or assessment for start-ups. Universities' motivations to engage in such activities mainly relate to the opportunity to access to new sources of funding, and to get new ideas which can be the basis for new fundamental research. From the firms' perspective, universities offer them a broad spectrum of expertise, human capital and training. Firms also use universities' research infrastructures as a way to save money and take advantage of their expertise and setting. Similarly, firms might outsource some R&D activities to universities because of the prohibitive costs. The benefits derived from third stream activities such as the ones described above are widely documented in the literature (Chang et al., 2009).

To cope with the challenges of adopting this third role, universities have developed new strategies and policies, and have been provided with new infrastructures to foster university-industry partnerships. Whereas the former may include the establishment of regulatory frameworks for the devolution of intellectual property rights, patents or licenses, the latter considers the creation of technology transfer offices (TTO), business incubator centres, or the establishment of science parks affiliated to universities.

3. Data and method

3.1. Data

Three outcomes are considered: patents (cutting-edge discoveries), spin-offs (new venture creation), and income from R&D contracts (how active is the university in establishing

lucrative R&D agreements with firms). There is a widespread agreement in considering these outcomes as relevant metrics of technology transfer activities conducted at universities (Berbegal-Mirabent et al., 2012). A fourth outcome is obtained as an "OR" combination of all them.

Antecedent conditions are classified into three dimensions: human capital, financial resources, and support infrastructure. Descriptive statistics are shown in Table 1. Human capital represents the knowledge, abilities and capabilities provided by individuals. This construct is operationalised through two indicators: staff engaged in research and teaching activities (faculty members), and the technical staff from the TTO that gives specific support to technology transfer activities. Financial resources give universities the opportunity to support new research activities. This effect is particularly relevant in both knowledge creation and diffusion, as they tend to require huge investments (Landry et al., 2007). In this study we employ the budget of the TTO. Lastly, we examine the role of specific infrastructures, which are expected to positively contribute to strengthen ties with the business sector (Phan et al., 2005): business incubators and science parks.

Data come from the CRUE (Conferencia de Rectores de Universidades Españolas) and the RedOTRI (Network of Spanish Technology Transfer Offices) reports, and contains information for all Spanish public universities for the period 2004-2011.

Table 1. Descriptive statistics of the outcome and antecedent conditions

Variables			**Mean**	**Std. dev.**	**Min.**	**Max.**
Outcomes	Patents		7.91	8.12	0.00	46.00
	Spin-offs		2.46	3.13	0.00	20.00
	R&D contracts income*		9,872.24	11,817.18	380.22	8,6170.00
Antecedent conditions	Human Resources	Faculty members	1,267.75	841.15	124.00	4,027.00
		TTO staff	16.91	14.25	3.00	94.00
	Financial resources	TTO budget*	716.38	713.12	24.00	3,995.00
	Support infrastructures	Business incubator	0.58	0.49	0.00	1.00
		Science Park	0.89	0.31	0.00	1.00

* Units in thousand €

3.2. Method

Because the interest of this research is not so much which factors are necessary but which combinations of factors are sufficient to explain an outcome, this study uses qualitative comparative analysis (QCA). This method assumes complex causality and focuses on asymmetric relationships that detect configurations that are minimally necessary and/or

sufficient for obtaining a specific outcome (Meyer et al., 1993). Configurations consist of conditions or factors that can be positive, negative, or absent.

To perform QCA, variables were transformed into a scale from 0 (full non-membership) to 1 (full membership) indicating their level of belongingness. For continuous variables we used a fuzzy-set transformation, while for dichotomous ones, crisp-set was preferred (Ragin, 2008). This process known as calibration and is reported in Table 2. In the next step, the truth table was built, followed by a reduction of the number of rows included in this table. Using Boolean algebra, the Quine-McCluskey algorithm (Quine, 1952) returned a set of combinations of causal conditions, each combination minimally sufficient to produce the outcome. Rows were reduced based on two criteria: coverage and consistency.

Table 2. Calibration values

Variable definition		Membership threshold values [a]		
		Full non-membership (0.05)	Crossover point (0.5)	Full membership (0.95)
Outcomes	Patents	1.00	4.00	17.00
	Spin-offs	0.00	1.00	6.00
	R&D contracts income	1,616.80	6,611	18,907.60
Antecedent conditions	Faculty members	422.60	1,046.00	2,478.00
	TTO staff	5.00	10.00	34.60
	TTO budget	144.00	485.00	1,706.40
	Business incubator [b]	0		1
	Science Park [b]	0		1

[a] Observations falling in the percentile-90 to represent full set membership. Percentile-10 is the threshold value for indicating full non-membership. The crossover point is defined by the median.
[b] Expressed in crisp-set terms.

4. Empirical results and discussion

We first tested whether any of the antecedent conditions was "necessary". To do this we computed the consistency scores. As any of the variables displayed values higher than 0.9 (Shneider et al., 2010), we concluded that none of the variables was a necessary condition to cause the outcome.

Table 3 shows the results for the intermediate solution. Different configurations explain the technology transfer outcomes considered, all of them exhibiting acceptable consistency indices (≥0.80). Raw coverage values also validate our approach, being particularly high in almost all of the recipes, except for those in model S, where values range from 0.02 to 0.35.

Table 3. Sufficient configurations of antecedent conditions for the different outcomes

Model	Configuration	Antecedent conditions					Coverage		Consistency
		Faculty members	TTO staff	TTO budget	Business incubator	Science Park	Raw coverage	Unique coverage	
Patents	P_1	●			○	●	0.2474	0.0369	0.8647
	P_2	●	●			●	0.5659	0.3554	0.8895
	P_3		●	●	○	●	0.2175	0.0331	0.9201
	Solution coverage: 0.6359 Solution consistency: 0.8502								
Spin-offs	S_1	●	●			○	0.0493	0.0278	0.9118
	S_2	●	●		●		0.3519	0.2389	0.8120
	S_3	●		○	●	○	0.0217	0.0078	0.9088
	S_4	○	○	●	●	●	0.1419	0.0504	0.8324
	S_5	●	○	●	○	●	0.0858	0.0859	0.8292
	Solution coverage: 0.5236 Solution consistency: 0.8136								
R&D contracts income	RD_1	●	●	●	●		0.4098	0.0195	0.9316
	RD_2	●	●	●		●	0.5772	0.1870	0.9151
	RD_3	●	●		●	●	0.4210	0.0307	0.9057
	Solution coverage: 0.6274 Solution consistency: 0.8975								
Technology Transfer (patents or spin-offs or R&D contracts income)	TT_1		●			●	0.6243	0.0792	0.9002
	TT_2		○	○	●		0.2595	0.0255	0.8526
	TT_3	●	●	○			0.2720	0.0155	0.9867
	TT_4	●			○	●	0.2006	0.0364	0.9362
	TT_5	○			●	●	0.2895	0.0205	0.7584
	TT_6		●		●	●	0.3209	0.0058	0.9754
	Solution coverage: 0.8146 Solution consistency: 0.8380								

Frequency threshold = 1.
Consistency threshold = 0.86 (Model P), = 0.83 (Model S), = 0.88 (Models RD and TT).
Following Ragin and Fiss (2008) notation, black circles ("●") indicate the presence of a condition, white circles ("○") denote its absence, and blank cells represent ambiguous conditions.

Results from the models where outcomes are assessed in an individual fashion (P, S and RD) suggest that there is no unique formula to produce them. In general terms, we can conclude that human capital is a highly valuable attribute, both in terms of faculty members and TTO staff. While the firsts apply their knowledge to produce cutting-edge discoveries that are sound to the industry, the latters are in charge of supporting the commercialization

process and bring their expertise in establishing university-business collaborations. In those cases where TTO staff is scarce (S_4 and S_5, Table 3) this absence can be compensated with TTO large budgets, so that it is possible to ask for external help in this process.

Financial resources are also found to be key antecedents of technology transfer outcomes. Specifically, they are an important ingredient for R&D contracts. Universities have to compete in a globalized market where other corporations and firms can act as R&D providers. Therefore, marketing investments are needed in order to raise external awareness of the research conducted at universities. As for the case of spin-offs, financial support tends to come from business angels and other external agents. Yet, there are still few universities managing large amounts of seed capital.

As for the influence of support infrastructures, results suggest that science parks are useful mechanisms for bringing together businesses and research centers. Said differently, their role is relevant for patenting and the establishment of lucrative R&D contracts. However, when it comes to create new academic ventures, the importance dilutes. On the other hand, business incubators are more relevant for spin-offs purposes.

Lastly, results from the fourth model report interesting findings. First, it is noteworthy to point out the positive effect of the science park. In four out of six configurations, a geographic enclave where firms and researchers can interact seems to play a significant role. Business incubators are another mechanism that boosts technology transfer processes. A critical mass of human capital can help overcome the absence of such advanced infrastructures. A key finding is the role of the TTO budget. Results reveal that a shortage in financial resources is not an impediment for transferring technology if there is a skilled pool of researchers that benefit from the technical expertise and help of TTO staff (configurations TT_2 and TT_3).

Following Ragin's (2008) recommendation, the two causal paths with greater raw coverage (configurations TT_1 and TT_6) deserve further attention. Both TT_1 and TT_6 are very similar. The difference relies in the role of the business incubator, which appears to be contributing to the outcome in configuration 6, but has an imprecise role in configuration 1.

5. Concluding remarks

The variety of recipes obtained suggests that there is no unique magic recipe that drives to multiple technology transfer outcomes. On the contrary, different pathways are envisioned revealing that Spanish universities can use different formulas to accomplish with the third mission. These findings reinforce the idea that universities should follow the path that best suits their strategic vision.

To the best of the authors' knowledge, this study is one of the first examining how different technology transfer outcomes can be obtained simultaneously. The results reported here might undoubtedly bring fresh insights to both university managers and scholars in the field of technology transfer. Although measures used are reliable, future research should consider examining other combinations of technology transfer outcomes and antecedent conditions.

References

Berbegal-Mirabent, J., Lafuente, E., & Solé, F. (2013). The pursuit of knowledge transfer activities: An efficiency analysis of Spanish universities. *Journal of Business Research*, 66(10), 2051-2059.

Berbegal-Mirabent, J., Sabaté, F., & Cañabate, A. (2012). Brokering knowledge from universities to the marketplace: The role of knowledge transfer offices. *Management Decision*, 50(7), 1285-1307.

Chang, Y., Yang, P.Y., & Chen, M. (2009). The determinants of academic research commercial performance: Towards an organisational ambidexterity perspective. *Research Policy*, 38(6), 936-946.

Landry, R., Amara, N., & Ouimet, M. (2007). Determinants of knowledge transfer: evidence from Canadian university researchers in natural sciences and engineering. *Journal of Technology Transfer*, 32(6), 561-592.

Lockett, A., & Wright, M. (2005). Resources, capabilities, risk capital and the creation of university spin-out companies. *Research Policy*, 34(7), 1043-1057.

Meyer, A. D., Tsui, A.S., & Hinings, C. R. (1993). Configurational approaches to organizational analysis. *Academy of Management Journal*, 36(6), 1175–1195.

Phan. P.. Siegel. D. S., Wright. M. (2005). "Science parks and incubators: Observations, synthesis and future research". *Journal of Business Venturing*. 20(2), 165-182.

Quine, W.V. (1952). The problem of simplifying truth functions. *The American Mathematical Monthly*, 59(8), 521–531.

Ragin, C. C. (2008). *Redesigning social inquiry: Fuzzy sets and beyond*. Chicago: University of Chicago Press.

Ragin, C. C., & Fiss, P. (2008). Net effects analysis versus configurational analysis: An empirical demonstration. In C.C. Ragin (ed.): *Redesigning Social Inquiry: Fuzzy Sets and Beyond* (pp. 190–212). Chicago, IL: University of Chicago Press.

Schneider, M. R., Schulze-Bentrop, C., & Paunescu, M. (2010). Mapping the institutional capital of high-tech firms: A fuzzy-set analysis of capitalist variety and export performance. *Journal of International Business Studies*, 41(2), 246-266.

Shattock, M. (2009). *Entrepreneurialism in universities and the knowledge economy*. Maidenhead: Society for Research into Higher Education and Open University Press.

Organizational design configurations in the early stages of firm's life cycle

Mosca, Luigi; Gianecchini, Martina; Campagnolo, Diego
Department of Economics and Management "M. Fanno", University of Padova, Italy.

Abstract

The liability of newness affirms that firms do face a higher risk of being selected out from their competitive environment more in the first years of their existence, than later (Stinchcombe, 1965; Henderson, 1999). The main reasons for the liability of newness are the problems of setting up an organizational structure and getting the new unit to work efficiently enough to keep pace with their competitors (Fritsch et al., 2006). Nevertheless, empirical studies about the survival rate of new firms mostly focus on such factors as: relationship with capital markets (Robb & Robinson, 2012), expectations of the entrepreneur (Sarasvathy et al., 2013), background of the employees (Andersson & Klepper, 2013), industry, region and time effects (Fritsch et al., 2006), and clusters role (Wennberg & Lindqvist, 2010). Conversely, a thorough understanding on the role played by the organizational solutions firms adopt in the early stages of their life is still missing. Using the fsQCA methodology, this paper fills this gap by analyzing the organizational configurations of those firms that have successfully overcome the liability of newness and have been able to grow.

Keywords: *Organizational configurations; young firms; life cycle; fsQCA.*

Entrepreneurial Orientation and Firm Performance in the Context of Upper Echelon Theory

Gali, Nazha[a]; Hughes, Mathew[a]; Mallet, Oliver[b]; Karam, Arze[c]
[a]Department of Management Entrepreneurship, Durham University, United Kingdom, [b]Department of Management, Newcastle University, United Kingdom, [c]Department of Economics and Finance, Durham University, United Kingdom

Abstract

Entrepreneurial Orientation (EO) is a firm-level phenomenon, which involves the firm's prospects to take risks, be proactive, and be innovative. Most of the research assumes a positive EO-performance relationship adopting the EO-as-advantage perspective without providing enough theoretical foundations of the way EO enhances performance. This paper provides insights into the EO and firm performance relationship looking into the EO-as-experimentation perspective. Through EO-as-experimentation perspective, we argue for the importance of looking into the differential effects of each of the EO dimensions on firm performance in active and inactive firms. We hypothesized that the effect of each of the proactiveness and innovativeness dimension of EO on firm performance is positive among active firms and negative among inactive firms. Whereas risk taking dimension of EO is negative among active and inactive firms. Based on the results of firm fixed effect regression some empirical support for the hypotheses is presented and discussed.

Keywords: *EO; risk taking; innovativeness; proactiveness; firm performance; EO-as-experimentation.*

Specific Characteristics of Actor Networks in the Social Innovation Process - A Comparative Analysis

Kleverbeck, Maria
Institute for Work and Technology, Westphalian University of Applied Science, Germany.

Abstract

The extended abstract presents the research which is the comparative study of the results obtained within the corporate governance efectiveness models. These are models, usually using the regression analysis, veryfing the correlation between the corporate governance standards and firm performance. The author's study shows that researchers often provide opposing results, that prevent precise determining, on a basis on quantitative evidence, the mentioned relationship. It also makes impossible to verify the rightness of a course of actions taken to rehabilitate and strengthen the corporate governance systems. The author's study presents these research differences and tries to explain them..

Keywords: *corporate governance; regression analysis; firm performance.*

A bibliometric overview of the Journal of Business Research

Merigó, José M.[a]; Mas-Tur, Alicia[b]; Roig-Tierno, Norat[c]; Ribeiro-Soriano, Domingo[b]
[a]Department of Management Control and Information Systems, University of Chile, Chile; [b]Department of Business Administration, University of Valencia, Spain; [c]ESIC Business and Marketing School, Valencia, Spain.

Abstract

The Journal of Business Research is a leading international journal in business research dating back to 1973. This study analyzes all the publications in the journal since its creation by using a bibliometric approach. The objective is to provide a complete overview of the main factors that affect the journal. This analysis includes key issues such as the publication and citation structure of the journal, the most cited articles, and the leading authors, institutions, and countries in the journal. Unsurprisingly, the USA is the leading region in the journal although a considerable dispersion exists, especially during the last years when European and Asian universities are taking a more significant position.

Keywords: *Business research; bibliometrics; Web of Science; journal analysis.*

Advanced Regression Methods

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3596

Structural Equation Modeling in Research on Brand Capital in Higher Education

Casanoves-Boix, Javier[a]; Küster-Boluda, Inés[b] and Vila-López, Natalia[c]
[a]Department of Marketing, ESIC Business and Marketing School, Spain, [b]Department of Marketing, Universitat de València, Spain, [c]Department of Marketing, Universitat de València, Spain.

Abstract

This research is done in order to examine the role of the brand capital in higher education. For this purpose, we analyze the main contributions of the literature related to the study of the brand capital and its application in the educational sector, identifying which variables determine the brand capital in the higher education sector. Once we establish the susceptible brand capital in the higher education sector, an empirical research is done by using a questionnaire developed in Spanish language with a Likert scale of grade 5 in which 1 point means "strongly disagree" and 5 mean "strongly agree", being based on the measurement scales proposed by Aaker (1992) and Keller (1993). Thus, we have a valid sample of 690 university professors (438 of them from public institutions and 356 from private ones). The results have been obtained by using a structural equation method, showing the relevance of each variable and determining those most discriminant for university professors.

Keywords*: Brand Capital, Brand Equity, Higher Education, Structural Equation Method, University Professors*

1. Introduction

Brands have evolved into a life experience for consumers, acquiring an emotional importance to them, that is reflected in their purchase satisfaction (Camacho, 2008). However, in the search for new forms of differentiation involving the actual creation of customer value, companies must be economically efficient (Kuster, and Aldás Vila, 2011). In the context discussed, brands can play an important role in the educational sector because in the last years universities have integrated marketing strategies and policies in its business model, both private and public (Fernández, 2002). So, there are some evidences that confirm that the marketing's theories and concepts can be applied in the educational context and, especially, in the field of higher education (Küster, 2012).

However, and as it is pointed out by the authors, the literature in this area is incoherent, still incipient, with a lack of theoretical models that reflect the particular context and the nature of the services of higher education. Because of that, in this research we contemplate two objectives, such as: (1) to analyze what it is understood as brand capital and what are its variables and determining components, talking about its application in the education sector; (2) to analyze what are the most relevant and discriminator elements of brand capital of the Higher Education Institutions, from the professors point of view.

In the same way, the elaboration of this research may be helpful for the university managers, given that they can figure out the key points in the opinions of several employees, with respect to the brand capital of their institutions and then to produce suitable strategies in order to maintain or improve the brand capital. So, with the purpose of reaching the objectives described above, we are dividing this research into two large parts: (1) review of the scientific literature regarding to our field of study, the brand capital and its key elements, as well as the exhaustive analysis of the marketing inside the education sector in Spain and (2) empirical research with university professors from Spanish universities, with the goal of decode those variables of the brand capital are more discriminator by using a structural equation method.

2. A brand capital model applied to higher education

After having revised the 7 main proposals about brand capital models done by literature and several contributions in the field of higher education, we present in Table 1 those elements shared by the authors in their models, whose cross elements have been considered key for our investigation, keeping in mind the importance of previous studies. Thus, all the authors consider four elements that, even though they are named in different ways, we understand that they have the same importance for the brand capital. These elements are: (1) brand awareness, (2) brand image, (3) perceived quality and (4) brand loyalty. These contributions, and more done by the literature, allow us to establish the following hypothesis, showed in Figure 1.

H1: The perception of brand awareness influences in the perception of brand capital among university professors.

H2: The perception of brand image influences in the perception of brand capital among university professors.

H3: The perception of perceived quality influences in the perception of brand capital among university professors.

H4: The perception of brand loyaly influences in the perception of brand capital among university professors.

Table 1. Main elements of brand capital in higher education

Elements of Brand Capital	Tot Brand Capital Models							Proposals of Brand Capital in Higher Education
	Farquhar (1989)	Aaker (1992)	Keller (1993)	Faircloth, Capella y Alford (2001)	Yoo y Donthu (2001)	Delgado y Munuera (2002)	Buil, Martínez y De Chernatony (2010)	
BRAND AWARENESS		X	X		X	X	X	BRAND AWARENESS
Brand Recall			X					Koku, 1997; Morphew, 2001; Sevier, 2001; Toma, Dubrow y Hartley, 2005; Brunzel, 2007; Furey, Springer y Parsons, 2009; Brewer y Zhao, 2010; Pinar, Trapp, Girard y Boyt, 2014
Brand Recognition			X					
Brand Prominence			X					
Perceived Risk Reduction		X						
BRAND IMAGE	X	X	X	X	X	X	X	BRAND IMAGE
Brand as a Company		X						Smith y Ennew, 2000; Bosch, Venter, Han y Boshoff, 2006; Jevons, 2006; Hamann, Williams y Omar, 2007; Hemsley-Brown y Goonawardana, 2007; Chen, 2008; Denegri, Cabezas, Herrera, Páez y Vargas, 2009; Williams y Omar, 2009; Waeraas y Solbakk, 2009; Whisman, 2009; Gómez y Medina, 2010; Stripling, 2010; Mourad, Ennew y Kortam, 2011; Williams, Williams, y Omar, 2013; Pinar, Trapp, Girard y Boyt, 2014
Brand Associations	X		X	X	X		X	
Brand Differentiation		X						
Brand Imagination			X					
Brand Performance			X					
Brand Personality		X						
Brand Reputation		X				X		
PERCEIVED QUALITY		X	X		X	X	X	PERCEIVED QUALITY
Brand Esteem		X						Kissman y Van Tran, 1990; Ramsden, 1991; Byron, 1995; Athiyaman, 1997; Vorhies, 1997; Booth, 1999; Binsardi y Ekwulugo, 2003; Mai, 2005; Peltier, Schibrowsky y Drago, 2007; Chen, 2008; Billings, Engelberg, Curtis, Block y Sullivan, 2010; Mourad, Ennew y Kortam, 2011; Pinar, Trapp, Girard y Boyt, 2014
Brand Feelings			X					
Brand Judgements			X					
Brand Leadership		X						
Brand Popularity		X						
Brand Reviews			X					
Perceived Value		X						
Willingness to Pay More			X					
BRAND LOYALTY	X	X			X	X	X	BRAND LOYALTY
Brand Resonance			X					Nicholls, Harris, Morgan, Clarke y Sims, 1995; Nguyen y LeBlanc, 2001; Lerman y Garbarino, 2002; Helgesen, 2008; Bok, 2009; Brown y Mazzarol, 2009; Pawsan y Ganesh, 2009; Rojas, Vasquez, Kara y Cerdá, 2009; Pinar, Trapp, Girard y Boyt, 2014
Brand Trust		X						
Customer Appreciation		X						
Customer Satisfaction	X	X						
Price Premium		X						

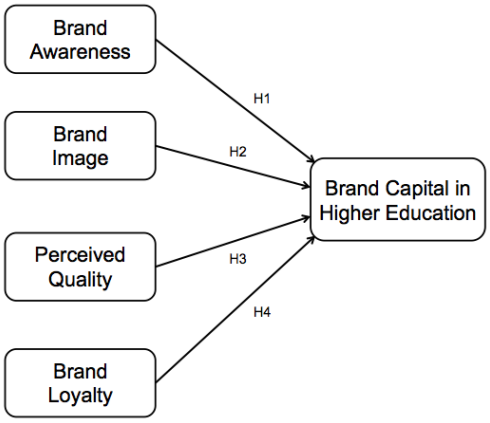

Figure 1. Theoretical model proposed for this research

3. Methodology

In order to corroborate the established hypothesis, we accomplished an empirical research quantitative in nature, by means of a survey in Spanish aimed at a group of 438 professors from the Faculty of Economics from two public universities, and at a group of 356 from six private ones in Valencia (Spain), obtaining an amount of 690 useful surveys for our investigation.

If we base on Churchill (1979) recommendations, and in the measuring scale suggested by Aaker (1992) and Keller (1993), we can detect four segments of measurement, which are: (1) brand awareness, (2) brand image, (3) perceived quality and (4) brand loyalty. All of them have been measured with the Likert scale of grade 5 (Bozal, 2005).

The technics for the data analysis are based on the Descriptive Statistics and the Multivariate Analysis, using as a support tool the SPSS v.18 for Windows for the data descriptive techniques and the EQS 6.2 for implementing the multivariate techniques. The statistical processing followed in this research for the data implies the implementation of analysis methods dependent on the information we need to obtain, making a difference between: (1) data description and classification, and (2) hypothesis testing.

4. Results

On one hand, the scales were measured by using a confirmatory factor analysis, where all loads and t robust values were significant at $p < 0.01$ value. The reliability was checked by three methods of analysis. First, using the Cronbach Alpha (CA) and obtaining values higher than 0.70 in all cases, which helped us to accept this rule (Nunnally and Bernstein, 1994). Second, doing an analysis of composite reliability (CR) and also obtaining values higher than 0.70 (Carmines and Zeller, 1979). Finally, in order to determine the reliability, we also did an analysis of the average variance extracted (AVE), where the results were close or above to 0.50 (Fornell and Larcker, 1981).

Table 2. Reliability and convergent validity of the scales

FACTOR	INDICATOR	LOAD	*t* ROBUST	CA	CR	AVE
Brand Awareness	BA1	0.52***	6.23	0.85	0.86	0.56
	BA2	0.81***	10.78			
	BA3	0.84***	9.99			
	BA4	0.82***	10.98			
	BA5	0.70***	8.66			
Brand Loyalty	BI1	0.63***	7.26	0.91	0.91	0.48
	BI2	0.77***	9.31			
	BI3	0.87***	13.91			
	BI4	0.75***	12.24			
	BI5	0.57***	7.14			
	BI6	0.75***	11.92			
	BI7	0.61***	8.10			
	BI8	0.41***	4.59			
	BI9	0.67***	8.71			
	BI10	0.77***	12.11			
	BI11	0.69***	8.22			
Perceived Quality	PQ1	0.77***	9.09	0.95	0.95	0.48
	PQ2	0.62***	9.79			
	PQ3	0.80***	12.74			
	PQ4	0.84***	12.34			
	PQ5	0.69***	9.31			
	PQ6	0.51***	5.70			
	PQ7	0.63***	8.35			
	PQ8	0.74***	10.70			
	PQ9	0.61***	8.75			
	PQ10	0.48***	5.61			
	PQ11	0.60***	8.04			
	PQ12	0.70***	9.58			
	PQ13	0.84***	11.94			
	PQ14	0.73***	11.47			
	PQ15	0.63***	8.05			
	PQ16	0.55***	6.77			
	PQ17	0.65***	9.66			
	PQ18	0.79***	12.96			
	PQ19	0.65***	9.27			
	PQ20	0.80***	10.70			
	PQ21	0.71***	8.08			
Brand Loyalty	BL1	0.82***	11.85	0.95	0.94	0.60
	BL2	0.87***	13.34			
	BL3	0.80***	13.80			
	BL4	0.86***	13.59			
	BL5	0.86***	15.86			
	BL6	0.63***	8.60			
	BL7	0.82***	11.43			
	BL8	0.50***	6.08			
	BL9	0.80***	12.50			
	BL10	0.81***	12.32			
	BL11	0.63***	9.67			
Brand Capital	BC1	0.41***	4.55	0.76	0.74	0.43
	BC2	0.71***	8.32			
	BC3	0.85***	12.79			
	BC4	0.59***	6.65			

On the other hand, the results suggest that the model designed in this research applied to the professors staff is satisfactory to explain the four hypotheses, obtaining a significant effect in the brand capital model proposed. Moreover, the goodness of fit statistics suggest that the structural model fits well with the data structure (S-B $\chi 2$ (p) = 5.407,353 (0,0000), df= 1.263; NFI = 0,78; NNFI = 0,81; CFI = 0,82; IFI = 0,82; RMSEA = 0,07). Moreover, we can observe in the Table 3 that the university professors give more importance to the perceived quality, followed by the brnd loyalty, the brand image and, finally, the brand awareness.

Table 3. Validation of the proposed model

Hypothesis	Structural Relationship	β Estand.	t Robust	Result
H1	Perception of Brand Awareness --> Perception of Brand Capital	0,065***	3,37	Accepted
H2	Perception of Brand Image --> Perception of Brand Capital	0,114***	4,14	Accepted
H3	Perception of Perceived Quality --> Perception of Brand Capital	0,545***	7,49	Accepted
H4	Perception of Brand Loyaly --> Perception of Brand Capital	0,337***	6,37	Accepted

N = 690; ***p<0,01; **p<0,05; * p<0,1

Satorra-Bentler χ^2 (p) = 5.407,35 (0.0000), *df* = 1.263; NFI = 0,78; NNFI = 0,82; CFI = 0,82 ; IFI = 0,82; RMSEA = 0,07

5. Conclusions

Firstly, the profile indicates that the sample is a mixed group of men and women from public and private university (very heterogeneous sample), 40 years middle age, with responses from 9 different nationalities (Spain is the country with the highest number of responses). Moreover, there are an average of 10 years work experience, being the majority of the sample full-time workers with an average monthly salary from 2,000 and 3,000 euros (considering that over 60% of staff is studying a PhD program).

On the other hand, and observing the results of hypothesis test, we can observe that the overall model proposed show a positive and direct relationship for the 4 hypotheses. So, there is a positive relationship between perceptions of brand awareness, brand image, perceived quality, brand loyalty and brand capital

References

Aaker, D. A. (1992). The value of brand equity. Journal of business strategy, 13(4), 27-32.

Athiyaman, A. (1997). Linking student satisfaction and service quality perceptions: the case of university education. European Journal of Marketing, 31(7), 528-540.

Bok, D. (2003). Universities in the marketplace: The commercialization of higher education. Princeton University Press.

Bozal, M.G. (2005): Escala mixta Likert-Thurstone. Anduli: revista andaluza de ciencias sociales, (5), 81-96.

Bosch, J., Venter, E., Han, Y., & Boshoff, C. (2006). The impact of brand identity on the perceived brand image of a merged higher education institution: Part one. Management Dynamics: Journal of the Southern African Institute for Management Scientists, 15(2), 10-30.

Brewer, A., & Zhao, J. (2010). The impact of a pathway college on reputation and brand awareness for its affiliated university in Sydney. International Journal of Educational Management, 24(1), 34-47.

Brown, R.M., & Mazzarol, T.W. (2009). The importance of institutional image to student satisfaction and loyalty within higher education. Higher Education, 58(1), 81-95.

Brunzel, D.L. (2007): Universities sell their brands, Journal of Product & Brand Management, 16(2), 152-3.

Buil, I., Martínez, E., & De Chernatony, L. (2010). Medición del valor de marca desde un enfoque formativo.

Camacho, J. (2008). El valor de la marca: Brand Equity. Datos, Diagnóstico y Tendencias. Nielsen. http://mx.nielsen.com/press.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Vol. 17. Sage publications.

Chen, L.H. (2008). Internationalization or international marketing? Two frameworks for understanding international students' choice of Canadian universities, Journal of Marketing for Higher Education, 18(1), 1-33.

Churchill, G.A. (1979). A Paradigm for Developing Better Measures of Marketing Constructs. Journal of Marketing Research. 16(1), 64-73.

Delgado-Ballester, E. & Munuera, J.L. (2002). Medición del capital de marca con indicadores formativos. Investigación y Markerting, 759, 16-20.

Denegri, M., Etchebarne, M.S., Geldres, V., Cabezas, D., & Herrera, V. (2009). Personalidad de marca de las carreras de ciencias empresariales: un análisis corporativo entre universidad pública y privada.

Faircloth, J.B., Capella, L.M., & Alford, B.L. (2001). The effect of brand attitude and brand image on brand equity. Journal of Marketing Theory and Practice, 61-75.

Farquhar, P.H. (1989). Managing brand equity. Marketing research, 1(3).

Fernández, C. (2002). Introducción al marketing para centros de enseñanza. ESIC Editorial, 2002.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. Journal of marketing research, pp. 382-388.

Furey, S., Springer, P. & Parsons, C. (2009). University Brand Promises, Presentation at Academy of Marketing 2009 Conference.

Gómez, D.F.H., & Medina, R.Z. (2010). Diagnóstico de la imagen de marca de las instituciones universitarias en España.

Hamann, D., Williams, R., & Omar, M. (2007). Branding Strategy and Consumer High-Technology Product, The Journal of Product & Brand Management, 16, (Winter/Spring), 2, 98–111.

Hemsley-Brown, J. & Goonawardana, S. (2007). Brand harmonization on the international higher education. Journal of Business Research, 60(9), 942-948.

Jevons, C. (2006). Universities: a prime example of branding gone wrong, Journal of Product & Brand Management, 15(7), 466-447.

Keller, K.L. (1993). Conceptualizing, measuring, and managing customer-based Brand equity, Journal of Marketing, 57, 1-22.

Kissman, K., & Van Tran, T. (1990). Perceived quality of field placement education among graduate social work students. Journal of Continuing Social Work Education, 5(2), 27-30.

Koku, P. (1997). What Is in a Name? The Impact of Strategic Name Change on Student Enrollment in Colleges and Universities, Journal of Marketing for Higher Education, 8(2), 53– 71.

Küster, I., Vila, N., & Aldás, J. (2011). Brand Equity Innovation: el uso de las nuevas tecnologías en el sector del vino para el incremento del valor de marca. Distribución y Consumo, 116, 67.

Küster, I. (2012). El Docente Universitario desde una perspectiva de mercado: Influencia en el rendimiento del estudiante. Alicante: 3 ciencias, 1-118.

Lerman, D. & Garbarino, E. (2002). Recall and Recognition of Brand Names: A Comparison of Word and Nonword Name Types, Psychology & Marketing, 19 (7/8), 621.

Morphew, C. (2001). A Rose by Any Other Name? Which Colleges Become Universities, The Review of Higher Education, 25 (2).

Mourad, M., Ennew, C., & Kortam, W. (2011). Brand equity in higher education.Marketing Intelligence & Planning, 29(4), 403-420.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory. Third Edition. McGraw-Hill. New York.

Paswan, A.K., & Ganesh, G. (2009). Higher education institutions: Satisfaction and loyalty among international students. Journal of Marketing for Higher Education, 19(1), 65-84.

Peltier, J.W., Schibrowsky, J.A., & Drago, W. (2007). The interdependence of the factors influencing the perceived quality of the online learning experience: A causal model. Journal of Marketing Education, 29(2), 140-153.

Pinar, M., Trapp, P., Girard, T., & Boyt, A. T. (2014). University Brand Equity: An Empirical Investigation of its Dimensions, Staying Current with Media & Millennials, 50.

Rojas-Méndez, J.I., Vasquez-Parraga, A.Z., Kara, A.L.I., & Cerda-Urrutia, A. (2009). Determinants of student loyalty in higher education: A tested relationship approach in Latin America. Latin American Business Review, 10(1), 21-39.

Smith, R. & Ennew, C. (2000). Service quality and its impact on word of mouth communication in higher education, paper presented at the Academy of Marketing Annual Conference, University of Derby, Derby, 5-7 July.

Stripling, J. (2010). Brand new dilemma, available at: http://insidehighered.com/news/2010/10/ 19/branding (accessed June 27, 2013).

Toma, J.D., Dubrow, G., & Hartley, M. (2005). The uses of institutional culture: Strengthening identification and building brand equity in higher education. Wiley Periodicals, Inc.

Waeraas, A., & Solbakk, M. (2009): Defining the essence of a university: lessons from higher education branding, Higher Education, 57(4), 449-462.

Whisman, R. (2009). Internal branding: a university's most valuable intangible asset. Journal of Product & Brand Management, 18(5), 367-370.

Williams, R.L., & Omar, M. (2009). Renaming Service Organizations for Growth, Presentation at Academy of Marketing 5th International Colloquium: Brand, Identity and Corporate Reputation, University of Cambridge, U.K.

Williams, R.L., Williams, H.A., & Omar, M. (2013). The Marketing Impact of the Principles of Renaming Within a Higher Education Service Organization, American Marketing Association.

Yoo, B., & Donthu, N. (2001). Developing and validating a multidimensional consumer-based brand equity scale. Journal of business research, 52(1), 1-14.

First International Conference on Advanced Research Methods and Analytics, CARMA2016
Universitat Politècnica de València, València, 2016
DOI: http://dx.doi.org/10.4995/CARMA2016.2016.3114

Scale of Attitudes Towards ICT (SATICT): Factor Structure and Factorial Invariance in Distance University Students

Ordóñez, Xavier Giovanni[a] and Romero, Sonia Janeth[b]
[a]Department of Research Methods and Diagnosis in Education, Complutense University of Madrid, Spain. [b]Department of Health and Education Sciences, Madrid Open University, Spain.

Abstract

Attitudes towards Information and Communication Technology (ICT) are preconceived beliefs about influence of the ICT tools in the process of learning. Studies carried out in this area shows that attitudes may influence cognitive and learning processes and also motivation of students. Despite its importance, very few instruments have been proposed to measure student attitudes, and none analyzes the factor structure or the factorial invariance of the scores, for that reason, the aim of the present study is to analyze the factor structure of SATICT, a new instrument proposed to measure the attitudes towards ICT on distance university students. A second aim is to test the factorial invariance across gender and educational level in a sample of 1080 university students of Madrid Open University using multi-group CFA. The results provide high support of the proposed factor structure with significant loadings and adequate model fit, however, the results also showed that factor structure couldn't be considered invariant across groups.

Keywords: *Factor structure; Factorial Invariance; Attitudes, Information and Communication Technology, Distance University Students.*

1. Introduction

At present, the Information and Communications Technology (ICT) are omnipresent in the teaching-learning processes, specially in distance education (Harasim, 1990; Mclsaac & Gunawardena, 1996). The appropriate use of technology in education is conditioned not only by the technological knowledge about the tools, but also by their attitudes. Is for that reason that the goal of the present study is to propose a new instrument to measure the attitudes towards incorporation of ICT in the learning process of distance university students called SATICT (Scale of Attitudes Towards ICT). Secondary goals were to study the factor structure and factorial invariance of the test scores across gender and educational level. To reach this goals the test SATICT were applied to a sample of 1080 university students of Madrid Open University, then, confirmatory factor analysis and multi-group analysis were made in order to meet the secondary objectives. This paper was structured as follows: first, a brief theoretical framework was presented, second, methodology aspects, including sample and instrument description and data analysis procedures were exposed, third, the main results are shown and finally, a brief discussion of the implications, limitations and future research was made.

2. Theoretical Framework

Several studies of national and international levels has been developed in the last years with the aim to know the attitudes of students towards ICT and to elaborate instruments (mainly surveys) to measure this construct. For example, Ahamed and Adulaziz (2004) examining the relationship between student performance and attitudes towards ICT in a virtual and conventional settings. Their findings revealed no significant differences in attitudes between virtual and conventional students. Francovicová and Prokop (2008) found in a sample of 214 elementary school students that attitudes toward ICT were positive and gender differences are weak. Siragusa and Dixon (2008) made a pilot study with 30 undergraduate students in higher education and their results showed some experienced feeling of anxiety and intimidation when students work with ICT.

Kubiatko (2010) focuses on differences of attitudes towards ICT among 316 Czech university students, he found that male, sophomores, and students living in town showed more positive attitudes in comparison to other groups. Kar, Saha and Mondal (2014) measured the attitude towards e-learning in a sample of 308 university students of four universities in West Bengal. Their results showed that students have high attitude scores and this scores did not differ significantly with their personal variables as gender, stream of study and residence.

Garcia, Escofet and Gros (2009), compared the attitudes towards ICT in 1042 students of blended and virtual universities in Spain. Their results showed statistically significant differences among students in virtual and face-to-face mode in all aspects of attitude.

Edmuns, Thorpe and Conole (2012) studied the influence of work, social context and course study on attitudes towards ICT in a sample of 421 students of an open university in the UK, their results showed that usefulness and ease of use are important variables that influence the attitudes. Rhema and Miliszewska (2014) studied the relationships between demographic characteristics and attitudes towards e-learning, they found no differences between men and women students or between urban and regional students with respect to their attitude towards ICT in a sample of 348 Libian's university students.

A common feature of the above studies is the use of surveys created at hoc to measure attitudes of students, however, the psychometric properties of those instruments are not explored or reported in the published research. On the other hand, although there are some recent studies that analyses different aspects of attitudes toward ICT in university students none of them uses factor analysis techniques to explore or confirm the underlying factors of the instruments used. Others studies are focused on the differences in the attitudes according to interest groups as gender, education level or education settings (virtual vs conventional), however, none of them check the factorial invariance, a step that is essential before any comparison between groups. The results of the present study are an initial approach to this topic.

3. Methodology

3.1 Participants

A total of 1231 students participated voluntarily (with informed consent) in this study, 600 females and 631 males. The instrument was sent to all the students enrolled in the course 2014-2015 both, in undergraduate and master studies (N = 5776), therefore, response rate was 21,31%. All students were recruited from Madrid Open University (MOU) in Spain. 63,44% of the sample are undergraduate and 36,56% are master students. 40,76% of the students works in ICT related areas and 57,66% has completed undergraduate studies previously. All the participants are between 18 and 69 years old (mean= 36,01, SD=9,59). 149 responses were discarded for being incomplete.

3.2 Instruments

The Scale of Attitudes Towards ICT (SATICT) is an instrument composed by 20 items distributed on three scales: affective (8 items, 2 inversely scored), behavioral (5 items, 1 inversely scored) and cognitive (7 items, 2 inversely scored). The format used is 5-point Likert type from 1 (totally disagree) to 5 (totally agree). High test scores indicates positive attitude towards the incorporation of ICT in the learning process. Descriptive results of the total scores and sub-scales are, for the total score: *min*=22; *max*=100; *m*=83,22; *s.d.*=12,24. Cronbach's Alpha (α=,903) indicating high overall reliability of the test scores. For the affective scale *min*=9; *max*=40; *m*=31,55; *s.d.*=5,44; α=,735. For the behavioral scale

min=5; *max*=25; *m*=20,57; *s.d.*=3,63; α=,767. For the cognitive scale *min*=7; *max*=35; *m*=31,09; *s.d.*=4,21; α=,801.

Previous evidence of face and content validity of the test was gathered by percentages of congruence item-factor. The questionnaire was reviewed by a group of 5 experts in educational technology, they were informed of the definition of the three factors that compose the instrument (affective, cognitive and behavioral) and they were asked to pair up each factor with the item they considered it measured. The percentage of items that is matched correctly with the factors vary from 45,8 to 83: rater 1 (11/24 = 45,8%); rater 2 (15/24 = 62,5%); rater 3 (20/24 = 83%); rater 4 (12/24 = 50%) and rater 5 (17/24 = 70.8%) indicating only partial evidence of content validity. Complete test (in Spanish Language) can be requested to the first author of the present paper.

3.3 Data analysis

Data analysis includes Confirmatory Factor Analysis (CFA) and multi-group analysis (MACs). As input for the CFA the asymptotic covariance matrix was used. Analysis was made with the WLS method. To study the fit of the model several fit indices were considered, following the proposed by Brown and Moore (2014): Satorra-Bentler Scaled χ^2, RMSEA (Root Mean Square Error of Approximation), 90% confidence interval of RMSEA, Comparative Fit Index (CFI) and Non-Normed Fit Index (NNFI). Values recommended as acceptable are: RMSEA under 0,05, CFI and NNFI over 0,95.

The factorial invariance across groups formed by gender (female-male) and educational level (undergraduate-master) was studied following the approach MACs (Little, 1997; Cheung & Rensvold, 2002) that comparatively evaluates the fit of four nested models (configural invariance, weak measurement invariance, strong and strict invariance models). Following the criteria proposed by Cheung & Rensvold, 2002, the change in χ^2 ($\Delta\chi^2$) was analyzed: if there is a significantly increase in the value of this statistic is indicative of lack of factorial invariance.

4. Results

4.1 Confirmatory Factor Analysis

A CFA with the whole sample was made in order to confirm the proposed three-dimensional structure for the SATICT. A correlation between errors of items 7 and 8 was suggested by the modification indices to improve the fit of the model. Tables 1 and 2 exhibit the factorial loadings and fit indices, respectively.

Table 1. Factorial loadings.

Item	A	B	C
1	-	-	.96*
2	-	.97*	-
3	-	-	.98*
4	-	-	.98*
5	.48*	-	-
6	1*	-	-
7	.66*	-	-
8	-	.73*	-
9	.64*	-	-
10	-	-	.77*
11	.75*	-	-
12	-	.96*	-
13	.84*	-	-
14	-	-	.93*
15	.94*	-	-
16	-	.90*	-
17	-	.93*	-
18	.71*	-	-
19	-	-	.81*
20	-	-	.84*

* Sign 1% and 5%.

Table 2. Fit indices.

χ^2	*Df*	*P*	CFI	NNFI	RMSEA	90% CI
693,83	166	,00	,97	,97	,054	,050-,059

4. 2 *Factorial invariance according to gender and educational level*

One objective of the present paper has been to analyze whether the measurement of attitudes was equivalent according to gender and educational level. For this purpose, as first step, the model was applied for each sample independently. Then, the configural model (without constraints) was estimated to demonstrate that the pattern of fixed and free parameters was equivalent across subsamples. As the configural model do not fit well it was impossible to introduce more constraints to the model. The fit indices of different models are summarized in the Table 3.

Table 3. Fit indices. Comparison of models to analyze factorial invariance by gender and educational level

Model	χ^2	df	*p*	CFI	NNFI	RMSEA	90% CI
Females	508,29	165	,00	,99	,98	,063	,057-,069
Males	501,45	165	,00	,98	,98	,061	,055-,067
Configural	5015,72	336	,00	,91	,90	,18	,17-,18
Undergraduate	552,90	165	,00	,98	,98	,059	,053-,064
Master	487,36	165	,00	,98	,98	,070	,063-,078
Configural	5466,21	336	,00	,90	,89	,19	,18-0,19

Table 3 showed that configural invariance is not achieved, indicating that the model do not fit across the groups. Although the model is the same across groups, the unknown parameters of the model are assumed to be different across the groups, and this assumption was not met in this case. The (global) Chi-square test statistic, CFI, NNFI and RMSEA

values for this multiple group model showed that configural invariance was not met and for that reason not make sense to introduce more restrictions to the model.

5. Conclusions

This study was made to examine the factorial structure of the SATICT test, a measure of the attitudes towards ICT in online/virtual education settings. Although several studies have used different surveys to evaluate the attitudes towards ICT, the present study deepens on the factorial structure of a new test instrument composed by three sub-scales (dimensions) that frequently compose the attitude (affective, cognitive and behavioral). Additionally, is the only test in Spanish currently available to measure the attitudes towards ICT for distance university students. The results of CFA supports the proposed three-dimensional structure, with an excellent model fit and significant factor loadings, however, the results also showed that configural invariance can not be confirmed on the groups formed by gender and educational level indicating that the pattern of fixed and free parameters was not equivalent across subsamples of interest in this sample.

This study also present some limitations: first, even though the sample is large it not has been taken probabilistically, which limits the generalizability of results, second, the paper focuses only on the analysis of the factor structure, leaving aside the analysis of other psuchometric properties. For that reason, future research may be focused in other psychometric properties as concurrent or predictive validity by taking a probabilistic sample that represents students of different online universities.

References

Ahmed, A., & Abdulaziz, E. (2004), Examining students performance and attitudes towards the use of information technology in a virtual and conventional setting. *Journal of Interactive Online Learning*, *2*(3), 1-9.

Brown, T. A., & Moore, M. T. (2014). Conformatory Factor Analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361-379). New York: Guilford.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing MI. *Structural Equation Modeling, 9*, 235-55. doi: 10.1207/S15328007SEM0902_5

Edmunds, R., Thorpe, M., & Conole, G. (2012). Student attitudes towards and use of ICT in course study, work and social activity: a technology acceptance model approach. *British Journal of Educational Technology, 43*(1), 71-84. doi: 10.1111/j.1467-8535.2010.01142.x

Fancovicová, J., & Prokop, P. (2008). Students' attitude toward computer use in Slovakia. *Eurasia Journal of Mathematics, Science and Technology Education, 4*(3), 255-262.

García, I., Escofet, A., & Grof, B. (2009). Students`attitude towards ICT learning uses: a comparison between digital learners in blended and virtual universities. *European Journal of Open, Distance and E-Learning,* Retrieved from: http://www.eurodl.org/?p=special&sp=articles&inum=5&article=624

Harasim, L. (1990). Online education: An environment for collaboration and intelectual amplification. In: L. Harasim (Ed): Online education. *Perspectives on a New Environment* (p. 33-66). Preager: New York.

Kar, D., Saha, B., & Mondal, B. C. (2014). Attitude of university students towards E-Learning in West Bengal. *American Journal of Educational Research, 2*(8), 669-673. doi: 10.12691/education-2-8-16

Kubiatko, M. (2010). Czech university students`attitudes towards ICT used in science education. *Journal of Technology and Information Education, 2*(3), 20-25. doi: 10.5507/jtie.2010.042

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76. doi: 10.1207/s15327906mbr3201_3

Mclsaac, M. S., & Gunawardena, C. N. (1996). Distance Education. In D. H. Jonassen (Ed) *Handbook on Research for Education Communications and Technology.* McMillan: New York.

Rhema, A., & Miliszewska, I. (2014). Analysis of student attitudes towards E-Learning: The case of engineering students in Lybia. *Issues in Informing Science and Information Technology, 11*. 169-190.

Siragusa, L., & Dixon, K. (2008). Planned behaviour: Student attitudes towards the use of ICT interactions in higher education. In: *Hello¡ Where are you in the landscape of educational technology? Proceeedings ascilite*, Melbourne, Australia.

The Effects of Cognitive Appraisal and Emotion on Consumer Behavior: The Critical Role of Recollection in the Luxury Cruise Setting

Joo, Eunkyung[a]; Shin, Hyejin[a]; Kim, Insin[a]; Choi, Jinsung[b]; Jang, Junhwa[b]; Hyun, Sunghyup Sean[b]

[a]Department of Tourism and Convention, Pusan National Univ, Republic of Korea; [b]School of Tourism, Hanyang Univ., Republic of Korea

Abstract

The purposes of this study were: (1) to integrate the cognitive appraisal theory and script theory; (2) to examine the bonding character of recollection; and (3) to assess the relationships between consumers 'appraisals, positive/negative emotions, recollection, storytelling and repurchase intention. A review of previous studies revealed 14 theoretical hypotheses. The proposed hypotheses were tested utilizing data collected from 300 luxury cruise passengers. Confirmatory factor analysis and structural equation modeling were utilized to test the proposed theoretical relationships. According to the results, this work was the first to integrate the cognitive appraisal approach and script theory and also depicted a new angle from which marketers can better understand cruise travelers' behavior.

Keywords: *Congnitive appraisal theory, script theory, emotions, recollection, luxury cruise.*

Using Eye Tracking and Electroencephalograph to Understand the Efficacy of Digital and Static Outdoor Advertisements

Reas, Brandon; Dishman, Paul; McCarter, Aaron; Jolley, Alan Dale
Vivint Neuromarketing SMARTLab, Utah Valley University, USA.

Abstract

This study discusses the contributions of advanced eye tracking research combined with electroencephalography (EEG) as a method of understanding the cognitive processing of digital vs. static outdoor advertisements. Subjects were exposed to a variety of billboard advertisements on a section of Interstate freeway in a suburban area in the western United States. Results showed that visual fixation time was higher for digital advertisements compared to static advertisements. In particular, the eye-tracking data revealed which advertisements received the most attention. This was mainly dependent upon location (i.e. distance from driver, distance from adjacent traffic signs, etc.).

As eye-tracking systems have become more sophisticated and affordable, there has been an increasing interest in the use of eye tracking within the traffic safety and outdoor advertising domain (Perez & Bertola, 2010). Eye tracking studies that have focused on web-based and driving stimuli have gathered eye-movement data while participants were engaged in low-attention settings (Lee and Ahn, 2012). The findings of these studies have indicated that digital and animated advertisements, in low attention settings, reduce the likelihood of mental recall and result in overall decreased cognitive engagement.

Twenty-five subjects between the ages of 18 and 45 participated in the study. A 16 mile (25.75 kilometer) freeway drive was videotaped and then projected onto a four-by-six-foot screen. Subjects then viewed the projected video from inside a stationary car to simulate a driving environment. Using Tobii2 Glasses eye tracking system, subjects' eye-movements and gaze patterns were recorded during the simulation. EEG data was also collected to measure the subject's emotional response, and to gain additional insight into how they felt about the advertisements. In addition, participants were asked immediately after their drive to list any advertisements they recalled.

These findings provide indications for best practices of effective outdoor advertising using gaze pattern analysis. These include positioning, layout, color schemes, etc. A potential implication for digital advertisements could be identifying the optimal length of time to display digital signage. Additionally, the results may suggest improvements in specific industry ads in order to maximize cognitive influence on consumer action (e.g. best times to display food and beverage advertisements).

From a traffic safety consideration, these results will provide a psychological understanding of whether or not outdoor advertisements present safety implications to drivers. Overall, findings provide a better understanding of digital and static outdoor advertising as it relates to safety and consumer behavior. The results of this study may have significant implications in both the private and public sectors.

Keywords: *Outdoor Advertising; Eye tracking; Electroencephography; Billboards.*
